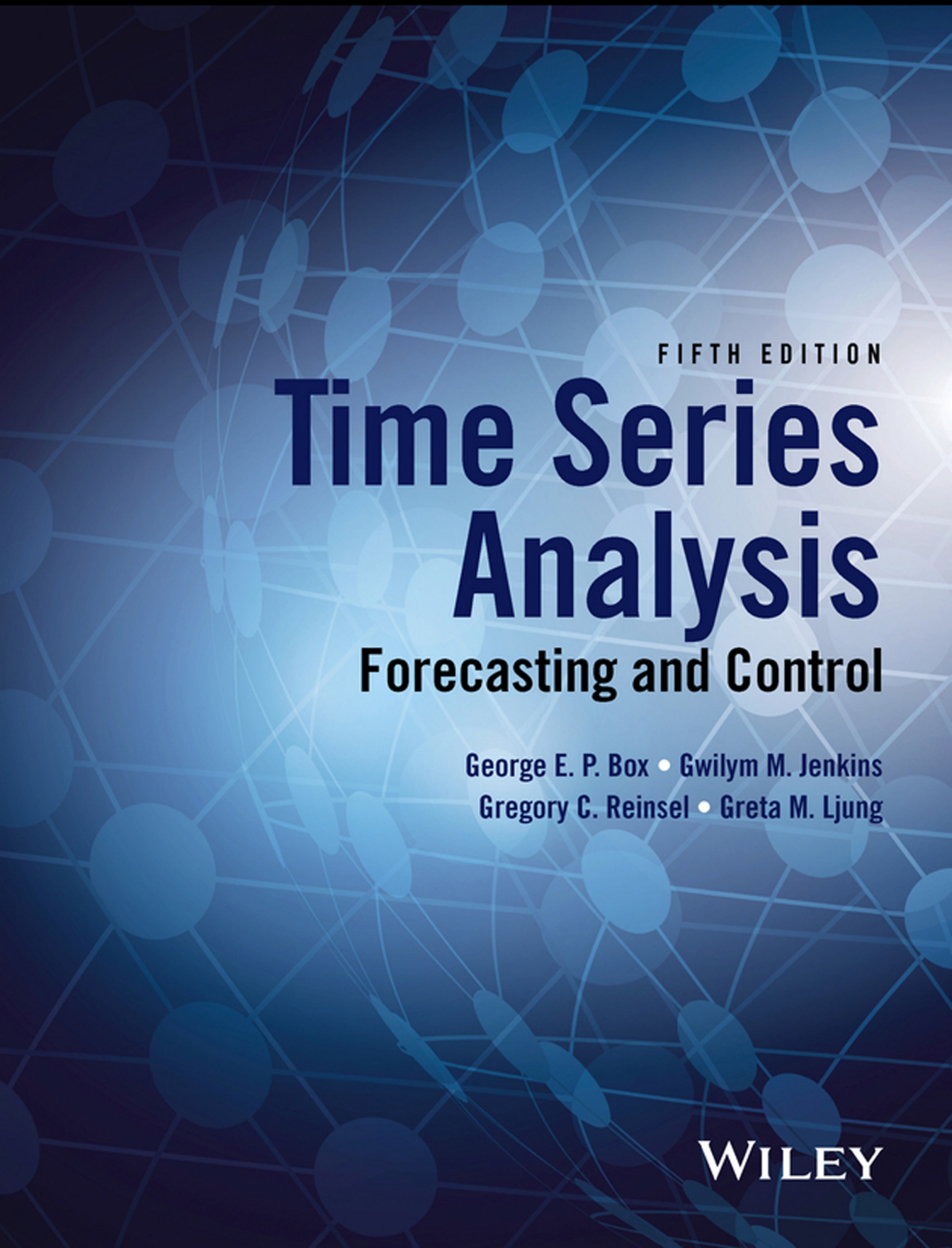
WILEY SERIES IN PROBABILITY AND STATISTICS



FIFTH EDITION

Time Series Analysis

Forecasting and Control

George E. P. Box • Gwilym M. Jenkins
Gregory C. Reinsel • Greta M. Ljung

WILEY

TIME SERIES ANALYSIS

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Geof H. Givens, Harvey Goldstein, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*

Editors Emeriti: *J. Stuart Hunter, Iain M. Johnstone, Joseph B. Kadane, Jozef L. Teugels*

A complete list of the titles in this series appears at the end of this volume.

TIME SERIES ANALYSIS

Forecasting and Control

Fifth Edition

GEORGE E. P. BOX
GWILYM M. JENKINS
GREGORY C. REINSEL
GRETA M. LJUNG

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

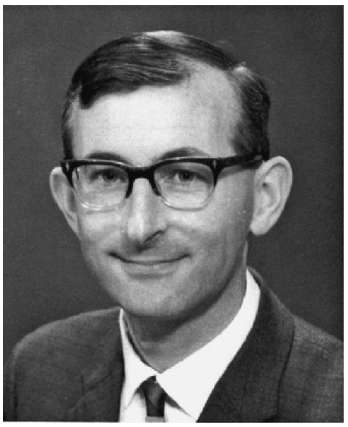
Box, George E. P.
Time series analysis : forecasting and control. – Fifth edition / George E.P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Greta M. Ljung.
pages cm
Includes bibliographical references and index.
ISBN 978-1-118-67502-1 (cloth : alk. paper) 1. Time-series analysis. 2. Prediction theory. 3. Transfer functions. 4. Feedback control systems–Mathematical models. I. Jenkins, Gwilym M. II. Reinsel, Gregory C. III. Ljung, Greta M., 1941- IV. Title.
QA280.B67 2016
519.5'5–dc23
2015015492

Printed in the United States of America

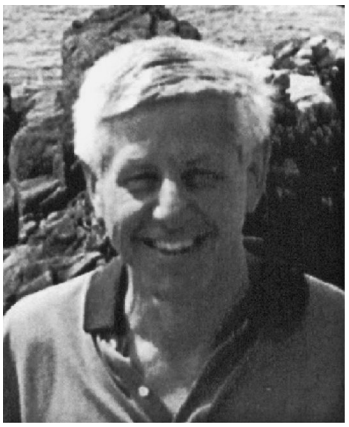
10 9 8 7 6 5 4 3 2 1

To the memory of

George E. P. Box



Gwilym M. Jenkins



Gregory C. Reinsel

CONTENTS

PREFACE TO THE FIFTH EDITION

This book describes statistical models and methods for analyzing discrete time series and presents important applications of the methodology. The models considered include the class of autoregressive integrated moving average (ARIMA) models and various extensions of these models. The properties of the models are examined and statistical methods for model specification, parameter estimation, and model checking are presented. Applications to forecasting nonseasonal as well as seasonal time series are described. Extensions of the methodology to transfer function modeling of dynamic relationships between two or more time series, modeling the effects of intervention events, multivariate time series modeling, and process control are discussed. Topics such as state-space and structural modeling, nonlinear models, long-memory models, and conditionally heteroscedastic models are also covered. The goal has been to provide a text that is practical and of value to both academicians and practitioners.

The first edition of this book appeared in 1970 and around that time there was a great upsurge in research on time series analysis and forecasting. This generated a large influx of new ideas, modifications, and improvements by many authors. For example, several new research directions began to emerge in econometrics around that time, leading to what is now known as time series econometrics. Many of these developments were reflected in the fourth edition of this book and have been further elaborated upon in this new edition.

The main goals of preparing a new edition have been to expand and update earlier material, incorporate new literature, enhance and update numerical illustrations through the use of R, and increase the number of exercises in the book. Some of the chapters in the previous edition have been reorganized. For example, Chapter 14 on multivariate time series analysis has been reorganized and expanded, placing more emphasis on vector autoregressive (VAR) models. The VAR models are by far the most widely used multivariate time series models in applied work. This edition provides an expanded treatment of these models that includes software demonstrations.

Chapter 10 has also been expanded and updated. This chapter covers selected topics in time series analysis that either extend or supplement material discussed in earlier chapters.

This includes unit roots testing, modeling of conditional heteroscedasticity, nonlinear models, and long memory models. A section of unit root testing that appeared in Chapter 7 of the previous edition has been expanded and moved to Section 10.1 in this edition. Section 10.2 deals with autoregressive conditionally heteroscedastic models, such as the ARCH and GARCH models. These models focus on the variability in a time series and are useful for modeling the volatility or variability in economic and financial series, in particular. The treatment of the ARCH and GARCH models has been expanded and several extensions have been added.

Elsewhere in the text, the exposition has been enhanced by revising, modifying, and omitting text as appropriate. Several tables have either been edited or replaced by graphs to make the presentation more effective. The number of exercises has been increased throughout the text and they now appear at the end of each chapter.

A further enhancement to this edition is the use of the statistical software R for model building and forecasting. The R package is available as a free download from the R Project for Statistical Computing at www.r-project.org. A brief description of the software is given in Appendix A1.1 of Chapter 1. Graphs generated using R now appear in many of the chapters along with R code that will help the reader reconstruct the graphs. The software is also used for numerical illustration in many of the examples in the text.

The fourth edition of this book was published by Wiley in 2008. Plans for a new edition began during the fall of 2012. I was deeply honored when George Box asked me to help him with this update. George was my Ph.D. advisor at the University of Wisconsin-Madison and remained a dear friend to me over the years as he did to all his students. Sadly, he was rather ill when the plans for this new edition were finalized towards the end of 2012. He did not have a chance to see the project completed as he passed away in March of 2013. I am deeply grateful for the opportunity to work with him and for the confidence he showed in assigning me this task. The book is dedicated to his memory and to the memory of his distinguished co-authors Gwilym Jenkins and Gregory Reinsel. Their contributions were many and they are all missed.

I also want to express my gratitude to several friends and colleagues in the time series community who have read the manuscript and provided helpful comments and suggestions. These include Ruey Tsay, William Wei, Sung Ahn, and Raja Velu who have read Chapter 14 on multivariate time series analysis, and David Dickey, Johannes Ledolter, Timo Teräsvirta, and Niels Haldrup who have read Chapter 10 on special topics. Their constructive comments and suggestions are much appreciated. Assistance and support from Paul Lindholm in Finland is also gratefully acknowledged. The use of R in this edition includes packages developed for existing books on time series analysis such as Cryer and Chan (2010), Shumway and Stoffer (2011), and Tsay (2014). We commend these authors for making their code and datasets available for public use through the R Project.

Research for the original version of this book was supported by the Air Force Office of Scientific Research and by the British Science Research Council. Research incorporated in the third edition was partially supported by the Alfred P. Sloan Foundation and by the National Aeronautics and Space Administration. Permission to reprint selected tables from *Biometrika Tables for Statisticians*, Vol. 1, edited by E. S. Pearson and H. O. Hartley is also acknowledged. On behalf of my co-authors, I would like to thank George Tiao, David Mayne, David Pierce, Granville Tunnicliffe Wilson, Donald Watts, John Hampton, Elaine Hodkinson, Patricia Blant, Dean Wichern, David Bacon, Paul Newbold, Hiro Kanemasu, Larry Haugh, John MacGregor, Bovas Abraham, Johannes Ledolter, Gina Chen, Raja Velu, Sung Ahn, Michael Wincek, Carole Leigh, Mary Esser, Sandy Reinsel, and

Meg Jenkins, for their help, in many different ways, in preparing the earlier editions. A very special thanks is extended to Claire Box for her long-time help and support.

The guidance and editorial support of Jon Gurstelle and Sari Friedman at Wiley is gratefully acknowledged. We also thank Stephen Quigley for his help in setting up the project, and Katrina Maceda and Shikha Pahuja for their help with the production.

Finally, I want to express my gratitude to my husband Bert Beander for his encouragement and support during the preparation of this revision.

GRETA M. LJUNG

Lexington, MA
May 2015

PREFACE TO THE FOURTH EDITION

It may be of interest to briefly recount how this book came to be written. Gwilym Jenkins and I first became friends in the late 1950s. We were intrigued by an idea that a chemical reactor could be designed that optimized itself automatically and could follow a moving maximum. We both believed that many advances in statistical theory came about as a result of interaction with researchers who were working on real scientific problems. Helping to design and build such a reactor would present an opportunity to further demonstrate this concept.

When Gwilym Jenkins came to visit Madison for a year, we discussed the idea with the famous chemical engineer Olaf Hougen, then in his eighties. He was enthusiastic and suggested that we form a small team in a joint project to build such a system. The National Science Foundation later supported this project. It took 3 years, but suffice it to say, that after many experiments, several setbacks, and some successes the reactor was built and it worked.

As expected, this investigation taught us a lot. In particular, we acquired proficiency in the manipulation of difference equations that were needed to characterize the dynamics of the system. It also gave us a better understanding of nonstationary time series required for realistic modeling of system noise. This was a happy time. We were doing what we most enjoyed doing: interacting with experimenters in the evolution of ideas and the solution of real problems, with real apparatus and real data.

Later there was fallout in other contexts, for example, advances in time series analysis, in forecasting for business and economics, and also developments in statistical process control (SPC) using some notions learned from the engineers.

Originally Gwilym came for a year. After that I spent each summer with him in England at his home in Lancaster. For the rest of the year, we corresponded using small reel-to-reel tape recorders. We wrote a number of technical reports and published some papers but eventually realized we needed a book. The first two editions of this book were written during a period in which Gwilym was, with extraordinary courage, fighting a debilitating illness to which he succumbed sometime after the book had been completed.

Later Gregory Reinsel, who had profound knowledge of the subject, helped to complete the third edition. Also in this fourth edition, produced after his untimely death, the new material is almost entirely his. In addition to a complete revision and updating, this fourth edition resulted in two new chapters: Chapter 10 on nonlinear and long memory models and Chapter 12 on multivariate time series.

This book should be regarded as a tribute to Gwilym and Gregory.

I was especially blessed to work with two such gifted colleagues.

GEORGE E. P. BOX

Madison, Wisconsin
March 2008

PREFACE TO THE THIRD EDITION

This book is concerned with the building of stochastic (statistical) models for time series and their use in important areas of application. This includes the topics of forecasting, model specification, estimation, and checking, transfer function modeling of dynamic relationships, modeling the effects of intervention events, and process control. Coincident with the first publication of *Time Series Analysis: Forecasting and Control,* there was a great upsurge in research in these topics. Thus, while the fundamental principles of the kind of time series analysis presented in that edition have remained the same, there has been a great influx of new ideas, modifications, and improvements provided by many authors.

The earlier editions of this book were written during a period in which Gwilym Jenkins was, with extraordinary courage, fighting a slowly debilitating illness. In the present revision, dedicated to his memory, we have preserved the general structure of the original book while revising, modifying, and omitting text where appropriate. In particular, Chapter 7 on estimation of ARMA models has been considerably modified. In addition, we have introduced entirely new sections on some important topics that have evolved since the first edition. These include presentations on various more recently developed methods for model specification, such as canonical correlation analysis and the use of model selection criteria, results on testing for unit root nonstationarity in ARIMA processes, the state-space representation of ARMA models and its use for likelihood estimation and forecasting, score tests for model checking, structural components, and deterministic components in time series models and their estimation based on regression-time series model methods. A new chapter (12) has been developed on the important topic of *intervention and outlier* analysis, reflecting the substantial interest and research in this topic since the earlier editions.

Over the last few years, the new emphasis on industrial quality improvement has strongly focused attention on the role of control both in process *monitoring* and in process *adjustment.* The control section of this book has, therefore, been completely rewritten to serve as an introduction to these important topics and to provide a better understanding of their relationship.

The objective of this book is to provide practical techniques that will be available to most of the wide audience who could benefit from their use. While we have tried to remove the inadequacies of earlier editions, we have not attempted to produce here a rigorous mathematical treatment of the subject.

We wish to acknowledge our indebtedness to Meg (Margaret) Jenkins and to our wives, Claire and Sandy, for their continuing support and assistance throughout the long period of preparation of this revision.

Research on which the original book was based was supported by the Air Force Office of Scientific Research and by the British Science Research Council. Research incorporated in the third edition was partially supported by the Alfred P. Sloan Foundation and by the National Aeronautics and Space Administration. We are grateful to Professor E. S. Pearson and the Biometrika Trustees for permission to reprint condensed and adapted forms of Tables 1, 8, and 12 of *Biometrika Tables for Statisticians*, Vol. 1, edited by E. S. Pearson and H. O. Hartley, to Dr. Casimer Stralkowski for permission to reproduce and adapt three figures from his doctoral thesis, and to George Tiao, David Mayne, Emanuel Parzen, David Pierce, Granville Wilson, Donald Watts, John Hampton, Elaine Hodkinson, Patricia Blant, Dean Wichern, David Bacon, Paul Newbold, Hiro Kanemasu, Larry Haugh, John MacGregor, Bovas Abraham, Gina Chen, Johannes Ledolter, Greta Ljung, Carole Leigh, Mary Esser, and Meg Jenkins for their help, in many different ways, in preparing the earlier editions.

GEORGE BOX AND GREGORY REINSEL

1

INTRODUCTION

A *time series* is a sequence of observations taken sequentially in time. Many sets of data appear as time series: a monthly sequence of the quantity of goods shipped from a factory, a weekly series of the number of road accidents, daily rainfall amounts, hourly observations made on the yield of a chemical process, and so on. Examples of time series abound in such fields as economics, business, engineering, the natural sciences (especially geophysics and meteorology), and the social sciences. Examples of data of the kind that we will be concerned with are displayed as time series plots in Figures 2.1 and 4.1. An intrinsic feature of a time series is that, typically, adjacent observations are *dependent*. The nature of this dependence among observations of a time series is of considerable practical interest. *Time series analysis* is concerned with techniques for the analysis of this dependence. This requires the development of stochastic and dynamic models for time series data and the use of such models in important areas of application.

In the subsequent chapters of this book, we present methods for building, identifying, fitting, and checking models for time series and dynamic systems. The methods discussed are appropriate for discrete (sampled-data) systems, where observation of the system occurs at equally spaced intervals of time.

We illustrate the use of these time series and dynamic models in five important areas of application:

1. The *forecasting* of future values of a time series from current and past values.
2. The determination of the *transfer function* of a system subject to inertia—the determination of a dynamic input–output model that can show the effect on the output of a system of any given series of inputs.
3. The use of indicator input variables in transfer function models to represent and assess the effects of unusual *intervention* events on the behavior of a time series.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

4. The examination of interrelationships among several related time series variables of interest and determination of appropriate *multivariate* dynamic models to represent these joint relationships among the variables over time.
5. The design of simple *control schemes* by means of which potential deviations of the system output from a desired target may, so far as possible, be compensated by adjustment of the input series values.

1.1 FIVE IMPORTANT PRACTICAL PROBLEMS

1.1.1 Forecasting Time Series

The use at time t of available observations from a time series to forecast its value at some future time $t + l$ can provide a basis for (1) economic and business planning, (2) production planning, (3) inventory and production control, and (4) control and optimization of industrial processes. As originally described by Holt et al. (1963), Brown (1962), and the Imperial Chemical Industries (ICI) monograph on short term forecasting (Coutie, 1964), forecasts are usually needed over a period known as the *lead time*, which varies with each problem. For example, the lead time in the inventory control problem was defined by Harrison (1965) as a period that begins when an order to replenish stock is placed with the factory and lasts until the order is delivered into stock.

We will assume that observations are available at *discrete*, equispaced intervals of time. For example, in a sales forecasting problem, the sales z_t in the current month t and the sales $z_{t-1}, z_{t-2}, z_{t-3}, \ldots$ in previous months might be used to forecast sales for lead times $l = 1, 2, 3, \ldots, 12$ months ahead. Denote by $\hat{z}_t(l)$ the forecast made at *origin* t of the sales z_{t+l} at some future time $t + l$, that is, at *lead time* l. The function $\hat{z}_t(l)$, which provides the forecasts at origin t for all future lead times, based on the available information from the current and previous values $z_t, z_{t-1}, z_{t-2}, z_{t-3}, \ldots$ through time t, will be called the *forecast function* at origin t. Our objective is to obtain a forecast function such that the mean square of the deviations $z_{t+l} - \hat{z}_t(l)$ between the actual and forecasted values is as small as possible *for each lead time* l.

In addition to calculating the best forecasts, it is also necessary to specify their accuracy, so that, for example, the risks associated with decisions based upon the forecasts may be calculated. The accuracy of the forecasts may be expressed by calculating *probability limits* on either side of each forecast. These limits may be calculated for any convenient set of probabilities, for example, 50 and 95%. They are such that the realized value of the time series, when it eventually occurs, will be included within these limits with the stated probability. To illustrate, Figure 1.1 shows the last 20 values of a time series culminating at time t. Also shown are forecasts made from origin t for lead times $l = 1, 2, \ldots, 13$, together with the 50% probability limits.

Methods for obtaining forecasts and estimating probability limits are discussed in detail in Chapter 5. These forecasting methods are developed based on the assumption that the time series z_t follows a *stochastic* model of known form. Consequently, in Chapters 3 and 4 a useful class of such time series models that might be appropriate to represent the behavior of a series z_t, called autoregressive integrated moving average (ARIMA) models, are introduced and many of their properties are studied. Subsequently, in Chapters 6, 7, and 8 the practical matter of how these models may be developed for actual time series data is explored, and the methods are described through the three-stage procedure of tentative

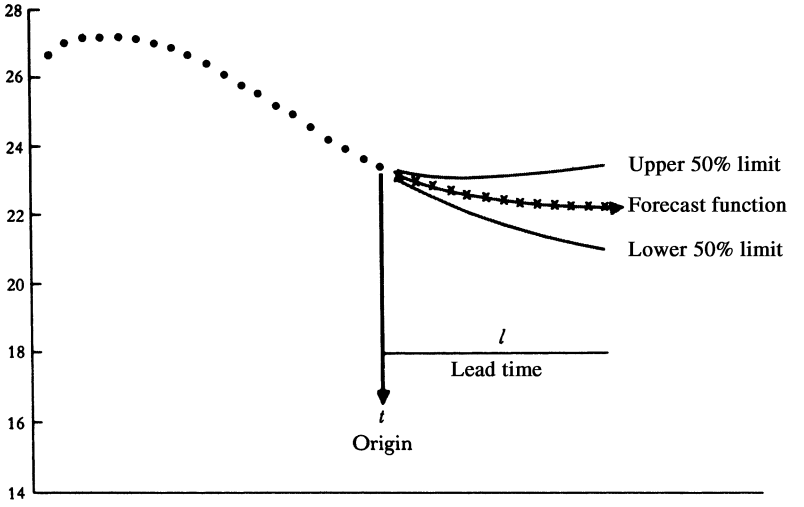


FIGURE 1.1 Values of a time series with forecast function and 50% probability limits.

model identification or specification, estimation of model parameters, and model checking and diagnostics.

1.1.2 Estimation of Transfer Functions

A topic of considerable industrial interest is the study of process dynamics discussed, for example, by Aström and Bohlin (1966, pp. 96–111) and Hutchinson and Shelton (1967). Such a study is made (1) to achieve better control of existing plants and (2) to improve the design of new plants. In particular, several methods have been proposed for estimating the transfer function of plant units from process records consisting of an input time series X_t and an output time series Y_t. Sections of such records are shown in Figure 1.2, where the input X_t is the rate of air supply and the output Y_t is the concentration of carbon dioxide produced in a furnace. The observations were made at 9-second intervals. A hypothetical impulse response function v_j, $j = 0, 1, 2, \ldots$, which determines the *transfer function* for the system through a dynamic linear relationship between input X_t and output Y_t of the form $Y_t = \sum_{j=0}^{\infty} v_j X_{t-j}$, is also shown in the figure as a bar chart. Transfer function models that

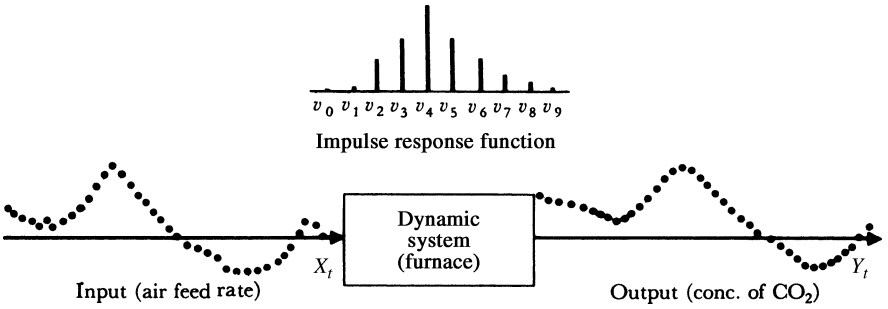


FIGURE 1.2 Input and output time series in relation to a dynamic system.

relate an input process X_t to an output process Y_t are introduced in Chapter 11 and many of their properties are examined.

Methods for estimating transfer function models based on deterministic perturbations of the input, such as step, pulse, and sinusoidal changes, have not always been successful. This is because, for perturbations of a magnitude that are relevant and tolerable, the response of the system may be masked by uncontrollable disturbances referred to collectively as *noise*. Statistical methods for estimating transfer function models that make allowance for noise in the system are described in Chapter 12. The estimation of dynamic response is of considerable interest in economics, engineering, biology, and many other fields.

Another important application of transfer function models is in forecasting. If, for example, the dynamic relationship between two time series Y_t and X_t can be determined, past values of *both* series may be used in forecasting Y_t. In some situations, this approach can lead to a considerable reduction in the errors of the forecasts.

1.1.3 Analysis of Effects of Unusual Intervention Events to a System

In some situations, it may be known that certain exceptional external events, *intervention events*, could have affected the time series z_t under study. Examples of such intervention events include the incorporation of new environmental regulations, economic policy changes, strikes, and special promotion campaigns. Under such circumstances, we may use transfer function models, as discussed in Section 1.1.2, to account for the effects of the intervention event on the series z_t, but where the ''input'' series will be in the form of a simple indicator variable taking only the values 1 and 0 to indicate (qualitatively) the presence or absence of the event.

In these cases, the intervention analysis is undertaken to obtain a quantitative measure of the impact of the intervention event on the time series of interest. For example, Box and Tiao (1975) used intervention models to study and quantify the impact of air pollution controls on smog-producing oxidant levels in the Los Angeles area and of economic controls on the consumer price index in the United States. Alternatively, the intervention analysis may be undertaken to adjust for any unusual values in the series z_t that might have resulted as a consequence of the intervention event. This will ensure that the results of the time series analysis of the series, such as the structure of the fitted model, estimates of model parameters, and forecasts of future values, are not seriously distorted by the influence of these unusual values. Models for intervention analysis and their use, together with consideration of the related topic of detection of outlying or unusual values in a time series, are presented in Chapter 13.

1.1.4 Analysis of Multivariate Time Series

For many problems in business, economics, engineering, and physical and environmental sciences, time series data may be available on several related variables of interest. A more informative and effective analysis is often possible by considering individual series as components of a multivariate or vector time series and analyzing the series jointly. For k-related time series variables of interest in a dynamic system, we may denote the series as $z_{1t}, z_{2t}, \ldots, z_{kt}$, and let $\boldsymbol{Z}_t = (z_{1t}, \ldots, z_{kt})'$ denote the $k \times 1$ time series vector at time t.

Methods of *multivariate* time series analysis are used to study the dynamic relationships among the several time series that comprise the vector $\boldsymbol{Z}_t$. This involves the development of statistical models and methods of analysis that adequately describe the interrelationships

among the series. Two main purposes for analyzing and modeling the vector of time series *jointly* are to gain an understanding of the dynamic relationships over time among the series and to improve accuracy of forecasts for individual series by utilizing the additional information available from the related series in the forecasts for each series. Multivariate time series models and methods for analysis and forecasting of multivariate series based on these models are considered in Chapter 14.

1.1.5 Discrete Control Systems

In the past, to the statistician, the words ''process control'' have usually meant the *quality control techniques* developed originally by Shewhart (1931) in the United States (see also Dudding and Jennet, 1942). Later on, the sequential aspects of quality control were emphasized, leading to the introduction of *cumulative sum charts* by Page (1957, 1961) and Barnard (1959) and the *geometric moving average* charts of Roberts (1959). Such basic charts are frequently employed in industries concerned with the manufacture of discrete ''parts'' as one aspect of what is called *statistical process control* (SPC). In particular (see Deming, 1986), they are used for continuous monitoring of a process. That is, they are used to supply a continuous screening mechanism for detecting assignable (or special) causes of variation. Appropriate display of plant data ensures that significant changes are quickly brought to the attention of those responsible for running the process. Knowing the answer to the question ''*when* did a change of this particular kind occur?'' we may be able to answer the question ''*why* did it occur?'' Hence a continuous incentive for process stabilization and improvement can be achieved.

By contrast, in the process and chemical industries, various forms of *feedback and feedforward* adjustment have been used in what we will call *engineering process control* (EPC). Because the adjustments made by engineering process control are usually computed and applied automatically, this type of control is sometimes called *automatic process control* (APC). However, the *manner* in which these adjustments are made is a matter of convenience. This type of control is necessary when there are inherent *disturbances* or *noise* in the system inputs that are impossible or impractical to remove. When we can measure fluctuations in an input variable that can be observed but not changed, it may be possible to make appropriate compensatory changes in some other control variable. This is referred to as *feedforward control*. Alternatively, or in addition, we may be able to use the deviation from target or ''error signal'' of the output characteristic itself to calculate appropriate compensatory changes in the control variable. This is called *feedback control*. Unlike feedforward control, this mode of correction can be employed even when the source of the disturbances is not accurately known or the magnitude of the disturbance is not measured.

In Chapter 15, we draw on the earlier discussions in this book, on time series and transfer function models, to provide insight into the statistical aspects of these control methods and to appreciate better their relationships and different objectives. In particular, we show how some of the ideas of feedback control can be used to design simple charts for *manually adjusting* processes. For example, the upper chart of Figure 1.3 shows hourly measurements of the viscosity of a polymer made over a period of 42 hours. The viscosity is to be controlled about a target value of 90 units. As each viscosity measurement comes to hand, the process operator uses the nomogram shown in the middle of the figure to compute the adjustment to be made in the manipulated variable (gas rate). The lower chart of Figure 1.3 shows the adjustments made in accordance with the nomogram.

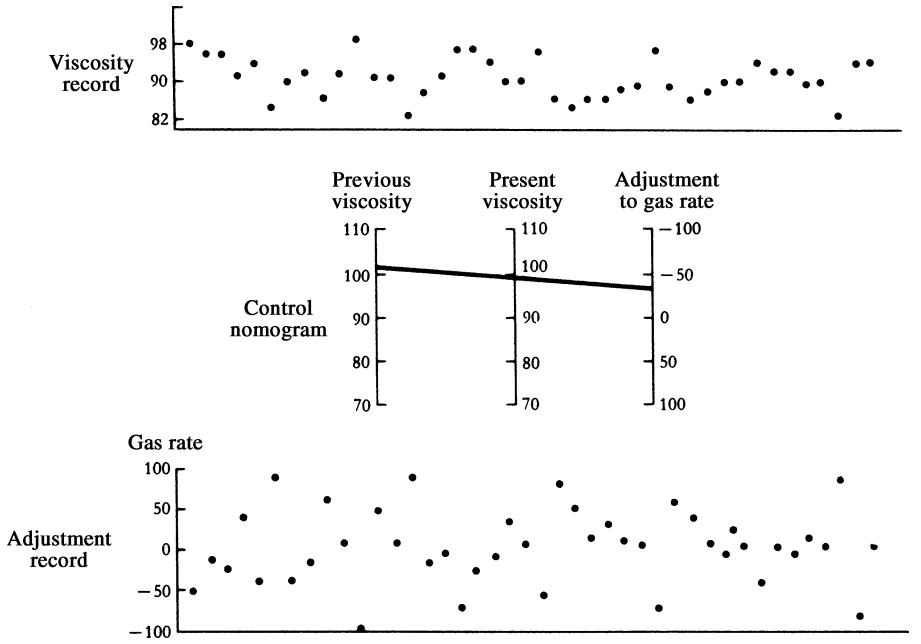


FIGURE 1.3 Control of viscosity. Record of observed viscosity and of adjustments in gas rate made using nomogram.

1.2 STOCHASTIC AND DETERMINISTIC DYNAMIC MATHEMATICAL MODELS

The idea of using a mathematical model to describe the behavior of a physical phenomenon is well established. In particular, it is sometimes possible to derive a model based on physical laws, which enables us to calculate the value of some time-dependent quantity nearly exactly at any instant of time. Thus, we might calculate the trajectory of a missile launched in a known direction with known velocity. If exact calculation were possible, such a model would be entirely *deterministic*.

Probably no phenomenon is totally deterministic, however, because unknown factors can occur such as a variable wind velocity that can throw a missile slightly off course. In many problems, we have to consider a time-dependent phenomenon, such as monthly sales of newsprint, in which there are many unknown factors and for which it is not possible to write a deterministic model that allows exact calculation of the future behavior of the phenomenon. Nevertheless, it may be possible to derive a model that can be used to calculate the *probability* of a future value lying between two specified limits. Such a model is called a probability model or a *stochastic model*. The models for time series that are needed, for example, to achieve optimal forecasting and control, are in fact stochastic models. It is necessary in what follows to distinguish between the probability model or stochastic process, as it is sometimes called, and the actually observed time series. Thus, a time series $z_1, z_2, \ldots, z_N$ of N successive observations is regarded as a sample realization from an infinite population of such time series that could have been generated by the stochastic

process. Very often we will omit the word ''stochastic'' from ''stochastic process'' and talk about the ''process.''

1.2.1 Stationary and Nonstationary Stochastic Models for Forecasting and Control

An important class of stochastic models for describing time series, which has received a great deal of attention, comprises what are called *stationary* models. Stationary models assume that the process remains in *statistical equilibrium* with probabilistic properties that do not change over time, in particular varying about a fixed *constant mean level* and with *constant variance*. However, forecasting has been of particular importance in industry, business, and economics, where many time series are often better represented as nonstationary and, in particular, as having no natural constant mean level over time. It is not surprising, therefore, that many of the economic forecasting methods originally proposed by Holt (1957, 1963), Winters (1960), Brown (1962), and the ICI monographs (Coutie, 1964) that used exponentially weighted moving averages can be shown to be appropriate for a particular type of *nonstationary* process. Although such methods are too narrow to deal efficiently with all time series, the fact that they often give the right kind of forecast function supplies a clue to the *kind of nonstationary* model that might be useful in these problems.

The stochastic model for which the exponentially weighted moving average forecast yields minimum mean square error (Muth, 1960) is a member of a class of *nonstationary* processes called autoregressive integrated moving average processes, which are discussed in Chapter 4. This wider class of processes provides a range of models, stationary and nonstationary, that adequately represent many of the time series met in practice. Our approach to forecasting has been first to derive an adequate stochastic model for the particular time series under study. As shown in Chapter 5, once an appropriate model has been determined for the series, the optimal forecasting procedure follows immediately. These forecasting procedures include the exponentially weighted moving average forecast as a special case.

Some Simple Operators. We employ extensively the *backward shift operator* B, which is defined by $Bz_t = z_{t-1}$; hence $B^m z_t = z_{t-m}$. The inverse operation is performed by the *forward shift operator* $F = B^{-1}$ given by $Fz_t = z_{t+1}$; hence $F^m z_t = z_{t+m}$. Another important operator is the *backward difference operator*, ∇, defined by $\nabla z_t = z_t - z_{t-1}$. This can be written in terms of B, since

$$\nabla z_t = z_t - z_{t-1} = (1 - B)z_t$$

Linear Filter Model. The stochastic models we employ are based on an idea originally due to Yule (1927) that an observable time series z_t in which successive values are highly dependent can frequently be regarded as generated from a series of *independent* ''shocks'' a_t. These shocks are *random* drawings from a fixed distribution, usually assumed normal and having mean zero and variance σ_a^2. Such a sequence of independent random variables $a_t, a_{t-1}, a_{t-2}, \ldots$ is called *a white noise* process.

The white noise process a_t is supposed transformed to the process z_t by what is called a *linear filter*, as shown in Figure 1.4. The linear filtering operation simply takes a weighted

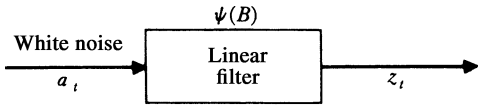


FIGURE 1.4 Representation of a time series as the output from a linear filter.

sum of previous random shocks a_t, so that

$$\begin{aligned} z_t &= \mu + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots \\ &= \mu + \psi(B) a_t \end{aligned} \tag{1.2.1}$$

In general, μ is a parameter that determines the ''level'' of the process, and

$$\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \cdots$$

is the linear operator that transforms a_t into z_t and is called the *transfer function* of the filter. The model representation (1.2.1) can allow for a flexible range of patterns of dependence among values of the process $\{z_t\}$ expressed in terms of the independent (unobservable) random shocks a_t.

The sequence $\psi_1, \psi_2, \ldots$ formed by the weights may, theoretically, be finite or infinite. If this sequence is finite, or infinite and *absolutely summable* in the sense that $\sum_{j=0}^{\infty} |\psi_j| < \infty$, the filter is said to be *stable* and the process z_t is stationary. The parameter μ is then the mean about which the process varies. Otherwise, z_t is nonstationary and μ has no specific meaning except as a reference point for the level of the process.

Autoregressive Models. A stochastic model that can be extremely useful in the representation of certain practically occurring series is the *autoregressive* model. In this model, the current value of the process is expressed as a finite, linear aggregate of *previous values of the process* and a random shock a_t. Let us denote the values of a process at equally spaced times $t,\ t-1,\ t-2,\ \ldots$ by $z_t,\ z_{t-1},\ z_{t-2}, \ldots$. Also let $\tilde{z}_t = z_t - \mu$ be the series of deviations from μ. Then

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \cdots + \phi_p \tilde{z}_{t-p} + a_t \tag{1.2.2}$$

is called an *autoregressive* (AR) *process of order p*. The reason for this name is that a linear model

$$\tilde{z} = \phi_1 \tilde{x}_1 + \phi_2 \tilde{x}_2 + \cdots + \phi_p \tilde{x}_p + a$$

relating a ''dependent'' variable z to a set of ''independent'' variables $x_1, x_2, \ldots, x_p$, plus a random error term a, is referred to as a *regression* model, and z is said to be ''regressed'' on $x_1, x_2, \ldots, x_p$. In (1.2.2) the variable z is regressed on previous values of itself; hence the model is *auto*regressive. If we define an *autoregressive operator* of order p in terms of the backward shift operator B by

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

the autoregressive model (1.2.2) may be written economically as

$$\phi(B)\tilde{z}_t = a_t$$

The model contains $p+2$ unknown parameters $\mu, \phi_1, \phi_2, \ldots, \phi_p, \sigma_a^2$, which in practice have to be estimated from the data. The additional parameter σ_a^2 is the variance of the white noise process a_t.

It is not difficult to see that the autoregressive model is a special case of the linear filter model of (1.2.1). For example, we can eliminate $\tilde{z}_{t-1}$ from the right-hand side of (1.2.2) by substituting

$$\tilde{z}_{t-1} = \phi_1 \tilde{z}_{t-2} + \phi_2 \tilde{z}_{t-3} + \cdots + \phi_p \tilde{z}_{t-p-1} + a_{t-1}$$

Similarly, we can substitute for $\tilde{z}_{t-2}$, and so on, to yield eventually an infinite series in the a's. Consider, specifically, the simple first-order ($p = 1$) AR process, $\tilde{z}_t = \phi \tilde{z}_{t-1} + a_t$. After m successive substitutions of $\tilde{z}_{t-j} = \phi \tilde{z}_{t-j-1} + a_{t-j}$, $j = 1, \ldots, m$ in the right-hand side we obtain

$$\tilde{z}_t = \phi^{m+1} \tilde{z}_{t-m-1} + a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \cdots + \phi^m a_{t-m}$$

In the limit as $m \to \infty$ this leads to the convergent infinite series representation $\tilde{z}_t = \sum_{j=0}^{\infty} \phi^j a_{t-j}$ with $\psi_j = \phi^j$, $j \geq 1$, provided that $|\phi| < 1$. Symbolically, in the general AR case we have that

$$\phi(B)\tilde{z}_t = a_t$$

is equivalent to

$$\tilde{z}_t = \phi^{-1}(B) a_t = \psi(B) a_t$$

with $\psi(B) = \phi^{-1}(B) = \sum_{j=0}^{\infty} \psi_j B^j$.

Autoregressive processes can be stationary or nonstationary. For the process to be stationary, the ϕ's must be such that the weights $\psi_1, \psi_2, \ldots$ in $\psi(B) = \phi^{-1}(B)$ form a convergent series. The necessary requirement for stationarity is that the autoregressive operator, $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$, considered as a polynomial in B of degree p, must have all roots of $\phi(B) = 0$ greater than 1 in absolute value; that is, all roots must lie outside the unit circle. For the first-order AR process $\tilde{z}_t = \phi \tilde{z}_{t-1} + a_t$ this condition reduces to the requirement that $|\phi| < 1$, as the argument above has already indicated.

Moving Average Models. The autoregressive model (1.2.2) expresses the deviation $\tilde{z}_t$ of the process as a *finite* weighted sum of p previous deviations $\tilde{z}_{t-1}, \tilde{z}_{t-2}, \ldots, \tilde{z}_{t-p}$ of the process, plus a random shock a_t. Equivalently, as we have just seen, it expresses $\tilde{z}_t$ as an *infinite* weighted sum of the a's.

Another kind of model, of great practical importance in the representation of observed time series, is the finite *moving average* process. Here we take $\tilde{z}_t$, linearly dependent on a *finite* number q of previous a's. Thus,

$$\tilde{z}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \tag{1.2.3}$$

is called a *moving average* (MA) *process of order q*. The name ''moving average'' is somewhat misleading because the weights $1, -\theta_1, -\theta_2, \ldots, -\theta_q$, which multiply the a's, need not total unity nor need they be positive. However, this nomenclature is in common use, and therefore we employ it.

If we define a *moving average operator* of order q by

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

the moving average model may be written economically as

$$\tilde{z}_t = \theta(B)a_t$$

It contains $q + 2$ unknown parameters μ, $\theta_1, \ldots, \theta_q$, σ_a^2, which in practice have to be estimated from the data.

Mixed Autoregressive–Moving Average Models. To achieve greater flexibility in fitting of actual time series, it is sometimes advantageous to include both autoregressive and moving average terms in the model. This leads to the mixed *autoregressive–moving average* (ARMA) model:

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \cdots + \phi_p \tilde{z}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{1.2.4}$$

or

$$\phi(B)\tilde{z}_t = \theta(B)a_t$$

The model employs $p + q + 2$ unknown parameters μ, $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$, σ_a^2, that are estimated from the data. This model may also be written in the form of the linear filter (1.2.1) as $\tilde{z}_t = \phi^{-1}(B)\theta(B)a_t = \psi(B)a_t$, with $\psi(B) = \phi^{-1}(B)\theta(B)$. In practice, it is frequently true that an adequate representation of actually occurring stationary time series can be obtained with autoregressive, moving average, or mixed models, in which p and q are not greater than 2 and often less than 2. We discuss the classes of autoregressive, moving average, and mixed models in much greater detail in Chapters 3 and 4.

Nonstationary Models. Many series actually encountered in industry or business (e.g., stock prices and sales figures) exhibit nonstationary behavior and in particular do not vary about a fixed mean. Such series may nevertheless exhibit homogeneous behavior over time of a kind. In particular, although the general level about which fluctuations are occurring may be different at different times, the broad behavior of the series, when differences in level are allowed for, may be similar over time. We show in Chapter 4 and later chapters that such behavior may often be represented by a model in terms of a generalized autoregressive operator $\varphi(B)$, in which one or more of the zeros of the polynomial $\varphi(B)$ [i.e., one or more of the roots of the equation $\varphi(B) = 0$] lie on the unit circle. In particular, if there are d unit roots and all other roots lie outside the unit circle, the operator $\varphi(B)$ can be written

$$\varphi(B) = \phi(B)(1 - B)^d$$

where $\phi(B)$ is a stationary autoregressive operator. Thus, a model that can represent homogeneous nonstationary behavior is of the form

$$\varphi(B)z_t = \phi(B)(1 - B)^d z_t = \theta(B)a_t$$

that is,

$$\phi(B)w_t = \theta(B)a_t \tag{1.2.5}$$

where

$$w_t = (1 - B)^d z_t = \nabla^d z_t \tag{1.2.6}$$

Thus, homogeneous nonstationary behavior can sometimes be represented by a model that calls for the dth difference of the process to be stationary. In practice, d is usually 0, 1, or at most 2, with $d = 0$ corresponding to stationary behavior.

The process defined by (1.2.5) and (1.2.6) provides a powerful model for describing stationary and nonstationary time series and is called an *autoregressive integrated moving average process*, of order (p, d, q), or ARIMA(p, d, q) process. The process is defined by

$$w_t = \phi_1 w_{t-1} + \cdots \phi_p w_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{1.2.7}$$

with $w_t = \nabla^d z_t$. Note that if we replace w_t, by $z_t - \mu$, when $d = 0$, the model (1.2.7) includes the stationary mixed model (1.2.4), as a special case, and also the pure autoregressive model (1.2.2) and the pure moving average model (1.2.3).

The reason for inclusion of the word ''integrated'' (which should perhaps more appropriately be ''summed'') in the ARIMA title is as follows. The relationship, which is the inverse to (1.2.6), is $z_t = S^d w_t$, where $S = \nabla^{-1} = (1 - B)^{-1} = 1 + B + B^2 + \cdots$ is the *summation operator* defined by

$$Sw_t = \sum_{j=0}^{\infty} w_{t-j} = w_t + w_{t-1} + w_{t-2} + \cdots$$

Thus, the general ARIMA process may be generated by summing or ''integrating'' the stationary ARMA process $w_t d$ times. In Chapter 9, we describe how a special form of the model (1.2.7) can be employed to represent seasonal time series. The chapter also includes a discussion of regression models where the errors are autocorrelated and follow an ARMA process.

Chapter 10 includes material that may be considered more specialized and that either supplements or extends the material presented in the earlier chapters. The chapter begins with a discussion of unit root testing that may be used as a supplementary tool to determine if a time series is nonstationary and can be made stationary through differencing. This is followed by a discussion of conditionally heteroscedastic models such as the ARCH and GARCH models. These models assume that the conditional variance of an observation given its past vary over time and are useful for modeling time varying volatility in economic and financial time series, in particular. In Chapter 10, we also discuss nonlinear time series models and fractionally integrated long-memory processes that allow for certain more general features in a time series than are possible using the linear ARIMA models.

1.2.2 Transfer Function Models

An important type of dynamic relationship between a continuous input and a continuous output, for which many physical examples can be found, is that in which the deviations of input X and output Y, from appropriate mean values, are related by a *linear* differential equation. In a similar way, for discrete data, in Chapter 11 we represent the transfer relationship between an output Y and an input X, each measured at equispaced times, by

the difference equation

$$(1 + \xi_1\nabla + \cdots + \xi_r\nabla^r)Y_t = (\eta_0 + \eta_1\nabla + \cdots + \eta_s\nabla^s)X_{t-b} \tag{1.2.8}$$

in which the differential operator $D = d/dt$ is replaced by the difference operator $\nabla = 1 - B$. An expression of the form (1.2.8), containing only a few parameters ($r \leq 2, s \leq 2$), may often be used as an approximation to a dynamic relationship whose true nature is more complex.

The linear model (1.2.8) may be written equivalently in terms of past values of the input and output by substituting $B = 1 - \nabla$ in (1.2.8), that is,

$$\begin{aligned}(1 - \delta_1 B - \cdots - \delta_r B^r)Y_t &= (\omega_0 - \omega_1 B - \cdots - \omega_s B^s)X_{t-b} \\ &= (\omega_0 B^b - \omega_1 B^{b+1} - \cdots - \omega_s B^{b+s})X_t\end{aligned} \tag{1.2.9}$$

or

$$\delta(B)Y_t = \omega(B)B^b X_t = \Omega(B)X_t$$

Alternatively, we can say that the output Y_t and the input X_t are linked by a linear filter

$$\begin{aligned}Y_t &= v_0 X_t + v_1 X_{t-1} + v_2 X_{t-2} + \cdots \\ &= v(B)X_t\end{aligned} \tag{1.2.10}$$

for which the transfer function

$$v(B) = v_0 + v_1 B + v_2 B^2 + \cdots \tag{1.2.11}$$

can be expressed as a ratio of two polynomial operators,

$$v(B) = \frac{\Omega(B)}{\delta(B)} = \delta^{-1}(B)\Omega(B)$$

The linear filter (1.2.10) is said to be *stable* if the series (1.2.11) converges for $|B| \leq 1$, equivalently, if the coefficients $\{v_j\}$ are absolutely summable, $\sum_{j=0}^{\infty} |v_j| < \infty$. The sequence of weights $v_0, v_1, v_2, \ldots$, which appear in the transfer function (1.2.11), is called the *impulse response function*. We note that for the model (1.2.9), the first b weights $v_0, v_1, \ldots, v_{b-1}$, are zero. A hypothetical impulse response function for the system of Figure 1.2 is shown in the center of that diagram.

Models with Superimposed Noise. We have seen that the problem of estimating an appropriate model, linking an output Y_t and an input X_t, is equivalent to estimating the transfer function $v(B) = \delta^{-1}(B)\Omega(B)$, for example, specifying the parametric form of the transfer function $v(B)$ and estimating its parameters. However, this problem is complicated in practice by the presence of noise N_t, which we assume corrupts the true relationship between input and output according to

$$Y_t = v(B)X_t + N_t$$

where N_t and X_t are independent processes. Suppose, as indicated by Figure 1.5, that the noise N_t can be described by a stationary or nonstationary stochastic model of the form

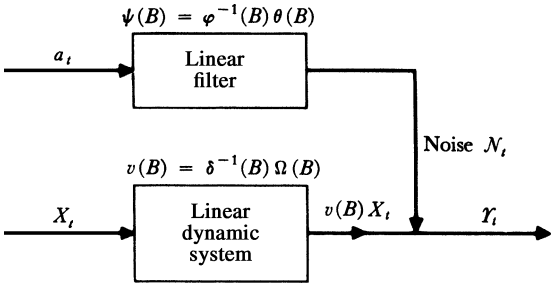


FIGURE 1.5 Transfer function model for dynamic system with superimposed noise model.

(1.2.5) or (1.2.7), that is,

$$N_t = \psi(B)a_t = \varphi^{-1}(B)\theta(B)a_t$$

Then the observed relationship between output and input will be

$$\begin{aligned} Y_t &= v(B)X_t + \psi(B)a_t \\ &= \delta^{-1}(B)\Omega(B)X_t + \varphi^{-1}(B)\theta(B)a_t \end{aligned} \tag{1.2.12}$$

In practice, it is necessary to estimate the transfer function

$$\psi(B) = \varphi^{-1}(B)\theta(B)$$

of the linear filter describing the noise, in addition to the transfer function $v(B) = \delta^{-1}(B)\Omega(B)$, which describes the dynamic relationship between the input and the output. Methods for doing this are discussed in Chapter 12.

1.2.3 Models for Discrete Control Systems

As stated in Section 1.1.5, control is an attempt to compensate for disturbances that infect a system. Some of these disturbances are measurable; others are not measurable and only manifest themselves as unexplained deviations from the target of the characteristic to be controlled. To illustrate the general principles involved, consider the special case where unmeasured disturbances affect the output Y_t of a system, and suppose that feedback control is employed to bring the output as close as possible to the desired target value by adjustments applied to an input variable X_t. This is illustrated in Figure 1.6. Suppose that N_t represents the effect at the output of various unidentified disturbances within the system, which in the absence of control could cause the output to drift away from the desired target value or *set point* T. Then, despite adjustments that have been made to the process, an error

$$\begin{aligned} \varepsilon_t &= Y_t - T \\ &= v(B)X_t + N_t - T \end{aligned}$$

will occur between the output and its target value T. The object is to choose a control equation so that the errors ε have the smallest possible mean square. The control equation expresses the adjustment $x_t = X_t - X_{t-1}$ to be taken at time t, as a function of the present deviation ε_t, previous deviations $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$, and previous adjustments $x_{t-1}, x_{t-2}, \ldots$. The mechanism (human, electrical, pneumatic, or electronic) that carries out the control action called for by the control equation is called the *controller*.

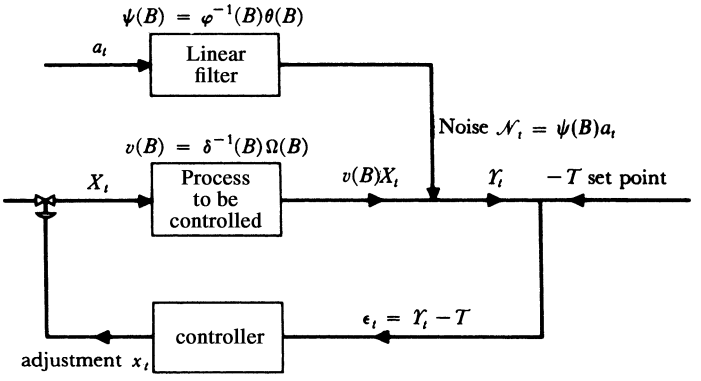


FIGURE 1.6 Feedback control scheme to compensate an unmeasured disturbance N_t.

One procedure for designing a controller is equivalent to forecasting the deviation from target which would occur *if no control were applied*, and then calculating the adjustment that would be necessary to cancel out this deviation. It follows that the forecasting and control problems are closely linked. In particular, if a minimum mean square error forecast is used, the controller will produce minimum mean square error control. To forecast the deviation from target that could occur if no control were applied, it is necessary to build a model

$$N_t = \psi(B)a_t = \varphi^{-1}(B)\theta(B)a_t$$

for the disturbance. Calculation of the adjustment x_t that needs to be applied to the input at time t to cancel out a predicted change at the output requires the building of a dynamic model with transfer function

$$v(B) = \delta^{-1}(B)\Omega(B)$$

which links the input with output. The resulting adjustment x_t will consist, in general, of a linear aggregate of previous adjustments and current and previous control errors. Thus the control equation will be of the form

$$x_t = \zeta_1 x_{t-1} + \zeta_2 x_{t-2} + \cdots + \chi_0 \varepsilon_t + \chi_1 \varepsilon_{t-1} + \chi_2 \varepsilon_{t-2} + \cdots \tag{1.2.13}$$

where $\zeta_1, \zeta_2, \ldots, \chi_0, \chi_1, \chi_2, \ldots$ are constants.

It turns out that, in practice, minimum mean square error control sometimes results in unacceptably large adjustments x_t to the input variable. Consequently, modified control schemes are employed that restrict the amount of variation in the adjustments. Some of these issues are discussed in Chapter 15.

1.3 BASIC IDEAS IN MODEL BUILDING

1.3.1 Parsimony

We have seen that the mathematical models we need to employ contain certain constants or parameters whose values must be estimated from the data. It is important, in practice, that

we employ the *smallest possible* number of parameters for adequate representations. The central role played by this principle of *parsimony* (Tukey, 1961) in the use of parameters will become clearer as we proceed. As a preliminary illustration, we consider the following simple example.

Suppose we fitted a dynamic model (1.2.9) of the form

$$Y_t = (\omega_0 - \omega_1 B - \omega_2 B^2 - \cdots - \omega_s B^s)X_t \tag{1.3.1}$$

when dealing with a system that was adequately represented by

$$(1 - \delta B)Y_t = \omega_0 X_t \tag{1.3.2}$$

The model (1.3.2) contains only two parameters, δ and ω_0, but for s sufficiently large, it could be represented approximately by the model (1.3.1), through

$$Y_t = (1 - \delta B)^{-1}\omega_0 X_t = \omega_0(1 + \delta B + \delta^2 B^2 + \cdots)X_t$$

with $|\delta| < 1$. Because of experimental error, we could easily fail to recognize the relationship between the coefficients in the fitted equation. Thus, we might needlessly fit a relationship like (1.3.1), containing $s + 1$ parameters, where the much simpler form (1.3.2), containing only two, would have been adequate. This could, for example, lead to unnecessarily poor estimation of the output Y_t for given values of the input $X_t, X_{t-1}, \ldots$.

Our objective, then, must be to obtain adequate but parsimonious models. Forecasting and control procedures could be seriously deficient if these models were either inadequate or unnecessarily prodigal in the use of parameters. Care and effort is needed in selecting the model. The process of selection is necessarily iterative; that is, it is a process of evolution, adaptation, or trial and error and is outlined briefly below.

1.3.2 Iterative Stages in the Selection of a Model

If the physical mechanism of a phenomenon were completely understood, it would be possible theoretically to write down a mathematical expression that described it exactly. This would result in a *mechanistic* or *theoretical* model. In most instances the complete knowledge or large experimental resources needed to produce a mechanistic model are not available, and we must resort to an empirical model. Of course, the exact mechanistic model and the exclusively empirical model represent extremes. Models actually employed usually lie somewhere in between. In particular, we may use incomplete theoretical knowledge to indicate a suitable class of mathematical functions, which will then be fitted empirically (e.g., Box and Hunter, 1965); that is, the number of terms needed in the model and the numerical values of the parameters are estimated from experimental data. This is the approach that we adopt in this book. As we have indicated previously, the stochastic and dynamic models we describe can be justified, at least partially, on theoretical grounds as having the right general properties.

It is normally supposed that successive values of the time series under consideration or of the input–output data are available for analysis. If possible, at least 50 and preferably 100 observations or more should be used. In those cases where a past history of 50 or more observations is not available, one proceeds by using experience and past information to derive a preliminary model. This model may be updated from time to time as more data become available.

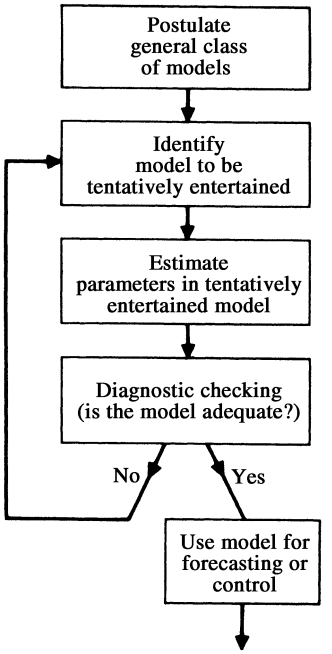


FIGURE 1.7 Stages in the iterative approach to model building.

In fitting dynamic models, a theoretical analysis can sometimes tell us not only the appropriate form for the model, but may also provide us with good estimates of the numerical values of its parameters. These values can then be checked later by analysis of data.

Figure 1.7 summarizes the iterative approach to model building for forecasting and control, which is employed in this book.

1. From the interaction of theory and practice, a *useful class of models* for the purposes at hand is considered.
2. Because this class is too extensive to be conveniently fitted directly to data, rough methods for *identifying* subclasses of these models are developed. Such methods of model identification employ data and knowledge of the system to suggest an appropriate parsimonious subclass of models that may be tentatively entertained. In addition, the identification process can be used to yield rough preliminary estimates of the parameters in the model.
3. The tentatively entertained model is *fitted* to data and its parameters *estimated*. The rough estimates obtained during the identification stage can now be used as starting values in more refined iterative methods for estimating the parameters, such as the nonlinear least squares and maximum likelihood methods.
4. *Diagnostic checks* are applied with the goal of uncovering possible lack of fit and diagnosing the cause. If no lack of fit is indicated, the model is ready to use. If any inadequacy is found, the iterative cycle of identification, estimation, and diagnostic checking is repeated until a suitable representation is found.

Identification, estimation, and diagnostic checking are discussed for univariate time series models in Chapters 6, 7, 8, and 9, for transfer function models in Chapter 12, for intervention models in Chapter 13, and for multivariate time series models in Chapter 14.

The model building procedures will be illustrated using actual time series with numerical calculations performed using the R software and other tools. A brief description of the R software is included in Appendix A1.1 along with references for further study. Exercises at the end of the chapters also make use of the software.

APPENDIX A1.1 USE OF THE R SOFTWARE

The R software for statistical computing and graphics is a common choice for data analysis and development of new statistical methods. R is available as Free Software under the terms of the Free Software Foundations's GNU General Public License in source code form. It compiles and runs on all common operating systems including Windows, MacOS X, and Linux. The main website for the R project is http://www.r-project.org.

The R environment consists of a base system, which is developed and maintained by the R Core Team, and a large set of user contributed packages. The base system provides the source code that implements the basic functionality of R. It also provides a set of standard packages that include commonly used probability distributions, graphical tools, classic datasets from the literature, and a set of statistical methods that include regression analysis and time series analysis. In addition to these base packages, there are now thousands of contributed packages developed by researchers around the world. Packages useful for time series modeling and forecasting include the stats package that is part of the base distribution and several contributed packages that are available for download. These include the TSA package by K-S Chan and Brian Ripley, the astsa package by David Stoffer, the Rmetrics packages fGarch and fUnitRoots for financial time series analysis by Diethelm Wuertz and associates, and the MTS package for multivariate time series analysis by Ruey Tsay. We use many functions from these packages in this book. We also use datasets available for download from the R datasets package, and the TSA and astsa packages.

Both the base system and the contributed packages are distributed through a network of servers called the Comprehensive R Archive Network (CRAN) that can be accessed from the official R website. Contributed packages that are not part of the base distribution can be installed directly from the R prompt ''>'' using the command install.package(). Under the Windows system, the installation can also be done from a drop-down list. The command will prompt the user to select a *CRAN Mirror*, after which a list of packages available for installation appears. To use a specific package, it also needs to be loaded into the system at the start of each session. For example, the TSA package can be loaded using the commands library(TSA) or require(TSA). The command data() will list all datasets available in the loaded packages. The command data(airquality) will load the dataset airquality from the R datasets package into memory. Data stored in a text file can be read into R using the command is read.table. For a .csv file, the command is read.csv. To get help on specific functions, e.g. the arima function which fits an ARIMA model to a time series, type help(arima) or ?arima.

R is object-oriented software and allows the user to create many objects. For example, the command ts() will create a time series object. This has advantages for plotting the time series and for certain other applications. However, it is not necessary to create a time series

object for many of the applications discussed in this book. The structure of the data in R can be examined using commands such as class(), str(), and summary().

The data used for illustration in this book, as well as in some of the exercises, include a set of time series listed in Part Five of the book. These series are also available at http://pages.cs.wisc.edu/ reinsel/bjr-data/index.html. At least three of the series are also included in the R datasets package and can be accessed using the data() command described above. Some of the exercises require the use of R and it will be assumed that the reader is already familiar with the basics of R, which can be obtained by working through relevant chapters of texts such as Crawley (2007) and Adler (2010). Comprehensive documentation in the form of manuals, contributed documents, online help pages, and FAQ sheets is also available on the R website. Since R builds on the S language, a useful reference book is also Venables and Ripley (2002).

EXERCISES

1.1. The dataset airquality in the R datasets package includes information on daily air quality measurements in New York, May to September 1973. The variables included are mean ozone levels at Roosevelt Island, solar radiation at Central Park, average wind speed at LaGuardia Airport, and maximum daily temperature at LaGuardia Airport; see help(airquality) for details.

(a) Load the dataset into R.

(b) Investigate the structure of the dataset.

(c) Plot each of the four series mentioned above using the plot() command in R; see help(plot) for details and examples.

(d) Comment on the behavior of the four series. Do you see any issues that may require special attention in developing a time series model for each of the four series.

1.2. Monthly totals of international airline passengers (in thousands of passengers), January 1949–December 1960, are available as Series G in Part Five of this book. The data are also available as series AirPassengers in the R datasets package.

(a) Load the dataset into R and examine the structure of the data.

(b) Plot the data using R and describe the behavior of the series.

(c) Perform a log transformation of the data and plot the resulting series. Compare the behavior of the original and log-transformed series. Do you see an advantage in using a log transformation for modeling purposes?

1.3. Download a time series of your choosing from the Internet. Note that financial and economic time series are available from sources such as Google Finance and the Federal Reserve Economic Data (FRED) of Federal Reserve Bank in St. Louis, Missouri, while climate data is available from from NOAA's National Climatic Data Center (NCDC).

(a) Store the data in a text file or a .csv file and read the data into R.

(b) Examine the properties of your series using plots or other appropriate tools.

(c) Does your time series appear to be stationary? If not, would differencing and/or some other transformation make the series stationary?

PART ONE

STOCHASTIC MODELS AND THEIR FORECASTING

In the first part of this book, which includes Chapters 2, 3, 4, and 5, a valuable class of stochastic models is described and its use in forecasting discussed.

A model that describes the probability structure of a sequence of observations is called a *stochastic process*. A time series of N successive observations $\mathbf{z}' = (z_1, z_2, \ldots, z_N)$ is regarded as a sample realization, from an infinite population of such samples, which could have been generated by the process. A major objective of statistical investigation is to infer properties of the population from those of the sample. For example, to make a forecast is to infer the probability distribution of a *future observation* from the population, given a sample $\mathbf{z}$ of past values. To do this, we need ways of describing stochastic processes and time series, and we also need classes of stochastic models that are capable of describing practically occurring situations. An important class of stochastic processes discussed in Chapter 2 is the *stationary* processes. They are assumed to be in a specific form of statistical equilibrium, and in particular, vary over time in a stable manner about a fixed mean. Useful devices for describing the behavior of stationary processes are the *autocorrelation function* and the *spectrum*.

Particular stationary stochastic processes of value in modeling time series are the autoregressive (AR), moving average (MA), and mixed autoregressive–moving average (ARMA) processes. The properties of these processes, in particular their autocorrelation structures, are described in Chapter 3.

Because many practically occurring time series (e.g., stock prices and sales figures) have nonstationary characteristics, the stationary models introduced in Chapter 3 are developed further in Chapter 4 to give a useful class of nonstationary processes called autoregressive integrated moving-average (ARIMA) models. The use of all these models in forecasting time series is discussed in Chapter 5 and is illustrated with examples.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

2

AUTOCORRELATION FUNCTION AND SPECTRUM OF STATIONARY PROCESSES

A central feature in the development of time series models is an assumption of some form of *statistical equilibrium*. A particularly useful assumption of this kind (but an unduly restrictive one, as we will see later) is that of *stationarity*. Usually, a stationary time series can be usefully described by its mean, variance, and *autocorrelation function* or equivalently by its mean, variance, and *spectral density function*. In this chapter, we consider the properties of these functions and, in particular, the properties of the autocorrelation function, which will be used extensively in developing models for actual time series.

2.1 AUTOCORRELATION PROPERTIES OF STATIONARY MODELS

2.1.1 Time Series and Stochastic Processes

Time Series. A time series is a set of observations generated sequentially over time. If the set is continuous, the time series is said to be *continuous*. If the set is discrete, the time series is said to be *discrete*. Thus, the observations from a discrete time series made at times $\tau_1, \tau_2, \ldots, \tau_t, \ldots, \tau_N$ may be denoted by $z(\tau_1), z(\tau_2), \ldots, z(\tau_t), \ldots, z(\tau_N)$. In this book, we consider only discrete time series where observations are made at a fixed interval h. When we have N *successive* values of such a series available for analysis, we write $z_1, z_2, \ldots, z_t, \ldots, z_N$ to denote observations made at equidistant time intervals $\tau_0 + h, \tau_0 + 2h, \ldots, \tau_0 + th, \ldots, \tau_0 + Nh$. For many purposes, the values of τ_0 and h are unimportant, but if the observation times need to be defined exactly, these two values can be specified. If we adopt τ_0 as the origin and h as the unit of time, we can regard z_t as the observation *at time t*.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

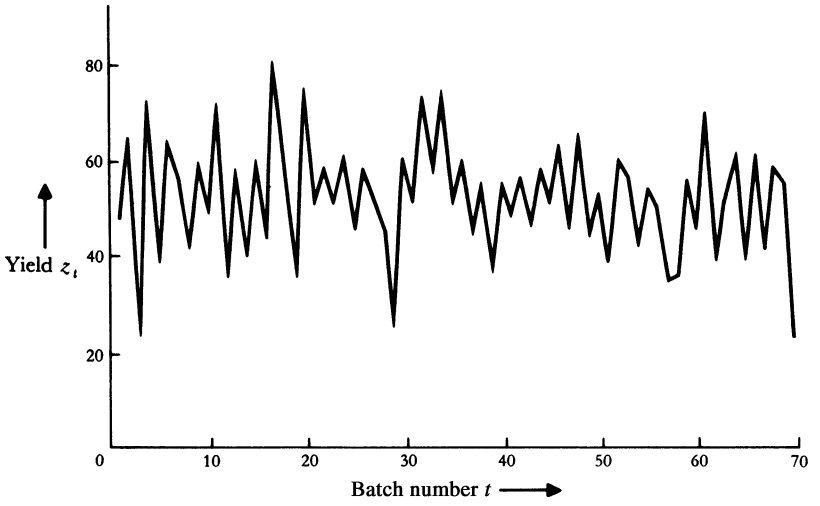


FIGURE 2.1 Yields of 70 consecutive batches from a chemical process.

Discrete time series may arise in two ways:

1. By *sampling* a continuous time series: For example, in the situation shown in Figure 1.2, where the continuous input and output from a gas furnace was sampled at intervals of 9 seconds.
2. By *accumulating* a variable over a period of time: Examples are rainfall, which is usually accumulated over a period such as a day or a month, and the yield from a batch process, which is accumulated over the batch time. For example, Figure 2.1 shows a time series consisting of the yields from 70 consecutive batches of a chemical process. The series shown here is included as Series F in Part Five of this book.

Deterministic and Statistical Time Series. If future values of a time series are exactly determined by some mathematical function such as

$$z_t = \cos(2\pi f t)$$

the time series is said to be *deterministic*. If future values can be described only in terms of a probability distribution, the time series is said to be nondeterministic or simply a *statistical time series*. The batch data of Figure 2.1 provide an example of a statistical time series. Thus, although there is a well-defined high–low pattern in the series, it is impossible to forecast the exact yield for the next batch. It is with such statistical time series that we are concerned in this book.

Stochastic Processes. A statistical phenomenon that evolves in time according to probabilistic laws is called a *stochastic process*. We will often refer to it simply as a *process*, omitting the word ''stochastic.'' The time series to be analyzed may then be thought of as a particular *realization*, produced by the underlying probability mechanism, of the system under study. In other words, *in analyzing a time series we regard it as a realization of a stochastic process.*

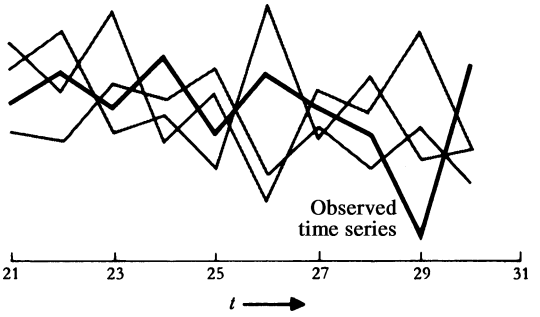


FIGURE 2.2 Observed time series (thick line), with other time series representing realizations of the same stochastic process.

For example, to analyze the batch data in Figure 2.1, we can imagine other sets of observations (other realizations of the underlying stochastic process), which might have been generated by the same chemical system, in the same $N = 70$ batches. Thus, Figure 2.2 shows the yields from batches $t = 21$ to $t = 30$ (thick line), together with other time series that *might* have been obtained from the population of time series defined by the underlying stochastic process. It follows that we can regard the observation z_t at a given time t, say $t = 25$, as a realization of a random variable z_t with probability density function $p(z_t)$. Similarly, the observations at any two times, say $t_1 = 25$ and $t_2 = 27$, may be regarded as realizations of two random variables z_{t_1} and z_{t_2} with joint probability density function $p(z_{t_1}, z_{t_2})$. For illustration Figure 2.3 shows contours of constant density for such a joint distribution, together with the marginal distribution at time t_1. In general, the observations making up an equispaced time series can be described by an N-dimensional random variable $(z_1, z_2, \ldots, z_N)$ with probability distribution $p(z_1, z_2, \ldots, z_N)$.

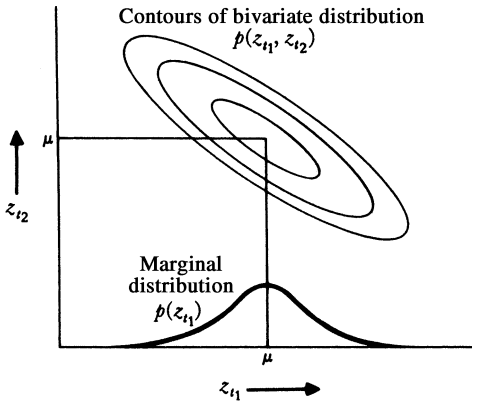


FIGURE 2.3 Contours of constant density of a bivariate probability distribution describing a stochastic process at two times t_1, t_2, together with the marginal distribution at time t_1.

2.1.2 Stationary Stochastic Processes

A very special class of stochastic processes, called *stationary processes*, is based on the assumption that the process is in a particular state of *statistical equilibrium*. A stochastic process is said to be *strictly stationary* if its properties are unaffected by a change of time origin, that is, if the joint probability distribution associated with m observations $z_{t_1}, z_{t_2}, \ldots, z_{t_m}$, made at *any* set of times $t_1, t_2, \ldots, t_m$, is the same as that associated with m observations $z_{t_1+k}, z_{t_2+k}, \ldots, z_{t_m+k}$, made at times $t_1 + k, t_2 + k, \ldots, t_m + k$. Thus, for a discrete process to be strictly stationary, the joint distribution of any set of observations must be unaffected by shifting all the times of observation forward or backward by any integer amount k.

Mean and Variance of a Stationary Process. When $m = 1$, the stationarity assumption implies that the probability distribution $p(z_t)$ is the same for all times t and may be written as $p(z)$. Hence, the stochastic process has a constant mean

$$\mu = E[z_t] = \int_{-\infty}^{\infty} z p(z) dz \tag{2.1.1}$$

which defines the level about which it fluctuates, and a constant variance

$$\sigma_z^2 = E[(z_t - \mu)^2] = \int_{-\infty}^{\infty} (z - \mu)^2 p(z) dz \tag{2.1.2}$$

which measures its *spread* about this level. Since the probability distribution $p(z)$ is the same for all times t, its shape can be inferred by forming the histogram of the observations $z_1, z_2, \ldots, z_N$, making up the observed time series. In addition, the mean μ of the stochastic process can be estimated by the sample mean

$$\bar{z} = \frac{1}{N} \sum_{t=1}^{N} z_t \tag{2.1.3}$$

of the time series, and the variance σ_z^2 of the stochastic process can be estimated by the sample variance

$$\hat{\sigma}_z^2 = \frac{1}{N} \sum_{t=1}^{N} (z_t - \bar{z})^2 \tag{2.1.4}$$

of the time series.

Autocovariance and Autocorrelation Coefficients. The stationarity assumption also implies that the joint probability distribution $p(z_{t_1}, z_{t_2})$ is the same for all times t_1, t_2, which are a constant interval apart. In particular, it follows that the covariance between values z_t and z_{t+k}, separated by k intervals of time, or by *lag* k, must be the same for all t under the stationarity assumption. This covariance is called the *autocovariance* at lag k and is defined by

$$\gamma_k = \text{cov}[z_t, z_{t+k}] = E[(z_t - \mu)(z_{t+k} - \mu)] \tag{2.1.5}$$

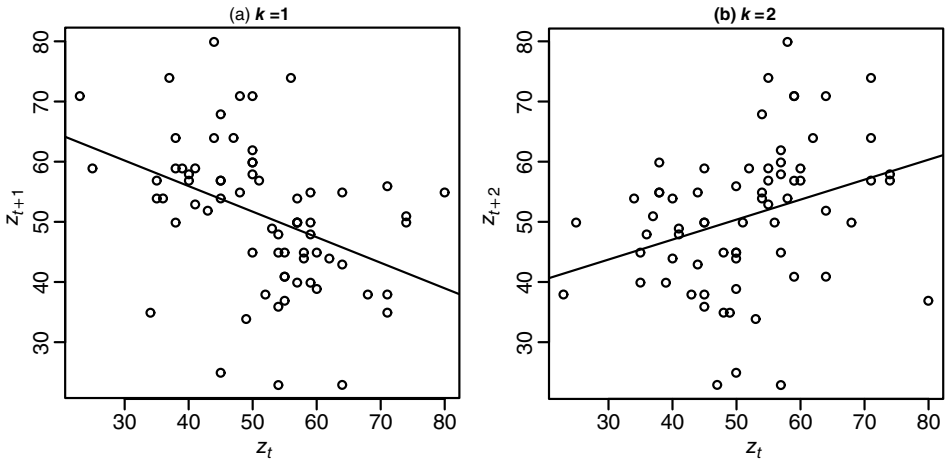


FIGURE 2.4 Scatter diagrams at lags (a) $k = 1$ and (b) $k = 2$ for the batch data of Figure 2.1.

Similarly, the *autocorrelation* at lag k is

$$\rho_k = \frac{E[(z_t - \mu)(z_{t+k} - \mu)]}{\sqrt{E[(z_t - \mu)^2]E[(z_{t+k} - \mu)^2]}} = \frac{E[(z_t - \mu)(z_{t+k} - \mu)]}{\sigma_z^2}$$

since, for a stationary process, the variance $\sigma_z^2 = \gamma_0$ is the same at time $t + k$ as at time t. Thus, the autocorrelation at lag k, that is, the correlation between z_t and z_{t+k}, is

$$\rho_k = \frac{\gamma_k}{\gamma_0} \tag{2.1.6}$$

which implies, in particular, that $\rho_0 = 1$.

It also follows for a stationary process that the nature of the joint probability distribution $p(z_t, z_{t+k})$ of values separated by k intervals of time can be inferred by plotting a scatter diagram using pairs of values (z_t, z_{t+k}) of the time series, separated by a constant interval or lag k. For the batch data displayed in Figure 2.1, Figure 2.4(a) shows a scatter diagram for lag $k = 1$, obtained by plotting z_{t+1} versus z_t, while Figure 2.4(b) shows a scatter diagram for lag $k = 2$, obtained by plotting z_{t+2} versus z_t. We see that neighboring values of the time series are correlated. The correlation between z_t and z_{t+1} appears to be negative and the correlation between z_t and z_{t+2} positive. Figure 2.4 was generated in R as follows:

```
> Yield = read.table("SeriesF.txt",header=TRUE)
> y1=Yield[2:70]
> x1=Yield[1:69]
> y2=Yield[3:70]
> x2=Yield[1:68]
> win.graph(width=5,height=2.7,pointsize=5)
> par(mfrow=c(1,2))  % Places two graphs side-by-side
> plot(y=y1,x=x1,ylab=expression(z[t+1]),xlab=expression(z[t]),
```

```
        main="(a): k=1",type='p')
> abline(lsfit(x1,y1))
> plot(y=y2,x=x2,ylab=expression(z[t+2]),xlab=expression(z[t]),
        main="(b): k=2",type='p')
> abline(lsfit(x2,y2))
```

2.1.3 Positive Definiteness and the Autocovariance Matrix

The covariance matrix associated with a stationary process for observations $(z_1, z_2, \ldots, z_n)$ made at n successive times is

$$\boldsymbol{\Gamma}_n = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{n-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{n-3} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \gamma_{n-3} & \cdots & \gamma_0 \end{bmatrix}$$
$$= \sigma_z^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix} = \sigma_z^2 \mathbf{P}_n \tag{2.1.7}$$

A covariance matrix $\boldsymbol{\Gamma}_n$ of this form, which is symmetric with constant elements on any diagonal, is called an *autocovariance matrix*, and the corresponding correlation matrix $\mathbf{P}_n$ is called an *autocorrelation matrix*. Now, consider any linear function of the random variables $z_t, z_{t-1}, \ldots, z_{t-n+1}$:

$$L_t = l_1 z_t + l_2 z_{t-1} + \cdots + l_n z_{t-n+1} \tag{2.1.8}$$

Since $\text{cov}[z_i, z_j] = \gamma_{|j-i|}$ for a stationary process, the variance of L_t is

$$\text{var}[L_t] = \sum_{i=1}^{n} \sum_{j=1}^{n} l_i l_j \gamma_{|j-i|}$$

which is necessarily greater than zero if the l's are not all zero. It follows that both the autocovariance matrix and the autocorrelation matrix are positive definite for any stationary process. Correspondingly, it is seen that both the autocovariance function $\{\gamma_k\}$ and the autocorrelation function $\{\rho_k\}$, viewed as functions of the lag k, are positive-definite functions in the sense that $\sum_{i=1}^{n} \sum_{j=1}^{n} l_i l_j \gamma_{|j-i|} > 0$ for every positive integer n and all constants $l_1, \ldots, l_n$.

Conditions Satisfied by the Autocorrelations of a Stationary Process. The positive definiteness of the autocorrelation matrix (2.1.7) implies that its determinant and all principal minors are greater than zero. In particular, for $n = 2$,

$$\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix} > 0$$

so that

$$1 - \rho_1^2 > 0$$

and hence

$$-1 < \rho_1 < 1$$

Similarly, for $n = 3$, we must have

$$\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix} > 0 \qquad \begin{vmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{vmatrix} > 0$$

$$\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix} > 0$$

which implies that

$$\begin{aligned} -1 &< \rho_1 < 1 \\ -1 &< \rho_2 < 1 \\ -1 &< \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} < 1 \end{aligned}$$

and so on. Since $\mathbf{P}_n$ must be positive definite for *all* values of n, the autocorrelations of a stationary process must satisfy a very large number of conditions. As will be shown in Section 2.2.3, all of these conditions can be brought together in the definition of the spectrum.

Stationarity of Linear Functions. It follows from the definition of stationarity that the process L_t, obtained by performing the linear operation (2.1.8) on a stationary process z_t for fixed n and fixed coefficients $l_1, \ldots, l_n$, is also stationary. The autocovariance of the process L_t, at a general lag $k \geq 0$, is given by

$$\text{cov}[L_t, L_{t-k}] = \sum_{i=1}^{n} \sum_{j=1}^{n} l_i l_j \text{cov}[z_{t+1-i}, z_{t+1-k-j}] = \sum_{i=1}^{n} \sum_{j=1}^{n} l_i l_j \gamma_{|k+j-i|}$$

In particular, the first difference $\nabla z_t = z_t - z_{t-1}$ and higher differences $\nabla^d z_t$ are stationary. This result is of particular importance to the discussion of nonstationary time series presented in Chapter 4.

The result also extends to infinite linear operations or infinite linear (time-invariant) filters applied to a stationary process $\{z_t\}$, under a condition of absolute summability. That is, if $\{z_t\}$ is a stationary process and $\{y_t\}$ is defined by the infinite linear (time-invariant) filter

$$y_t = \psi_0 z_t + \psi_1 z_{t-1} + \psi_2 z_{t-2} + \cdots = \sum_{i=0}^{\infty} \psi_i z_{t-i} \tag{2.1.9}$$

with fixed coefficients $\{\psi_i\}$ such that $\sum_{i=0}^{\infty} |\psi_i| < \infty$, then $\{y_t\}$ is also stationary. The absolute summability condition, $\sum_{i=0}^{\infty} |\psi_i| < \infty$, guarantees that the variables y_t in (2.1.9) are well-defined finite random variables (with probability one) and represent the limit of the sequence $\sum_{i=0}^{n} \psi_i z_{t-i}$ as $n \to \infty$. The variance of y_t in (2.1.9) (taking $E[z_t] = 0$ for convenience) is

$$\text{var}[y_t] = E[y_t^2] = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \psi_i \psi_j \gamma_{|j-i|}$$

This variance is finite since $|\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \psi_i \psi_j \gamma_{|j-i|}| \le \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} |\psi_i||\psi_j||\gamma_{|j-i|}| \le \gamma_0 \{\sum_{i=0}^{\infty} |\psi_i|\}^2 < \infty$. The autocovariance of y_t at any lag $k \ge 0$ is then

$$\text{cov}[y_t, y_{t-k}] = \lim_{n \to \infty} \sum_{i=0}^{n} \sum_{j=0}^{n} \psi_i \psi_j \gamma_{|k+j-i|} = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \psi_i \psi_j \gamma_{|k+j-i|} \tag{2.1.10}$$

which converges by the dominated convergence result.

Gaussian Processes. If the probability distribution of observations associated with *any* set of times is a multivariate normal distribution, the process is called a *normal* or *Gaussian* process. Since the multivariate normal distribution is fully characterized by its moments of first and second order, the existence of a fixed mean μ and an autocovariance matrix $\mathbf{\Gamma}_n$ of the form (2.1.7) for all n would be sufficient to ensure the stationarity of a Gaussian process.

Weak Stationarity. We have seen that for a process to be strictly stationary, the whole probability structure must depend only on time differences. A less restrictive requirement, called *weak stationarity* of order f, is that the moments up to some order f depend only on time differences. For example, the existence of a fixed mean μ and an autocovariance matrix $\mathbf{\Gamma}_n$ of the form (2.1.7) is sufficient to ensure stationarity up to second order. That is, a process $\{z_t\}$ is *weakly stationary* (of order 2), or second-order stationary, if the mean $E[z_t] = \mu$ is a fixed constant for all t and the autocovariances $\text{cov}[z_t, z_{t+k}] = \gamma_k$ depend only on the time difference or time lag k for all t. Thus, second-order stationarity and an assumption of normality are sufficient to produce strict stationarity.

White Noise Process. The most fundamental example of a stationary process is a sequence of *independent and identically distributed* random variables, denoted as $a_1, \ldots, a_t, \ldots$, which we also assume to have mean zero and variance σ_a^2. This process is strictly stationary and is referred to as a *white noise* process. Because independence implies that the a_t are uncorrelated, its autocovariance function is simply

$$\gamma_k = E[a_t a_{t+k}] = \begin{cases} \sigma_a^2 & k = 0 \\ 0 & k \ne 0 \end{cases}$$

If one concentrates only on the second-order properties, then a sequence of random variables a_t, which are *uncorrelated*, have mean zero, and common variance σ_a^2 has the same autocovariance function γ_k as above, and is weakly (second-order) stationary. Such a process may also be referred to as a white noise process (in the weak sense), when the focus

is only on the second-order properties. Although the white noise process has very basic properties, this process plays an important role in the building of processes with much more interesting and more complicated properties through linear filtering operations as in (2.1.8) and (2.1.9).

2.1.4 Autocovariance and Autocorrelation Functions

It was seen in Section 2.1.2 that the autocovariance coefficient γ_k, at lag k, measures the covariance between two values z_t and z_{t+k} a distance k apart. The plot of γ_k versus lag k is called the *autocovariance function* $\{\gamma_k\}$ of the stochastic process. Similarly, the plot of the autocorrelation coefficient ρ_k as a function of the lag k is called the *autocorrelation function* $\{\rho_k\}$ of the process. Note that the autocorrelation function is dimensionless, that is, independent of the scale of measurement of the time series. Since $\gamma_k = \rho_k \sigma_z^2$, knowledge of the autocorrelation function $\{\rho_k\}$ and the variance σ_z^2 is equivalent to knowledge of the autocovariance function $\{\gamma_k\}$.

The autocorrelation function, shown in Figure 2.5 as a plot of the diagonals of the autocorrelation matrix, reveals how the correlation between any two values of the series changes as their separation changes. Since $\rho_k = \rho_{-k}$, the autocorrelation function is

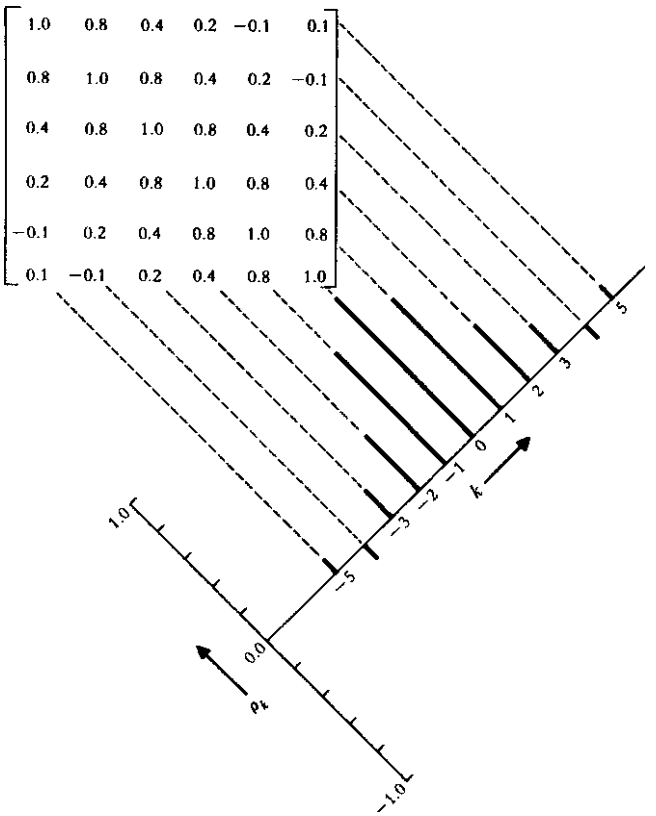


FIGURE 2.5 Autocorrelation matrix and corresponding autocorrelation function of a stationary process.

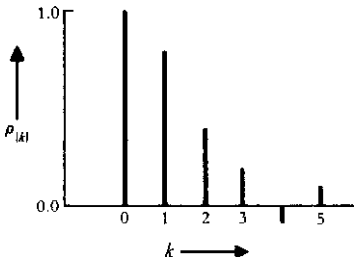


FIGURE 2.6 Positive half of the autocorrelation function of Figure 2.5.

necessarily symmetric about zero, and in practice it is only necessary to plot the positive half of this function. Figure 2.6 shows the positive half of the autocorrelation function given in Figure 2.5. When we speak of the autocorrelation function, we typically mean the positive half. In the past, the autocorrelation function has sometimes been called the *correlogram*.

From what has previously been shown, a *normal* stationary process z_t is completely characterized by its mean μ and its autocovariance function $\{\gamma_k\}$, or equivalently by its mean μ, variance σ_z^2, and autocorrelation function $\{\rho_k\}$.

2.1.5 Estimation of Autocovariance and Autocorrelation Functions

Up to now, we have only considered the theoretical autocorrelation function that describes a stochastic process. In practice, we have a finite time series $z_1, z_2, \ldots, z_N$ of N observations, from which we can only obtain *estimates* of the mean μ and the autocorrelations. The mean $\mu = E[z_t]$ is estimated as in (2.1.3) by the sample mean $\bar{z} = \sum_{t=1}^{N} z_t/N$. It is easy to see that $E[\bar{z}] = \mu$, so that $\bar{z}$ is an unbiased estimator of μ. As a measure of precision of $\bar{z}$ as an estimator of μ, we find that

$$\operatorname{var}[\bar{z}] = \frac{1}{N^2}\sum_{t=1}^{N}\sum_{s=1}^{N}\gamma_{t-s} = \frac{\gamma_0}{N}\left[1 + 2\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\rho_k\right]$$

A ''large-sample'' approximation for this variance is given by

$$\operatorname{var}[\bar{z}] = \left(\frac{\gamma_0}{N}\right)\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right)$$

in the sense that $N\operatorname{var}[\bar{z}] \to \gamma_0\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right)$ as $N \to \infty$, assuming that $\sum_{k=-\infty}^{\infty}|\rho_k| < \infty$. Notice that the first factor in $\operatorname{var}[\bar{z}]$, γ_0/N, is the familiar expression for the variance of $\bar{z}$ obtained from independent random samples of size N, but the presence of autocorrelation among the z_t values can substantially affect the precision of $\bar{z}$. For example, in the case where a stationary process has autocorrelations $\rho_k = \phi^{|k|}$, $|\phi| < 1$, the large-sample approximation for the variance of $\bar{z}$ becomes $\operatorname{var}[\bar{z}] = (\gamma_0/N)[(1+\phi)/(1-\phi)]$, and the second factor can obviously differ substantially from 1.

A number of estimates of the autocorrelation function have been suggested in the literature, and their properties are discussed by Jenkins and Watts (1968), among others. It

TABLE 2.1 Estimated Autocorrelation Function of Batch Data

k	r_k	k	r_k	k	r_k
1	−0.39	6	−0.05	11	0.11
2	0.30	7	0.04	12	−0.07
3	−0.17	8	−0.04	13	0.15
4	0.07	9	0.00	14	0.04
5	−0.10	10	0.01	15	−0.01

is concluded that the most satisfactory estimate of the kth lag autocorrelation ρ_k is

$$r_k = \hat{\rho}_k = \frac{c_k}{c_0} \tag{2.1.11}$$

where

$$c_k = \hat{\gamma}_k = \frac{1}{N}\sum_{t=1}^{N-k}(z_t - \bar{z})(z_{t+k} - \bar{z}) \qquad k = 0, 1, 2, \ldots, K \tag{2.1.12}$$

is the estimate of the autocovariance γ_k and $\bar{z}$ is the sample mean of the time series. The values r_k in (2.1.11) may be called the *sample* autocorrelation function. To obtain a useful estimate of the autocorrelation function in practice, we would typically need at least 50 observations, and the estimated autocorrelations r_k would be calculated for $k = 0, 1, \ldots, K$, where K was not larger than, say, $N/4$.

The estimated autocorrelation function r_k of the batch data in Figure 2.1 is given in Table 2.1 and plotted in Figure 2.7. The autocorrelation function is characterized by correlations that alternate in sign and tend to damp out with increasing lag. Autocorrelation functions of this kind are not uncommon in production data and can arise because of ''carryover'' effects. In this particular example, a high-yielding batch tended to produce tarry residues, which were not entirely removed from the vessel and adversely affected the yield of the next batch.

Figure 2.7 and the autocorrelations shown in Table 2.1 were generated in R as follows:

```
> Yield = read.table("SeriesF.txt",header=TRUE)
> ACF = acf(Yield,15)
> ACF    % retrieves the values shown in Table 2.1
```

2.1.6 Standard Errors of Autocorrelation Estimates

To identify a model for a time series, using methods to be described in Chapter 6, it is useful to have a rough check on whether ρ_k is effectively zero beyond a certain lag. For this purpose, we can use the following expression for the approximate variance of the estimated autocorrelation coefficient of a stationary normal process given by Bartlett (1946):

$$\text{var}[r_k] \simeq \frac{1}{N}\sum_{v=-\infty}^{\infty}(\rho_v^2 + \rho_{v+k}\rho_{v-k} - 4\rho_k\rho_v\rho_{v-k} + 2\rho_v^2\rho_k^2) \tag{2.1.13}$$

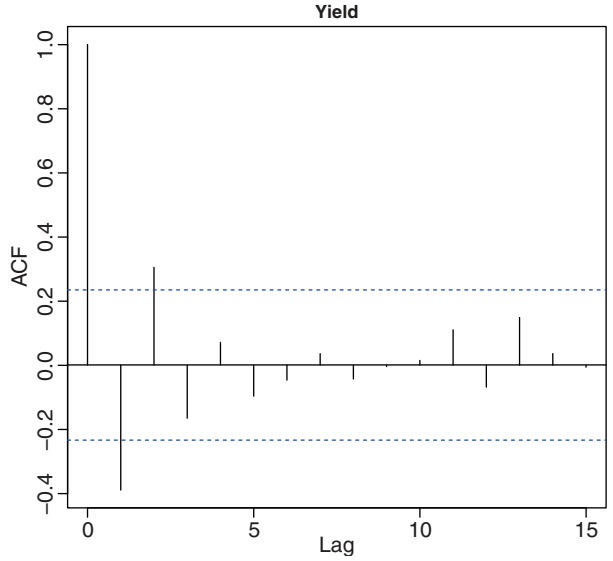


FIGURE 2.7 Estimated autocorrelation function of batch data.

For example, if $\rho_k = \phi^{|k|}$ $(-1 < \phi < 1)$, that is, the autocorrelation function damps out exponentially, (2.1.13) gives

$$\text{var}[r_k] \simeq \frac{1}{N}\left[\frac{(1+\phi^2)(1-\phi^{2k})}{1-\phi^2} - 2k\phi^{2k}\right] \tag{2.1.14}$$

and in particular

$$\text{var}[r_1] \simeq \frac{1}{N}(1-\phi^2)$$

For any process for which all the autocorrelations ρ_v are zero for $v > q$, all terms except the first appearing in the right-hand side of (2.1.13) are zero when $k > q$. Thus, for the variance of the estimated autocorrelation r_k, at lags k greater than some value q beyond which the theoretical autocorrelation function may be deemed to have ''*died out*'', Bartlett's approximation gives

$$\text{var}[r_k] \simeq \frac{1}{N}\left(1 + 2\sum_{v=1}^{q}\rho_v^2\right) \qquad k > q \tag{2.1.15}$$

To use this result in practice, the estimated autocorrelations r_k $(k = 1, 2, \ldots, q)$ are substituted for the theoretical autocorrelations ρ_k, and when this is done, we refer to the square root of (2.1.15) as the *large-lag standard error*. On the assumption that the ρ_k are all zero beyond some lag $k = q$, the large-lag standard error approximates the standard deviation of r_k for suitably large lags $(k > q)$. We will show in Chapter 3 that the moving average (MA) process in (1.2.3) has a correlation structure such that the approximation (2.1.15) applies to this process.

Similar expressions for the approximate covariance between the estimated autocorrelations r_k and r_{k+s} at two different lags k and $k + s$ were also given by Bartlett (1946). In

particular, the large-lag approximation reduces to

$$\text{cov}[r_k, r_{k+s}] \simeq \frac{1}{N} \sum_{v=-q}^{q} \rho_v \rho_{v+s} \qquad k > q \tag{2.1.16}$$

This result shows that care is required in the interpretation of individual autocorrelations because large covariances can exist between neighboring values. This effect can sometimes distort the visual appearance of the sample autocorrelation function, which may fail to damp out according to expectation.

A case of particular interest occurs for $q = 0$, that is, when the ρ_k are taken to be zero for all lags (other than lag 0), and hence the series is completely random or white noise. Then, the standard errors from (2.1.15) for estimated autocorrelations r_k take the simple form

$$\text{se}[r_k] \simeq \frac{1}{\sqrt{N}} \qquad k > 0$$

In addition, in this case the result in (2.1.16) indicates that estimated autocorrelations r_k and r_{k+s} at two different lags are not correlated, and since the r_k are also known to be approximately normally distributed for large N, a collection of estimated autocorrelations for different lags will tend to be independently and normally distributed with mean 0 and variance $1/N$.

Two standard error limits determined under the assumption that the series is completely random are included for the autocorrelation function of the batch data in Figure 2.7. Since N equals 70 in this case, the two standard errors limits are around ± 0.24. The magnitude of the estimated autocorrelation coefficients are clearly inconsistent with the assumption that the series is white noise.

Example. For further illustration, assume that the following estimated autocorrelations were obtained from a time series of length $N = 200$ observations, generated from a stochastic process for which it was *known* that $\rho_1 = -0.4$ and $\rho_k = 0$ for $k \geq 2$:

k	r_k	k	r_k
1	−0.38	6	0.00
2	−0.08	7	0.00
3	0.11	8	0.00
4	−0.08	9	0.07
5	0.02	10	−0.08

On the assumption that the series is completely random, that is, white noise, we have $q = 0$. Then, for *all lags*, (2.1.15) yields

$$\text{var}[r_k] \simeq \frac{1}{N} = \frac{1}{200} = 0.005$$

The corresponding standard error is $0.07 = (0.005)^{1/2}$. Since the value of -0.38 for r_1 is over five times this standard error, it can be concluded that ρ_1 is nonzero. Moreover, the estimated autocorrelations for lags greater than 1 are all small. Therefore, it might be reasonable to ask next whether the series was compatible with a hypothesis (whose relevance will be discussed later) whereby ρ_1 was nonzero, but $\rho_k = 0$ $(k \geq 2)$. Using

(2.1.15) with $q = 1$ and substituting r_1 for ρ_1, the estimated large-lag variance under this assumption is

$$\text{var}[r_k] \simeq \frac{1}{200}[1 + 2(-0.38)^2] = 0.0064 \qquad k > 1$$

yielding a standard error of 0.08. Since the estimated autocorrelations for lags greater than 1 are small compared with this standard error, there is no reason to doubt the adequacy of the model $\rho_1 \neq 0$, $\rho_k = 0$ $(k \geq 2)$.

Remark. The limits shown in Figure 2.7, which assume that the series is white noise, are generated by default in R. Alternative limits, consistent with the assumptions underlying (2.1.15), can be obtained by adding the argument ci.type="ma" to the acf() command.

2.2 SPECTRAL PROPERTIES OF STATIONARY MODELS

2.2.1 Periodogram of a Time Series

Another way of analyzing a time series is based on the assumption that it is made up of sine and cosine waves with different frequencies. A device that uses this idea, introduced by Schuster (1898), is the *periodogram*. The periodogram was originally used to detect and estimate the amplitude of a sine component, of known frequency, buried in noise. We will use it later to provide a check on the randomness of a series (usually, a series of residuals after fitting a particular model), where we consider the possibility that periodic components of unknown frequency may remain in the series.

To illustrate the calculation of the periodogram, suppose that the number of observations $N = 2q + 1$ is odd. We consider fitting the Fourier series model

$$z_t = \alpha_0 + \sum_{i=1}^{q}(\alpha_i c_{it} + \beta_i s_{it}) + e_t \tag{2.2.1}$$

where $c_{it} = \cos(2\pi f_i t)$, $s_{it} = \sin(2\pi f_i t)$, and $f_i = i/N$, which is the ith harmonic of the fundamental frequency $1/N$ associated with the ith sine wave component in (2.2.1) with frequency f_i and period $1/f_i = N/i$. The least squares estimates of the coefficients α_0 and (α_i, β_i) will be

$$a_0 = \bar{z} \tag{2.2.2}$$

$$a_i = \frac{2}{N}\sum_{t=1}^{N} z_t c_{it} \tag{2.2.3}$$

$$i = 1, 2, \ldots, q$$

$$b_i = \frac{2}{N}\sum_{t=1}^{N} z_t s_{it} \tag{2.2.4}$$

since $\sum_{t=1}^{N} c_{it}^2 = \sum_{t=1}^{N} s_{it}^2 = N/2$, and all terms in (2.2.1) are mutually orthogonal over $t = 1, \ldots, N$. The periodogram then consists of the $q = (N-1)/2$ values

$$I(f_i) = \frac{N}{2}(a_i^2 + b_i^2) \qquad i = 1, 2, \ldots, q \tag{2.2.5}$$

where $I(f_i)$ is called the *intensity* at frequency f_i. When N is even, we set $N = 2q$ and (2.2.2)–(2.2.5) apply for $i = 1, 2, \ldots, (q-1)$, but

$$a_q = \frac{1}{N}\sum_{t=1}^{N}(-1)^t z_t$$
$$b_q = 0$$

and

$$I(f_q) = I(0.5) = N a_q^2$$

Note that the highest frequency is 0.5 cycle per time interval because the smallest period is two intervals.

2.2.2 Analysis of Variance

In an analysis of variance table associated with the fitted regression (2.2.1), when N is odd, we can isolate $(N-1)/2$ pairs of degrees of freedom, after eliminating the mean. These are associated with the pairs of coefficients $(a_1, b_1), (a_2, b_2), \ldots, (a_q, b_q)$, and hence with the frequencies $1/N, 2/N, \ldots, q/N$. The periodogram $I(f_i) = (N/2)(a_i^2 + b_i^2)$ is seen to be simply the ''sum of squares'' associated with the pair of coefficients (a_i, b_i) and hence with the frequency $f_i = i/N$ or period $p_i = N/i$. Thus,

$$\sum_{t=1}^{n}(z_t - \bar{z})^2 = \sum_{i=1}^{q} I(f_i) \tag{2.2.6}$$

When N is even, there are $(N-2)/2$ pairs of degrees of freedom and a further single degree of freedom associated with the coefficient a_q.

If the series were truly random, containing no systematic sinusoidal component, that is,

$$z_t = \alpha_0 + e_t$$

with α_0 the fixed mean, and the e's independent and normal, with mean zero and variance σ^2, each component $I(f_i)$ would have expectation $2\sigma^2$ and would be distributed[1] as $\sigma^2\chi^2(2)$, independently of all other components. By contrast, if the series contained a systematic sine component having frequency f_i, amplitude A, and phase angle F, so that

$$z_t = \alpha_0 + \alpha\cos(2\pi f_i t) + \beta\sin(2\pi f_i t) + e_t$$

with $A\sin F = \alpha$ and $A\cos F = \beta$, the sum of squares $I(f_i)$ would tend to be inflated since its expected value would be $2\sigma^2 + N(\alpha^2 + \beta^2)/2 = 2\sigma^2 + NA^2/2$.

In practice, it is unlikely that the frequency f of an unknown systematic sine component would exactly match any of the frequencies f_i for which intensities have been calculated. In this case the periodogram would show an increase in the intensities in the immediate vicinity of f.

[1]It is to be understood that $\chi^2(m)$ refers to a random variable having a chi-square distribution with m degrees of freedom, defined explicitly, for example, in Appendix A7.1.

TABLE 2.2 Mean Monthly Temperatures for Central England in 1964

t	z_t	c_{1t}	t	z_t	c_{1t}
1	3.4	0.87	7	16.1	−0.87
2	4.5	0.50	8	15.5	−0.50
3	4.3	0.00	9	14.1	0.00
4	8.7	−0.50	10	8.9	0.50
5	13.3	−0.87	11	7.4	0.87
6	13.8	−1.00	12	3.6	1.00

Example. A large number of observations would generally be used in calculation of the periodogram. However, to illustrate the details of the calculation, we use the set of 12 mean monthly temperatures (in degrees Celsius) for central England during 1964, given in Table 2.2. The table gives $c_{it} = \cos(2\pi t/12)$, which is required in the calculation of a_1, obtained from

$$\begin{aligned} a_1 &= \frac{1}{6}[(3.4)(0.87) + \cdots + (3.6)(1.00)] \\ &= -5.30 \end{aligned}$$

The values of the a_i, b_i, $i = 1, 2, \ldots, 6$, are given in Table 2.3 and yield the analysis of variance of Table 2.4. As would be expected, the major component of these temperature data has a period of 12 months, that is, a frequency of 1/12 cycle per month.

2.2.3 Spectrum and Spectral Density Function

For completeness, we add here a brief discussion of the spectrum and spectral density function. The use of these important tools is described more fully by Jenkins and Watts (1968), Bloomfield (2000), and Shumway and Stoffer (2011, Chapter 4), among others. We do not apply them to the analysis of time series in this book, and this section can be omitted on first reading.

Sample Spectrum. The definition of the periodogram in (2.2.5) assumes that the frequencies $f_i = i/N$ are harmonics of the fundamental frequency $1/N$. By way of introduction to the spectrum, we relax this assumption and allow the frequency f to vary continuously

TABLE 2.3 Amplitudes of Sines and Cosines at Different Harmonics for Temperature Data

i	a_i	b_i
1	−5.30	−3.82
2	0.05	0.17
3	0.10	0.50
4	0.52	−0.52
5	0.09	−0.58
6	−0.30	

TABLE 2.4 Analysis of Variance Table for Temperature Data

	Frequency				
i	f_i	Period	Periodogram $I(f_i)$	Degrees of Freedom	Mean Square
1	1/12	12	254.96	2	127.48
2	1/6	6	0.19	2	0.10
3	1/4	4	1.56	2	0.78
4	1/3	3	3.22	2	1.61
5	5/12	12/5	2.09	2	1.05
6	1/2	2	1.08	1	1.08
			263.10	11	23.92

in the range of 0–0.5 cycle. The definition (2.2.5) of the periodogram may be modified to

$$I(f) = \frac{N}{2}(a_f^2 + b_f^2) \qquad 0 \le f \le \frac{1}{2} \tag{2.2.7}$$

and $I(f)$ is then referred to as the *sample spectrum* (Jenkins and Watts, 1968). Like the periodogram, it can be used to detect and estimate the amplitude of a sinusoidal component of unknown frequency f buried in noise and is, indeed, a more appropriate tool for this purpose if it is known that the frequency f is not harmonically related to the length of the series. Moreover, it provides a starting point for the theory of spectral analysis, using a result given in Appendix A2.1. This result shows that the sample spectrum $I(f)$ and the estimate c_k of the autocovariance function are linked by the important relation

$$I(f) = 2\left[c_0 + 2\sum_{k=1}^{N-1} c_k \cos(2\pi f k)\right] \qquad 0 \le f \le \frac{1}{2} \tag{2.2.8}$$

That is, the sample spectrum is the Fourier cosine transform of the estimate of the autocovariance function.

Spectrum. The periodogram and sample spectrum are appropriate tools for analyzing time series made up of mixtures of sine and cosine waves, at *fixed* frequencies buried in noise. However, stationary time series of the kind described in Section 2.1 are characterized by random changes of frequency, amplitude, and phase. For this type of series, the sample spectrum $I(f)$ fluctuates wildly and is not capable of any meaningful interpretation.

However, suppose that the sample spectrum was calculated for a time series of N observations, which is a realization of a stationary normal process. As already mentioned, such a process would not have any cosine or sine deterministic components, but we could formally carry through the Fourier analysis and obtain values of (a_f, b_f) for any given frequency f. If repeated realizations of N observations were taken from the stochastic process, we could build up a population of values for a_f, b_f, and $I(f)$. Thus, we could calculate the mean value of $I(f)$ in repeated realizations of size N, namely,

$$E[I(f)] = 2\left[E[c_0] + 2\sum_{k=1}^{N-1} E[c_k]\cos(2\pi f k)\right] \tag{2.2.9}$$

For large N, it may be shown (e.g., Jenkins and Watts, 1968) that the mean value of the estimate c_k of the autocovariance coefficient in repeated realizations tends to the theoretical autocovariance γ_k, that is,

$$\lim_{N\to\infty} E[c_k] = \gamma_k$$

On taking the limit of (2.2.9) as N tends to infinity, the *power spectrum* $p(f)$ is defined by

$$p(f) = \lim_{N\to\infty} E[I(f)] = 2\left[\gamma_0 + 2\sum_{k=1}^{\infty}\gamma_k \cos(2\pi f k)\right] \qquad 0 \le f \le \frac{1}{2} \tag{2.2.10}$$

We note that since

$$\begin{aligned} |p(f)| &\le 2\left[|\gamma_0| + 2\sum_{k=1}^{\infty}|\gamma_k||\cos(2\pi f k)|\right] \\ &\le 2\left(|\gamma_0| + 2\sum_{k=1}^{\infty}|\gamma_k|\right) \end{aligned} \tag{2.2.11}$$

a sufficient condition for the spectrum to converge is that γ_k damps out rapidly enough for the series (2.2.11) to converge. *Since the power spectrum is the Fourier cosine transform of the autocovariance function*, knowledge of the autocovariance function is mathematically equivalent to knowledge of the spectrum, and vice versa. From now on, we refer to the power spectrum as simply the spectrum.

On integrating (2.2.10) between the limits 0 and $\frac{1}{2}$, the variance of the process z_t is

$$\gamma_0 = \sigma_z^2 = \int_0^{1/2} p(f)\, df \tag{2.2.12}$$

Hence, in the same way that the periodogram $I(f)$ shows how the variance (2.2.6) of a series, consisting of mixtures of sines and cosines, is distributed between the various distinct harmonic frequencies, the spectrum $p(f)$ shows how the variance of a stochastic process is distributed between a continuous range of frequencies. One can interpret $p(f)\,df$ as measuring approximately the variance of the process in the frequency range of f to $f + df$. In addition, from the definition in (2.2.10), the spectral representation for the autocovariance function $\{\gamma_k\}$ can be obtained as

$$\gamma_k = \int_0^{1/2} \cos(2\pi f k) p(f)\, df$$

which together with (2.2.10) directly exhibits the one-to-one correspondence between the power spectrum and the autocovariance function of a process. Conversely, since the γ_k form a positive-definite sequence, provided the series (2.2.11) converges, it follows from Herglotz's theorem (see, e.g., Loève, 1977) that a unique function $p(f)$ exists such that γ_k have the spectral representation $\gamma_k = \frac{1}{2}\int_{-1/2}^{1/2} e^{i2\pi f k} p(f)\, df$. Consequently, the power spectrum $p(f)$ of a stationary process, for which (2.2.11) converges, can be defined as this unique function, which is guaranteed to exist and must have the form of the right-hand side of (2.2.10) by the spectral representation.

The fundamental property of the spectrum that $p(f) \geq 0$ for all $0 \leq f \leq \frac{1}{2}$ follows from $I(f) \geq 0$ and the definition in (2.2.10). In fact, a function $p(f)$ defined on $0 \leq f \leq \frac{1}{2}$ can be the spectrum of a stationary process if and only if it satisfies $p(f) \geq 0$ for $0 \leq f \leq \frac{1}{2}$ and $\int_0^{1/2} p(f)\, df \leq \infty$. Conversely, a sequence $\{\gamma_k\}_{k=0}^{\infty}$ can be the autocovariance function of a stationary process if and only if $\{\gamma_k\}$ is a nonnegative-definite sequence, and this is equivalent to the condition that $p(f) \geq 0$, $0 \leq f \leq \frac{1}{2}$, with $p(f)$ defined by (2.2.10).

Spectral Density Function. It is sometimes more convenient to base the definition (2.2.10) of the spectrum on the autocorrelations ρ_k rather than on the autocovariances γ_k. The resulting function

$$\begin{aligned} g(f) &= \frac{p(f)}{\sigma_z^2} \\ &= 2\left[1 + 2\sum_{k=1}^{\infty} \rho_k \cos(2\pi f k)\right] \qquad 0 \leq f \leq \frac{1}{2} \end{aligned} \tag{2.2.13}$$

is called the *spectral density function*. Using (2.2.12), it is seen that the spectral density function has the property

$$\int_0^{1/2} g(f)\, df = 1$$

Since $g(f)$ is also positive, it has the same properties as an ordinary probability density function. This analogy extends to the estimation properties of these two functions, as we discuss next.

Estimation of the Spectrum. One would expect that an estimate of the spectrum could be obtained from (2.2.10), by replacing the theoretical autocovariances γ_k with their estimates c_k. Because of (2.2.8), this corresponds to taking the sample spectrum as an estimate of $p(f)$. However, it can be shown (e.g., Jenkins and Watts, 1968) that the sample spectrum of a stationary time series fluctuates violently about the theoretical spectrum. An intuitive explanation of this fact is that the sample spectrum corresponds to using an interval, in the frequency domain, whose width is too small. This is analogous to using too small a group interval for the histogram when estimating an ordinary probability distribution. By using a modified or *smoothed* estimate

$$\hat{p}(f) = 2\left[c_0 + 2\sum_{k=1}^{N-1} \lambda_k c_k \cos(2\pi f k)\right] \tag{2.2.14}$$

where the λ_k are suitably chosen weights called a *lag window*, it is possible to increase the *bandwidth* of the estimate and to obtain a smoother estimate of the spectrum. The weights λ_k in (2.2.14) are typically chosen so that they die out to zero for lags $k > M$, where M is known as the truncation point and $M < N$ is moderately small in relation to series length N. As an alternative computational form, one can also obtain an estimate of the spectrum smoother than the sample spectrum $I(f)$ by forming a weighted average of a number of periodogram values $I(f_{i+j})$ in a small neighborhood of frequencies around a

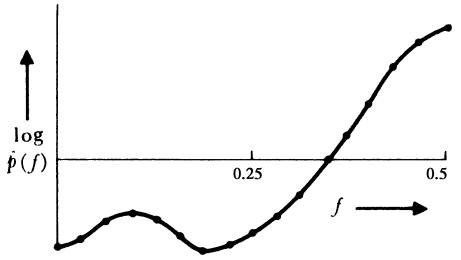


FIGURE 2.8 Estimated power spectrum of batch data.

given frequency f_i. Specifically, a smoothed periodogram estimator of $p(f_i)$ takes the form

$$\hat{p}(f_i) = \sum_{j=-m}^{m} W(f_j) I\left(f_i + \frac{j}{N}\right)$$

where $\sum_{j=-m}^{m} W(f_j) = 1$, the symmetric weighting function $W(f_i)$ is referred to as the *spectral window*, and m is chosen to be much smaller than $N/2$.

Figure 2.8 shows an estimate of the spectrum of the batch data. It is seen that most of the variance of the series is concentrated at high frequencies. This is due to the rapid oscillations in the original series, shown in Figure 2.1.

Remark. The command spectrum() can be used to estimate the power spectrum in R. To use this command, a smoothing window must be specified; see help(spectrum) and the references therein for details. The following command will generate a graph roughly similar to Figure 2.8:

```
spectrum(Yield,spans=c(7,7),taper=0)
```

As an alternative, the R program spec.ar() fits an autoregressive model of order p to the series and computes the spectral density of the fitted model. The lag order p is selected using a model selection criterion such as the AIC to be discussed in Chapter 6.

2.2.4 Simple Examples of Autocorrelation and Spectral Density Functions

For illustration, we now show equivalent representations of two simple stationary stochastic processes based on:

1. Their theoretical models
2. Their theoretical autocorrelation functions
3. Their theoretical spectra

Consider the two processes

$$z_t = 10 + a_t + a_{t-1} \qquad z_t = 10 + a_t - a_{t-1}$$

where $a_t, a_{t-1}, \ldots$ are a sequence of uncorrelated normal random variables with mean zero and variance σ_a^2, that is, Gaussian white noise. From the result in Section 2.1.3 on

stationarity of linear functions, it is clear that the two processes above are stationary. Using the definition (2.1.5),

$$\gamma_k = \text{cov}[z_t, z_{t+k}] = E[(z_t - \mu)(z_{t+k} - \mu)]$$

where $E[z_t] = E[z_{t+k}] = \mu = 10$, and the autocovariances of these two stochastic processes are obtained from

$$\begin{aligned}\gamma_k &= \text{cov}[a_t + a_{t-1}, a_{t+k} + a_{t+k-1}] \\ &= \text{cov}[a_t, a_{t+k}] + \text{cov}[a_t, a_{t+k-1}] + \text{cov}[a_{t-1}, a_{t+k}] + \text{cov}[a_{t-1}, a_{t+k-1}]\end{aligned}$$

and $\gamma_k = \text{cov}[a_t - a_{t-1}, a_{t+k} - a_{t+k-1}]$, respectively. Hence, the autocovariances are

$$\gamma_k = \begin{cases} 2\sigma_a^2 & k = 0 \\ \sigma_a^2 & k = 1 \\ 0 & k \geq 2 \end{cases} \qquad \gamma_k = \begin{cases} 2\sigma_a^2 & k = 0 \\ -\sigma_a^2 & k = 1 \\ 0 & k \geq 2 \end{cases}$$

Thus, the theoretical autocorrelation functions are

$$\rho_k = \begin{cases} 0.5 & k = 1 \\ 0.0 & k \geq 2 \end{cases} \qquad \rho_k = \begin{cases} -0.5 & k = 1 \\ 0.0 & k \geq 2 \end{cases}$$

and using (2.2.13), the theoretical spectral density functions are

$$g(f) = 2[1 + \cos(2\pi f)] \qquad g(f) = 2[1 - \cos(2\pi f)]$$

The autocorrelation functions and spectral density functions are plotted in Figure 2.9 together with a sample time series from each process.

1. It should be noted that for these two stationary processes, knowledge of either the autocorrelation function or the spectral density function, with the mean and variance of the process, is equivalent to knowledge of the model (given the normality assumption).
2. It will be seen that the autocorrelation function reflects one aspect of the behavior of the series. The comparatively smooth nature of the first series is accounted for by the positive association between successive values. The alternating tendency of the second series, in which positive deviations usually follow negative ones, is accounted for by the negative association between successive values.
3. The spectral density throws light on a different but equivalent aspect. The predominance of low frequencies in the first series and high frequencies in the second is shown by the spectra.

Remark. The two models considered in Figure 2.9 are special cases of the moving average model defined in (1.2.3). Specifically, the models are first-order moving average, or MA(1), models with parameters $\theta = -1$ and $\theta = +1$, respectively. As such, they are also special cases of the more general autoregressive integrated moving average (ARIMA) model defined in (1.2.7), where the order now is (0, 0, 1). Figure 2.9 was generated in R by taking advantage of special functions for simulating ARIMA processes and for computing the

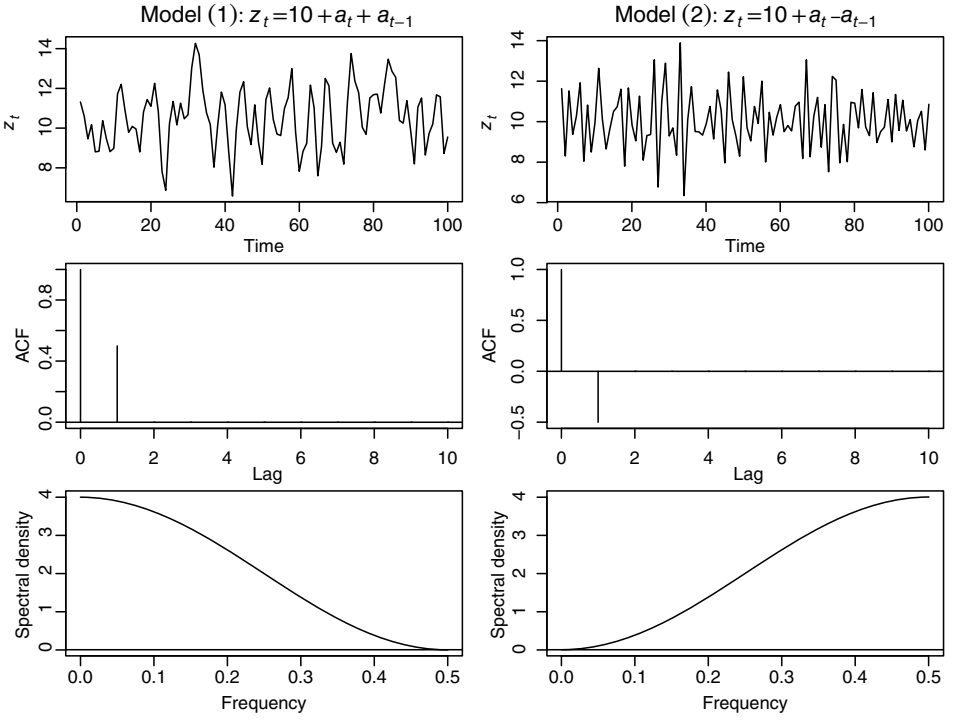


FIGURE 2.9 Two simple stochastic models with their corresponding theoretical autocorrelation functions and spectral density functions.

theoretical autocorrelation function and power spectrum for these processes. The function arima.sim() simulates a time series from a specified model, while ARMAacf() computes its theoretical autocorrelation. Both functions are available in the stats library of R. The TSA library includes a function ARMAspec() that computes and plots the theoretical spectrum of an autoregressive–moving average (ARMA) process. The commands used to generate Figure 2.9 are given below. Note, however, that the MA parameters are entered as +1 and −1, since R uses a definition that has positive signs of the MA parameters in (1.2.3).

```
> library(TSA)
> set.seed(12345)
> par(mfrow=c(3,2)) % Specifies panels in three rows and two columns
> plot(10+arima.sim(list(order=c(0,0,1), ma = +1.0), n=100),ylab =
  expression(z[t]),main=(expression(Model~(1):z[t] == 10+a[t]+a[t-1])))
> plot(10+arima.sim(list(order=c(0,0,1), ma = -1.0), n=100),ylab =
  expression(z[t]),main=(expression(Model~(2):z[t] == 10+a[t]-a[t-1])))
> plot(ARMAacf(ar=0,ma=1.0,10),type="h",x=(0:10),xlab="lag",ylab="ACF")
> abline(h=0)
> plot(ARMAacf(ar=0,ma=-1.0,10),type="h",x=(0:10),xlab="lag",ylab="ACF")
> abline(h=0)
> ARMAspec(model=list(ma=1.0),freq=seq(0,0.5,0.001),plot=TRUE)
> ARMAspec(model=list(ma=-1.0),freq=seq(0,0.5,0.001),plot=TRUE)
```

2.2.5 Advantages and Disadvantages of the Autocorrelation and Spectral Density Functions

Because the autocorrelation function and the spectrum are transforms of each other, they are mathematically equivalent, and therefore any discussion of their advantages and disadvantages turns not on mathematical questions but on the representational value. Because, as we have seen, each sheds light on a different aspect of the data, they should be regarded not as rivals but as allies. Each contributes something to an understanding of the stochastic process in question.

The obtaining of sample estimates of the autocorrelation function and of the spectrum are nonstructural approaches, analogous to the representation of an empirical distribution function by a histogram. They are both ways of letting data from stationary series ''speak for themselves'' and provide a first step in the analysis of time series, just as a histogram can provide a first step in the distributional analysis of data, pointing the way to some parametric model on which subsequent analysis will be based.

Parametric time series models such as those of Section 2.2.4, are not necessarily associated with a simple autocorrelation function or a simple spectrum. Working with either of these nonstructural methods, we may be involved in the estimation of many lag correlations and many spectral ordinates, even when a parametric model containing only one or two parameters could represent the data. Each correlation and each spectral ordinate is a parameter to be estimated, so that these nonstructural approaches might be very prodigal with parameters, when the approach via the model could be parsimonious. On the other hand, initially, we probably do not know what type of model may be appropriate, and initial use of one or the other of these nonstructural approaches is necessary to *identify* the type of model that is needed (in the same way that plotting a histogram helps to indicate which family of distributions may be appropriate). The choice between the spectrum and the autocorrelation function as a tool in model building depends upon the nature of the models that turn out to be practically useful. The models that we have found useful, which we consider in later chapters of this book, are simply described in terms of the autocorrelation function, and it is this tool that we will employ for model specification.

APPENDIX A2.1 LINK BETWEEN THE SAMPLE SPECTRUM AND AUTOCOVARIANCE FUNCTION ESTIMATE

Here, we derive the result (2.2.8):

$$I(f) = 2\left[c_0 + 2\sum_{k=1}^{N-1} c_k \cos(2\pi f k)\right] \qquad 0 \le f \le \frac{1}{2}$$

which links the sample spectrum $I(f)$ and the estimate c_k of the autocovariance function. Suppose that the least square estimates a_f and b_f of the cosine and sine components, at frequency f, in a series are combined according to $d_f = a_f - ib_f$, where $i = -\sqrt{-1}$; then

$$\begin{aligned} I(f) &= \frac{N}{2}(a_f - ib_f)(a_f + ib_f) \\ &= \frac{N}{2} d_f d_f^* \end{aligned} \tag{A2.1.1}$$

where d_f^* is the complex conjugate of d_f. Then, using (2.2.3) and (2.2.4), we obtain

$$\begin{aligned} d(f) &= \frac{2}{N}\sum_{t=1}^{N} z_t[\cos(2\pi ft) - i\sin(2\pi ft)] \\ &= \frac{2}{N}\sum_{t=1}^{N} z_t e^{-i2\pi ft} \\ &= \frac{2}{N}\sum_{t=1}^{N}(z_t - \bar{z})e^{-i2\pi ft} \end{aligned} \tag{A2.1.2}$$

Substituting (A2.1.2) in (A2.1.1) yields

$$I(f) = \frac{2}{N}\sum_{t=1}^{N}\sum_{t'=1}^{N}(z_t - \bar{z})(z_{t'} - \bar{z})e^{-i2\pi f(t-t')} \tag{A2.1.3}$$

Since

$$c_k = \frac{1}{N}\sum_{t=1}^{N-k}(z_t - \bar{z})(z_{t+k} - \bar{z})$$

the transformation $k = t - t'$ transforms (A2.1.3) into the following required result:

$$\begin{aligned} I(f) &= 2\sum_{k=-N+1}^{N-1} c_k e^{-i2\pi fk} \\ &= 2\left[c_0 + 2\sum_{k=1}^{N-1} c_k \cos(2\pi fk)\right] \qquad 0 \le f \le \frac{1}{2} \end{aligned}$$

EXERCISES

2.1. The following are temperature measurements z_t made every minute on a chemical reactor:

200, 202, 208, 204, 204, 207, 207, 204, 202, 199, 201, 198, 200,
202, 203, 205, 207, 211, 204, 206, 203, 203, 201, 198, 200, 206,
207, 206, 200, 203, 203, 200, 200, 195, 202, 204.207, 206, 200

(a) Plot the time series.
(b) Plot z_{t+1} versus z_t.
(c) Plot z_{t+2} versus z_t.
After inspecting the graphs, do you think that the series is autocorrelated?

2.2. State whether or not a stationary stochastic process can have the following values of autocorrelations:

(a) $\rho_1 = 0.80,\ \rho_2 = 0.55,\ \rho_k = 0,$ for $k > 2$

(b) $\rho_1 = 0.80,\ \rho_2 = 0.28,\ \rho_k = 0,$ for $k > 2$

2.3. Two stationary stochastic processes z_{1t} and z_{2t} have the following autocovariance functions:

$$z_{1t}: \quad \gamma_0 = 0.5,\ \gamma_1 = 0.2,\ \gamma_j = 0 \quad (j \geq 2)$$
$$z_{2t}: \quad \gamma_0 = 2.30,\ \gamma_1 = -1.43,\ \gamma_2 = 0.30,\ \gamma_j = 0 \quad (j \geq 3)$$

Calculate the autocovariance function of the process $z_{3t} = z_{1t} + 2z_{2t}$ and verify that it is a valid stationary process.

2.4. Calculate $c_0, c_1, c_2, c_3, r_1, r_2, r_3$ for the series given in Exercise 2.1. Make a graph of $r_k, k = 0, 1, 2, 3$.

2.5. On the assumption that $\rho_j = 0$ for $j > 2$, obtain the following:

(a) Approximate standard errors for r_1, r_2, and $r_j, j > 2$.

(b) The approximate correlation between r_4 and r_5.

2.6. The annual sales of mink furs by a North American company during 1911–1950 are included as Series N in Part Five of this book. The series is also available at http://pages.stat.wisc.edu/ reinsel/bjr-data/.

(a) Plot the time series using R. Calculate and plot the sample autocorrelation function of the series.

(b) Repeat the analysis in part (a) for the logarithm of the series. Do you see an advantage in using the log transformation in this case?

2.7. Repeat the calculations in Exercise 2.6 for the annual sunspot series given as Series E in Part Five of this book. Use a square root transformation of the data in part (b) in Exercise 2.6. (*Note:* This series is also available for a slightly longer time period as series sunspot.year in the datasets package of R).

2.8. Calculate and plot the theoretical autocorrelation function and the spectral density function for the AR(1) process $z_t = 0.95z_{t-1} + a_t$. (*Hint:* See the R code provided for Figure 2.9). Based on the results, how would you expect a time series generated from this model to fluctuate relative to its mean?

2.9. Calculate and plot the theoretical autocorrelation function and the spectral density function for the AR(2) process $z_t + 0.35z_{t-1} - 0.20z_{t-2} = a_t$.

2.10. Simulate a time series of length $N = 300$ from the AR(2) model specified in Exercise 2.9 and plot the resulting series.

(a) Estimate and plot the autocorrelation function for the simulated series. Compare the results with the theoretical autocorrelation function derived in Exercise 2.9.

(b) Repeat the calculations performed above for a series of length $N = 70$ generated from the same process and compare the results with those for $N = 200$.

(c) Do the estimated autocorrelation functions derived above show any similarity to autocorrelation function of the chemical yield series shown in Figure 2.7. If so, what would you conclude?

2.11. Using the data of Exercise 2.1, calculate the periodogram for periods 36, 18, 12, 9, 36/5, and 6 and construct an analysis of variance table showing the mean squares associated with these periods and the residual mean square.

2.12. A *circular* stationary stochastic process with period N is defined by $z_t = z_{t+N}$.

(a) Show that (see, e.g., Brockwell and Davis, 1991; Fuller, 1996; Jenkins and Watts, 1968) when $N = 2n$, the latent roots of the $N \times N$ autocorrelation matrix of z_t are

$$\lambda_k = 1 + 2\sum_{i=1}^{n-1} \rho_i \cos\left(\frac{\pi i k}{n}\right) + \rho_n \cos(\pi k)$$

$k = 1, 2, \ldots, N$ and the latent vectors corresponding to λ_k, λ_{N-k} (with $\lambda_k = \lambda_{N-k}$) are

$$\ell'_k = \left(\cos\left(\frac{\pi k}{n}\right), \cos\left(\frac{2\pi k}{n}\right), \ldots, \cos(2\pi k)\right)$$
$$\ell'_{N-k} = \left(\sin\left(\frac{\pi k}{n}\right), \sin\left(\frac{2\pi k}{n}\right), \ldots, \sin(2\pi k)\right)$$

(b) Verify that as N tends to infinity, with k/N fixed, λ_k tends to $g(k/N)/2$, where $g(f)$ is the spectral density function, showing that in the limit the latent roots of the autocorrelation matrix trace out the spectral curve.

3

LINEAR STATIONARY MODELS

In this chapter, we describe a general linear stochastic model that assumes that the time series is generated by a linear aggregation of random shocks. For practical representation, it is desirable to employ models that use parameters parsimoniously. Parsimony may often be achieved by representation of the linear process in terms of a small number of autoregressive–moving average (ARMA) terms. The properties of the resulting ARMA models are discussed in preparation for their use in model building in subsequent chapters.

3.1 GENERAL LINEAR PROCESS

3.1.1 Two Equivalent Forms for the Linear Process

In Section 1.2.1, we discussed the representation of a stochastic process as the output from a linear filter, whose input is white noise a_t, that is,

$$\begin{aligned} \tilde{z}_t &= a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots \\ &= a_t + \sum_{j=1}^{\infty} \psi_j a_{t-j} \end{aligned} \tag{3.1.1}$$

where $\tilde{z}_t = z_t - \mu$ is the deviation of the process from some origin, or from its mean, if the process is stationary. The *general linear process* (3.1.1) allows us to represent $\tilde{z}_t$ as a weighted sum of present and past values of the ''white noise'' process a_t. Important early references on the development of linear stochastic models include Yule (1927), Walker (1931), Slutsky (1937), Wold (1938), Kendall (1945), Bartlett (1946), Quenouille (1952, 1957), Doob (1953), Grenander and Rosenblatt (1957), Hannan (1960), Robinson (1967),

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

among others. The usefulness of these models is well-documented in subsequent literature. The white noise process a_t may be regarded as a *series of shocks* that drive the system. It consists of a sequence of uncorrelated random variables with mean zero and constant variance, that is,

$$E[a_t] = 0 \qquad \text{var}[a_t] = \sigma_a^2$$

Since the random variables a_t are assumed uncorrelated, it follows that their autocovariance function is

$$\gamma_k = E[a_t a_{t+k}] = \begin{cases} \sigma_a^2 & k = 0 \\ 0 & k \neq 0 \end{cases} \tag{3.1.2}$$

Thus, the autocorrelation function of white noise has a particularly simple form

$$\rho_k = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases} \tag{3.1.3}$$

A fundamental result in the development of stationary processes is that of Wold (1938), who established that any zero-mean purely nondeterministic stationary process $\tilde{z}_t$ possesses a linear representation as in (3.1.1) with $\sum_{j=0}^{\infty} \psi_j^2 < \infty$. The a_t are uncorrelated with common variance σ_a^2 but *need not be independent*. We will reserve the term *linear processes* for processes $\tilde{z}_t$ of the form of (3.1.1) in which the a_t are independent random variables.

For $\tilde{z}_t$ defined by (3.1.1) to represent a valid stationary process, it is necessary for the coefficients ψ_j to be *absolutely summable*, that is, for $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Under suitable conditions (see Koopmans, 1974, p. 254), $\tilde{z}_t$ is also a weighted sum of past $\tilde{z}_t$'s and an added shock a_t, that is,

$$\begin{aligned} \tilde{z}_t &= \pi_1 \tilde{z}_{t-1} + \pi_2 \tilde{z}_{t-2} + \cdots + a_t \\ &= \sum_{j=1}^{\infty} \pi_j \tilde{z}_{t-j} + a_t \end{aligned} \tag{3.1.4}$$

In this alternative form, the current deviation $\tilde{z}_t$ from the level μ may be thought of as being ''regressed'' on past deviations $\tilde{z}_{t-1}, \tilde{z}_{t-2}, \ldots$ of the process.

Relationships between the ψ Weights and the π Weights. The relationships between the ψ weights and the π weights may be obtained by using the previously defined *backward shift operator* B, such that

$$Bz_t = z_{t-1} \quad \text{and hence} \quad B^j z_t = z_{t-j}$$

Later, we will also need to use the forward shift operator $F = B^{-1}$, such that

$$Fz_t = z_{t+1} \quad \text{and} \quad F^j z_t = z_{t+j}$$

As an example of the use of the operator B, consider the following model

$$\tilde{z}_t = a_t - \theta a_{t-1} = (1 - \theta B) a_t$$

in which $\psi_1 = -\theta$, $\psi_j = 0$ for $j > 1$. Expressing a_t in terms of the $\tilde{z}_t$'s, we obtain

$$(1 - \theta B)^{-1}\tilde{z}_t = a_t$$

Hence, for $|\theta| < 1$,

$$(1 + \theta B + \theta^2 B^2 + \theta^3 B^3 + \cdots)\tilde{z}_t = a_t$$

and the deviation $\tilde{z}_t$ expressed in terms of previous deviations, as in (3.1.4), is

$$\tilde{z}_t = -\theta\tilde{z}_{t-1} - \theta^2\tilde{z}_{t-2} - \theta^3\tilde{z}_{t-3} - \cdots + a_t$$

so that for this model, $\pi_j = -\theta^j$.

Using the backshift operator B, the model (3.1.1) can be written as

$$\tilde{z}_t = \left(1 + \sum_{j-1}^{\infty} \psi_j B^j\right) a_t$$

or

$$\tilde{z}_t = \psi(B)a_t \tag{3.1.5}$$

where

$$\psi(B) = 1 + \sum_{j=1}^{\infty} \psi_j B^j = \sum_{j=0}^{\infty} \psi_j B^j$$

with $\psi_0 = 1$. As mentioned in Section 1.2.1, $\psi(B)$ is called the *transfer function* of the linear filter relating $\tilde{z}_t$ to a_t. It can be regarded as the *generating function* of the ψ weights, with B now treated simply as a variable whose jth power is the coefficient of ψ_j.

Similarly, (3.1.4) may be written as

$$\left(1 - \sum_{j=1}^{\infty} \pi_j B^j\right)\tilde{z}_t = a_t$$

or

$$\pi(B)\tilde{z}_t = a_t \tag{3.1.6}$$

Thus,

$$\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$$

is the generating function of the π weights. After operating on both sides of this expression by $\psi(B)$, we obtain

$$\psi(B)\pi(B)\tilde{z}_t = \psi(B)a_t = \tilde{z}_t$$

Hence, $\psi(B)\pi(B) = 1$, so that

$$\pi(B) = \psi^{-1}(B) \tag{3.1.7}$$

This relationship may be used to derive the π weights, knowing the ψ weights, and vice versa.

3.1.2 Autocovariance Generating Function of a Linear Process

A basic data analysis tool for identifying models in Chapter 6 will be the autocorrelation function. Therefore, it is important to know the autocorrelation function of a linear process. It is shown in Appendix A3.1 that the autocovariance function of the linear process (3.1.1) is given by

$$\gamma_k = \sigma_a^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k} \tag{3.1.8}$$

In particular, by setting $k = 0$, we find that its variance is

$$\gamma_0 = \sigma_z^2 = \sigma_a^2 \sum_{j=0}^{\infty} \psi_j^2 \tag{3.1.9}$$

It follows that the stationarity condition of absolute summability of the coefficients ψ_j, $\sum_{j=0}^{\infty} |\psi_j| < \infty$, implies that the series on the right of this equation converges, and hence guarantees that the process will have a finite variance.

Another way of obtaining the autocovariances of a linear process is via the *autocovariance generating function*

$$\gamma(B) = \sum_{k=-\infty}^{\infty} \gamma_k B^k \tag{3.1.10}$$

where γ_0, the variance of the process, is the coefficient of $B^0 = 1$, while γ_k, the autocovariance of lag k, is the coefficient of both B^j and $B^{-j} = F^j$. It is shown in Appendix A3.1 that

$$\gamma(B) = \sigma_a^2 \psi(B)\psi(B^{-1}) = \sigma_a^2 \psi(B)\psi(F) \tag{3.1.11}$$

For example, suppose that $\tilde{z}_t = a_t - \theta a_{t-1} = (1 - \theta B)a_t$ so that $\psi(B) = (1 - \theta B)$. Then,

$$\begin{aligned} \gamma(B) &= \sigma_a^2 (1 - \theta B)(1 - \theta B^{-1}) \\ &= \sigma_a^2 [-\theta B^{-1} + (1 + \theta^2) - \theta B] \end{aligned}$$

Comparing with (3.1.10), the autocovariances are

$$\begin{aligned} \gamma_0 &= (1 + \theta^2)\sigma_a^2 \\ \gamma_1 &= -\theta \sigma_a^2 \\ \gamma_k &= 0 \qquad k \geq 2 \end{aligned}$$

In the development that follows, when treated as a variable in a generating function, B will be able to take on complex values. In particular, it will often be necessary to consider the different cases when $|B| < 1$, $|B| = 1$, or $|B| > 1$, that is, when the complex number B lies inside, on, or outside the unit circle.

3.1.3 Stationarity and Invertibility Conditions for a Linear Process

Stationarity. The convergence of the series (3.1.9) ensures that the process has a finite variance. Also, we have seen in Section 2.1.3 that the autocovariances and autocorrelations must satisfy a set of conditions to ensure stationarity. For a linear process (3.1.1), these conditions are guaranteed by the single condition that $\sum_{j=0}^{\infty} |\psi_j| < \infty$. This condition can also be embodied in the condition that the series $\psi(B)$, which is the generating function of the ψ weights, must converge for $|B| \leq 1$, that is, on or within the unit circle. This result is discussed in Appendix A3.1.

Spectrum of a Linear Stationary Process. It is shown in Appendix A3.1 that if we substitute $B = e^{-i2\pi f}$, where $i = \sqrt{-1}$, in the autocovariance generating function (3.1.11), we obtain one half of the power spectrum. Thus, the spectrum of a linear process is

$$\begin{aligned} p(f) &= 2\sigma_a^2 \psi(e^{-i2\pi f})\psi(e^{i2\pi f}) \\ &= 2\sigma_a^2 |\psi(e^{-i2\pi f})|^2 \qquad 0 \leq f \leq \frac{1}{2} \end{aligned} \tag{3.1.12}$$

In fact, this is the well-known expression (e.g., Jenkins and Watts, 1968) that relates the spectrum $p(f)$ of the output from a linear system to the uniform spectrum $2\sigma_a^2$ of a white noise input by multiplying it with the squared gain $G^2(f) = |\psi(e^{-i2\pi f})|^2$ of the system.

Invertibility. We have seen that the ψ weights of a linear process must satisfy the condition that $\psi(B)$ converges on or within the unit circle if the process is to be stationary. We now consider a similar restriction applied to the π weights to ensure what is called *invertibility*. This invertibility condition is independent of the stationarity condition and is also applicable to the nonstationary linear models, which we introduce in Chapter 4.

To illustrate the basic idea of invertibility, consider again the special case

$$\tilde{z}_t = (1 - \theta B)a_t \tag{3.1.13}$$

Expressing the a_t's in terms of the present and past $\tilde{z}_t$'s, this model becomes

$$a_t = (1 - \theta B)^{-1}\tilde{z}_t = (1 + \theta B + \theta^2 B^2 + \cdots + \theta^k B^k)(1 - \theta^{k+1} B^{k+1})^{-1}\tilde{z}_t$$

that is,

$$\tilde{z}_t = -\theta\tilde{z}_{t-1} - \theta^2\tilde{z}_{t-2} - \cdots - \theta^k\tilde{z}_{t-k} + a_t - \theta^{k+1}a_{t-k-1} \tag{3.1.14}$$

If $|\theta| < 1$, on letting k tend to infinity, we obtain the infinite series

$$\tilde{z}_t = -\theta\tilde{z}_{t-1} - \theta^2\tilde{z}_{t-2} - \cdots + a_t \tag{3.1.15}$$

and the π weights of the model in the form of (3.1.4) are $\pi_j = -\theta^j$. Whatever the value of θ, $\tilde{z}_t = (1 - \theta B)a_t$ defines a perfectly proper stationary process. However, if $|\theta| \geq 1$, the current deviation $\tilde{z}_t$ in (3.1.14) depends on $\tilde{z}_{t-1}, \tilde{z}_{t-2}, \ldots, \tilde{z}_{t-k}$, with weights that increase as k increases. We avoid this situation by requiring that $|\theta| < 1$. We then say that the series is *invertible.* We see that this condition is equivalent to $\sum_{j=0}^{\infty} |\theta|^j \equiv \sum_{j=0}^{\infty} |\pi_j| < \infty$, so

that the series

$$\pi(B) = (1 - \theta B)^{-1} = \sum_{j=0}^{\infty} \theta^j B^j$$

converges for all $|B| \leq 1$, that is, on or within the unit circle. The invertibility requirement is needed to associate present events with *past* values in a sensible manner.

The general linear process (3.1.1) is invertible and can be written in the form

$$\pi(B)\tilde{z}_t = a_t$$

if the weights π_j are absolutely summable, that is, if $\sum_{j=0}^{\infty} |\pi_j| < \infty$, which implies that the series $\pi(B)$ converges on or within the unit circle.

Thus, to summarize, a linear process (3.1.1) is *stationary* if $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and is *invertible* if $\sum_{j=0}^{\infty} |\pi_j| < \infty$, where $\pi(B) = \psi^{-1}(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$.

3.1.4 Autoregressive and Moving Average Processes

The representations (3.1.1) and (3.1.4) of the general linear process would not be very useful in practice if they contained an infinite number of parameters ψ_j and π_j. We now describe a way to introduce parsimony and arrive at models that are representationally useful for practical applications.

Autoregressive Processes. Consider first the special case of (3.1.4) in which only the first p of the weights are nonzero. The model may be written as

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + \cdots + \phi_p \tilde{z}_{t-p} + a_t \tag{3.1.16}$$

where we now use the symbols $\phi_1, \phi_2, \ldots, \phi_p$ for the *finite* set of weight parameters. The resulting process is called an *autoregressive* process of order p, or more succinctly, an AR(p) process. In particular, the AR(1) and AR(2) models

$$\begin{aligned} \tilde{z}_t &= \phi_1 \tilde{z}_{t-1} + a_t \\ &= \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + a_t \end{aligned}$$

are of considerable practical importance.

The AR(p) model can be written in the equivalent form

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)\tilde{z}_t = a_t$$

or

$$\phi(B)\tilde{z}_t = a_t \tag{3.1.17}$$

This implies that

$$\tilde{z}_t = \frac{1}{\phi(B)} a_t = \phi^{-1}(B) a_t \equiv \psi(B) a_t$$

Hence, the autoregressive process can be thought of as the output $\tilde{z}_t$ from a linear filter with transfer function $\phi^{-1}(B) = \psi(B)$ when the input is white noise a_t.

Moving Average Processes. Next consider the special case of (3.1.1), when only the first q of the ψ weights are nonzero. The process may be written as

$$\tilde{z}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \tag{3.1.18}$$

where we now use the symbols $-\theta_1, -\theta_2, \ldots, -\theta_q$ for the *finite* set of weight parameters. This process is called a *moving average* process[1] of order q, which we often abbreviate as MA(q). The special cases of MA(1) and MA(2) models

$$\begin{aligned}\tilde{z}_t &= a_t - \theta_1 a_{t-1} \\ &= a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}\end{aligned}$$

are again particularly important in practice.

Using the backshift operator $Ba_t = a_{t-1}$, the MA(q) model can be written in the equivalent form as

$$\tilde{z}_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)a_t$$

or more succinctly as

$$\tilde{z}_t = \theta(B)a_t \tag{3.1.19}$$

Hence, the moving average process can be thought of as the output $\tilde{z}_t$ from a linear filter with transfer function $\theta(B)$ when the input is white noise a_t.

Mixed Autoregressive–Moving Average Processes. As discussed in Section 3.1.1, the *finite* moving average process

$$\tilde{z}_t = a_t - \theta_1 a_{t-1} = (1 - \theta_1 B)a_t \qquad |\theta_1| < 1$$

can also be written as an *infinite* autoregressive process

$$\tilde{z}_t = -\theta_1 \tilde{z}_{t-1} - \theta_1^2 \tilde{z}_{t-2} - \cdots + a_t$$

However, if the process really was MA(1), we would not obtain a parsimonious representation using an autoregressive model. Conversely, an AR(1) process could not be parsimoniously represented using a moving average model. In practice, to obtain parsimonious parameterization, it is often useful to include both autoregressive and moving average terms in the model. The resulting model

$$\tilde{z}_1 = \phi_1 \tilde{z}_{t-1} + \cdots + \phi_p \tilde{z}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

or

$$\phi(B)\tilde{z}_t = \theta(B)a_t \tag{3.1.20}$$

is called the *mixed autoregressive–moving average* process of order (p, q), which we abbreviate as ARMA(p, q). For example, the ARMA(1, 1) process is

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + a_t - \theta_1 a_{t-1}$$

[1] As we remarked in Chapter 1, the term ''moving average'' is somewhat misleading since the weights do not sum to unity. However, this nomenclature is now well established and we will use it here.

Now writing

$$\begin{aligned}\tilde{z}_t &= \phi^{-1}(B)\theta(B)a_t \\ &= \frac{\theta(B)}{\phi(B)}a_t = \frac{1-\theta_1 B-\cdots-\theta_q B^q}{1-\phi_1 B-\cdots-\phi_p B^p}a_t\end{aligned}$$

we see that the mixed ARMA process can be thought of as the output $\tilde{z}_t$ from a linear filter, whose transfer function is the ratio of two polynomial operators $\theta(B)$ and $\phi(B)$, when the input is white noise a_t. Furthermore, since $\tilde{z}_t = z_t - \mu$, where $\mu = E[z_t]$ is the mean of the process in the stationary case, the general ARMA(p, q) process can also be written in terms of the original process z_t as

$$\phi(B)z_t = \theta_0 + \theta(B)a_t \tag{3.1.21}$$

where the constant term θ_0 is

$$\theta_0 = (1-\phi_1-\phi_2-\cdots-\phi_p)\mu \tag{3.1.22}$$

In the next sections, we discuss some important characteristics of autoregressive, moving average, and mixed models. In particular, we study their variances, autocorrelation functions, spectra, and the stationarity and invertibility conditions that must be imposed on their parameters.

3.2 AUTOREGRESSIVE PROCESSES

3.2.1 Stationarity Conditions for Autoregressive Processes

The parameters $\phi_1, \phi_2, \ldots, \phi_p$ of an AR(p) process

$$\tilde{z}_t = \phi_1\tilde{z}_{t-1} + \cdots + \phi_p\tilde{z}_{t-p} + a_t$$

or

$$(1-\phi_1 B-\cdots-\phi_p B^p)\tilde{z}_t = \phi(B)\tilde{z}_t = a_t$$

must satisfy certain conditions for the process to be stationary. For illustration, the AR(1) process

$$(1-\phi_1 B)\tilde{z}_t = a_t$$

may be written as

$$\tilde{z}_t = (1-\phi_1 B)^{-1}a_t = \sum_{j=0}^{\infty}\phi_1^j a_{t-j}$$

provided that the infinite series on the right converges in an appropriate sense. Hence,

$$\psi(B) = (1-\phi_1 B)^{-1} = \sum_{j=0}^{\infty}\phi_1^j B^j \tag{3.2.1}$$

We have seen in Section 3.1.3 that for stationarity, $\psi(B)$ must converge for $|B| \leq 1$, or equivalently that $\sum_{j=0}^{\infty} |\phi_1|^j < \infty$. This implies that the parameter ϕ_1 of an AR(1) process must satisfy the condition $|\phi_1| < 1$ to ensure stationarity. Since the root of $1 - \phi_1 B = 0$ is $B = \phi_1^{-1}$, this condition is equivalent to saying that the root of $1 - \phi_1 B = 0$ must lie *outside* the unit circle.

The general AR(p) process $\phi(B)\tilde{z}_t = a_t$ can be written as

$$\tilde{z}_t = \phi^{-1}(B)a_t \equiv \psi(B)a_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}$$

provided that the right-side expression is convergent. Using the factorization

$$\phi(B) = (1 - G_1 B)(1 - G_2 B) \cdots (1 - G_p B)$$

where $G_1^{-1}, \ldots, G_p^{-1}$ are the roots of $\phi(B) = 0$, and expanding $\theta^{-1}(B)$ in partial fractions yields

$$\tilde{z}_t = \phi^{-1}(B)a_t = \sum_{i=1}^{p} \frac{K_i}{1 - G_i B} a_t$$

Hence, if $\psi(B) = \phi^{-1}(B)$ is to be a convergent series for $|B| \leq 1$, that is, if the weights $\psi_j = \sum_{i=1}^{p} K_i G_i^j$ are to be absolutely summable so that the AR(p) process is stationary, we must have $|G_i| < 1$, for $i = 1, \ldots, p$. Equivalently, the roots of the $\phi(B) = 0$ must lie *outside* the unit circle. The roots of the equation $\phi(B) = 0$ may be referred to as the zeros of the polynomial $\phi(B)$. Thus, for stationarity, the zeros of $\phi(B)$ must lie *outside* the unit circle. A similar argument may be applied when the zeros of $\phi(B)$ are not all distinct. The equation $\phi(B) = 0$ is called the *characteristic equation* for the process.

Note also that the roots of $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p = 0$ are the reciprocals to the roots of the polynomial equation in m,

$$m^p - \phi_1 m^{p-1} - \cdots - \phi_p = 0$$

Hence, the stationarity condition that all roots of $\phi(B) = 0$ must lie outside the unit circle, that is, be greater than 1 in absolute value, is equivalent to the condition that all roots of $m^p - \phi_1 m^{p-1} - \cdots - \phi_p = 0$ must lie *inside* the unit circle, that is, be less than 1 in absolute value.

Since the series $\pi(B) = \phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ is finite, no restrictions are required on the parameters of an autoregressive process to ensure invertibility.

ψ Weights. Since $\psi(B) = 1/\phi(B)$ so that $\phi(B)\psi(B) = 1$, it readily follows that the weights ψ_j for the AR(p) process satisfy the difference equation

$$\psi_j = \phi_1 \psi_{j-1} + \phi_2 \psi_{j-2} + \cdots + \phi_p \psi_{j-p} \qquad j > 0$$

with $\psi_0 = 1$ and $\psi_j = 0$ for $j < 0$, from which the weights ψ_j can easily be computed recursively in terms of the ϕ_i. In fact, as seen from the principles of linear difference equations as discussed in Appendix A4.1, the fact that the weights ψ_j satisfy the difference equation discussed earlier implies that they have an explicit representation in the form of $\psi_j = \sum_{i=1}^{p} K_i G_i^j$ for the case of distinct roots.

3.2.2 Autocorrelation Function and Spectrum of Autoregressive Processes

Autocorrelation Function. An important recurrence relation for the autocorrelation function of a stationary autoregressive process is found by multiplying throughout in

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \cdots + \phi_p \tilde{z}_{t-p} + a_t$$

by $\tilde{z}_{t-k}$, for $k \geq 0$, to obtain

$$\tilde{z}_{t-k}\tilde{z}_t = \phi_1 \tilde{z}_{t-k}\tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-k}\tilde{z}_{t-2} + \cdots + \phi_p \tilde{z}_{t-k}\tilde{z}_{t-p} + \tilde{z}_{t-k}a_t \tag{3.2.2}$$

Now, on taking expected values, we obtain the difference equation

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \cdots + \phi_p \gamma_{k-p} \qquad k > 0 \tag{3.2.3}$$

Note that the expectation $E[\tilde{z}_{t-k}a_t]$ is zero for $k > 0$, since $\tilde{z}_{t-k}$ can only involve the shocks a_j up to time $t-k$, which are uncorrelated with a_t. On dividing throughout in (3.2.3) by γ_0, we see that the autocorrelation function satisfies the same form of difference equation

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p} \qquad k > 0 \tag{3.2.4}$$

Note that this is analogous to the difference equation satisfied by the process $\tilde{z}_t$ itself, but without the random shock input a_t.

Now suppose that this equation is written as

$$\phi(B)\rho_k = 0$$

where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and B now operates on k and not t. Then, writing

$$\phi(B) = \prod_{i=1}^{p}(1 - G_i B)$$

the general solution for ρ_k in (3.2.4) (see, e.g., Appendix A4.1) is

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \cdots + A_p G_p^k \tag{3.2.5}$$

where $G_1^{-1}, G_2^{-1}, \ldots, G_p^{-1}$ are the roots of the *characteristic equation*

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p = 0$$

or equivalently, $G_1, G_2, \ldots, G_p$ are the roots of $m^p - \phi_1 m^{p-1} - \cdots - \phi_p = 0$.

For stationarity, we require that $|G_i| < 1$. Thus, two situations can arise in practice if we assume that the roots G_i are distinct.

1. A root G_i is real, in which case a term $A_i G_i^k$ in (3.2.5) decays to zero geometrically as k increases. We often refer to this as a damped exponential.
2. A pair of roots G_i and G_j are complex conjugates, in which case they contribute a term

   $$D^k \sin(2\pi f k + F)$$

   to the autocorrelation function (3.2.5), which follows a damped sine wave, with damping factor $D = |G_i| = |G_j|$ and frequency f such that $2\pi f = \cos^{-1}[|\mathrm{Re}(G_i)|/D]$.

In general, the autocorrelation function of a stationary autoregressive process will consist of a mixture of damped exponentials and damped sine waves.

Autoregressive Parameters in Terms of the Autocorrelations: Yule–Walker Equations. If we substitute $k = 1, 2, \ldots, p$ in (3.2.4), we obtain a set of linear equations for $\phi_1, \phi_2, \ldots, \phi_p$ in terms of $\rho_1, \rho_2, \ldots, \rho_p$, that is,

$$\begin{aligned} \rho_1 &= \phi_1 + \phi_2\rho_1 + \cdots + \phi_p\rho_{p-1} \\ \rho_2 &= \phi_1\rho_1 + \phi_2 + \cdots + \phi_p\rho_{p-2} \\ &\vdots \quad \vdots \quad \cdots \quad \vdots \\ \rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \cdots + \phi_p \end{aligned} \tag{3.2.6}$$

These are the well-known *Yule–Walker* equations (Yule, 1927; Walker, 1931). We obtain *Yule--Walker estimates* of the parameters by replacing the theoretical autocorrelations ρ_k by the estimated autocorrelations r_k. Note that, if we write

$$\boldsymbol{\phi} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} \quad \boldsymbol{\rho}_p = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{bmatrix} \quad \mathbf{P}_p = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & 1 \end{bmatrix}$$

the solution of (3.2.6) for the parameters $\boldsymbol{\phi}$ in terms of the autocorrelations may be written as

$$\boldsymbol{\phi} = \mathbf{P}_p^{-1}\boldsymbol{\rho}_p \tag{3.2.7}$$

Variance. When $k = 0$, the contribution from the term $E[\tilde{z}_{t-k}a_t]$, on taking expectations in (3.2.2), is $E[a_t^2] = \sigma_a^2$, since the only part of $\tilde{z}_t$ that will be correlated with a_t is the most recent shock, a_t. Hence, when $k = 0$,

$$\gamma_0 = \phi_1\gamma_{-1} + \phi_2\gamma_{-2} + \cdots + \phi_p\gamma_{-p} + \sigma_a^2$$

On substituting $\gamma_{-k} = \gamma_k$ and writing $\gamma_k = \gamma_0\rho_k$, the variance $\gamma_0 = \sigma_z^2$ may be written as

$$\sigma_z^2 = \frac{\sigma_a^2}{1 - \phi_1\rho_1 - \phi_2\rho_2 - \cdots - \phi_p\rho_p} \tag{3.2.8}$$

Spectrum. For the AR(p) process, $\psi(B) = \phi^{-1}(B)$ and

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

Therefore, using (3.1.12), the spectrum of an autoregressive process is

$$p(f) = \frac{2\sigma_a^2}{|1 - \phi_1 e^{-i2\pi f} - \phi_2 e^{-i4\pi f} - \cdots - \phi_p e^{-i2p\pi f}|^2} \qquad 0 \le f \le \tfrac{1}{2} \tag{3.2.9}$$

We now discuss two particularly important autoregressive processes, those of first and second order.

3.2.3 The First-Order Autoregressive Process

The first-order autoregressive process is

$$\begin{aligned}\tilde{z}_t &= \phi_1 \tilde{z}_{t-1} + a_t \\ &= a_t + \phi_1 a_{t-1} + \phi_1^2 a_{t-2} + \cdots \end{aligned} \tag{3.2.10}$$

where it has been shown in Section 3.2.1 that ϕ_1 must satisfy the condition $-1 < \phi_1 < 1$ for the process to be stationary.

Autocorrelation Function. Using (3.2.4), the autocorrelation function satisfies the first-order difference equation

$$\rho_k = \phi_1 \rho_{k-1} \qquad k > 0 \tag{3.2.11}$$

which, with $\rho_0 = 1$, has the solution

$$\rho_k = \phi_1^k \qquad k \geq 0 \tag{3.2.12}$$

Since $-1 < \phi < 1$, the autocorrelation function decays exponentially to zero when ϕ_1 is positive but decays exponentially to zero and oscillates in sign when ϕ_1 is negative. In particular, we note that

$$\rho_1 = \phi_1 \tag{3.2.13}$$

Variance. Using (3.2.8), the variance of the process is

$$\begin{aligned}\sigma_z^2 &= \frac{\sigma_a^2}{1 - \rho_1 \phi_1} \\ &= \frac{\sigma_a^2}{1 - \phi_1^2} \end{aligned} \tag{3.2.14}$$

on substituting $\rho_1 = \phi_1$

Spectrum. Finally, using (3.2.9), the spectrum is

$$\begin{aligned} p(f) &= \frac{2\sigma_a^2}{|1 - \phi_1 e^{-i2\pi f}|^2} \\ &= \frac{2\sigma_a^2}{1 + \phi_1^2 - 2\phi_1 \cos(2\pi f)} \qquad 0 \leq f \leq \tfrac{1}{2} \end{aligned} \tag{3.2.15}$$

Example. Figure 3.1 shows realizations from two AR(1) processes with $\phi_1 = 0.8$ and $\phi_1 = -0.8$, and the corresponding theoretical autocorrelation functions and spectra. Thus, when the parameter has the large positive value $\phi_1 = 0.8$, neighboring values in the series are similar and the series exhibits marked trends. This is reflected in the autocorrelation function, which slowly decays to zero, and in the spectrum, which is dominated by low frequencies. When the parameter has the large negative value $\phi_1 = -0.8$, the series tends to oscillate rapidly, and this is reflected in the autocorrelation function, which alternates

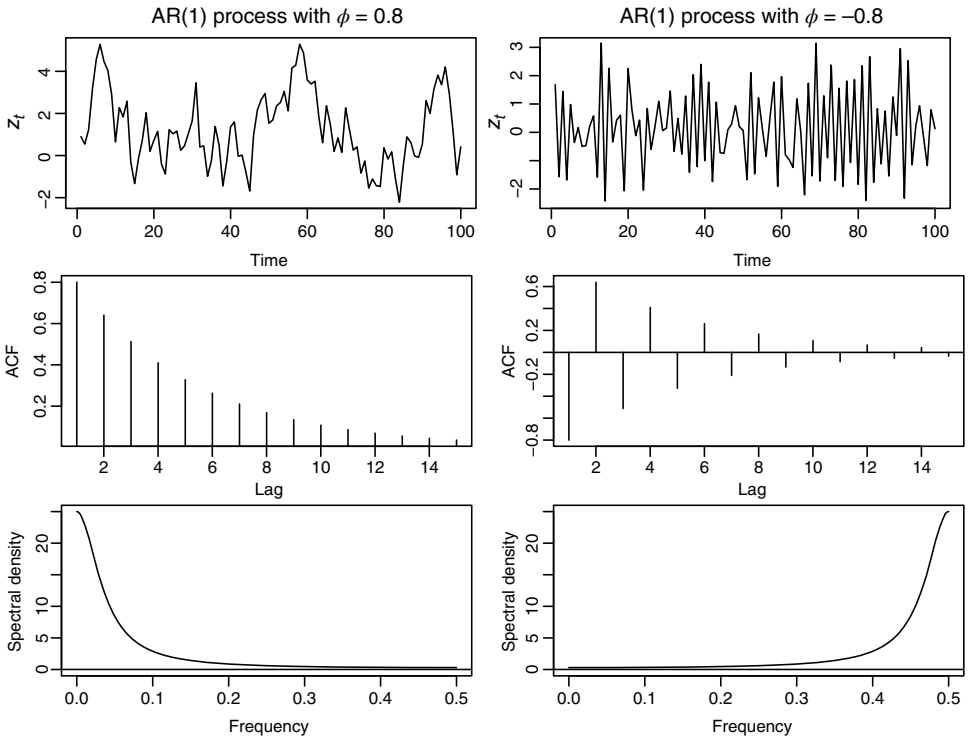


FIGURE 3.1 Realizations from two first-order autoregressive processes and their corresponding theoretical autocorrelation functions and spectral density functions.

in sign as it decays to zero, and in the spectrum, which is dominated by high frequencies. Figure 3.1 was generated in R and can be reproduced as follows:

```
>library(TSA)
>set.seed(12345)
>par(mfrow=c(3,2))
>plot(arima.sim(list(order=c(1,0,0),ar = 0.8), n=100),ylab=
 expression(z[t]),main=expression("AR(1) process with "*phi*"=0.8"))
>plot(arima.sim(list(order=c(1,0,0),ar = -0.8), n=100), ylab=
 expression(z[t]),main=expression("AR(1) process with "*phi*"=-0.8"))
>plot(ARMAacf(ar=0.8,ma=0,15)[-1],type="h",ylab="ACF",xlab="lag")
>abline(h=0)
>plot(ARMAacf(ar=-0.8,ma=0,15)[-1],type="h",ylab="ACF",xlab="lag")
>abline(h=0)
>ARMAspec(model=list(ar=0.8),freq=seq(0,0.5,0.001),plot=TRUE)
>ARMAspec(model=list(ar=-0.8),freq=seq(0,0.5,0.001),plot=TRUE)
```

3.2.4 Second-Order Autoregressive Process

Stationarity Condition. The second-order autoregressive process can be written as

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + a_t \tag{3.2.16}$$

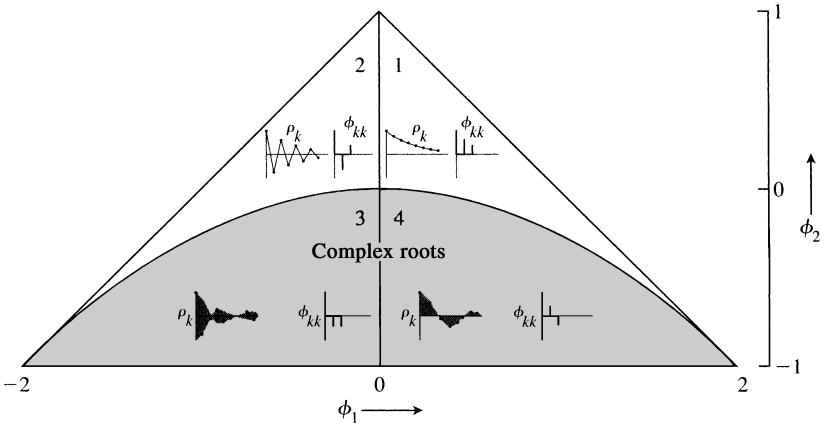


FIGURE 3.2 Typical autocorrelation and partial autocorrelation functions ρ_k and ϕ_{kk} for various stationary AR(2) models (*Source:* Stralkowski, 1968).

For stationarity, the roots of

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 = 0 \tag{3.2.17}$$

must lie outside the unit circle, which implies that the parameters ϕ_1 and ϕ_2 must lie in the triangular region

$$\begin{aligned} \phi_2 + \phi_1 &< 1 \\ \phi_2 - \phi_1 &< 1 \\ -1 < \phi_2 &< 1 \end{aligned} \tag{3.2.18}$$

as shown in Figure 3.2.

Autocorrelation Function. Using (3.2.4), the autocorrelation function satisfies the second-order difference equation

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} \qquad k > 0 \tag{3.2.19}$$

with starting values $\rho_0 = 1$ and $\rho_1 = \phi_1/(1 - \phi_2)$. From (3.2.5), the general solution to this difference equation is

$$\begin{aligned} \rho_k &= A_1 G_1^k + A_2 G_2^k \\ &= \frac{G_1(1 - G_2^2)G_1^k - G_2(1 - G_1^2)G_2^k}{(G_1 - G_2)(1 + G_1 G_2)} \end{aligned} \tag{3.2.20}$$

where G_1^{-1} and G_2^{-1} are the roots of the characteristic equation $\phi(B) = 0$. When the roots are real, the autocorrelation function consists of a mixture of damped exponentials. This occurs when $\phi_1^2 + 4\phi_2 \geq 0$ and corresponds to regions 1 and 2, which lie above the parabolic boundary in Figure 3.2. Specifically, in region 1, the autocorrelation function remains positive as it damps out, corresponding to a positive dominant root in (3.2.20). In region 2, the autocorrelation function alternates in sign as it damps out, corresponding to a negative dominant root.

If the roots G_1 and G_2 are complex ($\phi_1^2 + 4\phi_2 < 0$), a second-order autoregressive process displays *pseudoperiodic behavior*. This behavior is reflected in the autocorrelation function, for on substituting $G_1 = De^{i2\pi f_0}$ and $G_2 = De^{-i2\pi f_0}$ ($0 < f_0 < \frac{1}{2}$) in (3.2.20), we obtain

$$\rho_k = \frac{D^k \sin(2\pi f_0 k + F)}{\sin F} \tag{3.2.21}$$

We refer to this as a *damped sine wave* with damping factor D, *frequency* f_0, and *phase* F. These factors are related to the process parameters as follows:

$$D = |G_i| = \sqrt{-\phi_2} \tag{3.2.22}$$

where the positive square root is taken,

$$\cos(2\pi f_0) = \frac{\mathrm{Re}(G_i)}{D} = \frac{\phi_1}{2\sqrt{-\phi_2}} \tag{3.2.23}$$

$$\tan F = \frac{1 + D^2}{1 - D^2} \tan(2\pi f_0) \tag{3.2.24}$$

Again referring to Figure 3.2, the autocorrelation function is a damped sine wave in regions 3 and 4, the phase angle F being less than 90° in region 4 and lying between 90° and 180° in region 3. This means that the autocorrelation function starts with a positive value throughout region 4 but always switches sign from lag 0 to lag 1 in region 3.

Yule–Walker Equations. For the AR(2) model, the Yule–Walker equations become

$$\begin{aligned} \rho_1 &= \phi_1 + \phi_2 \rho_1 \\ \rho_2 &= \phi_1 \rho_1 + \phi_2 \end{aligned} \tag{3.2.25}$$

which, when solved for ϕ_1 and ϕ_2, give

$$\begin{aligned} \phi_1 &= \frac{\rho_1(1 - \rho_2)}{1 - \rho_1^2} \\ \phi_2 &= \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \end{aligned} \tag{3.2.26}$$

These equations can also be solved to express ρ_1 and ρ_2 in terms of ϕ_1 and ϕ_2 to give

$$\begin{aligned} \rho_1 &= \frac{\phi_1}{1 - \phi_2} \\ \rho_2 &= \phi_2 + \frac{\phi_1^2}{1 - \phi_2} \end{aligned} \tag{3.2.27}$$

which provide the starting values for the recursions in (3.2.19). Expressions (3.2.20) and (3.2.21) are useful for explaining the different patterns for ρ_k that may arise in practice. However, for computing the autocorrelations of an AR(2) process, it is simplest to make direct use of the recursions implied by (3.2.19).

Using the stationarity condition (3.2.18) and the expressions for ρ_1 and ρ_2 in (3.2.27), it can be seen that the admissible values of ρ_1 and ρ_2, for a stationary AR(2) process, must

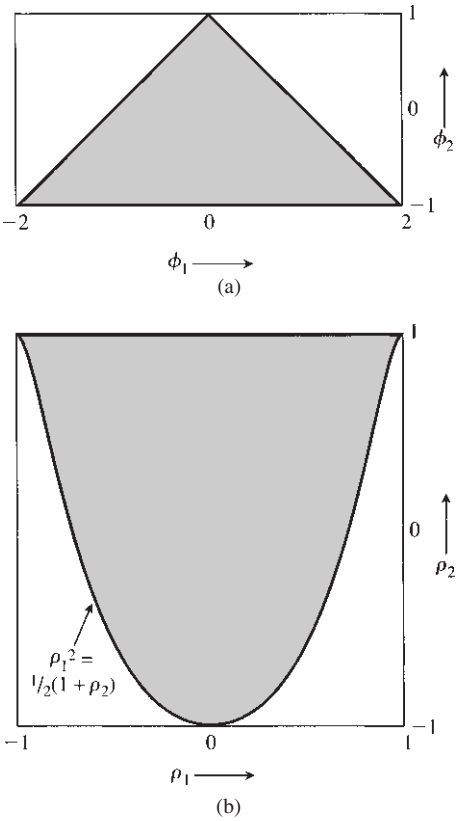


FIGURE 3.3 Admissible regions for (a) ϕ_1, ϕ_2 and (b) ρ_1, ρ_2, for a stationary AR(2) process.

lie in the region

$$\begin{aligned} -1 < \rho_1 < 1 \\ -1 < \rho_2 < 1 \\ \rho_1^2 < \tfrac{1}{2}(\rho_2 + 1) \end{aligned}$$

The admissible region for the parameters ϕ_1 and ϕ_2 is shown in Figure 3.3(a), while Figure 3.3(b) shows the corresponding admissible region for ρ_1 and ρ_2.

Variance. From (3.2.8), the variance of the AR(2) process is

$$\begin{aligned} \sigma_z^2 &= \frac{\sigma_a^2}{1 - \rho_1\phi_1 - \rho_2\phi_2} \\ &= \frac{1 - \phi_2}{1 + \phi_2} \frac{\sigma_a^2}{(1 - \phi_2)^2 - \phi_1^2} \end{aligned} \tag{3.2.28}$$

Spectrum. From (3.2.9), the spectrum is

$$p(f) = \frac{2\sigma_a^2}{|1 - \phi_1 e^{-i2\pi f} - \phi_2 e^{-i4\pi f}|^2}$$
$$= \frac{2\sigma_a^2}{1 + \phi_1^2 + \phi_2^2 - 2\phi_1(1 - \phi_2)\cos(2\pi f) - 2\phi_2 \cos(4\pi f)} \qquad 0 \le f \le \tfrac{1}{2} \tag{3.2.29}$$

The spectrum also reflects the pseudoperiodic behavior that the series exhibits when the roots of the characteristic equation are complex. For illustration, Figure 3.4(a) shows 70 values of a series generated from the AR(2) model

$$\tilde{z}_t = 0.75\tilde{z}_{t-1} - 0.50\tilde{z}_{t-2} + a_t$$

Figure 3.4(b) shows the corresponding theoretical autocorrelation function. The roots of the characteristic equation

$$1 - 0.75B + 0.5B^2 = 0$$

are complex, so that the pseudoperiodic behavior observed in the series is to be expected. We clearly see this behavior reflected in the theoretical autocorrelation function of Figure 3.4(b), the average apparent period being about 6. The damping factor D and frequency

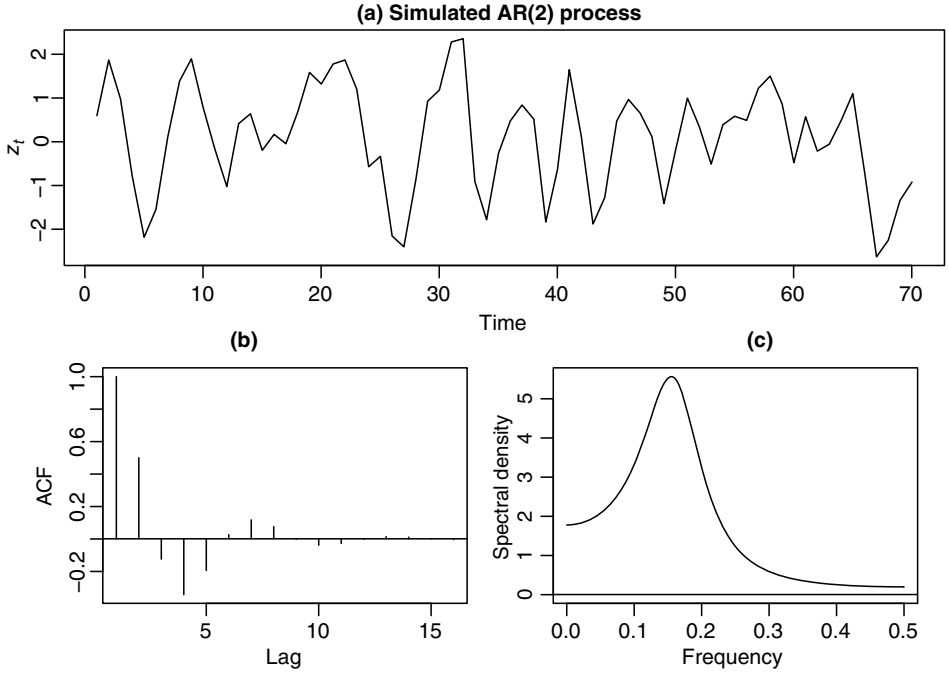


FIGURE 3.4 (a) Time series generated from a second-order autoregressive process $\tilde{z}_t = 0.75\,\tilde{z}_{t-1} - 0.50\,\tilde{z}_{t-2} + a_t$, along with (b) the theoretical autocorrelation function, and (c) the spectral density function for the same process.

f_0, from (3.2.22) and (3.2.23), are

$$D = \sqrt{0.50} = 0.71 \qquad f_0 = \frac{\cos^{-1}(0.5303)}{2\pi} = \frac{1}{6.2}$$

Thus, the fundamental period of the autocorrelation function is 6.2. In addition, the theoretical spectral density function in Figure 3.4(c) shows that a large proportion of the variance of the series is accounted for by frequencies in the neighborhood of f_0.

Figure 3.4 was generated in R using the following commands:

```
> library(TSA)
> ar.acf=ARMAacf(model=list(ar=c(0.75,-0.5)))
> ar.spec=ARMAspec(model=list(ar=c(0.75,-0.5),freq=seq(0,0.5,0.0005)))
> layout(matrix(c(1,1,2,3),2,2,byrow=TRUE))
> plot(arima.sim(list(order=c(2,0,0),ar=c(0.75,-0.5)), n=70), ylab=
      expression(z[t]),xlab="Time",main=("Simulated AR(2) process"))
> plot(ar.acf, main="b")
> plot(ar.spec, main="c")
```

3.2.5 Partial Autocorrelation Function

In practice, we typically do not know the order of the autoregressive process initially, and the order has to be specified from the data. The problem is analogous to deciding on the number of independent variables to be included in a multiple regression. The partial autocorrelation function is a tool that exploits the fact that, whereas an AR(p) process has an autocorrelation function that is infinite in extent, the partial autocorrelations are zero beyond lag p.

The partial autocorrelations can be described in terms of p nonzero *functions* of the autocorrelations. Denote by ϕ_{kj} the jth coefficient in an autoregressive representation of order k, so that ϕ_{kk} is the last coefficient. From (3.2.4), the ϕ_{kj} satisfy the set of equations

$$\rho_j = \phi_{k1}\rho_{j-1} + \cdots + \phi_{k(k-1)}\rho_{j-k+1} + \phi_{kk}\rho_{j-k} \qquad j = 1, 2, \ldots, k \tag{3.2.30}$$

leading to the Yule–Walker equations (3.2.6), which may be written as

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{bmatrix} \tag{3.2.31}$$

or

$$\mathbf{P}_k \boldsymbol{\phi}_k = \boldsymbol{\rho}_k \tag{3.2.32}$$

Solving these equations for $k = 1, 2, 3, \ldots$, successively, we obtain

$$\phi_{11} = \rho_1$$

$$\phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \tag{3.2.33}$$

$$\phi_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}}$$

In general, for ϕ_{kk}, the determinant in the numerator has the same elements as that in the denominator, but with the last column replaced by $\boldsymbol{\rho}_k$. The quantity ϕ_{kk}, regarded as a function of the lag k, is called the *partial autocorrelation* function.

For an AR(p) process, the partial autocorrelations ϕ_{kk} will be nonzero for $k \leq p$ and *zero for* $k > p$. In other words, the partial autocorrelation function of the AR(p) process has a *cutoff* after lag p. For the AR(2) process, partial autocorrelation functions ϕ_{kk} are shown in each of the four regions of Figure 3.2. As a numerical example, the partial autocorrelations of the AR(2) process $\tilde{z}_t = 0.75\tilde{z}_{t-1} - 0.50\tilde{z}_{t-2} + a_t$ considered in Figure 3.4 are $\phi_{11} = \rho_1 = 0.5$, $\phi_{22} = (\rho_2 - \rho_1^2)/(1 - \rho_1^2) = -0.5 \equiv \phi_2$, and $\phi_{kk} = 0$, for all $k > 2$.

The quantity ϕ_{kk} is called the *partial autocorrelation* of the process $\{z_t\}$ at lag k, since it equals the partial correlation between the variables z_t and z_{t-k} adjusted for the intermediate variables $z_{t-1}, z_{t-2}, \ldots, z_{t-k+1}$ (or the correlation between z_t and z_{t-k} not accounted for by $z_{t-1}, z_{t-2}, \ldots, z_{t-k+1}$). Now, it is easy to establish from least squares theory that the values $\phi_{k1}, \phi_{k2}, \ldots, \phi_{kk}$, which are the solutions to (3.2.31), are the regression coefficients in the linear regression of z_t on $z_{t-1}, \ldots, z_{t-k}$, that is, they are the values of coefficients $b_1, \ldots, b_k$, which minimize $E[(z_t - b_0 - \sum_{i=1}^{k} b_i z_{t-i})^2]$. Hence, assuming for convenience that the process $\{z_t\}$ has mean zero, the best linear predictor, in the mean squared error sense, of z_t based on $z_{t-1}, z_{t-2}, \ldots, z_{t-k+1}$ is

$$\hat{z}_t = \phi_{k-1,1} z_{t-1} + \phi_{k-1,2} z_{t-2} + \cdots + \phi_{k-1,k-1} z_{t-k+1}$$

whether the process is an AR or not. Similarly, the best linear predictor of z_{t-k} based on the (future) values $z_{t-1}, z_{t-2}, \ldots, z_{t-k+1}$ is

$$\hat{z}_{t-k} = \phi_{k-1,1} z_{t-k+1} + \phi_{k-1,2} z_{t-k+2} + \cdots + \phi_{k-1,k-1} z_{t-1}$$

Then, the lag k partial autocorrelation of $\{z_t\}$, ϕ_{kk}, can be defined as the correlation between the residuals from these two regressions on $z_{t-1}, \ldots, z_{t-k+1}$, that is,

$$\phi_{kk} = \text{corr}[z_t - \hat{z}_t, z_{t-k} - \hat{z}_{t-k}] \tag{3.2.34}$$

TABLE 3.1 Estimated Partial Autocorrelation Function for the Chemical Yield Data in Figure 2.1

k	$\hat{\phi}_{kk}$	k	$\hat{\phi}_{kk}$	k	$\hat{\phi}_{kk}$
1	−0.39	6	−0.12	11	0.14
2	0.18	7	0.02	12	−0.01
3	0.00	8	0.00	13	0.09
4	−0.04	9	− 0.06	14	0.17
5	−0.07	10	0.00	15	0.00

As examples, we find that $\phi_{11} = \text{corr}[z_t, z_{t-1}] = \rho_1$, while

$$\begin{aligned}\phi_{22} &= \text{corr}[z_t - \rho_1 z_{t-1}, z_{t-2} - \rho_1 z_{t-1}] \\ &= \frac{\gamma_2 - 2\rho_1\gamma_1 + \rho_1^2\gamma_0}{[(\gamma_0 + \rho_1^2\gamma_0 - 2\rho_1\gamma_1)^2]^{1/2}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}\end{aligned}$$

which agrees with the results in (3.2.33) derived from the Yule–Walker equations. Higher order partial autocorrelations ϕ_{kk} defined through (3.2.34) can similarly be shown to be the solution to the appropriate set of Yule–Walker equations.

3.2.6 Estimation of the Partial Autocorrelation Function

The partial autocorrelations may be estimated by fitting successively autoregressive models of orders $1, 2, 3, \ldots$ by least squares and picking out the estimates $\hat{\phi}_{11}, \hat{\phi}_{22}, \hat{\phi}_{33}, \ldots$ of the last coefficient fitted at each stage. Alternatively, if the values of the parameters are not too close to the nonstationary boundaries, approximate Yule–Walker estimates of the successive autoregressive models may be employed. The estimated partial autocorrelations can then be obtained by substituting estimates r_j for the theoretical autocorrelations in (3.2.30), to yield

$$r_j = \hat{\phi}_{k1} r_{j-1} + \hat{\phi}_{k2} r_{j-2} + \cdots + \hat{\phi}_{k(k-1)} r_{j-k+1} + \hat{\phi}_{kk} r_{j-k} \qquad j = 1, 2, \ldots, k \qquad (3.2.35)$$

and solving the resultant equations for $k = 1, 2, \ldots$. This can be done using a simple recursive method due to Levinson (1947) and Durbin (1960), which we describe in Appendix A3.2. However, these estimates obtained from (3.2.35) become very sensitive to rounding errors and should not be used if the values of the parameters are close to the nonstationary boundaries.

3.2.7 Standard Errors of Partial Autocorrelation Estimates

It was shown by Quenouille (1949) that on the hypothesis that the process is autoregressive of order p, the estimated partial autocorrelations of order $p + 1$, and higher, are approximately independently and normally distributed with zero mean. Also, if n is the number of observations used in fitting,

$$\text{var}[\hat{\phi}_{kk}] \simeq \frac{1}{n} \qquad k \geq p + 1$$

Thus, the standard error (SE) of the estimated partial autocorrelation $\hat{\phi}_{kk}$ is

$$\text{SE}[\hat{\phi}_{kk}] = \hat{\sigma}[\hat{\phi}_{kk}] \simeq \frac{1}{\sqrt{n}} \qquad k \geq p+1 \tag{3.2.36}$$

3.2.8 Calculations in R

The estimation of the partial autocorrelation function is conveniently performed in R. For example, the command pacf(Yield) in the stats package gives the estimated partial autocorrelations shown in Table 3.1 for the chemical yield data plotted in Figure 2.1. An alternative is to use the command acf2() in the R package astsa. This command has the advantage that it produces plots of the autocorrelation and partial autocorrelation functions in a single graph. This allows easy comparison of the two functions, which will be useful for specifying a model for the time series. Figure 3.5 shows a graph of the 15 first autocorrelations and partial autocorrelations for the chemical yield data produced using this routine. The patterns of the two functions resemble those of an AR(1) process with a negative value of ϕ_1, or possibly an AR(2) process with a dominant negative root (see region 2 of Figure 3.2). Also shown in Figure 3.5 by dashed lines are the two SE limits calculated on the assumption that the process is white noise. Since $\hat{\phi}_{22}$ is the second biggest partial autocorrelation, the possibility that the process is AR(2) should be kept in mind.

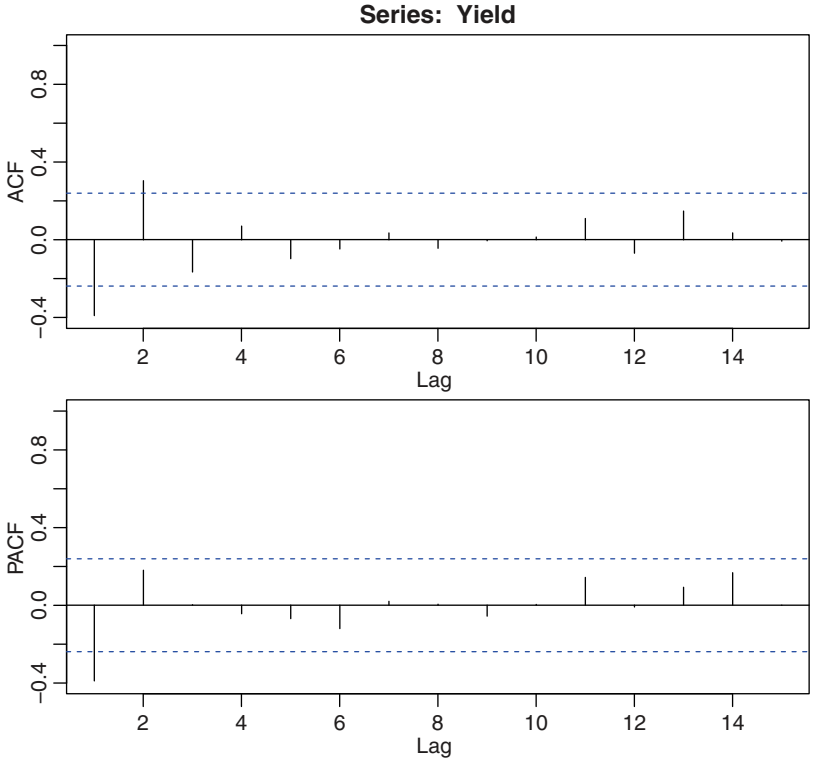


FIGURE 3.5 Estimated autocorrelation and partial autocorrelation functions for the chemical yield data in Series F.

The use of the autocorrelation and partial autocorrelation functions for model specification will be discussed more fully in Chapter 6. Figure 3.5 was generated using the following R commands:

```
> library(astsa)
> seriesF=read.table("SeriesF.txt,header=TRUE)
> Yield=ts(seriesF)
> acf2(Yield,15)
```

3.3 MOVING AVERAGE PROCESSES

3.3.1 Invertibility Conditions for Moving Average Processes

We now derive the conditions that the parameters $\theta_1, \theta_2, \ldots, \theta_q$ must satisfy to ensure the invertibility of the MA(q) process:

$$\begin{aligned}\tilde{z}_t &= a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \\ &= (1 - \theta_1 B - \cdots - \theta_q B^q) a_t \\ &= \theta(B) a_t \end{aligned} \tag{3.3.1}$$

We have already seen in Section 3.1.3 that the first-order moving average process

$$\tilde{z}_t = (1 - \theta_1 B) a_t$$

is invertible if $|\theta_1| < 1$; that is,

$$\pi(B) = (1 - \theta_1 B)^{-1} = \sum_{j=0}^{\infty} \theta_1^j B^j$$

converges on or within the unit circle. However, this is equivalent to saying that the root, $B = \theta_1^{-1}$ of $(1 - \theta_1 B) = 0$, lies *outside* the unit circle.

The invertibility condition for higher order MA processes may be obtained by writing $\tilde{z}_t = \theta(B) a_t$ as

$$a_t = \theta^{-1}(B) \tilde{z}_t$$

Hence, if

$$\theta(B) = \prod_{i=1}^{q} (1 - H_i B)$$

where $H_1^{-1}, \ldots, H_q^{-1}$ are the roots of $\theta(B) = 0$, then, on expanding in partial fractions, we obtain

$$\pi(B) = \theta^{-1}(B) = \sum_{i=1}^{q} \left(\frac{M_i}{1 - H_i B} \right)$$

which converges, or equivalently, the weights $\pi_j = -\sum_{i=1}^{q} M_i H_i^j$ are absolutely summable, if $|H_i| < 1$, for $i = 1, 2, \ldots, q$. It follows that the invertibility condition for

an MA(q) process is that the roots H_i^{-1} of the characteristic equation

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q = 0 \tag{3.3.2}$$

lie *outside* the unit circle. From the relation $\theta(B)\pi(B) = 1$, it follows that the weights π_j satisfy the difference equation

$$\pi_j = \theta_1 \pi_{j-1} + \theta_2 \pi_{j-2} + \cdots + \theta_q \pi_{j-q} \qquad j > 0$$

with the convention that $\pi_0 = -1$ and $\pi_j = 0$ for $j < 0$, from which the weights π_j can easily be computed recursively in terms of the θ_i.

Note that since the series

$$\psi(B) = \theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

is finite, no restrictions are needed on the parameters of the moving average process to ensure stationarity.

3.3.2 Autocorrelation Function and Spectrum of Moving Average Processes

Autocorrelation Function. The autocovariance function of an MA(q) process is

$$\begin{aligned}\gamma_k &= E[(a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q})(a_{t-k} - \theta_1 a_{t-k-1} - \cdots - \theta_q a_{t-k-q})] \\ &= -\theta_k E[a_{t-k}^2] + \theta_1\theta_{k+1} E[a_{t-k-1}^2] + \cdots + \theta_{q-k}\theta_q E[a_{t-q}^2]\end{aligned}$$

since the a_t are uncorrelated, and $\gamma_k = 0$ for $k > q$. Hence, the variance of the process is

$$\gamma_0 = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma_a^2 \tag{3.3.3}$$

and

$$\gamma_k = \begin{cases} (-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q)\sigma_a^2 & k = 1, 2, \ldots, q \\ 0 & k > q \end{cases}$$

Thus, the autocorrelation function is

$$\rho_k = \begin{cases} \dfrac{-\theta_k + \theta_1\theta_{k+1} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \cdots + \theta_q^2} & k = 1, 2, \ldots, q \\ 0 & k > q \end{cases} \tag{3.3.4}$$

We see that the autocorrelation function of an MA(q) process is zero, beyond the order q of the process. In other words, the autocorrelation function of a moving average process has a *cutoff* after lag q.

Moving Average Parameters in Terms of Autocorrelations. If $\rho_1, \rho_2, \ldots, \rho_q$ are known, the q equations (3.3.4) may be solved for the parameters $\theta_1, \theta_2, \ldots, \theta_q$. However, unlike the Yule–Walker equations (3.2.6) for an autoregressive process, the equations (3.3.4) are nonlinear. Hence, except in the simple case where $q = 1$, which is discussed shortly, these equations have to be solved iteratively. Estimates of the moving average parameters may be obtained by substituting estimates r_k for ρ_k and solving the resulting equations. However, unlike the autoregressive estimates obtained from the Yule–Walker equations,

the resulting moving average estimates may not have high statistical efficiency. Nevertheless, they can provide useful rough estimates at the model identification stage discussed in Chapter 6. Furthermore, they provide useful starting values for an iterative parameter estimation procedure, discussed in Chapter 7, which converges to the efficient maximum likelihood estimates.

Spectrum. For the MA(q) process,

$$\psi(B) = \theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

Therefore, using (3.1.12), the spectrum of an MA(q) process is

$$p(f) = 2\sigma_a^2 |1 - \theta_1 e^{-i2\pi f} - \theta_2 e^{-i4\pi f} - \cdots - \theta_q e^{-i2q\pi f}|^2 \qquad 0 \le f \le \frac{1}{2} \tag{3.3.5}$$

We now discuss in greater detail the moving average processes of first and second order, which are of considerable practical importance.

3.3.3 First-Order Moving Average Process

We have already introduced the MA(1) process

$$\begin{aligned} \tilde{z}_t &= a_t - \theta_1 a_{t-1} \\ &= (1 - \theta_1 B) a_t \end{aligned}$$

and we have seen that θ_1 must lie in the range $-1 < \theta_1 < 1$ for the process to be invertible. The process is, of course, stationary for all values of θ_1.

Autocorrelation Function. It is easy to see that the variance of this process equals

$$\gamma_0 = (1 + \theta_1^2)\sigma_a^2$$

The autocorrelation function is

$$\rho_k = \begin{cases} \dfrac{-\theta_1}{1 + \theta_1^2} & k = 1 \\ 0 & k > 1 \end{cases} \tag{3.3.6}$$

from which it is noted that ρ_1 must satisfy $|\rho_1| = |\theta_1|/(1 + \theta_1^2) \le \frac{1}{2}$. Also, for $k = 1$, we find that

$$\rho_1 \theta_1^2 + \theta_1 + \rho_1 = 0 \tag{3.3.7}$$

with roots for θ_1 equal to $\theta_1 = (-1 \pm \sqrt{1 - 4\rho_1^2})/(2\rho_1)$. Since the product of the roots is unity, we see that if θ_1 is a solution, so is θ_1^{-1}. Furthermore, if θ_1 satisfies the invertibility condition $|\theta_1| < 1$, the other root θ_1^{-1} will be greater than unity and will not satisfy the condition. For example, if $\rho_1 = -0.4$, the two solutions are $\theta_1 = 0.5$ and $\theta_1 = 2.0$. However, only the solution $\theta_1 = 0.5$ corresponds to an invertible model.

Spectrum. Using (3.3.5), the spectrum of the MA(1) process is

$$\begin{aligned} p(f) &= 2\sigma_a^2 |1 - \theta_1 e^{-i2\pi f}|^2 \\ &= 2\sigma_a^2 [1 + \theta_1^2 - 2\theta_1 \cos(2\pi f)] \qquad 0 \le f \le \frac{1}{2} \end{aligned} \tag{3.3.8}$$

In general, when θ_1 is negative, ρ_1 is positive, and the spectrum is dominated by low frequencies. Conversely, when θ_1 is positive, ρ_1 is negative, and the spectrum is dominated by high frequencies.

Partial Autocorrelation Function. Using (3.2.31) with $\rho_1 = -\theta_1/(1 + \theta_1^2)$ and $\rho_k = 0$, for $k > 1$, we obtain after some algebraic manipulation

$$\phi_{kk} = \frac{-\theta_1^k (1 - \theta_1^2)}{1 - \theta_1^{2(k+1)}}$$

Thus, $|\phi_{kk}| < |\theta_1|^k$, and the partial autocorrelation function is dominated by a damped exponential. If ρ_1 is positive, so that θ_1 is negative, the partial autocorrelations alternate in sign. If, however, ρ_1 is negative, so that θ_1 is positive, the partial autocorrelations are negative. From (3.1.15), it has been seen that the weights π_j for the MA(1) process are $\pi_j = -\theta_1^j$, and hence since these are coefficients in the infinite autoregressive form of the process, it makes sense that the partial autocorrelation function ϕ_{kk} for the MA(1) essentially mimics the exponential decay feature of the weights π_j.

We now note a duality between the AR(1) and the MA(1) processes. Thus, whereas the autocorrelation function of an MA(1) process has a cutoff after lag 1, the autocorrelation function of an AR(1) process tails off exponentially. Conversely, whereas the partial autocorrelation function of an MA(1) process tails off and is dominated by a damped exponential, the partial autocorrelation function of an AR(1) process has a cutoff after lag 1. It turns out that a corresponding approximate duality of this kind occurs in general in the autocorrelation and partial autocorrelation functions between AR and MA processes.

3.3.4 Second-Order Moving Average Process

Invertibility Conditions. The second-order moving average process is defined by

$$\begin{aligned} \tilde{z}_t &= a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \\ &= (1 - \theta_1 B - \theta_2 B^2) a_t \end{aligned}$$

and is stationary for all values of θ_1 and θ_2. However, it is invertible only if the roots of the characteristic equation

$$1 - \theta_1 B - \theta_2 B^2 = 0 \tag{3.3.9}$$

lie outside the unit circle, that is,

$$\begin{aligned} \theta_2 + \theta_1 &< 1 \\ \theta_2 - \theta_1 &< 1 \\ -1 < \theta_2 &< 1 \end{aligned} \tag{3.3.10}$$

These are parallel to conditions (3.2.18) required for the *stationarity* of an AR(2) process.

Autocorrelation Function. Using (3.3.3), the variance of the process is

$$\gamma_0 = \sigma_a^2(1 + \theta_1^2 + \theta_2^2)$$

and using (3.3.4), the autocorrelation function is

$$\begin{aligned} \rho_1 &= \frac{-\theta_1(1-\theta_2)}{1+\theta_1^2+\theta_2^2} \\ \rho_2 &= \frac{-\theta_2}{1+\theta_1^2+\theta_2^2} \\ \rho_k &= 0 \qquad k > 2 \end{aligned} \tag{3.3.11}$$

Thus, the autocorrelation function has a cutoff after lag 2.

It follows from (3.3.10) and (3.3.11) that the first two autocorrelations of an invertible MA(2) process must lie within the area bounded by segments of the curves

$$\begin{aligned} \rho_2 + \rho_1 &= -0.5 \\ \rho_2 - \rho_1 &= -0.5 \\ \rho_1^2 &= 4\rho_2(1 - 2\rho_2) \end{aligned} \tag{3.3.12}$$

The invertibility region (3.3.10) for the parameters is shown in Figure 3.6(a) and the corresponding admissible region (3.3.12) for the autocorrelations in Figure 3.6(b). The latter shows whether a given pair of autocorrelations ρ_1 and ρ_2 is consistent with the assumption that the model is an MA(2) process. If they are consistent, the values of the parameters θ_1 and θ_2 can be obtained by solving the nonlinear equations (3.3.11). To facilitate this calculation, Chart C in the Collection of Tables and Charts in Part Five has been prepared so that the values of θ_1 and θ_2 can be read off directly, given ρ_1 and ρ_2.

Spectrum. Using (3.3.5), the spectrum of the MA(2) process is

$$\begin{aligned} p(f) &= 2\sigma_a^2|1 - \theta_1 e^{-i2\pi f} - \theta_2 e^{-i4\pi f}|^2 \\ &= 2\sigma_a^2[1 + \theta_1^2 + \theta_2^2 - 2\theta_1(1-\theta_2)\cos(2\pi f) - 2\theta_2\cos(4\pi f)] \\ & \qquad\qquad 0 < f < \frac{1}{2} \end{aligned} \tag{3.3.13}$$

and is the reciprocal of the spectrum (3.2.29) of a second-order autoregressive process, apart from the constant $2\sigma_a^2$.

Partial Autocorrelation Function. The exact expression for the partial autocorrelation function of an MA(2) process is complicated, but it is dominated by the sum of two exponentials if the roots of the characteristic equation $1 - \theta_1 B - \theta_2 B^2 = 0$ are real, and by a damped sine wave if the roots are complex. Thus, it behaves like the autocorrelation function of an AR(2) process. The autocorrelation functions and partial autocorrelation functions for various values of the parameters within the invertible region are shown in Figure 3.7. Comparison of Figure 3.7 with Figure 3.2, which shows the corresponding

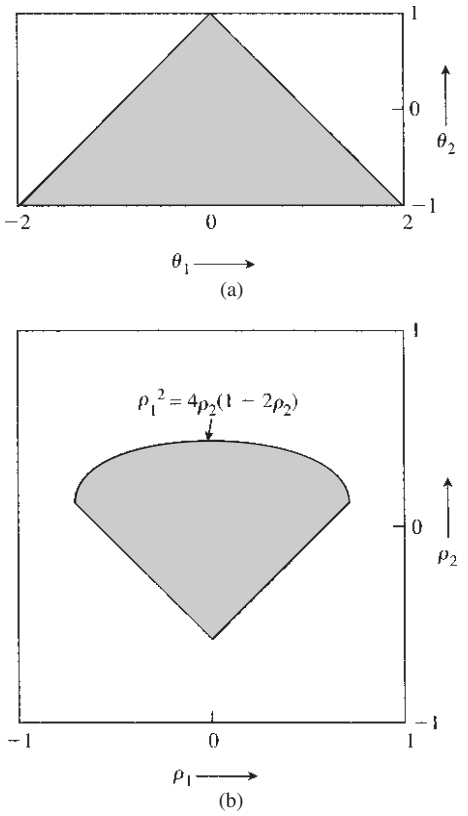


FIGURE 3.6 Admissible regions for (a) θ_1, θ_2 and (b) ρ_1, ρ_2 for an invertible MA(2) process.

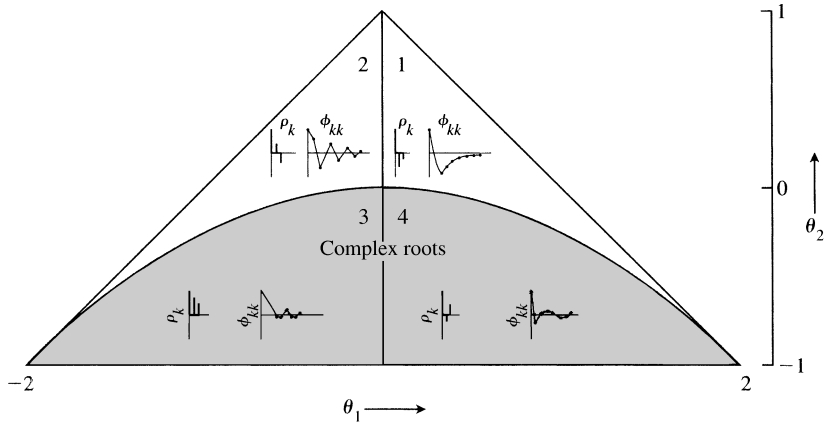


FIGURE 3.7 Autocorrelation and partial autocorrelation functions ρ_k and ϕ_{kk} for various MA(2) models.

autocorrelations and partial autocorrelations for an AR(2) process, illustrates the duality between the MA(2) and the AR(2) processes.

Example. For illustration, consider the second-order moving average model

$$\tilde{z}_t = a_t - 0.8a_{t-1} + 0.5a_{t-2}$$

The variance of the process is $\gamma_0 = \sigma_a^2(1 + (0.8)^2 + (-0.5)^2) = 1.89\sigma_a^2$, and from (3.3.11) the theoretical autocorrelations are

$$\rho_1 = \frac{-0.8(1-(-0.5))}{1+(0.8)^2+(-0.5)^2} = \frac{-1.20}{1.89} = -0.635 \qquad \rho_2 = \frac{-(-0.5)}{1.89} = 0.265$$

and $\rho_k = 0$, for $k > 2$. The theoretical partial autocorrelations are obtained by solving (3.2.31) successively; the first several values are $\phi_{11} = \rho_1 = -0.635$, $\phi_{22} = (\rho_2 - \rho_1^2)/(1 - \rho_1^2) = -0.232$, $\phi_{33} = 0.105$, $\phi_{44} = 0.191$, and $\phi_{55} = 0.102$.

Figure 3.8 shows the autocorrelation and partial autocorrelation functions up to 15 lags for this example. Note the partial autocorrelations ϕ_{kk} display an approximate damped sinusoidal behavior with moderate rate of damping, similar to the behavior depicted for region 4 in Figure 3.7. This is consistent with the fact that the roots of $\theta(B) = 0$ are complex with modulus (damping factor) $D = \sqrt{0.5} \simeq 0.71$ and frequency $f_0 = \cos^{-1}(0.5657)/(2\pi) = 1/6.48$ in this example.

The autocorrelation and partial autocorrelation functions shown in Figure 3.8 were generated using the function ARMAacf() in the R stats package. The commands needed to reproduce the graph are shown below. Note that the moving average parameters in the ARMAacf() function are again entered with their signs reversed since R uses positive signs in defining the moving average operator, rather than the negative signs used here.

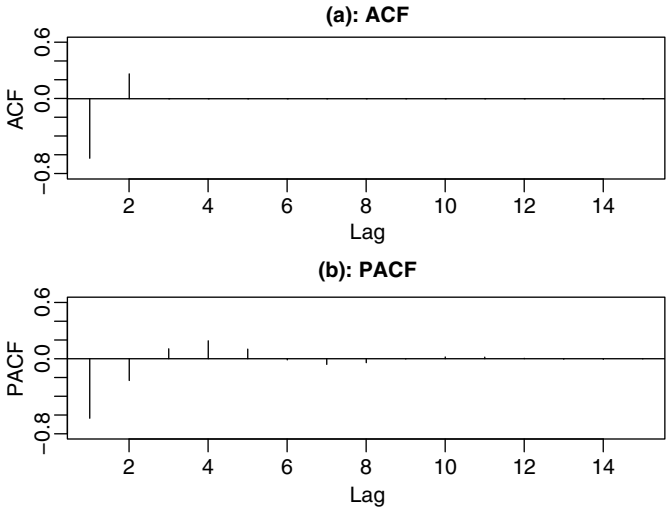


FIGURE 3.8 (a) Autocorrelation function and (b) partial autocorrelation function for the MA(2) model $\tilde{z}_t = a_t - 0.8a_{t-1} + 0.5a_{t-2}$.

```
> ACF=ARMAacf(ar=0,ma=c(-0.8,+0.5),lag.max=15,pacf=FALSE)[-1]
> PACF=ARMAacf(ar=0,ma=c(-0.8,+0.5),lag.max=15,pacf=TRUE)
> par(mfrow=c(2,1))
> plot(ACF,type='h',ylim=c(-0.8,0.6),xlab='lag',main='(a): ACF')
> abline(h=0)
> plot(PACF,type='h',ylim=c(-0.8,0.6),xlab='lag',main='(b):PACF')
> abline(h=0)
> ACF    % Retrieves the autocorrelation coefficients
> PACF   % Retrieves the partial autocorrelation coefficients
```

3.3.5 Duality Between Autoregressive and Moving Average Processes

The previous sections have examined the properties of autoregressive and moving average processes and discussed the *duality* between these processes. As illustrated in Table 3.2 at the end of this chapter, this duality has the following consequences:

1. In a stationary autoregressive process of order p, a_t can be represented as a *finite* weighted sum of previous $\tilde{z}$'s, or $\tilde{z}_t$ as an infinite weighted sum
$$\tilde{z}_t = \phi^{-1}(B)a_t$$
of previous a's. Conversely, an invertible moving average process of order q, $\tilde{z}_t$, can be represented as a finite weighted sum of previous a's, or a_t as an infinite weighted sum
$$\theta^{-1}(B)\tilde{z}_t = a_t$$
of previous $\tilde{z}$'s.
2. The finite MA process has an autocorrelation function that is zero beyond a certain point, but since it is equivalent to an infinite AR process, its partial autocorrelation function is infinite in extent and is dominated by damped exponentials and/or damped sine waves. Conversely, the AR process has a partial autocorrelation function that is zero beyond a certain point, but its autocorrelation function is infinite in extent and consists of a mixture of damped exponentials and/or damped sine waves.
3. For an autoregressive process of finite order p, the parameters are not required to satisfy any conditions to ensure invertibility. However, for stationarity, the roots of $\phi(B) = 0$ must lie outside the unit circle. Conversely, the parameters of the MA process are not required to satisfy any conditions to ensure stationarity. However, for invertibility, the roots of $\theta(B) = 0$ must lie outside the unit circle.
4. The spectrum of a moving average process has an inverse relationship to the spectrum of the corresponding autoregressive process.

3.4 MIXED AUTOREGRESSIVE–MOVING AVERAGE PROCESSES

3.4.1 Stationarity and Invertibility Properties

We have noted earlier that to achieve parsimony it may be necessary to include both autoregressive and moving average terms. Thus, we may need to employ the mixed ARMA model

$$\tilde{z}_t = \phi_1\tilde{z}_{t-1} + \cdots + \phi_p\tilde{z}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{3.4.1}$$

that is,

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)\tilde{z}_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)a_t$$

or

$$\phi(B)\tilde{z}_t = \theta(B)a_t$$

where $\phi(B)$ and $\theta(B)$ are polynomial operators in B of degrees p and q.

We subsequently refer to this process as an ARMA(p, q) process. It may be thought of in two ways:

1. As a pth-order autoregressive process

$$\phi(B)\tilde{z}_t = e_t$$

with e_t following the qth-order moving average process $e_t = \theta(B)a_t$.

2. As a qth-order moving average process

$$\tilde{z}_t = \theta(B)b_t$$

with b_t following the pth-order autoregressive process $\phi(B)b_t = a_t$ so that

$$\phi(B)\tilde{z}_t = \theta(B)\phi(B)b_t = \theta(B)a_t$$

It is obvious that moving average terms on the right of (3.4.1) will not affect the earlier arguments, which establish conditions for stationarity of an autoregressive process. Thus, $\phi(B)\tilde{z}_t = \theta(B)a_t$ will define a stationary process provided that the characteristic equation $\phi(B) = 0$ has all its roots outside the unit circle. Similarly, the roots of $\theta(B) = 0$ must lie outside the unit circle if the process is to be invertible.

Thus, the stationary and invertible ARMA(p, q) process (3.4.1) has both the infinite moving average representation

$$\tilde{z}_t = \psi(B)a_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}$$

where $\psi(B) = \phi^{-1}(B)\theta(B)$, and the infinite autoregressive representation

$$\pi(B)\tilde{z}_t = \tilde{z}_t - \sum_{j=1}^{\infty} \pi_j \tilde{z}_{t-j} = a_t$$

where $\pi(B) = \theta^{-1}(B)\phi(B)$, with both the ψ_j weights and the π_j weights being absolutely summable. The weights ψ_j are determined from the relation $\phi(B)\psi(B) = \theta(B)$ to satisfy

$$\psi_j = \phi_1\psi_{j-1} + \phi_2\psi_{j-2} + \cdots + \phi_p\psi_{j-p} - \theta_j \qquad j > 0$$

with $\psi_0 = 1$, $\psi_j = 0$ for $j < 0$, and $\theta_j = 0$ for $j > q$, while from the relation $\theta(B)\pi(B) = \phi(B)$ the π_j are determined to satisfy

$$\pi_j = \theta_1\pi_{j-1} + \theta_2\pi_{j-2} + \cdots + \theta_q\pi_{j-q} + \phi_j \qquad j > 0$$

with the $\pi_0 = -1$, $\pi_j = 0$ for $j < 0$, and $\phi_j = 0$ for $j > p$. From these relations, the ψ_j and π_j weights can readily be computed recursively in terms of the ϕ_i and θ_i coefficients.

3.4.2 Autocorrelation Function and Spectrum of Mixed Processes

Autocorrelation Function. The autocorrelation function of the mixed process may be derived by a method similar to that used for autoregressive processes in Section 3.2.2. On multiplying throughout in (3.4.1) by $\tilde{z}_{t-k}$ and taking expectations, we see that the autocovariance function satisfies the difference equation

$$\gamma_k = \phi_1\gamma_{k-1} + \cdots + \phi_p\gamma_{k-p} + \gamma_{za}(k) - \theta_1\gamma_{za}(k-1) - \cdots - \theta_q\gamma_{za}(k-q)$$

where $\gamma_{za}(k)$ is the cross-covariance function between z and a and is defined by $\gamma_{za}(k) = E[\tilde{z}_{t-k}a_t]$. Since z_{t-k} depends only on shocks that have occurred up to time $t-k$ through the infinite moving average representation $\tilde{z}_{t-k} = \psi(B)a_{t-k} = \sum_{j=0}^{\infty}\psi_j a_{t-k-j}$, it follows that

$$\gamma_{za}(k) = \begin{cases} 0 & k > 0 \\ \psi_{-k}\sigma_a^2 & k \le 0 \end{cases}$$

Hence, the preceding equation for γ_k may be expressed as

$$\gamma_k = \phi_1\gamma_{k-1} + \cdots + \phi_p\gamma_{k-p} - \sigma_a^2(\theta_k\psi_0 + \theta_{k+1}\psi_1 + \cdots + \theta_q\psi_{q-k}) \tag{3.4.2}$$

with the convention that $\theta_0 = -1$. We see that this implies

$$\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + \cdots + \phi_p\gamma_{k-p} \qquad k \ge q+1$$

and hence

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \cdots + \phi_p\rho_{k-p} \qquad k \ge q+1 \tag{3.4.3}$$

or

$$\phi(B)\rho_k = 0 \qquad k \ge q+1$$

Thus, for the ARMA(p, q) process, there will be q autocorrelations $\rho_1, \ldots, \rho_q$ whose values depend directly on the choice of the q moving average parameters θ_i, as well as on the p autoregressive parameters ϕ_j. Also, the p values $\rho_{q-p+1}, \ldots, \rho_q$ provide the necessary starting values for the difference equation $\phi(B)\rho_k = 0$, where $k \ge q+1$, which then entirely determines the autocorrelations at higher lags. If $q - p < 0$, the whole autocorrelation function ρ_j, for $j = 0, 1, 2, \ldots$, will consist of a mixture of damped exponentials and/or damped sine waves, whose nature is dictated by (the roots of) the polynomial $\phi(B)$ and the starting values. If, however, $q - p \ge 0$, there will be $q - p + 1$ initial values $\rho_0, \rho_1, \ldots, \rho_{q-p}$, which do not follow this general pattern. These facts are useful in identifying mixed series.

Variance. When $k = 0$, we have

$$\gamma_0 = \phi_1\gamma_1 + \cdots + \phi_p\gamma_p + \sigma_a^2(1 - \theta_1\psi_1 - \cdots - \theta_q\psi_q) \tag{3.4.4}$$

which has to be solved along with the p equations (3.4.2) for $k = 1, 2, \ldots p$ to obtain $\gamma_0, \gamma_1, \ldots, \gamma_p$.

Spectrum. Using (3.1.12), the spectrum of the mixed ARMA(p, q) process is

$$\begin{aligned} p(f) &= 2\sigma_a^2 \frac{|\theta(e^{-i2\pi f})|^2}{|\phi(e^{-i2\pi f})|^2} \\ &= 2\sigma_a^2 \frac{|1 - \theta_1 e^{-i2\pi f} - \cdots - \theta_q e^{-i2q\pi f}|^2}{|1 - \phi_1 e^{-i2\pi f} - \cdots - \phi_p e^{-i2p\pi f}|^2} \qquad 0 \le f \le \frac{1}{2} \end{aligned} \tag{3.4.5}$$

Partial Autocorrelation Function. The mixed process $\phi(B)\tilde{z}_t = \theta(B)a_t$ can be written as

$$a_t = \theta^{-1}(B)\phi(B)\tilde{z}_t$$

where $\theta^{-1}(B)$ is an infinite series in B. Hence, the partial autocorrelation function of a mixed process is infinite in extent. It behaves eventually like the partial autocorrelation function of a pure moving average process, being dominated by a mixture of damped exponentials and/or damped sine waves, depending on the order of the moving average and the values of the parameters it contains.

3.4.3 First Order Autoregressive First-Order Moving Average Process

A mixed ARMA process of considerable practical importance is the ARMA(1, 1) process

$$\tilde{z}_t - \phi_1 \tilde{z}_{t-1} = a_t - \theta_1 a_{t-1} \tag{3.4.6}$$

that is,

$$(1 - \phi_1 B)\tilde{z}_t = (1 - \theta_1 B)a_t$$

We now derive some of its more important properties.

Stationarity and Invertibility Conditions. First, we note that the process is stationary if $-1 < \phi_1 < 1$, and invertible if $-1 < \theta_1 < 1$. Hence, the admissible parameter space is the square shown in Figure 3.9(a). In addition, from the relations $\psi_1 = \phi_1\psi_0 - \theta_1 = \phi_1 - \theta_1$ and $\psi_j = \phi_1\psi_{j-1}$ for $j > 1$, we find that the ψ_j weights are given by $\psi_j = (\phi_1 - \theta_1)\phi_1^{j-1}$, $j \ge 1$, and similarly it is easily seen that $\pi_j = (\phi_1 - \theta_1)\theta_1^{j-1}$, $j \ge 1$, for the stationary and invertible ARMA(1, 1) process.

Autocorrelation Function. From (3.4.2) and (3.4.4) we obtain

$$\begin{aligned} \gamma_0 &= \phi_1\gamma_1 + \sigma_a^2(1 - \theta_1\psi_1) \\ \gamma_1 &= \phi_1\gamma_0 - \theta_1\sigma_a^2 \\ \gamma_k &= \phi_1\gamma_{k-1} \qquad k \ge 2 \end{aligned}$$

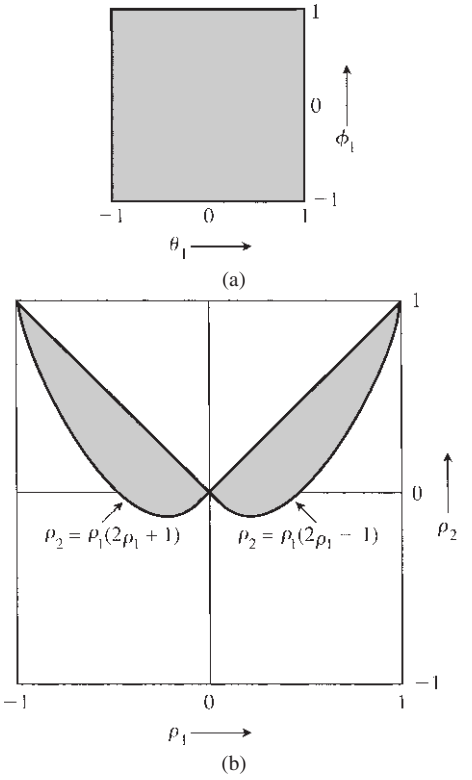


FIGURE 3.9 Admissible regions for (a) ϕ_1, θ_1 and (b) ρ_1, ρ_2 for a stationary and invertible ARMA(1, 1) process.

with $\psi_1 = \phi_1 - \theta_1$. Hence, solving the first two equations for γ_0 and γ_1, the autocovariance function of the process is

$$\begin{aligned}\gamma_0 &= \frac{1 + \theta_1^2 - 2\phi_1\theta_1}{1 - \phi_1^2}\sigma_a^2 \\ \gamma_1 &= \frac{(1 - \phi_1\theta_1)(\phi_1 - \theta_1)}{1 - \phi_1^2}\sigma_a^2 \\ \gamma_k &= \phi_1\gamma_{k-1} \qquad k \geq 2\end{aligned} \tag{3.4.7}$$

The last equation gives $\rho_k = \phi_1\rho_{k-1}$, $k \geq 2$, so that $\rho_k = \rho_1\phi_1^{k-1}$, $k > 1$. Thus, the autocorrelation function decays exponentially from the starting value ρ_1, which depends on θ_1 and ϕ_1.[2] This exponential decay is smooth if ϕ_1 is positive and alternates if ϕ_1 is negative. Furthermore, the sign of ρ_1 is determined by the sign of $(\phi_1 - \theta_1)$ and dictates from which side of zero the exponential decay takes place.

[2]By contrast, the autocorrelation function for the AR(1) process decays exponentially from the starting value $\rho_0 = 1$.

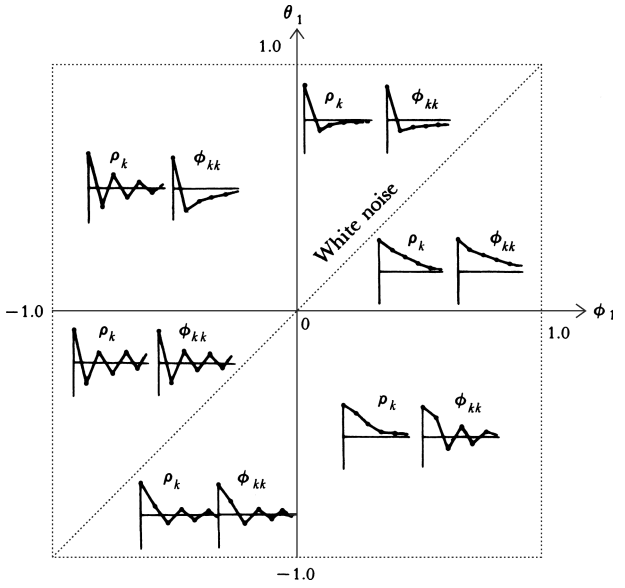


FIGURE 3.10 Autocorrelation and partial autocorrelation functions ρ_k and ϕ_{kk} for various ARMA(1, 1) models.

The first two autocorrelations may be expressed in terms of the parameters of the ARMA(1,1) process, as follows:

$$\rho_1 = \frac{(1 - \phi_1\theta_1)(\phi_1 - \theta_1)}{1 + \theta_1^2 - 2\phi_1\theta_1} \qquad (3.4.8)$$
$$\rho_2 = \phi_1\rho_1$$

Using these expressions and the stationarity and invertibility conditions, it may be shown that ρ_1 and ρ_2 must lie in the region

$$|\rho_2| < |\rho_1|$$
$$\rho_2 > \rho_1(2\rho_1 + 1) \qquad \rho < 0 \qquad (3.4.9)$$
$$\rho_2 > \rho_1(2\rho_1 - 1) \qquad \rho_1 > 0$$

Figure 3.9(b) shows the admissible space for ρ_1 and ρ_2; that is, it indicates which combinations of ρ_1 and ρ_2 are possible for a mixed (1, 1) stationary, invertible process.

Partial Autocorrelation Function. The partial autocorrelation function of the mixed ARMA(1, 1) process consists of a single initial value $\phi_{11} = \rho_1$. Thereafter, it behaves like the partial autocorrelation function of a pure MA(1) process and is dominated by a damped exponential. Thus, as shown in Figure 3.10, when θ_1 is positive, it is dominated by a smoothly damped exponential that decays from a value of ρ_1, with sign determined by the sign of $(\phi_1 - \theta_1)$. Similarly, when θ_1 is negative, it is dominated by an exponential that oscillates as it decays from a value of ρ_1, with sign determined by the sign of $(\phi_1 - \theta_1)$.

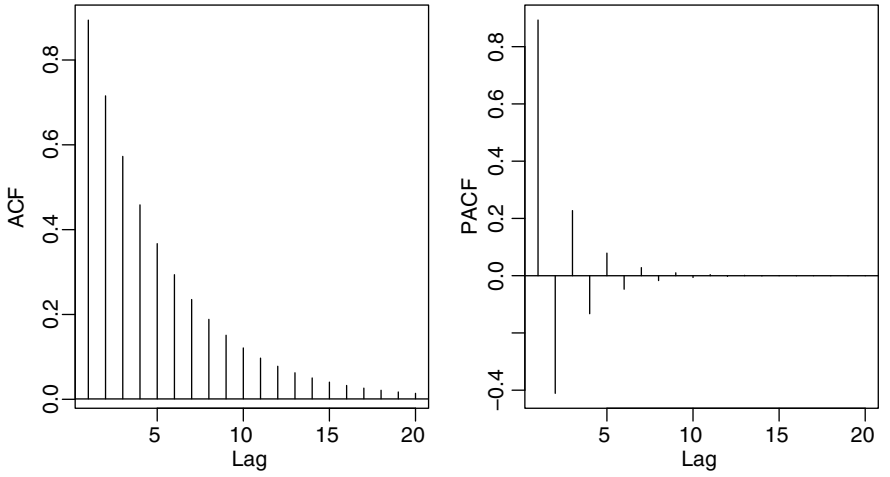


FIGURE 3.11 Theoretical autocorrelation and partial autocorrelation functions of an ARMA(1,1) process with $\phi = 0.8$ and $\theta = -0.6$.

Numerical Example. For numerical illustration, consider the ARMA(1, 1) process,

$$(1 - 0.8B)\tilde{z}_t = (1 + 0.6B)a_t$$

so that $\phi = 0.8$ and $\theta = -0.6$. Further assuming $\sigma_a^2 = 1$, we find from (3.4.7) and (3.4.8) that the variance of $\tilde{z}_t$ is $\gamma_0 = 6.444$, and $\rho_1 = 0.893$. Also, the autocorrelation function satisfies $\rho_j = 0.8\rho_{j-1}$, $j \geq 2$, so that $\rho_j = 0.893(0.8)^{j-1}$, for $j \geq 2$.

The autocorrelation and partial autocorrelation functions are shown in Figure 3.11. The exponential decay in the autocorrelation function is clearly evident from the graph. The partial autocorrelation function also exhibits an exponentially decaying pattern that oscillates in sign due to the negative value of θ. The figure was generated in R using the commands included below. Notice again that the parameter θ, although negative in this example, is entered as + 0.6 since R defines the MA operator $\theta(B)$ as $(1 + \theta B)$ rather that $(1 - \theta B)$ as done in this text.

```
> ACF=ARMAacf(ar=0.8,ma=0.6,20)[-1]
> PACF=ARMAacf(ar=0.8,ma=0.6,20,pacf=TRUE)
> win.graph(width=8,height=4)
> par(mfrow=c(1,2))
> plot(ACF,type="h",xlab="lag");abline(h=0)
> plot(PACF,type="h",xlab="lag");abline(h=0)
```

3.4.4 Summary

Figure 3.12 brings together the admissible regions for the parameters and for the autocorrelations ρ_1, ρ_2 for AR(2), MA(2), and ARMA(1, 1) processes, which are restricted to being both stationary and invertible. Table 3.2 summarizes the properties of mixed ARMA processes and brings together all the important results for autoregressive, moving average, and mixed processes, which will be needed in Chapter 6 to identify models for observed

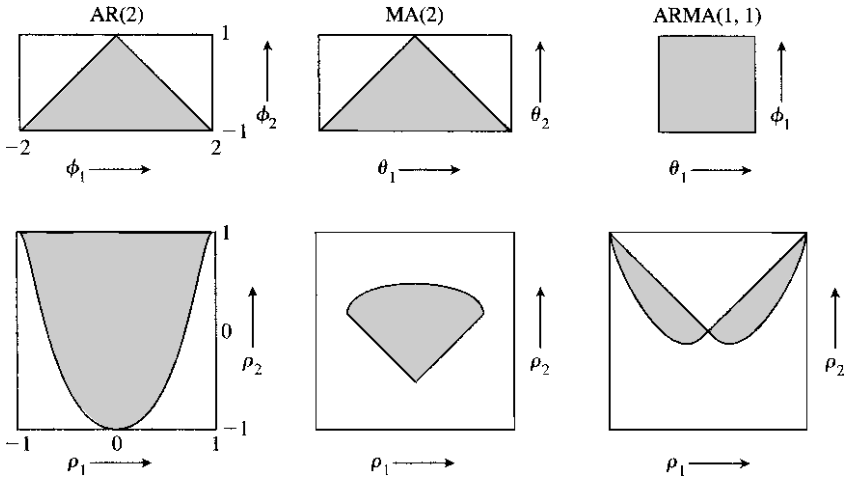


FIGURE 3.12 Admissible regions for the parameters and ρ_1, ρ_2 for AR(2), MA(2), and ARMA(1, 1) processes that are restricted to being both stationary and invertible.

time series. In the next chapter, we extend the mixed ARMA model to produce models that can describe nonstationary behavior of the kind that is frequently met in practice.

APPENDIX A3.1 AUTOCOVARIANCES, AUTOCOVARIANCE GENERATING FUNCTION, AND STATIONARITY CONDITIONS FOR A GENERAL LINEAR PROCESS

Autocovariances. The autocovariance at lag k of the linear process

$$\tilde{z}_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}$$

with $\psi_0 = 1$ is clearly

$$\begin{aligned} \gamma_k &= E[\tilde{z}_t \tilde{z}_{t+k}] \\ &= E\left[\sum_{j=0}^{\infty}\sum_{h=0}^{\infty} \psi_j \psi_h a_{t-j} a_{t+k-h}\right] \\ &= \sigma_a^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k} \end{aligned} \tag{A3.1.1}$$

using the property (3.1.2) for the autocovariance function of white noise.

Autocovariance Generating Function. The result (A3.1.1) may be substituted in the autocovariance generating function

$$\gamma(B) = \sum_{k=-\infty}^{\infty} \gamma_k B^k \tag{A3.1.2}$$

TABLE 3.2 Summary of Properties of Autoregressive, Moving Average, and Mixed ARMA Processes

	Autoregressive Process	Moving Average Processes	Mixed Processes
Model in terms of previous $\tilde{z}'s$	$\phi(B)\tilde{z}_t = a_t$	$\theta^{-1}(B)\tilde{z}_t = a_t$	$\theta^{-1}(B)\phi(B)\tilde{z}_t = a_t$
Model in terms of previous a's	$\tilde{z}_t = \phi^{-1}(B)a_t$	$\tilde{z}_t = \theta(B)a_t$	$\tilde{z}_t = \phi^{-1}(B)\theta(B)a_t$
π weights	Finite series	Infinite series	Infinite series
ψ weights	Infinite series	Finite series	Infinite series
Stationarity condition	Roots of $\phi(B) = 0$ lie outside the unit circle	Always stationary	Roots of $\phi(B) = 0$ lie outside the unit circle
Invertibility condition	Always invertible	Roots of $\theta(B) = 0$ lie outside the unit circle	Roots of $\theta(B) = 0$ lie outside the unit circle
Autocorrelation function	Infinite (damped exponentials and/or damped sine waves)	Finite	Infinite (damped exponentials and/or damped sine waves after first $q - p$ lags)
	Tails off	Cuts off after lag q	Tails off
Partial autocorrelation function	Finite	Infinite (dominated by damped exponentials and/or damped sine waves)	Infinite (dominated by damped exponentials and/or damped sine waves after first $p - q$ lags)
	Cuts off after lag p	Tails off	Tails off

to give

$$\gamma(B) = \sigma_a^2 \sum_{k=-\infty}^{\infty} \sum_{j=0}^{\infty} \psi_j \psi_{j+k} B^k$$
$$= \sigma_a^2 \sum_{j=0}^{\infty} \sum_{k=-j}^{\infty} \psi_j \psi_{j+k} B^k$$

since $\psi_h = 0$ for $h < 0$. Writing $j + k = h$, so that $k = h - j$, we have

$$\gamma(B) = \sigma_a^2 \sum_{j=0}^{\infty} \sum_{h=0}^{\infty} \psi_j \psi_h B^{h-j}$$
$$= \sigma_a^2 \sum_{h=0}^{\infty} \psi_h B^h \sum_{j=0}^{\infty} \psi_j B^{-j}$$

that is,

$$\gamma(B) = \sigma_a^2 \psi(B)\psi(B^{-1}) = \sigma_a^2 \psi(B)\psi(F) \tag{A3.1.3}$$

which is the result (3.1.11) quoted in the text.

Stationarity Conditions. If we substitute $B = e^{-i2\pi f}$ and $F = B^{-1} = e^{i2\pi f}$ in the autocovariance generating function (A3.1.2), we obtain half the power spectrum. Hence, the power spectrum of a linear process is

$$p(f) = 2\sigma_a^2 \psi(e^{-i2\pi f})\psi(e^{i2\pi f})$$
$$= 2\sigma_a^2 |\psi(e^{-i2\pi f})|^2 \qquad 0 \le f \le \tfrac{1}{2} \tag{A3.1.4}$$

It follows that the variance of the process is

$$\sigma_z^2 = \int_0^{1/2} p(f)\,df = 2\sigma_a^2 \int_0^{1/2} \psi(e^{-i2\pi f})\psi(e^{i2\pi f})\,df \tag{A3.1.5}$$

Now if the integral (A3.1.5) is to converge, it may be shown (Grenander and Rosenblatt, 1957) that the infinite series $\psi(B)$ must converge for B on or within the unit circle. More directly, for the linear process $\tilde{z}_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}$, the condition $\sum_{j=0}^{\infty} |\psi_j| < \infty$ of absolute summability of the coefficients ψ_j implies (see Brockwell and Davis, 1991; Fuller, 1996) that the sum $\sum_{j=0}^{\infty} \psi_j a_{t-j}$ converges with probability 1 and hence represents a valid stationary process.

APPENDIX A3.2 RECURSIVE METHOD FOR CALCULATING ESTIMATES OF AUTOREGRESSIVE PARAMETERS

We now show how Yule–Walker estimates for the parameters of an AR($p + 1$) model may be obtained from the estimates for an AR(p) model fitted to the same time series. This recursive method of calculation, which is due to Levinson (1947) and Durbin (1960), can be used to approximate the partial autocorrelation function, as described in Section 3.2.6.

To illustrate the recursion, consider equations (3.2.35). Yule–Walker estimates are obtained for $k = 2, 3$ from

$$\begin{aligned} r_2 &= \hat{\phi}_{21} r_1 + \hat{\phi}_{22} \\ r_1 &= \hat{\phi}_{21} + \hat{\phi}_{22} r_1 \end{aligned} \tag{A3.2.1}$$

and

$$\begin{aligned} r_3 &= \hat{\phi}_{31} r_2 + \hat{\phi}_{32} r_1 + \hat{\phi}_{33} \\ r_2 &= \hat{\phi}_{31} r_1 + \hat{\phi}_{32} + \hat{\phi}_{33} r_1 \\ r_1 &= \hat{\phi}_{31} + \hat{\phi}_{32} r_1 + \hat{\phi}_{33} r_2 \end{aligned} \tag{A3.2.2}$$

The coefficients $\hat{\phi}_{31}$ and $\hat{\phi}_{32}$ may be expressed in terms of $\hat{\phi}_{33}$ using the last two equations of (A3.2.2). The solution may be written in matrix form as

$$\begin{pmatrix} \hat{\phi}_{31} \\ \hat{\phi}_{32} \end{pmatrix} = \mathbf{R}_2^{-1} \begin{pmatrix} r_2 - \hat{\phi}_{33} r_1 \\ r_1 - \hat{\phi}_{33} r_2 \end{pmatrix} \tag{A3.2.3}$$

where

$$\mathbf{R}_2 = \begin{bmatrix} r_1 & 1 \\ 1 & r_1 \end{bmatrix}$$

Now, (A3.2.3) may be written as

$$\begin{bmatrix} \hat{\phi}_{31} \\ \hat{\phi}_{32} \end{bmatrix} = \mathbf{R}_2^{-1} \begin{bmatrix} r_2 \\ r_1 \end{bmatrix} - \hat{\phi}_{33} \mathbf{R}_2^{-1} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \tag{A3.2.4}$$

Using the fact that (A3.2.1) may also be written as

$$\begin{bmatrix} \hat{\phi}_{21} \\ \hat{\phi}_{22} \end{bmatrix} = \mathbf{R}_2^{-1} \begin{bmatrix} r_2 \\ r_1 \end{bmatrix}$$

it follows that (A3.2.4) becomes

$$\begin{bmatrix} \hat{\phi}_{31} \\ \hat{\phi}_{32} \end{bmatrix} = \begin{bmatrix} \hat{\phi}_{21} \\ \hat{\phi}_{22} \end{bmatrix} - \hat{\phi}_{33} \begin{bmatrix} \hat{\phi}_{22} \\ \hat{\phi}_{21} \end{bmatrix}$$

that is,

$$\begin{aligned} \hat{\phi}_{31} &= \hat{\phi}_{21} - \hat{\phi}_{33} \hat{\phi}_{22} \\ \hat{\phi}_{32} &= \hat{\phi}_{22} - \hat{\phi}_{33} \hat{\phi}_{21} \end{aligned} \tag{A3.2.5}$$

To complete the calculation of $\hat{\phi}_{31}$ and $\hat{\phi}_{32}$, we need an expression for $\hat{\phi}_{33}$. On substituting (A3.2.5) in the first of the equations (A3.2.2), we obtain

$$\hat{\phi}_{33} = \frac{r_3 - \hat{\phi}_{21} r_2 - \hat{\phi}_{22} r_1}{1 - \hat{\phi}_{21} r_1 - \hat{\phi}_{22} r_2} \tag{A3.2.6}$$

Thus, the partial autocorrelation $\hat{\phi}_{33}$ is first calculated from $\hat{\phi}_{21}$ and $\hat{\phi}_{22}$, using (A3.2.6), and then the other two coefficients, $\hat{\phi}_{31}$ and $\hat{\phi}_{32}$, may be obtained from (A3.2.5).

In general, the recursive formulas are

$$\hat{\phi}_{p+1,j} = \hat{\phi}_{pj} - \hat{\phi}_{p+1,p+1}\hat{\phi}_{p,p+1-j} \qquad j = 1, 2, \ldots, p \tag{A3.2.7}$$

$$\hat{\phi}_{p+1,p+1} = \frac{r_{p+1} - \sum_{j=1}^{p} \hat{\phi}_{pj} r_{p+1-j}}{1 - \sum_{j=1}^{p} \hat{\phi}_{pj} r_j} \tag{A3.2.8}$$

EXERCISES

3.1 Write the following models in B notation:

(1) $\tilde{z}_t - 0.5\tilde{z}_{t-1} = a_t$

(2) $\tilde{z}_t = a_t - 1.3a_{t-1} + 0.4a_{t-2}$

(3) $\tilde{z}_t - 0.5\tilde{z}_{t-1} = a_t - 1.3a_{t-1} + 0.4a_{t-2}$

3.2 For each of the models of Exercise 3.1 and also for the following models, state whether it is (**a**) stationary or (**b**) invertible.

(4) $\tilde{z}_t - 1.5\tilde{z}_{t-1} + 0.6\tilde{z}_{t-2} = a_t$

(5) $\tilde{z}_t - \tilde{z}_{t-1} = a_t - 0.5a_{t-1}$

(6) $\tilde{z}_t - \tilde{z}_{t-1} = a_t - 1.3a_{t-1} + 0.3a_{t-2}$

3.3. For each of the models in Exercise 3.1, obtain:

(a) The first four ψ_j weights

(b) The first four π_j weights

(c) The autocovariance generating function

(d) The first four autocorrelations ρ_j

(e) The variance of $\tilde{z}_t$ assuming that $\sigma_a^2 = 1.0$

3.4. Calculate the first fifteen ψ_j weights for each of the three models in Exercise 3.2 using the function ARMAtoMA in R. See help(ARMAtoMA) for details.

3.5. Classify each of the models (1) to (4) in Exercises 3.1 and 3.2 as a member of the class of ARMA(p, q) processes.

3.6. **(a)** Write down the Yule–Walker equations for models (1) and (4) considered in Exercises 3.1 and 3.2.

(b) Solve these equations to obtain ρ_1 and ρ_2 for the two models.

(c) Obtain the partial autocorrelation function for the two models.

3.7. Consider the first-order autoregressive model $z_t = \theta_0 + \phi z_{t-1} + a_t$, where the constant θ_0 is a function of the mean of the series.

(a) Derive the autocovariances $\gamma_k = E([z_t - \mu][z_{t-k} - \mu])$ for this series.

(b) Calculate and plot the autocorrelation function for $\phi = 0.8$ using the R command ARMAacf(); see help(ARMAacf) for details.

(c) Calculate and plot the partial autocorrelation function for the same process.

3.8. Consider the mixed ARMA(1,1) model $z_t - \phi z_{t-1} = a_t - \theta a_{t-1}$, where $-1 < \phi < 1$ and $E(z_t)$ is assumed to be zero for convenience.

(a) Derive the autocovariances $\gamma_k = E([z_t - \mu][z_{t-k} - \mu])$ for this series.

(b) Calculate and plot the autocorrelation function for $\phi = 0.9$ and $\theta = -0.3$ using R (see Exercise 3.7).

(c) Calculate and plot the partial autocorrelation function for the same process.

3.9. For the AR(2) process $\tilde{z}_t - 1.0\tilde{z}_{t-1} + 0.5\tilde{z}_{t-2} = a_t$:

(a) Calculate ρ_1.

(b) Using ρ_0 and ρ_1 as starting values and the difference equation form for the autocorrelation function, calculate the values of ρ_k for $k = 2, \ldots, 15$.

(c) Use the plotted autocorrelation function to estimate the period and damping factor of the autocorrelation function.

(d) Check the values in (c) by direct calculation using the parameter values and the related roots G_1^{-1} and G_2^{-1} of $\phi(B) = 1 - 1.0B + 0.5B^2$.

3.10. **(a)** Plot the power spectrum $g(f)$ of the autoregressive process of Exercise 3.9, and show that it has a peak at a period that is close to the period in the autocorrelation function.

(b) Graphically, or otherwise, estimate the proportion of the variance of the series in the frequency band between $f = 0.0$ and $f = 0.2$ cycle per data interval.

3.11. **(a)** Why is it important to factorize the autoregressive and moving average operators after fitting a model to an observed series?

(b) It was shown by Jenkins (1975) that the number of mink skins z_t traded annually between 1848 and 1909 in North Canada is adequately represented by the AR(4) model

$$(1 - 0.82B + 0.22B^2 + 0.28B^4)[\ln(z_t) - \mu] = a_t$$

Factorize the autoregressive operator and explain what the factors reveal about the autocorrelation function and the underlying nature of the mink series. The data for the period 1850–1911 are listed as Series N in Part Five of this book. Note that the roots of $\phi(B) = 0$ can be calculated using the R commond polyroot(), where the autoregressive parameters are entered with their signs reversed; see help(polyroot) for details.

3.12. Calculate and plot the theoretical autocorrelation function and partial autocorrelation function for the AR(4) model specified in Exercise 3.11(b).

4

LINEAR NONSTATIONARY MODELS

Many empirical time series (e.g., stock price series) behave as though they had no fixed mean. Even so, they exhibit homogeneity in the sense that apart from local level, or perhaps local level and trend, one part of the series behaves much like any other part. Models that describe such homogeneous nonstationary behavior can be obtained by assuming that some suitable *difference* of the process is stationary. In this chapter, we examine the properties of the important class of models for which the dth difference of the series is a stationary mixed autoregressive–moving average process. These models are called autoregressive integrated moving average (ARIMA) processes.

4.1 AUTOREGRESSIVE INTEGRATED MOVING AVERAGE PROCESSES

4.1.1 Nonstationary First-Order Autoregressive Process

Figure 4.1 shows four time series that have arisen in forecasting and control problems. All of them exhibit behavior suggestive of nonstationarity. Series A, C, and D represent ''uncontrolled'' outputs (concentration, temperature, and viscosity, respectively) from three different chemical processes. These series were collected to show the effect on these outputs of uncontrolled and unmeasured disturbances such as variations in feedstock and ambient temperature. The temperature Series C was obtained by temporarily disconnecting the controllers on the pilot plant involved and recording the subsequent temperature fluctuations. Both A and D were collected on full-scale processes, where it was necessary to maintain some output quality characteristic as close as possible to a fixed level. To achieve this control, another variable had been manipulated to approximately cancel out variations in the output. However, the effect of these manipulations on the output was accurately

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

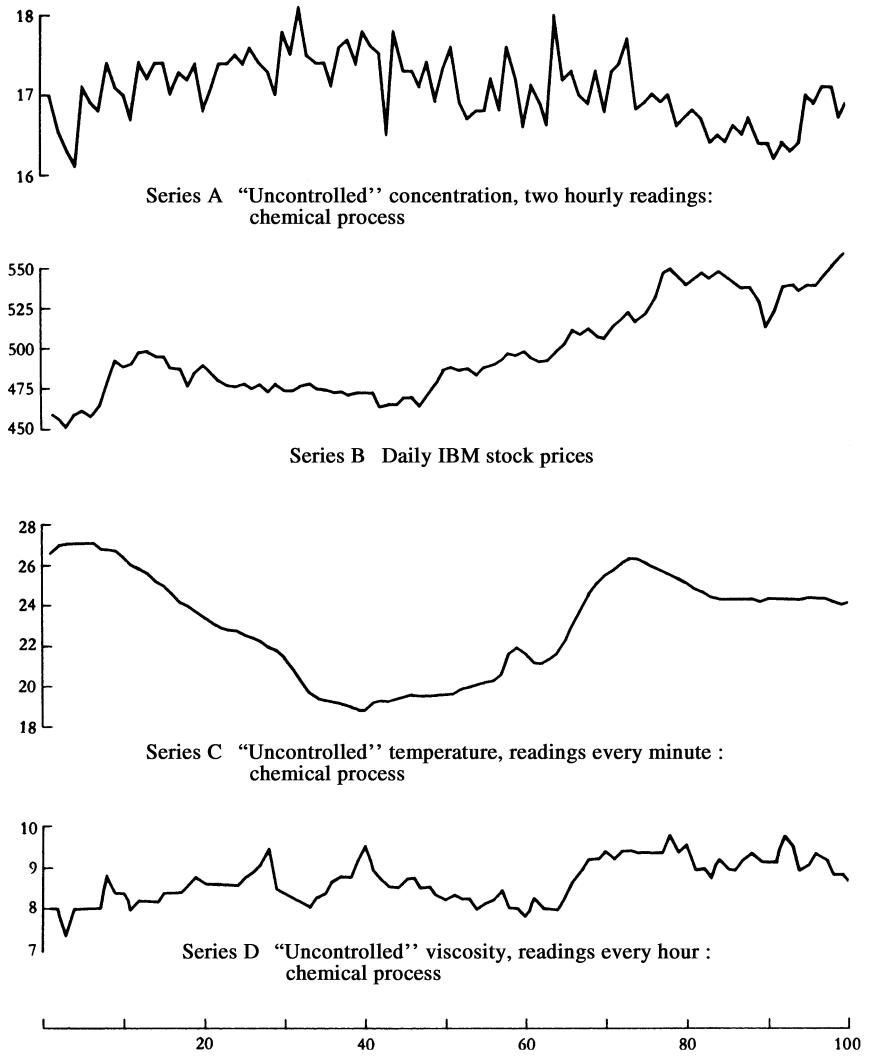


FIGURE 4.1 Typical time series arising in forecasting and control problems.

known in each case, so that it was possible to compensate numerically for the control action. That is, it was possible to calculate very nearly the values of the series that would have been obtained if no corrective action had been taken. It is these compensated values that are recorded here and referred to as the "uncontrolled" series. Series B consists of the daily IBM stock prices during a period beginning in May 1961. A complete list of all the series is given in the collection of time series at the end of this book. In Figure 4.1, 100 successive observations have been plotted from each series and the points joined by straight lines.

There are an unlimited number of ways in which a process can be nonstationary. However, the types of economic and industrial series that we wish to analyze frequently exhibit a particular kind of homogeneous nonstationary behavior that can be represented by a stochastic model, which is a modified form of the autoregressive–moving average

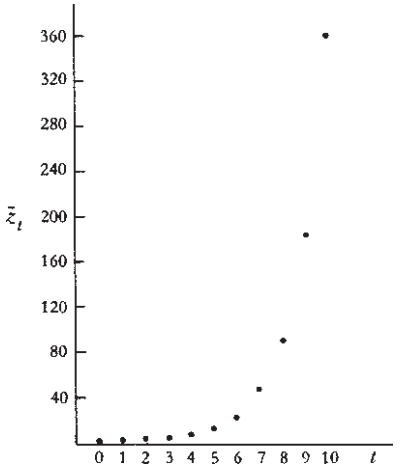


FIGURE 4.2 Realization of the nonstationary first-order autoregressive process $\tilde{z}_t = 2\tilde{z}_{t-1} + a_t$ with $\sigma_a^2 = 1$.

(ARMA) model. In Chapter 3, we considered the mixed ARMA model

$$\phi(B)\tilde{z}_t = \theta(B)a_t \tag{4.1.1}$$

with $\phi(B)$ and $\theta(B)$ polynomial operators in B, of degree p and q, respectively. To ensure stationarity, the roots of $\phi(B) = 0$ must lie outside the unit circle. A natural way of obtaining nonstationary processes is to relax this restriction.

To gain some insight into the possibilities, consider the first-order autoregressive model,

$$(1 - \phi B)\tilde{z}_t = a_t \tag{4.1.2}$$

which is stationary for $|\phi| < 1$. Let us study the behavior of this process for $\phi = 2$, a value outside the stationary range. Figure 4.2 shows a series $\tilde{z}_t$ generated by the model $\tilde{z}_t = 2\tilde{z}_{t-1} + a_t$ using a set of unit random normal deviates a_t and setting $\tilde{z}_0 = 0.7$. It is seen that after a short induction period, the series ''breaks loose'' and essentially follows an exponential curve, with the generating a_t's playing almost no further part. The behavior of series generated by processes of higher order, which violate the stationarity condition, is similar. Furthermore, this behavior is essentially the same whether or not moving average terms are introduced on the right of the model.

4.1.2 General Model for a Nonstationary Process Exhibiting Homogeneity

Autoregressive Integrated Moving Average Model. Although nonstationary models of the kind described above are of value to represent explosive or evolutionary behavior (such as bacterial growth), the applications that we describe in this book are not of this type. So far, we have seen that an ARMA process is stationary if the roots of $\phi(B) = 0$ lie *outside* the unit circle, and exhibits explosive nonstationary behavior if the roots lie *inside* the unit circle. The only case remaining is that the roots of $\phi(B) = 0$ lie *on* the unit circle. It turns out that the resulting models are of great value in representing homogeneous nonstationary

time series. In particular, nonseasonal series are often well represented by models in which one or more of these roots are *unity* and these are considered in the present chapter[1].

Let us consider the model

$$\varphi(B)\tilde{z}_t = \theta(B)a_t \tag{4.1.3}$$

where $\varphi(B)$ is a nonstationary autoregressive operator such that d of the roots of $\varphi(B) = 0$ are unity and the remainder lie outside the unit circle. Then the model can be written as

$$\varphi(B)\tilde{z}_t = \phi(B)(1-B)^d\tilde{z}_t = \theta(B)a_t \tag{4.1.4}$$

where $\phi(B)$ is a *stationary* autoregressive operator. Since $\nabla^d\tilde{z}_t = \nabla^d z_t$, for $d \geq 1$, where $\nabla = 1 - B$ is the differencing operator, we can write the model as

$$\phi(B)\nabla^d z_t = \theta(B)a_t \tag{4.1.5}$$

Equivalently, the process is defined by the two equations

$$\phi(B)w_t = \theta(B)a_t \tag{4.1.6}$$

and

$$w_t = \nabla^d z_t \tag{4.1.7}$$

Thus, we see that the model corresponds to assuming that the dth difference of the series can be represented by a stationary, invertible ARMA process. An alternative way of looking at the process for $d \geq 1$ results from inverting (4.1.7) to give

$$z_t = S^d w_t \tag{4.1.8}$$

where S is the infinite summation operator defined by

$$\begin{aligned} Sx_t = \sum_{h=-\infty}^{t} x_h &= (1 + B + B^2 + \cdots)x_t \\ &= (1-B)^{-1}x_t = \nabla^{-1}x_t \end{aligned}$$

Thus,

$$S = (1-B)^{-1} = \nabla^{-1}$$

The operator S^2 is similarly defined as

$$\begin{aligned} S^2x_t &= Sx_t + Sx_{t-1} + Sx_{t-2} + \cdots \\ &= \sum_{i=-\infty}^{t}\sum_{h=-\infty}^{i} x_h = (1 + 2B + 3B^2 + \cdots)x_t \end{aligned}$$

and so on for higher order d. Equation (4.1.8) implies that the process (4.1.5) can be obtained by summing (or ''integrating'') the stationary process (4.1.6) d times. Therefore, we call the process (4.1.5) an *autoregressive integrated moving average (ARIMA) process*.

[1]In Chapter 9, we consider models, capable of representing seasonality of period s, for which the characteristic equation has roots lying on the unit circle that are the sth roots of unity.

The ARIMA models for nonstationary time series, which were also considered earlier by Yaglom (1955), are of fundamental importance for forecasting and control as discussed by Box and Jenkins (1962, 1963, 1965, 1968a, 1968b, 1969) and Box et al. (1967a). Nonstationary processes were also discussed by Zadeh and Ragazzini (1950), Kalman (1960), and Kalman and Bucy (1961). An earlier procedure for time series analysis that employed differencing was the *variate difference method* (see Tintner (1940) and Rao and Tintner (1963)). However, the motivation, methods, and objectives of this procedure were quite different from those discussed here.

Technically, the *infinite* summation operator $S = (1 - B)^{-1}$ in (4.1.8) cannot actually be used in defining the nonstationary ARIMA processes, since the infinite sums involved will not be convergent. Instead, we can consider the finite summation operator S_m for any positive integer m, given by

$$S_m = (1 + B + B^2 + \cdots + B^{m-1}) \equiv \frac{1 - B^m}{1 - B}$$

Similarly, the finite double summation operator can be defined as

$$\begin{aligned} S_m^{(2)} &= \sum_{j=0}^{m-1} \sum_{i=j}^{m-1} B^i = (1 + 2B + 3B^2 + \cdots + mB^{m-1}) \\ &\equiv \frac{1 - B^m - mB^m(1 - B)}{(1 - B)^2} \end{aligned}$$

since $(1 - B)S_m^{(2)} = S_m - mB^m$, and so on. Then the relation between an integrated ARMA process z_t with $d = 1$, for example, and the corresponding stationary ARMA process $w_t = (1 - B)z_t$, in terms of values back to some earlier time origin $k < t$, can be expressed as

$$z_t = \frac{S_{t-k}}{1 - B^{t-k}} w_t = \frac{1}{1 - B^{t-k}} (w_t + w_{t-1} + \cdots + w_{k+1})$$

so that $z_t = w_t + w_{t-1} + \cdots + w_{k+1} + z_k$ can be thought of as the sum of a finite number of terms from the stationary process w plus an initializing value of the process z at time k. Hence, in the formal definition of the stochastic properties of a nonstationary ARIMA process as generated in (4.1.3), it would typically be necessary to specify initializing conditions for the process at some time point k in the finite (but possibly remote) past. However, these initial condition specifications will have little effect on most of the important characteristics of the process, and such specifications will for the most part not be emphasized in this book.

As mentioned in Chapter 1, the model (4.1.5) is equivalent to representing the process z_t as the output from a linear filter (unless $d = 0$, this is an *unstable* linear filter), whose input is white noise a_t. Alternatively, we can regard it as a device *for transforming the highly dependent, and possibly nonstationary process z_t, to a sequence of uncorrelated random variables a_t*, that is, for transforming the process to white noise.

If in (4.1.5), the autoregressive operator $\phi(B)$ is of order p, the dth difference is taken, and the moving average operator $\theta(B)$ is of order q, we say that we have an ARIMA model of order (p, d, q), or simply an ARIMA(p, d, q) process.

Two Interpretations of the ARIMA Model. We now show that the ARIMA model is an intuitively reasonable model for many time series that occur in practice. First, we note that the local behavior of a stationary time series is heavily dependent on the *level* of $\tilde{z}_t$. This is to be contrasted with the behavior of series such as those in Figure 4.1, where the local behavior of the series appears to be independent of its level.

If we are to use models for which the behavior of the process is independent of its level, we must choose the autoregressive operator $\varphi(B)$ such that

$$\varphi(B)(\tilde{z}_t + c) = \varphi(B)\tilde{z}_t$$

where c is any constant. Thus $\varphi(B)$ must be of the form

$$\varphi(B) = \phi_1(B)(1 - B) = \phi_1(B)\nabla$$

Therefore, a class of processes having the desired property will be of the form

$$\phi_1(B)w_t = \theta(B)a_t$$

where $w_t = \nabla\tilde{z}_t = \nabla z_t$. Required homogeneity excludes the possibility that w_t should increase explosively. This means that either $\phi_1(B)$ is a stationary autoregressive operator or $\phi_1(B) = \phi_2(B)(1 - B)$, so that $\phi_2(B)w_t = \theta(B)a_t$, where now $w_t = \nabla^2 z_t$. In the latter case, the same argument can be applied to the second difference, and so on.

Eventually, we arrive at the conclusion that for the representation of time series that are nonstationary but nevertheless exhibit homogeneity, the operator on the left of (4.1.3) should be of the form $\phi(B)\nabla^d$, where $\phi(B)$ is a *stationary* autoregressive operator. Thus, we are led back to the model (4.1.5).

To approach the model from a somewhat different viewpoint, consider the situation where $d = 0$ in (4.1.4), so that $\phi(B)\tilde{z}_t = \theta(B)a_t$. The requirement that the zeros of $\phi(B)$ lie outside the unit circle would ensure not only that the process $\tilde{z}_t$ was stationary with mean zero, but also that $\nabla z_t, \nabla^2 z_t, \nabla^3 z_t, \ldots$ were each stationary with mean zero. Figure 4.3(a) shows one kind of nonstationary series we would like to represent. This series is homogeneous except in level, in that except for a vertical translation, one part of it looks much the same as another. We can represent such behavior by retaining the requirement that each of the differences be stationary with zero mean, but letting the level ''go free.'' We do this by using the model

$$\phi(B)\nabla z_t = \theta(B)a_t$$

Figure 4.3(b) shows a second kind of nonstationarity or fairly common occurrence. The series has neither a fixed level nor a fixed slope, but its behavior is homogeneous if we allow for differences in these characteristics. We can represent such behavior by the model

$$\phi(B)\nabla^2 z_t = \theta(B)a_t$$

which ensures stationarity and zero mean for all differences after the first and second but allows the level and the slope to ''go free.''

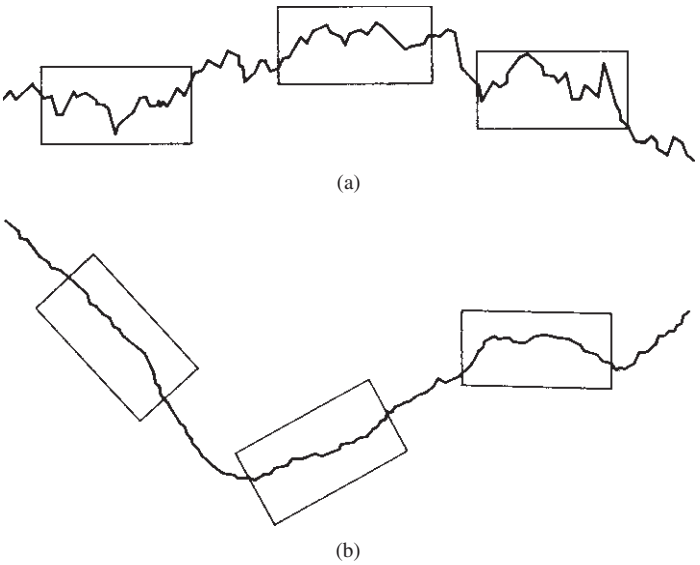


FIGURE 4.3 Two kinds of homogeneous nonstationary behavior. (a) A series showing nonstationarity in level such as can be represented by the model $\phi(B)\nabla z_t = \theta(B)a_t$. (b) A series showing nonstationarity in level and in slope such as can be represented by the model $\phi(B)\nabla^2 z_t = \theta(B)a_t$.

4.1.3 General Form of the ARIMA Model

For reasons to be given below, it is sometimes useful to consider a slight extension of the ARIMA model in (4.1.5), by adding a constant term θ_0, yielding the more general form

$$\varphi(B)z_t = \phi(B)\nabla^d z_t = \theta_0 + \theta(B)a_t \tag{4.1.9}$$

where

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$
$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

In what follows:

1. $\phi(B)$ will be called the *autoregressive operator*; it is assumed to be stationary, that is, the roots of $\phi(B) = 0$ lie outside the unit circle.
2. $\varphi(B) = \phi(B)\nabla^d$ will be called the *generalized autoregressive* operator; it is a nonstationary operator with d of the roots of $\varphi(B) = 0$ equal to unity, that is, d unit roots.
3. $\theta(B)$ will be called the *moving average operator*; it is assumed to be invertible, that is, the roots of $\theta(B) = 0$ lie outside the unit circle.

When $d = 0$, this model represents a stationary process. The requirements of stationarity and invertibility apply independently, and, in general, the operators $\phi(B)$ and $\theta(B)$ will not be of the same order. Examples of the stationarity regions for the simple cases of $p = 1, 2$ and the identical invertibility regions for $q = 1, 2$ were given in Chapter 3.

Stochastic and Deterministic Trends. When the constant term θ_0 is omitted, the model (4.1.9) is capable of representing series that have *stochastic* trends, as typified, for example, by random changes in the level and slope of the series. In general, however, we may wish to include a *deterministic* function of time $f(t)$ in the model. In particular, automatic allowance for a deterministic polynomial trend, of degree d, can be made by permitting θ_0 to be nonzero. For example, when $d = 1$, we may use the model with $\theta_0 \neq 0$ to represent a possible deterministic linear trend in the presence of nonstationary noise. Since, from (3.1.22), to allow θ_0 to be nonzero is equivalent to permitting

$$E[w_t] = E[\nabla^d z_t] = \mu_w = \frac{\theta_0}{1 - \phi_1 - \phi_2 - \cdots - \phi_p}$$

to be nonzero, an alternative way of expressing this more general model (4.1.9) is in the form of a stationary invertible ARMA process in $\tilde{w}_t = w_t - \mu_w$. That is,

$$\phi(B)\tilde{w}_t = \theta(B)a_t \tag{4.1.10}$$

Notice, when $d = 1$, for example, $\nabla z_t = w_t = \tilde{w}_t + \mu_w$ implies that $z_t = \tilde{z}_t + \mu_w t + \alpha$, where α is an intercept constant and the process $\tilde{z}_t$ is such that $\nabla \tilde{z}_t = \tilde{w}_t$, which has zero mean. Thus, $\theta_0 \neq 0$ allows for a deterministic linear trend component in z_t with slope $\mu_w = \theta_0/(1 - \phi_1 - \cdots - \phi_p)$.

In many applications, where no physical reason for a deterministic component exists, the mean of w can be assumed to be zero unless such an assumption is inconsistent with the data. In many cases, the assumption of a stochastic trend is more realistic than the assumption of a deterministic trend. This is of special importance in forecasting, since a stochastic trend does not require the series to follow the trend pattern seen in the past. In what follows, when $d > 0$, we will often assume that $\mu_w = 0$, or equivalently, that $\theta_0 = 0$, unless it is clear from the data or from the nature of the problem that a nonzero mean, or more generally a deterministic component of known form, is needed.

Some Important Special Cases of the ARIMA Model. In Chapter 3, we examined some important special cases of the model (4.1.9), corresponding to the stationary situation, $d = 0$. The following models represent some special cases of the nonstationary model ($d \geq 1$), which seem to be common in practice.

1. *The (0, 1, 1) process:*

$$\begin{aligned} \nabla z_t &= a_t - \theta_1 a_{t-1} \\ &= (1 - \theta_1 B)a_t \end{aligned}$$

corresponding to $p = 0$, $d = 1$, $q = 1$, $\phi(B) = 1$, $\theta(B) = 1 - \theta_1 B$.

2. *The (0, 2, 2) process:*

$$\begin{aligned} \nabla^2 z_t &= a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \\ &= (1 - \theta_1 B - \theta_2 B^2)a_t \end{aligned}$$

corresponding to $p = 0$, $d = 2$, $q = 2$, $\phi(B) = 1$, $\theta(B) = 1 - \theta_1 B - \theta_2 B^2$.

3. *The (1, 1, 1) process:*

$$\nabla z_t - \phi_1 \nabla z_{t-1} = a_t - \theta_1 a_{t-1}$$

TABLE 4.1 Summary of Simple Nonstationary Models Fitted to Time Series of Figure 4.1

Series	Model	Order of Model
A	$\nabla z_t = (1 - 0.7B)a_t$	(0, 1, 1)
B	$\nabla z_t = (1 + 0.1B)a_t$	(0, 1, 1)
C	$(1 - 0.8B)\nabla z_t = a_t$	(1, 1, 0)
D	$\nabla z_t = (1 - 0.1B)a_t$	(0, 1, 1)

or

$$(1 - \phi_1 B)\nabla z_t = (1 - \theta_1 B)a_t$$

corresponding to $p = 1$, $d = 1$, $q = 1$, $\phi(B) = 1 - \phi_1 B$, $\theta(B) = 1 - \theta_1 B$.

For the representation of nonseasonal time series (seasonal models are considered in Chapter 9), we rarely seem to meet situations for which either p, d, or q need to be greater than 2. Frequently, values of zero or unity will be appropriate for one or more of these orders. For example, we show later that Series A, B, C, and D given in Figure 4.1 are reasonably well represented[2] by the simple models shown in Table 4.1.

Nonlinear Transformation of z. The range of useful applications of the model (4.1.9) widens considerably if we allow the possibility of transformation. Thus, we may substitute $z_t^{(\lambda)}$ for z_t, in (4.1.9), where $z_t^{(\lambda)}$ is some nonlinear transformation of z_t, involving one or more parameters λ. A suitable transformation may be suggested by the application, or in some cases it can be estimated from the data. For example, if we were interested in the sales of a recently introduced commodity, we might find that the sales volume was increasing at a rapid rate and that it was the *percentage* fluctuation that showed nonstationary stability (homogeneity) rather than the absolute fluctuation. This would support the analysis of the logarithm of sales since

$$\nabla \log(z_t) = \log\left(\frac{z_t}{z_{t-1}}\right) = \log\left(1 + \frac{\nabla z_t}{z_{t-1}}\right) \simeq \frac{\nabla z_t}{z_{t-1}}$$

where $\nabla z_t / z_{t-1}$ are the relative or percentage changes, the approximation holding if the relative changes are not excessively large. When the data cover a wide range and especially for seasonal data, estimation of the transformation using the approach of Box and Cox (1964) may be helpful (for an example, see Section 9.3.5). This approach considers the family of power transformations of the form $z_t^{(\lambda)} = (z_t^{\lambda} - 1)/\lambda$ for $\lambda \neq 0$ and $z_t^{(0)} = \log(z_t)$ for $\lambda = 0$.

Software to estimate the parameter λ in the Box–Cox power transformation is available in the TSA and MASS libraries of R. For example, the function BoxCox.ar() in the TSA package finds a power transformation so that the transformed series is approximately a Gaussian AR process.

[2]As is discussed more fully later, there are certain advantages in using a nonstationary rather than a stationary model in cases of doubt. In particular, none of the fitted models above assume that z_t has a fixed mean. However, we show in Chapter 7 that it is possible in certain cases to obtain stationary models of slightly better fit.

4.2 THREE EXPLICIT FORMS FOR THE ARIMA MODEL

We now consider three different ''explicit'' forms for the general model (4.1.9). Each of these allows some special aspect to be appreciated. Thus, the current value z_t of the process can be expressed

1. In terms of previous values of the z's and current and previous values of the a's, by direct use of the *difference equation*,
2. In terms of *current and previous shocks* a_{t-j} only, and
3. In terms of a weighted sum of *previous values* z_{t-j} of the process and the current shock a_t.

In this chapter, we are concerned primarily with *nonstationary* models in which $\nabla^d z_t$ is a stationary process and d is greater than zero. For such models, we can, without loss of generality, omit μ from the specification or equivalently replace $\tilde{z}_t$ by z_t. The results of this chapter and the next will, however, apply to stationary models for which $d = 0$, provided that z_t is then interpreted as the *deviation* from the mean μ.

4.2.1 Difference Equation Form of the Model

Direct use of the difference equation permits us to express the current value z_t of the process in terms of previous values of the z's and of the current and previous values of the a's. Thus, if

$$\varphi(B) = \phi(B)(1 - B)^d = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_{p+d} B^{p+d}$$

the general model (4.1.9), with $\theta_0 = 0$, may be written as

$$z_t = \varphi_1 z_{t-1} + \cdots + \varphi_{p+d} z_{t-p-d} - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} + a_t \tag{4.2.1}$$

For example, consider the process represented by the model of order (1, 1, 1)

$$(1 - \phi B)(1 - B)z_t = (1 - \theta B)a_t$$

where, for convenience, we drop the subscript 1 on ϕ_1 and θ_1. Then this process may be written as

$$[1 - (1 + \phi)B + \phi B^2]z_t = (1 - \theta B)a_t$$

that is,

$$z_t = (1 + \phi)z_{t-1} - \phi z_{t-2} + a_t - \theta a_{t-1} \tag{4.2.2}$$

with $\varphi_1 = 1 + \phi$ and $\varphi_2 = -\phi$ in the notation introduced above. For many purposes, and, in particular, for calculating forecasts, the difference equation (4.2.1) is the most convenient form to use.

4.2.2 Random Shock Form of the Model

Model in Terms of Current and Previous Shocks. As discussed in Chapter 3, a linear model can be written as the output z_t from the linear filter

$$\begin{aligned} z_t &= a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots \\ &= a_t + \sum_{j=1}^{\infty} \psi_j a_{t-j} \\ &= \psi(B) a_t \end{aligned} \tag{4.2.3}$$

whose input is white noise, or a sequence of uncorrelated shocks a_t with mean 0 and common variance σ_a^2. It is sometimes useful to express the ARIMA model in this form, and, in particular, the ψ weights will be needed in Chapter 5 to calculate the variance of the forecast errors. However, since the nonstationary ARIMA processes are not in statistical equilibrium over time, they cannot be assumed to extend infinitely into the past, and hence an infinite representation as in (4.2.3) will not be possible. But a related finite truncated form, which will be discussed subsequently, always exists. We now show that the ψ weights for an ARIMA process may be obtained directly from the difference equation form of the model.

General Expression for the ψ Weights. If we operate on both sides of (4.2.3) with the generalized autoregressive operator $\varphi(B)$, we obtain

$$\varphi(B) z_t = \varphi(B)\psi(B) a_t$$

However, since $\varphi(B) z_t = \theta(B) a_t$, it follows that

$$\varphi(B)\psi(B) = \theta(B) \tag{4.2.4}$$

Therefore, the ψ weights may be obtained by equating coefficients of B in the expansion

$$\begin{aligned} (1 - \varphi_1 B - \cdots - \varphi_{p+d} B^{p+d})(1 + \psi_1 B + \psi_2 B^2 + \cdots) \\ = (1 - \theta_1 B - \cdots - \theta_q B^q) \end{aligned} \tag{4.2.5}$$

Thus, we find that the ψ_j weights of the ARIMA process can be determined recursively through the equations

$$\psi_j = \varphi_1 \psi_{j-1} + \varphi_2 \psi_{j-2} + \cdots + \varphi_{p+d} \psi_{j-p-d} - \theta_j \quad j > 0$$

with $\psi_0 = 1$, $\psi_j = 0$ for $j < 0$, and $\theta_j = 0$ for $j > q$. We note that for j greater than the larger of $p + d - 1$ and q, the ψ weights satisfy the homogeneous difference equation defined by the generalized autoregressive operator, that is,

$$\varphi(B)\psi_j = \phi(B)(1 - B)^d \psi_j = 0 \tag{4.2.6}$$

where B now operates on the subscript j. Thus, for sufficiently large j, the weights ψ_j are represented by a mixture of polynomials, damped exponentials, and damped sinusoids in the argument j.

Example. For illustration, consider the (1, 1, 1) process (4.2.2), for which

$$\begin{aligned}\varphi(B) &= (1 - \phi B)(1 - B) \\ &= 1 - (1 + \phi)B + \phi B^2\end{aligned}$$

and

$$\theta(B) = 1 - \theta B$$

Substituting in (4.2.5) gives

$$[1 - (1 + \phi)B + \phi B^2](1 + \psi_1 B + \psi_2 B^2 + \cdots) = 1 - \theta B$$

and hence the ψ_j satisfy the recursion $\psi_j = (1 + \phi)\psi_{j-1} - \phi\psi_{j-2}$, $j \geq 2$ with $\psi_0 = 1$ and $\psi_1 = (1 + \phi) - \theta$. Thus, since the roots of $\varphi(B) = (1 - \phi B)(1 - B) = 0$ are $G_1^{-1} = 1$ and $G_2^{-1} = \phi^{-1}$, we have, in general,

$$\psi_j = A_0 + A_1\phi^j \tag{4.2.7}$$

where the constants A_0 and A_1 are determined from the initial values $\psi_0 = A_0 + A_1 = 1$ and $\psi_1 = A_0 + A_1\phi = 1 + \phi - \theta$ as

$$A_0 = \frac{1 - \theta}{1 - \phi} \quad A_1 = \frac{\theta - \phi}{1 - \phi}$$

Thus, informally, we may wish to express model (4.2.2) in the equivalent form

$$z_t = \sum_{j=0}^{\infty}(A_0 + A_1\phi^j)a_{t-j} \tag{4.2.8}$$

Since $|\phi| < 1$, the weights ψ_j tend to A_0 for large j, so that shocks a_{t-j}, which entered in the remote past, receive a constant weight A_0. However, the representation in (4.2.8) is strictly not valid because the infinite sum on the right does not converge in any sense; that is, the weights ψ_j are not absolutely summable as in the case of a stationary process. A related truncated version of the random shock form of the model is always valid, as we discuss in detail shortly. Nevertheless, for notational convenience, we will often refer to the infinite random shock form (4.2.3) of an ARIMA process, even though this form is strictly not convergent, as a simple notational device to represent the valid truncated form in (4.2.14), in situations where the distinction between the two forms is not important.

Truncated Form of the Random Shock Model. For technical purposes, it is necessary and in some cases convenient to consider the model in a form slightly different from (4.2.3). Suppose that we wish to express the current value z_t of the process in terms of the $t - k$ shocks $a_t, a_{t-1}, \ldots, a_{k+1}$, which have entered the system since some time origin $k < t$. This time origin k might, for example, be the time at which the process was first observed.

The general model

$$\varphi(B)z_t = \theta(B)a_t \tag{4.2.9}$$

is a difference equation with the solution

$$z_t = C_k(t - k) + I_k(t - k) \tag{4.2.10}$$

A short discussion of linear difference equations is given in Appendix A4.1. We remind the reader that the solution of such equations closely parallels the solution of linear differential equations. The *complimentary function* $C_k(t-k)$ is the general solution of the homogeneous difference equation

$$\varphi(B)C_k(t-k) = 0 \tag{4.2.11}$$

In general, this solution will consist of a *linear* combination of certain functions of time. These functions are powers t^j, real geometric (exponential) terms G^t, and complex geometric (exponential) terms $D^t \sin(2\pi f_0 t + F)$, where the constants G, f_0, and F are functions of the parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$ of the model. The coefficients that form the linear combinations of these terms can be determined so as to satisfy a set of initial conditions defined by the values of the process before time $k+1$. The *particular integral* $I_k(t-k)$ is *any* function that satisfies

$$\varphi(B)I_k(t-k) = \theta(B)a_t \tag{4.2.12}$$

It should be carefully noted that in this expression B operates on t and *not on* k. It is shown in Appendix A4.1 that this equation is satisfied for $t-k > q$ by

$$I_k(t-k) = \sum_{j=0}^{t-k-1} \psi_j a_{t-j} = a_t + \psi_1 a_{t-1} + \cdots + \psi_{t-k-1} a_{k+1} \quad t > k \tag{4.2.13}$$

with $I_k(t-k) = 0, t \le k$. This particular integral $I_k(t-k)$, thus, represents the finite truncated form of the infinite random shock form (4.2.3), while the complementary function $C_k(t-k)$ embodies the ''initializing'' features of the process z in the sense that $C_k(t-k)$ is already determined or specified by the time $k+1$. Hence, the truncated form of the random shock model for the ARIMA process (4.1.3) is given by

$$z_t = \sum_{j=0}^{t-k-1} \psi_j a_{t-j} + C_k(t-k) \tag{4.2.14}$$

For illustration, consider Figure 4.4. The above discussion implies that any observation z_t can be considered in relation to any previous time k and can be divided up into two additive parts. The first part $C_k(t-k)$ is the component of z_t, *already determined at time* k, and indicates what the observations prior to time $k+1$ had to tell us about the value of the series at time t. It represents the course that the process would take if at time k, the source of shocks a_t had been ''switched off.'' The second part, $I_k(t-k)$, represents an additional component, *unpredictable at time* k, which embodies the entire effect of shocks entering the system at time k. Hence, to specify an ARIMA process, one must specify the initializing component $C_k(t-k)$ in (4.2.14) for some time origin k in the finite (but possibly remote) past, with the remaining course of the process being determined through the truncated random shock terms in (4.2.14).

Example. For illustration, consider again the example

$$(1-\phi B)(1-B)z_t = (1-\theta B)a_t$$

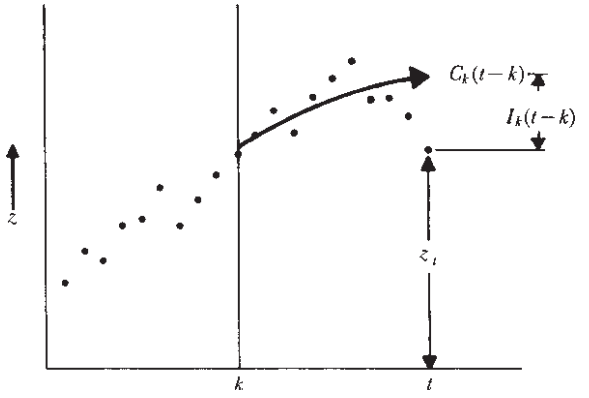


FIGURE 4.4 Role of the complementary function $C_k(t-k)$ and of the particular integral $I_k(t-k)$ in describing the behavior of a time series.

The complementary function is the solution of the difference equation

$$(1-\phi B)(1-B)C_k(t-k)=0$$

that is,

$$C_k(t-k)=b_0^{(k)}+b_1^{(k)}\phi^{t-k}$$

where $b_0^{(k)}$, $b_1^{(k)}$ are coefficients that depend on the past history of the process and, it will be noted, change with the origin k.

Making use of the ψ weights (4.2.7), a particular integral (4.2.13) is

$$I_k(t-k)=\sum_{j=0}^{t-k-1}(A_0+A_1\phi^j)a_{t-j}$$

so that, finally, we can write the model (4.2.8) in the equivalent form

$$z_t=b_0^{(k)}+b_1^{(k)}\phi^{t-k}+\sum_{j=0}^{t-k-1}(A_0+A_1\phi^j)a_{t-j} \tag{4.2.15}$$

Note that since $|\phi|<1$, if $t-k$ is chosen sufficiently large, the term involving ϕ^{t-k} in this expression is negligible and may be ignored.

Link Between the Truncated and Nontruncated Forms of the Random Shock Model. Returning to the general case, we can always think of the process with reference to some (possibly remote) finite origin k, with the process having the truncated random shock form as in (4.2.14). By comparison with the nontruncated form in (4.2.3), one can see that we might, informally, make the correspondence of representing the complementary function $C_k(t-k)$ in terms of the ψ weights as

$$C_k(t-k)=\sum_{j=t-k}^{\infty}\psi_j a_{t-j} \tag{4.2.16}$$

even though, formally, the infinite sum on the right of (4.2.16) does not converge. As mentioned earlier, for notational simplicity, we will often use this correspondence.

In summary, then, for the general model (4.2.9),

1. We can express the value z_t of the process, informally, as an infinite weighted sum of current and previous shocks a_{t-j} , according to

$$z_t = \sum_{j=0}^{\infty} \psi_j a_{t-j} = \psi(B)a_t$$

2. The value of z_t can be expressed, more formally, as a weighted finite sum of the $t-k$ current and previous shocks occurring after some origin k, plus a complementary function $C_k(t-k)$. This finite sum consists of the first $t-k$ terms of the infinite sum, so that

$$z_t = C_k(t-k) + \sum_{j=0}^{t-k-1} \psi_j a_{t-j} \tag{4.2.17}$$

Finally, the complementary function $C_k(t-k)$ can be taken, for notational convenience, to be represented as the truncated infinite sum, so that

$$C_k(t-k) = \sum_{j=t-k}^{\infty} \psi_j a_{t-j} \tag{4.2.18}$$

For illustration, consider once more the model

$$(1-\phi B)(1-B)z_t = (1-\theta B)a_t$$

We can write z_t either, informally, as an infinite sum of the a_{t-j}'s

$$z_t = \sum_{j=0}^{\infty} (A_0 + A_1\phi^j)a_{t-j}$$

or, more formally, in terms of the weighted finite sum as

$$z_t = C_k(t-k) + \sum_{j=0}^{t-k-1} (A_0 + A_1\phi^j)a_{t-j}$$

Furthermore, the complementary function can be written as

$$C_k(t-k) = b_0^{(k)} + b_1^{(k)}\phi^{t-k}$$

where $b_0^{(k)}$ and $b_1^{(k)}$, which satisfy the initial conditions through time k, are

$$b_0^{(k)} = \frac{z_k - \phi z_{k-1} - \theta a_k}{1-\phi} \quad b_1^{(k)} = \frac{-\phi(z_k - z_{k-1}) + \theta a_k}{1-\phi}$$

The complementary function can also be represented, informally, as the truncated infinite sum

$$C_k(t-k) = \sum_{j=t-k}^{\infty} (A_0 + A_1\phi^j)a_{t-j}$$

from which it can be seen that $b_0^{(k)}$ and $b_1^{(k)}$ may be represented as

$$b_0^{(k)} = A_0 \sum_{j=t-k}^{\infty} a_{t-j} = \frac{1-\theta}{1-\phi} \sum_{j=t-k}^{\infty} a_{t-j}$$

$$b_1^{(k)} = A_1 \sum_{j=t-k}^{\infty} \phi^{j-(t-k)} a_{t-j} = \frac{\theta-\phi}{1-\phi} \sum_{j=t-k}^{\infty} \phi^{j-(t-k)} a_{t-j}$$

Complementary Function as a Conditional Expectation. One consequence of the truncated form (4.2.14) is that for $m > 0$,

$$\begin{aligned} C_k(t-k) = C_{k-m}(t-k+m) + \psi_{t-k}a_k + \psi_{t-k+1}a_{k-1} + \cdots \\ + \psi_{t-k+m-1}a_{k-m+1} \end{aligned} \tag{4.2.19}$$

which shows how the complementary function changes as the origin k is changed. Now denote by $E_k[z_t]$ the *conditional expectation of* z_t, *at time* k. That is the expectation given complete historical knowledge of the process up to, but not beyond time k. To calculate this expectation, note that

$$E_k[a_j] = \begin{cases} 0 & j > k \\ a_j & j \le k \end{cases}$$

That is, *standing at time* k, the expected values of the future a's are zero and of those that have happened already are their actually realized values.

By taking conditional expectations at time k on both sides of (4.2.17), we obtain $E_k[z_t] = C_k(t-k)$. Thus, for $(t-k) > q$, the complementary function provides the expected value of the future value z_t of the process, *viewed from time* k and based on knowledge of the past. The particular integral shows how that expectation is modified by *subsequent* events represented by the shocks $a_{k+1}, a_{k+2}, \ldots, a_t$. In the problem of forecasting, which we discuss in Chapter 5, it will turn out that $C_k(t-k)$ is the minimum mean square error forecast of z_t made at time k. Equation (4.2.19) may be used in ''updating'' this forecast.

4.2.3 Inverted Form of the Model

Model in Terms of Previous z's and the Current Shock a_t. We have seen in Section 3.1.1 that the model

$$z_t = \psi(B)a_t$$

may also be written in the inverted form

$$\psi^{-1}(B)z_t = a_t$$

or

$$\pi(B)z_t = \left(1 - \sum_{j=1}^{\infty} \pi_j B^j\right) z_t = a_t \tag{4.2.20}$$

Thus, z_t is an infinite weighted sum of previous values of z, plus a random shock:

$$z_t = \pi_1 z_{t-1} + \pi_2 z_{t-2} + \cdots + a_t$$

Because of the invertibility condition, the π weights must form a convergent series; that is, $\pi(B)$ must converge on or within the unit circle.

General Expression for the π Weights. To derive the π weights for the general ARIMA model, we can substitute (4.2.20) in

$$\varphi(B)z_t = \theta(B)a_t$$

to obtain

$$\varphi(B)z_t = \theta(B)\pi(B)z_t$$

Hence, the π weights can be obtained explicitly by equating coefficients of B in

$$\varphi(B) = \theta(B)\pi(B) \tag{4.2.21}$$

that is,

$$\begin{aligned}(1 - \varphi_1 B - \cdots - \varphi_{p+d} B^{p+d}) = (1 - \theta_1 B - \cdots - \theta_q B^q) \\ \times (1 - \pi_1 B - \pi_2 B^2 - \cdots)\end{aligned} \tag{4.2.22}$$

Thus, we find that the π_j weights of the ARIMA process can be determined recursively through

$$\pi_j = \theta_1 \pi_{j-1} + \theta_2 \pi_{j-2} + \cdots + \theta_q \pi_{j-q} + \varphi_j \quad j > 0$$

with the convention $\pi_0 = -1$, $\pi_j = 0$ for $j < 0$, and $\varphi_j = 0$ for $j > p + d$. It will be noted that for j greater than the larger of $p + d$ and q, the π weights satisfy the homogeneous difference equation defined by the *moving average operator*

$$\theta(B)\pi_j = 0$$

where B now operates on j. Hence, for sufficiently large j, the π weights will exhibit similar behavior as the autocorrelation function (3.2.5) of an autoregressive process; that is, they follow a mixture of damped exponentials and damped sine waves.

Another interesting fact is that if $d \geq 1$, the π weights in (4.2.20) sum to unity. This may be verified by substituting $B = 1$ in (4.2.21). Thus, $\varphi(B) = \phi(B)(1 - B)^d$ is zero when $B = 1$ and $\theta(1) \neq 0$, because the roots of $\theta(B) = 0$ lie outside the unit circle. Hence, it follows from (4.2.21) that $\pi(1) = 0$, that is,

$$\sum_{j=1}^{\infty} \pi_j = 1 \tag{4.2.23}$$

Therefore, if $d \geq 1$, the process may be written in the form

$$z_t = \bar{z}_{t-1}(\pi) + a_t \tag{4.2.24}$$

where

$$\bar{z}_{t-1}(\pi) = \sum_{j=1}^{\infty} \pi_j z_{t-j}$$

is a *weighted average* of previous values of the process.

Example. We again consider, for illustration, the ARIMA(1, 1, 1) process:

$$(1 - \phi B)(1 - B)z_t = (1 - \theta B)a_t$$

Then, using (4.2.21),

$$\pi(B) = \varphi(B)\theta^{-1}(B) = [1 - (1 + \phi)B + \phi B^2](1 + \theta B + \theta^2 B^2 + \cdots)$$

so that

$$\pi_1 = \phi + (1 - \theta) \quad \pi_2 = (\theta - \phi)(1 - \theta) \quad \pi_j = (\theta - \phi)(1 - \theta)\theta^{j-2}, \quad j \geq 3.$$

The first seven π weights corresponding to $\phi = -0.3$ and $\theta = 0.5$ are given in Table 4.2. Thus, z_t would be generated by a weighted average of previous values, plus an additional shock, according to

$$z_t = (0.2z_{t-1} + 0.4z_{t-2} + 0.2z_{t-3} + 0.1z_{t-4} + \cdots) + a_t$$

We notice, in particular, that the π weights die out as more and more remote values of z_{t-j} are involved. This happens when $-1 < \theta < 1$, so that the series is invertible.

We mention in passing that, for models fitted to actual time series, the convergent π weights usually die out rather quickly. Thus, although z_t may be theoretically dependent on the remote past, the representation

$$z_t = \sum_{j=1}^{\infty} \pi_j z_{t-j} + a_t$$

will usually show that z_t is dependent *to an important extent* only on recent past values z_{t-j} of the time series. This is still true even though for nonstationary models with $d > 0$, the ψ weights in the ''weighted shock'' representation

$$z_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}$$

do not die out to zero. What happens, of course, is that all the information that remote values of the shocks a_{t-j} supply about z_t is contained in recent values $z_{t-1}, z_{t-2}, \cdots$ of the series. In particular, the expectation $E_k[z_t]$, which in theory is conditional on complete history of the process up to time k, can usually be computed to sufficient accuracy from *recent* values of the time series. This fact is particularly important in forecasting applications.

TABLE 4.2 First Seven π Weights for an ARIMA(1, 1, 1) Process with $\phi = -0.3$, $\theta = 0.5$

j	1	2	3	4	5	6	7
π_j	0.2	0.4	0.2	0.1	0.05	0.025	0.0125

4.3 INTEGRATED MOVING AVERAGE PROCESSES

A nonstationary model that is useful in representing some commonly occurring series is the (0, 1, 1) process:

$$\nabla z_t = a_t - \theta a_{t-1}$$

The model contains only two parameters, θ and σ_a^2. Figure 4.5 shows two time series generated by this model from the same sequence of random normal deviates a_t. For the first series, $\theta = 0.6$, and for the second, $\theta = 0$. Models of this kind have often been found useful in inventory control problems, in representing certain kinds of disturbances occurring in industrial processes, and in econometrics. We will show in Chapter 7 that this simple process can, with suitable parameter values, supply useful representations of Series A, B, and D shown in Figure 4.1. Another valuable model is the (0, 2, 2) process

$$\nabla^2 z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$$

which contains three parameters, θ_1, θ_2, and σ_a^2. Figure 4.6 shows two series generated from this model using the same set of normal deviates. For the first series, the parameters $(\theta_1, \theta_2) = (0, 0)$ and for the second $(\theta_1, \theta_2) = (1.5, -0.8)$. The series tend to be much smoother than those generated by the (0, 1, 1) process. The (0, 2, 2) models are useful in representing disturbances (such as Series C) in systems with a large degree of inertia. Both the (0, 1, 1) and the (0, 2, 2) models are special cases of the class

$$\nabla^d z_t = \theta(B) a_t \tag{4.3.1}$$

We call these models *integrated moving average* (IMA) processes, of order $(0, d, q)$, and consider their properties in the following section.

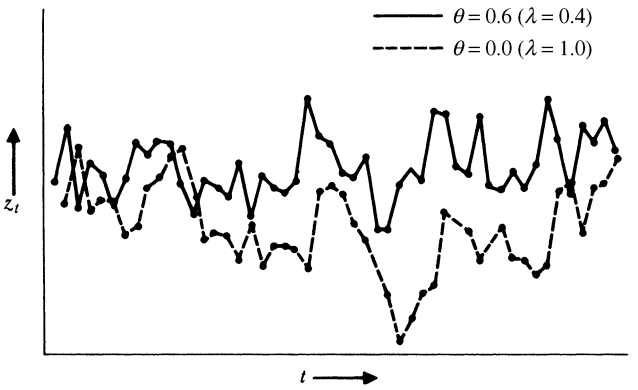


FIGURE 4.5 Two time series generated from IMA(0, 1, 1) models.

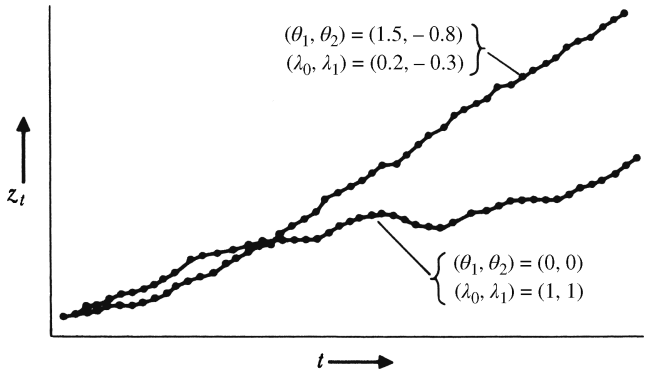


FIGURE 4.6 Two time series generated from IMA(0, 2, 2) models.

4.3.1 Integrated Moving Average Process of Order (0, 1, 1)

Difference Equation Form. The IMA(0, 1, 1) process

$$\nabla z_t = (1 - \theta B)a_t \quad -1 < \theta < 1$$

possesses useful representational capability, and we now study its properties in more detail. The model can be written in terms of the z's and the a's in the form

$$z_t = z_{t-1} + a_t - \theta a_{t-1} \tag{4.3.2}$$

Random Shock Form of Model. Alternatively, we can obtain z_t in terms of the a's alone by summing on both sides of (4.3.2). Before doing this, there is some advantage in expressing the right-hand operator in terms of ∇ rather than B. Thus, we can write

$$1 - \theta B = (1 - \theta)B + (1 - B) = (1 - \theta)B + \nabla = \lambda B + \nabla$$

where $\lambda = 1 - \theta$, and the invertibility region in terms of λ is defined by $0 < \lambda < 2$. Hence

$$\nabla z_t = \lambda a_{t-1} + \nabla a_t$$

Relative to some time origin $k < t$, applying the finite summation operator $S_{t-k} = 1 + B + \cdots + B^{t-k-1} = (1 - B^{t-k})/(1 - B)$, we obtain

$$(1 - B^{t-k})z_t = \lambda S_{t-k} a_{t-1} + (1 - B^{t-k})a_t \tag{4.3.3}$$

so that

$$z_t = a_t + \lambda(a_{t-1} + a_{t-2} + \cdots + a_{k+1}) + (z_k - \theta a_k) \tag{4.3.4}$$

In comparison to $z_t = \sum_{j=0}^{t-k-1} \psi_j a_{t-j} + C_k(t-k)$, the weights are $\psi_0 = 1, \psi_j = \lambda$ for $j \geq 1$. Also, the complementary function is $C_k(t-k) = z_k - \theta a_k = b_0^{(k)}$ (a ''constant'' b_0 for each k), which is the solution of the difference equation (1 - B)$C_k(t-k) = 0$. Moreover, in the infinite form $z_t = a_t + \lambda \sum_{j=1}^{\infty} a_{t-j}$, we may identify $b_0^{(k)}$ with $\lambda \sum_{j=t-k}^{\infty} a_{t-j}$. For this model, then, the complementary function is simply a constant (i.e., a polynomial in t of degree zero) representing the current ''level'' of the process and associated with the particular origin of

reference k. If the origin is changed from $k-1$ to k, then b_0 is ''updated'' according to

$$b_0^{(k)} = b_0^{(k-1)} + \lambda a_k$$

since using (4.3.2), $b_0^{(k)} = z_k + (\lambda - 1)a_k = z_{k-1} - \theta a_{k-1} + \lambda a_k$.

Inverted Form of Model. Finally, we can consider the model in the form

$$\pi(B)z_t = a_t$$

or equivalently, in the form

$$z_t = \sum_{j=1}^{\infty} \pi_j z_{t-j} + a_t = \bar{z}_{t-1}(\pi) + a_t$$

where $\bar{z}_{t-1}(\pi)$ is a weighted moving average of previous values of the process.

Using (4.2.21), the π weights for the IMA(0, 1, 1) process are given by

$$(1 - \theta B)\pi(B) = 1 - B$$

that is,

$$\begin{aligned}\pi(B) = \frac{1-B}{1-\theta B} &= \frac{1 - \theta B - (1-\theta)B}{1-\theta B} \\ &= 1 - (1-\theta)(B + \theta B^2 + \theta^2 B^3 + \cdots)\end{aligned}$$

so that

$$\pi_j = (1-\theta)\theta^{j-1} = \lambda(1-\lambda)^{j-1} \quad j \geq 1$$

Thus, the process may be written as

$$z_t = \bar{z}_{t-1}(\lambda) + a_t \tag{4.3.5}$$

The weighted moving average of previous values of the process

$$\bar{z}_{t-1}(\lambda) = \lambda \sum_{j=1}^{\infty} (1-\lambda)^{j-1} z_{t-j} \tag{4.3.6}$$

is, in this case, an *exponentially weighted moving average* (EWMA). This term reflects the fact that the weights

$$\lambda \quad \lambda(1-\lambda) \quad \lambda(1-\lambda)^2 \quad \lambda(1-\lambda)^3 \cdots$$

fall off exponentially (i.e., as a geometric progression) as j increases. The weight function for an IMA(0, 1, 1) process, with $\lambda = 0.4$ (or $\theta = 0.6$), is shown in Figure 4.7.

Although the invertibility condition is satisfied for $0 < \lambda < 2$, in practice, we are most often concerned with values of λ between zero and 1 (i.e., $0 < \theta < 1$). We note that if λ had a value equal to 1, the weight function would consist of a single spike ($\pi_1 = 1, \pi_j = 0$ for $j > 1$). As the value λ approaches zero, the exponential weights die out more and more slowly and the EWMA stretches back further into past values of the process. Finally, with

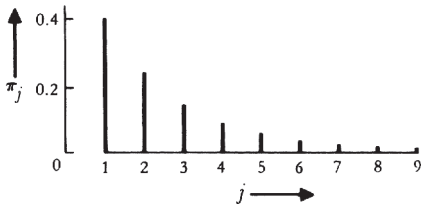


FIGURE 4.7 The π weights for an IMA process of order (0, 1, 1) with $\lambda = 1 - \theta = 0.4$.

$\lambda = 0$ and $\theta = 1$, the model $(1 - B)z_t = (1 - B)a_t$ is equivalent to $z_t = \theta_0 + a_t$, with θ_0 being given by the mean of all past values.

Since $b_0^{(k)} = z_k - \theta a_k = z_{k+1} - a_{k+1}$, or $z_{k+1} = b_0^{(k)} + a_{k+1}$, on comparison with (4.3.5) it follows that for this process, the complementary function $b_0^{(k)} = C_k(t - k)$ in (4.3.4) is

$$b_0^{(k)} = \bar{z}_k(\lambda) \tag{4.3.7}$$

an exponentially weighted average of values up to the origin k. In fact, (4.3.4) may be written as

$$z_t = \bar{z}_k(\lambda) + \lambda \sum_{j=1}^{t-k-1} a_{t-j} + a_t$$

We have seen that the complementary function $C_k(t - k)$ can be thought of as telling us what is known about the future value of the process at time t, based on knowledge of the past when *we are standing at time k*. For the IMA(0, 1, 1) process, this takes the form of information about the ''level'' or location of the process $b_0^{(k)} = \bar{z}_k(\lambda)$. At time k, our knowledge of the future behavior of the process is that it will diverge from this level in accordance with the ''random walk'' represented by $\lambda \sum_{j=1}^{t-k-1} a_{t-j} + a_t$, whose expectation is zero and whose behavior we cannot predict. As soon as a new observation is available, that is, as soon as we move our origin to time $k + 1$, the level will be updated to $b_0^{(k+1)} = \bar{z}_{k+1}(\lambda)$.

Important Properties of the IMA(0, 1, 1) Process. Since the process is nonstationary, it does not vary in a stable manner about a fixed mean. However, the exponentially weighted moving average $\bar{z}_t(\lambda)$ can be regarded as measuring the local level of the process at time t. From its definition (4.3.6), we obtain the well-known recursion formula for the EWMA:

$$\bar{z}_t(\lambda) = \lambda z_t + (1 - \lambda)\bar{z}_{t-1}(\lambda) \tag{4.3.8}$$

This expression shows that for the IMA(0, 1, 1) model, each new level is arrived at by interpolating between the new observation and the previous level. If λ is equal to unity, $\bar{z}_t(\lambda) = z_t$ which would ignore all evidence concerning location coming from previous observations. On the other hand, if λ had some value close to zero, $\bar{z}_1(\lambda)$ would rely heavily on the previous value $\bar{z}_{t-1}(\lambda)$, which would have weight $1 - \lambda$. Only the small weight λ would be given to the new observation.

Now consider the two equations

$$\begin{aligned} z_t &= \bar{z}_{t-1}(\lambda) + a_t \\ \bar{z}_t(\lambda) &= \bar{z}_{t-1}(\lambda) + \lambda a_t \end{aligned} \tag{4.3.9}$$

the latter being obtained by substituting (4.3.5) in (4.3.8) and is also directly derivable from (4.3.7).

It was pointed out by Muth (1960) that the two equations (4.3.9) provide a useful way of thinking about the generation of the process. The first equation shows how, with the level of the system at $\bar{z}_{t-1}(\lambda)$, a shock a_t is added at time t and produces the value z_t. However, the second equation shows that only a proportion λ of the shock is actually absorbed into the level and has a lasting influence, the remaining proportion $\theta = 1 - \lambda$ of the shock being dissipated. Now a new level $\bar{z}_t(\lambda)$ having been established by the absorption of a_t, a new shock a_{t+1} enters the system at time $t + 1$. Equations (4.3.9), with subscripts increased by unity, will then show how this shock produces z_{t+1} and how a proportion λ of it is absorbed into the system to produce the new level $\bar{z}_{t+1}(\lambda)$, and so on.

Equation (4.3.4) can be used to obtain variance and correlation features of the IMA(0, 1, 1) process directly. For example, with reference to the origin k and treating the initializing function $b_0^{(k)}$ as constant, we find that

$$\text{var}[z_t] = \sigma_a^2[1 + (t - k - 1)\lambda^2] \tag{4.3.10}$$

which does not converge as t increases. We might also view this variance as, essentially, the variance of the difference $z_t - z_k$, treating $a_k = 0$ in (4.3.4). In particular, in the case of a random walk process, $z_t = z_{t-1} + a_t$, we have $\lambda = 1$, and this variance function grows proportionally with $t - k$, whereas for more common situations with $0 < \lambda < 1$ (i.e., $0 < \theta < 1$) and especially for λ close to zero, the variance function of $z_t - z_k$ grows much more slowly with $t - k$. In addition, for $s > 0$, $\text{cov}[z_t, z_{t+s}] = \sigma_a^2[\lambda + (t - k - 1)\lambda^2]$, which implies that $\text{corr}[z_t, z_{t+s}]$ will be close to 1 for $t - k$ large relative to s (and λ not close to zero). Hence, it follows that adjacent values of the process will be highly positively correlated, so the process will tend to exhibit rather smooth behavior (unless λ is close to zero).

The properties of the IMA(0, 1, 1) process with deterministic drift

$$\nabla z_t = \theta_0 + (1 - \theta_1 B)a_t$$

are discussed in Appendix A4.2.

4.3.2 Integrated Moving Average Process of Order (0, 2, 2)

Difference Equation Form. The IMA(0, 2, 2) process

$$\nabla^2 z_t = (1 - \theta_1 B - \theta_2 B^2)a_t \tag{4.3.11}$$

can be used to represent series exhibiting stochastic trends (e.g., see Fig. 4.6), and we now study its general properties within the invertibility region:

$$-1 < \theta_2 < 1 \quad \theta_2 + \theta_1 < 1 \quad \theta_2 - \theta_1 < 1$$

Proceeding as before, z_t can be written explicitly in terms of z's and a's as

$$z_t = 2z_{t-1} - z_{t-2} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$$

Alternatively, we can rewrite the right-hand operator in terms of differences:

$$1 - \theta_1 B - \theta_2 B^2 = (\lambda_0 \nabla + \lambda_1)B + \nabla^2$$

and on equating coefficients, we find expressions for the θ's in terms of the λ's, and vice versa, as follows:

$$\begin{aligned} \theta_1 &= 2 - \lambda_0 - \lambda_1 & \lambda_0 &= 1 + \theta_2 \\ \theta_2 &= \lambda_0 - 1 & \lambda_1 &= 1 - \theta_1 - \theta_2 \end{aligned} \tag{4.3.12}$$

The IMA(0, 2, 2) model may then be rewritten as

$$\nabla^2 z_t = (\lambda_0 \nabla + \lambda_1) a_{t-1} + \nabla^2 a_t \tag{4.3.13}$$

There is an important advantage in using this form of the model, as compared with (4.3.11). This stems from the fact that if we set $\lambda_1 = 0$ in (4.3.13), we obtain

$$\nabla z_t = [1 - (1 - \lambda_0)B]a_t$$

which corresponds to a (0, 1, 1) process, with $\theta = 1 - \lambda_0$. However, if we set $\theta_2 = 0$ in (4.3.11), we obtain

$$\nabla^2 z_t = (1 - \theta_1 B)a_t$$

As will be shown in Chapter 5, for a series generated by the (0, 2, 2) model, the optimal forecasts lie along a straight line, the level and *slope* of which are continually updated as new data become available. By contrast, a series generated by a (0, 1, 1) model can supply no information about slope but only about a continually updated level. It can be an important question whether a linear trend, as well as the level, can be forecasted and updated. When the choice is between these two models, this question turns on whether or not λ_1 in (4.3.13) is zero.

The invertibility region for an IMA(0, 2, 2) process is the same as that given for an MA(2) process in Chapter 3. It may be written in terms of the θ's and λ's as follows:

$$\begin{aligned} \theta_2 + \theta_1 &< 1 & 0 < 2\lambda_0 + \lambda_1 &< 4 \\ \theta_2 - \theta_1 &< 1 & \lambda_1 &> 0 \\ -1 < \theta_2 &< 1 & \lambda_0 &> 0 \end{aligned} \tag{4.3.14}$$

The triangular region for the θ's was shown in Figure 3.6 and the corresponding region for the λ's is shown in Figure 4.8.

Truncated and Infinite Random Shock Forms of Model. On applying the finite double summation operator $S^{(2)}_{t-k}$, relative to a time origin k, to (4.3.13), we find that

$$\begin{aligned} [1 - B^{t-k} - (t-k)B^{t-k}(1-B)]z_t = {} & [\lambda_0(S_{t-k} - (t-k)B^{t-k}) + \lambda_1 S^{(2)}_{t-k}]a_{t-1} \\ & + [1 - B^{t-k} - (t-k)B^{t-k}(1-B)]a_t \end{aligned}$$

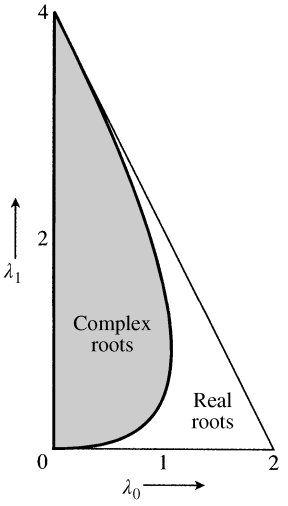


FIGURE 4.8 Invertibility region for parameters λ_0 and λ_1 of an IMA(0, 2, 2) process.

Hence, we obtain the truncated form of the random shock model as

$$\begin{aligned} z_t &= \lambda_0 S_{t-k-1} a_{t-1} + \lambda_1 S^{(2)}_{t-k-1} a_{t-1} + a_t + b_0^{(k)} + b_1^{(k)}(t-k) \\ &= \lambda_0 \sum_{j=1}^{t-k-1} a_{t-j} + \lambda_1 \sum_{j=1}^{t-k-1} j a_{t-j} + a_t + C_k(t-k) \end{aligned} \tag{4.3.15}$$

So, for this process, the ψ weights are

$$\psi_0 = 1 \quad \psi_1 = (\lambda_0 + \lambda_1) \cdots \quad \psi_j = (\lambda_0 + j\lambda_1) \cdots$$

The complementary function is the solution of

$$(1 - B)^2 C_k(t-k) = 0$$

that is,

$$C_k(t-k) = b_0^{(k)} + b_1^{(k)}(t-k) \tag{4.3.16}$$

which is a polynomial in $(t-k)$ of degree 1 whose coefficients depend on the location of the origin k. From (4.3.15), we find that these coefficients are given explicitly as

$$\begin{aligned} b_0^{(k)} &= z_k - (1 - \lambda_0) a_k \\ b_1^{(k)} &= z_k - z_{k-1} - (1 - \lambda_1) a_k + (1 - \lambda_0) a_{k-1} \end{aligned}$$

Also, by considering the differences $b_0^{(k)} - b_0^{(k-1)}$ and $b_1^{(k)} - b_1^{(k-1)}$, it follows that if the origin is updated from $k-1$ to k, then b_0 and b_1 are updated according to

$$\begin{aligned} b_0^{(k)} &= b_0^{(k-1)} + b_1^{(k-1)} + \lambda_0 a_k \\ b_1^{(k)} &= b_1^{(k-1)} + \lambda_1 a_k \end{aligned} \tag{4.3.17}$$

We see that when this model is appropriate, our expectation of the future behavior of the series, judged from origin k, would be represented by the straight line (4.3.16), having location $b_0^{(k)}$ and slope $b_1^{(k)}$. In practice, the process will, by time t, have diverged from this line because of the influence of the random component

$$\lambda_0 \sum_{j=1}^{t-k-1} a_{t-j} + \lambda_1 \sum_{j=1}^{t-k-1} j a_{t-j} + a_t$$

which at time k is unpredictable. Moreover, on moving from origin $k-1$ to origin k, the intercept and slope are updated according to (4.3.17).

Informally, through (4.3.15) we may also obtain the infinite random shock form as

$$z_t = \lambda_0 \sum_{j=1}^{\infty} a_{t-j} + \lambda_1 \sum_{j=1}^{\infty} j a_{t-j} + a_t = \lambda_0 S a_{t-1} + \lambda_1 S^2 a_{t-1} + a_t \tag{4.3.18}$$

So by comparison with (4.3.15), the complementary function can be represented informally as

$$C_k(t-k) = \lambda_0 \sum_{j=t-k}^{\infty} a_{t-j} + \lambda_1 \sum_{j=t-k}^{\infty} j a_{t-j} = b_0^{(k)} + b_1^{(k)}(t-k)$$

By writing the second infinite sum above in the form

$$\sum_{j=t-k}^{\infty} j a_{t-j} = (t-k) \sum_{j=t-k}^{\infty} a_{t-j} + \sum_{j=t-k}^{\infty} [j-(t-k)] a_{t-j}$$

we see that the coefficients $b_0^{(k)}$ and $b_1^{(k)}$ can be associated with

$$b_0^{(k)} = \lambda_0 S a_k + \lambda_1 S^2 a_{k-1} = (\lambda_0 - \lambda_1) S a_k + \lambda_1 S^2 a_k$$
$$b_1^{(k)} = \lambda_1 S a_k$$

Inverted Form of Model. Finally, we consider the model in the inverted form:

$$z_t = \sum_{j=1}^{\infty} \pi_j z_{t-j} + a_t = \bar{z}_{t-1}(\pi) + a_t$$

Using (4.2.22), we find on equating coefficients in

$$1 - 2B + B^2 = (1 - \theta_1 B - \theta_2 B^2)(1 - \pi_1 B - \pi_2 B^2 - \cdots)$$

that the π weights of the IMA(0, 2, 2) process are

$$\begin{aligned} \pi_1 &= 2 - \theta_1 = \lambda_0 + \lambda_1 \\ \pi_2 &= \theta_1(2 - \theta_1) - (1 + \theta_2) = \lambda_0 + 2\lambda_1 - (\lambda_0 + \lambda_1)^2 \\ (1 - \theta_1 B - \theta_2 B^2)\pi_j &= 0 \qquad j \geq 3 \end{aligned} \tag{4.3.19}$$

where B now operates on j.

If the roots of the characteristic equation $1 - \theta_1 B - \theta_2 B^2 = 0$ are real, the π weights are a mixture of two damped exponentials. If the roots are complex, the weights follow a

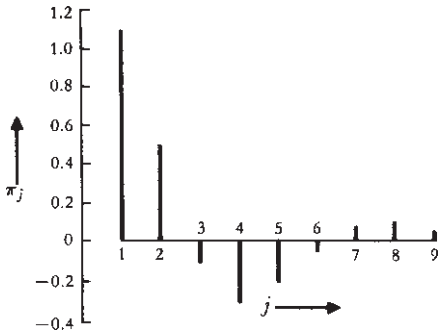


FIGURE 4.9 The π weights for an IMA(0, 2, 2) process with $\lambda_0 = 0.5$, $\lambda_1 = 0.6$.

damped sine wave. Figure 4.9 shows the weights for a process with $\theta_1 = 0.9$ and $\theta_2 = -0.5$, that is, $\lambda_0 = 0.5$ and $\lambda_1 = 0.6$. For these parameter values, the characteristic equation has complex roots (the discriminant $\theta_1^2 + 4\theta_2 = -1.19$ is less than zero). Hence, the weights in Figure 4.9 follow a damped sine wave, as expected.

4.3.3 General Integrated Moving Average Process of Order $(0, d, q)$

Difference Equation Form. The general integrated moving average process of order $(0, d, q)$ is

$$\nabla^d z_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) a_t = \theta(B) a_t \tag{4.3.20}$$

where the zeros of $\theta(B)$ must lie outside the unit circle for the process to be invertible. This model may be written explicitly in terms of past z's and a's in the form

$$z_t = d z_{t-1} - \frac{1}{2} d(d-1) z_{t-2} + \cdots + (-1)^{d+1} z_{t-d} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

Random Shock Form of Model. To obtain z_t in terms of the a_t's, we write the right-hand operator in (4.3.20) in terms of $\nabla = 1 - B$. In this way, we obtain

$$(1 - \theta_1 B - \cdots - \theta_q B^q) = (\lambda_{d-q} \nabla^{q-1} + \cdots + \lambda_0 \nabla^{d-1} + \cdots + \lambda_{d-1}) B + \nabla^d \tag{4.3.21}$$

where, as before, the λ's may be written explicitly in terms of the θ's, by equating coefficients of B.

On substituting (4.3.21) in (4.3.20) and summing d times, informally, we obtain

$$z_t = (\lambda_{d-q} \nabla^{q-d-1} + \cdots + \lambda_0 S + \cdots + \lambda_{d-1} S^d) a_{t-1} + a_t \tag{4.3.22}$$

Thus, for $q > d$, we notice that in addition to the d sums, we pick up $q - d$ additional terms $\nabla^{q-d-1} a_{t-1}, \ldots$ involving $a_{t-1}, a_{t-2}, \ldots, a_{t+d-q}$.

If we write this solution in terms of finite sums of a's entering the system after some origin k, we obtain the same form of equation, but with an added complementary function, which is the solution of

$$\nabla^d C_k(t-k) = 0$$

that is, the polynomial

$$C_k(t-k) = b_0^{(k)} + b_1^{(k)}(t-k) + b_2^{(k)}(t-k)^2 + \cdots + b_{d-1}^{(k)}(t-k)^{d-1}$$

As before, the complementary function $C_k(t-k)$ represents the finite behavior of the process, which is predictable at time k. Similarly, the coefficients $b_j^{(k)}$ may be expressed, informally, in terms of the infinite sums up to origin k, that is, $Sa_k, S^2a_k, \ldots, S^da_k$. Accordingly, we can discover how the coefficients $b_j^{(k)}$ change as the origin is changed, from $k-1$ to k.

Inverted Form of Model. Finally, the model can be expressed in the inverted form

$$\pi(B)z_t = a_t$$

or

$$z_t = \bar{z}_{t-1}(\pi) + a(t)$$

The π weights may be obtained by equating coefficients in (4.2.22), that is,

$$(1-B)^d = (1-\theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)(1-\pi_1 B - \pi_2 B^2 - \cdots) \tag{4.3.23}$$

This expression implies that for j greater than the larger of d and q, the π weights satisfy the homogeneous difference equation

$$\theta(B)\pi_j = 0$$

defined by the moving average operator. Hence, for sufficiently large j, the weights π_j follow a mixture of damped exponentials and sine waves.

IMA Process of Order (0, 2, 3). One final special case of sufficient interest to merit comment is the IMA process of order (0, 2, 3):

$$\nabla^2 z_t = (1-\theta_1 B - \theta_2 B^2 - \theta_3 B^3)a_t$$

Proceeding as before, if we apply the finite double summation operator, this model can be written in truncated random shock form as

$$z_t = \lambda_{-1}a_{t-1} + \lambda_0 \sum_{j=1}^{t-k-1} a_{t-j} + \lambda_1 \sum_{j=1}^{t-k-1} j a_{t-j} + a_t + b_0^{(k)} + b_1^{(k)}(t-k)$$

where the relations between the λ's and θ's are

$$\begin{aligned} \theta_1 &= 2 - \lambda_{-1} - \lambda_0 - \lambda_1 & \lambda_{-1} &= -\theta_3 \\ \theta_2 &= \lambda_0 - 1 + 2\lambda_{-1} & \lambda_0 &= 1 + \theta_2 + 2\theta_3 \\ \theta_3 &= -\lambda_{-1} & \lambda_1 &= 1 - \theta_1 - \theta_2 - \theta_3 \end{aligned}$$

Alternatively, it can be written, informally, in the infinite integrated form as

$$z_t = \lambda_{-1}a_{t-1} + \lambda_0 S a_{t-1} + \lambda_1 S^2 a_{t-1} + a_t$$

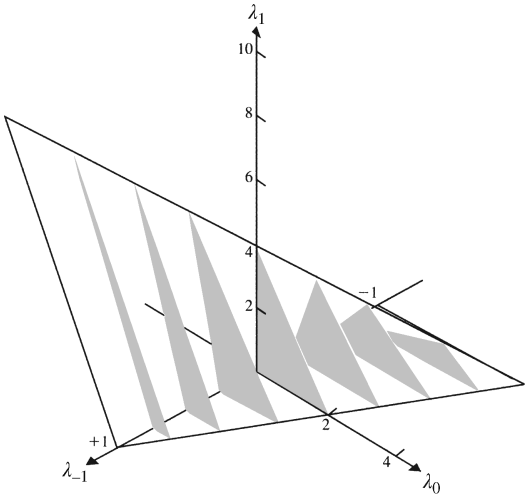


FIGURE 4.10 Invertibility region for parameters λ_{-1}, λ_0, and λ_1 and of an IMA(0, 2, 3) process.

Finally, the invertibility region is defined by

$$\begin{aligned} \theta_1 + \theta_2 + \theta_3 &< 1 & \lambda_1 &> 0 \\ -\theta_1 + \theta_2 - \theta_3 &< 1 & 2\lambda_0 + \lambda_1 &< 4(1 - \lambda_{-1}) \\ \theta_3(\theta_3 - \theta_1) - \theta_2 &< 1 & \lambda_0(1 + \lambda_{-1}) &> -\lambda_1\lambda_{-1} \\ -1 < \theta_3 &< 1 & -1 < \lambda_{-1} &< 1 \end{aligned}$$

as is shown in Figure 4.10.

In Chapter 5, we show how forecasts of future values of a time series can be generated in an optimal manner when the model is an ARIMA process. In studying these forecasts, we make considerable use of the various model forms discussed in this chapter.

APPENDIX A4.1 LINEAR DIFFERENCE EQUATIONS

In this book, we are often concerned with linear difference equations. In particular, the ARIMA model relates an output z_t to an input a_t in terms of the difference equation

$$\begin{aligned} &z_t - \varphi_1 z_{t-1} - \varphi_2 z_{t-2} - \cdots - \varphi_{p'} z_{t-p'} \\ &\quad = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} \end{aligned} \tag{A4.1.1}$$

where $p' = p + d$.

Alternatively, we may write (A4.1.1) as

$$\varphi(B) z_t = \theta(B) a_t$$

where

$$\begin{aligned} \varphi(B) &= 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_{p'} B^{p'} \\ \theta(B) &= 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \end{aligned}$$

We now derive an expression for the general solution of the difference equation (A4.1.1) relative to an origin $k < t$.

1. We show that the general solution may be written as

$$z_t = C_k(t-k) + I_k(t-k)$$

   where $C_k(t-k)$ is the complementary function and $I_k(t-k)$ is a ''particular integral.''
2. We then derive a general expression for the complementary function $C_k(t-k)$.
3. Finally, we derive a general expression for a particular integral $I_k(t-k)$.

General Solution. The argument is identical to that for the solution of linear differential or linear algebraic equations. Suppose that z_t' is any particular solution of

$$\varphi(B)z_t = \theta(B)a_t \tag{A4.1.2}$$

that is, it satisfies

$$\varphi(B)z_t' = \theta(B)a_t \tag{A4.1.3}$$

On subtracting (A4.1.3) from (A4.1.2), we obtain

$$\varphi(B)(z_t - z_t') = 0$$

Thus $z_t'' = z_t - z_t'$ satisfies

$$\varphi(B)z_t'' = 0 \tag{A4.1.4}$$

Now

$$z_t = z_t' + z_t''$$

and hence the general solution of (A4.1.2) is the sum of the complementary function z_t'', which is the general solution of the homogeneous difference equation (A4.1.4), and a particular integral z_t', which is any particular solution of (A4.1.2). Relative to any origin $k < t$, we denote the complementary function z_t'' by $C_k(t-k)$ and the particular integral z_t' by $I_k(t-k)$.

Evaluation of the Complementary Function.

Distinct Roots. Consider the homogeneous difference equation

$$\varphi(B)z_t = 0 \tag{A4.1.5}$$

where

$$\varphi(B) = (1 - G_1B)(1 - G_2B)\cdots(1 - G_{p'}B) \tag{A4.1.6}$$

and where we assume in the first instance that $G_1, G_2, \ldots, G_{p'}$ are *distinct*. Then, it is shown below that the general solution of (A4.1.5) at time t, when the series is referred to an origin

at time k, is

$$z_t = A_1 G_1^{t-k} + A_2 G_2^{t-k} + \cdots + A_{p'} G_{p'}^{t-k} \tag{A4.1.7}$$

where the A_i's are constants. Thus, a real root G of $\varphi(B) = 0$ contributes a damped exponential term G^{t-k} to the complementary function. A pair of complex roots contributes a damped sine wave term $D^{t-k}\sin(2\pi f_0 t + F)$.

To see that the expression given in (A4.1.7) does satisfy (A4.1.5), we can substitute (A4.1.7) in (A4.1.5) to give

$$\varphi(B)(A_1 G_1^{t-k} + A_2 G_2^{t-k} + \cdots + A_{p'} G_{p'}^{t-k}) = 0 \tag{A4.1.8}$$

Now consider

$$\begin{aligned}\varphi(B) G_i^{t-k} &= (1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_{p'} B^{p'}) G_i^{t-k} \\ &= G_i^{t-k-p'}(G_i^{p'} - \varphi_1 G_i^{p'-1} - \cdots - \varphi_{p'})\end{aligned}$$

We see that $\varphi(B) G_i^{t-k}$ vanishes for each value of i if

$$G_i^{p'} - \varphi_1 G_i^{p'-1} - \cdots - \varphi_{p'} = 0$$

that is, if $B_i = 1/G_i$ is a root of $\varphi(B) = 0$. Now, since (A4.1.6) implies that the roots of $\varphi(B) = 0$ are $B_i = 1/G_i$, it follows that $\varphi(B) G_i^{t-k}$ is zero for all i and hence (A4.1.8) holds, confirming that (A4.1.7) is a general solution of (A4.1.5).

To prove (A4.1.7) directly, consider the special case of the second-order equation:

$$(1 - G_1 B)(1 - G_2 B) z_t = 0$$

which we can write as

$$(1 - G_1 B) y_t = 0 \tag{A4.1.9}$$

where

$$y_t = (1 - G_2 B) z_t \tag{A4.1.10}$$

Now (A4.1.9) implies that

$$y_t = G_1 y_{t-1} = G_1^2 y_{t-2} = \cdots = G_1^{t-k} y_k$$

and hence

$$y_t = D_1 G_1^{t-k}$$

where $D_1 = y_k$ is a constant determined by the starting value y_k. Hence (A4.1.10) may be written as

$$\begin{aligned} z_t &= G_2 z_{t-1} + D_1 G_1^{t-k} \\ &= G_2(G_2 z_{t-2} + D_1 G_1^{t-k-1}) + D_1 G_1^{t-k} \\ &\quad \vdots \\ &= G_2^{t-k} z_k + D_1(G_1^{t-k} + G_2 G_1^{t-k-1} + \cdots + G_2^{t-k-1} G_1) \\ &= G_2^{t-k} z_k + \frac{D_1}{1 - G_2/G_1}(G_1^{t-k} - G_2^{t-k}) \\ &= A_1 G_1^{t-k} + A_2 G_2^{t-k} \end{aligned} \tag{A4.1.11}$$

where A_1, A_2 are constants determined by the starting values of the series. By an extension of the argument above, it may be shown that the general solution of (A4.1.5), when the roots of $\varphi(B) = 0$ are distinct, is given by (A4.1.7).

Equal Roots. Suppose that $\varphi(B) = 0$ has d equal roots G_0^{-1}, so that $\varphi(B)$ contains a factor $(1 - G_0 B)^d$. In particular, consider the solution (A4.1.11) for the second-order equation when both G_1 and G_2 are equal to G_0. Then, (A4.1.11) reduces to

$$z_t = G_0^{t-k} z_k + D_1 G_0^{t-k}(t-k)$$

or

$$z_t = [A_0 + A_1(t-k)] G_0^{t-k}$$

In general, if there are d equal roots G_0, it may be verified by direct substitution in (A4.1.5) that the general solution is

$$\begin{aligned} z_t = [&A_0 + A_1(t-k) + A_2(t-k)^2 + \cdots \\ &+ A_{d-1}(t-k)^{d-1}] G_0^{t-k} \end{aligned} \tag{A4.1.12}$$

In particular, when the equal roots G_0 are all equal to unity as in the IMA $(0, d, q)$ process, the solution is

$$z_t = A_0 + A_1(t-k) + A_2(t-k)^2 + \cdots + A_{d-1}(t-k)^{d-1} \tag{A4.1.13}$$

that is, a polynomial in $t - k$ of degree $d - 1$.

In general, when $\varphi(B)$ factors according to

$$(1 - G_1 B)(1 - G_2 B) \cdots (1 - G_p B)(1 - G_0 B)^d$$

the complementary function is

$$C_k(t-k) = G_0^{t-k} \sum_{j=0}^{d-1} A_j (t-k)^j + \sum_{i=1}^{p} D_i G_i^{t-k} \tag{A4.1.14}$$

Thus, in general, the complementary function consists of a mixture of damped exponential terms G^{t-k}, polynomial terms $(t-k)^j$, damped sine wave terms of the form $D^{t-k} \sin(2\pi f_0 t + F)$, and combinations of these functions.

Evaluation of the "Particular Integral". We now show that a particular integral $I_k(s-k)$, satisfying

$$\varphi(B)I_k(t-k) = \theta(B)a_t \quad t-k > q \tag{A4.1.15}$$

is a function defined as follows:

$$\begin{aligned} I_k(s-k) &= 0 \quad s \le k \\ I_k(1) &= a_{k+1} \\ I_k(2) &= a_{k+2} + \psi_1 a_{k+1} \\ &\vdots \\ I_k(t-k) &= a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots + \psi_{t-k-1} a_{k+1} \quad t > k \end{aligned} \tag{A4.1.16}$$

where the ψ weights are those appearing in the form (4.2.3) of the model. Thus, the ψ weights satisfy

$$\varphi(B)\psi(B)a_t = \theta(B)a_t \tag{A4.1.17}$$

Now the terms on the left-hand side of (A4.1.17) may be set out as follows:

$$\begin{array}{r|l} a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots + \psi_{t-k-1} a_{k+1} & +\psi_{t-k} a_k + \cdots \\ -\varphi_1(a_{t-1} + \psi_1 a_{t-2} + \cdots + \psi_{t-k-2} a_{k+1} & +\psi_{t-k-1} a_k + \cdots) \\ -\varphi_2(\quad \cdots & \cdots) \\ \vdots & \vdots \\ -\varphi_{p'}(a_{t-p'} + \cdots + \psi_{t-k-p'-1} a_{k+1} & +\psi_{t-k-p'} a_k + \cdots \end{array} \tag{A4.1.18}$$

Since the right-hand side of (A4.1.17) is

$$a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

it follows that the first $q+1$ columns in this array sum to $a_t, -\theta_1 a_{t-1}, \ldots, -\theta_q a_{t-q}$. Now the left-hand term in (A4.1.15), where $I_k(s-k)$ is given by (A4.1.16), is equal to the sum of the terms in the first $(t-k)$ columns of the array, that is, those to the left of the vertical line. Therefore, if $t-k < q$, that is, the vertical line is drawn after $q+1$ columns, the sum of all terms up to the vertical line is equal to $\theta(B)a_t$. This shows that (A4.1.16) is a particular integral of the difference equation.

Example. Consider the IMA(0, 1, 1) process

$$z_t - z_{t-1} = a_t - \theta a_{t-1} \tag{A4.1.19}$$

for which $\psi_j = 1-\theta$ for $j \ge 1$. Then

$$\begin{aligned} I_k(0) &= 0 \\ I_k(1) &= a_{k+1} \\ &\vdots \end{aligned} \tag{A4.1.20}$$

$$I_k(t-k) = a_t + (1-\theta)\sum_{j=1}^{t-k-1} a_{t-j} \quad t-k>1$$

Now if $z_t = I_k(t-k)$ is a solution of (A4.1.19), then

$$I_k(t-k) - I_k(t-k-1) = a_t - \theta a_{t-1}$$

and as is easily verified, while this is not satisfied by (A4.1.20) for $t-k=1$, it is satisfied by (A4.1.20) for $t-k>1$, that is, for $t-k>q$.

APPENDIX A4.2 IMA(0, 1, 1) PROCESS WITH DETERMINISTIC DRIFT

The general model $\phi(B)\nabla^d z_t = \theta_0 + \theta(B)a_t$ can also be written as

$$\phi(B)\nabla^d z_t = \theta(B)\varepsilon_t$$

with the shocks ε_t having a nonzero mean $\xi = \theta_0/(1-\theta_1-\cdots-\theta_q)$. For example, the IMA(0, 1, 1) model is then

$$\nabla z_t = (1-\theta B)\varepsilon_t$$

with $E[\varepsilon_t] = \xi = \theta_0/(1-\theta)$. In this form, z_t could represent, for example, the outlet temperature from a reactor when heat was being supplied from a heating element at a fixed rate. Now if

$$\varepsilon_t = \xi + a_t \tag{A4.2.1}$$

where a_t is white noise with zero mean, then with reference to a time origin k, the integrated form of the model is

$$z_t = b_0^{(k)} + \lambda\sum_{j=1}^{t-k-1}\varepsilon_{t-j} + \varepsilon_t \tag{A4.2.2}$$

with $\lambda = 1-\theta$. Substituting for (A4.2.1) in (A4.2.2), the model written in terms of the a's is

$$z_t = b_0^{(k)} + \lambda\xi(t-k-1) + \xi + \lambda\sum_{j=1}^{t-k-1} a_{t-j} + a_t \tag{A4.2.3}$$

Thus, we see that z_t contains a deterministic slope or drift due to the term $\lambda\xi(t-k-1)$, with the slope of the deterministic linear trend equal to $\lambda\xi = \theta_0$. Moreover, if we denote the ''level'' of the process at time $t-1$ by l_{t-1}, where

$$z_t = l_{t-1} + a_t$$

we see that the level is changed from time $t-1$ to time t, according to

$$l_t = l_{t-1} + \lambda\xi + \lambda a_t$$

The change in the level, thus, contains a deterministic component $\lambda\xi = \theta_0$, as well as a stochastic component λa_t.

APPENDIX A4.3 ARIMA PROCESSES WITH ADDED NOISE

In this appendix, we consider the effect of added noise (e.g., measurement error) to a general ARIMA(p, d, q) process. The results are also relevant to determine the nature of the reduced form ARIMA model of an observed process in structural component models (see Section 9.4), in which an observed series Z_t is presumed to be represented as the sum of two unobservable component processes that follow specified ARIMA models.

A4.3.1 Sum of Two Independent Moving Average Processes

As a necessary preliminary to what follows, consider a stochastic process w_t, which is the sum of two *independent* moving average processes of orders q_1 and q_2, respectively. That is,

$$w_t = w_{1t} + w_{2t} = \theta_1(B)a_t + \theta_2(B)b_t \tag{A4.3.1}$$

where $\theta_1(B)$ and $\theta_2(B)$ are polynomials in B, of orders q_1 and q_2, and the white noise processes a_t and b_t have zero means, variances σ_a^2 and σ_b^2, and are mutually independent. Suppose that $q = \max(q_1, q_2)$; then since

$$\gamma_j(w) = \gamma_j(w_1) + \gamma_j(w_2)$$

it is clear that the autocovariance function $\gamma_j(w)$ for w_t must be zero for $j > q$. It follows that there exists a representation of w_t as a single MA(q) process:

$$w_t = \theta(B)u_t \tag{A4.3.2}$$

where u_t is a white noise process with mean zero and variance σ_u^2. Thus, the sum of two independent moving average processes is another moving average process, whose order is the same as that of the component process of higher order.

The parameters in the MA(q) model can be deduced by equating the autocovariances of w_t, as determined from the representation in (A4.3.1), with the autocovariances of the basic MA(q) model (A4.3.2), as given in Section 3.3.2. For an example, suppose that $w_{1t} = \theta_1(B)a_t = (1 - \theta_{1,1}B)a_t$ is MA(1) and $w_{2t} = \theta_2(B)b_t = (1 - \theta_{1,2}B - \theta_{2,2}B^2)b_t$ is MA(2), so that $w_t = \theta(B)u_t$ is MA(2) with

$$\begin{aligned} w_t &= (1 - \theta_{1,1}B)a_t + (1 - \theta_{1,2}B - \theta_{2,2}B^2)b_t \\ &= (1 - \theta_1 B - \theta_2 B^2)u_t \end{aligned}$$

The parameters of the MA(2) model for w_t can be determined by considering

$$\begin{aligned} \gamma_0(w) &= (1 + \theta_{1,1}^2)\sigma_a^2 + (1 + \theta_{1,2}^2 + \theta_{2,2}^2)\sigma_b^2 \equiv (1 + \theta_1^2 + \theta_2^2)\sigma_u^2 \\ \gamma_1(w) &= -\theta_{1,1}\sigma_a^2 + (-\theta_{1,2} + \theta_{1,2}\theta_{2,2})\sigma_b^2 \equiv (-\theta_1 + \theta_1\theta_2)\sigma_u^2 \\ \gamma_2(w) &= -\theta_{2,2}\sigma_b^2 \equiv -\theta_2\sigma_u^2 \end{aligned}$$

and solving for θ_1, θ_2, and σ_u^2 in terms of given values for the autocovariances $\gamma_0(w)$, $\gamma_1(w)$, $\gamma_2(w)$ as determined from the left-hand-side expressions for these.

A4.3.2 Effect of Added Noise on the General Model

Correlated Noise. Consider the general nonstationary model for the process z_t of order (p, d, q):

$$\phi(B)\nabla^d z_t = \theta(B)a_t \tag{A4.3.3}$$

Suppose that we cannot observe z_t itself, but only $Z_t = z_t + b_t$, where b_t represents some extraneous noise (e.g., measurement error) or simply some additional unobserved component that together with z_t forms the observed process Z_t, and b_t may be autocorrelated. We wish to determine the nature of the model for the observed process Z_t. In general, applying $\phi(B)\nabla^d$ to both sides of $Z_t = z_t + b_t$, we have

$$\phi(B)\nabla^d Z_t = \theta(B)a_t + \phi(B)\nabla^d b_t$$

If the noise b_t follows a stationary ARMA process of order $(p_1, 0, q_1)$,

$$\phi_1(B)b_t = \theta_1(B)\alpha_t \tag{A4.3.4}$$

where α_t is a white noise process independent of the a_t process, then

$$\underbrace{\phi_1(B)\phi(B)\nabla^d}_{p_1+p+d} Z_t = \underbrace{\phi_1(B)\theta(B)}_{p_1+q} a_t + \underbrace{\phi(B)\theta_1(B)\nabla^d}_{p+q_1+d} \alpha_t \tag{A4.3.5}$$

where the values below the braces indicate the degrees of the various polynomials in B. Now the right-hand side of (A4.3.5) is of the form (A4.3.1). Let $P = p_1 + p$ and Q be equal to whichever of $(p_1 + q)$ and $(p + q_1 + d)$ is larger. Then we can write

$$\phi_2(B)\nabla^d Z_t = \theta_2(B)u_t$$

with u_t a white noise process, and the Z_t process is seen to be an ARIMA of order (P, d, Q). The stationary AR operator in the ARIMA model for Z_t is determined as $\phi_2(B) = \phi_1(B)\phi(B)$, and the parameters of the MA operator $\theta_2(B)$ and σ_u^2 are determined in the same manner as described in Section A4.3.1, that is, by equating the nonzero autocovariances from the representations:

$$\phi_1(B)\theta(B)a_t + \phi(B)\theta_1(B)\nabla^d\alpha_t = \theta_2(B)u_t$$

Added White Noise. If, as might be true in some applications, the added noise is white, then $\phi_1(B) = \theta_1 B = 1$ in (A4.3.4), and we obtain

$$\phi(B)\nabla^d Z_t = \theta_2(B)u_t \tag{A4.3.6}$$

with

$$\theta_2(B)u_t = \theta(B)a_t + \phi(B)\nabla^d b_t$$

which is of order (p, d, Q) where Q is the larger of q and $(p + d)$. If $p + d \leq q$, the order of the process with error is the same as that of the original process. The only effect of the added white noise is to change the values of the θ's (but not the ϕ's).

Effect of Added White Noise on an Integrated Moving Average Process. In particular, an IMA process of order $(0, d, q)$, with white noise added, remains an IMA of order $(0, d, q)$ if $d \leq q$; otherwise, it becomes an IMA of order $(0, d, d)$. In either case, the parameters of the process are changed by the addition of noise, with the representation $\nabla^d Z_t = \theta_2(B)u_t$ as in (A4.3.6). The nature of these changes can be determined by equating the autocovariances of the dth differences of the process, with added noise, to those of the dth differences of a simple IMA process, that is, as a special case of the above, by equating the nonzero autocovariances in the representation

$$\theta(B)a_t + \nabla^d b_t = \theta_2(B)u_t$$

The procedure will now be illustrated with an example.

A4.3.3 Example for an IMA(0, 1, 1) Process with Added White Noise

Consider the properties of the process $Z_t = z_t + b_t$ when

$$z_t = z_{t-1} - (1 - \lambda)a_{t-1} + a_t \tag{A4.3.7}$$

and the b_t and a_t are mutually independent white noise processes. The Z_t process has first difference $W_t = Z_t - Z_{t-1}$ given by

$$W_t = [1 - (1 - \lambda)B]a_t + (1 - B)b_t \tag{A4.3.8}$$

The autocovariances for the first differences W_t are

$$\begin{aligned} \gamma_0 &= \sigma_a^2[1 + (1 - \lambda)^2] + 2\sigma_b^2 \\ \gamma_1 &= -\sigma_a^2(1 - \lambda) - \sigma_b^2 \\ \gamma_j &= 0 \qquad j \geq 2 \end{aligned} \tag{A4.3.9}$$

The fact that the γ_j are zero beyond the first lag confirms that the process with added noise is, as expected, an IMA process of order (0, 1, 1). To obtain explicitly the parameters of the IMA that represents the noisy process, we suppose that it can be written as

$$Z_t = Z_{t-1} - (1 - \Lambda)u_{t-1} + u_t \tag{A4.3.10}$$

where u_t is a white noise process. The process (A4.3.10) has first differences $W_t = Z_t - Z_{t-1}$ with autocovariances

$$\begin{aligned} \gamma_0 &= \sigma_u^2[1 + (1 - \Lambda)^2] \\ \gamma_1 &= -\sigma_u^2(1 - \Lambda) \\ \gamma_j &= 0 \qquad j \geq 2 \end{aligned} \tag{A4.3.11}$$

Equating (A4.3.9) and (A4.3.11), we can solve for Λ and σ_u^2 explicitly. Thus

$$\frac{\Lambda^2}{1-\Lambda^2} = \frac{\lambda^2}{1-\lambda+\sigma_b^2/\sigma_a^2}$$

$$\sigma_u^2 = \sigma_a^2 \frac{\lambda^2}{\Lambda^2} \tag{A4.3.12}$$

Suppose, for example, that the original series has $\lambda = 0.5$ and $\sigma_b^2 = \sigma_a^2$; then, $\Lambda = 0.333$ and $\sigma_u^2 = 2.25\sigma_a^2$.

A4.3.4 Relation between the IMA(0, 1, 1) Process and a Random Walk

The process

$$z_t = z_{t-1} + a_t \tag{A4.3.13}$$

which is an IMA(0, 1, 1) process, with $\lambda = 1(\theta = 0)$, is called a *random walk*. If the a_t are steps taken forward or backward at time t, then z_t will represent the position of the walker at time t.

Any IMA(0, 1, 1) process can be thought of as a random walk buried in white noise b_t, uncorrelated with the shocks a_t associated with the random walk process. If the noisy process is $Z_t = z_t + b_t$, where z_t is defined by (A4.3.13), then using (A4.3.12), we have

$$Z_t = Z_{t-1} - (1-\Lambda)u_{t-1} + u_t$$

with

$$\frac{\Lambda^2}{1-\Lambda^2} = \frac{\sigma_a^2}{\sigma_b^2} \qquad \sigma_u^2 = \frac{\sigma_a^2}{\Lambda^2} \tag{A4.3.14}$$

A4.3.5 Autocovariance Function of the General Model with Added Correlated Noise

Suppose that the basic process is an ARIMA process of order (p, d, q):

$$\phi(B)\nabla^d z_t = \theta(B)a_t$$

and that $Z_t = z_t + b_t$ is observed, where the stationary process b_t, which has autocovariance function $\gamma_j(b)$, is independent of the process a_t, and hence of z_t. Suppose that $\gamma_j(w)$ is the autocovariance function for $w_t = \nabla^d z_t = \phi^{-1}(B)\theta(B)a_t$ and that $W_t = \nabla^d Z_t$. We require the autocovariance function for W_t. Now

$$\nabla^d(Z_t - b_t) = \phi^{-1}(B)\theta(B)a_t$$

$$W_t = w_t + v_t$$

where

$$v_t = \nabla^d b_t = (1-B)^d b_t$$

Hence

$$\begin{aligned}\gamma_j(W) &= \gamma_j(w) + \gamma_j(v)\\ \gamma_j(v) &= (1-B)^d(1-F)^d\gamma_j(b)\\ &= (-1)^d(1-B)^{2d}\gamma_{j+d}(b)\end{aligned}$$

and

$$\gamma_j(W) = \gamma_j(w) + (-1)^d(1-B)^{2d}\gamma_{j+d}(b) \tag{A4.3.15}$$

For example, suppose that correlated noise b_t is added to an IMA(0, 1, 1) process defined by $w_t = \nabla z_t = (1-\theta B)a_t$. Then the autocovariances of the first difference W_t of the ''noisy'' process will be

$$\begin{aligned}\gamma_0(W) &= \sigma_a^2(1+\theta^2) + 2[\gamma_0(b) - \gamma_1(b)]\\ \gamma_1(W) &= -\sigma_a^2\theta + [2\gamma_1(b) - \gamma_0(b) - \gamma_2(b)]\\ \gamma_j(W) &= [2\gamma_j(b) - \gamma_{j-1}(b) - \gamma_{j+1}(b)] \qquad j \geq 2\end{aligned}$$

In particular, if b_t was first-order autoregressive, so that $b_t = \phi b_{t-1} + \alpha_t$,

$$\begin{aligned}\gamma_0(W) &= \sigma_a^2(1+\theta^2) + 2\gamma_0(b)(1-\phi)\\ \gamma_1(W) &= -\sigma_a^2\theta - \gamma_0(b)(1-\phi)^2\\ \gamma_j(W) &= -\gamma_0(b)\phi^{j-1}(1-\phi)^2 \qquad j \geq 2\end{aligned}$$

where $\gamma_0(b) = \sigma_\alpha^2/(1-\phi^2)$. In fact, from (A4.3.5), the resulting noisy process $Z_t = z_t + b_t$ is in this case defined by

$$(1-\phi B)\nabla Z_t = (1-\phi B)(1-\theta B)a_t + (1-B)\alpha_t$$

which will be of order (1, 1, 2), and for the associated ARMA(1, 2) process $W_t = \nabla Z_t$, we know that the autocovariances satisfy $\gamma_j(W) = \phi\gamma_{j-1}(W)$ for $j \geq 3$ [e.g., see (3.4.3)] as is shown explicitly above.

EXERCISES

4.1. For each of the models

(1) $(1-B)z_t = (1-0.5B)a_t$

(2) $(1-B)z_t = (1-0.2B)a_t$

(3) $(1-0.5B)(1-B)z_t = a_t$

(4) $(1-0.2B)(1-B)z_t = a_t$

(5) $(1-0.2B)(1-B)z_t = (1-0.5B)a_t$

(a) Obtain the first seven ψ_j weights.

(b) Obtain the first seven π_j weights.

(c) Classify as a member of the class of ARIMA(*p, d, q*) processes.

4.2. For the five models of Exercise 4.1, and using where appropriate the results there obtained,

(a) Write each model in random shock form.

(b) Write each model as a complementary function plus a particular integral in relation to an origin $k = t - 3$.

(c) Write each model in inverted form.

4.3. Consider the IMA(0, 2, 2) process with parameters $\theta_1 = 0.8$ and $\theta_2 = -0.4$.

(a) Is the process invertible? If so, what is the expected pattern of the π weights?

(b) Calculate and plot the first ten π weights for the original series z_t and comment.

(c) Calculate and plot the first ten π weights for the differenced series $w_t = (1 - B)^2 z_t$.

4.4 Given the following series of random shocks a_t, and given that $z_0 = 20, z_{-1} = 19$,

t	a_t	t	a_t	t	a_t
0	−0.3	5	−0.6	10	−0.4
1	0.6	6	1.7	11	0.9
2	0.9	7	−0.9	12	0.0
3	0.2	8	−1.3	13	−1.4
4	0.1	9	−0.6	14	−0.6

(a) Use the difference equation form of the model to obtain $z_1, z_2, \ldots, z_{14}$ for each of the five models in Exercise 4.1.

(b) Plot the resulting series.

4.5. Using the inverted forms of each of the models in Exercise 4.1, obtain z_{12}, z_{13}, and z_{14}, using only the values $z_1, z_2, \ldots, z_{11}$ derived in Exercise 4.4 and a_{12}, a_{13}, and a_{14}. Confirm that the values agree with those obtained in Exercise 4.4.

4.6. Consider the IMA(0, 1, 1) model $(1 - B)z_t = (1 - \theta)a_t$, where the a_t are i.i.d. $N(0, \sigma_a^2)$.

(a) Derive the expected value and variance of z_t, $t = 1, 2, \ldots$, assuming that the process starts at time $t = 1$ with $z_0 = 10$.

(b) Derive the correlation coefficient ρ_k between z_t and z_{t-k}, conditioning on $z_0 = 10$. Assume that t is much larger than the lag k.

(c) Provide an approximate value for the autocorrelation coefficient ρ_k derived in part (c).

4.7. If $\bar{z}_t = \sum_{j=1}^{\infty} \pi_j z_{t+1-j}$, then for models (1) and (2) of Exercise 4.1, which are of the form $(1 - B)z_t = (1 - \theta B)a_t$, $\bar{z}_t$ is an exponentially weighted moving average. For these two models, by actual calculation, confirm that $\bar{z}_{11}, \bar{z}_{12}$, and $\bar{z}_{13}$ satisfy

the relations

$$
\begin{aligned}
z_t &= \bar{z}_{t-1} + a_t \quad \text{(see Exercise 4.5)} \\
\bar{z}_t &= \bar{z}_{t-1} + (1-\theta)a_t \\
&= (1-\theta)z_t + \theta \bar{z}_{t-1}
\end{aligned}
$$

4.8. If $w_{1t} = (1 - \theta_1 B)a_{1t}$ and $w_{2t} = (1 - \theta_2 B)a_{2t}$, show that $w_{3t} = w_{1t} + w_{2t}$ may be written as $w_{3t} = (1 - \theta_3 B)a_{3t}$, and derive an expression for θ_3 and $\sigma^2_{a_3}$ in terms of the parameters of the other two processes. State your assumptions.

4.9. Suppose that $Z_t = z_t + b_t$, where z_t is a first-order autoregressive process $(1 - \phi B)z_t = a_t$ and b_t is a white noise process with variance σ^2_b. What model does the process Z_t follow? State your assumptions.

4.10. **(a)** Simulate a time series of $N = 200$ observations from an IMA(0, 2, 2) model with parameters $\theta_1 = 0.8$ and $\theta_2 = -0.4$ using the arima.sim() function in R; type help(arima.sim) for details. Plot the resulting series and comment on its behavior.

(b) Estimate and plot the autocorrelation function of the simulated time series.

(c) Estimate and plot the autocorrelation functions of the first and second differences of the series.

(d) Comment on the patterns of the autocorrelation functions generated above. Are the results consistent with what you would expect to see for this IMA(0, 2, 2) process?

4.11. Download the daily S& P 500 Index stock price values for the period January 2, 2014 to present from the Internet (e.g., http://research.stlouisfed.org).

(a) Plot the series using R. Calculate and graph the autocorrelation and partial autocorrelation functions for this series. Does the series appear to be stationary?

(b) Repeat the calculations in part **(a)** for the first and second differences of the series. Describe the effects of differencing in this case. Can you suggest a model that might be appropriate for this series?

(c) The return or relative gain on a stock can be calculated as $(z_t - z_{t-1})/z_t$ or $\log(z_t) - \log(z_{t-1})$. Perform this calculation and comment on the stationarity of the resulting series.

4.12. Repeat the analysis in Exercise 11 for the Dow Jones Industrial Average, or for a time series of your own choosing.

5

FORECASTING

In Chapter 4, we discussed the properties of autoregressive integrated moving average (ARIMA) models and examined in detail some special cases that appear to be common in practice. We will now show how these models may be used to forecast future values of an observed time series. In Part Two, we will consider the problem of selecting a suitable model of this form and fitting it to actual data. For the present, however, we proceed as if the model were known *exactly,* bearing in mind that estimation errors in the parameters will not seriously affect the forecasts unless the time series is relatively short.

This chapter will focus on nonseasonal time series. The forecasting, as well as model fitting, of seasonal time series is described in Chapter 9. We show how minimum mean square error (MSE) forecasts may be generated directly from the *difference equation* form of the model. A further recursive calculation yields probability limits for the forecasts. It is emphasized that for practical computation of the forecasts, this approach via the difference equation is the simplest and most elegant. However, to provide insight into the nature of the forecasts, we also consider them from other viewpoints. As a computational tool, we also demonstrate how to generate forecasts and associated probability limits using the R software.

5.1 MINIMUM MEAN SQUARE ERROR FORECASTS AND THEIR PROPERTIES

In Section 4.2, we discussed three explicit forms for the general ARIMA model:

$$\varphi(B)z_t = \theta(B)a_t \tag{5.1.1}$$

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

where $\varphi(B) = \phi(B)\nabla^d$. We begin by recalling these three forms since each one sheds light on a different aspect of the forecasting problem.

We will consider forecasting a value z_{t+l}, $l \geq 1$, when we are currently at time t. This forecast is said to be made at *origin* t for lead *time* l. We now summarize the results of Section 4.2, but writing $t + l$ for t and t for k.

Three Explicit Forms for the Model. An observation z_{t+l} generated by the ARIMA process may be expressed as follows:

1. Directly in terms of the difference equation by

$$z_{t+l} = \varphi_1 z_{t+l-1} + \cdots + \varphi_{p+d} z_{t+l-p-d} - \theta_1 a_{t+l-1} - \cdots - \theta_q a_{t+l-q} + a_{t+l} \tag{5.1.2}$$

2. As an infinite weighted sum of current and previous shocks a_j:

$$z_{t+l} = \sum_{j=0}^{\infty} \psi_j a_{t+l-j} \tag{5.1.3}$$

where $\psi_0 = 1$ and, as in (4.2.5), the ψ weights may be obtained by equating coefficients in

$$\varphi(B)(1 + \psi_1 B + \psi_2 B^2 + \cdots) = \theta(B) \tag{5.1.4}$$

Equivalently, for positive l, with reference to origin $k < t$, the model may be written in the truncated form:

$$\begin{aligned} z_{t+l} &= a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} \\ &\quad + \psi_l a_t + \cdots + \psi_{t+l-k-1} a_{k+1} + C_k(t + l - k) \\ &= a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} + C_t(l) \end{aligned} \tag{5.1.5}$$

where $C_k(t + l - k)$ is the complementary function relative to the finite origin k of the process. From (4.2.19), we recall that the complementary function relative to the forecast origin t can be expressed as $C_t(l) = C_k(t + l - k) + \psi_l a_t + \psi_{l+1} a_{t-1} + \cdots + \psi_{t+l-k-1} a_{k+1}$. Informally, $C_t(l)$ is associated with the truncated infinite sum:

$$C_t(l) = \sum_{j=l}^{\infty} \psi_j a_{t+l-j} \tag{5.1.6}$$

3. As an infinite weighted sum of previous observations, plus a random shock,

$$z_{t+l} = \sum_{j=1}^{\infty} \pi_j z_{t+l-j} + a_{t+l} \tag{5.1.7}$$

Also, if $d \geq 1$,

$$\bar{z}_{t+l-1}(\pi) = \sum_{j=1}^{\infty} \pi_j z_{t+l-j} \tag{5.1.8}$$

will be a weighted average, since then $\sum_{j=1}^{\infty} \pi_j = 1$. As in (4.2.22), the π weights may be obtained from

$$\varphi(B) = (1 - \pi_1 B - \pi_2 B^2 - \cdots)\theta(B) \tag{5.1.9}$$

5.1.1 Derivation of the Minimum Mean Square Error Forecasts

Now suppose, at origin t, that we are to make a forecast $\hat{z}_t(l)$ of z_{t+l}, which is to be a linear function of current and previous observations $z_t, z_{t-1}, z_{t-2}, \ldots$. Then, it will also be a linear function of current and previous shocks $a_t, a_{t-1}, a_{t-2}, \ldots$.

Suppose, then, that the best forecast is

$$\hat{z}_t(l) = \psi_l^* a_t + \psi_{l+1}^* a_{t-1} + \psi_{l+2}^* a_{t-2} + \cdots$$

where the weights $\psi_l^*, \psi_{l+1}^*, \ldots$ are to be determined. Then, using (5.1.3), the mean square error of the forecast is

$$\begin{aligned} E[z_{t+l} - \hat{z}_t(l)]^2 =& (1 + \psi_1^2 + \cdots + \psi_{l-1}^2)\sigma_a^2 \\ &+ \sum_{j=0}^{\infty} (\psi_{l+j} - \psi_{l+j}^*)^2 \sigma_a^2 \end{aligned} \tag{5.1.10}$$

which is minimized by setting $\psi_{l+j}^* = \psi_{l+j}$. This conclusion is a special case of more general results in prediction theory (Wold, 1938; Kolmogoroff (1939, 1941a, 1941b), Wiener, 1949; Whittle, 1963). We then have

$$\begin{aligned} z_{t+l} &= (a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1}) \\ &\quad + (\psi_l a_t + \psi_{l+1} a_{t-1} + \cdots) \qquad (5.1.11) \\ &= e_t(l) + \hat{z}_t(l) \qquad (5.1.12) \end{aligned}$$

where $e_t(l)$ is the error of the forecast $\hat{z}_t(l)$ at lead time l.

Certain important facts emerge. As before, denote $E[z_{t+l}|z_t, z_{t-1}, \ldots]$, the conditional expectation of z_{t+l} given the knowledge of all the z's up to time t, by $E_t[z_{t+l}]$. We will assume that a_t are a sequence of independent random variables.

1. Then, $E[a_{t+j}|z_t, z_{t-1}, \ldots] = 0$, $j > 0$, and so from (5.1.3),

$$\hat{z}_t(l) = \psi_l a_t + \psi_{l+1} a_{t-1} + \cdots = E_t[z_{t+l}] \tag{5.1.13}$$

Thus, the minimum mean square error forecast at origin t, for lead time l, is the conditional expectation of z_{t+l} at time t. When $\hat{z}_t(l)$ is regarded as a function of l for fixed t, it will be called the *forecast function* for origin t. We note that a minimal requirement on the random shocks a_t in the model (5.1.1) in order for the conditional expectation $E_t[z_{t+l}]$, which always equals the minimum mean square error forecast, to coincide with the minimum mean square error *linear* forecast is that $E_t[a_{t+j}] = 0$, $j > 0$. This property may not hold for certain types of nonlinear processes studied, for example, by Priestley (1988), Tong (1983, 1990), and many subsequent authors. Such processes may, in fact, possess a linear representation as in (5.1.1), but the shocks a_t will not be independent, only uncorrelated, and the best forecast $E_t[z_{t+l}]$ may not coincide with the best linear forecast $\hat{z}_t(l)$ as obtained in (5.1.11).

2. The forecast error for lead time l is

$$e_t(l) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} \tag{5.1.14}$$

Since

$$E_t[e_t(l)] = 0 \tag{5.1.15}$$

the forecast is unbiased. Also, the variance of the forecast error is

$$V(l) = \text{var}[e_t(l)] = (1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{l-1}^2)\sigma_a^2 \tag{5.1.16}$$

3. It is readily shown that not only is $\hat{z}_t(l)$ the minimum mean square error forecast of z_{t+l}, but that any linear function $\sum_{l=1}^{L} w_l \hat{z}_t(l)$ of the forecasts is also a minimum mean square error forecast of the corresponding linear function $\sum_{l=1}^{L} w_l z_{t+l}$ of the future observations. For example, suppose that using (5.1.13), we have obtained, from monthly data, minimum mean square error forecasts $\hat{z}_t(1)$, $\hat{z}_t(2)$, and $\hat{z}_t(3)$ of the sales of a product 1, 2, and 3 months ahead. Then, it is true that $\hat{z}_t(1) + \hat{z}_t(2) + \hat{z}_t(3)$ is the minimum mean square error forecast of the sales $z_{t+1} + z_{t+2} + z_{t+3}$ during the next quarter.

4. *The Shocks as One-Step-Ahead Forecast Errors*. Using (5.1.14), the one-step-ahead forecast error is

$$e_t(1) = z_{t+1} - \hat{z}_t(1) = a_{t+1} \tag{5.1.17}$$

Hence, the shocks a_t, which generate the process, and which have been introduced so far merely as a set of independent random variables or shocks, turn out to be the *one-step-ahead forecast errors*.

It follows that for a minimum mean square error forecast, the one-step-ahead forecast errors must be uncorrelated. This makes sense, for if the one-step-ahead errors were correlated, the forecast error a_{t+1} could, to some extent, be predicted from available forecast errors $a_t, a_{t-1}, a_{t-2}, \ldots$. If the prediction so obtained was $\hat{a}_{t+1}$, then $\hat{z}_t(1) + \hat{a}_{t+1}$ would be a better forecast of z_{t+1} than was $\hat{z}_t(1)$.

5. *Correlation between the Forecast Errors*. Although the optimal forecast errors at lead time 1 will be uncorrelated, the forecast errors for longer lead times in general will be correlated. In Section A5.1.1, we derive a general expression for the correlation between the forecast errors $e_t(l)$ and $e_{t-j}(l)$, made at the *same* lead time l from *different* origins t and $t - j$.

Now, it is also true that forecast errors $e_t(l)$ and $e_t(l + j)$, made at different lead times from the same origin t, are correlated. One consequence of this is that there will often be a tendency for the forecast function to lie either wholly above or below the values of the series, when they eventually come to hand. In Section A5.1.2, we give a general expression for the correlation between the forecast errors $e_t(l)$ and $e_t(l + j)$, made from the *same* origin.

5.1.2 Three Basic Forms for the Forecast

We have seen that the minimum mean square error forecast $\hat{z}_t(l)$ for lead time l is the conditional expectation $E_t[z_{t+l}]$, of z_{t+l}, at origin t. Using this fact, we can write expressions for the forecast in any one of three different ways, corresponding to the three ways of

expressing the model summarized earlier in this section. To simplify the notation, we will temporarily adopt the convention that square brackets imply that the conditional expectation, at time t, is to be taken. Thus,

$$[a_{t+l}] = E_t[a_{t+l}] \qquad [z_{t+l}] = E_t[z_{t+l}]$$

For $l > 0$, the following are three different ways of expressing the forecasts:

1. *Forecasts from Difference Equation.* Taking conditional expectations at time t in (5.1.2), we obtain

$$\begin{aligned}[z_{t+l}] = \hat{z}_t(l) = \varphi_1[z_{t+l-1}] + \cdots + \varphi_{p+d}[z_{t+l-p-d}] - \theta_1[a_{t+l-1}] \\ - \cdots - \theta_q[a_{t+l-q}] + [a_{t+l}]\end{aligned} \tag{5.1.18}$$

2. *Forecasts in Integrated Form.* Use of (5.1.3) gives

$$\begin{aligned}[z_{t+l}] = \hat{z}_t(l) = [a_{t+l}] + \psi_1[a_{t+l-1}] + \cdots + \psi_{l-1}[a_{t+l}] \\ + \psi_l[a_t] + \psi_{l+1}[a_{t-1}] + \cdots\end{aligned} \tag{5.1.19}$$

yielding the form (5.1.13) discussed above. Alternatively, using the truncated form of the model (5.1.5), we have

$$\begin{aligned}[z_{t+l}] = \hat{z}_t(l) &= [a_{t+l}] + \psi_1[a_{t+l-1}] + \cdots \\ &\quad + \psi_{t+l-k-1}[a_{k+1}] + C_k(t + l - k) \\ &= [a_{t+l}] + \psi_1[a_{t+l-1}] + \cdots + \psi_{l-1}[a_{t+1}] + C_t(l)\end{aligned} \tag{5.1.20}$$

where $C_t(l)$ is the complementary function at origin t.

3. *Forecasts as a Weighted Average of Previous Observations and Forecasts Made at Previous Lead Times from the Same Origin.* Finally, taking conditional expectations in (5.1.7) yields

$$[z_{t+l}] = \hat{z}_t(l) = \sum_{j=1}^{\infty} \pi_j[z_{t+l-j}] + [a_{t+l}] \tag{5.1.21}$$

It is to be noted that the minimum mean square error forecast is defined in terms of the conditional expectation

$$[z_{t+l}] = E_t[z_{t+l}] = E[z_{t+l}|z_t, z_{t-1}, \ldots]$$

which theoretically requires knowledge of the z's stretching back into the infinite past. However, the requirement of invertibility imposed on the ARIMA model ensures that the π weights in (5.1.21) form a convergent series. Hence, for the computation of a forecast, the dependence on z_{t-j} for $j > k$ can typically be ignored. In practice, the π weights usually decay rather quickly, so whatever form of the model is employed, only a moderate length of series $z_t, z_{t-1}, \ldots, z_{t-k}$ is needed to calculate the forecasts to sufficient accuracy. The methods we discuss are easily modified to calculate the exact finite sample forecasts, $E[z_{t+l}|z_t, z_{t-1}, \ldots, z_1]$, based on the finite length of data $z_t, z_{t-1}, \ldots, z_1$.

To calculate the conditional expectations in expressions (5.1.18–5.1.21), we note that if j is a nonnegative integer,

$$\begin{aligned}
[z_{t-j}] &= E_t[z_{t-j}] = z_{t-j} && j = 0, 1, 2, \ldots \\
[z_{t+j}] &= E_t[z_{t+j}] = \hat{z}_t(j) && j = 1, 2, \ldots \\
[a_{t-j}] &= E_t[a_{t-j}] = a_{t-j} = z_{t-j} - \hat{z}_{t-j-1}(1) && j = 0, 1, 2, \ldots \\
[a_{t+j}] &= E_t[a_{t+j}] = 0 && j = 1, 2, \ldots
\end{aligned} \tag{5.1.22}$$

Therefore, to obtain the forecast $\hat{z}_t(l)$, one writes down the model for z_{t+l} in any one of the three explicit forms above and treats the terms on the right according to the following rules:

1. The $z_{t-j}(j = 0, 1, 2, \ldots)$, which have already occurred at origin t, are left unchanged.
2. The $z_{t+j}(j = 1, 2, \ldots)$, which have not yet occurred, are replaced by their forecasts $\hat{z}_t(j)$ at origin t.
3. The $a_{t-j}(j = 0, 1, 2, \ldots)$, which have occurred, are available from $z_{t-j} - \hat{z}_{t-j-1}(1)$.
4. The $a_{t+j}(j = 1, 2, \ldots)$, which have not yet occurred, are replaced by zeros.

For routine calculation, it is easiest to work directly with the difference equation form (5.1.18). Hence, the forecasts for $l = 1, 2, \ldots$ are calculated recursively as

$$\hat{z}_t(l) = \sum_{j=1}^{p+d} \varphi_j \hat{z}_t(l-j) - \sum_{j=l}^{q} \theta_j a_{t+l-j}$$

where $\hat{z}_t(-j) = [z_{t-j}]$ denotes the observed value z_{t-j} for $j \geq 0$, and the moving average terms are not present for lead times $l > q$.

Example: Forecasting Using the Difference Equation Form. We will show in Chapter 7 that the viscosity data in Series C can be represented by the model

$$(1 - 0.8B)(1 - B)z_{t+1} = a_{t+1}$$

that is,

$$(1 - 1.8B + 0.8B^2)z_{t+1} = a_{t+1}$$

or

$$z_{t+l} = 1.8z_{t+l-1} - 0.8z_{t+l-2} + a_{t+l}$$

The forecasts at origin t are given by

$$\begin{aligned}
\hat{z}_t(1) &= 1.8z_t - 0.8z_{t-1} \\
\hat{z}_t(2) &= 1.8\hat{z}_t(1) - 0.8z_t \\
\hat{z}_t(l) &= 1.8\hat{z}_t(l-1) - 0.8\hat{z}_t(l-2) \qquad l = 3, 4, \ldots
\end{aligned} \tag{5.1.23}$$

yielding in a simple recursive calculation.

There are no moving average terms in this model. However, such terms produce no added difficulties. Later in this chapter, we have a series arising in a control problem, for

which the model at time $t + l$ is

$$\nabla^2 z_{t+l} = (1 - 0.9B + 0.5B^2)a_{t+l}$$

or, equivalently, $z_{t+l} = 2z_{t+l-1} - z_{t+l-2} + a_{t+l} - 0.9a_{t+l-1} + 0.5a_{t+l-2}$. Then,

$$\begin{aligned}
\hat{z}_t(1) &= 2z_t - z_{t-1} - 0.9a_t + 0.5a_{t-1} \\
\hat{z}_t(2) &= 2\hat{z}_t(1) - z_t + 0.5a_t \\
\hat{z}_t(l) &= 2\hat{z}_t(l-1) - \hat{z}_t(l-2) \qquad l = 3, 4, \ldots
\end{aligned}$$

In these expressions, we remember that $a_t = z_t - \hat{z}_{t-1}(1)$, $a_{t-1} = z_{t-1} - \hat{z}_{t-2}(1)$, and the forecasting process may be started off initially by setting unknown a values equal to their unconditional expected values of zero. Thus, assuming by convention that data are available starting from time $s = 1$, the necessary a_s's are computed recursively from the difference equation form (5.1.2) of the model:

$$a_s = z_s - \hat{z}_{s-1}(1) = z_s - \left(\sum_{j=1}^{p+d} \varphi_j z_{s-j} - \sum_{j=1}^{q} \theta_j a_{s-j} \right) \qquad s = p + d + 1, \ldots, t$$

setting initial a_s's equal to zero, for $s < p + d + 1$. Alternatively, it is possible to estimate the necessary initial a_s's, as well as the initial z_s's, using *back-forecasting*. This technique, which essentially determines the conditional expectations of the presample a_s's and z_s's, given the available data, is discussed in Chapter 7 with regard to parameter estimation of ARIMA models. However, provided that a sufficient length of data series $z_t, z_{t-1}, \ldots, z_1$ is available, the two different treatments of the initial values will have a negligible effect on the forecasts $\hat{z}_1(l)$.

5.2 CALCULATING FORECASTS AND PROBABILITY LIMITS

5.2.1 Calculation of ψ Weights

It is often the case that forecasts are needed for several lead times $1, 2, \ldots, L$. As already shown, the difference equation form of the model allows the forecasts to be generated recursively in the order $\hat{z}_t(1)$, $\hat{z}_t(2)$, $\hat{z}_t(3)$, and so on. To obtain probability limits for these forecasts, it is necessary to calculate the weights $\psi_1, \psi_2, \ldots, \psi_{L-1}$. This is accomplished using the relation

$$\varphi(B)\psi B = \theta(B) \tag{5.2.1}$$

that is, by equating coefficients of powers of B in

$$\begin{aligned}
(1 - \varphi_1 B - \cdots - \varphi_{p+d} B^{p+d})\ (1 + \psi_1 B + \psi_2 B^2 + \cdots) \\
= (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)
\end{aligned} \tag{5.2.2}$$

Knowing the values of the φ's and the θ's, the values of ψ may be obtained as follows:

$$\begin{aligned}\psi_1 &= \varphi_1 - \theta_1 \\ \psi_2 &= \varphi_1\psi_1 + \varphi_2 - \theta_2 \\ &\vdots \\ \psi_j &= \varphi_j\psi_{j-1} + \cdots + \varphi_{p+d}\psi_{j-p-d} - \theta_j\end{aligned} \tag{5.2.3}$$

where $\psi_0 = 1$, $\psi_j = 0$ for $j < 0$, and $\theta_j = 0$ for $j > q$. If K is the greater of the integers $p + d - 1$ and q, then for $j > K$ the ψ's satisfy the difference equation:

$$\psi_j = \varphi_1\psi_{j-1} + \varphi_2\psi_{j-2} + \cdots + \varphi_{p+d}\psi_{j-p-d} \tag{5.2.4}$$

Thus, the ψ's are easily calculated recursively. For example, for the model $(1 - 1.8B + 0.8B^2)z_t = a_t$, appropriate to Series C, we have

$$(1 - 1.8B + 0.8B^2)(1 + \psi_1 B + \psi_2 B^2 + \cdots) = 1$$

Hence, with $\varphi_1 = 1.8$ and $\varphi_2 = -0.8$, we obtain

$$\begin{aligned}\psi_0 &= 1 \\ \psi_1 &= 1.8 \\ \psi_j &= 1.8\psi_{j-1} - 0.8\psi_{j-2} \qquad j = 2, 3, 4, \ldots\end{aligned}$$

so that $\psi_2 = (1.8 \times 1.8) - (0.8 \times 1.0) = 2.44$ and $\psi_3 = (1.8 \times 2.44) - (0.8 \times 1.8) = 2.95$, and so on.

Before proceeding to discuss the probability limits, we briefly mention the use of the ψ weights for updating of forecasts as new data become available.

5.2.2 Use of the ψ Weights in Updating the Forecasts

Using (5.1.13), we can express the forecasts $\hat{z}_{t+1}(l)$ and $\hat{z}_t(l + 1)$ of the future observation z_{t+l+1} made at origins $t + 1$ and t as

$$\begin{aligned}\hat{z}_{t+1}(l) &= \psi_l a_{t+1} + \psi_{l+1}a_t + \psi_{l+2}a_{t-1} + \cdots \\ \hat{z}_t(l + 1) &= \psi_{l+1}a_t + \psi_{l+2}a_{t-1} + \cdots\end{aligned}$$

On subtraction, it follows that

$$\hat{z}_{t+1}(l) = \hat{z}_t(l + 1) + \psi_l a_{t+1} \tag{5.2.5}$$

Explicitly, the t-origin forecast of z_{t+l+1} can be updated to become the $t + 1$ origin forecast of the same z_{t+l+1}, by adding a constant multiple of the one-step-ahead forecast error $a_{t+l} \equiv z_{t+l} - \hat{z}_t(l)$ with multiplier ψ_l.

This leads to a rather remarkable conclusion. Suppose that we currently have forecasts at origin t for lead times $1, 2, \ldots, L$. Then, as soon as z_{t+l} becomes available, we can calculate $a_{t+1} \equiv z_{t+l} - \hat{z}_t(1)$ and proportionally update to obtain forecasts $\hat{z}_{t+1}(l) = \hat{z}_t(l + 1) + \psi_l a_{t+1}$ at origin $t + 1$, for lead times $1, 2, \ldots, L - 1$. The new forecast $\hat{z}_{t+1}(L)$, for lead time L, cannot be calculated by this means but is easily obtained from the forecasts at shorter lead times, using the difference equation.

TABLE 5.1 Variance Function for Series C

l	1	2	3	4	5	6	7	8	9	10
$V(l)/\sigma_a^2$	1.00	4.24	10.19	18.96	30.24	43.86	59.46	76.79	95.52	115.41

5.2.3 Calculation of the Probability Limits at Different Lead Times

The expression (5.1.16) shows that the variance of the l-steps-ahead forecast error for any origin t is the expected value of

$$e_t^2(l) = [z_{t+l} - \hat{z}_t(l)]^2$$

and is given by

$$V(l) = \left(1 + \sum_{j=1}^{l-1} \psi_j^2\right) \sigma_a^2$$

For example, using the ψ weights calculated above, the function $V(l)/\sigma_a^2$ for Series C is shown in Table 5.1.

Assuming that the a's are normal, it follows that given information up to time t, the conditional probability distribution $p(z_{t+l}|z_t, z_{t-1}, \ldots)$ of a future value z_{t+l} of the process will be normal with mean $\hat{z}_t(l)$ and standard deviation

$$\sigma(l) = \left(1 + \sum_{j=1}^{l-1} \psi_j^2\right)^{1/2} \sigma_a$$

Thus, the variate $(z_{t+l} - \hat{z}_t(l))/\sigma(l)$ will have a unit normal distribution and so $\hat{z}_t(l) \pm u_{\varepsilon/2}\sigma(l)$ provides limits of an interval such that z_{t+l} will lie within the interval with probability $1 - \varepsilon$, where $u_{\varepsilon/2}$ is the deviate exceeded by a proportion $\varepsilon/2$ of the unit normal distribution. Figure 5.1 shows the conditional probability distributions of future values z_{21}, z_{22}, z_{23} for Series C, given information up to origin $t = 20$.

We show in Chapter 7 how an estimate s_a^2, of the variance σ_a^2, may be obtained from time series data. When the number of observations on which this estimate is based is, say, at least 50, s_a may be substituted for σ_a and approximate $1 - \varepsilon$ probability limits $z_{t+l}(-)$ and $z_{t+l}(+)$ for z_{t+l} will be given by

$$z_{t+l}(\pm) = \hat{z}_t(l) \pm u_{\varepsilon/2} \left(1 + \sum_{j=1}^{l-1} \psi_j^2\right)^{1/2} s_a \tag{5.2.6}$$

It follows from Table 7.6 that for Series C, $s_a = 0.134$; hence, the 50 and 95% limits, for z_{t+2}, for example, are given by

$$\text{50\% limits} : \hat{z}_t(2) \pm (0.674)(1 + 1.8^2)^{1/2}(0.134) = \hat{z}_t(2) \pm 0.19$$
$$\text{95\% limits} : \hat{z}_t(2) \pm (1.960)(1 + 1.8^2)^{1/2}(0.134) = \hat{z}_t(2) \pm 0.55$$

Figure 5.2 shows a section of Series C together with the several-steps-ahead forecasts (indicated by crosses) from origins $t = 20$ and $t = 67$. Also shown are the 50 and 95% probability limits for z_{20+l}, for $l = 1$ to 14. The interpretation of the limits $z_{t+l}(-)$ and

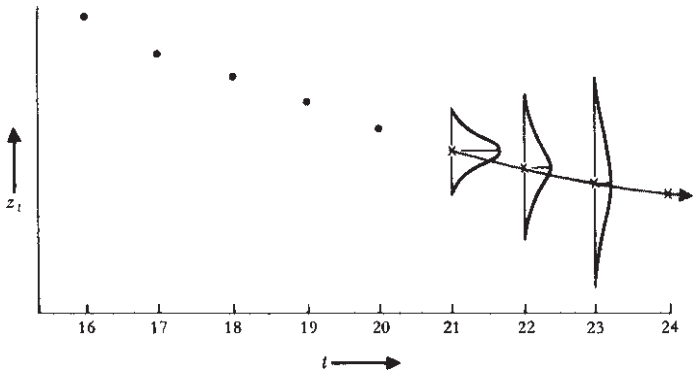


FIGURE 5.1 Conditional probability distributions of future values z_{21}, z_{22}, and z_{23} for Series C, given information up to origin $t = 20$.

$z_{t+l}(+)$ should be noted carefully. These limits are such that *given the information available at origin t*, there is a probability of $1 - \varepsilon$ that the actual value z_{t+l}, when it occurs, will be within them, that is,

$$\Pr\{z_{t+l}(-) < z_{t+l} < z_{t+l}(+)\} = 1 - \varepsilon$$

Also, the probabilities quoted apply to *individual* forecasts and not jointly to the forecasts at different lead times. For example, it is true that with 95% probability, the limits for lead time 10 will include the value z_{t+10} when it occurs. It is not true that the series can be expected to remain within *all* the limits simultaneously with this probability.

5.2.4 Calculation of Forecasts Using R

Forecasts of future values of a time series that follows an ARIMA(p, d, q) can be calculated using R. A convenient option is to use the function sarima.for() in the astsa package. For example, if z represents the observed time series, the command

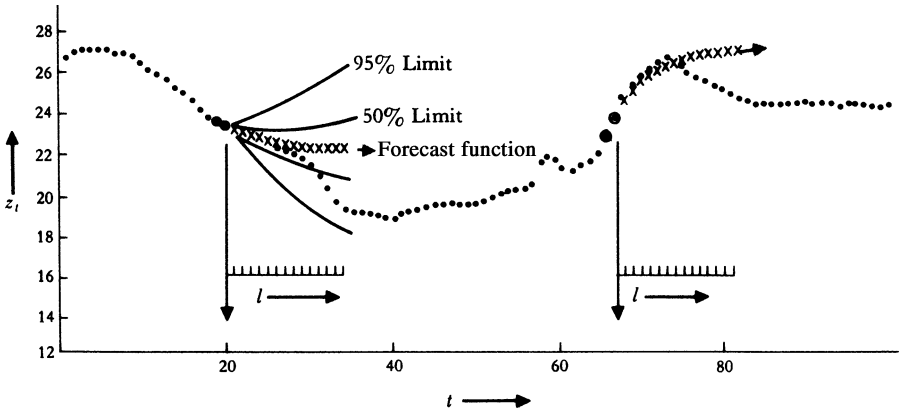


FIGURE 5.2 Forecasts for Series C and probability limits.

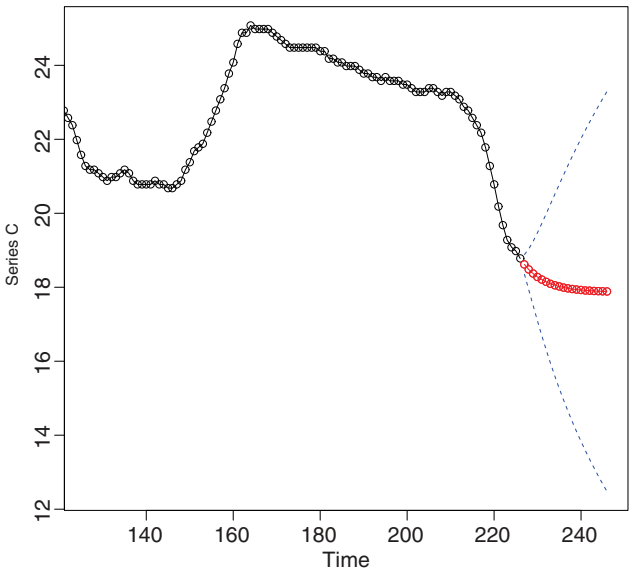


FIGURE 5.3 Forecasts for Series C with ±2 prediction error limits generated using R.

sarima.for(z,n.ahead,p,d,q,no.constant=TRUE) will fit the ARIMA(p, d, q) model without a constant term to the series and generate forecasts from the fitted model. The argument n.ahead specifies the number of forecasts to be generated. The output gives the forecasts and the standard errors of the forecasts, and supplies a graph of the forecasts along with their $+/-$ 2 prediction error limits. Thus, forecasts up to 20 steps ahead for Series C based on the ARIMA(1, 1, 0) model $(1 - \phi B)(1 - B) = a_t$ are generated as follows:

```
> library(astsa)
> seriesC=read.table("SeriesC.txt,header=TRUE)
> m1=sarima.for(seriesC,20,1,1,0,no.constant=FALSE)
> m1   % prints output from file m1
```

This code generates an output file ‘‘m1’’ that includes the forecasts (‘‘pred’’) and the prediction errors (‘‘se’’) of the forecasts. These can be accessed as m1$pred and m1$se, if needed for further analysis. Figure 5.3 shows a graph of the forecasts and their associated ±2 prediction error limits for Series C. We note that the limits become wider as the lead time increases, reflecting the increased uncertainty due to the fact that the series is nonstationary and does not vary around a fixed mean level.

5.3 FORECAST FUNCTION AND FORECAST WEIGHTS

Forecasts are calculated most simply by direct use of the difference equation. From the purely *computational* standpoint, the other model forms are less convenient. However, from the point of view of studying the nature of the forecasts, it is useful to consider in greater detail the alternative forms discussed in Section 5.1.2 and, in particular, to consider the explicit form of the forecast function.

5.3.1 Eventual Forecast Function Determined by the Autoregressive Operator

At time $t + l$, the ARIMA model may be written as

$$z_{t+l} - \varphi_1 z_{t+l-1} - \cdots - \varphi_{p+d} z_{t+l-p-d} = a_{t+l} - \theta_1 a_{t+l-1} - \cdots - \theta_q a_{t+l-q} \tag{5.3.1}$$

Taking the conditional expectations at time t, we have, for $l > q$,

$$\hat{z}_t(l) - \varphi_1 \hat{z}_t(l-1) - \cdots - \varphi_{p+d} \hat{z}_t(l-p-d) = 0 \qquad l > q \tag{5.3.2}$$

where it is understood that $\hat{z}_t(-j) = z_{t-j}$ for $j \geq 0$. This difference equation has the solution

$$\hat{z}_t(l) = b_0^{(t)} f_0(l) + b_1^{(t)} f_1(l) + \cdots + b_{p+d-1}^{(t)} f_{p+d-1}(l) \tag{5.3.3}$$

for $l > q - p - d$. Note that the forecast $\hat{z}_t(l)$ is the complementary function introduced in Chapter 4. In (5.3.3), $f_0(l), f_1(l), \ldots, f_{p+d-1}(l)$ are functions of the lead time l. In general, they could include polynomials, exponentials, sines and cosines, and products of these functions. The functions $f_0(l), f_1(l), \ldots, f_{p+d-1}(l)$ consist of d polynomial terms l^i, $i = 0, \ldots, d-1$, of degree $d-1$, associated with the nonstationary operator $\nabla^d = (1-B)^d$, and p damped exponential and damped sinusoidal terms of the form G^l and $D^l \sin(2\pi f l + F)$, respectively, associated with the roots of $\phi(B) = 0$ for the stationary autoregressive operator. That is, the forecast function has the form

$$\hat{z}_t(l) = b_0^{(t)} + b_1^{(t)} l + \cdots + b_{d-1}^{(t)} l^{d-1} + b_d^{(t)} f_d(l) + b_{d+1}^{(t)} f_{d+1}(l) + \cdots + b_{p+d-1}^{(t)} f_{p+d-1}(l)$$

For instance, if $\phi(B) = 0$ has p distinct real roots $G_1^{-1}, \ldots, G_p^{-1}$, then the last p terms in $\hat{z}_t(l)$ are $b_d^{(t)} G_1^l + b_{d+1}^{(t)} G_2^l + \cdots + b_{p+d-1}^{(t)} G_p^l$. Since the operator $\phi(B)$ is stationary, we have $|G| < 1$ and $D < 1$ and the last p terms in $\hat{z}_t(l)$ are transient and decay to zero as l increases. Hence, the forecast function is dominated by the remaining polynomial terms, $\sum_{i=0}^{d-1} b_i^{(t)} l^i$, as l increases. For a *given origin* t, the coefficients $b_j^{(t)}$ are constants applying to all lead times l, but they change from one origin to the next, *adapting* themselves appropriately to the particular part of the series being considered. From now on we call the function defined by (5.3.3) the *eventual forecast function*; ''eventual'' because when it occasionally happens that $q > p + d$, it supplies the forecasts only for lead times $l > q - p - d$.

We see from (5.3.2) that it is the general autoregressive operator $\varphi(B)$ that determines the mathematical form of the forecast function, that is, the nature of the f's in (5.3.3). Specifically, it determines whether the forecast function is to be a polynomial, a mixture of sines and cosines, a mixture of exponentials, or a combination of these functions.

5.3.2 Role of the Moving Average Operator in Fixing the Initial Values

While the autoregressive operator determines the nature of the eventual forecast function, the moving average operator is influential in determining how that function is to be ''fitted'' to the data and hence how the coefficients $b_0^{(t)}, b_1^{(t)}, \ldots, b_{p+d-1}^{(t)}$ in (5.3.3) are to be calculated and updated.

For example, consider the IMA(0, 2, 3) process:

$$z_{t+l} - 2z_{t+l-1} + z_{t+l-2} = a_{t+l} - \theta_1 a_{t+l-1} - \theta_2 a_{t+l-2} - \theta_3 a_{t+l-3}$$

Taking the conditional expectation, the forecast function becomes

$$\begin{aligned}
\hat{z}_t(1) &= 2z_t - z_{t-1} - \theta_1 a_t - \theta_2 a_{t-1} - \theta_3 a_{t-2} \\
\hat{z}_t(2) &= 2\hat{z}_t(1) - z_t - \theta_2 a_t - \theta_3 a_{t-1} \\
\hat{z}_t(3) &= 2\hat{z}_t(2) - \hat{z}_t(1) - \theta_3 a_t \\
\hat{z}_t(l) &= 2\hat{z}_t(l-1) - \hat{z}_t(l-2) \qquad l > 3
\end{aligned}$$

Therefore, since $\varphi(B) = (1 - B)^2$ in this model, the eventual forecast function is the unique straight line

$$\hat{z}_t(l) = b_0^{(t)} + b_1^{(t)} l \qquad l > 1$$

which passes through $\hat{z}_t(2)$ and $\hat{z}_t(3)$ as shown in Figure 5.4. However, note that if the θ_3 term had not been included in the model, then $q - p - d = 0$, and the forecast would have been given at *all lead times* by the straight line passing through $\hat{z}_t(1)$ and $\hat{z}_t(2)$.

In general, since only one function of the form (5.3.3) can pass through $p + d$ points, the eventual forecast function is that unique curve of the form required by $\varphi(B)$, which passes through the $p + d$ ''pivotal'' values $\hat{z}_t(q), \hat{z}_t(q-1), \ldots, \hat{z}_t(q - p - d + 1)$, where $\hat{z}_t(-j) = z_{t-j}$ $(j = 0, 1, 2, \ldots)$. In the extreme case where $q = 0$, so that the model is of the purely autoregressive form $\varphi(B)z_t = a_t$, the curve passes through the points $z_t, z_{t-1}, \ldots, z_{t-p-d+1}$. Thus, the pivotal values can consist of forecasts or of actual values of the series; they are indicated in the figures by circled points.

The moving average terms help to decide the way in which we ''reach back'' into the series to fit the forecast function determined by the autoregressive operator $\varphi(B)$. Figure 5.5 illustrates the situation for the model of order (1,1,3) given by $(1 - \phi B)\nabla z_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3)a_t$. The (hypothetical) weight functions indicate the linear functional dependence of the three forecasts, $\hat{z}_t(1)$, $\hat{z}_t(2)$, and $\hat{z}_t(3)$, on $z_t, z_{t-1}, z_{t-2}, \ldots$. Since the forecast function contains $p + d = 2$ coefficients, it is uniquely determined by the forecasts $\hat{z}_t(3)$ and $\hat{z}_t(2)$, that is, by $\hat{z}_t(q)$ and $\hat{z}_t(q-1)$. We next consider how the forecast weight functions, referred to above, are determined.

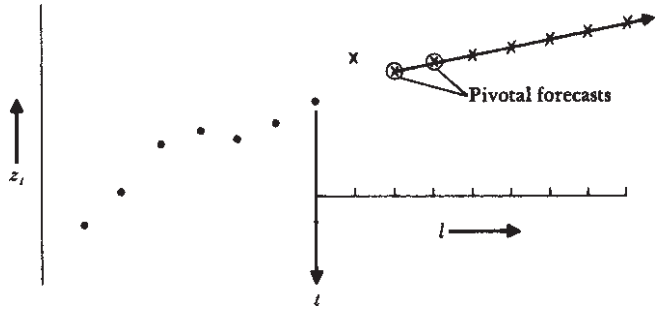


FIGURE 5.4 Eventual forecast function for an IMA(0, 2, 3) process.

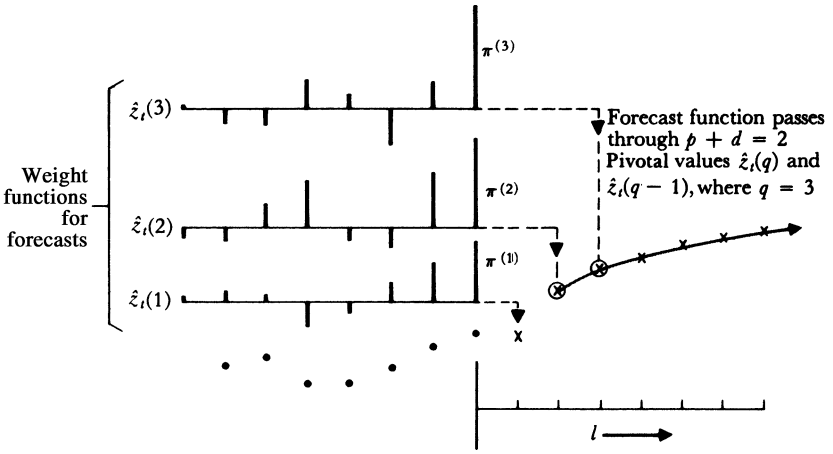


FIGURE 5.5 Dependence of forecast function on observations for a (1, 1, 3) process $(1 - \phi B)\nabla z_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3)a_t$.

5.3.3 Lead *l* Forecast Weights

The fact that the general model may also be written in inverted form,

$$a_t = \pi(B)z_t = (1 - \pi_1 B - \pi_2 B^2 - \pi_3 B^3 - \cdots)z_t \tag{5.3.4}$$

allows us to write the forecast as in (5.1.21). On substituting for the conditional expectations in (5.1.21), we obtain

$$\hat{z}_t(l) = \sum_{j=1}^{\infty} \pi_j \hat{z}_t(l-j) \tag{5.3.5}$$

where, as before, $\hat{z}_t(-h) = z_{t-h}$ for $h = 0, 1, 2, \ldots$. Thus, in general,

$$\hat{z}_t(l) = \pi_1 \hat{z}_t(l-1) + \cdots + \pi_{l-1}\hat{z}_t(1) + \pi_l z_t + \pi_{l+1} z_{t-1} + \cdots \tag{5.3.6}$$

and, in particular,

$$\hat{z}_t(1) = \pi_1 z_t + \pi_2 z_{t-1} + \pi_3 z_{t-2} + \cdots$$

The forecasts for higher lead times may also be expressed directly as linear functions of the observations $z_t, z_{t-1}, z_{t-2}, \ldots$. For example, the lead 2 forecast at origin t is

$$\begin{aligned}\hat{z}_t(2) &= \pi_1 \hat{z}_t(1) + \pi_2 z_t + \pi_3 z_{t-1} + \cdots \\ &= \pi_1 \sum_{j=1}^{\infty} \pi_j z_{t+1-j} + \sum_{j=1}^{\infty} \pi_{j+1} z_{t+1-j} \\ &= \sum_{j=1}^{\infty} \pi_j^{(2)} z_{t+1-j}\end{aligned}$$

where

$$\pi_j^{(2)} = \pi_1\pi_j + \pi_{j+1} \qquad j = 1, 2, \ldots \tag{5.3.7}$$

Proceeding in this way, it is readily shown that

$$\hat{z}_t(l) = \sum_{j=1}^{\infty} \pi_j^{(l)} z_{t+1-j} \tag{5.3.8}$$

where

$$\pi_j^{(l)} = \pi_{j+l-1} + \sum_{h=1}^{l-1} \pi_h \pi_j^{(l-h)} \qquad j = 1, 2, \ldots \tag{5.3.9}$$

and $\pi_j^{(1)} = \pi_j$. Alternative methods for computing these weights are given in Appendix A5.2.

As seen earlier, the π_j's themselves may be obtained explicitly by equating coefficients in

$$\theta(B)(1 - \pi_1 B - \pi_2 B^2 - \cdots) = \varphi(B)$$

Given these values, the $\pi_j^{(l)}$'s may readily be obtained, if so desired, using (5.3.9) or the results of Appendix A5.2. As an example, consider again the model

$$\nabla^2 z_t = (1 - 0.9B + 0.5B^2)a_t$$

which was fitted to a series, a part of which is shown in Figure 5.6. Equating coefficients in

$$(1 - 0.9B + 0.5B^2)(1 - \pi_1 B - \pi_2 B^2 - \cdots) = 1 - 2B + B^2$$

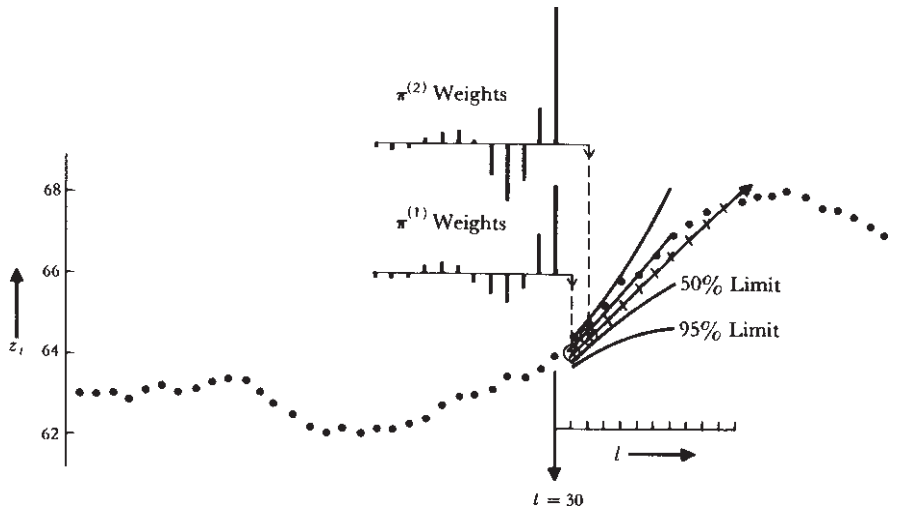


FIGURE 5.6 Part of a series fitted by $\nabla^2 z_t = (1 - 0.9B + 0.5B^2)a_t$ with forecast function for origin $t = 30$, forecast weights, and probability limits.

TABLE 5.2 π Weights for the Model $\nabla^2 z_t = (1 - 0.9B + 0.5B^2)a_t$

j	$\pi_j^{(1)}$	$\pi_j^{(2)}$
1	1.100	1.700
2	0.490	0.430
3	−0.109	−0.463
4	−0.343	−0.632
5	−0.254	−0.336
6	−0.057	0.013
7	0.076	0.181
8	0.097	0.156
9	0.049	0.050
10	−0.004	−0.032
11	−0.028	−0.054
12	−0.023	−0.026

yields the weights $\pi_j = \pi_j^{(1)}$, from which the weights $\pi_j^{(2)}$ may be computed using (5.3.7). The two sets of weights are given for $j = 1, 2, \ldots, 12$ in Table 5.2. In this example, the lead 1 and lead 2 forecasts, expressed in terms of the observations $z_t, z_{t-1}, \ldots$, are

$$\hat{z}_t(1) = 1.10z_t + 0.49z_{t-1} - 0.11z_{t-2} - 0.34z_{t-3} - 0.25z_{t-4} - \cdots$$

and

$$\hat{z}_t(2) = 1.70z_t + 0.43z_{t-1} - 0.46z_{t-2} - 0.63z_{t-3} - 0.34z_{t-4} + \cdots$$

In fact, the weights follow damped sine waves as shown in Figure 5.6.

5.4 EXAMPLES OF FORECAST FUNCTIONS AND THEIR UPDATING

The forecast functions for some special cases of the general ARIMA model will now be considered. We exhibit these in the three forms discussed in Section 5.1.2. While the forecasts are most easily computed from the difference equation itself, the other forms provide insight into the nature of the forecast function in particular cases.

5.4.1 Forecasting an IMA(0, 1, 1) Process

Difference Equation Approach. We first consider the model $\nabla z_t = (1 - \theta B)a_t$. At time $t + l$, the model may be written as

$$z_{t+l} = z_{t+l-1} + a_{t+l} - \theta a_{t+l-1}$$

Taking conditional expectations at origin t yields

$$\begin{aligned} \hat{z}_t(1) &= z_t - \theta a_t \\ \hat{z}_t(l) &= \hat{z}_t(l-1) \qquad l \geq 2 \end{aligned} \tag{5.4.1}$$

Hence, for all lead times, the forecasts at origin t will follow a straight line parallel to the time axis. Using the fact that $z_t = \hat{z}_{t-1}(1) + a_t$, we can write (5.4.1) in either of two useful forms.

The first of these is

$$\hat{z}_t(l) = \hat{z}_{t-l}(l) + \lambda a_t \tag{5.4.2}$$

where $\lambda = 1 - \theta$. This form is identical to the general updating form (5.2.5) for this model, since $\psi_l \equiv \lambda$ and $\hat{z}_{t-1}(l+1) = \hat{z}_{t-1}(l)$ for all $l \geq 1$. This form implies that having seen that our previous forecast $\hat{z}_{t-1}(l)$ falls short of the realized value by a_t, we adjust it by an amount λa_t. It will be recalled from Section 4.3.1 that λ measures the proportion of any given shock a_t, which is permanently absorbed by the ''level'' of the process. Therefore, it is reasonable to increase the forecast by that part λa_t of a_t, which we expect to be absorbed.

The second way of rewriting (5.4.1) is to write $a_t = z_t - \hat{z}_{t-1}(1) = z_t - \hat{z}_{t-1}(l)$ in (5.4.2) to obtain

$$\hat{z}_t(l) = \lambda z_t + (1 - \lambda)\hat{z}_{t-1}(l) \tag{5.4.3}$$

This form implies that the new forecast is a linear interpolation at argument λ between old forecast and new observation. Thus, if λ is very small, we rely principally on a weighted average of past data and heavily discounting the new observation z_t. By contrast, if $\lambda = 1$ ($\theta = 0$), the evidence of past data is completely ignored, $\hat{z}_t(l) = z_t$, and the forecast for all future time is the current value. With $\lambda > 1$, we induce an extrapolation rather than an interpolation between $\hat{z}_{t-1}(l)$ and z_t. The forecast error must now be *magnified* in (5.4.2) to indicate the change in the forecast.

Forecast Function in Integrated Form. The eventual forecast function is the solution of $(1 - B)\hat{z}_t(l) = 0$. Thus, $\hat{z}_t(l) = b_0^{(t)}$, and since $q - p - d = 0$, it provides the forecast for all lead times, that is,

$$\hat{z}_t(l) = b_0^{(t)} \qquad l > 0 \tag{5.4.4}$$

For *any fixed origin*, $b_0^{(t)}$ is a constant, and the forecasts for all lead times will follow a straight line parallel to the time axis. However, the coefficient $b_0^{(t)}$ will be updated as a new observation becomes available and the origin advances. Thus, the forecast function can be thought of as a polynomial of degree zero in the lead time l, with a coefficient that is adaptive with respect to the origin t.

A comparison of (5.4.4) with (5.4.1) shows that

$$b_0^{(t)} = \hat{z}_t(l) = z_t - \theta a_t$$

Equivalently, by referring to (4.3.4), since the truncated integrated form of the model, relative to an initial origin k, is

$$\begin{aligned} z_t &= \lambda S_{t-k-1} a_{t-1} + a_t + (z_k - \theta a_k) \\ &= \lambda(a_{t-1} + \cdots + a_{k+1}) + a_t + (z_k - \theta a_k) \end{aligned}$$

it follows that

$$\hat{z}_t(l) = b_0^{(t)} = \lambda S_{t-k} a_t + (z_k - \theta a_k) = \lambda(a_t + \cdots + a_{k+1}) + (z_k - \theta a_k)$$

Also, $\psi_j = \lambda\,(j = 1, 2, \ldots)$ and hence the adaptive coefficient $b_0^{(t)}$ can be updated from origin t to origin $t + 1$ according to

$$b_0^{(t+1)} = b_0^{(t)} + \lambda a_{t+1} \tag{5.4.5}$$

similar to (5.4.2).

Forecast as a Weighted Average of Previous Observations. Since, for this process, the $\pi_j^{(l)}$ weights of (5.3.8) are also the weights for the one-step-ahead forecast, we can also write, using (4.3.6),

$$\hat{z}_t(l) = b_0^{(t)} = \lambda z_t + \lambda(1-\lambda)z_{t-1} + \lambda(1-\lambda)^2 z_{t-2} + \cdots \tag{5.4.6}$$

Thus, for the IMA(0, 1, 1) model, the forecast for all future time is an *exponentially weighted moving average* of current and past z's.

Example: Forecasting Series A. It will be shown in Chapter 7 that Series A is closely fitted by the model

$$(1 - B)z_t = (1 - 0.7B)a_t$$

In Figure 5.7, the forecasts at origins $t = 39, 40, 41, 42$, and 43 and also at origin $t = 79$ are shown for lead times $1, 2, \ldots, 20$. The weights π_j, which for this model are forecast weights for any lead time, are given in Table 5.3. These weights are shown diagrammatically in their appropriate positions for the forecast $\hat{z}_{39}(l)$ in Figure 5.7.

Variance Functions. Since for this model, $\psi_j = \lambda (j = 1, 2, \ldots)$, the expression (5.1.16) for the variance of the lead l forecast errors is

$$V(l) = \sigma_a^2[1 + (l-1)\lambda^2] \tag{5.4.7}$$

Using the estimate $s_a^2 = 0.101$, appropriate for Series A, in (5.4.7). 50 and 95% probability limits were calculated and are shown in Figure 5.7 for origin $t = 79$.

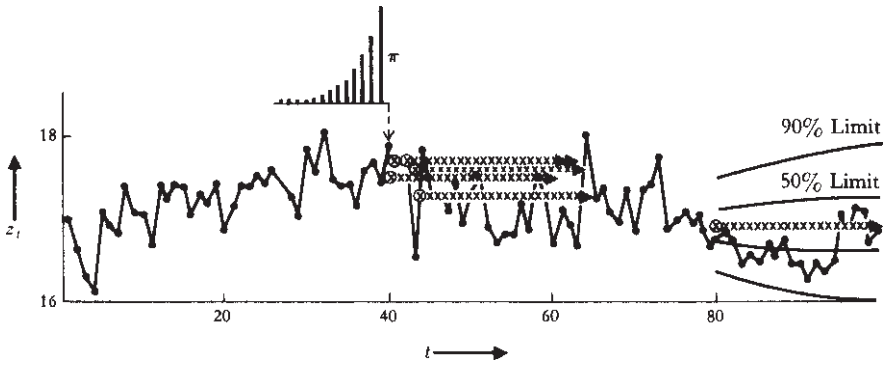


FIGURE 5.7 Part of Series A with forecasts at origins $t = 39, 40, 41, 42, 43$ and at $t = 79$.

TABLE 5.3 Forecast Weights Applied to Previous z's for Any Lead Time Used in Forecasting Series A with Model $\nabla z_t = (1 - 0.7B)a_t$

j	π_j	j	π_j
1	0.300	7	0.035
2	0.210	8	0.025
3	0.147	9	0.017
4	0.103	10	0.012
5	0.072	11	0.008
6	0.050	12	0.006

5.4.2 Forecasting an IMA(0, 2, 2) Process

Difference Equation Approach. We now consider the model $\nabla^2 z_t = (1 - \theta_1 B - \theta_2 B^2)a_t$. At time $t + l$, the model may be written as

$$z_{t+l} = 2z_{t+l-1} - z_{t+l-2} + a_{t+l} - \theta_1 a_{t+l-1} - \theta_2 a_{t+l-2}$$

On taking conditional expectations at time t, we obtain

$$\begin{aligned}
\hat{z}_t(1) &= 2z_t - z_{t-1} - \theta_1 a_t - \theta_2 a_{t-1} \\
\hat{z}_t(2) &= 2\hat{z}_t(1) - z_t - \theta_2 a_t \\
\hat{z}_t(l) &= 2\hat{z}_t(l-1) - \hat{z}_t(l-2) \qquad l \geq 3
\end{aligned}$$

from which the forecasts may be calculated. Forecasting of the series of Figure 5.6 in this way was illustrated in Section (5.1.2). An alternative way of generating the first $L - 1$ of L forecasts is via the updating formula (5.2.5),

$$\hat{z}_{t+1}(l) = \hat{z}_t(l+1) + \psi_l a_{t+1} \tag{5.4.8}$$

The truncated integrated model, as in (4.3.15), is

$$z_t = \lambda_0 S_{t-k-1} a_{t-1} + \lambda_1 S^{(2)}_{t-k-1} a_{t-1} + a_t + b_0^{(k)} + b_1^{(k)}(t-k) \tag{5.4.9}$$

where $\lambda_0 = 1 + \theta_2$ and $\lambda_1 = 1 - \theta_1 - \theta_2$, so that $\psi_j = \lambda_0 + j\lambda_1 (j = 1, 2, \ldots)$. Therefore, the updating function for this model is

$$\hat{z}_{t+1}(l) = \hat{z}_t(l+1) + (\lambda_0 + l\lambda_1)a_{t+1} \tag{5.4.10}$$

Forecast in Integrated Form. The eventual forecast function is the solution of $(1 - B)^2 \hat{z}_t(l) = 0$, that is, $\hat{z}_t(l) = b_0^{(t)} + b_1^{(t)} l$. Since $q - p - d = 0$, the eventual forecast function provides the forecast for all lead times, that is,

$$\hat{z}_t(l) = b_0^{(t)} + b_1^{(t)} l \qquad l > 0 \tag{5.4.11}$$

Thus, the forecast function is a linear function of the lead time l, with coefficients that are adaptive with respect to the origin t. The stochastic model in truncated integrated form is

$$z_{t+l} = \lambda_0 S_{t+l-k-1} a_{t+l-1} + \lambda_1 S^{(2)}_{t+l-k-1} a_{t+l-1} + a_{t+l} + b_0^{(k)} + b_1^{(k)}(t+l-k)$$

and taking expectations at origin t, we obtain

$$\begin{aligned}\hat{z}_t(l) &= \lambda_0 S_{t-k} a_t + \lambda_1 (l a_t + (l+1)a_{t-1} + \cdots + (l+t-k-1)a_{k+1}) \\ &\quad + b_0^{(k)} + b_1^{(k)}(t+l-k) \\ &= [\lambda_0 S_{t-k} a_t + \lambda_1 S^{(2)}_{t-k-1} a_{t-1} + b_0^{(k)} + b_1^{(k)}(t-k)] + (\lambda_1 S_{t-k} a_t + b_1^{(k)})l\end{aligned}$$

The adaptive coefficients may thus be identified as

$$\begin{aligned}b_0^{(t)} &= \lambda_0 S_{t-k} a_t + \lambda_1 S^{(2)}_{t-k-1} a_{t-1} + b_0^{(k)} + b_1^{(k)}(t-k) \\ b_1^{(t)} &= \lambda_1 S_{t-k} a_t + b_1^{(k)}\end{aligned} \tag{5.4.12}$$

or informally based on the infinite integrated form as $b_0^{(t)} = \lambda_0 S a_t + \lambda_1 S^2 a_{t-1}$ and $b_1^{(t)} = \lambda_1 S a_t$. Hence, their updating formulas are

$$\begin{aligned}b_0^{(t)} &= b_0^{(t-1)} + b_1^{(t-1)} + \lambda_0 a_t \\ b_1^{(t)} &= b_1^{(t-1)} + \lambda_1 a_t\end{aligned} \tag{5.4.13}$$

similar to relations (4.3.17). The additional slope term $b_1^{(t-1)}$, which occurs in the updating formula for $b_0^{(t)}$, is an adjustment to change the location parameter b_0 to a value appropriate to the new origin. It will also be noted that λ_0 and λ_1 are the fractions of the shock a_t, which are transmitted to the location parameter and the slope parameter, respectively.

Forecasts as a Weighted Average of Previous Observations. For this model, then, the forecast function is a straight line that passes through the forecasts $\hat{z}_t(1)$ and $\hat{z}_t(2)$. This is illustrated for the series in Figure 5.6, which shows the forecasts made at origin $t = 30$, with appropriate weight functions. It will be seen how dependence of the entire forecast function on previous z's in the series is a reflection of the dependence of $\hat{z}_t(1)$ and $\hat{z}_t(2)$ on these values. The weight functions for $\hat{z}_t(1)$ and $\hat{z}_t(2)$, plotted in the figure, have been given in Table 5.2.

The example illustrates once more that while the AR operator $\varphi(B)$ determines the form of function to be used (a straight line in this case), the MA operator is of importance in determining the way in which that function is ''fitted'' to previous data.

Dependence of the Adaptive Coefficients in the Forecast Function on Previous z's. Since for the general model, the values of the adaptive coefficients in the forecast function are determined by $\hat{z}_t(q), \hat{z}_t(q-1), \ldots, \hat{z}_t(q-p-d+1)$, which can be expressed as functions of the observations, it follows that the same is true for the adaptive coefficients themselves.

For instance, in the case of the model $\nabla^2 z_t = (1 - 0.9B + 0.5B^2)a_t$ of Figure 5.6,

$$\hat{z}_t(1) = b_0^{(t)} + b_1^{(t)} = \sum_{j=1}^{\infty} \pi_j^{(1)} z_{t+1-j}$$

$$\hat{z}_t(2) = b_0^{(t)} + 2b_1^{(t)} = \sum_{j=1}^{\infty} \pi_j^{(2)} z_{t+1-j}$$

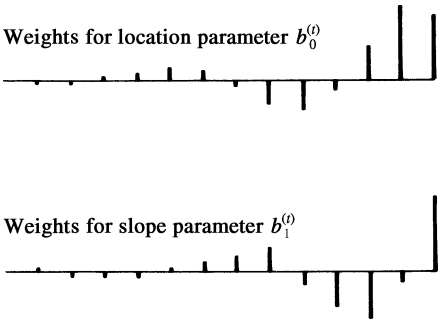


FIGURE 5.8 Weights applied to previous z's determining location and slope for the model $\nabla^2 z_t = (1 - 0.9B + 0.5B^2)a_t$.

so that

$$b_0^{(t)} = 2\hat{z}_t(1) - \hat{z}_t(2) = \sum_{j=1}^{\infty}(2\pi_j^{(1)} - \pi_j^{(2)})z_{t+1-j}$$

and

$$b_1^{(t)} = \hat{z}_t(2) - \hat{z}_t(1) = \sum_{j=1}^{\infty}(\pi_j^{(2)} - \pi_j^{(1)})z_{t+1-j}$$

These weight functions are plotted in Figure 5.8.

Variance of the Forecast Error. Using (5.1.16) and the fact that $\psi_j = \lambda_0 + j\lambda_1$, the variance of the lead l forecast error is

$$V(l) = \sigma_a^2[1 + (l-1)\lambda_0^2 + \frac{1}{6}l(l-1)(2l-1)\lambda_1^2 + \lambda_0\lambda_1 l(l-1)] \tag{5.4.14}$$

Using the estimate $s_a^2 = 0.032$, $\lambda_0 = 0.5$, and $\lambda_1 = 0.6$, the 50 and 95% limits are shown in Figure 5.6 for the forecast at origin $t = 30$.

5.4.3 Forecasting a General IMA(0, d, q) Process

As an example, consider the process of order (0, 1,3):

$$(1 - B)z_{t+l} = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3)a_{t+1}$$

Taking conditional expectations at time t, we obtain

$$\begin{aligned}
\hat{z}_t(1) - z_t &= -\theta_1 a_t - \theta_2 a_{t-1} - \theta_3 a_{t-2} \\
\hat{z}_t(2) - \hat{z}_t(1) &= -\theta_2 a_t - \theta_3 a_{t-1} \\
\hat{z}_t(3) - \hat{z}_t(2) &= -\theta_3 a_t \\
\hat{z}_t(l) - \hat{z}_t(l-1) &= 0 \qquad l = 4, 5, 6, \ldots
\end{aligned}$$

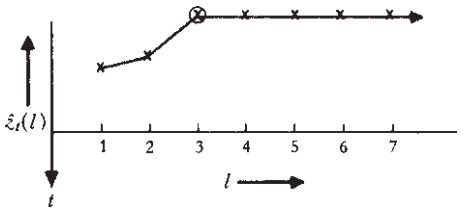


FIGURE 5.9 Forecast function for an IMA(0, 1, 3) process.

Hence, $\hat{z}_t(l) = \hat{z}_t(3) = b_0^{(t)}$ for all $l > 2$, as expected, since $q - p - d = 2$. As shown in Figure 5.9, the forecast function makes two initial ''jumps,'' depending on previous a's, before leveling out to the eventual forecast function.

For the IMA$(0, d, q)$ process, the eventual forecast function satisfies the difference equation $(1 - B)^d \hat{z}_t(l) = 0$, and has for its solution, a polynomial in l of degree $d - 1$:

$$\hat{z}_t(l) = b_0^{(t)} + b_1^{(t)} l + b_2^{(t)} l^2 + \cdots + b_{d-1}^{(t)} l^{d-1}$$

This will provide the forecasts $\hat{z}_t(l)$ for $l - q - d$. The coefficients $b_0^{(t)}, b_1^{(t)}, \ldots, b_{d-1}^{(t)}$ must be updated progressively as the origin advances. The forecast for origin t will make $q - d$ initial ''jumps,'' which depend on $a_t, a_{t-1}, \ldots, a_{t-q+1}$, and after this, will follow the polynomial above.

5.4.4 Forecasting Autoregressive Processes

Consider a process of order $(p, d, 0)$, $\varphi(B)z_t = a_t$. The eventual forecast function is the solution of $\varphi(B)\hat{z}_t(l) = 0$. It applies for all lead times and passes through the last $p + d$ available values of the series. For example, the model for the IBM stock series (Series B) is very nearly

$$(1 - B)z_t = a_t$$

so that

$$\hat{z}_t(l) \approx z_t$$

The best forecast for all future time is very nearly the current value of the stock. The weight function for $\hat{z}_t(l)$ is a spike at time t and there is no averaging over past history.

Stationary Autoregressive Models. The stationary AR(p) process $\phi(B)\tilde{z}_t = a_t$ will in general produce a forecast function that is a mixture of exponentials and damped sines. In particular, for $p = 1$, the model

$$(1 - \phi B)\tilde{z}_t = a_t \qquad -1 < \phi < 1$$

has a forecast function that, for all $l > 0$, is the solution of $(1 - \phi B)\hat{\tilde{z}}_t(l) = 0$. Thus,

$$\hat{\tilde{z}}_t(l) = b_0^{(t)} \phi^l \qquad l > 0 \tag{5.4.15}$$

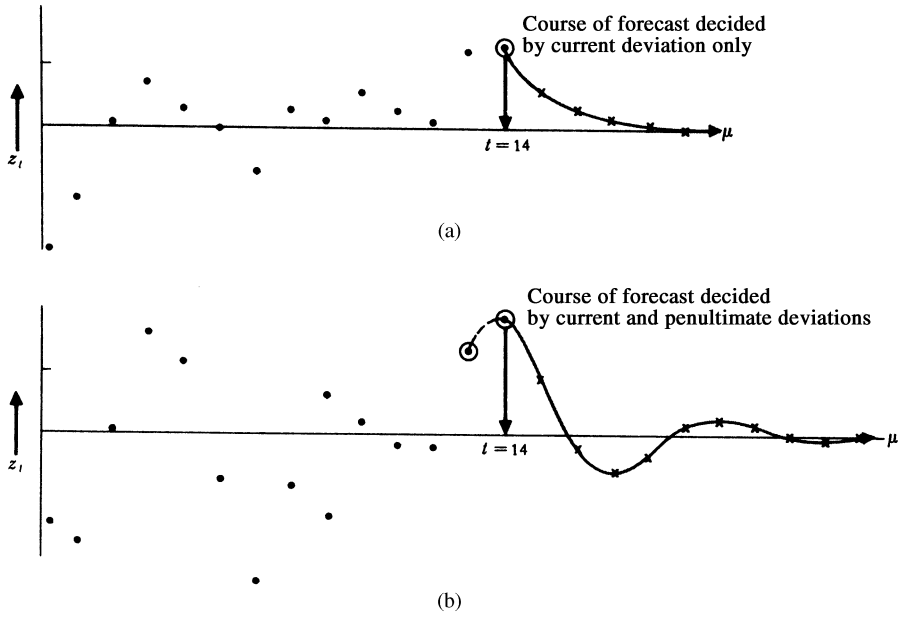


FIGURE 5.10 Forecast functions for (a) the AR(1) process $(1 - 0.5B)\tilde{z}_t = a_t$, and (b) the AR(2) process $(1 - 0.75B + 0.5B^2)\tilde{z}_t = a_t$ from a time origin $t = 14$.

Also, $\hat{\tilde{z}}_t(1) = \phi\tilde{z}_t$, so that $b_0^{(t)} = \tilde{z}_t$ and

$$\hat{\tilde{z}}_t(l) = \tilde{z}_t\phi^l$$

So, the forecasts for the original process z_t are $\hat{z}_t(l) = \mu + \phi^l(z_t - \mu)$.

Hence, the minimum mean square error forecast predicts the current deviation from the mean decaying exponentially to zero. In Figure 5.10(a) a time series is shown that is generated from the process $(1 - 0.5B)\tilde{z}_t = a_t$, with the forecast function at origin $t = 14$. The course of this function is seen to be determined entirely by the single deviation $\tilde{z}_{14}$. Similarly, the minimum mean square error forecast for a second-order autoregressive process is such that the current deviation from the mean is predicted to decay to zero via a damped sine wave or a mixture of two exponentials. Figure 5.10(b) shows a time series generated from the process $(1 - 0.75B + 0.50B^2)\tilde{z}_t = a_t$ and the forecast at origin $t = 14$. Here the course of the forecast function at origin t is determined entirely by the last two deviations, $\tilde{z}_{14}$ and $\tilde{z}_{13}$.

Variance Function for the Forecast from an AR(1) Process. Since the AR(1) process at time $t + l$ may be written as

$$\tilde{z}_{t+l} = a_{t+l} + \phi a_{t+l-1} + \cdots + \phi^{l-1}a_{t+1} + \phi^l\tilde{z}_t$$

it follows from (5.4.15) that

$$e_t(l) = \tilde{z}_{t+l} - \hat{\tilde{z}}_t(l) = a_{t+l} + \phi a_{t+l-1} + \cdots + \phi^{l-1}a_{t+1}$$

Hence,

$$V(l) = \text{var}[e_t(l)] = \sigma_a^2(1 + \phi^2 + \cdots + \phi^{2(l-1)}) = \frac{\sigma_a^2(1 - \phi^{2l})}{1 - \phi^2} \tag{5.4.16}$$

We see that for this stationary process, as l tends to infinity the variance increases to a constant value $\gamma_0 = \sigma_a^2/(1 - \phi^2)$, associated with the variation of the process about the ultimate forecast μ. This is in contrast to the behavior of forecast variance functions for nonstationary models that ''blow up'' for large lead times.

Nonstationary Autoregressive Models of Order (p, d, 0). For the model

$$\phi(B)\nabla^d z_t = a_t$$

the dth difference of the process decays back to its mean when projected several steps ahead. The mean of $\nabla^d z_t$ will usually be assumed to be zero unless contrary evidence is available. When needed, it is possible to introduce a nonzero mean by replacing $\nabla^d z_t$ by the deviation $(\nabla^d z_t - \mu_w)$ in the model. For example, consider the model

$$(1 - \phi B)(\nabla z_t - \mu_w) = a_t \tag{5.4.17}$$

After substituting $t + j$ for t and taking conditional expectations at origin t, we readily obtain [compare with (5.4.15) et seq.]

$$\hat{z}_t(j) - \hat{z}_t(j-1) - \mu_w = \phi^j(z_t - z_{t-1} - \mu_w)$$

or $\hat{w}_t(j) - \mu_w = \phi^j(w_t - \mu_w)$, where $w_t = \nabla z_t$. This shows how the forecasted *difference* decays exponentially from the initial value $w_t = z_t - z_{t-1}$ to its mean value μ_w. On summing this expression from $j = 1$ to $j = l$, that is, using $\hat{z}_t(l) = \hat{w}_t(l) + \cdots + \hat{w}_t(1) + z_t$, we obtain the forecast function

$$\hat{z}_t(l) = z_t + \mu_w l + (z_t - z_{t-1} - \mu_w)\frac{\phi(1 - \phi^l)}{1 - \phi} \qquad l \geq 1$$

that approaches asymptotically the straight line

$$f(l) = z_t + \mu_w l + (z_t - z_{t-1} - \mu_w)\frac{\phi}{1 - \phi}$$

with deterministic slope μ_w. If the forecasts are generated using the function sarima.for() in the astsa package in R, a deterministic slope can be incorporated into the forecast function by setting the argument no.constant=FALSE. The treatment of the constant term can have a big impact on the forecasts and should be considered carefully when a possible trend might be present.

We now consider the forecasting of some important mixed models.

5.4.5 Forecasting a (1, 0, 1) Process

Difference Equation Approach. Consider the stationary model

$$(1 - \phi B)\tilde{z}_t = (1 - \theta B)a_t$$

The forecasts are readily obtained from

$$\begin{aligned} \hat{\tilde{z}}_t(1) &= \phi\tilde{z}_t - \theta a_t \\ \hat{\tilde{z}}_t(l) &= \phi\hat{\tilde{z}}_t(l-1) \qquad l \geq 2 \end{aligned} \tag{5.4.18}$$

The forecasts decay geometrically to the mean, as in the first-order autoregressive process, but with a lead 1 forecast modified by a factor depending on $a_t = z_t - \hat{z}_{t-1}(1)$. The ψ weights are

$$\psi_j = (\phi - \theta)\phi^{j-1} \qquad j = 1, 2, \ldots$$

and hence, using (5.2.5), the updated forecasts for lead times $1, 2, \ldots, L-1$ could be obtained from previous forecasts for lead times $2, 3, \ldots, L$ according to

$$\hat{\tilde{z}}_{t+1}(l) = \hat{\tilde{z}}_t(l+1) + (\phi - \theta)\phi^{l-1}a_t + 1$$

Integrated Form. The eventual forecast function for all $l > 0$ is the solution of $(1 - \phi B)\hat{\tilde{z}}_t(l) = 0$, that is,

$$\hat{\tilde{z}}_t(l) = b_0^{(t)}\phi^l \qquad l > 0$$

However,

$$\hat{\tilde{z}}_t(l) = b_0^{(t)}\phi = \phi\tilde{z}_t - \theta a_t = \left[\left(1 - \frac{\theta}{\phi}\right)\tilde{z}_t + \frac{\theta}{\phi}\hat{\tilde{z}}_{t-1}(1)\right]\phi$$

Thus,

$$\hat{\tilde{z}}_t(1) = \left[\left(1 - \frac{\theta}{\phi}\right)\tilde{z}_t + \frac{\theta}{\phi}\hat{\tilde{z}}_{t-1}(1)\right]\phi^l \tag{5.4.19}$$

Hence, the forecasted deviation at lead l decays exponentially from an initial value, which is a linear interpolation between the previous lead 1 forecasted deviation and the current deviation. When ϕ is equal to unity, the forecast for all lead times becomes the familiar exponentially weighted moving average and (5.4.19) becomes equal to (5.4.3).

Weights Applied to Previous Observations. The π weights, and hence the weights applied to previous observations to obtain the lead 1 forecasts, as

$$\pi_j = (\phi - \theta)\theta^{j-1} \qquad j = 1, 2, \ldots$$

Note that the weights for this stationary process sum to $(\phi - \theta)/(1 - \theta)$ and not to unity. If ϕ were equal to 1, the process would become a nonstationary IMA(0, 1, 1) process, the weights would then sum to unity, and the behavior of the generated series would be independent of the level of z_t.

For example, Series A is later fitted to a $(1, 0, 1)$ model with $\phi = 0.9$ and $\theta = 0.6$, and hence the weights are $\pi_1 = 0.30$, $\pi_2 = 0.18$, $\pi_3 = 0.11$, $\pi_4 = 0.07$, …, which sum to

0.75. The forecasts (5.4.19) decay very slowly to the mean, and for short lead times are practically indistinguishable from the forecasts obtained from the alternative IMA(0, 1, 1) model $\nabla z_t = a_t - 0.7a_{t-1}$, for which the weights are $\pi_1 = 0.30$, $\pi_2, = 0.21$, $\pi_3 = 0.15$, $\pi_4 = 0.10$, and so on, and sum to unity. The latter model has the advantage that it does not tie the process to a fixed mean.

Variance Function. Since the ψ weights are given by

$$\psi_j = (\phi - \theta)\phi^{j-1} \qquad j = 1, 2, \ldots$$

it follows that the variance function is

$$V(l) = \sigma_a^2 \left[1 + (\phi - \theta)^2 \frac{1 - \phi^{2(l-1)}}{1 - \phi^2}\right] \tag{5.4.20}$$

which increases asymptotically to the value $\sigma_a^2(1 - 2\phi\theta + \theta^2)/(1 - \phi^2)$, the variance γ_0 of the process.

5.4.6 Forecasting a (1, 1, 1) Process

Another important mixed model is the nonstationary (1, 1, 1) process:

$$(1 - \phi B)(1 - B)z_t = (1 - \theta B)a_t$$

Difference Equation Approach. At time $t + 1$, the model may be written

$$z_{t+1} = (1 + \phi)z_{t+l-1} - \phi z_{t+l-2} + a_{t+l} - \theta a_{t+l-1}$$

On taking conditional expectations, we obtain

$$\begin{aligned} \hat{z}_t(1) &= (1 + \phi)z_t - \phi z_{t-1} - \theta a_t \\ \hat{z}_t(l) &= (1 + \phi)\hat{z}_t(l - 1) - \phi\hat{z}_t(l - 2) \qquad l > 1 \end{aligned} \tag{5.4.21}$$

Integrated Form. Since $q < p + d$, the eventual forecast function for all $l > 0$ is the solution of $(1 - \phi B)(1 - B)\hat{z}_t(l) = 0$, which is

$$\hat{z}_t(l) = b_0^{(t)} + b_1^{(t)}\phi^l$$

Substituting for $\hat{z}_t(1)$ and $\hat{z}_t(2)$ in (5.4.21), we find explicitly that

$$\begin{aligned} b_0^{(t)} &= z_t + \frac{\phi}{1 - \phi}(z_t - z_{t-1}) - \frac{\theta}{1 - \theta}a_t \\ b_1^{(t)} &= \frac{\theta a_t - \phi(z_t - z_{t-1})}{1 - \phi} \end{aligned}$$

Thus, finally,

$$\hat{z}_t(l) = z_t + \phi\frac{1-\phi^l}{1-\phi}(z_t - z_{t-1}) - \theta\frac{1-\phi^l}{1-\phi}a_t \tag{5.4.22}$$

It is evident that for large l, the forecast tends to $b_0^{(t)}$.

Weights Applied to Previous Observations. Eliminating a_t from (5.4.22), we obtain the alternative form for the forecast in terms of previous z's:

$$\hat{z}_t(l) = \left[1 - \frac{\theta-\phi}{1-\phi}(1-\phi^l)\right] z_t + \left[\frac{\theta-\phi}{1-\phi}(1-\phi^l)\right] \bar{z}_{t-1}(\theta) \tag{5.4.23}$$

where $\bar{z}_{t-1}(\theta)$ is an exponentially weighted moving average with parameter θ, that is, $\bar{z}_{t-1}(\theta) = (1-\theta)\sum_{j=1}^{\infty}\theta^{j-1}z_{t-j}$. Thus, the π weights for the process consist of a ''spike'' at time t and an EWMA starting at time $t-1$. If we refer to $(1-\alpha)x + \alpha y$ as a linear interpolation between x and y at argument α, the forecast (5.4.23) is a linear interpolation between z and $\bar{z}_{t-1}(\theta)$. The argument for lead time 1 is $\theta - \phi$, but as the lead time is increased, the argument approaches $(\theta-\phi)/(1-\phi)$. For example, when $\theta = 0.9$ and $\phi = 0.5$, the lead 1 forecast is

$$\hat{z}_t(1) = 0.6z_t + 0.4\bar{z}_{t-1}(\theta)$$

and for long lead times, the forecast approaches

$$\hat{z}_t(\infty) = 0.2z_t + 0.8\bar{z}_{t-1}(\theta)$$

5.5 USE OF STATE-SPACE MODEL FORMULATION FOR EXACT FORECASTING

5.5.1 State-Space Model Representation for the ARIMA Process

The use of state-space models for time series analysis began with the work of Kalman (1960) and many of the early developments took place in the field of engineering. These models consist of a state equation that describes the evolution of a dynamic system in time, and a measurement equation that represents the observations as linear combinations of the unobserved state variable corrupted by additive noise. In engineering applications, the state variable generally represents a well-defined set of physical variables, but these variables are not directly observable, and the state equation represents the dynamics that govern the system. In statistical applications, the state-space model is a convenient form to represent many types of models, including autoregressive–moving average (ARMA) models, structural component models of ''signal-plus-noise'' form, or time-varying parameter models. In the literature, state-space models have been used for forecasting, maximum likelihood estimation of parameters, signal extraction, seasonal adjustments, and other applications (see, for example, Durbin and Koopman, 2012). In this section, we introduce the state-space form of an ARIMA model and discuss its use in exact finite sample forecasting. Other applications involving the use of state-space models for likelihood calculations, estimation of structural components, treatment of missing values, and applications related to vector ARMA models will be discussed in Sections 7.4, 9.4, 13.3, and 14.6.

For an ARIMA(p, d, q) process $\varphi(B)z_t = \theta(B)a_t$, define the forecasts $\hat{z}_t(j) = E_t[z_t + j]$ as in Section 5.1, for $j = 0, 1, \ldots, r$, with $r = \max(p + d, q + 1)$, and $\hat{z}_t(0) = z_t$. From the updating equations (5.2.5), we have $\hat{z}_t(j-1) = \hat{z}_{t-1}(j) + \psi_{j-1}a_t$, $j = 1, 2, \ldots, r-1$. Also for $j = r > q$, recall from (5.3.2) that

$$\hat{z}_t(j-1) = \hat{z}_{t-1}(j) + \psi_{j-1}a_t = \sum_{i=1}^{p+d} \varphi_i \hat{z}_{t-1}(j-i) + \psi_{j-1}a_t$$

So we define the ''state'' vector at time t, $\boldsymbol{Y}_t$, with r components as $\boldsymbol{Y}_t = (z_t, \hat{z}_t(1), \ldots, \hat{z}_t(r-1))'$. Then from the relations above, we find that the vector $\boldsymbol{Y}_t$ satisfies the first-order system of equations:

$$\boldsymbol{Y}_t = \begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 1 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 1 \\ \varphi_r & \varphi_{r-1} & \cdot & \cdot & \cdot & \varphi_1 \end{bmatrix} \boldsymbol{Y}_{t-1} + \begin{bmatrix} 1 \\ \psi_1 \\ \cdot \\ \cdot \\ \cdot \\ \psi_{r-1} \end{bmatrix} a_t \tag{5.5.1}$$

where $\varphi_i = 0$ if $i > p + d$. So we have

$$\boldsymbol{Y}_t = \boldsymbol{\Phi}\boldsymbol{Y}_{t-1} + \boldsymbol{\Psi}a_t \tag{5.5.2}$$

together with the observation equation

$$Z_t = z_t + N_t = [1, 0, \ldots, 0]\boldsymbol{Y}_t + N_t = \mathbf{H}\boldsymbol{Y}_t + N_t \tag{5.5.3}$$

where the additional noise N_t would be present only if the process z_t is observed subject to additional white noise; otherwise, we simply have $z_t = \mathbf{H}\boldsymbol{Y}_t$. The last two equations above constitute what is known as a state-space representation of the model, which consists of a state or transition equation (5.5.2) and an observation equation (5.5.3), and $\boldsymbol{Y}_t$ is known as the state vector. We note that there are many other constructions of the state vector $\boldsymbol{Y}_t$ that will give rise to state-space equations of the general form of (5.5.2) and (5.5.3); that is, the state-space form of an ARIMA model is not unique. The two equations of the form above, in general, represent what is known as a state-space model, with unobservable state vector $\boldsymbol{Y}_t$ and observations Z_t, and can arise in time series settings more general than the context of ARIMA models.

Consider a state-space model of a slightly more general form, with state equation

$$\boldsymbol{Y}_t = \boldsymbol{\Phi}_t\boldsymbol{Y}_{t-1} + \boldsymbol{a}_t \tag{5.5.4}$$

and observation equation

$$Z_t = \mathbf{H}_t\boldsymbol{Y}_t + N_t \tag{5.5.5}$$

where it is assumed that $\boldsymbol{a}_t$ and N_t are independent white noise processes, $\boldsymbol{a}_t$ is a vector white noise process with covariance matrix $\boldsymbol{\Sigma}_a$, and N_t has variance σ_N^2. In this model, the (unobservable) state vector $\boldsymbol{Y}_t$ summarizes the state of the dynamic system through time t, and the state equation (5.5.4) describes the evolution of the dynamic system in time, while the measurement equation (5.5.5) indicates that the observations Z_t consist of linear

combinations of the state variables corrupted by additive white noise. The matrix $\mathbf{\Phi}_t$ in (5.5.4) is an $r \times r$ transition matrix and $\mathbf{H}_t$ in (5.5.5) is a $1 \times \boldsymbol{r}$ vector, which are allowed to vary with time t. Often, in applications these are constant matrices, $\mathbf{\Phi}_t \equiv \mathbf{\Phi}$ and $\mathbf{H}_t \equiv \mathbf{H}$ for all t, that do not depend on t, as in the state-space form (5.5.2) and (5.5.3) of the ARIMA model. In this case, the system or model is said to be *time invariant*. The minimal dimension r of the state vector $\boldsymbol{Y}_t$ in a state-space model needs to be sufficiently large so that the dynamics of the system can be represented by the simple Markovian (first-order) structure as in (5.5.4).

5.5.2 Kalman Filtering Relations for Use in Prediction

For the general state-space model (5.5.4) and (5.5.5), define the finite sample optimal (minimum mean square error matrix) estimate of the state vector $\boldsymbol{Y}_{t+l}$ based on observations $Z_t, \ldots, Z_1$ over the finite past time period, as

$$\hat{\boldsymbol{Y}}_{t+l|t} = E[\boldsymbol{Y}_{t+l} | Z_t, \ldots, Z_1]$$

with

$$\mathbf{V}_{t+1|t} = E[(\boldsymbol{Y}_{t+l} - \hat{\boldsymbol{Y}}_{t+l|t})(\boldsymbol{Y}_{t+l} - \hat{\boldsymbol{Y}}_{t+l|t})']$$

equal to the error covariance matrix. A convenient computational procedure, known as the *Kalman filter* equations, is then available to obtain the current estimate $\hat{\boldsymbol{Y}}_{t|t}$, in particular. It is known that, starting from some appropriate initial values $\boldsymbol{Y}_0 \equiv \hat{\boldsymbol{Y}}_{0|0}$ and $\mathbf{V}_0 \equiv \mathbf{V}_{0|0}$, the optimal filtered estimate, $\hat{\boldsymbol{Y}}_{t|t}$, is given through the following recursive relations:

$$\hat{\boldsymbol{Y}}_{t|t} = \hat{\boldsymbol{Y}}_{t|t-1} + \boldsymbol{K}_t \left(Z_t - \mathbf{H}_t \hat{\boldsymbol{Y}}_{t|t-1} \right) \tag{5.5.6}$$

where

$$\mathbf{K}_t = \mathbf{V}_{t|t-1} \mathbf{H}_t' [\mathbf{H}_t \mathbf{V}_{t|t-1} \mathbf{H}_t' + \sigma_N^2]^{-1} \tag{5.5.7}$$

with

$$\hat{\boldsymbol{Y}}_{t|t-1} = \mathbf{\Phi}_t \hat{\boldsymbol{Y}}_{t-1|t-1} \qquad \mathbf{V}_{t|t-1} = \mathbf{\Phi}_t \mathbf{V}_{t-1|t-1} \mathbf{\Phi}_t' + \mathbf{\Sigma}_a \tag{5.5.8}$$

and

$$\begin{aligned} \mathbf{V}_{t|t} &= [\mathbf{1} - \mathbf{K}_t \mathbf{H}_t] \mathbf{V}_{t|t-1} \\ &= \mathbf{V}_{t|t-1} - \mathbf{V}_{t|t-1} \mathbf{H}_t' [\mathbf{H}_t \mathbf{V}_{t|t-1} \mathbf{H}_t' + \sigma_N^2]^{-1} \mathbf{H}_t \mathbf{V}_{t|t-1} \end{aligned} \tag{5.5.9}$$

for t = 1, 2, ….

In (5.5.6), the quantity $a_{t|t-1} = Z_t - \mathbf{H}_t \hat{\boldsymbol{Y}}_{t|t-1} \equiv Z_t - \hat{Z}_{t|t-1}$ is called the (finite sample) innovation at time t, because it is the new information provided by the measurement Z_t that was not available from the previous observed (finite) history of the system. The factor $\mathbf{K}_t$ is called the *Kalman gain* matrix. The filtering procedure in (5.5.6) has the recursive ''prediction–correction'' or ''updating'' form, and the validity of these equations as representing the minimum mean square error predictor can readily be verified through the principles of updating. For example, verification of (5.5.6) follows from the principle,

for linear prediction, that

$$E[\boldsymbol{Y}_t|Z_t,\ldots,Z_1] = E[\boldsymbol{Y}_t|Z_t - \hat{Z}_{t|t-1}, Z_{t-1},\ldots,Z_1]$$
$$= E[\boldsymbol{Y}_t|Z_{t-1},\ldots,Z_1] + E[\boldsymbol{Y}_t|Z_t - \hat{Z}_{t|t-1}]$$

since $a_{t|t-1} = Z_t - \hat{Z}_{t|t-1}$ is independent of $Z_{t-1},\ldots,Z_1$. From (5.5.6), it is seen that the estimate of $\boldsymbol{Y}_t$ based on observations through time t equals the prediction of $\boldsymbol{Y}_t$ from observations through time $t-1$ updated by the factor $\boldsymbol{K}_t$ times the innovation $a_{t|t-1}$. Equation (5.5.7) indicates that $\boldsymbol{K}_t$ can be interpreted as the regression coefficients of $\boldsymbol{Y}_t$ on the innovation $a_{t|t-1}$, with $\text{var}[a_{t|t-1}] = \mathbf{H}_t\mathbf{V}_{t|t-1}\mathbf{H}_t' + \sigma_N^2$ and $\text{cov}[\boldsymbol{Y}_t, a_{t|t-1}] = \mathbf{V}_{t|t-1}\mathbf{H}_t'$ following directly from (5.5.5) since $a_{t|t-1} = \mathbf{H}_t(Z_t - \hat{Z}_{t|t-1}) + N_t$. Thus, the general *updating relation* is

$$\hat{\boldsymbol{Y}}_{t|t} = \hat{\boldsymbol{Y}}_{t|t-1} + \text{cov}[\boldsymbol{Y}_t, a_{t|t-1}]\{\text{var}[a_{t|t-1}]\}^{-1}a_{t|t-1}$$

where $a_{t|t-1} = Z_t - \hat{Z}_{t|t-1}$, and the relation in (5.5.9) is the usual updating of the error covariance matrix to account for the new information available from the innovation $a_{t|t-1}$, while the *prediction relations* (5.5.8) follow directly from (5.5.4).

In general, forecasts of future state values are available directly as $\hat{\boldsymbol{Y}}_{t+l|t} = \boldsymbol{\Phi}_{t+l}\hat{\boldsymbol{Y}}_{t+l-1|t}$ for $l = 1, 2, \ldots$, with the covariance matrix of the forecast errors generated recursively essentially through (5.5.8) as

$$\mathbf{V}_{t+l|t} = \boldsymbol{\Phi}_{t+l}\mathbf{V}_{t+l-1|t}\boldsymbol{\Phi}_{t+l}' + \boldsymbol{\Sigma}_a$$

Finally, forecasts of future observations, $Z_{t+l} = \mathbf{H}_{t+l}\boldsymbol{Y}_{t+l} + N_{t+l}$, are then available as $\hat{Z}_{t+l|t} = \mathbf{H}_{t+l}\hat{\boldsymbol{Y}}_{t+l|t}$ with forecast error variance

$$v_{t+l|t} = E[(Z_{t+l} - \hat{Z}_{t+l|t})^2] = \mathbf{H}_{t+l}\mathbf{V}_{t+l|t}\mathbf{H}_{t+l}' + \sigma_N^2$$

Use for Exact Forecasting in ARIMA Models. For ARIMA models, with state-space representation (5.5.2) and (5.5.3) and $Z_t = z_t = \mathbf{H}\boldsymbol{Y}_t$ with $\mathbf{H} = [1, 0, \ldots, 0]$, the Kalman filtering procedure constitutes an alternative method to obtain exact finite sample forecasts, based on data $z_t, z_{t-1}, \ldots, z_1$, for future values in the ARIMA process, subject to specification of appropriate initial conditions to use in (5.5.6) to (5.5.9). For stationary zero-mean processes z_t, the appropriate initial values are $\hat{\boldsymbol{Y}}_{0|0} = \mathbf{0}$, a vector of zeros, and $\mathbf{V}_{0|0} = \text{cov}[\boldsymbol{Y}_0] \equiv \mathbf{V}_*$, the covariance matrix of $\boldsymbol{Y}_0$, which can easily be determined under stationarity through the definition of $\boldsymbol{Y}_t$. Specifically, since the state vector $\boldsymbol{Y}_t$ follows the stationary vector AR(1) model $\boldsymbol{Y}_t = \boldsymbol{\Phi}\boldsymbol{Y}_{t-1} + \Psi a_t$, its covariance matrix $\mathbf{V}_* = \text{cov}[\boldsymbol{Y}_t]$ satisfies $\mathbf{V}_* = \boldsymbol{\Phi}\mathbf{V}_*\boldsymbol{\Phi}' + \sigma_a^2\boldsymbol{\Psi}\boldsymbol{\Psi}'$, which can be readily solved for $\mathbf{V}_*$. For nonstationary ARIMA processes, additional assumptions need to be specified (see, for example, Ansley and Kohn (1985) and Bell and Hillmer (1987)).

The forecasts of the ARIMA process z_t are obtained recursively as indicated above, with l-step-ahead forecast $\hat{z}_{t+l|t} = \mathbf{H}\hat{\boldsymbol{Y}}_{t+l|t}$, the first element of the vector $\hat{\boldsymbol{Y}}_{t+l|t}$, where

$$\hat{\boldsymbol{Y}}_{t+l|t} = \boldsymbol{\Phi}\hat{\boldsymbol{Y}}_{t+l-1|t}$$

with forecast error variance $v_{t+l|t} = \mathbf{H}\mathbf{V}_{t+l|t}\mathbf{H}'$. The ''steady-state'' values of the Kalman filtering procedure l-step-ahead forecasts $\hat{z}_{t+l|t}$ and their forecast error variances $v_{t+l|t}$,

which are rapidly approached as t increases, will be identical to the expressions given in Sections 5.1 and 5.2, $\hat{z}_t(l)$ and $V(l) = \sigma_a^2(1 + \sum_{j=1}^{l-1} \psi_j^2)$.

In particular, for the ARIMA process in state-space form, we can obtain the exact (finite sample) one-step-ahead forecasts:

$$\hat{z}_{t|t-1} = E[z_t | z_{t-1}, \ldots, z_1] = \mathbf{H}\hat{\boldsymbol{Y}}_{t|t-1}$$

and their error variances $v_t \equiv \mathbf{H}\mathbf{V}_{t|t-1}\mathbf{H}'$, conveniently through the Kalman filtering equations (5.5.6)–(5.5.9). This can be particularly useful for evaluation of the likelihood function, based on n observations $z_1, \ldots, z_n$ from the ARIMA process, applied to the problem of maximum likelihood estimation of model parameters (see, for example, Jones (1980) and Gardner et al. (1980)). This will be discussed again in Section 7.4.

Innovations Form of State-Space Model and Steady State for Time-Invariant Models. One particular alternative form of the general state variable model, referred to as the innovations or prediction error representation, is worth noting. If we set $\boldsymbol{Y}_t^* = \hat{\boldsymbol{Y}}_{t|t-1}$ and $a_t^* = a_{t|t-1} = Z_t - \mathbf{H}_t\hat{\boldsymbol{Y}}_{t|t-1}$, then from (5.5.6) and (5.5.8) we have

$$\boldsymbol{Y}_{t+1}^* = \boldsymbol{\Phi}_{t+1}\boldsymbol{Y}_t^* + \boldsymbol{\Phi}_{t+l}\mathbf{K}_t a_t^* \equiv \boldsymbol{\Phi}_{t+1}\boldsymbol{Y}_t^* + \boldsymbol{\Psi}_t^* a_t^* \qquad \text{and} \qquad Z_t = \mathbf{H}_t\boldsymbol{Y}_t^* + a_t^*$$

which is also of the general form of a state-space model but with the same white noise process a_t^* (the one-step-ahead prediction errors) involved in both the transition and observation equations.

In the ''stationary case'' (i.e., time-invariant and stable case) of the state-space model, where $\boldsymbol{\Phi}_t \equiv \boldsymbol{\Phi}$ and $\mathbf{H}_t \equiv \mathbf{H}$ in (5.5.4) and (5.5.5) are constant matrices and $\boldsymbol{\Phi}$ has all eigenvalues less than 1 in absolute value, we can obtain the steady-state form of the innovations representation by setting $\boldsymbol{Y}_t^* = E[\boldsymbol{Y}_t | Z_{t-1}, Z_{t-2}, \ldots]$, the projection of $\boldsymbol{Y}_t$ based on the infinite past of $\{Z_t\}$. In this case, in the Kalman filter relations (5.5.7) to (5.5.9), the error covariance matrix $\mathbf{V}_{t+1|t}$ approaches the steady-state matrix $\mathbf{V} = \lim_{t\to\infty} \mathbf{V}_{t+1|t}$ as $t \to \infty$, which satisfies

$$\mathbf{V} = \boldsymbol{\Phi}\mathbf{V}\boldsymbol{\Phi}' - \boldsymbol{\Phi}\mathbf{V}\mathbf{H}'[\mathbf{H}\mathbf{V}\mathbf{H}' + \sigma_N^2]^{-1}\mathbf{H}\mathbf{V}\boldsymbol{\Phi}' + \boldsymbol{\Sigma}_a$$

Then, also, the Kalman gain matrix $\mathbf{K}_t$ in (5.5.7) approaches the steady-state matrix, $\mathbf{K}_t \to \mathbf{K}$, where $\mathbf{K} = \mathbf{V}\mathbf{H}'[\mathbf{H}\mathbf{V}\mathbf{H}' + \sigma_N^2]^{-1}$, $a_t^* = a_{t|t-1}$ tends to $a_t = Z_t - \mathbf{H}\boldsymbol{Y}_t^* \equiv Z_t - E[Z_t | Z_{t-1}, Z_{t-2}, \ldots]$, the one-step-ahead prediction errors, and $\sigma_{t|t-1}^2 = \text{var}[a_{t|t-1}] \to \sigma_a^2 = \text{var}[a_t]$, where $\sigma_a^2 = \mathbf{H}\mathbf{V}\mathbf{H}' + \sigma_N^2$, as $t \to \infty$. These steady-state filtering results for the time-invariant model case also hold under slightly weaker conditions than stability of the transition matrix $\boldsymbol{\Phi}$ (e.g., Harvey (1989), Section 3.3), such as in the nonstationary random walk plus noise model discussed in the example of Section 5.5.3. Hence, in the time-invariant situation, the state variable model can be expressed in the steady-state innovation or prediction error form as

$$\boldsymbol{Y}_{t+l}^* = \boldsymbol{\Phi}\boldsymbol{Y}_t^* + \boldsymbol{\Phi}\mathbf{K}a_t \equiv \boldsymbol{\Phi}\boldsymbol{Y}_t^* + \boldsymbol{\Psi}^* a_t \quad \text{and} \quad Z_t = \mathbf{H}\boldsymbol{Y}_t^* + a_t \tag{5.5.10}$$

In particular, for the ARIMA process $\varphi(B)z_t = \theta(B)a_t$ with no additional observation error so that $Z_t = z_t$, a prediction error form (5.5.10) of the state-space model can be given with state vector $\boldsymbol{Y}_{t+1}^* = (\hat{z}_t(1), \ldots, \hat{z}_t(r^*))'$ of dimension $r^* = \max(p + d, q)$, $\boldsymbol{\Psi}^* = (\psi_1, \ldots, \psi_{r^*})'$, and observation equation $z_t = \hat{z}_{t-1}(1) + a_t$. For example, consider the ARMA(1, 1) process $(1 - \phi B)z_t = (1 - \theta B)a_t$. In addition to the state-space form

with state equation given by (5.5.1) and $\boldsymbol{Y}_t = (z_t, \hat{z}_t(1))'$, we have the innovations form of its state-space representation simply as $\hat{z}_t(1) = \phi\hat{z}_{t-1}(1) + \psi^* a_t$ and $z_t = \hat{z}_{t-1}(1) + a_t$, or $Y^*_{t+1} = \phi Y^*_t + \psi^* a_t$ and $z_t = Y^*_t + a_t$ with the (single) state variable $Y^*_{t+1} = \hat{z}_t(1)$ and $\psi^* = \psi_1 = \phi - \theta$.

5.5.3 Smoothing Relations in the State Variable Model

Another problem of interest within the state variable model framework, particularly in applications to economics and business, is to obtain ''smoothed'' estimates of past values of the state vector $\boldsymbol{Y}_t$ given the observations $Z_1, \ldots, Z_n$ through some fixed time n. One convenient method to obtain the desired estimates, known as the *fixed-interval smoothing* algorithm, makes use of the Kalman filter estimates $\hat{\boldsymbol{Y}}_{t|t}$ obtainable through (5.5.6)–(5.5.9). The smoothing algorithm produces the minimum MSE estimator (predictor) of the state value $\boldsymbol{Y}_t$ given the observations through time n, $\hat{\boldsymbol{Y}}_{t|n} = E[\boldsymbol{Y}_t|Z_1, \ldots, Z_n]$. In general, define $\hat{\boldsymbol{Y}}_{t|T} = E[\boldsymbol{Y}_t|Z_1, \ldots, Z_T]$ and $\mathbf{V}_{t|T} = E[(\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|T})(\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|T})']$. We assume that the filtered estimates $\hat{\boldsymbol{Y}}_{t|t}$ and their error covariance matrices $\mathbf{V}_{t|t}$, for $t = 1, \ldots, n$, have already been obtained by the Kalman filter equations. Then, the optimal smoothed estimates are obtained by the (backward) recursive relations, in which the filtered estimate $\hat{\boldsymbol{Y}}_{t|t}$ is updated, as

$$\hat{\boldsymbol{Y}}_{t|n} = \hat{\boldsymbol{Y}}_{t|t} + \mathbf{A}_t(\hat{\boldsymbol{Y}}_{t+1|n} - \hat{\boldsymbol{Y}}_{t+1|t}) \tag{5.5.11}$$

where

$$\mathbf{A}_t = \mathbf{V}_{t|t}\boldsymbol{\Phi}'_{t+l|t}\mathbf{V}^{-1}_{t+1|t} \equiv \text{cov}[\boldsymbol{Y}_t, \boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t}]\{\text{cov}[\boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t}]\}^{-1} \tag{5.5.12}$$

and

$$\mathbf{V}_{t|n} = \mathbf{V}_{t|t} - \mathbf{A}_t(\mathbf{V}_{t+1|t} - \mathbf{V}_{t+1|n})\mathbf{A}'_t \tag{5.5.13}$$

The result (5.5.11) is established from the following argument. First, consider $\boldsymbol{u}_t = E[\boldsymbol{Y}_t|Z_1, \ldots, Z_t, \boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t}, N_{t+1}, \boldsymbol{a}_{t+2}, N_{t+2}, \ldots, \boldsymbol{a}_n, N_n]$. Then, because $\{\boldsymbol{\alpha}_{t+j}, j \geq 2\}$ and $\{N_{t+j},\ j \geq 1\}$ are independent of the other conditioning variables in the definition of $\boldsymbol{u}_t$ and are also independent of $\boldsymbol{Y}_t$, we have $\boldsymbol{u}_t = \hat{\boldsymbol{Y}}_{t|t} + E[\boldsymbol{Y}_t|\boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t}] = \hat{\boldsymbol{Y}}_{t|t} + \mathbf{A}_t(\boldsymbol{Y}_{t+1} - \hat{\boldsymbol{Y}}_{t+1|t})$, where $\mathbf{A}_t$ is given by (5.5.12). Thus, because the conditioning variables in $\boldsymbol{u}_t$ generate $Z_1, \ldots, Z_n$, it follows that

$$\begin{aligned}\hat{\boldsymbol{Y}}_{t|n} &= \boldsymbol{E}[\boldsymbol{Y}_t|Z_1, \ldots, Z_n] \\ &= E[\boldsymbol{u}_t|Z_1, \ldots, Z_n] = \hat{\boldsymbol{Y}}_{t|t} + \mathbf{A}_t(\hat{\boldsymbol{Y}}_{t+1|n} - \hat{\boldsymbol{Y}}_{t+1|t})\end{aligned}$$

as in (5.5.11). The relation (5.5.13) for the error covariance matrix follows from rather straightforward calculations. This derivation of the fixed-interval smoothing relations is given by Ansley and Kohn (1982).

Thus, it is seen from (5.5.11)–(5.5.13) that the optimal smoothed estimates $\hat{\boldsymbol{Y}}_{t|n}$ are obtained by first obtaining the filtered values $\hat{\boldsymbol{Y}}_{t|t}$ through the forward recursion of the Kalman filter relations, followed by the backward recursions of (5.5.11)–(5.5.13) for $t = n - 1, \ldots, 1$. This type of smoothing procedure has applications for estimation of trend and seasonal components (seasonal adjustment) in economic time series, as will be discussed in Section 9.4. When smoothed estimates $\hat{\boldsymbol{Y}}_{t|n}$ are desired only at a fixed time point (or

only at a few fixed points), for example, in relation to problems that involve the estimation of isolated missing values in a time series, then an alternative ''fixed-point'' smoothing algorithm may be useful (e.g., see Anderson and Moore (1979) or Brockwell and Davis (1991)).

Example. As a simple example of the state-space model and associated Kalman filtering and smoothing, consider a basic structural model in which an observed series Z_t is viewed as the sum of unobserved trend and noise components. To be specific, assume that the observed process can be represented as

$$Z_t = \mu_t + N_t \quad \text{where} \quad \mu_t = \mu_{t-1} + a_t$$

so that μ_t is a random walk process and N_t is an independent (white) noise process. This is a simple example of a time-invariant state-space model with $\mathbf{\Phi} = 1$ and $\mathbf{H} = 1$ in (5.5.4) and (5.5.5) and with the state vector $\boldsymbol{Y}_t = \mu_t$ representing an underlying (unobservable) ''trend or level'' process (or ''permanent'' component). For this model, application of the Kalman filter and associated smoothing algorithm can be viewed as the estimation of the underlying trend process μ_t based on the observed process Z_t. The Kalman filtering relations (5.5.6)–(5.5.9) for this basic model reduce to

$$\hat{\mu}_{t|t} = \hat{\mu}_{t-1|t-1} + K_t(Z_t - \hat{\mu}_{t-1|t-1}) = K_t Z_t + (1 - K_t)\hat{\mu}_{t-1|t-1}$$

where the gain is $K_t = V_{t|t-1}[V_t|t-1 + \sigma_N^2]^{-1}$, with

$$V_{t+1|t} = V_{t|t-1} - V_{t|t-1}[V_{t|t-1} + \sigma_N^2]^{-1}V_{t|t-1} + \sigma_a^2$$

Then $\hat{\mu}_{t|t}$ represents the current estimate of the trend component μ_t given the observations $Z_1, \ldots, Z_t$ through time t. The steady-state solution to the Kalman filter relations is obtained as $t \to \infty$ for V ($V = \lim_{t\to\infty} V_{t+1|t}$), which satisfies $V = V - V[V + \sigma_N^2]^{-1}V + \sigma_a^2$, that is, $V[V + \sigma_N^2]^{-1}V = \sigma_a^2$, and the corresponding steady-state gain is $K = V[V + \sigma_N^2]^{-1}$. In addition, the recursion (5.5.11) for the smoothed estimate of the trend component μ_t becomes

$$\begin{aligned}\hat{\mu}_{t|n} &= \hat{\mu}_{t|t} + A_t(\hat{\mu}_{t+1|n} - \hat{\mu}_{t+1|t}) \\ &= (1 - A_t)\hat{\mu}_{t|t} + A_t\hat{\mu}_{t+1|n} \qquad t = n-1, \ldots, 1\end{aligned}$$

noting that $\hat{\mu}_{t+1|t} = \hat{\mu}_{t|t}$, where $A_t = V_{t|t}V_{t+1|r}^{-1} = V_{t|t}\{V_{t|t} + \sigma_a^2\}^{-1}$ and $V_{t|t} = (1 - K_t)V_{t|t-1}$, with the recursion for the calculation of $V_{t|t-1}$ being as given above. Thus, the smoothed value is a weighted average of the filtered estimate $\hat{\mu}_{t|t}$ at time t and the smoothed estimate $\hat{\mu}_{t+1|n}$ at time $t + 1$. The steady-state form of this smoothing recursion is the same as above with a constant $A = \lim_{t\to\infty} A_t$, which can be found to equal $A = 1 - K$. Hence, the steady-state (backward) smoothing relation (5.5.11) for this example has the same form as the steady-state filter relation already mentioned; that is, they both have the form of an exponential weighted moving average (EWMA) with the same weight.

5.6 SUMMARY

The results of this chapter may be summarized as follows: Let $\bar{z}_t$ be the deviation of an observed time series from any known deterministic function of time $f(t)$. In particular, for a stationary series, $f(t)$ could be equal to μ, the mean of the series, or it could be equal to zero, so that $\bar{z}_t$ was the observed series. Then, consider the general ARIMA model

$$\phi(B)\nabla^d \bar{z}_t = \theta(B)a_t$$

or

$$\varphi(B)\bar{z}_t = \theta(B)a_t$$

Minimum Mean Square Error Forecast. Given the knowledge of the series up to some origin t, the minimum mean square error forecast $\bar{z}_t(l)(l > 0)$ of $\bar{z}_{t+1}$ is the conditional expectation

$$\hat{\bar{z}}_t(l) = [\bar{z}_{t+l}] = E[\bar{z}_{t+l}|\bar{z}_t, \bar{z}_{t-1}, \ldots]$$

Lead 1 Forecast Errors. A necessary consequence is that the lead 1 forecast errors are the generating a_t's in the model and are uncorrelated.

Calculation of the Forecasts. It is usually simplest in practice to compute the forecasts directly from the difference equation to give

$$\begin{aligned}\hat{\bar{z}}_1(l) =& \varphi_1[\bar{z}_{t+l-1}] + \cdots + \varphi_{p+d}[\bar{z}_{t+l-p-d}] + [a_{t+l}] - \theta_1[a_{t+l-1}] \\ & - \cdots - \theta_q[a_t + l - q]\end{aligned} \tag{5.6.1}$$

The conditional expectations in (5.6.1) are evaluated by inserting actual $\bar{z}$'s when these are known, forecasted $\bar{z}$'s for future values, actual a's when these are known, and zeros for future a's. The forecasting process may be initiated by approximating a's by zeros and, in practice, the appropriate form for the model and suitable estimates for the parameters are obtained by methods set out in Chapters 6–8.

Probability Limits for Forecasts. The probability limits may be obtained as follows:

1. By first calculating the ψ weights from

$$\begin{aligned}\psi_0 &= 1 \\ \psi_1 &= \varphi_1 - \theta_1 \\ \psi_2 &= \varphi_1\psi_1 + \varphi_2 - \theta_2 \\ &\vdots \\ \psi_j &= \varphi_1\psi_{j-1} + \cdots + \varphi_{p+d}\psi_{j-p-d} - \theta_j\end{aligned} \tag{5.6.2}$$

where $\theta_j = 0, j > q$.

2. For each desired level of probability ϵ, and for each lead time l, substituting in

$$\tilde{z}_{t+l}(\pm) = \hat{\tilde{z}}_t(l) \pm u_{\epsilon/2}\left(1 + \sum_{j=1}^{l-1} \psi_j^2\right)^{1/2} \sigma_a \tag{5.6.3}$$

where in practice σ_a is replaced by an estimate s_a, of the standard deviation of the white noise process a_t, and $u_{\epsilon/2}$ is the deviate exceeded by a proportion $\epsilon/2$ of the unit normal distribution.

Updating the Forecasts. When a new deviation $\tilde{z}_{t+1}$ comes to hand, the forecasts may be updated to origin $t+1$, by calculating the new forecast error $a_{t+1} = \tilde{z}_{t+1} - \tilde{z}_t(1)$ and using the difference equation (5.6.1) with $t+1$ replacing t. However, an *alternative* method is to use the forecasts $\hat{\tilde{z}}_t(1), \hat{\tilde{z}}_t(2), \ldots, \hat{\tilde{z}}(L)$ at origin t, to obtain the first $L-1$ forecasts $\hat{\tilde{z}}_{t+1}(1), \hat{\tilde{z}}_{t+1}(2), \ldots, \hat{\tilde{z}}_{t+1}(L-1)$ at origin $t+1$, from

$$\hat{\tilde{z}}_{t+1}(l) = \hat{\tilde{z}}_t(l+1) + \psi_l a_{t+1} \tag{5.6.4}$$

and then generate the last forecast $\hat{\tilde{z}}_{t+1}(L)$ using the difference equation (5.6.1).

Other Ways of Expressing the Forecasts. The above is all that is needed for *practical* utilization of the forecasts. However, the following alternative forms provide theoretical insight into the nature of the forecasts generated by different models:

1. *Forecasts in Integrated Form.* For $l > q - p - d$, the forecasts lie on the unique curve

$$\hat{\tilde{z}}_t(l) = b_0^{(t)} f_0(l) + b_1^{(t)} f_1(l) + \cdots + b_{p+d-1}^{(t)} f_{p+d-1}(l) \tag{5.6.5}$$

determined by the ''pivotal'' values $\hat{\tilde{z}}_t(q), \hat{\tilde{z}}_t(q-1), \ldots, \hat{\tilde{z}}_t(q-p-d+1)$, where $\hat{\tilde{z}}_t(-j) = \tilde{z}_{t-j}$ $(j = 0, 1, 2, \ldots)$. If $q > p + d$, the first $q - p - d$ forecasts do not lie on this curve. In general, the stationary autoregressive operator contributes damped exponential and damped sine wave terms to (5.6.5), and the nonstationary operator ∇^d contributes polynomial terms up to degree $d-1$.

The adaptive coefficients $b_j^{(t)}$ in (5.6.5) may be updated from origin t to $t+1$ by amounts depending on the last lead 1 forecast error a_{t+1}, according to the general formula

$$\mathbf{b}^{(t+1)} = \mathbf{L}'\mathbf{b}^{(t)} + \mathbf{g}a_{t+1} \tag{5.6.6}$$

given in Appendix A5.3. Specific examples of the updating are given in (5.4.5) and (5.4.13) for the IMA(0, 1, 1) and IMA(0, 2, 2) processes, respectively.

2. *Forecasts as a Weighted Sum of Past Observations.* It is instructive from a theoretical point of view to express the forecasts as a weighted sum of past observations. Thus, if the model is written in inverted form,

$$a_t = \pi(B)\tilde{z}_t = (1 - \pi_1 B - \pi_2 B^2 - \cdots)\tilde{z}_t$$

the lead 1 forecast is

$$\hat{\tilde{z}}_t(1) = \pi_1 \tilde{z}_t + \pi_2 \tilde{z}_{t-1} + \cdots \tag{5.6.7}$$

and the forecasts for longer lead times may be obtained from

$$\hat{\tilde{z}}_t(l) = \pi_1[\tilde{z}_{t+l-1}] + \pi_2[\tilde{z}_{t+l-2}] + \cdots \tag{5.6.8}$$

where the conditional expectations in (5.6.8) are evaluated by replacing $\tilde{z}$'s by actual values when known, and by forecasted values when unknown.

Alternatively, the forecast for any lead time may be written as a linear function of the available observations. Thus,

$$\hat{\tilde{z}}_t(l) = \sum_{j=1}^{\infty} \pi_j^{(l)} \tilde{z}_{t+l-j}$$

where the $\pi_j^{(l)}$ are functions of the π_j's.

Role of Constant Term in Forecasts. The forecasts will be impacted by the allowance of a nonzero constant term θ_0 in the ARIMA(p, d, q) model, $\varphi(B)z_t = \theta_0 + \theta(B)a_t$, where $\varphi(B) = \phi(B)\nabla^d$. Then, in (5.3.3) and (5.6.5), an additional deterministic polynomial term of degree d, $(\mu_w/d!)l^d$ with $w_t = \nabla^d z_t$ and $\mu_w = E[w_t] = \theta_0/(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$, will be present. This follows because in place of the relation $\varphi(B)\hat{z}_t(l) = \theta_0$ in (5.3.2), the forecasts now satisfy $\varphi(B)\hat{z}_t(l) = \theta_0$, $1 > q$, and the deterministic polynomial term of degree d represents a particular solution to this nonhomogeneous difference equation. Hence, in the instance of a nonzero constant term θ_0, the ARIMA model is also expressible as $\phi(B)(\nabla^d z_t - \mu_w) = \theta(B)a_t$, $\mu_w \neq 0$, and the forecast in the form (5.6.5) may be viewed as representing the forecast value of $\tilde{z}_{t+l} = z_{t+l} - f(t + l)$, where $f(t + l) = (\mu_w/d!)(t + l)^d + g(t + l)$ and $g(t)$ is any fixed deterministic polynomial in t of degree less than or equal to $d - 1$ (including the possibility $g(t) = 0$). For example, in an ARIMA model with $d = 1$ such as the ARIMA(1, 1, 1) model example of Section 5.4.6, but with $\theta_0 \neq 0$, the eventual forecast function of the form $\hat{z}_t(l) = b_0^{(t)} + b_1^{(t)}\phi^l$ will now contain the additional deterministic linear trend term $\mu_w l$, where $\mu_w = \theta_0/(1 - \phi)$, similar to the result in the example for the ARIMA$(1, 1, 0)$ model in (5.4.17). Note that in the special case of a stationary process z_t, with $d = 0$, the additional deterministic term in (5.3.3) reduces to the mean of the process z_t, $\mu = E[z_t]$.

APPENDIX A5.1 CORRELATION BETWEEN FORECAST ERRORS

A5.1.1 Autocorrelation Function of Forecast Errors at Different Origins

Although it is true that for an optimal forecast the forecast errors for lead time 1 will be uncorrelated, this will not generally be true of forecasts at longer lead times. Consider forecasts for lead times l, made at origins t and $t - j$, respectively, where j is a positive integer. Then, if $j = l, l + 1, l + 2, \ldots$, the forecast errors will contain no common component, but for $j = 1, 2, \ldots, l - 1$, certain of the a's will be included in both forecast errors. Specifically,

$$e_t(l) = z_{t+l} - \hat{z}_t(l) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1}a_{t+1}$$
$$e_{t-j}(l) = z_{t-j+l} - \hat{z}_{t-j}(l) = a_{t-j+l} + \psi_1 a_{t-j+l-1} + \cdots + \psi_{l-1}a_{t-j+1}$$

TABLE A5.1 Autocorrelations of Forecast Errors at Lead 6 for Series C

j	0	1	2	3	4	5	6
$\rho[e_t(6), e_{t-j(6)}]$	1.00	0.81	0.61	0.41	0.23	0.08	0.00

and for $j < l$, the lag j autocovariance of the forecast errors for lead time l is

$$E[e_t(l)e_{t-j}(l)] = \sigma_a^2 \sum_{i=j}^{l-1} \psi_i \psi_{i-j} \tag{A5.1.1}$$

where $\psi_0 = 1$. The corresponding autocorrelations are

$$\rho[e_t(l), e_{t-j}(l)] = \begin{cases} \dfrac{\sum_{i=j}^{l-1} \psi_i \psi_{i-j}}{\sum_{i=0}^{l-1} \psi_i^2} & 0 \le j \le l \\ 0 & j \ge l \end{cases} \tag{A5.1.2}$$

We show in Chapter 7 that Series C of Figure 4.1 is well fitted by the $(1, 1, 0)$ model $(1 - 0.8B)\nabla z_t = a_t$. To illustrate (A5.1.2), we calculate the autocorrelation function of the forecast errors at lead time 6 for this model. It follows from Section 5.2.1 that the ψ weights $\psi_1, \psi_2, \ldots, \psi_5$ for this model are 1.80, 2.44. 2.95, 3.36, and 3.69, respectively. Thus, for example, the lag 1 autocovariance is

$$\begin{aligned} E[e_t(6)e_{t-1}(6)] &= \sigma_a^2[(1.80 \times 1.00) + (2.44 \times 1.80) + \cdots + (3.69 \times 3.36)] \\ &= 35.70\sigma_a^2 \end{aligned}$$

On dividing by $E[e_t^2(6)] = 43.86\sigma_a^2$, we obtain $\rho[e_t(6), e_{t-1}(6)] = 0.81$. The first six autocorrelations are shown in Table A5.1 and plotted in Figure A5.l(a). As expected, the autocorrelations beyond the fifth are zero.

A5.1.2 Correlation Between Forecast Errors at the Same Origin with Different Lead Times

Suppose that we make a series of forecasts for different lead times from the *same* fixed origin t. Then, the errors for these forecasts will be correlated. We have for $j = 1, 2, 3, \ldots,$

$$\begin{aligned} e_t(l) = z_{t+1} - \hat{z}_t(l) &= a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} \\ e_t(l+j) = z_{t+l+j} - \hat{z}_t(l+j) &= a_{t+l+j} + \psi_1 a_{t+l+j-1} + \cdots + \psi_{j-1} a_{t+l+1} \\ &\quad + \psi_j a_{t+l} + \psi_{j+1} a_{t+l-1} + \cdots + \psi_{l+j-1} a_{t+1} \end{aligned}$$

so that the covariance between the t-origin forecast errors at lead times l and $l + j$ is $\sigma_a^2 \sum_{i=0}^{l-1} \psi_i \psi_{i+j}$, where $\psi_0 = 1$.

Thus, the correlation coefficient between the t-origin forecast errors at lead times l and $l + j$ is

$$\rho[e_t(l), e_t(l+j)] = \frac{\sum_{i=0}^{l-1} \psi_i \psi_{i+j}}{\left(\sum_{h=0}^{l-1} \psi_h^2 \sum_{g=0}^{l+j-1} \psi_g^2\right)^{1/2}} \tag{A5.1.3}$$

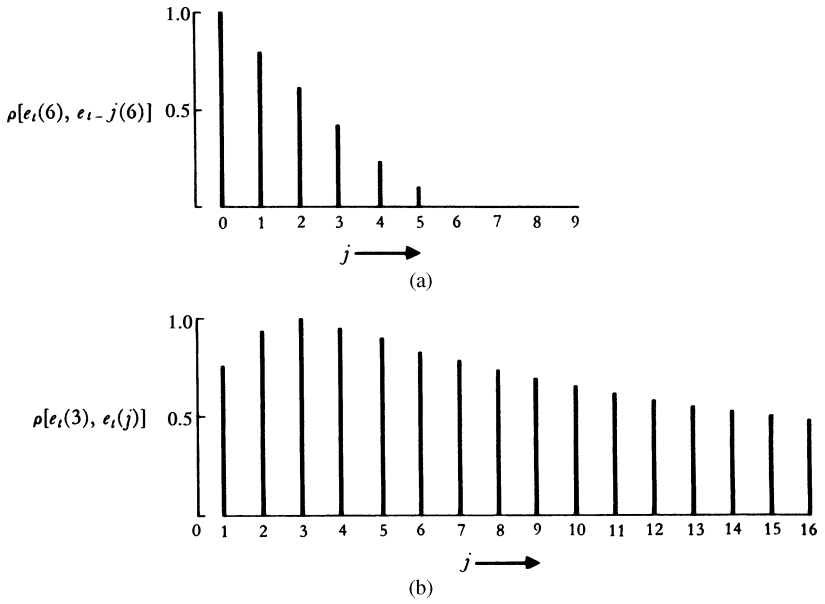


FIGURE A5.1 Correlations between various forecast errors for Series C. (a) Autocorrelations of forecast errors for Series C from different origins at lead time $l = 6$ (b) Correlations between forecast errors for Series C from the same origin at lead time 3 and lead time j.

To illustrate (A5.1.3), we compute, for forecasts made from the same origin, the correlation between the forecast error at lead time 3 and the forecast errors at lead times $j = 1, 2, 3, 4, \ldots, 16$ for Series C. For example, using (A5.1.3) and the ψ weights given in Section 5.2.2,

$$\begin{aligned} E[e_t(3)e_t(5)] &= \sigma_a^2 \sum_{i=0}^{2} \psi_i \psi_{i+2} = \sigma_a^2[\psi_0\psi_2 + \psi_1\psi_3 + \psi_2\psi_4] \\ &= \sigma_a^2[(1.00 \times 2.44) + (1.80 \times 2.95) + (2.44 \times 3.36)] \\ &= 15.94\sigma_a^2 \end{aligned}$$

The correlations for lead times $j = 1, 2, \ldots, 16$ are shown in Table A5.2 and plotted in Figure A5.1(b). As is to be expected, forecasts made from the same origin at different lead times are highly correlated.

APPENDIX A5.2 FORECAST WEIGHTS FOR ANY LEAD TIME

In this appendix we consider an alternative procedure for calculating the forecast weights $\pi_j^{(l)}$ applied to previous z's for any lead time l. To derive this result, we make use of the identity (3.1.7), namely,

$$(1 + \psi_1 B + \psi_2 B^2 + \cdots)(1 - \pi_1 B - \pi_2 B^2 - \cdots) = 1$$

from which the π weights may be obtained in terms of the ψ weights, and vice versa.

TABLE A5.2 Correlation Between Forecast Errors at Lead 3 and at Lead j Made from a Fixed Origin for Series C

j	$\rho[e_t(3), e_t(j)]$	j	$\rho[e_t(3), e_t(j)]$
1	0.76	9	0.71
2	0.94	10	0.67
3	1.00	11	0.63
4	0.96	12	0.60
5	0.91	13	0.57
6	0.85	14	0.54
7	0.80	15	0.52
8	0.75	16	0.50

On equating coefficients, we find, for $j \geq 1$,

$$\psi_j = \sum_{i=1}^{j} \pi_i \psi_{j-i} \qquad (\psi_0 = 1) \tag{A5.2.1}$$

Thus, for example,

$$\begin{aligned} \psi_1 &= \pi_1 & \pi_1 &= \psi_1 \\ \psi_2 &= \pi_1\psi_1 + \pi_2 & \pi_2 &= \psi_2 - \psi_1\pi_1 \\ \psi_3 &= \pi_1\psi_2 + \pi_2\psi_1 + \pi_3 & \pi_3 &= \psi_3 - \psi_1\pi_2 - \psi_2\pi_1 \end{aligned}$$

Now from (5.3.6),

$$\hat{z}_t(l) = \pi_1\hat{z}_t(l-1) + \pi_2\hat{z}_t(l-2) + \cdots + \pi_{l-1}\hat{z}_t(1) + \pi_l z_t + \pi_{l+1}z_{t-1} + \cdots \tag{A5.2.2}$$

Since each of the forecasts in (A5.2.1) is itself a function of the observations $z_t, z_{t-1}, z_{t-2}, \ldots$, we can write

$$\hat{z}_t(l) = \pi_1^{(l)} z_t + \pi_2^{(l)} z_{t-1} + \pi_3^{(l)} z_{t-2} + \cdots$$

where the lead l forecast weights may be calculated from the lead 1 forecast weights $\pi_j^l = \pi_j$. We now show that the weights $\pi_j^{(l)}$ can be obtained using the identity

$$\pi_j^{(l)} = \sum_{i=1}^{l} \psi_{l-i}\pi_{i+j-1} = \pi_{j+l-1} + \psi_1\pi_{j+l-2} + \cdots + \psi_{l-1}\pi_j \tag{A5.2.3}$$

For example, the weights for the forecast at lead time 3 are

$$\begin{aligned} \pi_1^{(3)} &= \pi_3 + \psi_1\pi_2 + \psi_2\pi_1 \\ \pi_2^{(3)} &= \pi_4 + \psi_1\pi_3 + \psi_2\pi_2 \\ \pi_3^{(3)} &= \pi_5 + \psi_1\pi_4 + \psi_2\pi_3 \end{aligned}$$

and so on. To derive (A5.2.3), we write

$$\hat{z}_t(l) = \psi_l a_t + \psi_{l+1} a_{t-1} + \cdots$$
$$\hat{z}_{t+l-1}(l) = \psi_1 a_{t+l-1} + \cdots + \psi_l a_t + \psi_{l+1} a_{t-1} + \cdots$$

On subtraction, we obtain

$$\hat{z}_t(l) = \hat{z}_{t+l-1}(1) - \psi_1 a_{t+l-1} - \psi_2 a_{t+l-2} - \cdots - \psi_{l-1} a_{t+1}$$

Hence,

$$\begin{aligned}\hat{z}_t(l) =& \pi_1 z_{t+l-1} + \pi_2 z_{t+l-2} + \cdots + \pi_{l-1} z_{t+1} + \pi_l z_t + \pi_{l+1} z_{t-1} + \cdots \\ &+ \psi_1(-z_{t+l-1} + \pi_1 z_{t+l-2} + \cdots + \pi_{l-2} z_{t+1} + \pi_{l-1} z_t + \pi_l z_{t-1} + \cdots) \\ &+ \psi_2(-z_{t+l-2} + \pi_1 z_{t+l-3} + \cdots + \pi_{l-3} z_{t+1} + \pi_{l-2} z_t + \pi_{l-1} z_{t-1} + \cdots) \\ &+ \cdots \\ &+ \psi_{l-1}(-z_{t+1} + \pi_1 z_t + \pi_2 z_{t-1} + \cdots)\end{aligned}$$

Using the relation (A5.2.1), each one of the coefficients of $z_{t+l-1}, \ldots, z_{t+1}$ is seen to vanish, as they should, and on collecting terms, we obtain the required result (A5.2.3). Alternatively, we may use the formula in the recursive form

$$\pi_j^{(l)} = \pi_{j+1}^{(l-1)} + \psi_{l-1}\pi_j \tag{A5.2.4}$$

Using the model $\nabla^2 z_t = (1 - 0.9B + 0.5B^2)a_t$ for illustration, we calculate the weights for lead time 2. Equation (A5.2.4) gives

$$\pi_j^{(2)} = \pi_{j+1} + \psi_1 \pi_j$$

and using the weights in Table 5.2, with $\psi_1 = 1.1$ we have, for example,

$$\pi_1^{(2)} = \pi_2 + \psi_1 \pi_1 = 0.490 + (1.1)(1.1) = 1.700$$
$$\pi_2^{(2)} = \pi_3 + \psi_1 \pi_2 = -0.109 + (1.1)(0.49) = 0.430$$

and so on. The first 12 weights have been given in Table 5.2.

APPENDIX A5.3 FORECASTING IN TERMS OF THE GENERAL INTEGRATED FORM

A5.3.1 General Method of Obtaining the Integrated Form

We emphasize once more that for practical computation of the forecasts, the difference equation procedure is by far the simplest. The following general treatment of the integrated form is given only to elaborate further on the forecasts obtained. In this treatment, rather than solving explicitly for the forecast function as we did in the examples given in Section 5.4, it will be appropriate to write down the general form of the eventual forecast function involving $p + d$ adaptive coefficients. We then show how the eventual forecast function needs to be modified to deal with the first $q - p - d$ forecasts if $q > p + d$. Finally, we show how to update the adaptive coefficients from origin t to origin $t + 1$.

If it is understood that $\hat{z}_t(-j) = z_{t-j}$ for $j = 0, 1, 2, \ldots$, then using the conditional expectation argument of Section 5.1.1, the forecasts satisfy the difference equation:

$$\begin{aligned}
&\hat{z}_t(1) - \varphi_1 \hat{z}_t(0) - \cdots - \varphi_{p+d} \hat{z}_t(1-p-d) = -\theta_1 a_t - \cdots - \theta_q a_{t-q+1} \\
&\hat{z}_t(2) - \varphi_1 \hat{z}_t(1) - \cdots - \varphi_{p+d} \hat{z}_t(2-p-d) = -\theta_2 a_t - \cdots - \theta_q a_{t-q+2} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \vdots \qquad\qquad\qquad\qquad\qquad\qquad \text{(A5.3.1)} \\
&\hat{z}_t(q) - \varphi_1 \hat{z}_t(q-1) - \cdots - \varphi_{p+d} \hat{z}_t(q-p-d) = -\theta_q a_t \\
&\hat{z}_t(l) - \varphi_1 \hat{z}_t(l-1) - \cdots - \varphi_{p+d} \hat{z}_t(l-p-d) = 0 \qquad l > q
\end{aligned}$$

The eventual forecast function is the solution of the last equation and may be written as

$$\hat{z}_t(l) = b_0^{(t)} f_0(l) + b_1^{(t)} f_1(l) + \cdots + b_{p+d-1}^{(t)} f_{p+d-1}(l) = \sum_{i=0}^{p+d-1} b_i^{(t)} f_i(l) \qquad l > q - p - d \qquad \text{(A5.3.2)}$$

When q is less than or equal to $p + d$, the eventual forecast function will provide forecasts $\hat{z}_t(1), \hat{z}_t(2), \hat{z}_t(3), \ldots$ for all lead times $l \geq 1$.

As an example of such a model with $q \leq p + d$, suppose that

$$(1 - B)(1 - \sqrt{3}B + B^2)^2 z_t = (1 - 0.5B)a_t$$

so that $p + d = 5$ and $q = 1$. Then,

$$(1 - B)(1 - \sqrt{3}B + B^2)^2 \hat{z}_t(l) = 0 \qquad l = 2, 3, 4, \ldots$$

where B now operates on l and not on t. Solution of this difference equation yields the forecast function

$$\begin{aligned}
\hat{z}_t(l) = b_0^{(t)} + b_1^{(t)} \cos\left(\frac{2\pi l}{12}\right) + b_2^{(t)} l \cos\left(\frac{2\pi l}{12}\right) \\
+ b_3^{(t)} \sin\left(\frac{2\pi l}{12}\right) + b_4^{(t)} l \sin\left(\frac{2\pi l}{12}\right) \qquad l = 1, 2, \ldots
\end{aligned}$$

If q is greater than $p + d$, then for lead times $l \leq q - p - d$, the forecast function will have additional terms containing a_{t-i}'s. Thus,

$$\hat{z}_t(l) = \sum_{i=0}^{p+d-1} b_i^{(t)} f_i(l) + \sum_{i=0}^{j} d_{li} a_{t-i} \qquad l \leq q - p - d \qquad \text{(A5.3.3)}$$

where $j = q - p - d - l$ and the d's may be obtained explicitly by substituting (A5.3.3) in (A5.3.1). For example, consider the stochastic model

$$\nabla^2 z_t = (1 - 0.8B + 0.5B^2 - 0.4B^3 + 0.1B^4)a_t$$

in which $p + d = 2, q = 4, q - p - d = 2$ and $\varphi_1 = 2, \varphi_2 = -1, \theta_1 = 0.8, \theta_2 = -0.5, \theta_3 = 0.4$, and $\theta_4 = -0.1$. Using the recurrence relation (5.2.3), we obtain $\psi_1 = 1.2, \psi_2 = 1.9$,

$\psi_3 = 2.2$, and $\psi_4 = 2.6$. Now, from (A5.3.3),

$$\begin{aligned}\hat{z}_t(1) &= b_0^{(t)} + b_1^{(t)} + d_{10}a_t + d_{11}a_{t-1}\\ \hat{z}_t(2) &= b_0^{(t)} + 2b_1^{(t)} + d_{20}a_t\\ \hat{z}_t(l) &= b_0^{(t)} + b_1^{(t)}l \qquad l > 2\end{aligned} \tag{A5.3.4}$$

Using (A5.3.1) gives

$$\hat{z}_t(4) - 2\hat{z}_t(3) + \hat{z}_t(2) = 0.1a_t$$

so that from (A5.3.4)

$$d_{20}a_t = 0.1a_t$$

and hence $d_{20} = 0.1$. Similarly, from (A5.3.1),

$$\hat{z}_t(3) - 2\hat{z}_t(2) + \hat{z}_t(l) = -0.4a_t + 0.1a_{t-1}$$

and hence using (A5.3.4),

$$-0.2a_t + d_{10}a_t + d_{11}a_{t-1} = -0.4a_t + 0.1a_{t-1}$$

yielding

$$d_{10} = -0.2 \qquad d_{11} = 0.1$$

Hence, the forecast function is

$$\begin{aligned}\hat{z}_t(1) &= b_0^{(t)} + b_1^{(t)} - 0.2a_t + 0.1a_{t-1}\\ \hat{z}_t(2) &= b_0^{(t)} + 2b_1^{(t)} + 0.1a_t\\ \hat{z}_t(l) &= b_0^{(t)} + b_1^{(t)}l \qquad l > 2\end{aligned}$$

A5.3.2 Updating the General Integrated Form

Updating formulas for the coefficients may be obtained using the identity (5.2.5) with $t + 1$ replaced by t:

$$\hat{z}_t(l) = \hat{z}_{t-1}(l + 1) + \psi_l a_t$$

Then, for $l > q - p - d$,

$$\sum_{i=0}^{p+d-1} b_i^{(t)} f_i(l) = \sum_{i=0}^{p+d-1} b_i^{(t-1)} f_i(l + 1) + \psi_l a_t \tag{A5.3.5}$$

By solving $p + d$ such equations for different values of l, we obtain the required updating formula for the individual coefficients, in the form

$$b_i^{(t)} = \sum_{j=0}^{p+d-1} L_{ij} b_j^{(t-1)} + g_i a_t$$

Note that the updating of each of the coefficients of the forecast function depends only on the lead 1 forecast error $a_t = z_t - \hat{z}_{t-1}(1)$.

A5.3.3 Comparison with the Discounted Least-Squares Method

Although to work with the integrated form is an unnecessarily complicated way of computing forecasts, it allows us to compare the present mean square error forecast with another type of forecast that has received considerable attention. Let us write

$$\mathbf{F}_l = \begin{bmatrix} f_0(l) & f_1(l) & \cdots & f_{p+d-1}(l) \\ f_0(l+1) & f_1(l+1) & \cdots & f_{p+d-1}(l+1) \\ \vdots & \vdots & \cdots & \vdots \\ f_0(l+p+d-1) & f_1(l+p+d-1) & \cdots & f_{p+d-1}(l+p+d-1) \end{bmatrix}$$

$$\mathbf{b}^{(t)} = \begin{bmatrix} b_0^{(t)} \\ b_1^{(t)} \\ \vdots \\ b_{p+d-1}^{(t)} \end{bmatrix} \qquad \boldsymbol{\psi}_l = \begin{bmatrix} \psi_l \\ \psi_{l+1} \\ \vdots \\ \psi_{l+p+d-1} \end{bmatrix}$$

Then, using (A5.3.5) for $l, l+1, \ldots, l+p+d+-1$, we obtain for $l > q-p-d$,

$$\mathbf{F}_l \mathbf{b}^{(t)} = \mathbf{F}_{l+1} \mathbf{b}^{t-1} + \boldsymbol{\psi}_l a_t$$

yielding

$$\mathbf{b}^{(t)} = (\mathbf{F}_l^{-1} \mathbf{F}_{l+1}) \mathbf{b}^{(t-1)} + (\mathbf{F}_l^{-1} \boldsymbol{\psi}_l) a_t$$

or

$$\mathbf{b}^{(t)} = \mathbf{L}' \mathbf{b}^{(t-1)} + \mathbf{g} a_t \tag{A5.3.6}$$

Equation (A5.3.6) is of the same algebraic *form* as the updating function given by the ''discounted least-squares'' procedure of Brown (1962) and Brown and Meyer (1961). For comparison, if we denote the forecast error given by that method by e_t, then Brown's updating formula may be written as

$$\boldsymbol{\beta}^{(t)} = \mathbf{L}' \boldsymbol{\beta}^{(t-1)} + \mathbf{h} e_t \tag{A5.3.7}$$

where $\boldsymbol{\beta}^{(t)}$ is his vector of adaptive coefficients. The same matrix $\mathbf{L}$ appears in (A5.3.6) and (A5.3.7). This is inevitable, for this first factor merely allows for changes in the coefficients arising from translation to the new origin and would have to occur in any such formula. For example, consider the straight line forecast function:

$$\hat{z}_{t-1}(l) = b_0^{(t-1)} + b_l^{(t-1)} l$$

where $b_0^{(t-1)}$ is the ordinate at time $t-1$, the origin of the forecast. This can equally well be written as

$$\hat{z}_{t-1}(l) = (b_0^{(t-1)} + b_1^{(t-1)}) + b_1^{(t-1)}(l-1)$$

where now $(b_0^{(t-1)} + b_1^{(t-1)})$ is the ordinate at time t. Obviously, if we update the forecast to origin t, the coefficient b_0 must be suitably adjusted even if the forecast function were to remain unchanged.

In general, the matrix **L** does not change the forecast function, it merely relocates it. The actual updating is done by the vector of coefficients **g** and **h**. We will see that the coefficients **g**, which yield the minimum mean square error forecasts, and the coefficients **h** given by Brown are in general completely different.

Brown's Method of Forecasting.

1. A forecast function is selected from the general class of linear combinations and products of polynomials, exponentials, and sines and cosines.
2. The selected forecast function is fitted to past values by a ''discounted least-squares'' procedure. In this procedure, the coefficients are estimated and updated so that the sum of squares of weighted discrepancies

$$s_w = \sum_{j=0}^{\infty} \omega_j [z_{t-j} - \hat{\hat{z}}_t(-j)]^2 \tag{A5.3.8}$$

between past values of the series and the value given by the forecast function at the corresponding past time are minimized. The weight function ω_j is chosen arbitrarily to fall off geometrically, so that $\omega_j = (1-\alpha)^j$, where the constant α, usually called the *smoothing constant*, is (again arbitrarily) set equal to a value in the range 0.1–0.3.

Difference between the Minimum Mean Square Error Forecasts and those of Brown. To illustrate these comments, consider the forecasting of IBM stock prices, discussed by Brown (1962, p. 141). In this study, he used a quadratic model that would be, in the present notation,

$$\hat{\hat{z}}_t(l) = \beta_0^{(t)} + \beta_1^{(t)} l + \frac{1}{2}\beta_2^{(t)} l^2$$

With this model, he employed his method of discounted least squares to forecast stock prices 3 days ahead. The results obtained from this method are shown for a section of the IBM series in Figure A5.2, where they are compared with the minimum mean square error forecasts.

The discounted least-squares method can be criticized on the following grounds:

1. The nature of the forecast function ought to be decided by the autoregressive operator $\varphi(B)$ in the stochastic model, and not arbitrarily. In particular, it cannot be safely chosen by visual inspection of the time series itself. For example, consider the IBM stock prices plotted in Figure A5.2. It will be seen that a quadratic function might well be used to *fit* short pieces of this series to values already available. If such fitting were relevant to forecasting, we might conclude, as did Brown, that a polynomial forecast function of degree 2 was indicated. The most general linear process for which a quadratic function would produce *minimum mean square error* forecasts at every lead time $l = 1, 2, \ldots$ is defined by the (0, 3, 3) model

$$\nabla^3 z_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3) a_t$$

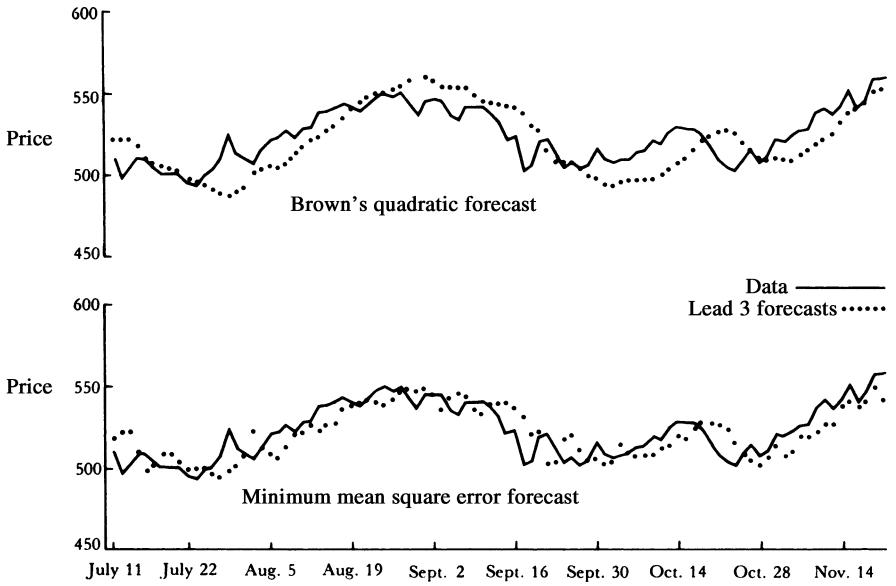


FIGURE A5.2 IBM stock price series with comparison of lead 3 forecasts obtained from best IMA(0, 1, 1) model and Brown's quadratic forecast for a period beginning from July 11, 1960.

which, arguing as in Section 4.3.3, can be written as

$$\nabla^3 z_t = \nabla^3 a_t + \lambda_0 \nabla^2 a_{t-1} + \lambda_1 \nabla a_{t-1} + \lambda_2 a_{t-1}$$

However, we show in Chapter 7 that if this model is correctly fitted, the least-squares estimates of the parameters are $\lambda_1 = \lambda_2 = 0$ and $\lambda_0 \simeq 1.0$. Thus, $\nabla z_t = (1 - \theta B)a_t$, with $\theta = 1 - \lambda_0$ close to zero, is the appropriate stochastic model, and the appropriate forecasting polynomial is $\hat{z}_t(l) = \beta_0^{(t)}$, which is of degree 0 in l and not of degree 2.

2. The choice of the weight function ω_j in (A5.3.8) must correspondingly be decided by the stochastic model, and not arbitrarily. The use of the discounted least-squares fitting procedure would produce minimum mean square error forecasts in the very restricted case, where

a. the process was of order (0, 1, 1), so $\nabla z_t = (1 - \theta B)a_t$,

b. a polynomial of degree 0 was fitted, and

c. the smoothing constant α was set equal to our $\lambda = 1 - \theta$.

In the present example, even if the correct polynomial model of degree 0 had been chosen, the value $\alpha = \lambda = 0.1$, actually used by Brown, would have been quite inappropriate. The correct value λ for this series is close to unity.

3. The exponentially discounted weighted least-squares procedure forces all the $p + d$ coefficients in the updating vector $\mathbf{h}$ to be functions of the single smoothing parameter α. In fact, they should be functions of the $p + q$ independent parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$.

Thus, the differences between the two methods are not trivial, and it is interesting to compare their performances on the IBM data. The minimum mean square error forecast is

TABLE A5.3 Comparison of Mean Square Error of Forecasts Obtained at Various Lead Times Using Best IMA(0, 1, 1) Model and Brown's Quadratic Forecasts

	Lead Time l									
	1	2	3	4	5	6	7	8	9	10
MSE (Brown)	102	158	218	256	363	452	554	669	799	944
MSE ($\lambda = 0.9$)	42	91	136	180	282	266	317	371	427	483

$\hat{z}_t(l) = b_0(t)$, with updating $b_0^{(t)} = b_0^{(t-1)} + \lambda a_t$, where $\lambda \simeq 1.0$. If λ is taken to be exactly equal to unity, this is equivalent to using

$$\hat{z}_t(l) = z_t$$

which implies that the best forecast of the stock price for all future time is the present price.[1] The suggestion that stock prices behave in this way is, of course, not new and goes back to Bachelier (1900). Since $z_t = Sa_t$ when $\lambda = 1$, this implies that z_t is a random walk.

To compare the minimum mean square error forecast with Brown's quadratic forecasts, a direct comparison was made using the IBM stock price series from July 11, 1960 to February 10, 1961, for 150 observations. For this stretch of the series, the minimum MSE forecast is obtained using the model $\nabla z_t = a_t - \theta a_{t-1}$, with $\theta = 0.1$, or $\lambda = 1 - \theta = 0.9$. Figure A5.2 shows the minimum MSE forecasts for lead time 3 and the corresponding values of Brown's quadratic forecasts. It is seen that the minimum MSE forecasts, which are virtually equivalent to using today's price to predict that 3 days ahead, are considerably better than those obtained using Brown's more complicated procedure.

The mean square errors for the forecast at various lead times, computed by direct comparison of the value of the series and their lead l forecasts, are shown in Table A5.3 for the two types of forecasts. It is seen that Brown's quadratic forecasts have mean square errors that are much larger than those obtained by the minimum mean square error method.

EXERCISES

5.1. For the models

(1) $\tilde{z}_t - 0.5\tilde{z}_{t-1} = a_t$

(2) $\nabla z_t = a_t - 0.5a_{t-1}$

(3) $(1 - 0.6B)\nabla z_t = a_t$

write down the forecasts for lead times $l = 1$ and $l = 2$:

(a) From the difference equation

(b) In integrated form (using the ψ_j weights)

(c) As a weighted average of previous observations

[1] This result is approximately true supposing that no relevant information except past values of the series itself is available and that fairly short forecasting periods are being considered. For longer periods, growth and inflationary factors would become important.

5.2. The following observations represent values $z_{91}, z_{92}, \ldots, z_{100}$ from a series fitted by the model $\nabla z_t = a_t - 1.1a_{t-1} + 0.28a_{t-2}$:

$$166, 172, 172, 169, 164, 168, 171, 167, 168, 172$$

(a) Generate the forecasts $\hat{z}_{100}(l)$ for $l = 1, 2, \ldots, 12$ and draw a graph of the series values and the forecasts (assume $a_{90} = 0, a_{91} = 0$).

(b) With $\hat{\sigma}_a^2 = 1.103$, calculate the estimated standard deviations $\hat{\sigma}(l)$ of the forecast errors and use them to calculate 80% probability limits for the forecasts. Insert these probability limits on the graph, on either side of the forecasts.

5.3. Suppose that the data of Exercise 5.2 represent monthly sales.

(a) Calculate the minimum mean square error forecasts for quarterly sales for 1, 2, 3, 4 *quarters* ahead, using the data up to $t = 100$.

(b) Calculate 80% probability limits for these forecasts.

5.4. Using the data and forecasts of Exercise 5.2, and given the further observation $z_{101} = 174$:

(a) Calculate the forecasts $\hat{z}_{101}(l)$ for $l = 1, 2, \ldots, 11$ using the updating formula

$$\hat{z}_{t+1}(l) = \hat{z}_t(l+1) + \psi_l a_{t+1}$$

(b) Verify these forecasts using the difference equation directly.

5.5. For the model $\nabla z_t = a_t - 1.1a_{t-1} + 0.28a_{t-2}$ of Exercise 5.2:

(a) Write down expressions for the forecast errors $e_t(1), e_t(2), \ldots, e_t(6)$, from the same origin t.

(b) Calculate and plot the autocorrelations of the series of forecast errors $e_t(3)$.

(c) Calculate and plot the correlations between the forecast errors $e_t(2)$ and $e_t(j)$ for $j = 1, 2, \ldots, 6$.

5.6. Let the vector $\boldsymbol{e}' = (e_1, e_2, \ldots, e_L)$ have for its elements the forecast errors made $1, 2, \ldots, L$ steps ahead, all from the same origin t. Then if $\boldsymbol{a}' = (a_{t+1}, a_{t+2}, \ldots, a_{t+L})$ are the corresponding uncorrelated random shocks, show that

$$\boldsymbol{e} = \mathbf{M}\boldsymbol{a} \quad \text{where} \quad \mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \psi_1 & 1 & 0 & \cdots & 0 \\ \psi_2 & \psi_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \psi_{L-1} & \psi_{L-2} & \psi_{L-3} & \cdots & 1 \end{bmatrix}$$

Also, show that (e.g., Box and Tiao, 1976; Tiao et al., 1975) $\boldsymbol{\Sigma}_e$, the covariance matrix of the e's, is $\boldsymbol{\Sigma}_e = \sigma_a^2 \mathbf{M}\mathbf{M}'$ and hence that a test to determine if a set of subsequently realized values $z_{t+1}, z_{t+2}, \ldots, z_{t+L}$ of the series taken jointly differ significantly from the forecasts made at the origin t is obtained by referring

$$\boldsymbol{e}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{e} = \frac{\boldsymbol{e}'(\mathbf{M}\mathbf{M}')^{-1}\boldsymbol{e}}{\sigma_a^2} = \frac{\boldsymbol{a}'\boldsymbol{a}}{\sigma_a^2} = \frac{1}{\sigma_a^2}\sum_{j=1}^{L} a_{t+j}^2$$

to a chi-square distribution with L degrees of freedom. Note that a_{t+j} is the *one*-step-ahead forecast error calculated from $z_{t+j} - \hat{z}_{t+j-1}(1)$.

5.7. Suppose that a quarterly economic time series is well represented by the model

$$\nabla z_t = 0.5 + (1 - 1.0B + 0.5B^2)a_t$$

with $\sigma_a^2 = 0.04$.

(a) Given $z_{48} = 130, a_{47} = -0.3, a_{48} = 0.2$, calculate and plot the forecasts $\hat{z}_{48}(l)$ for $l = 1, 2, \ldots, 12$.

(b) Calculate and insert the 80% probability limits on the graph.

(c) Express the series and forecasts in integrated form.

5.8. Consider the annual Wölfer sunspot numbers for the period 1770–1869 listed as Series E in Part Five of this text. The same series is available for the longer period 1700–1988 as "sunspot.year" in the datasets package of R. You can use either data set. Suppose that the series can be represented by an autoregressive model of order 3.

(a) Plot the time series and comment. Does the series look stationary?

(b) Generate forecasts and associated probability limits for up to 20 time periods ahead for the series.

(c) Perform a square root transformation of the data and repeat (a) and (b) above.

(d) Use the function BoxCox.ar() in the TSA package of R to show that the square root transformation is appropriate for this series; see help(BoxCox.ar) for details. (*Note:* Adding a small amount, for example, 1/2, to the series, eliminates zero values and allows the program to consider a log transformation as an option).

5.9. A time series representing a global mean land–ocean temperature index from 1880 to 2009 is available in a file called ''gtemp'' in the astsa package of R. The data are temperature deviations, measured in degree centigrades, from the 1951–1980 average temperature, as described by Shumway and Stoffer (2011, p. 5). Assume that a third-order autoregressive model is appropriate for the first differences $w_t = (1 - B)z_t$ of this series.

(a) Plot the time series z_t and the differenced series w_t using R.

(b) Generate forecasts and associated probability limits for up to 20 time periods ahead for this series using the function sarima.for() without including a constant term in the model.

(c) Generate the same forecasts and probability limits as in part (b) but with a constant term now added to the model. Discuss your findings and comment on the implications of including a constant in this case.

5.10. For the model $(1 - 0.6B)(1 - B)z_t = (1 + 0.3B)a_t$, express explicitly in the state-space form of (5.5.2) and (5.5.3), and write out precisely the recursive relations of the Kalman filter for this model. Indicate how the (exact) forecasts $\hat{z}_{t+l|t}$ and their forecast error variances $v_{t+l|t}$ are determined from these recursions.

PART TWO

STOCHASTIC MODEL BUILDING

We have seen that an ARIMA model of order (p, d, q) provides a class of models capable of representing time series that, although not necessarily stationary, are homogeneous and in statistical equilibrium in many respects.

The ARIMA model is defined by the equation

$$\phi(B)(1 - B)^d z_t = \theta_0 + \theta(B)a_t$$

where $\phi(B)$ and $\theta(B)$ are operators in B of degree p and q, respectively, whose zeros lie outside the unit circle. We have noted that the model is very general, including as special cases autoregressive models, moving average models, mixed autoregressive–moving average models, and the integrated forms of all three.

Iterative Approach to Model Building. The development of a model of this kind to describe the dependence structure in an observed time series is usually best achieved by a three-stage iterative procedure based on identification, estimation, and diagnostic checking.

1. By *identification* we mean the use of the data, and of any information on how the series was generated, to suggest a subclass of parsimonious models worthy to be entertained.
2. By *estimation* we mean efficient use of the data to make inferences about the parameters conditional on the adequacy of the model entertained.
3. By *diagnostic checking* we mean checking the fitted model in its relation to the data with intent to reveal model inadequacies and so to achieve model improvement.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

In Chapter 6, which follows, we discuss model identification, in Chapter 7 estimation of parameters, and in Chapter 8 diagnostic checking of the fitted model. In Chapter 9 we expand on the class of models developed in Chapters 3 and 4 to the *seasonal* ARIMA models, and all the model building techniques of the previous chapters are illustrated by applying them to modeling and forecasting seasonal time series. In Chapter 10 we consider some additional topics that represent extensions beyond the linear ARIMA class of models such as conditional heteroscedastic time series models, nonlinear time series models, and fractionally integrated long memory processes, which allow for certain more general features in the time series than are possible in the linear ARIMA models. Unit root testing is also discussed in this chapter.

6

MODEL IDENTIFICATION

In this chapter, we discuss methods for identifying nonseasonal autoregressive integrated moving average (ARIMA) time series models. Identification methods are rough procedures applied to a set of data to indicate the kind of model that is worthy of further investigation. The specific aim here is to obtain some idea of the values of p, d, and q needed in the general linear ARIMA model and to obtain initial estimates for the parameters. The tentative model specified provides a starting point for the application of the more formal and efficient estimation methods described in Chapter 7. The examples used to demonstrate the model-building process will include Series A–F that have been discussed in earlier chapters and are listed in the Collection of Time Series in Part Five of this book. The series are also available electronically at http://pages.stat.wisc.edu/ reinsel/bjr-data/.

6.1 OBJECTIVES OF IDENTIFICATION

It should first be said that identification and estimation necessarily overlap. Thus, we may estimate the parameters of a model, which is more elaborate than the one we expect to use, so as to decide *at what point* simplification is possible. Here we employ the estimation procedure to carry out part of the identification. It should also be explained that identification is necessarily inexact. It is inexact because the question of what types of models occur in practice and in what specific cases depends on the behavior of the physical world and therefore cannot be decided by purely mathematical argument. Furthermore, because at the identification stage no precise formulation of the problem is available, statistically ''inefficient'' methods must necessarily be used. It is a stage at which graphical methods are particularly useful and judgment must be exercised. However, it should be kept in mind

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

that the preliminary identification commits us to nothing except tentative consideration of a class of models that will later be efficiently fitted and checked.

6.1.1 Stages in the Identification Procedure

Our task, then, is to identify an appropriate subclass of models from the general ARIMA family

$$\phi(B)\nabla^d z_t = \theta_0 + \theta(B)a_t \tag{6.1.1}$$

which may be used to represent a given time series. Our approach will be as follows:

1. To assess the stationarity of the process z_t and, if necessary, to difference z_t as many times as is needed to produce stationarity, hopefully reducing the process under study to the mixed autoregressive–moving average process:

   $$\phi(B)w_t = \theta_0 + \theta(B)a_t$$

   where

   $$w_t = (1 - B)^d z_t = \nabla^d z_t$$

2. To identify the resulting autoregressive–moving average (ARMA) model for w_t.

Our principal tools for putting steps 1 and 2 into effect will be the sample autocorrelation function and the sample partial autocorrelation function. They are used not only to help guess the form of the model but also to obtain approximate estimates of the parameters. Such approximations are often useful at the estimation stage to provide starting values for iterative procedures employed at that stage. Some additional model identification tools may also be employed and are discussed in Section 6.2.4.

6.2 IDENTIFICATION TECHNIQUES

6.2.1 Use of the Autocorrelation and Partial Autocorrelation Functions in Identification

Identifying the Degree of Differencing. We have seen in Section 3.4.2 that for a stationary mixed autoregressive–moving average process of order $(p, 0, q)$, $\phi(B)\tilde{z}_t = \theta(B)a_t$, the autocorrelation function satisfies the difference equation

$$\phi(B)\rho_k = 0, k > q$$

Also, if $\phi(B) = \prod_{i=1}^{p}(1 - G_i B)$, the solution of this difference equation for the kth autocorrelation is, assuming distinct roots, of the form

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \cdots + A_p G_p^k \qquad k > q - p \tag{6.2.1}$$

The stationarity requirement that the zeros of $\phi(B)$ lie outside the unit circle implies that the roots $G_1, G_2, \ldots, G_p$ lie inside the unit circle.

This expression shows that in the case of a stationary model in which none of the roots lie close to the boundary of the unit circle, the autocorrelation function will quickly ''die out'' for moderate and large k. Suppose now that a single real root, say G_1, approaches unity, so that

$$G_1 = 1 - \delta$$

where δ is some small positive quantity. Then, since for k large

$$\rho_k \simeq A_1(1 - k\delta)$$

the autocorrelation function will not die out quickly and will fall off slowly and very nearly linearly. A similar argument may be applied if more than one of the roots approaches unity.

Therefore, a tendency for the autocorrelation function not to die out quickly is taken as an indication that a root close to unity may exist. The estimated autocorrelation function tends to follow the behavior of the theoretical autocorrelation function. Therefore, failure of the estimated autocorrelation function to die out rapidly might logically suggest that we should treat the underlying stochastic process as nonstationary in z_t, but possibly as stationary in ∇z_t, or in some higher difference.

However, even though failure of the estimated autocorrelation function to die out rapidly suggests nonstationarity, the estimated autocorrelations need not be extremely high even at low lags. This is illustrated in Appendix A6.1, where the expected behavior of the estimated autocorrelation function is considered for the nonstationary $(0, 1, 1)$ process $\nabla z_t = (1 - \theta B)a_t$. The ratio $E[c_k]/E[c_0]$ of expected values falls off only slowly, but depends initially on the value of θ and on the number of observations in the series, and need not be close to unity if θ is close to 1. We illustrate this point again in Section 6.3.4 for Series A.

For the reasons given, it is assumed that the degree of differencing d, necessary to achieve stationarity, has been reached when the autocorrelation function of $w_t = \nabla^d z_t$ dies out fairly quickly. In practice, d is normally 0, 1, or 2, and it is usually sufficient to inspect the first 20 or so estimated autocorrelations of the original series, and of its first and second differences, if necessary.

Overdifferencing. Once stationarity is achieved, further differencing should be avoided. Overdifferencing introduces extra serial correlation and increases model complexity. To illustrate this point, assume that the series z_t follows a random walk so that the differenced series $w_t = (1 - B)z_t = a_t$ is white noise and thus stationary. Further differencing of w_t leads to $(1 - B)w_t = (1 - B)a_t$, which is a MA(1) model for w_t with parameter $\theta = 1$. Thus, the resulting model for z_t would be an ARIMA(0, 2, 1) model instead of the simpler ARIMA(0, 1, 0) model. The model with $\theta = 1$ is noninvertible and the pure autoregressive representation does not exit. Noninvertibility also causes problems at the parameter estimation stage in that approximate maximum likelihood methods tends to produce biased estimates in this case.

Figure 6.1 shows the autocorrelation and partial autocorrelation functions of a time series of length 200 generated from a random walk model with innovations variance equal to 1. The first 1000 observations were discarded to eliminate potential start-up effects. The estimated autocorrelations up to lag 20 of the original series and its first and second differences are shown in the graph. The autocorrelations of the original series fail to damp out quickly,

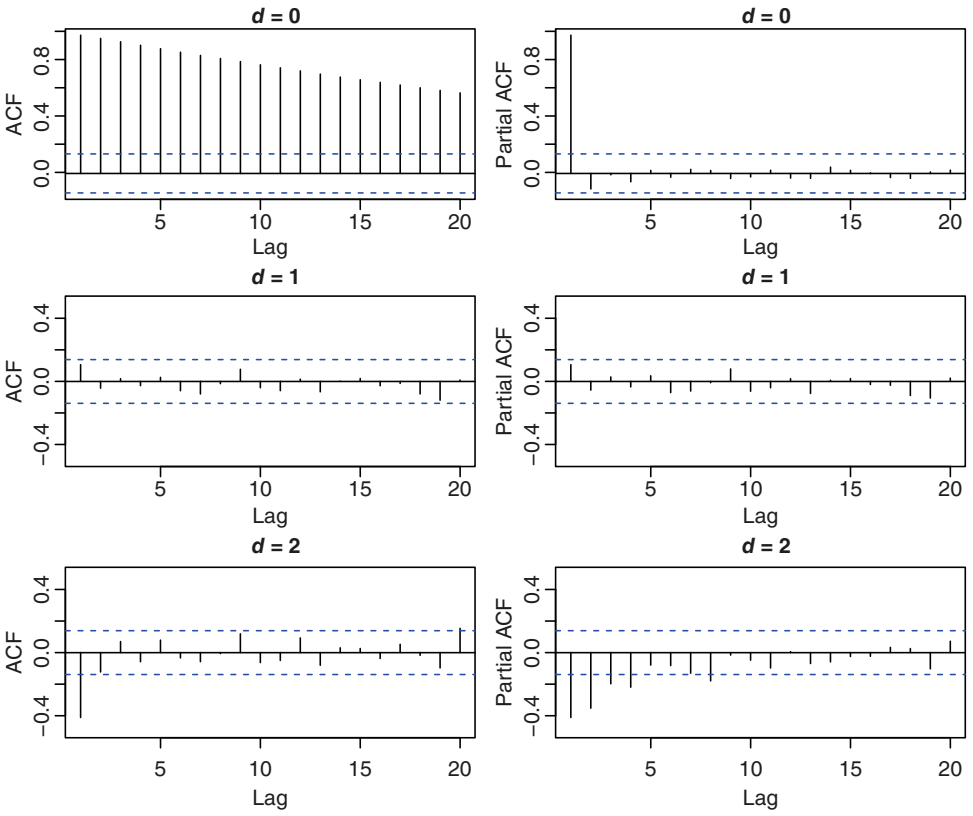


FIGURE 6.1 Estimated autocorrelation and partial autocorrelation functions for a simulated random walk process and its first ($d = 1$) and second ($d = 2$) differences.

indicating a need for differencing. The autocorrelations of $w_t = \nabla z_t$, on the other hand, are all small, demonstrating that stationarity has now been achieved. The autocorrelation function of the second differences $w_t = \nabla^2 z_t$ also indicates stationarity, but it has a spike at lag 1 showing the extra correlation that has emerged because of overdifferencing. The value of r_1 is close to -0.5, which is consistent with the lag 1 autocorrelation coefficient for an MA(1) model with $\theta = 1$. Figure 6.1 can be reproduced in R as follows:

```
> RW=arima.sim(list(order=c(0,1,0)),n=200,n.start=1000)
> acf0=acf(RW,20)
> pacf0=pacf(RW,20)
> acf1=acf(diff(RW),20)
> pacf1=pacf(diff(RW),20)
> acf2=acf(diff(diff(RW)),20)
> pacf2=pacf(diff(diff(RW)),20)
> par(mfrow=c(3,2))
> plot(acf0,main='d=0')
> plot(pacf0,main='d=0')
> plot(acf1,ylim=c(-0.5,0.5),main='d=1')
> plot(pacf1,ylim=c(-0.5,0.5),main='d=1')
```

```
> plot(acf2,ylim=c(-0.5,0.5),main='d=2')
> plot(pacf2,ylim=c(-0.5,0.5),main='d=2')
```

Identifying a Stationary ARMA Model for the Differenced Series. Having tentatively decided on the degree of differencing d, we examine the patterns of the estimated autocorrelation and partial autocorrelation functions of the differenced series, $w_t = (1 - B)^d z_t$, to determine a suitable choice for the orders p and q of the autoregressive and moving average operators. Here we recall the characteristic behavior of the theoretical autocorrelation and partial autocorrelation functions for moving average, autoregressive, and mixed processes, discussed in Chapter 3.

Briefly, whereas the autocorrelation function of an autoregressive process of order p tails off, its partial autocorrelation function has a cutoff after lag p. Conversely, the autocorrelation function of a moving average process of order q has a cutoff after lag q, while its partial autocorrelation function tails off. If both the autocorrelations and partial autocorrelations tail off, a mixed process is suggested. Furthermore, the autocorrelation function for a mixed process, containing a pth-order autoregressive component and a qth-order moving average component, is a mixture of exponentials and damped sine waves after the first $q-p$ lags. Conversely, the partial autocorrelation function for a mixed process is dominated by a mixture of exponentials and damped sine waves after the first $p-q$ lags (see Table 3.2).

In general, autoregressive (moving average) behavior, as measured by the autocorrelation function, tends to mimic moving average (autoregressive) behavior as measured by the partial autocorrelation function. For example, the autocorrelation function of a first-order autoregressive process decays exponentially, while the partial autocorrelation function cuts off after the first lag. Correspondingly, for a first-order moving average process, the autocorrelation function cuts off after the first lag. Although not precisely exponential, the partial autocorrelation function is dominated by exponential terms and has the general appearance of an exponential.

Of particular importance are the autoregressive and moving average processes of first and second order and the simple mixed $(1, d, 1)$ process. The properties of the theoretical autocorrelation and partial autocorrelation functions for these processes are summarized in Table 6.1, which requires careful study and provides a convenient reference table. The reader should also refer to Figures 3.2, 3.7, and 3.10, which show typical behavior of the autocorrelation function and the partial autocorrelation function for the second-order autoregressive process, the second-order moving average process, and the simple mixed ARMA(1, 1) process.

6.2.2 Standard Errors for Estimated Autocorrelations and Partial Autocorrelations

Estimated autocorrelations can have rather large variances and can be highly autocorrelated with each other. For this reason, *detailed* adherence to the theoretical autocorrelation function cannot be expected in the estimated function. In particular, moderately large estimated autocorrelations can occur after the theoretical autocorrelation function has damped out, and apparent ripples and trends can occur in the estimated function that have no basis in the theoretical function. In employing the estimated autocorrelation function as a tool for identification, it is usually possible to be fairly sure about broad characteristics, but more subtle indications may or may not represent real effects. For these reasons, two or more related

TABLE 6.1 Behavior of the Autocorrelation Functions for the dth Difference of an ARIMA Process of Order (p, d, q)[a]

	Order				
	$(1, d, 0)$	$(0, d, 1)$	$(2, d, 0)$	$(0, d, 2)$	$(1, d, 1)$
Behavior of ρ_k	Decays exponentially	Only ρ_1 nonzero	Mixture of exponentials or damped sine wave	Only ρ_1 and ρ_2 nonzero	Decays exponentially from first lag
Behavior of ϕ_{kk}	Only ϕ_{11} nonzero	Exponential dominates decay	Only ϕ_{11} and ϕ_{22} nonzero	Dominated by mixture of exponential or damped sine wave	Dominated by exponential decay from first lag
Preliminary estimates from	$\phi_1 = \rho_1$	$\rho_1 = \dfrac{-\theta_1}{1+\theta_1^2}$	$\phi_1 = \dfrac{\rho_1(1-\rho_2)}{1-\rho_1^2}$	$\rho_1 = \dfrac{-\theta_1(1-\theta_2)}{1+\theta_1^2+\theta_2^2}$	$\rho_1 = \dfrac{(1-\theta_1\phi_1)(\phi_1-\theta_1)}{1+\theta_1^2-2\phi_1\theta_1}$
			$\phi_2 = \dfrac{\rho_2-\rho_1^2}{1-\rho_1^2}$	$\rho_2 = \dfrac{-\theta_2}{1+\theta_1^2+\theta_2^2}$	$\rho_2 = \phi_1\rho_1$
Admissible region	$-1 < \phi_1 < 1$	$-1 < \theta_1 < 1$	$-1 < \phi_2 < 1$ $\phi_2 + \phi_1 < 1$ $\phi_2 - \phi_1 < 1$	$-1 < \theta_2 < 1$ $\theta_2 + \theta_1 < 1$ $\theta_2 - \theta_1 < 1$	$-1 < \phi_1 < 1$ $-1 < \theta_1 < 1$

[a] Table A and Charts B–D are included at the end of this book to facilitate the calculation of approximate estimates of the parameters for first-order moving average, second-order autoregressive, second-order moving average, and the mixed ARMA(1, 1) processes.

models may need to be entertained and investigated further at the estimation and diagnostic checking stages of model building.

In practice, it is important to have some indication of how far an estimated value may differ from the corresponding theoretical value. In particular, we need some means for judging whether the autocorrelations and partial autocorrelations are effectively zero after some specific lag q or p, respectively. For *larger lags*, on the hypothesis that the process is moving average of order q, we can compute standard errors of estimated autocorrelations from the simplified form of Bartlett's formula (2.1.15), with sample estimates replacing theoretical autocorrelations. Thus,

$$\hat{\sigma}[r_k] \simeq \frac{1}{n^{1/2}}[1 + 2(r_1^2 + r_2^2 + \cdots + r_q^2)]^{1/2} \qquad k > q \tag{6.2.2}$$

For the partial autocorrelations, we use the result quoted in (3.2.36) that, on the hypothesis that the process is autoregressive of order p, the standard error for estimated partial autocorrelations of order $p + 1$ and higher is

$$\hat{\sigma}[\hat{\phi}_{kk}] \simeq \frac{1}{n^{1/2}} \qquad k > p \tag{6.2.3}$$

It was shown by Anderson (1942) that for moderate n, the distribution of an estimated autocorrelation coefficient, whose theoretical value is zero, is approximately normal. Thus, on the hypothesis that the theoretical autocorrelation ρ_k is zero, the estimate r_k divided by its standard error will be approximately distributed as a unit normal variate. A similar result is true for the partial autocorrelations. These facts provide an informal guide as to whether theoretical autocorrelations and partial autocorrelations beyond a particular lag are essentially zero.

6.2.3 Identification of Models for Some Actual Time Series

Series A–D. In this section, the model specification tools described above are applied to some of the actual time series that we encountered in earlier chapters. We first discuss potential models for Series A to D plotted in Figure 4.1. As remarked in Chapter 4 on nonstationarity, we expect Series A, C, and D to possess nonstationary characteristics since they represent the ''uncontrolled'' behavior of certain process outputs. Similarly, we would expect the IBM stock price Series B to have no fixed level and to be nonstationary.

The estimated autocorrelations of z_t and the first differences ∇z_t for Series A–D are shown in Figure 6.2. Figure 6.3 shows the corresponding estimated partial autocorrelations. The two figures were generated in R using commands similar to those used to produce Figure 6.1. For the chemical process concentration readings in Series A, the autocorrelations for ∇z_t are small after the first lag. This suggests that this time series might be described by an IMA(0, 1, 1) model. However, from the autocorrelation function of z_t, it is seen that after lag 1 the correlations do decrease fairly regularly. Therefore, an alternative is that the series follows a mixed ARMA(1, 0, 1) model. The partial autocorrelation function of z_t seems to support this possibility. We will see later that the two alternatives result in virtually the same model. For the stock price Series B, the results confirm the nonstationarity of the original series and suggest that a random walk model $(1 - B)z_t = a_t$ is appropriate for this series.

The estimated autocorrelations of the temperature Series C also indicate nonstationarity. The roughly exponential decay in the autocorrelations for the first difference suggests a

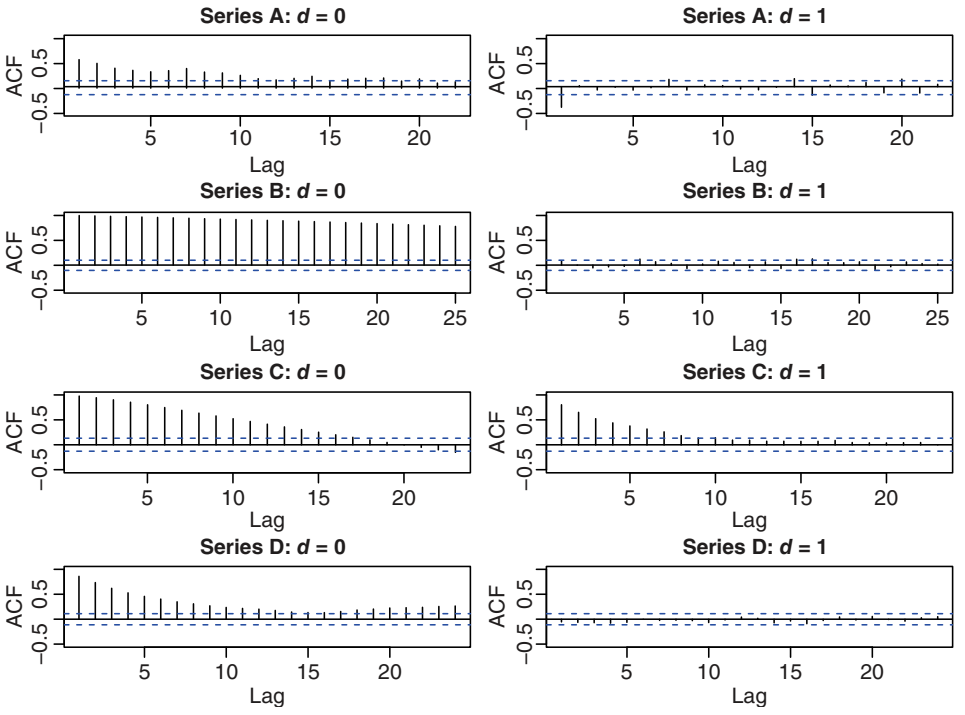


FIGURE 6.2 Estimated autocorrelation functions of the original series ($d = 0$), and their first differences ($d = 1$) for Series A–D.

process of order $(1, 1, 0)$, with an autoregressive parameter ϕ around 0.8. Alternatively, we notice that the autocorrelations of ∇z_t decay at a relatively slow rate, suggesting that further differencing might be needed. The autocorrelation and partial autocorrelation functions of the second differences $\nabla^2 z_t$ (not shown) were rather small, suggesting a white noise process for the second differences. This implies that an IMA$(0, 2, 0)$ model might also be appropriate for this series. Thus, the possibilities are

$$(1 - 0.8B)(1 - B)z_t = a_t$$
$$(1 - B)^2 z_t = a_t$$

The second model is very similar to the first, differing only in the choice of 0.8 rather than 1.0 for the autoregressive coefficient.

Finally, the autocorrelation and partial autocorrelation functions for the viscosity Series D suggest that an AR(1) model $(1 - \phi B)z_t = a_t$ with ϕ around 0.8 might be appropriate for this series. Alternatively, since the autocorrelation coefficients decay at a relatively slow rate, we will also consider the model $(1 - B)z_t = a_t$ for this series.

Series E and F. Series E shown in the top graph of Figure 6.4 represents the annual Wölfer sunspot numbers over the period 1770–1869. This series is likely to be stationary since the number of sunspots is expected to remain in equilibrium over long periods of time. The autocorrelation and partial autocorrelation functions in Figure 6.4 show characteristics similar to those of an AR(2) process. However, as will be seen later, a marginally better

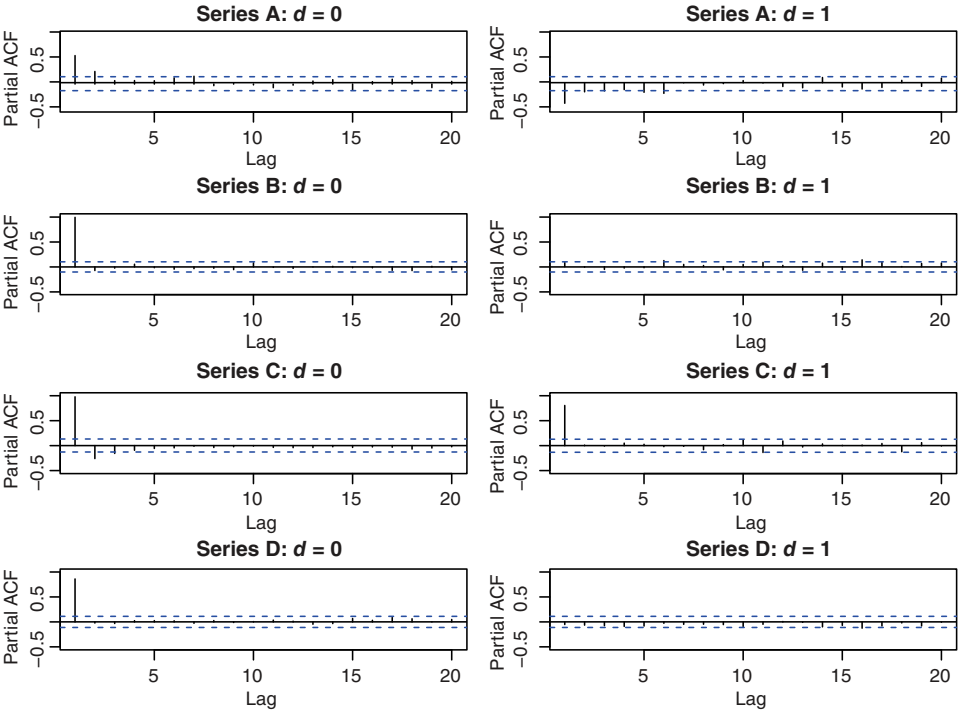


FIGURE 6.3 Estimated partial autocorrelation functions of the original series ($d = 0$), and their first differences ($d = 1$) for Series A–D.

fit is obtained using an AR(3) model. The fit can be improved further using a square root or log-transformation of the series. An autoregressive model of order nine is suggested by the order selection routine ar() in the R package that selects AR order based on the Akaike information criterion (AIC) to be discussed in Section 6.2.4. Other options considered in the literature include nonlinear time series models, such as bilinear or threshold autoregressive models, discussed briefly in Section 10.3.

Series F introduced in Chapter 2 represents the yields of a batch chemical process. The series is expected to be stationary since the batches are processed under uniformly controlled conditions. The stationarity is confirmed by Figure 6.5 that shows a graph of the series along with the autocorrelation and partial autocorrelation functions of the series and its first differences. The results for the undifferenced series suggest that a first-order autoregressive model might be appropriate for this series.

A summary of the models tentatively identified for Series A to F is given in Table 6.2. Note that for Series C and F, the alternative models suggested above have been made slightly more general for further illustrations later on.

Notes on the identification procedure. The graphs of the autocorrelation and partial autocorrelation functions shown above were generated using R. In assessing the estimated correlation functions, it is very helpful to plot one or two standard error limits around zero for the estimated coefficients. Limits from the R package are included in the graphs displayed above. These limits are approximate two standard error limits, $\pm 2/\sqrt{(n)}$, determined

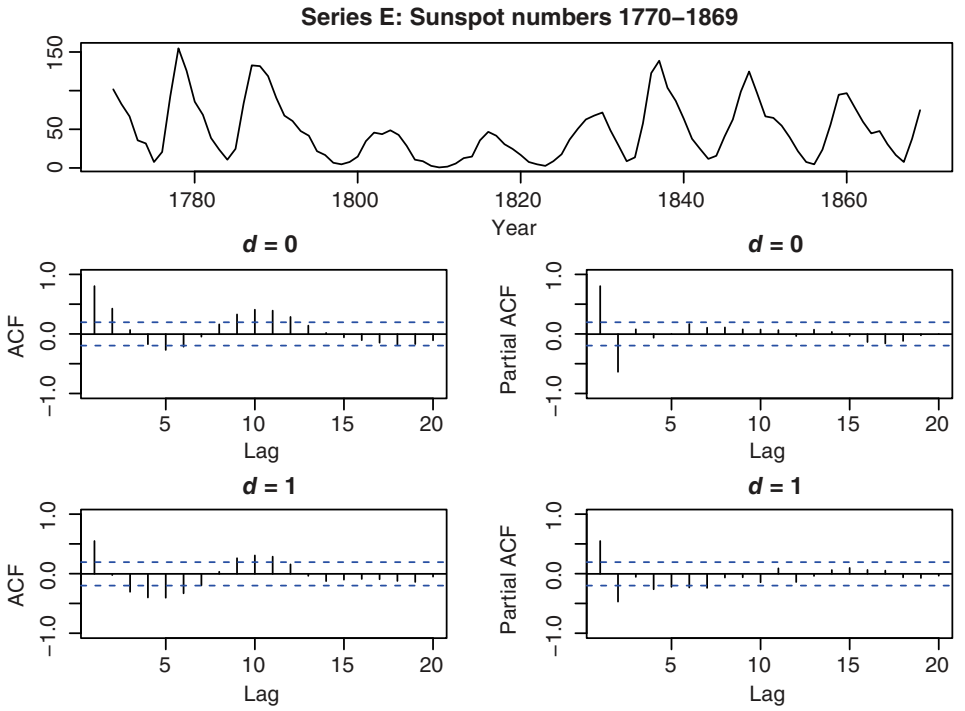


FIGURE 6.4 Estimated autocorrelation and partial autocorrelation functions of the sunspot series (Series E) and its first differences.

under the assumption that all the theoretical autocorrelation coefficients are zero so that the series is white noise. If a hypothesis about a specific model is postulated, alternative limits could be determined from Bartlett's formula as discussed above. When the calculations are performed in R, inclusion of the argument ci.type='ma' in the acf() function

TABLE 6.2 Tentative Identification of Models for Series A–F

Series	Degree of Differencing	Apparent Nature of Differenced Series	Identification for z_t
A	Either 0	Mixed first-order AR with first-order MA	(1, 0, 1)
	or 1	First-order MA	(0, 1, 1)
B	1	First-order MA	(0, 1, 1)
C	Either 1	First-order AR	(1, 1, 0)
	or 2	Uncorrelated noise	(0, 2, 2)[a]
D	Either 0	First-order AR	(1, 0, 0)
	or 1	Uncorrelated noise	(0, 1, 1)[a]
E	Either 0	Second-order AR	(2, 0, 0)
	or 0	Third-order AR	(3, 0, 0)
F	0	Second-order AR	(2, 0, 0)

[a] The order of the moving average operator appears to be zero, but the more general form is retained for subsequent consideration.

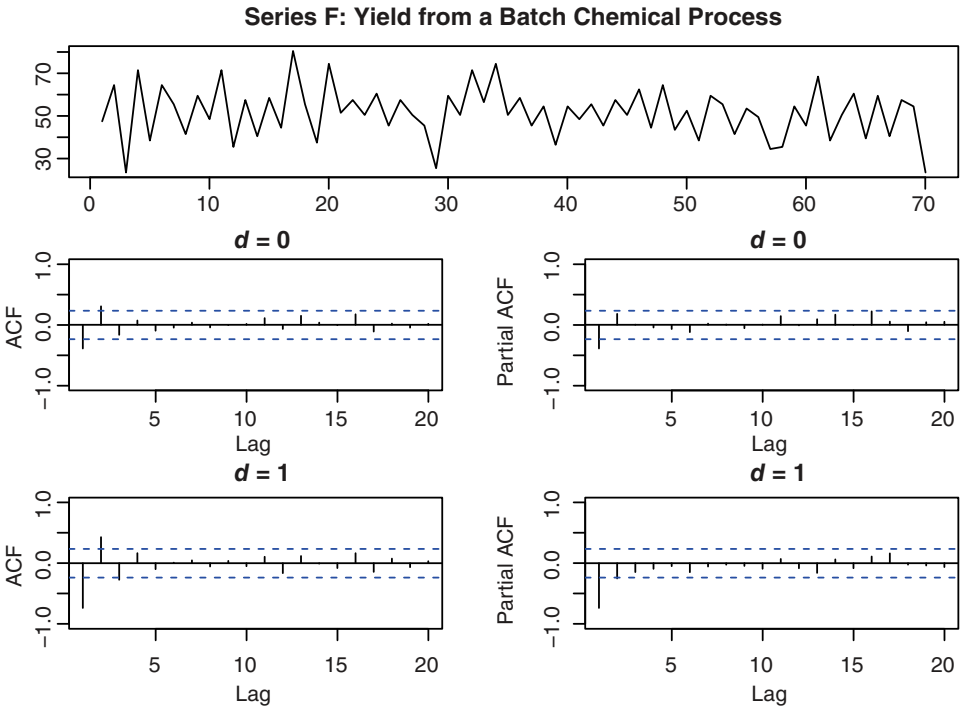


FIGURE 6.5 Estimated autocorrelation and partial autocorrelation functions for the yield of a batch chemical process (Series F) and its first differences.

yields confidence bounds computed based on the assumption that the true model is MA($k-1$).

Three other points concerning this identification procedure need to be mentioned:

1. Simple differencing of the kind we have used will not produce stationarity in series containing seasonal components. In Chapter 9, we discuss the appropriate modifications for such seasonal time series.
2. As discussed in Chapter 4, a nonzero value for θ_0 in (6.1.1) implies the existence of a systematic polynomial trend of degree d. For the nonstationary models in Table 6.2, a value of $\theta_0 = 0$ can perfectly well account for the behavior of the series. Occasionally, however, there will be some real physical phenomenon requiring the provision of such a component. In other cases, it might be uncertain whether or not such a provision should be made. Some indication of the evidence supplied by the data, for the inclusion of θ_0 in the model, can be obtained at the identification stage by comparing the mean $\overline{w}$ of $w_t = \nabla^d z_t$ with its approximate standard error, using $\sigma^2(\overline{w}) = n^{-1}\sigma_w^2[1 + 2\rho_1(w) + 2\rho_2(w) + \cdots]$.
3. It was noted in Section 3.4.2 that, for any ARMA(p, q) process with $p - q > 0$, the *whole positive half* of the autocorrelation function will be a mixture of damped sine waves and exponentials. This does not, of course, prevent us from tentatively identifying q, because (a) the partial autocorrelation function will show $p - q$ ''anomalous'' values before behaving like that of an MA(q) process, and (b) q must be such that the

autocorrelation function could take, as starting values following the general pattern, ρ_q back to $\rho_{-(p-q-1)}$.

6.2.4 Some Additional Model Identification Tools

Although the sample autocorrelation and partial autocorrelation functions are extremely useful in model identification, there are sometimes cases involving mixed models where they can provide ambiguous results. This may not be a serious problem since, as has been emphasized, model specification is always tentative and subject to further examination, diagnostic checking, and modification, if necessary. Nevertheless, there has been considerable interest in developing additional tools for use at the model identification stage. These include the R and S array approach proposed by Gray et al. (1978), the generalized partial autocorrelation function studied by Woodward and Gray (1981), the inverse autocorrelation function considered by Cleveland (1972) and Chatfield (1979), the extended sample autocorrelation function of Tsay and Tiao (1984), and the use of canonical correlation analysis as examined by Akaike (1976), Cooper and Wood (1982), and Tsay and Tiao (1985). Model selection criteria such as the AIC criterion introduced by Akaike (1974a) and the Bayesian Information Criterion (BIC) of Schwarz (1978) are also useful supplementary tools.

Canonical Correlation Methods. For illustration, we briefly discuss the use of canonical correlation analysis for model identification. In general, for two sets of random variables, $\boldsymbol{Y}_1 = (y_{11}, y_{12}, \ldots, y_{1k})'$ and $\boldsymbol{Y}_2 = (y_{21}, y_{22}, \ldots, y_{2l})'$, of dimensions k and l (assume $k \leq l$), canonical correlation analysis involves determining linear combinations $U_i = \boldsymbol{a}_i'\boldsymbol{Y}_1$ and $V_i = \boldsymbol{b}_i'\boldsymbol{Y}_2$, $i = 1, \ldots, k$, and corresponding correlations $\rho(i) = \text{corr}[U_i, V_i]$ with $\rho(1) \geq \rho(2) \geq \cdots \geq \rho(k) \geq 0$. The linear combinations are chosen so that the U_i and V_j are mutually uncorrelated for $i \neq j$, U_1 and V_1 have the maximum possible correlation $\rho(1)$ among all linear combinations of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, U_2 and V_2 have the maximum possible correlation $\rho(2)$ among all linear combinations of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ that are uncorrelated with U_1 and V_1, and so on. The resulting correlations $\rho(i)$ are called the *canonical correlations* between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, and the variables U_i and V_i are the corresponding canonical variates. If $\boldsymbol{\Omega} = \text{cov}[\boldsymbol{Y}]$ denotes the covariance matrix of $\boldsymbol{Y} = (\boldsymbol{Y}_1', \boldsymbol{Y}_2')'$, with $\boldsymbol{\Omega}_{ij} = \text{cov}[\boldsymbol{Y}_i, \boldsymbol{Y}_j]$, then it is known that the values $\rho^2(i)$ are the ordered eigenvalues of the matrix $\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}$ and the vectors $\boldsymbol{a}_i$, such that $U_i = \boldsymbol{a}_i'\boldsymbol{Y}_1$, are the corresponding (normalized) eigenvectors; that is, the $\rho^2(i)$ and $\boldsymbol{a}_i$ satisfy

$$[\rho^2(i)\mathbf{I} - \boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21}]\boldsymbol{a}_i = \mathbf{0} \qquad i = 1, \ldots, k \tag{6.2.4}$$

with $\rho^2(1) \geq \rho^2(2) \geq \cdots \geq \rho^2(k) \geq 0$ (e.g., Anderson (1984), p. 490). Similarly, one can define the notion of *partial canonical correlations* between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, given another set of variables $\boldsymbol{Y}_3$, as the canonical correlations between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ after they have been ''adjusted'' for the effects of $\boldsymbol{Y}_3$ by linear regression on Y_3, analogous to the definition of partial correlations as discussed in Section 3.2.5. A useful property to note is that if there exist (at least) $s \leq k$ linearly independent linear combinations of $\boldsymbol{Y}_1$ that are completely uncorrelated with $\boldsymbol{Y}_2$, say $\boldsymbol{U} = \mathbf{A}'\boldsymbol{Y}_1$ such that $\text{cov}[\boldsymbol{Y}_2, \boldsymbol{U}] = \boldsymbol{\Omega}_{21}\mathbf{A} = \mathbf{0}$, then there are (at least) s zero canonical correlations between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$. This follows easily from (6.2.4) since there will be (at least) s linearly independent eigenvectors satisfying (6.2.4) with

corresponding $\rho(i) = 0$. In effect, then, the number s of *zero* canonical correlations is equal to $s = k - r$, where $r = \text{rank}(\boldsymbol{\Omega}_{21})$.

In the ARMA time series model context, following the approach of Tsay and Tiao (1985), we consider $\boldsymbol{Y}_{m,t} = (\tilde{z}_t, \tilde{z}_{t-1}, \ldots, \tilde{z}_{t-m})'$ and examine the canonical correlation structure between the variables $\boldsymbol{Y}_{m,t}$ and

$$\boldsymbol{Y}_{m,t-j-1} = (\tilde{z}_{t-j-1}, \tilde{z}_{t-j-2}, \ldots, \tilde{z}_{t-j-1-m})'$$

for various combinations of $m = 0, 1, \ldots$ and $j = 0, 1, \ldots$ A key feature to recall is that the autocovariance function γ_k of an ARMA(p, q) process $\tilde{z}_t$ satisfies (3.4.2), and, in particular,

$$\gamma_k - \sum_{i=1}^{p} \phi_i \gamma_{k-i} = 0 \qquad k > q$$

Thus, for example, if $m \geq p$, there is (at least) one linear combination of $\boldsymbol{Y}_{m,t}$,

$$\tilde{z}_t - \sum_{i=1}^{p} \phi_i \tilde{z}_{t-1} = (1, -\phi_1, \ldots, -\phi_p, 0, \ldots, 0)\boldsymbol{Y}_{m,t} = \boldsymbol{a}'\boldsymbol{Y}_{m,t} \tag{6.2.5}$$

such that

$$\boldsymbol{a}'\boldsymbol{Y}_{m,t} = a_t - \sum_{i=1}^{q} \theta_i a_{t-i}$$

which is uncorrelated with $\boldsymbol{Y}_{m,t-j-1}$ for $j \geq q$. In particular, then, for $m = p$ and $j = q$, there is one zero canonical correlation between $\boldsymbol{Y}_{p,t}$ and $\boldsymbol{Y}_{p,t-q-1}$, as well as between $\boldsymbol{Y}_{p,t}$ and $\boldsymbol{Y}_{p,t-j-1}$, $j > q$, and between $\boldsymbol{Y}_{m,t}$ and $\boldsymbol{Y}_{m,t-q-1}$, $m > p$, while in general it is not difficult to establish that there are $s = \min(m + 1 - p, j + 1 - q)$ zero canonical correlations between $\boldsymbol{Y}_{m,t}$ and $\boldsymbol{Y}_{m,t-j-1}$ for $m > p$ and $j > q$. Hence, one can see that determination of the structure of the zero canonical correlations between $\boldsymbol{Y}_{m,t}$ and $\boldsymbol{Y}_{m,t-j-1}$ for various values of m and j will serve to characterize the orders p and q of the ARMA model, and so the canonical correlations will be useful in model identification. We note the special cases of these canonical correlations are as follows. First, when $m = 0$, we are simply examining the autocorrelations ρ_{j+1} between z_t and z_{t-j-1}, which will all equal zero in an MA(q) process for $j \geq q$. Second, when $j = 0$, we are examining the partial autocorrelations $\phi_{m+1,m+1}$ between z_t and z_{t-m-1}, given $z_{t-1}, \ldots, z_{t-m}$, and these will all equal zero in an AR(p) process for $m \geq p$. Hence, the canonical correlation analysis can be viewed as an extension of the analysis of the autocorrelation and partial autocorrelation functions of the process.

In practice, based on (6.2.4), one is led to consider the sample canonical correlations $\hat{\rho}(i)$, which are determined from the eigenvalues of the matrix:

$$\left(\sum\nolimits_t \boldsymbol{Y}_{m,t}\boldsymbol{Y}'_{m,t}\right)^{-1} \left(\sum\nolimits_t \boldsymbol{Y}_{m,t}\boldsymbol{Y}'_{m,t-j-1}\right) \times \left(\sum\nolimits_t \boldsymbol{Y}_{m,t-j-1}\boldsymbol{Y}'_{m,t-j-1}\right)^{-1} \left(\sum\nolimits_t \boldsymbol{Y}_{m,t-j-1}\boldsymbol{Y}'_{m,t}\right) \tag{6.2.6}$$

for various values of lag $j = 0, 1, \ldots$ and $m = 0, 1, \ldots$. Tsay and Tiao (1985) use a chi-squared test statistic approach based on the *smallest* eigenvalue (squared sample canonical correlation) $\hat{\lambda}(m, j)$ of (6.2.6). They propose the statistic $c(m, j) = -(n - m - j)\ln[1 -$

$\hat{\lambda}(m,j)/d(m,j)]$, where $d(m,j) = 1 + 2\sum_{i=1}^{j} r_i^2(w'), j > 0, r_i(w')$ denotes the sample autocorrelation at lag i of $w_t' = z_t - \hat{\phi}_1^{(j)} z_{t-1} - \cdots - \hat{\phi}_m^{(j)} z_{t-m}$, and the $\hat{\phi}_i^{(j)}$ are estimates of the ϕ_i's obtained from the eigenvector (see, for example, equation (6.2.5)) corresponding to $\hat{\lambda}(m,j)$. The statistic $c(m,j)$ has an asymptotic χ_1^2 distribution when $m = p$ and $j \geq q$ or when $m \geq p$ and $j = q$ and can be used to test whether there exists a zero canonical correlation in theory. Hence if the sample statistics exhibit a pattern such that they are all insignificant, relative to a χ_1^2 distribution, for $m \geq p$ and $j \geq q$ for some p and q values, then the model might reasonably be identified as an ARMA(p,q) for the smallest values (p,q) such that this pattern holds. Tsay and Tiao (1985) also show that this procedure is valid for nonstationary ARIMA models $\varphi(B)z_t = \theta(B)a_t$, in the sense that the overall order $p + d$ of the generalized AR operator $\varphi(B)$ can be determined by the procedure, without initially deciding on differencing of the original series z_t.

Canonical correlation methods were previously also proposed for ARMA modeling by Akaike (1976) and Cooper and Wood (1982). Their approach is to perform a canonical correlation analysis between the vector of present and past values, $\boldsymbol{P}_t \equiv \boldsymbol{Y}_{m,t} = (\tilde{z}_t, \tilde{z}_{t-1}, \ldots, \tilde{z}_{t-m})'$, and the vector of future values, $\boldsymbol{F}_{t+1} = (\tilde{z}_{t+1}, \tilde{z}_{t+2}, \ldots)'$. In practice, the finite lag m used to construct the vector of present and past values $\boldsymbol{P}_t$ may be fixed by use of an order determination criterion such as Akaike information criteria to be discussed a little later in this section, applied to fitting of AR models of various orders. The canonical correlation analysis is performed sequentially by adding elements to $\boldsymbol{F}_{t+1}$ one at a time, starting with $\boldsymbol{F}_{t+1}^* = (\tilde{z}_{t+1})$, until the first zero canonical correlation between $\boldsymbol{P}_t$ and the $\boldsymbol{F}_{t+1}$ is determined. Akaike (1976) uses an AIC-type criterion called deviance information criterion (DIC) to judge whether the smallest sample canonical correlation can be taken to be zero, while Cooper and Wood (1982) use a traditional chi-squared statistic approach to assess the significance of the smallest canonical correlation, although as pointed out by Tsay (1989a), to be valid in the presence of a moving average component, this statistic needs to be modified.

At a given stage in the procedure, when the smallest sample canonical correlation between $\boldsymbol{P}_t$ and $\boldsymbol{F}_{t+1}^*$ is first judged to be 0 and $\tilde{z}_{t+K+1}$ is the most recent variable to be included in $\boldsymbol{F}_{t+1}^*$, a linear combination of $\tilde{z}_{t+K+1}$ in terms of the remaining elements of $\boldsymbol{F}_{t+1}^*$ is identified that is uncorrelated with the past. Specifically, the linear combination $\tilde{z}_{t+K+1} - \sum_{j=1}^{K} \phi_j \tilde{z}_{t+K+1-j}$ of the elements in the vector $\boldsymbol{F}_{t+1}^*$ of future values is (in theory) determined to be uncorrelated with the past $\boldsymbol{P}_t$. Hence, this canonical correlation analysis procedure determines that the forecasts $\hat{z}_t(K+1)$ of the process satisfy

$$\hat{z}_t(K+1) - \sum_{j=1}^{K} \phi_j \hat{z}_t(K+1-j) = \theta_0$$

By reference to the relation (5.3.2) in Section 5.3, for a stationary process, this implies that an ARMA model is identified for the process, with $K = \max\{p,q\}$.

As can be seen, in the notation of Tsay and Tiao (1985), the methods of Akaike and Cooper and Wood represent canonical correlation analysis between $\boldsymbol{Y}_{m,t}$ and $\boldsymbol{Y}_{n-1,t+n}$ for various $n = 1, 2, \ldots$. Since the Tsay and Tiao method considers canonical correlation analysis between $\boldsymbol{Y}_{m,t}$ and $\boldsymbol{Y}_{m,t-j-1}$ for various combinations of $m = 0, 1, \ldots$ and $j = 0, 1, \ldots$, it is more general and, in principle, it is capable of providing information on the orders p and q of the AR and MA parts of the model separately, rather than just the maximum of these two values. In practice, when using the methods of Akaike and Cooper and Wood,

the more detailed information on the individual orders p and q would be determined at the stage of maximum likelihood estimation of the parameters of the ARMA(K, K) model.

Use of Model Selection Criteria. Another approach to model selection involves the use of information criteria such as AIC proposed by Akaike (1974a) or the Bayesian information criteria of Schwarz (1978). In the implementation of this approach, a range of potential ARMA models are estimated by maximum likelihood methods to be discussed in Chapter 7, and for each model, a criterion such as AIC (normalized by sample size n), given by

$$\text{AIC}_{p,q} = \frac{-2\ln(\text{maximized likelihood}) + 2r}{n} \approx \ln(\hat{\sigma}_a^2) + r\frac{2}{n} + \text{ constant}$$

or the related BIC given by

$$\text{BIC}_{p,q} = \ln(\hat{\sigma}_a^2) + r\frac{\ln(n)}{n}$$

is evaluated. Here, $\hat{\sigma}_a^2$ is the maximum likelihood estimate of σ_a^2, and $r = p + q + 1$ is the number of estimated parameters, including a constant term. In the above criteria, the first term essentially corresponds to $-2/n$ times the log of the maximized likelihood, while the second term is a ''penalty factor'' for inclusion of additional parameters in the model. In the information criteria approach, models that yield a minimum value for the criterion are to be preferred, and the AIC or BIC values are compared among various models as the basis for selection of the model. Hence, since the BIC criterion imposes a greater penalty for the number of estimated model parameters than does AIC, use of minimum BIC for model selection would always result in a chosen model whose number of parameters is no greater than that chosen under AIC.

Hannan and Rissanen (1982) proposed a two-step model selection procedure that avoids the need to maximize the likelihood function for multiple combinations of p and q. At the first step, one fits an AR model of sufficiently high order m^* to the series $\tilde{z}_t$. The residuals $\tilde{a}_t$ from the fitted AR(m^*) model provide estimates of the innovations a_t in the ARMA(p, q) model. At the second step, one regresses $\tilde{z}_t$ on $\tilde{z}_{t-1}, \ldots, \tilde{z}_{t-p}$ and $\tilde{a}_{t-1}, \ldots, \tilde{a}_{t-q}$, for various combinations of p and q. That is, one fits approximate models of the form

$$\tilde{z}_t = \sum_{j=1}^{p} \phi_j \tilde{z}_{t-j} - \sum_{j=1}^{q} \theta_j \tilde{a}_{t-j} + a_t \tag{6.2.7}$$

using ordinary least squares, and the estimated error variance, uncorrected for degrees of freedom, is denoted by $\hat{\sigma}_{p,q}^2$. Then, using the BIC criterion, the order (p, q) of the ARMA model is chosen as the one that minimizes $\ln(\hat{\sigma}_{p,q}^2) + (p + q)\ln(n)/n$. Hannnan and Rissanen show that, under very general conditions, the estimators of p and q chosen in this manner tend almost surely to the true values. The appeal of this procedure is that computation of maximum likelihood estimates over a wide range of possible ARMA models is avoided.

While these order selection procedures are useful, they should be viewed as supplementary tools to assist in the model selection process. In particular, they should not be used as a substitute for careful examination of the estimated autocorrelation and partial autocorrelation functions of the series, and critical examination of the residuals $\hat{a}_t$ from

a fitted model should always be included as a major part of the overall model selection process.

6.3 INITIAL ESTIMATES FOR THE PARAMETERS

6.3.1 Uniqueness of Estimates Obtained from the Autocovariance Function

While a given ARMA model has a unique autocovariance structure, the converse is not true without additional conventions imposed for uniqueness, as we discuss subsequently. At first sight this would seem to rule out the use of the estimated autocovariances as a means of identification. However, we show in Section 6.4 that the estimated autocovariance function may indeed be used for this purpose. The reason is that, although there exists a multiplicity of ARMA models possessing the same autocovariance function, there exists only one that expresses the current value of $w_t = \nabla^d z_t$, exclusively in terms of previous history and in stationary invertible form.

6.3.2 Initial Estimates for Moving Average Processes

As shown in Chapter 3, the first q autocorrelations of a MA(q) process are nonzero and can be written in terms of the parameters of the model as

$$\rho_k = \frac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2} \qquad k = 1, 2, \ldots, q \tag{6.3.1}$$

The expression (6.3.1) for $\rho_1, \rho_2, \ldots, \rho_q$, in terms of $\theta_1, \theta_2, \ldots, \theta_q$, supplies q equations in q unknowns. Preliminary estimates of the θ's can be obtained by substituting the estimates r_k for ρ_k in (6.3.1) and solving the resulting nonlinear equations. A preliminary estimate of σ_a^2 may then be obtained from

$$\gamma_0 = \sigma_a^2(1 + \theta_1^2 + \cdots + \theta_q^2)$$

by substituting the preliminary estimates of the θ's and replacing $\gamma_0 = \sigma_w^2$ by its estimate c_0. The numerical values of the estimated autocorrelation coefficients r_k for the series `z` are conveniently obtained from R as follows:

```
> ac=acf(z)
> ac
```

Preliminary Estimates for a (0, d, 1) Process. Table A in Part Five relates ρ_1 to θ_1, and by substituting $r_1(w)$ for ρ_1 can be used to provide initial estimates for any $(0, d, 1)$ process $w_t = (1 - \theta_1 B)a_t$, where $w_t = \nabla^d z_t$.

Preliminary Estimates for a (0, d, 2) Process. Chart C in Part Five relates ρ_1 and ρ_2 to θ_1 and θ_2, and by substituting $r_1(w)$ and $r_2(w)$ for ρ_1 and ρ_2 can be used to provide initial estimates for any $(0, d, 2)$ process.

In obtaining preliminary estimates in this way, the following points should be kept in mind:

1. The autocovariances are second moments of the joint distribution of the w's. Thus, the parameter estimates are obtained by equating sample moments to their theoretical values. It is well known that the *method of moments* is not necessarily efficient and can produce poor estimates for models that include moving average terms. However, the rough estimates obtained can be useful in obtaining fully efficient estimates, because they supply an approximate idea of ''where in the parameter space to look'' for the most efficient estimates.
2. In general, the equation (6.3.1), obtained by equating moments, will have multiple solutions. For instance, when $q = 1$,

$$\rho_1 = \frac{-\theta_1}{1 + \theta_1^2} \tag{6.3.2}$$

and hence from $\theta_1^2 + (1/\rho_1)\theta_1 + 1 = 0$, we see that both

$$\theta_1 = -\frac{1}{2\rho_1} + \left[\frac{1}{(2\rho_1)^2} - 1\right]^{1/2}$$

and

$$\theta_1 = -\frac{1}{2\rho_1} - \left[\frac{1}{(2\rho_1)^2} - 1\right]^{1/2} \tag{6.3.3}$$

are possible solutions. For illustration, the first lag autocorrelation of the first difference of Series A is about -0.4. Substitution in (6.3.3) yields the pair of solutions $\theta_1 \simeq 0.5$ and $\theta_1' \simeq 2.0$. However, the chosen value $\theta_1 \simeq 0.5$ is the only value that lies within the invertibility interval $-1 < \theta_1 < 1$. In fact, it is shown in Section 6.4.1 that it is always true that only one of the multiple solutions of (6.3.1) can satisfy the invertibility condition.

Examples. Series A, B, and D were all identified in Table 6.2 as possible IMA processes of order (0, 1, 1). We have seen in Section 4.3.1 that this model may be written in following the alternative forms:

$$\begin{aligned}
\nabla z_t &= (1 - \theta_1 B)a_t \\
\nabla z_t &= \lambda_0 a_{t-1} + \nabla a_t \qquad (\lambda_0 = 1 - \theta_1) \\
z_t &= \lambda_0 \sum_{j=1}^{\infty} (1 - \lambda_0)^{j-1} z_{t-j} + a_t
\end{aligned}$$

Using Table A in Part Five, the approximate estimates of the parameters shown in Table 6.3 were obtained.

Series C has been tentatively specified in Table 6.2 as an IMA(0, 2, 2) process:

$$\nabla^2 z_t = (1 - \theta_1 B - \theta_2 B^2)a_t$$

or equivalently,

$$\nabla^2 z_t = (\lambda_0 \nabla + \lambda_1)a_{t-1} + \nabla^2 a_t$$

TABLE 6.3 Initial Estimates of Parameters for Series A, B, and D

Series	r_1	$\hat{\theta}_1$	$\hat{\lambda}_0 = 1 - \hat{\theta}_1$
A	−0.41	0.5	0.5
B	0.09	−0.1	1.1
D	−0.05	0.1	0.9

Since the first two sample autocorrelations of $\nabla^2 z_t$ are very close to zero, Chart C in Part Five gives $\hat{\theta}_1 = 0, \hat{\theta}_2 = 0$, so that $\hat{\lambda}_0 = 1 + \hat{\theta}_2 = 1$ and $\hat{\lambda}_1 = 1 - \hat{\theta}_1 - \hat{\theta}_2 = 1$. On this basis, the series would be represented by

$$\nabla^2 Z_t = a_t \tag{6.3.4}$$

This would mean that the second difference, $\nabla^2 z_t$, was very nearly a random (white noise) series.

6.3.3 Initial Estimates for Autoregressive Processes

For an assumed AR process of order 1 or 2, initial estimates for ϕ_1 and ϕ_2 can be calculated by substituting estimates r_j for the theoretical autocorrelations ρ_j in the formulas of Table 6.1, which are obtained from the Yule–Walker equations (3.2.6). In particular, for an AR(1), $\hat{\phi}_{11} = r_1$, and for an AR(2),

$$\begin{aligned} \hat{\phi}_{21} &= \frac{r_1(1 - r_2)}{1 - r_1^2} \\ \hat{\phi}_{22} &= \frac{r_2 - r_1^2}{1 - r_1^2} \end{aligned} \tag{6.3.5}$$

where $\hat{\phi}_{pj}$ denotes the estimated jth autoregressive parameter in a process of order p. The corresponding formulas given by the Yule–Walker equations for higher order schemes may be obtained by substituting the r_j for the ρ_j in (3.2.7). Thus,

$$\hat{\boldsymbol{\phi}} = \mathbf{R}_p^{-1} \mathbf{r}_p \tag{6.3.6}$$

where $\mathbf{R}_p$ is an estimate of the $p \times p$ matrix $\mathbf{P}_p$, as depicted following (3.2.6) in 3.2.2, of autocorrelations up to order $p - 1$, and $\mathbf{r}_p = (r_1, r_2, \ldots, r_p)'$. For example, if $p = 3$, (6.3.6.) becomes

$$\begin{bmatrix} \hat{\phi}_{31} \\ \hat{\phi}_{32} \\ \hat{\phi}_{33} \end{bmatrix} = \begin{bmatrix} 1 & r_1 & r_2 \\ r_1 & 1 & r_1 \\ r_2 & r_1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} \tag{6.3.7}$$

A simple recursive method due to Levinson and Durbin for obtaining the estimates for an AR(p) from those of an AR($p - 1$) was discussed in Appendix A3.2.

It will be shown in Chapter 7 that in contrast to the situation for MA processes, the autoregressive parameters obtained from (6.3.6) approximate the fully efficient maximum likelihood estimates.

Example. Series E representing the sunspot data behaves in its undifferenced form like[1] an autoregressive process of second order:

$$(1 - \phi_1 B - \phi_2 B^2)\tilde{z}_t = a_t$$

Substituting the estimates $r_1 = 0.81$ and $r_2 = 0.43$, obtained using R, into (6.3.5), we have $\hat{\phi}_1 = 1.32$ and $\hat{\phi}_2 = -0.63$.

As a second example, consider again Series C identified as either of order $(1, 1, 0)$ or possibly $(0, 2, 2)$. The first possibility would give

$$(1 - \phi_1 B)\nabla Z_t = a_t$$

with $\hat{\phi}_1 = 0.81$, since r_1 for ∇z_t is 0.81.

This example is interesting because it makes clear that the two alternative models that have been identified for this series are closely related. On the supposition that the series is of order $(0, 2, 2)$, we found in (6.3.4) that this simplifies to

$$(1 - B)(1 - B)z_t = a_t \tag{6.3.8}$$

The alternative

$$(1 - 0.81B)(1 - B)z_t = a_t \tag{6.3.9}$$

is very similar.

6.3.4 Initial Estimates for Mixed Autoregressive–Moving Average Processes

It is often found, either initially or after suitable differencing, that $w_t = \nabla^d z_t$ is most economically represented by a mixed ARMA process:

$$\phi(B)w_t = \theta(B)a_t$$

As noted in Section 6.2.1, a mixed process is indicated if both the autocorrelation and partial autocorrelation functions tail off rather than either having a cutoff feature. Another helpful fact in identifying the mixed process is that after lag $q - p$, the theoretical autocorrelations of the mixed process behave like the autocorrelations of a pure autoregressive process $\phi(B)w_t = a_t$ (see (3.4.3)). In particular, if the autocorrelation function of the dth difference appears to be falling off exponentially from an aberrant first value r_1, we would suspect that we have a process of order $(1, d, 1)$ that is,

$$(1 - \phi_1 B)w_t = (1 - \theta_1 B)a_t \tag{6.3.10}$$

where $w_t = \nabla^d z_t$.

[1]The sunspot series has been the subject of much investigation. Early references include Schuster (1906), Yule (1927), and Moran (1954). The series does not appear to be adequately represented by a second-order autoregressive process. A model related to the underlying mechanism at work would, of course, be the most satisfactory. More recent work has suggested empirically that a second-order autoregressive model would provide a better fit if a suitable transformation such as log or square root were first applied to z. Inclusion of a higher order term, at lag 9, in the AR model also improves the fit. Other possibilities include the use of nonlinear time series models, such as bilinear or threshold autoregressive models (e.g., see Section 10.3), as has been investigated by Subba Rao and Gabr (1984), Tong and Lim (1980), and Tong (1983,1990).

Approximate values for the parameters of the process (6.3.10) are obtained by substituting the estimates $r_1(w)$ and $r_2(w)$ for ρ_1 and ρ_2 in the expression (3.4.8). This gives

$$r_1 = \frac{(1 - \hat{\phi}_1\hat{\theta}_1)(\hat{\phi}_1 - \hat{\theta}_1)}{1 + \hat{\theta}_1^2 - 2\hat{\phi}_1\hat{\theta}_1}$$

$$r_2 = r_1\hat{\phi}_1$$

Chart D in Part Five relates ρ_1 and ρ_2 to ϕ_1 and θ_1 can be used to provide initial estimates of the parameters for any $(1, d, 1)$ process.

For example, using Figure 6.2, Series A was identified as of order $(0, 1, 1)$, with θ_1 about 0.5. Looking at the autocorrelation function of z_t rather than that of $w_t = \nabla z_t$, we see from r_1 onward the autocorrelations decay roughly exponentially, although slowly. Thus, an alternative specification for Series A is that it is generated by a stationary process of order $(1, 0, 1)$. The estimated autocorrelations and the corresponding initial estimates of the parameters are then

$$r_1 = 0.57 \qquad r_2 = 0.50 \qquad \hat{\phi}_1 \simeq 0.87 \qquad \hat{\theta}_1 \simeq 0.48$$

This identification yields the approximate model of order $(1, 0, 1)$:

$$(1 - 0.9B)\tilde{z}_t = (1 - 0.5B)a_t$$

whereas the previously identified model of order $(0, 1, 1)$, given in Table 6.5, is

$$(1 - B)z_t = (1 - 0.5B)a_t$$

Again we see that the ''alternative'' models are nearly the same.

Compensation between Autoregressive and Moving Average Operators. The alternative models identified above are even more alike than they appear. This is because small changes in the autoregressive operator of a mixed model can be nearly compensated by corresponding changes in the moving average operator. In particular, if we have a model

$$[1 - (1 - \delta)B]\tilde{z}_t = (1 - \theta B)a_t$$

where δ is small and positive, we can write

$$\begin{aligned}(1 - B)\tilde{z}_t &= [1 - (1 - \delta)B]^{-1}(1 - B)(1 - \theta B)a_t \\ &= \{1 - \delta B[1 + (1 - \delta)B + (1 - \delta)^2B^2 + \cdots]\}(1 - \theta B)a_t \\ &= [1 - (\theta + \delta)B]a_t + \text{terms in } a_{t-2}, a_{t-3}, \cdots, \text{of order } \delta\end{aligned}$$

6.3.5 Initial Estimate of Error Variance

For comparison with the more efficient methods of estimation to be described in Chapter 7, it is interesting to see how much additional information about the model can be extracted at the identification stage. We have already shown how to obtain initial estimates $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ of the parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$ in the ARMA model, identified for an appropriate difference $w_t = \nabla^d z_t$ of the series. In this section we show how to obtain preliminary estimates of the error variance σ_a^2, and in Section 6.3.6 we show how to obtain an approximate standard error for the sample mean $\overline{w}$ of the appropriately differenced series.

An initial estimate of the error variance may be obtained by substituting an estimate c_0 in the expression for the variance γ_0 given in Chapter 3. Thus, substituting in (3.2.8), an initial estimate of σ_a^2 for an AR process may be obtained from

$$\hat{\sigma}_a^2 = c_0(1 - \hat{\phi}_1 r_1 - \hat{\phi}_2 r_2 - \cdots - \hat{\phi}_p r_p) \tag{6.3.11}$$

Similarly, from (3.3.3), an initial estimate for a MA process may be obtained from

$$\hat{\sigma}_a^2 = \frac{c_0}{1 + \hat{\theta}_1^2 + \cdots + \hat{\theta}_q^2} \tag{6.3.12}$$

The form of the estimate for a mixed process is, in general, more complicated. However, for the important ARMA(1,1) process, it takes the form (see (3.4.7))

$$\hat{\sigma}_a^2 = \frac{1 - \hat{\phi}_1^2}{1 + \hat{\theta}_1^2 - 2\hat{\phi}_1\hat{\theta}_1} c_0 \tag{6.3.13}$$

For example, consider the (1, 0, 1) model identified for Series A. Using (6.3.13) with $\hat{\phi}_1 = 0.87$, $\hat{\theta}_1 = 0.48$, and $c_0 = 0.1586$, we obtain the estimate $\hat{\sigma}_a^2 = 0.098$.

6.3.6 Approximate Standard Error for $\overline{w}$

The general ARIMA model, for which the mean μ_w of $w_t = \nabla^d z_t$ is not necessarily zero, may be written in any one of the three forms:

$$\phi(B)(w_t - \mu_w) = \theta(B)a_t \tag{6.3.14}$$

$$\phi(B)w_t = \theta_0 + \theta(B)a_t \tag{6.3.15}$$

$$\theta(B)w_t = \theta(B)(a_t + \xi) \tag{6.3.16}$$

where

$$\mu_w = \frac{\theta_0}{1 - \phi_1 - \phi_2 - \cdots - \phi_p} = \frac{(1 - \theta_1 - \theta_2 - \cdots - \theta_q)\xi}{1 - \phi_1 - \phi_2 - \cdots - \phi_p}$$

Hence, if $1 - \phi_1 - \phi_2 - \cdots - \phi_p \neq 0$ and $1 - \theta_1 - \theta_2 - \cdots - \theta_p \neq 0$, $\mu_w = 0$ implies that $\theta_0 = 0$ and $\xi = 0$. Now, in general, when $d = 0$, μ_z will not be zero. However, consider the eventual forecast function associated with the general model (6.3.14) when $d > 0$. With $\mu_w = 0$, this forecast function already contains an adaptive polynomial component of degree $d - 1$. The effect of allowing μ_w to be nonzero is to introduce a *fixed* polynomial term into this function of degree d. For example, if $d = 2$ and μ_w is nonzero, the forecast function $\hat{z}_t(l)$ includes a quadratic component in l, in which the coefficient of the quadratic term is fixed and does not adapt to the series. Because models of this kind are often inapplicable when $d > 0$, the hypothesis that $\mu_w = 0$ will frequently not be contradicted by the data. Indeed, as we have indicated, we usually assume that $\mu_w = 0$ unless evidence to the contrary presents itself.

At this, the identification stage of model building, an indication of whether or not a nonzero value for μ_w is needed may be obtained by comparison of $\overline{w} = \sum_{t=1}^{n} w_t/n$ with

its approximate standard error (see Section 2.1.5). With $n = N - d$ differences available,

$$\sigma^2(\overline{w}) = n^{-1}\gamma_0 \sum_{-\infty}^{\infty} \rho_j = n^{-1} \sum_{-\infty}^{\infty} \gamma_j$$

that is,

$$\sigma^2(\overline{w}) = n^{-1}\gamma(1) \tag{6.3.17}$$

where $\gamma(B)$ is the autocovariance generating function defined in (3.1.10) and $\gamma(1)$ is its value when $B = B^{-1} = 1$ is substituted.

For illustration, consider the process of order $(1, d, 0)$:

$$(1 - \phi B)(w_t - \mu_w) = a_t$$

with $w_t = \nabla^d z_t$. From (3.1.11), we obtain

$$\gamma(B) = \frac{\sigma_a^2}{(1 - \phi B)(1 - \phi F)}$$

so

$$\sigma^2(\overline{w}) = n^{-1}(1 - \phi)^{-2}\sigma_a^2$$

But $\sigma_a^2 = \sigma_w^2(1 - \phi^2)$, so

$$\sigma^2(\overline{w}) = \frac{\sigma_w^2}{n}\frac{1 - \phi^2}{(1 - \phi)^2} = \frac{\sigma_w^2}{n}\frac{1 + \phi}{1 - \phi}$$

and

$$\sigma(\overline{w}) = \sigma_w \left[\frac{1 + \phi}{n(1 - \phi)}\right]^{1/2}$$

Now ϕ and σ_w^2 are estimated by r_1 and c_0, respectively, as defined in (2.1.11) and (2.1.12). Thus, for a $(1, d, 0)$ process, the required standard error is given by

$$\hat{\sigma}(\overline{w}) = \left[\frac{c_0(1 + r_1)}{n(1 - r_1)}\right]^{1/2}$$

Proceeding in this way, the expressions for $\sigma(\overline{w})$ given in Table 6.4 may be obtained.

Tentative Identification of Models A–F. Table 6.5 summarizes the models tentatively identified for Series A to F, with the preliminary parameter estimates inserted. These parameter values are used as initial guesses for the more efficient estimation methods to be described in Chapter 7.

6.3.7 Choice Between Stationary and Nonstationary Models in Doubtful Cases

As the results in Tables 6.2 and 6.5 suggest, the preliminary identification of the need for differencing and of the degree of differencing is not always easily determined. The apparent ambiguity in identifying models for Series A, C, and D (particularly with regard

TABLE 6.4 Approximate Standard Error for $\overline{w}$, where $w_t = \nabla^d z_t$ and z_t is an ARIMA Process of Order (p, d, q)

$(1, d, 0)$	$(0, d, 1)$
$\left[\frac{c_0(1+r_1)}{n(1-r_1)}\right]^{1/2}$	$\left[\frac{c_0(1+2r_1)}{n}\right]^{1/2}$
$(2, d, 0)$	$(0, d, 2)$
$\left[\frac{c_0(1+r_1)(1-2r_1^2+r_2)}{n(1-r_1)(1-r_2)}\right]^{1/2}$	$\left[\frac{c_0(1+2r_1+2r_2)}{n}\right]^{1/2}$
$(1, d, 1)$	
$\left[\frac{c_0}{n}\left(1+\frac{2r_1^2}{r_1-r_2}\right)\right]^{1/2}$	

to the degree of differencing) is, of course, more apparent than real. It arises whenever the roots of $\phi(B) = 0$ approach unity. When this happens, it becomes less and less important whether a root near unity is included in $\phi(B)$ or an additional difference is included corresponding to a unit root. A more precise evaluation is possible using the estimation procedures discussed in Chapter 7 and, in particular, the more formal unit root testing procedures to be discussed in Chapter 10. However, the following should be kept in mind:

1. From time series that are necessarily of finite length, it is never possible to *prove* that a zero of the autoregressive operator is exactly equal to unity.
2. There is, of course, no sudden transition from stationary behavior to nonstationary behavior. This can be understood by considering the behavior of the simple mixed

TABLE 6.5 Summary of Models Identified for Series A–F, with Initial Estimates Inserted

Series	Differencing	$\overline{w} \pm \hat{\sigma}(\overline{w})^a$	$\hat{\sigma}_w^2 = c_0$	Identified Model	$\hat{\sigma}_a^2$
A	Either 0	17.06 ± 0.10	0.1586	$z_t - 0.87z_{t-1} = 2.45 + a_t - 0.48a_{t-1}$	0.098
	or 1	0.002 ± 0.011	0.1364	$\nabla z_t = a_t - 0.53a_{t-1}$	0.107
B	1	-0.28 ± 0.41	52.54	$\nabla z_t = a_t + 0.09a_{t-1}$	52.2
C	Either 1	-0.035 ± 0.047	0.0532	$\nabla z_t - 0.81\nabla z_{t-1} = a_t$	0.019
	or 2	-0.003 ± 0.008	0.0198	$\nabla^2 z_t = a_t - 0.09a_{t-1} - 0.07a_{t-2}$	0.020
D	Either 0	9.13 ± 0.04	0.3620	$z_t - 0.86z_{t-1} = 1.32 + a_t$	0.093
	or 1	0.004 ± 0.017	0.0965	$\nabla z_t = a_t - 0.05a_{t-1}$	0.096
E	Either 0	46.9 ± 5.4	1382.2	$z_t - 1.32z_{t-1} + 0.63z_{t-2} = 14.9 + a_t$	289.0
	or 0	46.9 ± 5.4	1382.2	$z_t - 1.37z_{t-1} + 0.74z_{t-2} - 0.08z_{t-3} = 13.7 + a_t$	287.0
F	0	51.1 ± 1.1	139.80	$z_t + 0.32z_{t-1} - 0.18z_{t-2} = 58.3 + a_t$	115.0

[a] When $d = 0$, read z for w.

model

$$(1 - \phi_1 B)(z_t - \mu) = (1 - \theta_1 B)a_t$$

Series generated by such a model behave in a more nonstationary manner as ϕ_1 increases toward unity. For example, a series with $\phi_1 = 0.99$ can wander away from its mean μ and not return for very long periods. It is as if the attraction that the mean exerts in the series becomes less and less as ϕ_1 approaches unity, and finally, when ϕ_1 is equal to unity, the behavior of the series is completely independent of μ.

In doubtful cases, there may be an advantage in employing the nonstationary model rather than the stationary alternative (e.g., in treating a ϕ_1, whose estimate is close to unity, as being *equal* to unity). This is particularly true in forecasting and control problems. Where ϕ_1 is close to unity, we do not really know whether the mean of the series has meaning or not. Therefore, it may be advantageous to employ the nonstationary model, which does not include a fixed mean μ. If we use such a model, forecasts of future behavior will not in any way depend on an estimated mean, calculated from a previous period, which may have no relevance to the future level of the series.

6.4 MODEL MULTIPLICITY

6.4.1 Multiplicity of Autoregressive–Moving Average Models

With the normal distribution assumption, knowledge of the first and second moments of a probability distribution implies complete knowledge of the distribution. In particular, knowledge of the mean of $w_t = \nabla^d z_t$ and of its autocovariance function uniquely determines the probability structure or w_t. We now show that although this unique probability structure can be represented by a *multiplicity* of linear ARMA models, uniqueness is achieved in the model when we introduce the appropriate stationarity and invertibility restrictions.

Suppose that w_t, having autocovariance generating function $\gamma(B)$, is represented by the linear ARMA model

$$\phi(B)w_t = \theta(B)a_t \tag{6.4.1}$$

where the zeros of $\phi(B)$ and of $\theta(B)$ lie outside the unit circle. Then, this model may also be written as

$$\prod_{i=1}^{p}(1 - G_i B)w_t = \prod_{j=1}^{q}(1 - H_j B)a_t \tag{6.4.2}$$

where the G^{-1} are the roots of $\phi(B) = 0$ and H_j^{-1} are the roots of $\theta(B) = 0$, and G_i, H_j lie inside the unit circle. Using (3.1.11), the autocovariance generating function for w is

$$\gamma(B) = \prod_{i=1}^{p}(1 - G_i B)^{-1}(1 - G_i F)^{-1}\prod_{j=1}^{q}(1 - H_j B)(1 - H_j F)\sigma_a^2$$

Multiple Choice of Moving Average Parameters. Since

$$(1 - H_j B)(1 - H_j F) = H_j^2(1 - H_j^{-1} B)(1 - H_j^{-1} F)$$

it follows that any one of the stochastic models

$$\prod_{i=1}^{p}(1 - G_i B)w_t = \prod_{j=1}^{q}(1 - H_j^{\pm 1} B)ka_t$$

can have the same autocovariance generating function if the constant k is chosen appropriately. In the above, it is understood that for complex roots, reciprocals of both members of the conjugate pair will be taken (so as to always obtain *real-valued* coefficients in the MA operator). However, if a real root H is inside the unit circle, H^{-1} will lie outside, or if a complex pair, say H_1 and H_2, are inside, then the pair H_1^{-1} and H_2^{-1} will lie outside. It follows that there will be only *one stationary invertible* model of the form (6.4.2), which has a given autocovariance function.

Backward Representations. Now $\gamma(B)$ also remains unchanged if in (6.4.2) we replace $1 - G_i B$ by $1 - G_i F$ or $1 - H_j B$ by $1 - H_j F$. Thus, all the stochastic models

$$\prod_{i=1}^{p}(1 - G_i B^{\pm 1})w_t = \prod_{j=1}^{q}(1 - H_j B^{\pm 1})a_t$$

have identical autocovariance structure. However, representations containing the operator $B^{-1} = F$ refer to future w's and/or future a's, so that although stationary and invertible representations exist in which w_t is expanded in terms of future w's and a's, only one such representation, (6.4.2), exists that relates w_t entirely to *past* history.

A model form that, somewhat surprisingly, is of practical interest is that in which *all* B's are replaced by F's in (6.4.1), so that

$$\phi(F)w_t = \theta(F)e_t$$

where e_t is a sequence of independently distributed random variables having mean zero and variance $\sigma_e^2 = \sigma_a^2$. This then is a stationary invertible representation in which w_t is expressed *entirely* in terms of future w's and e's. We refer to it as the *backward* form of the process, or more simply as the *backward process*.

Equation (6.4.2) is not the most general form of a stationary invertible linear ARMA model having the autocovariance generating function $\gamma(B)$. For example, the model (6.4.2) may be multiplied on both sides by any factor $1 - QB$. Thus, the process

$$(1 - QB)\prod_{i=1}^{p}(1 - G_i B)w_t = (1 - QB)\prod_{j=1}^{q}(1 - H_j B)a_t$$

has the same autocovariance structure as (6.4.2). This fact will present no particular difficulty at the identification stage, since we will be naturally led to choose the simplest representation, and so for uniqueness we require that there be *no common factors* between the AR and MA operators in the model. However, as discussed in Chapter 7, we need to be alert to the possibility of common factors in the estimated AR and MA operators when fitting the process.

Finally, we reach the conclusion that a stationary-invertible model, in which a current value w_t is expressed only in terms of *previous* history and which contains *no common factors* between the AR and MA operators, is uniquely determined by the autocovariance structure.

Proper understanding of model multiplicity is of importance for a number of reasons:

1. We are reassured by the foregoing argument that the autocovariance function can logically be used to identify a linear stationary-invertible ARMA model that expresses w_t in terms of previous history.
2. The nature of the multiple solutions for moving average parameters obtained by equating moments is clarified.
3. The backward process

$$\phi(F)w_t = \theta(F)e_t$$

obtained by replacing B by F in the linear ARMA model, is useful in estimating values of the series that have occurred before the first observation was made.

Now we consider reasons 2 and 3 in greater detail.

6.4.2 Multiple Moment Solutions for Moving Average Parameters

In estimating the q parameters $\theta_1, \theta_2, \ldots, \theta_q$ in the MA model by equating autocovariances, we have seen that multiple solutions are obtained. To each combination of roots, there will be a corresponding linear representation, but to only one such combination will there be an invertible representation in terms of past history.

For example, consider the MA(1) process in w_t:

$$w_t = (1 - \theta_1 B)a_t$$

and suppose that $\gamma_0(w)$ and $\gamma_1(w)$ are known and we want to deduce the values of θ_1 and σ_a^2. Since

$$\gamma_0 = (1 + \theta_1^2)\sigma_a^2 \qquad \gamma_1 = -\theta_1\sigma_a^2 \qquad \gamma_k = 0 \qquad k > 1 \tag{6.4.3}$$

then

$$-\frac{\gamma_0}{\gamma_1} = \theta_1^{-1} + \theta_1$$

and if $(\theta_1 = \theta, \sigma_a^2 = \sigma^2)$ is a solution for given γ_0 and γ_1, so is $(\theta_1 = \theta^{-1}, \sigma_a^2 = \theta^2\sigma^2)$. Apparently, then, for given values of γ_0 and γ_1, there are a *pair* of possible models:

$$w_t = (1 - \theta B)a_t$$

and

$$w_t = (1 - \theta^{-1}B)\alpha_t \tag{6.4.4}$$

with $\text{var}[a_t] = \sigma_a^2$ and $\text{var}[\alpha_t] = \sigma_\alpha^2 = \sigma_a^2\theta^2$. If $-1 < \theta < 1$, then (6.4.4) is not an invertible representation. However, this model may be written as

$$w_t = [(1 - \theta^{-1}B)(-\theta F)](-\theta^{-1}B\alpha_t)$$

Thus, after setting $e_t = -\alpha_{t-1}/\theta$, the model becomes

$$w_t = (1 - \theta F)e_t \tag{6.4.5}$$

where e_t has the same variance as a_t. Thus, (6.4.5) is simply the ''backward'' process, which is dual to the forward process:

$$w_t = (1 - \theta B)a_t \tag{6.4.6}$$

Just as the shock a_t in (6.4.6) is expressible as a convergent sum of current and *previous* values of w,

$$a_t = w_t + \theta w_{t-1} + \theta^2 w_{t-2} + \cdots$$

the shock e_t in (6.4.5) is expressible as a convergent sum of current and future values of w:

$$e_t = w_t + \theta w_{t+1} + \theta^2 w_{t+2} + \cdots$$

Thus, the root θ^{-1} *would* produce an ''invertible'' process, but only if a representation of the shock e_t in terms of future values of w were permissible. The invertibility regions shown in Table 6.1 delimit acceptable values of the parameters, *given* that we express the shock in terms of *previous* history.

6.4.3 Use of the Backward Process to Determine Starting Values

Suppose that a time series $w_1, w_2, \ldots, w_n$ is available from a process

$$\phi(B)w_t = \theta(B)a_t \tag{6.4.7}$$

In Chapter 7, problems arise where we need to estimate the values w_0, w_{-1}, w_{-2}, and so on, of the series that occurred *before* the first observation was made. This happens because ''starting values'' are needed for certain basic recursive calculations used for estimating the parameters in the model. Now, suppose that we require to estimate w_{-l}, given $w_1, \ldots, w_n$. The discussion of Section 6.4.1 shows that the probability structure of $w_1, \ldots, w_n$ is equally explained by the forward model (6.4.7), or by the backward model

$$\phi(F)w_t = \theta(F)e_t \tag{6.4.8}$$

The value w_{-l}, thus, bears exactly the same probability relationship to the sequence $w_1, w_2, \ldots, w_n$, as does the value w_{n+l+1} to the sequence $w_n, w_{n-1}, w_{n-2}, \ldots, w_1$. Thus, to estimate a value $l + 1$ periods before observations started, we can first consider what would be the optimal estimate or forecast $l + 1$ periods after the series

ended, and then apply this procedure to the *reversed* series. In other words, we ''forecast'' the reversed series. We call this ''back forecasting.''

APPENDIX A6.1 EXPECTED BEHAVIOR OF THE ESTIMATED AUTOCORRELATION FUNCTION FOR A NONSTATIONARY PROCESS

Suppose that a series of N observations $z_1, z_2, \ldots, z_N$ is generated by a nonstationary $(0, 1, 1)$ process

$$\nabla z_t = (1 - \theta B)a_t$$

and the estimated autocorrelations r_k are computed, where

$$r_k = \frac{c_k}{c_0} = \frac{\sum_{t=1}^{N-k}(z_t - \bar{z})(z_{t+k} - \bar{z})}{\sum_{t=1}^{N}(z_t - \bar{z})^2}$$

Some idea of the behavior of these estimated autocorrelations may be obtained by deriving expected values for the numerator and denominator of this expression and considering the ratio. We will write, following Wichern (1973),

$$\begin{aligned}\mathcal{E}[r_k] &= \frac{E[c_k]}{E[c_0]} \\ &= \frac{\sum_{t=1}^{N-k} E[(z_t - \bar{z})(z_{t+k} - \bar{z}]}{\sum_{t=1}^{N} E[(z_t - \bar{z})^2]}\end{aligned}$$

After straightforward but tedious algebra, we find that

$$\mathcal{E}[r_k] = \frac{(N-k)[(1-\theta)^2(N^2 - 1 + 2k^2 - 4kN) - 6\theta]}{N(N-1)[(N+1)(1-\theta)^2 + 6\theta]} \tag{A6.1.1}$$

For θ close to zero, $\mathcal{E}[r_k]$ will be close to unity, but for large values of θ, it can be considerably smaller than unity, even for small values of k. Figure A6.1 illustrates this fact by showing values of $\mathcal{E}[r_k]$ for $\theta = 0.8$ with $N = 100$ and $N = 200$. Although, as anticipated for a nonstationary process, the ratios $\mathcal{E}[r_k]$ of expected values fail to damp out quickly, it will be seen that they do not approach the value 1 even for small lags.

Similar effects may be demonstrated whenever the parameters approach values where cancellation on both sides of the model would produce a stationary process. For instance, in the example above we can write the model as

$$(1 - B)z_t = [(1 - B) + \delta B]a_t$$

where $\delta = 0.2$. As δ tends to zero, the behavior of the process would be expected to come closer and closer to that of the white noise process $z_t = a_t$, for which the autocorrelation function is zero for lags $k > 0$.

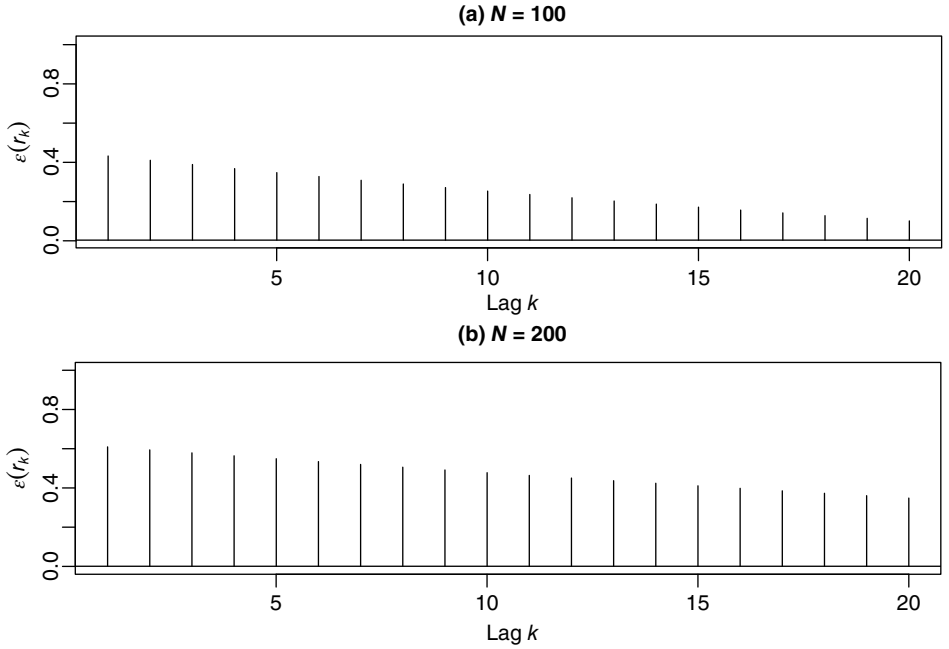


FIGURE A6.1 $\mathcal{E}[r_k] = E[c_k]/E[c_0]$ for series generated by $\nabla z_t = (1 - 0.8B)a_t$.

EXERCISES

6.1. Given the five identified models and the corresponding values of the estimated autocorrelations of $w_t = \nabla^d z_t$ in the following table:

	Identified Model			
	p	d	q	Estimated Autocorrelations
(1)	1	1	0	$r_1 = 0.72$
(2)	0	1	1	$r_1 = -0.41$
(3)	1	0	1	$r_1 = 0.40, r_2 = 0.32$
(4)	0	2	2	$r_1 = 0.62, r_2 = 0.13$
(5)	2	1	0	$r_1 = 0.93, r_2 = 0.81$

(a) Obtain preliminary estimates of the parameters analytically.

(b) Check these estimates using the charts and tables in Part Five of the book.

(c) Write down the identified models in backward shift operator notation with the preliminary estimates inserted.

6.2. For the (2, 1, 0) process considered on line (5) of Exercise 6.1, the sample mean and variance of $w_t = \nabla z_t$ are $\overline{w} = 0.23$ and $s_w^2 = 0.25$. If the series contains $N = 101$ observations,

(a) show that a constant term needs to be included in the model,

(b) express the model in the form $w_t - \phi_1 w_{t-1} - \phi_2 w_{t-2} = \theta_0 + a_t$ with numerical values inserted for the parameters, including an estimate of σ_a^2.

6.3. Consider the chemical process temperature readings referred to as Series C in this book.

(a) Plot the original series and the series of first differences using R.

(b) Use the R package to calculate and plot the ACF and PACF of this series. Repeat the calculation for the first and second differences of the series.

(c) Specify a suitable model, or models, for this series. Use the method of moments to obtain preliminary parameter estimates for the series.

6.4. Quarterly measurements of the gross domestic product (GDP) in the United Kingdom over the period 1955–1969 are included in Series P in Part Five of this book.

(a) Calculate and plot the ACF and PACF of this series.

(b) Repeat the analysis in part (a) for the first differences of the series.

(c) Identify a model for the series. Would a log transformation of the data be helpful?

(d) Obtain preliminary estimates for the parameters and for their standard errors.

(e) Obtain preliminary estimates for μ_z and σ_a^2.

6.5. Quarterly UK unemployment rate (in thousands) is part of Series P analyzed in Exercise 6.4. Repeat parts (a) to (e) of Exercise 6.4 for this series.

6.6. A time series defined by $z_t = 1000 \log_{10}(H_t)$, where H_t is the price of hogs recorded annually by the U.S. Census of Agriculture on January 1 for each of the 82 years, from 1867 to 1948 is listed as Series Q in the Collection of Time Series in Part Five. This is a well-known time series analyzed by Quenouille (1957), and others.

(a) Plot the series. Compute and plot the ACF and PACF of the series.

(b) Identify a time series model for the series.

6.7. Measurements of the annual flow of the river Nile at Ashwan from 1871 to 1970 are available as series ''Nile'' in the datasets package in R; type help(Nile) for details.

(a) Plot the series and compute the ACF and PACF for the series.

(b) Repeat the analysis in part (a) for the differenced series.

(c) Identify a model for the series. Are there any unusual features worth noting.

6.8. The file ''EuStockMarkets'' in the R datasets package contains the daily closing prices of four major European stock indices: Germany DAX (Ibis), Switzerland SMI, France CAC, and UK FTSE. The data are sampled in business time, so weekends and holidays are omitted.

(a) Plot each of the four series and compute the ACF and PACF for the series.

(b) Repeat the analysis in part (a) for the differenced series.

(c) Identify a model for the series. Are there any unusual features worth noting.

6.9. Download a time series of your choice from the Internet. Plot the time series and identify a suitable model for the series.

7

PARAMETER ESTIMATION

This chapter deals with the estimation of the parameters in ARIMA models and provides a general account of likelihood and Bayesian methods for parameter estimation. It is assumed that a suitable model of this form has been selected using the model specification tools described in Chapter 6. After the parameters have been estimated, the fitted model will be subjected to diagnostic checks and goodness-of-fit tests to be described in the next chapter. As pointed out by R. A. Fisher, for tests of goodness of fit to be relevant, it is necessary that efficient use of data should have been made in the fitting process. If this is not so, inadequacy of fit may simply arise because of the inefficient fitting and not because the form of the model is inadequate. This chapter examines in detail maximum likelihood estimation under the normality assumption and describes least-squares approximations that are suitable for many series.

It is assumed that the reader is familiar with certain basic ideas in estimation theory. Appendices A7.1 and A7.2 summarize some important results in normal distribution theory and linear least-squares that are useful for this chapter. Throughout the chapter, bold type is used to denote vectors and matrices. Thus, $\mathbf{X} = \{x_{ij}\}$ is a matrix with x_{ij} an element in the ith row and jth column, and $\mathbf{X}'$ is the transpose of the matrix $\mathbf{X}$.

7.1 STUDY OF THE LIKELIHOOD AND SUM-OF-SQUARES FUNCTIONS

7.1.1 Likelihood Function

Suppose that we have a sample of N observations, $\mathbf{z}$ with which we associate an N-dimensional random variable, whose known probability distribution $p(\mathbf{z}|\boldsymbol{\xi})$ depends on some unknown parameters $\boldsymbol{\xi}$. We use the vector $\boldsymbol{\xi}$ to denote a general set of parameters and, in particular, it could refer to the $p + q + 1$ parameters $(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2)$ of the ARIMA model.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

Before the data are available, $p(\mathbf{z}|\boldsymbol{\xi})$ will associate a density with each different outcome $\mathbf{z}$ of the experiment, for fixed $\boldsymbol{\xi}$. After the data have become available, we are led to contemplate the various values of $\boldsymbol{\xi}$ that might have given rise to the fixed set of observations $\mathbf{z}$ actually obtained. The appropriate function for this purpose is the *likelihood function* $L(\boldsymbol{\xi}|\mathbf{z})$, which is of the same form as $p(\mathbf{z}|\boldsymbol{\xi})$, but in which $\mathbf{z}$ is now fixed but $\boldsymbol{\xi}$ is variable. It is only the relative value of $L(\boldsymbol{\xi}|\mathbf{z})$ that is of interest, so that the likelihood function is usually regarded as containing an *arbitrary multiplicative constant.*

It is often convenient to work with the log-likelihood function $\ln[L(\boldsymbol{\xi}|\mathbf{z})] = l(\boldsymbol{\xi}|\mathbf{z})$, which contains an *arbitrary additive constant*. One reason that the likelihood function is of fundamental importance in estimation theory is because of the *likelihood principle*, urged on somewhat different grounds by Fisher (1956), Barnard (1949), and Birnbaum (1962). This principle says that, given that the assumed model is correct, all that the *data* have to tell us about the parameters is contained in the likelihood function, all other aspects of the data being irrelevant. From a Bayesian point of view, the likelihood function is equally important, since it is the component in the posterior distribution of the parameters that comes from the data.

For a complete understanding of the parameter estimation in a specific case, it is necessary to carefully study of the likelihood function, or in the Bayesian framework, the posterior distribution of the parameters, which in the cases we consider, is dominated by the likelihood. In many examples, for moderate and large samples, the log-likelihood function will be unimodal and can be approximated adequately over a sufficiently extensive region near the maximum by a quadratic function. In such cases, the log-likelihood function can be described by its maximum and its second derivatives at the maximum. The values of the parameters that maximize the likelihood function, or equivalently the log-likelihood function, are called *maximum likelihood (ML) estimates*. The second derivatives of the log-likelihood function provide measures of ''spread'' of the likelihood function and can be used to calculate approximate standard errors for the estimates.

The limiting properties of maximum likelihood estimates are usually established for independent observations. But as was shown by Whittle (1953), they may be extended to cover stationary time series. Other early literature on the parameter estimation in time series models includes Barnard et al. (1962), Bartlett (1955), Durbin (1960), Grenander and Rosenblatt (1957), Hannan (1960), and Quenouille (1942, 1957).

7.1.2 Conditional Likelihood for an ARIMA Process

Let us suppose that the $N = n + d$ original observations $\mathbf{z}$ form a time series that we denote by $z_{-d+1}, \ldots, z_0, z_1, z_2, \ldots, z_n$. We assume that this series is generated by an ARIMA(p, d, q) model. From these observations, we can generate a series $\mathbf{w}$ of $n = N - d$ differences $w_1, w_2, \ldots, w_n$ where $w_t = \nabla^d z_t$. Thus, the general problem of fitting the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ of the ARIMA model (6.1.1) is equivalent to fitting to the w_t's, the stationary and invertible[1] ARMA(p, q) model, which may be written as

$$a_t = \tilde{w}_t - \phi_1\tilde{w}_{t-1} - \phi_2\tilde{w}_{t-2} - \cdots - \phi_p\tilde{w}_{t-p} + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_q a_{t-q} \tag{7.1.1}$$

where $\tilde{w}_t = w_t - \mu$ are the mean-centered observations.

[1] Special care is needed to ensure that estimate lies in the invertible region. See Appendix A7.7.

For $d > 0$, it is often appropriate to assume that $\mu = 0$. When this is not appropriate, we assume that the series mean $\bar{w} = \sum_{t=1}^{n} w_t/n$ is substituted for μ. For many sample sizes common in practice, this approximation will be adequate. However, if desired, μ can be included as an additional parameter to be estimated.

The a_t's cannot be calculated immediately from (7.1.1) because of the difficulty of starting up the difference equation. However, suppose that the p values $\mathbf{w}_*$ of the w_t's and the q values $\mathbf{a}_*$ of the a_t's prior to the start of the w_t series were given. Then, for any choice of parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$, we could calculate successively a set of values $a_t(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{w}_*, \mathbf{a}_*, \mathbf{w})$, $t = 1, 2, \ldots, n$. Now, assuming that the a_t's are normally distributed, their probability density is

$$p(a_1, a_2, \ldots, a_n) \propto (\sigma_a^2)^{-n/2} \exp\left[-\left(\sum_{t=1}^{n} \frac{a_t^2}{2\sigma_a^2}\right)\right]$$

Given the data $\mathbf{w}$, the log-likelihood associated with the parameter values $(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2)$, *conditional* on the choice of $(\mathbf{w}_*, \mathbf{a}_*)$, would then be

$$l_*(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = -\frac{n}{2}\ln(\sigma_a^2) - \frac{S_*(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2} \tag{7.1.2}$$

where

$$S_*(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} a_t^2(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{w}_*, \mathbf{a}_*, \mathbf{w}) \tag{7.1.3}$$

In the above equations, a subscript asterisk is used on the likelihood and sum-of-squares functions to emphasize that they are conditional on the choice of the starting values. We notice that the conditional log-likelihood l_* involves the data only through the conditional *sum-of-squares function*. It follows that contours of l_* for any fixed value of σ_a^2 in the space of $(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2)$ are contours of S_*, that these maximum likelihood estimates are the same as the least-squares estimates, and that in general, we can, on the normal assumption, study the behavior of the conditional likelihood by studying the conditional sum-of-squares function. In particular for any fixed σ_a^2, l_* is a linear function of S_*. The parameter values obtained by minimizing the conditional sum-of-squares function $S_*(\boldsymbol{\phi}, \boldsymbol{\theta})$ will be called *conditional least-squares estimates.*

7.1.3 Choice of Starting Values for Conditional Calculation

We will shortly discuss the calculation of the unconditional likelihood, which, strictly, is what we need for parameter estimation. However, when n is moderate or large, a sufficient approximation to the unconditional likelihood is often obtained by using the conditional likelihood with suitable values substituted for the elements of $\mathbf{w}_*$ and $\mathbf{a}_*$ in (7.1.3). One procedure is to set the elements of $\mathbf{w}_*$ and of $\mathbf{a}_*$ equal to their unconditional expectations. The unconditional expectations of the elements of $\mathbf{a}_*$ are zero, and if the model contains no deterministic part, and in particular if $\mu = 0$, the unconditional expectations of the elements

TABLE 7.1 Sum-of-Squares Functions for the Model $\nabla z_t = (1 - \theta B)a_t$ Fitted to the IBM Data

θ	$\lambda = 1 - \theta$	$S_*(\theta)$	$S(\theta)$	θ	$\lambda = 1 - \theta$	$S_*(\theta)$	$S(\theta)$
−0.5	1.5	23,929	23,928	0.1	0.9	19,896	19,896
−0.4	1.4	21,595	21,595	0.2	0.8	20,851	20,851
−0.3	1.3	20,222	20,222	0.3	0.7	22,315	22,314
−0.2	1.2	19,483	19,483	0.4	0.6	24,471	24,468
−0.1	1.1	19,220	19,220	0.5	0.5	27,694	27,691
0.0	1.0	19,363	19,363				

of $\mathbf{w}_*$ will also be zero[2]. However, this approximation can be poor if some of the roots of $\phi(B) = 0$ are close to the boundary of the unit circle, so that the process approaches nonstationarity. This is also true if some of the roots of $\theta(B) = 0$ are close to the boundary of the invertibility region. Setting the presample values equal to zero could in these cases introduce a large transient, which is slow to die out. For a pure AR(p) model, a more reliable approximation procedure, and one we employ sometimes, is to use (7.1.1) to calculate the a_t's *from* a_{p+1} *onward*, thus using actual values of the w_t's throughout. Using this method, we have only $n - p = N - p - d$ values of a_t, but the slight loss of information will be unimportant for long series.

For seasonal series, discussed in Chapter 9, the conditional approximation is not always satisfactory and the unconditional calculation becomes necessary. Inclusion of the determinant in the unconditional likelihood function can also be important for seasonal time series.

Example: IMA(0, 1, 1) Process. To illustrate the recursive calculation of the conditional sum of squares S_*, we consider the IMA(0, 1, 1) model tentatively identified in Section 6.4 for the IBM data in Series B. The model is

$$\nabla z_t = (1 - \theta B)a_t \qquad -1 < \theta < 1 \tag{7.1.4}$$

so that $a_t = w_t + \theta a_{t-1}$, where $w_t = \nabla z_t$ and $E[w_t] = 0$. Thus, for the particular parameter value $\theta = 0.5$, the a_t's are calculated recursively from

$$a_t = w_t + 0.5a_{t-1}$$

setting the initial value a_0 equal to zero. Proceeding in this way, we find that

$$S_*(0.5) = \sum_{t=1}^{368} a_t^2(\theta = 0.5 | a_0 = 0) = 27,694$$

The conditional sums of squares $S_*(\theta)$ are shown in Table 7.1 for values of θ from −0.5 to +0.5 in steps of 0.1. We note that $S_*(\theta)$ has its minimum for $\theta = -0.1$. This is consistent with the preliminary moment estimate of −0.09 derived for this series in Chapter 6.

[2]If the assumption $E|w_t| = \mu \neq 0$ is appropriate, we can substitute $\bar{w}$ for each of the elements of $\mathbf{w}_*$.

7.1.4 Unconditional Likelihood, Sum-of-Squares Function, and Least-Squares Estimates

Assuming that the $N = n + d$ observations are generated by an ARIMA model, the unconditional log-likelihood is given by

$$l(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = f(\boldsymbol{\phi}, \boldsymbol{\theta}) - \frac{n}{2}\ln(\sigma_a^2) - \frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2} \tag{7.1.5}$$

where $f(\boldsymbol{\phi}, \boldsymbol{\theta})$ involves the determinant in the joint density of the w_t's and is a function of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. The *unconditional sum-of-squares function* is given by

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} [a_t|\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\theta}]^2 + [\mathbf{e}_*]'\boldsymbol{\Omega}^{-1}[\mathbf{e}_*] \tag{7.1.6}$$

where $[a_t|\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\theta}] = E[a_t|\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\theta}]$ denotes the expectation of a_t conditional on $\mathbf{w}$, $\boldsymbol{\phi}$, and $\boldsymbol{\theta}$. When the meaning is clear from the context, we will further abbreviate this conditional expectation to $[a_t]$. In (7.1.6),

$$\mathbf{e}_* = (\bar{w}_{1-p}, \ldots, \bar{w}_0, a_{1-q}, \ldots, a_0)'$$

represents the $p + q$ initial values of the $\bar{w}_t$ and a_t prior to $t = 1$, $\boldsymbol{\Omega}\sigma_a^2 = \text{cov}[\mathbf{e}_*]$ is the covariance matrix of $\mathbf{e}_*$, and

$$[\mathbf{e}_*] = ([\tilde{w}_{1-p}], \ldots, [\tilde{w}_0], [a_{1-q}], \ldots, [a_0])'$$

denotes the vector of conditional expectations (''back-forecasts'') of the initial values, given $\mathbf{w}$, $\boldsymbol{\phi}$, and $\boldsymbol{\theta}$. An alternative way to represent $S(\boldsymbol{\phi}, \boldsymbol{\theta})$ is as

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=-\infty}^{n} [a_t]^2$$

which in comparison with (7.1.6) indicates that $\Sigma_{t=-\infty}^{0}[a_t]^2 = [\mathbf{e}_*]'\boldsymbol{\Omega}^{-1}[\mathbf{e}_*]$.

Usually, $f(\boldsymbol{\phi}, \boldsymbol{\theta})$ is of importance only for small n. For moderate and large values of n, (7.1.5) is dominated by $S(\boldsymbol{\phi}, \boldsymbol{\theta})/2\sigma_a^2$, and thus the contours of the unconditional sum-of-squares function in the space of the parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$ are very nearly contours of the likelihood and log-likelihood. It follows, in particular, that the parameter estimates obtained by minimizing the sum of squares (7.1.6), which we call (*unconditional* or *exact*) *least-squares estimates*, will usually provide very close approximations to the maximum likelihood estimates. From a Bayesian viewpoint, on assumptions discussed in Section 7.5, for all AR(p) and MA(q), essentially the posterior density is a function only of $S(\boldsymbol{\phi}, \boldsymbol{\theta})$. Hence, very nearly the least-squares estimates are those with maximum posterior density. In the remainder of this section and in Section 7.1.5, the main emphasis will be on the unconditional sum-of-squares function $S(\boldsymbol{\phi}, \boldsymbol{\theta})$ in (7.1.6), and its use in calculating least-squares estimates. An alternate method for calculation of the unconditional sum of squares and likelihood functions based on the state-space model and innovations approach will be discussed in Section 7.4.

In the calculation of the unconditional sum of squares, the $[a_t]$'s are computed recursively by taking conditional expectations in (7.1.1). A preliminary back-calculation provides the

values $[w_{-j}]$ and $[a_{-j}]$, $j = 0, 1, 2, \ldots$ (i.e., the back-forecasts) needed to start off the forward recursion.

Calculation of the Unconditional Sum of Squares for a Moving Average Process. For illustration, we reconsider the IBM stock price example using only the first 10 values of the series.[3] For the IMA(0, 1, 1) model, the only back-forecast that is needed for $S(\theta)$ is $[a_0]$. We begin by describing an *approximate*, but nevertheless accurate, method to obtain $[a_0]$. Recall from Section 6.4.3 that the model for w_t may be written in either the forward or backward forms:

$$w_t = (1 - \theta B)a_t \qquad w_t = (1 - \theta F)e_t$$

and where again $\mu = E[w_t]$ is assumed equal to zero. Hence, we can write

$$[e_t] = [w_t] + \theta[e_{t+1}] \tag{7.1.7}$$

$$[a_t] = [w_t] + \theta[a_{t-1}] \tag{7.1.8}$$

where $[w_t] = w_t$ for $t = 1, 2, \ldots, n$ and is the back-forecast of w_t for $t \leq 0$. These are the two basic equations that we need in the computations. A convenient format for the calculations is shown in Table 7.2. We begin by entering in the table what we know:

1. The data values $z_0, z_1, \ldots, z_9$, from which we can calculate the first differences $w_1, w_2, \ldots, w_9$.
2. The values $[e_0], [e_{-1}], \ldots$, which are zero, since $e_0, e_{-1}, \ldots$ are distributed independently of **w**.
3. The values $[a_{-1}], [a_{-2}], \ldots$, which are zero, because for any MA(q) process, $a_{-q}, a_{-q-1}, \ldots$ are distributed independently of **w**. However, note that $[a_0], [a_{-1}], \ldots, [a_{-q+1}]$ will be nonzero and can be obtained by back-forecasting. Thus, in the present example, $[a_0]$ is computed this way.

Beginning at the end of the series, (7.1.7) is now used to compute the $[e_t]$'s for $t = 9, 8, 7, \ldots, 1$. We start the backward process by setting $[e_{10}] = 0$. The effect of this approximation will be to introduce a transient into the system. However, for series of moderate length, the effect will typically be negligible by the time the beginning of the series is reached and thus will not affect the calculation of the a_t's. If desired, the adequacy of this approximation can be checked in any given case by performing a second iterative cycle.

Thus, to start the recursion in Table 7.2, in the row corresponding to $t = 9$, we enter a zero in the sixth column for the unknown value $0.5[e_{10}]$. Then, using (7.1.7), we obtain

$$\begin{aligned}[e_9] &= [w_9] + 0.5[e_{10}] \\ &= w_9 + 0 = -3\end{aligned}$$

[3]In practice, of course, useful parameter estimates could not be obtained from as few as 10 observations. We utilize this data subset merely to illustrate the calculations.

TABLE 7.2 Calculation of the $[a]$'s from the First 10 Values of Series B, Using $\theta = 0.5$

t	z_t	$[a_t]$	$0.5[a_{t-1}]$	$[w_t]$	$0.5[e_{t+1}]$	$[e_t]$	u_t
−1	[458.4]	0	0	0	0	0	
0	460	1.6	0	1.6	−1.6	0	−2.1
1	457	−2.2	0.8	−3.0	−0.1	−3.1	−4.1
2	452	−6.1	−1.1	−5.0	4.8	−0.2	−2.3
3	459	3.9	−3.0	7.0	2.6	9.6	8.5
4	462	5.0	2.0	3.0	2.3	5.3	9.5
5	459	−0.5	2.5	−3.0	7.6	4.6	9.2
6	463	3.7	−0.2	4.0	11.1	15.1	19.4
7	479	17.9	1.9	16.0	6.2	22.2	31.4
8	493	22.9	9.0	14.0	−1.5	12.5	27.5
9	490	8.5	11.5	−3.0	0	−3.0	8.5

so $0.5[e_9] = -1.5$ can be entered in the line $t = 8$, which enables us to compute $[e_8]$, and so on. Finally, we obtain

$$[e_0] = [w_0] + \theta[e_1]$$

that is, $0 = [w_0] - 1.6$, which gives $[w_0] = 1.6$, and thereafter $[w_{-h}] = 0, h = 1, 2, 3, \ldots$. Now, using (7.1.8) with $t = 0$, we obtain

$$[a_0] = [w_0] + \theta[a_{-1}] = 1.6 + (0.5)(0) = 1.6$$

and we can then continue the forward calculations of the remaining $[a_t]$'s, leading to $S(0.5) = \Sigma_{t=0}^{9}[a_t|0.5, \mathbf{w}]^2 = 1016.406$.

An alternative method that yields exact estimates of the presample values is presented in Appendix A7.3. For the model considered above, this method involves first computing the values $a_t(a_0 = 0)$, which we abbreviate as a_t^0, by the conditional method as $a_t^0 = w_t + \theta a_{t-1}^0, t = 1, 2, \ldots, n$, using $a_0^0 = 0$ as the initial value. Then a backward recursion is performed to obtain $u_t = a_t^0 + \theta u_{t+1}$, beginning from $t = n$, down to $t = 0$, with $u_{n+1} = 0$ as the starting value. Finally, then, the exact estimate of $[a_0]$ is given by $[a_0] = -u_0(1 - \theta^2)/(1 - \theta^{2(n+1)})$. Using this starting value, the $[a_t]$ are computed from the forward recursion $[a_t] = w_t + \theta[a_{t-1}], t = 1, 2, \ldots, n$, as in (7.1.8) and the exact sum of squares becomes $S(\theta) = \Sigma_{t=0}^{n}[a_t]^2$.

In the above example, by first computing the a_t^0 using a forward recursion setting $a_0 = 0$, we obtain the values of u_t by the backward recursion for $t = 9, 8, \ldots, 0$, displayed in the final column of Table 7.2. Hence, we obtain the exact estimate of a_0 as $[a_0] = -u_0(1 - \theta^2)/(1 - \theta^{2(n+1)}) = 1.549$. This value is very close to the approximate value of 1.545 obtained by the backward model approach, and the small difference has essentially no effect on the calculation of the remaining values $[a_t]$. Using the exact method for the entire series, we find that the unconditional sum of squares for $\theta = 0.5$ is

$$S(0.5) = \sum_{t=0}^{368}[a_t|0.5, \mathbf{w}]^2 = 27,691$$

which for this particular example is very close to the conditional value $S_*(0.5) = 27,694$. The unconditional sum of squares $S(\theta)$, for values of θ between -0.5 and $+0.5$, have been added to Table 7.1 and are very close to the conditional values $S_*(\theta)$ computed earlier.

7.1.5 General Procedure for Calculating the Unconditional Sum of Squares

In the above example, w_t was a first-order moving average process, with zero mean. It followed that all forecasts for lead times greater than 1 were zero and consequently that only one preliminary value (the back-forecast $[w_0] = 1.6$) was required to start the recursive calculations using the approximate approach, and only one value $[a_0]$ in the exact approach. For a qth-order moving average process, q nonzero preliminary values $[w_0], [w_{-1}], \ldots, [w_{1-q}]$ would be needed, or equivalently, the q values $[a_0], [a_{-1}], \ldots, [a_{1-q}]$ in the exact approach, with $S(\boldsymbol{\theta}) = \sum_{t=1-q}^{n} [a_t]^2$. Special procedures, which we discuss in Section 7.3.1, are available for estimating parameters in autoregressive models. However, we show in Appendix A7.3 that the procedure described in this section can supply the unconditional sum of squares for any ARIMA model.

Specifically, suppose that the w_t's are generated by the stationary forward model

$$\phi(B)\tilde{w}_t = \theta(B)a_t \tag{7.1.9}$$

where $w_t = \nabla^d z_t$ and $\tilde{w}_t = w_t - \mu$. Then, they could equally well have been generated by the backward model

$$\phi(F)\tilde{w}_t = \theta(F)e_t \tag{7.1.10}$$

As before, in the approximate method that utilizes the backward model, we could first employ (7.1.10) to supply back-forecasts $[\tilde{w}_{-j}|\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\theta}]$. Theoretically, the presence of the autoregressive operator ensures a series of such estimates that is infinite in extent. However, assuming stationarity, the estimates $[\tilde{w}_t]$ at and beyond some point $t = -Q$, with Q of moderate size, become essentially equal to zero. Thus, to a sufficient approximation, we can write

$$\tilde{w}_t = \phi^{-1}(B)\theta(B)a_t = \sum_{j=0}^{\infty} \psi_j a_{t-j} \simeq \sum_{j=0}^{Q} \psi_j a_{t-j}$$

This means that the original mixed process could be replaced by a moving average process of order Q, and the procedure for moving averages outlined in Section 7.1.4 may be used.

Thus, in general, the dual set of equations for generating the conditional expectations $[a_t|\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{w}]$ is obtained by taking conditional expectations in (7.1.10) and (7.1.9). That is,

$$\phi(F)[\tilde{w}_t] = \theta(F)[e_t] \tag{7.1.11}$$

is first used to generate the backward forecasts and then

$$\phi(B)[\tilde{w}_t] = \theta(B)[a_t] \tag{7.1.12}$$

is used to generate the $[a_t]$'s. If we find that the forecasts are negligible in magnitude beyond some lead time Q, the recursive calculation goes forward with

$$\begin{aligned} [e_{-j}|\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{w}] &= 0 \qquad j = 0, 1, 2, \ldots \\ [a_{-j}|\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{w}] &= 0 \qquad j > Q - 1 \end{aligned} \tag{7.1.13}$$

and the sum of squares is approximated by $S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1-Q}^{n} [a_t]^2$. As mentioned earlier, a second iterative cycle in this approximate method could be used, if desired.

Alternatively, for the general model (7.1.9), the exact method discussed in Appendix A7.3 can be used to obtain the sum of squares as

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1-q}^{n} [a_t]^2 + ([\mathbf{w}_*] - \mathbf{C}'[\mathbf{a}_*])'\mathbf{K}^{-1}([\mathbf{w}_*] - \mathbf{C}'[\mathbf{a}_*]) \tag{7.1.14}$$

Here, the vectors $[\mathbf{w}_*]' = ([\tilde{w}_{1-p}], \ldots, [\tilde{w}_0])$ and $[\mathbf{a}_*]' = ([a_{1-q}], \ldots, [a_0])$ are the exact back-forecasted values obtained as in (A7.3.12). They are given by $[\mathbf{e}_*] = ([\mathbf{w}_*]', [\mathbf{a}_*]')' = \mathbf{D}^{-1}\mathbf{F}'\mathbf{u}$, where the values u_t, $t = 1, \ldots, n$ of the vector $\mathbf{u}$ are obtained through the backward recursion $u_t = a_t^0 + \theta_1 u_{t+1} + \cdots + \theta_q u_{t+q}$ with zero initial values $u_{n+1} = \cdots = u_{n+q} = 0$, and the a_t^0 are the conditional values of the a_t computed from (7.1.12) using zero initial values, $a_{1-q}^0 = \cdots = a_0^0 = 0$ and $\tilde{w}_{1-p}^0 = \cdots = \tilde{w}_0^0 = 0$. After solving the equations $\mathbf{D}[\mathbf{e}_*] = \mathbf{F}'\mathbf{u}$, as described in (A7.3.12), the exact $[a_t]$'s are then calculated through the recursion

$$[a_t] = [\tilde{w}_t] - \phi_1[\tilde{w}_{t-1}] - \cdots - \phi_p[\tilde{w}_{t-p}] + \theta_1[a_{t-1}] + \cdots + \theta_q[a_{t-q}] \tag{7.1.15}$$

for $t = 1, 2, \ldots, n$ using the exact back-forecasts as starting values, with $[\tilde{w}_t] = \tilde{w}_t$ for $1 \le t \le n$. The matrices $\mathbf{C}$, $\mathbf{K}$, $\mathbf{D}$, and $\mathbf{F}$ necessary for the computation in (7.1.14) are defined explicitly in Appendix A7.3.

Comment on the Approximation. We saw that for the IMA(0, 1, 1) model fitted to the IBM Series B, the *conditional* sums of squares provides a very close approximation to the unconditional value. This will generally be the case for sufficiently long nonseasonal time series. However, as is discussed further in Chapter 9, for seasonal series, in particular, the conditional approximation becomes less satisfactory and the unconditional sum of squares should ordinarily be computed. Moreover, including the determinant in the likelihood function to obtain exact maximum likelihood estimates of the parameters can be beneficial if the roots of the moving average operator are close to the unit circle.

Simulation studies have been performed by Dent and Min (1978) and Ansley and Newbold (1980) to empirically investigate and compare the performance of the conditional least-squares, unconditional least-squares, and maximum likelihood estimators for ARMA models. Generally, the conditional and unconditional least-squares estimators serve as satisfactory approximations to the maximum likelihood estimator for large-sample sizes. However, the simulation evidence suggests a preference for the maximum likelihood estimator for small- or moderate-sample sizes, especially if the moving average operator has a root close to the boundary of the invertibility region. Some additional information on the relative performance of the different estimators was provided by Hillmer and Tiao (1979) and Osborn (1982), who examined the *expected values* of the conditional sum of squares, the unconditional sum of squares, and the log-likelihood for an MA(1) model, as functions of the unknown parameter θ, for different sample sizes n. These studies provide an idea of

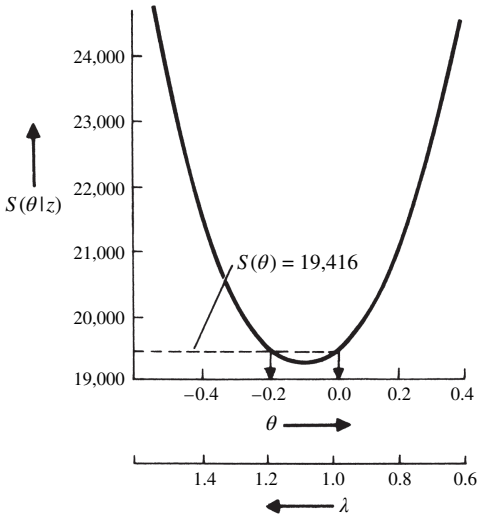


FIGURE 7.1 Plot of $S(\theta)$ for Series **B**.

how the corresponding estimators will behave for various sample sizes, and the results are consistent with those obtained from simulation studies.

7.1.6 Graphical Study of the Sum-of-Squares Function

The sum-of-squares function $S(\theta)$ for the IBM data given in Table 7.1 is plotted in Figure 7.1. The overall minimum sum of squares is at about $\theta = -0.09$ ($\lambda = 1.09$), which is the least-squares estimate and, on the assumption of normality, a close approximation to the *maximum likelihood estimate* of the parameter θ.

The graphical study of the sum-of-squares functions is readily extended to two parameters by evaluating the sum of squares over a suitable grid of parameter values and plotting contours. As discussed earlier, on the assumption of normality, the contours are very nearly likelihood contours. Figure 7.2 shows a grid of $S(\lambda_0, \lambda_1)$ values for Series B fitted with the IMA(0, 2, 2) model:

$$\begin{aligned}\nabla^2 z_t &= (1 - \theta_1 B - \theta_2 B^2)a_t \\ &= [1 - (2 - \lambda_0 - \lambda_1)B - (\lambda_0 - 1)B^2]a_t \end{aligned} \tag{7.1.16}$$

or in the form

$$\nabla^2 z_t = (\lambda_0 \nabla + \lambda_1)a_{t-1} + \nabla^2 a_t$$

The minimum sum of squares in Figure 7.2 is at about $\hat{\lambda}_0 = 1.09$ and $\hat{\lambda}_1 = 0.0$. The plot thus confirms that the preferred model in this case is an IMA(0, 1, 1) process. The device illustrated here, of fitting a model somewhat more elaborate than that expected to be needed, can provide a useful confirmation of the original identification. The elaboration of the model should be made, of course, in the direction ''feared'' to be necessary.

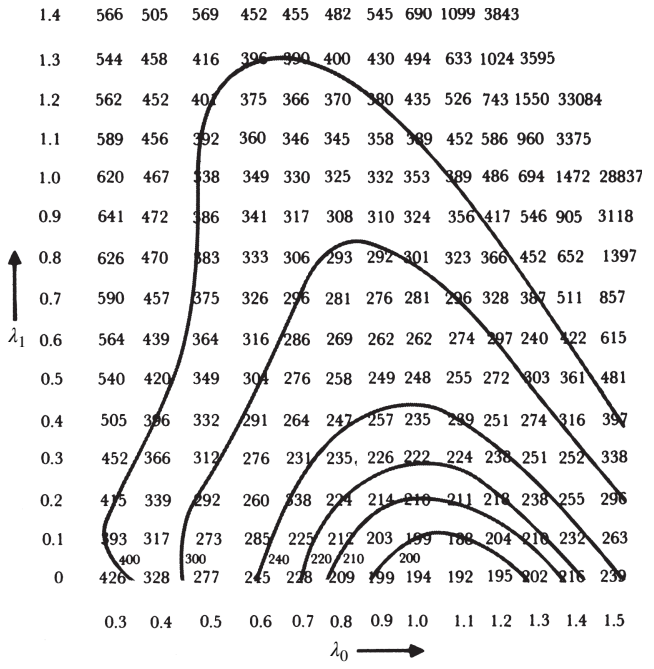


FIGURE 7.2 Values of $S(\lambda_0, \lambda_1) \times 10^{-2}$ for Series B on a grid of (λ_0, λ_1) values and approximate contours.

Three Parameters. When we wish to study models with three parameters, two-dimensional contour diagrams for a number of values of the third parameter can be drawn. For illustration, part of such a series of diagrams is shown in Figure 7.3 for Series A, C, and D. In each case, the ''elaborated'' model

$$\begin{aligned}\nabla^2 z_t &= (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3) a_t \\ &= [1 - (2 - \lambda_{-1} - \lambda_0 - \lambda_1) B - (\lambda_0 + 2\lambda_{-1} - 1) B^2 + \lambda_{-1} B^3] a_t\end{aligned}$$

or

$$\nabla^2 z_t = (\lambda_{-1} \nabla^2 + \lambda_0 \nabla + \lambda_1) a_{t-1} + \nabla^2 a_t$$

has been fitted, leading to the conclusion that the best-fitting models of this type[4] are as shown in Table 7.3.

The inclusion of additional parameters (particularly λ_{-1}) in this fitting process is not strictly necessary, but we have included them to illustrate the effect of overfitting and to show how closely our identification seems to be confirmed for these series.

[4]We show later in Section 7.2.5 that slightly better fits are obtained in some cases with closely related models containing ''stationary'' autoregressive terms.

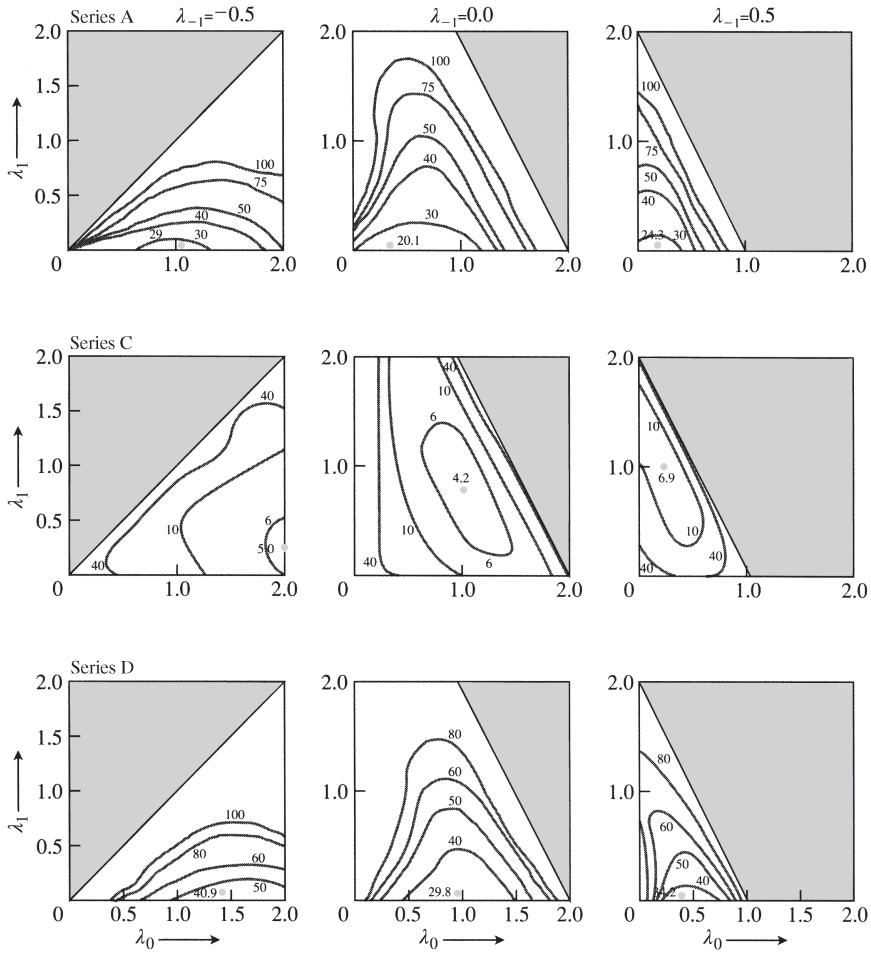


FIGURE 7.3 Sum-of-squares contours for Series A, C, and D (shaded lines indicate boundaries of the invertibility regions).

TABLE 7.3 IMA Models Fitted to Series A, C, and D

Series	$\hat{\lambda}_{-1}$	$\hat{\lambda}_0$	$\hat{\lambda}_1$	Fitted Series
A	0	0.3	0.0	$\nabla z_t = 0.3a_{t-1} + \nabla a_t$
C	0	1.1	0.8	$\nabla^2 z_t = 1.1\nabla a_{t-1} + 0.8a_{t-1} + \nabla^2 a_t$
D	0	0.9	0.0	$\nabla z_t = 0.9a_{t-1} + \nabla a_t$

7.1.7 Examination of the Likelihood Function and Confidence Regions

The likelihood function is not, of course, plotted merely to indicate maximum likelihood values. The graph of this function contains the totality of information that comes from the data. In some fields of study, cases can occur where the likelihood function has two or more peaks and also where the likelihood function contains sharp ridges and spikes. In each such

case, the likelihood function is trying to tell us something that we need to know. Thus, the existence of two peaks of approximately equal heights implies that there are two sets of parameter values that might explain the data. The existence of obliquely oriented ridges means that a value of one parameter, considerably different from its maximum likelihood value, could explain the data if accompanied by a value of the other parameter, which deviated appropriately. To understand the estimation fully, it is thus useful to examine the likelihood function both analytically and graphically.

Need for Care in Interpreting the Likelihood Function. Care is needed in interpreting the likelihood function. For example, results discussed later, which assume that the log-likelihood is approximately quadratic near its maximum, will clearly not apply to the three-parameter cases depicted in Figure 7.3. However, these examples are exceptional because here we are deliberately *overfitting* the model. If the simpler model is justified, we should *expect* to find the likelihood function contours truncated near its maximum by a boundary in the higher dimensional parameter space. However, quadratic approximations *could* be used if the simpler *identified* model rather than the overparameterized model was fitted.

Special care is needed when the maximum of the likelihood function may be on or near a boundary. Consider the situation shown in Figure 7.4 and suppose we know a priori that a parameter $\beta > \beta_0$. The maximum likelihood within the permissible range of β is at B, where $\beta = \beta_0$, not at A or at C. It will be noticed that the first derivative of the likelihood is in this case *nonzero* at the maximum likelihood value and that the quadratic approximation is certainly not an adequate representation of the likelihood.

When a class of estimation problems are examined initially, it is important to plot the likelihood function to identify potential issues. After the behavior of a potential model is well understood, and knowledge of the situation indicates that it is appropriate to do so, we may take certain shortcuts, which we now consider. We begin by considering expressions for the variances and covariances of maximum likelihood estimates, appropriate when the log-likelihood is approximately quadratic and the sample size is moderately large.

In what follows, it is convenient to define a vector $\boldsymbol{\beta}$ whose $k = p + q$ elements are the autoregressive and moving average parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Thus, the complete set of

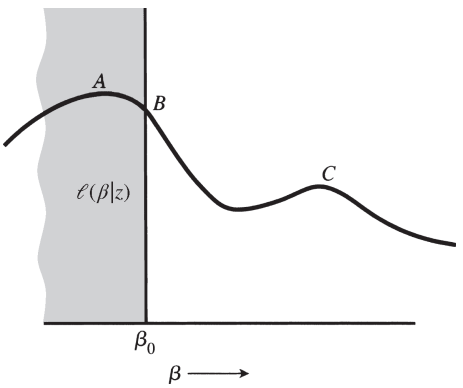


FIGURE 7.4 Hypothetical likelihood function with a constraint $\beta > \beta_0$.

$p + q + 1 = k + 1$ parameters of the ARMA process may be written as $\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2$; or as $\boldsymbol{\beta}, \sigma_a^2$; or simply as $\boldsymbol{\xi}$.

Variances and Covariances of ML Estimates. For the appropriately parameterized ARMA model, it will often happen that over the relevant[5] region of the parameter space, the log-likelihood is approximately quadratic in the elements of $\boldsymbol{\beta}$ (i.e., of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$), so that

$$l(\boldsymbol{\xi}) = l(\boldsymbol{\beta}, \sigma_a^2) \simeq l(\hat{\boldsymbol{\beta}}, \sigma_a^2) + \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} l_{ij}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j) \tag{7.1.17}$$

where, to the approximation considered, the derivatives

$$l_{ij} = \frac{\partial^2 l(\boldsymbol{\beta}, \sigma_a^2)}{\partial \beta_i \partial \beta_j} \tag{7.1.18}$$

are constant. For large n, the influence of the term $f(\boldsymbol{\phi}, \boldsymbol{\theta})$ in (7.1.5) can be ignored in most cases. Hence, $l(\boldsymbol{\beta}, \sigma_a^2)$ will be essentially quadratic in $\boldsymbol{\beta}$ if this is true for $S(\boldsymbol{\beta})$. Alternatively, $l(\boldsymbol{\beta}, \sigma_a^2)$ will be essentially quadratic in $\boldsymbol{\beta}$ if the conditional expectations $[a_t|\boldsymbol{\beta}, \mathbf{w}]$ in (7.1.6) are approximately locally linear in the elements of $\boldsymbol{\beta}$. Thus, for moderate- and large-sample sizes n, when the local quadratic approximation (7.1.17) is adequate, useful approximations to the variances and covariances of the estimates and approximate confidence regions may be obtained.

Information Matrix for the Parameters $\boldsymbol{\beta}$. The $(k \times k)$ matrix $-\{E[l_{ij}]\} = \mathbf{I}(\boldsymbol{\beta})$ is referred to (Fisher, 1956; Whittle, 1953) as the *information matrix* for the parameters $\boldsymbol{\beta}$, where the expectation is taken over the distribution of $\mathbf{w}$. For a given value of σ_a^2, the *variance–covariance* matrix $\mathbf{V}(\hat{\boldsymbol{\beta}})$ for the ML estimates $\hat{\boldsymbol{\beta}}$ is, for large samples, given by the inverse of this information matrix, that is,

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) \simeq \{-E[l_{ij}]\}^{-1} \equiv \mathbf{I}^{-1}(\boldsymbol{\beta}) \tag{7.1.19}$$

For example, if $k = 2$, the large-sample variance–covariance matrix is

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} V(\hat{\beta}_1) & \operatorname{cov}[\hat{\beta}_1, \hat{\beta}_2] \\ \operatorname{cov}[\hat{\beta}_1, \hat{\beta}_2] & V(\hat{\beta}_2) \end{bmatrix} \simeq - \begin{bmatrix} E[l_{11}] & E[l_{12}] \\ E[l_{12}] & E[l_{22}] \end{bmatrix}^{-1}$$

In addition, the ML estimates $\hat{\boldsymbol{\beta}}$ obtained from a stationary invertible ARMA process were shown to be asymptotically distributed as *multivariate normal* with mean vector $\boldsymbol{\beta}$ and covariance matrix $\mathbf{I}^{-1}(\boldsymbol{\beta})$ (e.g., Mann and Wald, 1943; Whittle, 1953; Hannan, 1960; Walker, 1964) in the sense that $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to the multivariate normal $N\{0, \mathbf{I}_*^{-1}(\boldsymbol{\beta})\}$ as $n \to \infty$, where $\mathbf{I}_*(\boldsymbol{\beta}) = \lim n^{-1}\mathbf{I}(\boldsymbol{\beta})$. The specific form of the information matrix $\mathbf{I}(\boldsymbol{\beta})$ and the limiting matrix $\mathbf{I}_*(\boldsymbol{\beta})$ for ARMA(p, q) models are described in Section 7.2.6, and details on the asymptotic normality of the estimator $\hat{\boldsymbol{\beta}}$ are examined for the special case of AR models in Appendix A7.5.

[5] Say over a 95% confidence region.

Now, using (7.1.5), we have

$$l_{ij} \simeq \frac{-S_{ij}}{2\sigma_a^2} \tag{7.1.20}$$

where

$$S_{ij} = \frac{\partial^2 S(\boldsymbol{\beta}|\mathbf{w})}{\partial\beta_i \partial\beta_j}$$

Furthermore, if for large samples, we approximate the expected values of l_{ij} or of S_{ij} by the values actually observed, then, using (7.1.19), we obtain

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) \simeq \{-E[l_{ij}]\}^{-1} \simeq 2\sigma_a^2\{E[S_{ij}]\}^{-1} \simeq 2\sigma_a^2\{S_{ij}\}^{-1} \tag{7.1.21}$$

Thus, for $k = 2$,

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) \simeq 2\sigma_a^2 \begin{bmatrix} \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_1^2} & \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_2} \\ \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_2} & \dfrac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_2^2} \end{bmatrix}^{-1}$$

If $S(\boldsymbol{\beta})$ were exactly quadratic in $\boldsymbol{\beta}$ over the relevant region of the parameter space, then all the derivatives S_{ij} would be constant over this region. In practice, the S_{ij} will vary somewhat, and we will usually assume that the derivatives are determined at or near the point $\hat{\boldsymbol{\beta}}$. Now, it is shown in the Appendices A7.3 and A7.4 that an estimate[6] of σ_a^2 is provided by

$$\hat{\sigma}_a^2 = \frac{S(\hat{\boldsymbol{\beta}})}{n} \tag{7.1.22}$$

and that for large samples, $\hat{\sigma}_a^2$ and $\hat{\boldsymbol{\beta}}$ are uncorrelated. Finally, the elements of (7.1.21) may be estimated from

$$\text{cov}[\hat{\beta}_i, \hat{\beta}_j] \simeq 2\hat{\sigma}_a^2 S^{ij} \tag{7.1.23}$$

where the $(k \times k)$ matrix $\{S^{ij}\}$ is given by $\{S^{ij}\} = \{S_{ij}\}^{-1}$ and the expression (7.1.23) is understood to define the variance $V(\hat{\beta}_i)$ when $j = i$.

Approximate Confidence Regions for the Parameters. In particular, these results allow us to obtain the approximate variances of our estimates. By taking the square root of these variances, we obtain approximate *standard errors* (SE) of the estimates. The standard error of an estimate $\hat{\beta}_i$ is denoted by $\text{SE}[\hat{\beta}_i]$. When we have to consider several parameters simultaneously, we need some means of judging the precision of the estimates *jointly*. One means of doing this is to determine a *confidence region*. If, for given σ_a^2, $l(\boldsymbol{\beta}, \sigma_a^2)$ is approximately quadratic in $\boldsymbol{\beta}$ in the neighborhood of $\hat{\boldsymbol{\beta}}$, then using (7.1.19) (see also

[6]Arguments can be advanced for using the divisor $n - k = n - p - q$ rather than n in (7.1.22), but for moderate-sample sizes, this modification does not make much difference.

TABLE 7.4 $S(\lambda)$ and Its First and Second Differences for Various Values of λ for Series B

$\lambda = 1 - \theta$	$S(\lambda)$	$\nabla(S)$	$\nabla^2(S)$
1.5	23,928	2,333	960
1.4	21,595	1,373	634
1.3	20,222	739	476
1.2	19,483	263	406
1.1	19,220	−143	390
1.0	19,363	−533	422
0.9	19,896	−955	508
0.8	20,851	−1,463	691
0.7	22,314	−2,154	1069
0.6	24,468	−3,223	
0.5	27,691		

Appendix A7.1), an approximate $1 - \varepsilon$ confidence region will be defined by

$$-\sum_i \sum_j E[l_{ij}](\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j) < \chi^2_\varepsilon(k) \tag{7.1.24}$$

where $\chi^2_\varepsilon(k)$ is the significance point exceeded by a proportion ε of the χ^2 distribution, having k degrees of freedom.

Alternatively, using the approximation (7.1.21) and substituting the estimate of (7.1.22) for σ_a^2, the approximate confidence region is given by[7]

$$\sum_i \sum_j S_{ij}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j) < 2\hat{\sigma}_a^2 \chi^2_\varepsilon(k) \tag{7.1.25}$$

However, for a quadratic $S(\boldsymbol{\beta})$ surface

$$S(\boldsymbol{\beta}) - S(\hat{\boldsymbol{\beta}}) = \frac{1}{2}\sum_i \sum_j S_{ij}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j) \tag{7.1.26}$$

Thus, using (7.1.22) and (7.1.25), we finally obtain the result that the approximate $1 - \varepsilon$ confidence region is bounded by the contour on the sum-of-squares surface, for which

$$S(\boldsymbol{\beta}) = S(\hat{\boldsymbol{\beta}})\left[1 + \frac{\chi^2_\varepsilon(k)}{n}\right] \tag{7.1.27}$$

Examples of the Calculation of Approximate Confidence Intervals and Regions.

1. *Example: Series B.* For Series B, values of $S(\lambda)$ and of its differences are shown in Table 7.4. The second difference of $S(\lambda)$ is not constant, and thus $S(\lambda)$ is not strictly quadratic. However, in the range from $\lambda = 0.85$ to $\lambda = 1.35$, $\nabla^2(S)$ does not change greatly, so that (7.1.27) can be expected to provide a reasonably close approx-

[7] A somewhat closer approximation based on the F distribution, which takes account of the approximate sampling distribution of $\hat{\sigma}_a^2$, may be employed. For moderate-sample sizes this refinement does not make much practical difference.

imation. With a minimum value $S(\hat{\lambda}) = 19,216$, the critical value $S(\lambda)$, defining an approximate 95% confidence interval, is then given by

$$S(\lambda) = 19,216\left(1 + \frac{3.84}{368}\right) = 19,416$$

Reading off the values of λ corresponding to $S(\lambda) = 19,416$ in Figure 7.1, we obtain an approximate confidence interval $0.98 < \lambda < 1.19$.

Alternatively, we can employ (7.1.25). Using the second difference at $\lambda = 1.1$, given in Table 7.4, to approximate the derivative, we obtain

$$S_{11} = \frac{\partial^2 S}{\partial \lambda^2} \simeq \frac{390}{(0.1)^2}$$

Also, using (7.1.22), $\hat{\sigma}_a^2 = 19,216/368 = 52.2$. Thus, the 95% confidence interval, defined by (7.1.25), is

$$\frac{390}{(0.1)^2}(\lambda - 1.09)^2 < 2 \times 52.2 \times 3.84$$

that is, $|\lambda - 1.09| < 0.10$. Thus, the interval is $0.99 < \lambda < 1.19$, which agrees closely with the previous calculation.

In this example, where there is only a single parameter λ, the use of (7.1.24) and (7.1.25) is equivalent to using an interval $\hat{\lambda} \pm u_{\varepsilon/2}\hat{\sigma}(\hat{\lambda})$, where $u_{\varepsilon/2}$ is the value, which excludes a proportion $\varepsilon/2$ in the upper tail of the standard normal distribution. An approximate standard error for $\hat{\lambda}$, $\hat{\sigma}(\hat{\lambda}) = \sqrt{2\hat{\sigma}_a^2 S_{11}^{-1}}$, is obtained from (7.1.23). In the present example,

$$V(\hat{\lambda}) = 2\hat{\sigma}_a^2 S_{11}^{-1} = \frac{2 \times 52.2 \times 0.1^2}{390} = 0.00268$$

and the approximate standard error is $\hat{\sigma}(\hat{\lambda}) = \sqrt{0.00268} = 0.052$. Thus, the approximate 95% confidence interval is $\hat{\lambda} \pm 1.96\hat{\sigma}(\hat{\lambda}) = 1.09 \pm 0.10$, as before.

Finally, we show later in Section 7.2.6 that it is possible to evaluate (7.1.19) analytically, for large samples from an MA(1) process, yielding

$$V(\hat{\lambda}) \simeq \frac{\lambda(2 - \lambda)}{n}$$

For the present example, substituting $\hat{\lambda} = 1.09$ for λ, we find that $V(\hat{\lambda}) \simeq 0.00269$, which agrees closely with the previous estimate and so yields the same standard error of 0.052 and the same confidence interval.

2. *Example: Series C.* In the identification of Series C, one model that was entertained was a (0, 2, 2) process. To illustrate the application of (7.1.27) for more than one parameter, Figure 7.5 shows an approximate 95% confidence region (shaded) for λ_0 and λ_1 of Series C. For this example, $S(\hat{\lambda}) = 4.20$, $n = 224$, and $\chi^2_{0.05}(2) = 5.99$,

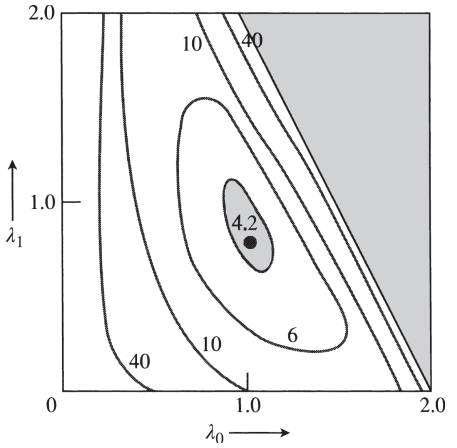


FIGURE 7.5 Sum-of-squares contours with shaded 95% confidence region for Series C, assuming a model of order (0, 2, 2).

so that the approximate 95% confidence region is bounded by the contour for which

$$S(\lambda_0, \lambda_1) = 4.20\left(1 + \frac{5.99}{224}\right) = 4.31$$

7.2 NONLINEAR ESTIMATION

7.2.1 General Method of Approach

The plotting of the sum-of-squares function is of particular importance in the study of new estimation problems because it ensures that any peculiarities in the estimation situation show up. When we are satisfied that anomalies are unlikely, other methods may be used.

We have seen that for most cases, the maximum likelihood estimates are closely approximated by the least-squares estimates, which minimize

$$S(\boldsymbol{\phi}, \theta) = \sum_{t=1}^{n} [a_t]^2 + [\mathbf{e}_*]' \boldsymbol{\Omega}^{-1} [\mathbf{e}_*]$$

and in practice, this function can be approximated by a finite sum $\sum_{t=1-Q}^{n} [a_t]^2$.

In general, considerable simplification occurs in the minimization with respect to $\boldsymbol{\beta}$, of a sum of squares $\Sigma_{t=1}^{n} [f_t(\boldsymbol{\beta})]^2$, if each $f_t(\boldsymbol{\beta})$ $(t = 1, 2, \ldots, n)$ is a *linear* function of the parameters $\boldsymbol{\beta}$. We now show that the autoregressive and moving average models differ with respect to the linearity of the $[a_t]$. For the purely autoregressive process, $[a_t] = \phi(B)[\tilde{w}_t] = [\tilde{w}_t] - \Sigma_{i=1}^{p} \phi_i [\tilde{w}_{t-i}]$ and

$$\frac{\partial [a_t]}{\partial \phi_i} = -[\tilde{w}_{t-i}] + \phi(B) \frac{\partial [\tilde{w}_t]}{\partial \phi_i}$$

Now for $u > 0, [\tilde{w}_u] = \tilde{w}_u$ and $\partial[\tilde{w}_u]/\partial\phi_i = 0$, while for $u \le 0, [\tilde{w}_u]$ and $\partial[\tilde{w}_u]/\partial\phi_i$ are both functions of $\boldsymbol{\phi}$. Thus, except for the effect of ''starting values,'' $[a_t]$ is linear in the ϕ's. By contrast, for the pure moving average process,

$$[a_t] = \theta^{-1}(B)[\tilde{w}_t] \qquad \frac{\partial[a_t]}{\partial\theta_j} = \theta^{-2}(B)[\tilde{w}_{t-j}] + \theta^{-1}(B)\frac{\partial[\tilde{w}_t]}{\partial\theta_j}$$

so that the $[a_t]$'s are always nonlinear functions of the moving average parameters.

We will see in Section 7.3 that special simplifications occur in obtaining least-squares and maximum likelihood estimates for the autoregressive process. We show in the present section how, by iterative application of linear least-squares, estimates may be obtained for any ARMA process.

Linearization of the Model. In what follows, we continue to use $\boldsymbol{\beta}$ as a general symbol for the $k = p + q$ parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$. We need, then, to minimize

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) \simeq \sum_{t=1-Q}^{n} [a_t|\tilde{\boldsymbol{w}}, \boldsymbol{\beta}]^2 = \sum_{t=1-Q}^{n} [a_t]^2$$

Expanding $[a_t]$ in a Taylor series about its value corresponding to some guessed set of parameter values $\boldsymbol{\beta}_0' = (\beta_{1,0}, \beta_{2,0}, \ldots, \beta_{k,0})$, we have approximately

$$[a_t] = [a_{t,0}] - \sum_{i=1}^{k} (\beta_i - \beta_{i,0}) x_{t,i} \tag{7.2.1}$$

where $[a_{t,0}] = [a_t|\mathbf{w}, \boldsymbol{\beta}_0]$ and

$$x_{t,i} = -\frac{\partial[a_t]}{\partial\beta_i}\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$$

Now, if $\mathbf{X}$ is the $(n + Q) \times k$ matrix $\{x_{t,i}\}$, then the $n + Q$ equations (7.2.1) may be expressed as

$$[\mathbf{a}_0] = \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + [\mathbf{a}]$$

where $[\mathbf{a}_0]$ and $[\mathbf{a}]$ are column vectors with $n + Q$ elements.

The adjustments $\boldsymbol{\beta} - \boldsymbol{\beta}_0$, which minimize $S(\boldsymbol{\beta}) = S(\boldsymbol{\phi}, \boldsymbol{\theta}) = [\mathbf{a}]'[\mathbf{a}]$, may now be obtained by linear least-squares, that is, by ''regressing'' the $[a_o]$'s onto the x's. This gives the usual linear least-squares estimates, as presented in Appendix A7.2.1, of the adjustments as $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{a}_0]$, hence, $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{a}_0]$. Because the $[a_t]$'s will not be exactly linear in the parameters $\boldsymbol{\beta}$, a single adjustment will not immediately produce the final least-squares values. Instead, the adjusted values $\hat{\boldsymbol{\beta}}$ are substituted as new guesses and the process is repeated until convergence occurs. Convergence is faster if reasonably good guesses, such as may be obtained at the identification stage, are used initially. If sufficiently bad initial guesses are used, the process may not converge at all.

7.2.2 Numerical Estimates of the Derivatives

The derivatives $x_{t,i}$ may be obtained directly, as we illustrate later. They can also be computed numerically using a general nonlinear least-squares routine. This is done

by perturbing the parameters ''one at a time.'' Thus, for a given model, the values $[a_t|\mathbf{w}, \beta_{1,0}, \beta_{2,0}, \ldots, \beta_{k,0}]$ for $t = 1 - Q, \ldots, n$ are calculated recursively, using whatever preliminary ''back-forecasts'' may be needed. The calculation is then repeated for $[a_t|\mathbf{w}, \beta_{1,0} + \delta_1, \beta_{2,0}, \ldots, \beta_{k,0}]$, then for $[a_t|\mathbf{w}, \beta_{1,0}, \beta_{2,0} + \delta_2, \ldots, \beta_{k,0}]$, and so on. The negative of the required derivative is then given to sufficient accuracy using

$$x_{t,i} = \frac{[a_t|\mathbf{w}, \beta_{1,0}, \ldots, \beta_{i,0}, \ldots, \beta_{k,0}] - [a_t|\mathbf{w}, \beta_{1,0}, \ldots, \beta_{i,0} + \delta_i, \ldots, \beta_{k,0}]}{\delta_i} \tag{7.2.2}$$

The numerical method described above has the advantage of universal applicability and requires us to program the calculation of the $[a_t]$'s only, not their derivatives. General nonlinear estimation routines, which essentially require only input instructions on how to compute the $[a_t]$'s, are generally available. In some versions, it is necessary to choose the δ's in advance. In others, the program itself carries through a preliminary iteration to find suitable δ's. Many programs include special features to avoid overshoot and to speed up convergence.

Provided that the least-squares solution is not on or near a constraining boundary, the value of $\mathbf{X} = \mathbf{X}_{\hat{\boldsymbol{\beta}}}$ from the final iteration may be used to compute approximate variances, covariances, and confidence intervals. Thus, similar to the usual linear least-squares results in Appendix A7.2.3,

$$(\mathbf{X}'_{\hat{\boldsymbol{\beta}}}\mathbf{X}_{\hat{\boldsymbol{\beta}}})^{-1}\sigma_a^2$$

will approximate the variance–covariance matrix of the $\hat{\beta}$'s, and σ_a^2 will be estimated by $\hat{\sigma}_a^2 = S(\hat{\boldsymbol{\beta}})/n$.

7.2.3 Direct Evaluation of the Derivatives

We now show that it is also possible to obtain derivatives directly, but additional recursive calculations are needed. To illustrate the method, it is sufficient to consider an ARMA(1, 1) process, which can be written in either of the forms as

$$e_t = w_t - \phi w_{t+1} + \theta e_{t+1}$$
$$a_t = w_t - \phi w_{t-1} + \theta a_{t-1}$$

We have seen in Section 7.1.4, how the two versions of the model may be used in alternation, one providing initial values with which to start off a recursion with the other. We assume that a first computation has already been made yielding values of $[e_t]$, of $[a_t]$, and of $[w_0], [w_{-1}], \ldots, [w_{1-Q}]$, as in Section 7.1.5, and that $[w_{-Q}], [w_{-Q-1}], \ldots$ and hence $[a_{-Q}], [a_{-Q-1}], \ldots$ are negligible. We now show that a similar dual calculation may be used in calculating derivatives.

Using the notation $a_t^{(\phi)}$ to denote the partial derivative $\partial[a_t]/\partial\phi$, we obtain

$$e_t^{(\phi)} = w_t^{(\phi)} - \phi w_{t+1}^{(\phi)} + \theta e_{t+1}^{(\phi)} - [w_{t+1}] \tag{7.2.3}$$

$$a_t^{(\phi)} = w_t^{(\phi)} - \phi w_{t-1}^{(\phi)} + \theta a_{t-1}^{(\phi)} - [w_{t-1}] \tag{7.2.4}$$

$$e_t^{(\theta)} = w_t^{(\theta)} - \phi w_{t+1}^{(\theta)} + \theta e_{t+1}^{(\theta)} + [e_{t+1}] \tag{7.2.5}$$

$$a_t^{(\theta)} = w_t^{(\theta)} - \phi w_{t-1}^{(\theta)} + \theta a_{t-1}^{(\theta)} + [a_{t-1}] \tag{7.2.6}$$

Now,

$$\left.\begin{aligned} [w_t] &= w_t \\ w_t^{(\phi)} &= w_t^{(\theta)} = 0 \end{aligned}\right\} \qquad t = 1, 2, \ldots, n \tag{7.2.7}$$

and

$$[e_{-j}] = 0 \qquad j = 0, 1, \ldots, n \tag{7.2.8}$$

Consider equations (7.2.3) and (7.2.4). By setting $e_{n+1}^{(\phi)} = 0$ in (7.2.3), we can begin a back recursion, which using (7.2.7) and (7.2.8) eventually allows us to compute $w_{-j}^{(\phi)}$ for $j = 0, 1, \ldots, Q-1$. Since $a_{-Q}^{(\phi)}, a_{-Q-1}^{(\phi)}, \ldots$ can be taken to be zero, we can now use (7.2.4) to compute recursively the required derivatives $a_t^{(\phi)}$. In a similar way, (7.2.5) and (7.2.6) can be used to calculate the derivatives $a_t^{(\theta)}$.

7.2.4 General Least-Squares Algorithm for the Conditional Model

An approximation that we have sometimes used with long series is to set starting values for the a_t's, and hence for the derivatives in the x_t's, equal to their unconditional expectations of zero and then to proceed directly with the forward recursions. The effect is to introduce a transient into both the a_t and the x_t series, the latter being slower to die out since the x_t's depend on the a_t's. In some instances, where there is an abundance of data (say, 200 or more observations), the effect of the approximation can be nullified at the expense of some loss of information, by discarding, say, the first 10 calculated values.

If we adopt the approximation, an interesting general algorithm for this conditional model results. The ARMA(p, q) model can be written as

$$a_t = \theta^{-1}(B)\phi(B)\tilde{w}_t$$

where $w_t = \nabla^d z_t$, $\tilde{w}_t = w_t - \mu$ and

$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_i B^i - \cdots \theta_q B^q$$
$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_j B^j - \cdots \phi_p B^p$$

If the first guesses for the parameters $\boldsymbol{\beta} = (\boldsymbol{\phi}, \boldsymbol{\theta})$ are $\boldsymbol{\beta}_0 = (\boldsymbol{\phi}_0, \boldsymbol{\theta}_0)$, then

$$a_{t,0} = \theta_0^{-1}(B)\phi_0(B)\tilde{w}_t$$

and

$$-\left.\frac{\partial a_t}{\partial \phi_j}\right|_{\beta_0} = u_{t,j} = u_{t-j} \qquad -\left.\frac{\partial a_t}{\partial \theta_i}\right|_{\beta_0} = v_{t,i} = v_{t-i}$$

where

$$u_t = \theta_0^{-1}(B)\tilde{w}_t = \phi_0^{-1}(B)a_{t,0} \tag{7.2.9}$$
$$v_t = -\theta_0^{-2}(B)\phi_0(B)\tilde{w}_t = -\theta_0^{-1}(B)a_{t,0} \tag{7.2.10}$$

The a_t's, u_t's, and v_t's may be calculated recursively, with starting values for a_t's, u_t's, and v_t's set equal to zero, as follows:

$$\begin{aligned} a_{t,0} &= \tilde{w}_t - \phi_{1,0}\tilde{w}_{t-1} - \cdots - \phi_{p,0}\tilde{w}_{t-p} + \theta_{1,0}a_{t-1,0} \\ &\quad + \cdots + \theta_{q,0}a_{t-q,0} \qquad (7.2.11) \\ u_t &= \theta_{1,0}u_{t-1} + \cdots + \theta_{q,0}u_{t-q} + \tilde{w}_t \qquad (7.2.12) \\ &= \phi_{1,0}u_{t-1} + \cdots + \phi_{p,0}u_{t-p} + a_{t,0} \qquad (7.2.13) \\ v_t &= \theta_{1,0}v_{t-1} + \cdots + \theta_{q,0}v_{t-q} - a_{t,0} \qquad (7.2.14) \end{aligned}$$

Corresponding to (7.2.1), the approximate linear regression equation becomes

$$a_{t,0} = \sum_{j=1}^{p}(\phi_j - \phi_{j,0})u_{t-j} + \sum_{i=1}^{q}(\theta_i - \theta_{i,0})v_{t-i} + a_t \qquad (7.2.15)$$

The adjustments are then the regression coefficients of $a_{t,0}$ on the u_{t-j} and the v_{t-i}. By adding the adjustments to the first guesses $(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0)$, a set of ''second guesses'' are formed and these now take the place of $(\boldsymbol{\phi}_0, \boldsymbol{\theta}_0)$ in a second iteration, in which new values of $a_{t,0}, u_t$, and v_t are computed, until convergence eventually occurs.

Alternative Form for the Algorithm. The approximate linear expansion (7.2.15) can be written in the form

$$\begin{aligned} a_{t,0} &= \sum_{j=1}^{p}(\phi_j - \phi_{j,0})B^j\phi_0^{-1}(B)a_{t,0} - \sum_{i=1}^{q}(\theta_i - \theta_{i,0})B^i\theta_0^{-1}(B)a_{t,0} + a_t \\ &= -[\phi(B) - \phi_0(B)]\phi_0^{-1}(B)a_{t,0} + [\theta(B) - \theta_0(B)]\theta_0^{-1}(B)a_{t,0} + a_t \end{aligned}$$

that is,

$$a_{t,0} = -\phi(B)[\theta_0^{-1}(B)a_{t,0}] + \theta(B)[\theta_0^{-1}(B)a_{t,0}] + a_t \qquad (7.2.16)$$

which presents the algorithm in an interesting form.

Application to an IMA(0, 2, 2) process. To illustrate the calculation with the conditional approximation, consider the estimation of least-squares values $\hat{\theta}_1, \hat{\theta}_2$ for Series C using the model of order (0, 2, 2):

$$w_t = (1 - \theta_1 B - \theta_2 B^2)a_t$$

with $w_t = \nabla^2 z_t$,

$$\begin{aligned} a_{t,0} &= w_t + \theta_{1,0}a_{t-1,0} + \theta_{2,0}a_{t-2,0} \\ v_t &= -a_{t,0} + \theta_{1,0}v_{t-1} + \theta_{2,0}v_{t-2} \end{aligned}$$

Using the initial values $\theta_{1,0} = 0.1$ and $\theta_{2,0} = 0.1$, the first adjustments to $\theta_{1,0}$ and $\theta_{2,0}$ are found by ''regressing'' $a_{t,0}$ on v_{t-1} and v_{t-2}. The process is repeated until convergence occurs. Successive parameter estimates are shown in Table 7.5.

TABLE 7.5 Convergence of Parameter Estimates for IMA(0, 2, 2) Process

Iteration	θ_1	θ_2
0	0.1000	0.1000
1	0.1247	0.1055
2	0.1266	0.1126
3	0.1286	0.1141
4	0.1290	0.1149
5	0.1292	0.1151
6	0.1293	0.1152
7	0.1293	0.1153
8	0.1293	0.1153

7.2.5 ARIMA Models Fitted to Series A–F

In Table 7.6, we summarize the models fitted by the iterative least-squares procedure of Sections 7.2.1 and 7.2.2 to Series A–F. The models fitted were identified in Chapter 6 and summarized in Tables 6.2 and 6.5. In the case of Series A, C, and D, two possible models were identified and subsequently fitted. For Series A and D, the alternative models involve the use of a stationary autoregressive operator $(1 - \phi B)$ instead of the unit-root operator $(1 - B)$. Examination of Table 7.6 shows that in both cases the autoregressive model results in a slightly smaller residual variance although the models are very similar. Even though a slightly better fit is possible with a stationary model, the IMA(0, 1, 1) model might be

TABLE 7.6 Summary of Models Fitted to Series A–F[a]

Series	Number of Observations	Fitted Models	Residual Variance[b]
A	197	$z_t - 0.92z_{t-1} = 1.45 + a_t - 0.58a_{t-1}$ (± 0.04) (± 0.08)	0.097
		$\nabla z_t = a_t - 0.70a_{t-1}$ (± 0.05)	0.101
B	369	$\nabla z_t = a_t + 0.09a_{t-1}$ (± 0.05)	52.2
C	226	$\nabla z_t - 0.82\nabla z_{t-1} = a_t$ (± 0.04)	0.018
		$\nabla^2 z_t = a_t - 0.13a_{t-1} - 0.12a_{t-2}$ (± 0.07) (± 0.07)	0.019
D	310	$z_t - 0.87z_{t-1} = 1.17 + a_t$ (± 0.03)	0.090
		$\nabla z_t = a_t - 0.06a_{t-1}$ (± 0.06)	0.096
E	100	$z_t = 14.35 + 1.42z_{t-1} - 0.73z_{t-2} + a_t$ (± 0.07) (± 0.07)	227.8
		$z_t = 11.31 + 1.57z_{t-1} - 1.02z_{t-2} + 0.21z_{t-3} + a_t$ (± 0.10) (± 0.15) (± 0.10)	218.1
F	70	$z_t = 58.87 - 0.342z_{t-1} + 0.19z_{t-2} + a_t$ (± 0.12) (± 0.12)	112.7

[a] The values $(\pm)$ under each estimate denote the standard errors of those estimates.

[b] Obtained from $S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})/n$.

preferable in these cases on the grounds that unlike the stationary model, it does not assume that the series has a fixed mean. This is especially important in predicting future values of the series. For if the level does change, a model with $d > 0$ will continue to track it, whereas a model for which $d = 0$ will be tied to a mean level that may have become out of date. It must be noted, however, that for Series D formal unit root testing to be discussed further in Section 10.1 does not support the need for differencing and suggests a preference for the stationary AR(1) model. Also, unit root testing for Series C indicates a preference for the ARIMA(1, 1, 0) model over a model in terms of second differences. Unit root testing for Series A within the ARMA(1, 1) model, though, does not reject the need for the nonstationary operator $(1 - B)$ for the autoregressive part.

The limits under the coefficients in Table 7.6 represent the standard errors of the estimates obtained from the covariance matrix $(\mathbf{X}'_{\hat{\boldsymbol{\beta}}}\mathbf{X}_{\hat{\boldsymbol{\beta}}})^{-1}\hat{\sigma}_a^2$, as described in Section 7.2.1. Note that the estimate $\hat{\phi}_3$ in the AR(3) model, fitted to the sunspot Series E, is 2.1 times its standard error, indicating that a marginally better fit is obtained by the third-order autoregressive process, as compared with the second-order autoregressive process. This is in agreement with a conclusion reached by Moran (1954).

Parameter Estimation Using R. Parameter estimation for ARIMA models based on the methods described above is available in the R software package. The relevant tools include the arima() command in the stats package and the sarima() command in the astsa package. Details of the commands are obtained by typing help(arima) and help(sarima) in R. Using the arima() command, the order of the model is specified using the argument order=c(p,d,q), and the estimation method is specified by method=c("CSS") for conditional least-squares and method=c("ML") for the full maximum likelihood method. The sarima() fits the ARIMA(p, d, q) model to a series z by maximum likelihood using the command sarima(z,p,d,q).

For illustration, we first use the arima() routine in the stats package to estimate the parameters the ARIMA(3, 0, 0) model for the sunspot data in Series E. The relevant command and a partial model output are provided below.

```
> arima(ts(seriesE),order=c(3,0,0),method=c("CSS"))

Coefficients:
         ar1      ar2     ar3  intercept
      1.5519  -1.0069  0.2076    46.7513
s.e.  0.0980   0.1540  0.0981     5.9932

sigma^2 estimated as 219.3:  log-likelihood = -411.42,  aic = NA
```

We see that the estimates of the autoregressive parameters are very close to the values provided in Table 7.6. However, using this routine, the intercept reported in the output is the mean of the series, so that the constant term in the model needs to be calculated as $\hat{\theta}_0 = \hat{\mu}(1 - \hat{\phi}_1 - \hat{\phi}_2 - \hat{\phi}_3)$. This gives an estimate for the constant of 11.57.

The commands and a partial output from performing the analysis using sarima() are as follows:

```
> library(astsa}
> sarima(ts(seriesE),3,0,0)
```

```
Coefficients:
         ar1      ar2     ar3    xmean
      1.5531  -1.0018  0.2063  48.4443
s.e.  0.0981   0.1544  0.0989   6.0706

sigma^2 estimated as 218.2:  log-likelihood=-412.49, aic 834.99
$AIC: [1] 6.465354, $AICc: [1] 6.491737, $BIC: [1] 5.569561
```

The results are close to the earlier ones. The sarima() command has an advantage in that model diagnostics of the type discussed in Chapter 8 below are provided automatically as part of the output (see, e.g., Figures 8.2 and 8.3). This allows the user to efficiently evaluate the adequacy of a fitted model and make comparisons between alternative models. For example, by fitting both the AR(2) and the AR(3) models to the sunspot series, it is readily seen that the AR(3) model provides a better fit to the data. Moreover, the fit can be improved by using a square root or log transformation of the series, although a Q–Q plot still indicates a departure from normality of the standardized residuals.

7.2.6 Large-Sample Information Matrices and Covariance Estimates

In this section, we examine in more detail the information matrix and the covariance matrix of the parameter estimates. Denote by $\mathbf{X} = [\mathbf{U} : \mathbf{V}]$, the $n \times (p+q)$ matrix of the time lagged $u_t's$ and $v_t's$ defined in (7.2.13) and (7.2.14), when the elements of $\boldsymbol{\beta}_0$ are the *true* values of the parameters, for a sample size n sufficiently large for end effects to be ignored. Then, since $x_{t,i} = -\partial[a_t]/\partial\beta_i$ and using (7.1.20),

$$E[l_{ij}] \simeq -\frac{1}{2\sigma_a^2} E\left[\frac{\partial^2 S(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j}\right] = -\frac{1}{\sigma_a^2} E\left[\sum_{t=1}^{n} \frac{\partial[a_t]}{\partial\beta_i}\frac{\partial[a_t]}{\partial\beta_j}\right] = -\frac{1}{\sigma_a^2} E\left[\sum_{t=1}^{n} x_{t,i}x_{t,j}\right]$$

the information matrix for $(\boldsymbol{\phi}, \boldsymbol{\theta})$ for the mixed ARMA model is

$$\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta}) = E[\mathbf{X}'\mathbf{X}]\sigma_a^{-2} = E\begin{bmatrix}\mathbf{U}'\mathbf{U} & \mathbf{U}'\mathbf{V} \\ \mathbf{V}'\mathbf{U} & \mathbf{V}'\mathbf{V}\end{bmatrix}\sigma_a^{-2} \tag{7.2.17}$$

that is,

$$= n\sigma_a^{-2}\left[\begin{array}{ccc|ccc} \gamma_{uu}(0) & \gamma_{uu}(1) & \cdots \gamma_{uu}(p-1) & \gamma_{uv}(0) & \gamma_{uv}(-1) & \cdots \gamma_{uv}(1-q) \\ \gamma_{uu}(1) & \gamma_{uu}(0) & \cdots \gamma_{uu}(p-2) & \gamma_{uv}(1) & \gamma_{uv}(0) & \cdots \gamma_{uv}(2-q) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{uu}(p-1) & \gamma_{uu}(p-2) & \cdots \gamma_{uu}(0) & \gamma_{uv}(p-1) & \gamma_{uv}(p-2) & \cdots \gamma_{uv}(p-q) \\ \hline \gamma_{uv}(0) & \gamma_{uv}(1) & \cdots \gamma_{uv}(p-1) & \gamma_{vv}(0) & \gamma_{vv}(1) & \cdots \gamma_{vv}(q-1) \\ \gamma_{uv}(-1) & \gamma_{uv}(0) & \cdots \gamma_{uv}(p-2) & \gamma_{vv}(1) & \gamma_{vv}(0) & \cdots \gamma_{vv}(q-2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma_{uv}(1-q) & \gamma_{uv}(2-q) & \cdots \gamma_{uv}(p-q) & \gamma_{vv}(q-1) & \gamma_{vv}(q-2) & \cdots \gamma_{vv}(0) \end{array}\right] \tag{7.2.18}$$

where $\gamma_{uu}(k)$ and $\gamma_{vv}(k)$ are the autocovariances for the u_t's and the v_t's, and $\gamma_{uv}(k)$ are the cross-covariances defined by

$$\gamma_{uv}(k) = \gamma_{vu}(-k) = E[u_t v_{t+k}] = E[v_t u_{t-k}]$$

The large-sample covariance matrix for the maximum likelihood estimates may be obtained using

$$\mathbf{V}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}) \simeq \mathbf{I}^{-1}(\boldsymbol{\phi}, \boldsymbol{\theta})$$

Estimates of $\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta})$ and hence of $\mathbf{V}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ may be obtained by evaluating the u_t's and v_t's with $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}$ and omitting the expectation sign in (7.2.17) leading to $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}) = (\boldsymbol{X}'_{\hat{\beta}} \boldsymbol{X}_{\hat{\beta}})^{-1} \hat{\sigma}_a^2$, or by substituting standard sample estimates of the autocovariances and cross-covariances in (7.2.18). Theoretical large-sample results can be obtained by noticing that, with the elements of $\boldsymbol{\beta}_0$ equal to the true values of the parameters, equations (7.2.13) and (7.2.14) imply that the derived series u_t and v_t follow *autoregressive* processes defined by

$$\phi(B) u_t = a_t \qquad \theta(B) v_t = -a_t$$

It follows that the autocovariances that appear in (7.2.18) are those for pure autoregressive processes, and the cross-covariances are the negative of those between two such processes generated by the same a_t's.

We illustrate the use of this result with a few examples.

Covariance Matrix of Parameter Estimates for AR(p) and MA(q) Processes. Let $\boldsymbol{\Gamma}_p(\boldsymbol{\phi})$ be the $p \times p$ autocovariance matrix of p successive observations from an AR(p) process with parameters $\boldsymbol{\phi}' = (\phi_1, \phi_2, \ldots, \phi_p)$. Then, using (7.2.18), the $p \times p$ covariance matrix of the estimates $\hat{\boldsymbol{\phi}}$ is given by

$$\mathbf{V}(\hat{\boldsymbol{\phi}}) \simeq n^{-1} \sigma_a^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi}) \tag{7.2.19}$$

Let $\boldsymbol{\Gamma}_q(\boldsymbol{\theta})$ be the $q \times q$ autocovariance matrix of q successive observations from an AR(q) process with parameters $\boldsymbol{\theta}' = (\theta_1, \theta_2, \ldots, \theta_q)$. Then, using (7.2.18), the $q \times q$ covariance matrix of the estimates $\hat{\boldsymbol{\theta}}$ in an MA(q) model is

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) \simeq n^{-1} \sigma_a^2 \boldsymbol{\Gamma}_q^{-1}(\boldsymbol{\theta}) \tag{7.2.20}$$

Covariances for the Zeros of an ARMA Process. It is occasionally useful to parameterize an ARMA process in terms of the zeros of $\phi(B)$ and $\theta(B)$. In this case, a particularly simple form is obtained for the covariance matrix of the parameter estimates.

Consider the ARMA(p, q) process parameterized in terms of its zeros (assumed to be real and distinct), so that

$$\prod_{i=1}^{p} (1 - G_i B) \tilde{w}_t = \prod_{j=1}^{q} (1 - H_j B) a_t$$

or

$$a_t = \prod_{i=1}^{p}(1 - G_i B) \prod_{j=1}^{q}(1 - H_j B)^{-1} \tilde{w}_t$$

The derivatives of the a_t's are then such that

$$u_{t,i} = -\frac{\partial a_t}{\partial G_i} = (1 - G_i B)^{-1} a_{t-1}$$
$$v_{t,j} = -\frac{\partial a_t}{\partial H_j} = -(1 - H_j B)^{-1} a_{t-1}$$

Hence, using (7.2.18), for large samples, the information matrix for the roots is such that

$n^{-1}\mathbf{I}(\mathbf{G}, \mathbf{H})$

$$= \left[\begin{array}{cccc|cccc} (1-G_1^2)^{-1} & (1-G_1G_2)^{-1} & \cdots & (1-G_1G_p)^{-1} & -(1-G_1H_1)^{-1} & \cdots & -(1-G_1H_q)^{-1} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ (1-G_1G_p)^{-1} & (1-G_2G_p)^{-1} & \cdots & (1-G_p^2)^{-1} & -(1-G_pH_1)^{-1} & \cdots & -(1-G_pH_q)^{-1} \\ \hline -(1-G_1H_1)^{-1} & -(1-G_2H_1)^{-1} & \cdots & -(1-G_pH_1)^{-1} & (1-H_1^2)^{-1} & \cdots & (1-H_1H_q)^{-1} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ -(1-G_1H_q)^{-1} & -(1-G_2H_q)^{-1} & \cdots & -(1-G_pH_q)^{-1} & (1-H_1H_q)^{-1} & \cdots & (1-H_q^2)^{-1} \end{array}\right] \tag{7.2.21}$$

Examples: For an AR(2) process $(1 - G_1 B)(1 - G_2 B)\tilde{w}_t = a_t$, we have

$$\mathbf{V}(\hat{G}_1, \hat{G}_2) \simeq n^{-1}\begin{bmatrix} (1-G_1^2)^{-1} & (1-G_1G_2)^{-1} \\ (1-G_1G_2)^{-1} & (1-G_2^2)^{-1} \end{bmatrix}^{-1}$$
$$= \frac{1}{n}\frac{1 - G_1G_2}{(G_1 - G_2)^2}\begin{bmatrix} (1-G_1^2)(1-G_1G_2) & -(1-G_1^2)(1-G_2^2) \\ -(1-G_1^2)(1-G_2^2) & (1-G_2^2)(1-G_1G_2) \end{bmatrix} \tag{7.2.22}$$

Exactly parallel results will be obtained for a second-order moving average process.

Similarly, for the ARMA(1,1) process $(1 - \phi B)\tilde{w}_t = (1 - \theta B)a_t$, on setting $\phi = G_1$ and $\theta = H_1$ in (7.2.21), we obtain

$$\mathbf{V}(\hat{\phi}, \hat{\theta}) \simeq n^{-1}\begin{bmatrix} (1-\phi^2)^{-1} & -(1-\phi\theta)^{-1} \\ -(1-\phi\theta)^{-1} & (1-\theta^2)^{-1} \end{bmatrix}^{-1}$$
$$= \frac{1}{n}\frac{1 - \phi\theta}{(\phi - \theta)^2}\begin{bmatrix} (1-\phi^2)(1-\phi\theta) & (1-\phi^2)(1-\theta^2) \\ (1-\phi^2)(1-\theta^2) & (1-\theta^2)(1-\phi\theta) \end{bmatrix} \tag{7.2.23}$$

The results for these two processes illustrate a duality property between the information matrices for the autoregressive model and the general ARMA(p, q) model. Namely, suppose

that the information matrix for parameters **(G, H)** of the ARMA(p, q) model

$$\prod_{i=1}^{p}(1 - G_i B)\tilde{w}_t = \prod_{j=1}^{q}(1 - H_j B)a_t$$

is denoted as $\mathbf{I}\{\mathbf{G}, \mathbf{H}|(p, q)\}$, and suppose, correspondingly, that the information matrix for the parameters **(G, H)** in the pure AR$(p + q)$ model

$$\prod_{i=1}^{p}(1 - G_i B)\prod_{j=1}^{q}(1 - H_j B)\tilde{w}_t = a_t$$

is denoted as

$$\mathbf{I}\{\mathbf{G}, \mathbf{H}|(p + q, 0)\} = \left[\begin{array}{c|c} \mathbf{I}_{GG} & \mathbf{I}_{GH} \\ \hline \mathbf{I}'_{GH} & \mathbf{I}_{HH} \end{array}\right]$$

where the matrix is partitioned after the pth row and column. Then, for moderate and large samples, we can see directly from (7.2.21) that

$$\mathbf{I}\{\mathbf{G}, \mathbf{H}|(p, q)\} \simeq \mathbf{I}\{\mathbf{G}, -\mathbf{H}|(p + q, 0)\} = \left[\begin{array}{c|c} \mathbf{I}_{GG} & -\mathbf{I}_{GH} \\ \hline -\mathbf{I}'_{GH} & \mathbf{I}_{HH} \end{array}\right] \tag{7.2.24}$$

Hence, since for moderate and large samples, the inverse of the information matrix provides a close approximation to the covariance matrix $\mathbf{V}(\hat{\mathbf{G}}, \hat{\mathbf{H}})$ of the parameter estimates, we have, correspondingly,

$$\mathbf{V}\{\hat{\mathbf{G}}, \hat{\mathbf{H}}|(p, q)\} \simeq \mathbf{V}\{\hat{\mathbf{G}}, -\hat{\mathbf{H}}|(p + q, 0)\} \tag{7.2.25}$$

7.3 SOME ESTIMATION RESULTS FOR SPECIFIC MODELS

In Appendices A7.3, A7.4, and A7.5, some estimation results for special cases are derived. These, and results obtained earlier in this chapter, are summarized here for reference.

7.3.1 Autoregressive Processes

It is possible to obtain estimates of the parameters of a pure autoregressive process by solving *certain linear* equations. We show in Appendix A7.4:

1. How exact least-squares estimates may be obtained by solving a linear system of equations (see also Section 7.5.3).
2. How, by slight modification of the coefficients in these equations, a close approximation to the exact maximum likelihood equations may be obtained.
3. How conditional least-squares estimates, as defined in Section 7.1.3, may be obtained by solving a system of linear equations of the form of the standard linear regression model normal equations.

4. How estimates that are approximations to the least-squares estimates and to the maximum likelihood estimates may be obtained using the estimated autocorrelations as coefficients in the linear Yule–Walker equations.

The estimates obtained in item 1 are, of course, identical with those given by direct minimization of $S(\boldsymbol{\phi})$, as described in general terms in Section 7.2. The estimates in 4 are the well-known approximations due to Yule and Walker. They are useful as first estimates at the identification stage but can differ appreciably from estimates 1, 2, or 3, in some cases. For instance, differences can occur for an AR(2) model if the parameter estimates $\hat{\phi}_1$ and $\hat{\phi}_2$ are highly correlated, as is the case for the AR(2) model fitted to Series E in Table 7.6.

Yule–Walker Estimates. The Yule–Walker estimates (6.3.6) are

$$\hat{\boldsymbol{\phi}} = \mathbf{R}^{-1}\boldsymbol{r}$$

where

$$\mathbf{R} = \begin{bmatrix} 1 & r_1 & \cdots & r_{p-1} \\ r_1 & 1 & \cdots & r_{p-2} \\ \vdots & \vdots & & \vdots \\ r_{p-1} & r_{p-2} & \cdots & 1 \end{bmatrix} \qquad \mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{bmatrix} \tag{7.3.1}$$

In particular, the estimates for the AR(1) and the AR(2) processes are

$$\begin{aligned} &\text{AR(1)}: \qquad \hat{\phi}_1 = r_1 \\ &\text{AR(2)}: \qquad \hat{\phi}_1 = \frac{r_1(1-r_2)}{1-r_1^2} \qquad \hat{\phi}_2 = \frac{r_2 - r_1^2}{1-r_1^2} \end{aligned} \tag{7.3.2}$$

It is shown in Appendix A7.4 that an approximation to $S(\hat{\boldsymbol{\phi}})$ is provided by

$$S(\hat{\boldsymbol{\phi}}) = \sum_{t=1}^{n} \tilde{w}_t^2 (1 - \mathbf{r}'\hat{\boldsymbol{\phi}}) \tag{7.3.3}$$

so that

$$\hat{\sigma}_a^2 = \frac{S(\hat{\boldsymbol{\phi}})}{n} = c_0(1 - \mathbf{r}'\hat{\boldsymbol{\phi}}) \tag{7.3.4}$$

where c_0 is the sample variance of the w_t's. A parallel expression relates σ_a^2 and γ_0, the theoretical variance of the w_t's [see (3.2.8)], namely,

$$\sigma_a^2 = \gamma_0(1 - \boldsymbol{\rho}'\boldsymbol{\phi})$$

where the elements of $\boldsymbol{\rho}$ and of $\boldsymbol{\phi}$ are the theoretical values. Thus, from (7.2.19) and Appendix A7.5, the covariance matrix for the estimates $\hat{\boldsymbol{\phi}}$ is

$$\boldsymbol{V}(\hat{\boldsymbol{\phi}}) \simeq n^{-1}\sigma_a^2\boldsymbol{\Gamma}^{-1} = n^{-1}(1 - \boldsymbol{\rho}'\boldsymbol{\phi})\boldsymbol{P}^{-1} \tag{7.3.5}$$

where $\boldsymbol{\Gamma}$ and $\boldsymbol{P} = (1/\gamma_0)\boldsymbol{\Gamma}$ are the autocovariance and autocorrelation matrices of p successive values of the AR(p) process.

In particular, for the AR(1) and AR(2) processes, we find that

$$\text{AR(1)}: \qquad V(\hat{\phi}) \simeq n^{-1}(1-\phi^2) \tag{7.3.6}$$

$$\text{AR(2)}: \boldsymbol{V}(\hat{\phi}_1, \hat{\phi}_2) \simeq n^{-1} \begin{bmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{bmatrix} \tag{7.3.7}$$

Estimates of the variances and covariances are obtained by substituting estimates of the parameters in (7.3.5). Thus,

$$\boldsymbol{V}(\hat{\boldsymbol{\phi}}) = n^{-1}(1 - \boldsymbol{r}'\hat{\boldsymbol{\phi}})\boldsymbol{R}^{-1} \tag{7.3.8}$$

Using (7.3.7) it is readily shown that the correlation between the estimates of the AR(2) parameters is approximately equal to $-\rho_1$. This implies, in particular, that a large lag-1 correlation in the series can give rise to unstable estimates, which may explain the differences between the Yule–Walker and the least squares estimates noted above.

7.3.2 Moving Average Processes

Maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ for moving average processes may, in simple cases, be obtained graphically, as illustrated in Section 7.1.6, or more generally, by the iterative calculation described in Section 7.2.1. From (7.2.20), it follows that for moderate and large samples, the covariance matrix for the estimates of the parameters of a qth-order moving average process is of the same form as the corresponding matrix for an autoregressive process of the same order. Thus, for the MA(1) and MA(2) processes, we find, corresponding to (7.3.6) and (7.3.7)

$$\text{MA(1)}: \qquad \mathbf{V}(\hat{\theta}) \simeq n^{-1}(1-\theta^2) \tag{7.3.9}$$

$$\text{MA(2)}: \quad \mathbf{V}(\hat{\theta}_1, \hat{\theta}_2) \simeq n^{-1} \begin{bmatrix} 1-\theta_2^2 & -\theta(1+\theta_2) \\ -\theta_1(1+\theta_2) & 1-\theta_2^2 \end{bmatrix} \tag{7.3.10}$$

7.3.3 Mixed Processes

Maximum likelihood estimates $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ for mixed processes, as for moving average processes, may be obtained graphically in simple cases, and more generally, by iterative calculation. For moderate and large samples, the covariance matrix may be obtained by evaluating and inverting the information matrix (7.2.18). In the important special case of the ARMA(1, 1) process

$$(1-\phi B)\tilde{w}_t = (1-\theta B)a_t$$

we obtain, as in (7.2.23),

$$\mathbf{V}\,(\hat{\phi}, \hat{\theta}) \simeq n^{-1} \frac{1-\phi\theta}{(\phi-\theta)^2} \begin{bmatrix} (1-\phi^2)(1-\phi\theta) & (1-\phi^2)(1-\theta^2) \\ (1-\phi^2)(1-\theta^2) & (1-\theta^2)(1-\phi\theta) \end{bmatrix} \tag{7.3.11}$$

It is noted that when $\phi = \theta$, the variances of $\hat{\phi}$ and $\hat{\theta}$ are infinite. This is to be expected, for in this case the factor $(1-\phi B) = (1-\theta B)$ cancels on both sides of the model, which becomes

$$\tilde{w}_t = a_t$$

This is a particular case of *parameter redundancy*, which we discuss further in Section 7.3.5.

7.3.4 Separation of Linear and Nonlinear Components in Estimation

It is occasionally of interest to make an analysis in which the estimation of the parameters of the mixed model is separated into its basic linear and nonlinear parts. Consider the general mixed model $\phi(B)\tilde{w}_t = \theta(B)a_t$, which we write as $a_t = \phi(B)\theta^{-1}(B)\tilde{w}_t$, or

$$a_t = \phi(B)(\varepsilon_t|\boldsymbol{\theta}) \tag{7.3.12}$$

where

$$(\varepsilon_t|\boldsymbol{\theta}) = \theta^{-1}(B)\tilde{w}_t$$

that is,

$$\tilde{w}_t = \theta(B)(\varepsilon_t|\boldsymbol{\theta}) \tag{7.3.13}$$

For any given set of θ's, the ε_t's may be calculated recursively from (7.3.13), which may be written as

$$\varepsilon_t = \tilde{w}_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}$$

The recursion may be started by setting unknown ε_t's equal to zero. Having calculated the ε_t's, the conditional estimates $\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}$ may readily be obtained. These are the estimated autoregressive parameters in the linear model (7.3.12), which may be written as

$$a_t = \varepsilon_t - \phi_1\varepsilon_{t-1} - \phi_2\varepsilon_{t-2} - \cdots - \phi_p\varepsilon_{t-p} \tag{7.3.14}$$

As discussed in Section 7.3.1, the least-squares estimates of the autoregressive parameters may be found by direct solution of a set of linear equations. In simple cases, we can examine the behavior of $S(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \boldsymbol{\theta})$ and find its minimum by computing $S(\hat{\boldsymbol{\phi}}_{\boldsymbol{\theta}}, \boldsymbol{\theta})$ on a grid of $\boldsymbol{\theta}$ values and plotting contours.

Example Using Series C. One possible model for Series C considered earlier is the ARIMA(1, 1, 0) model $(1 - \phi B)w_t = a_t$ with $w_t = \nabla z_t$ and $E[w_t] = 0$. Consider now the somewhat more elaborate model $(1 - \phi B)w_t = (1 - \theta_1 B - \theta_2 B^2)a_t$. Following the argument given above, the process may be thought of as resulting from a combination of the nonlinear model $\varepsilon_t = w_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$ and the linear model $a_t = \varepsilon_t - \phi\varepsilon_{t-1}$.

For each choice of the nonlinear parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$ within the invertibility region, a set of ε_t's was calculated recursively. Using the Yule–Walker approximation, an estimate $\hat{\phi}_{\boldsymbol{\theta}} = r_1(\varepsilon)$ could now be obtained together with

$$S(\hat{\phi}_{\boldsymbol{\theta}}, \boldsymbol{\theta}) \simeq \sum_{t=1}^{n} \varepsilon_t^2[1 - r_1^2(\varepsilon)]$$

This sum of squares was plotted for a grid of values of θ_1 and θ_2 and its contours are shown in Figure 7.6. We see that a minimum close to $\theta_1 = \theta_2 = 0$ is indicated, at which point $r_1(\varepsilon) = 0.805$. Thus, within the whole class of models of order (1, 1, 2), the simple (1, 1, 0) model $(1 - 0.8B)\nabla z_t = a_t$ is confirmed to provide an adequate representation.

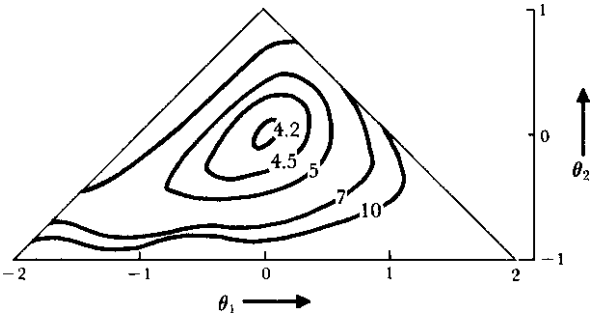


FIGURE 7.6 Counters of $S(\hat{\phi}_\theta, \theta)$ for Series C plotted over the admissible parameter space for the θ's.

7.3.5 Parameter Redundancy

The model $\phi(B)\tilde{w}_t = \theta(B)a_t$ is identical to the model

$$(1 - \alpha B)\phi(B)\tilde{w}_t = (1 - \alpha B)\theta(B)a_t$$

in which both autoregressive and moving average operators are multiplied by the same factor, $1 - \alpha B$. Serious difficulties in the estimation procedure will arise if a model is fitted that contains a redundant factor. Therefore, care is needed in avoiding the situation where redundant or near-redundant *common* factors occur. The existence of redundancy is not always obvious. For example, one can see the common factor in the ARMA(2, 1) model

$$(1 - 1.3B + 0.4B^2)\tilde{w}_t = (1 - 0.5B)a_t$$

only after factoring the left-hand side to obtain

$$(1 - 0.5B)(1 - 0.8B)\tilde{w}_t = (1 - 0.5B)a_t$$

that is, $(1 - 0.8B)\tilde{w}_t = a_t$.

In practice, it is not just exact cancellation that causes difficulties, but also near-cancellation. For example, suppose that the true model was

$$(1 - 0.4B)(1 - 0.8B)\tilde{w}_t = (1 - 0.5B)a_t \tag{7.3.15}$$

If an attempt was made to fit this model as ARMA(2, 1), extreme instability in the parameter estimates could arise because of near-cancellation of the factors $(1 - 0.4B)$ and $(1 - 0.5B)$, on the left- and right-hand sides. In this case, combinations of parameter values yielding similar $[a_t]$'s and so similar likelihoods can be found, and a change of parameter value on the left can be nearly compensated by a suitable change on the right. The sum-of-squares contour surfaces in the three-dimensional parameter space will thus approach obliquely oriented cylinders, and a line of ''near least-squares'' solutions rather than a clearly defined point minimum will be found.

From a slightly different viewpoint, we can write the model (7.3.15) in terms of an infinite autoregressive operator. Making the necessary expansion, we find that

$$(1 - 0.700B - 0.030B^2 - 0.015B^3 - 0.008B^4 - \cdots)\tilde{w}_t = a_t$$

Thus, very nearly, the model is

$$(1 - 0.7B)\tilde{w}_t = a_t \tag{7.3.16}$$

The instability of the estimates, obtained by attempting to fit an ARMA(2, 1) model, would occur because we would be trying to fit three parameters in a situation that could almost be represented by one.

A principal reason for going through the identification procedure prior to fitting the model is to avoid difficulties arising from parameter redundancy and to achieve *parsimony* in parameterization.

Redundancy in the ARMA(1,1) Model. The simplest model where the possibility occurs for direct cancellation of factors is the ARMA(1, 1) process:

$$(1 - \phi B)\tilde{w}_t = (1 - \theta B)a_t$$

In particular, if $\phi = \theta$, then whatever common value they have, $\tilde{w}_t = a_t$, so that $\tilde{w}_t$ is generated by a white noise process. The data then cannot supply information about the common parameter, and using (7.3.11), $\hat{\phi}$ and $\hat{\theta}$ have infinite variances. Furthermore, whatever the values of ϕ and θ, $S(\phi, \theta)$ must be constant on the line $\phi = \theta$. This is illustrated in Figure 7.7, which shows a sum-of-squares plot for the data of Series A. However, for these data, the least-squares values $\hat{\phi} = 0.92$ and $\hat{\theta} = 0.58$ correspond to a point that is not particularly close to the line $\phi = \theta$, and no difficulties occur in the estimation of these parameters.

In practice, if the identification technique we have recommended is adopted, these difficulties will be avoided. An ARMA(1, 1) process in which ϕ is very nearly equal to θ will normally be identified as white noise, or if the difference is nonnegligible, as an AR(1) or MA(1) process with a single small coefficient.

In summary:

1. We should avoid mixed models containing near common factors, and we should be alert to the difficulties that can result.

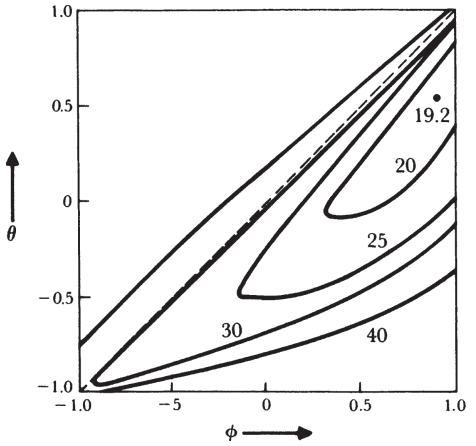


FIGURE 7.7 Sum-of-squares plot for Series A.

2. We will automatically avoid such models if we use identification and estimation procedures intelligently.

7.4 LIKELIHOOD FUNCTION BASED ON THE STATE-SPACE MODEL

In Section 5.5, we introduced the state-space model formulation of the ARMA process along with Kalman filtering and described its use for prediction. This approach also provides a convenient method to evaluate the exact likelihood function for an ARMA model. The use of this approach has been suggested by Jones (1980), Gardner et al. (1980), and others.

The state-space model form of the ARMA(p, q) model given in Section 5.5 is

$$\boldsymbol{Y}_t = \boldsymbol{\Phi}\boldsymbol{Y}_{t-1} + \boldsymbol{\Psi}a_t \quad \text{and} \quad w_t = \mathbf{H}\boldsymbol{Y}_t \tag{7.4.1}$$

where $\boldsymbol{Y}_t' = (w_t, \hat{w}_t(1), \ldots, \hat{w}_t(r-1))$, $r = \max(p, q+1)$, $\mathbf{H} = (1, 0, \ldots, 0)$,

$$\boldsymbol{\Phi} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ \phi_r & \phi_{r-1} & \cdots & \cdots & \phi_1 \end{bmatrix}$$

and $\boldsymbol{\Psi}' = (1, \psi_1, \ldots, \psi_{r-1})$. The Kalman filter equations (5.5.6)–(5.5.9) provide one-step-ahead forecasts $\hat{\boldsymbol{Y}}_{t|t-1} = E[\boldsymbol{Y}_t | w_{t-1}, \ldots, w_1]$ of the state vector $\boldsymbol{Y}_t$ and the error covariance matrix $\mathbf{V}_{t|t-1} = E[(\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1})(\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_{t|t-1})']$. Specifically, for the state–space form of the ARMA(p, q) model, these recursive equations are

$$\hat{\boldsymbol{Y}}_{t|t} = \hat{\boldsymbol{Y}}_{t|t-1} + \mathbf{K}_t(w_t - \hat{w}_{t|t-1}) \quad \text{with} \quad \mathbf{K}_t = \mathbf{V}_{t|t-1}\mathbf{H}'[\mathbf{H}\mathbf{V}_{t|t-1}\mathbf{H}']^{-1} \tag{7.4.2}$$

where $\hat{w}_{t|t-1} = \mathbf{H}\hat{\boldsymbol{Y}}_{t|t-1}$, and

$$\hat{\boldsymbol{Y}}_{t|t-1} = \boldsymbol{\Phi}\hat{\boldsymbol{Y}}_{t-1|t-1} \qquad \mathbf{V}_{t|t-1} = \boldsymbol{\Phi}\mathbf{V}_{t-1|t-1}\boldsymbol{\Phi}' + \sigma_a^2\boldsymbol{\Psi}\boldsymbol{\Psi}' \tag{7.4.3}$$

with

$$\mathbf{V}_{t|t} = [\mathbf{I} - \mathbf{K}_t\mathbf{H}]\mathbf{V}_{t|t-1} \tag{7.4.4}$$

for $t = 1, 2, \ldots, n$. In particular, then, the first component of the forecast vector is $\hat{w}_{t|t-1} = \mathbf{H}\hat{\boldsymbol{Y}}_{t|t-1} = E[w_t | w_{t-1}, \ldots, w_1]$, $a_{t|t-1} = w_t - \hat{w}_{t|t-1}$ is the one-step innovation, and the element $\sigma_a^2 v_t = \mathbf{H}\mathbf{V}_{t|t-1}\mathbf{H}' = E[(w_t - \hat{w}_{t|t-1})^2]$ is the one-step forecast error variance.

To obtain the exact likelihood function of the vector of n observations $\mathbf{w}' = (w_1, w_2, \ldots, w_n)$ using the above results, we note that the joint distribution of $\mathbf{w}$ can be factored as

$$p(\mathbf{w}|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = \prod_{t=1}^{n} p(w_t | w_{t-1}, \ldots, w_1; \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) \tag{7.4.5}$$

where $p(w_t|w_{t-1}, \ldots, w_1; \boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2)$ denotes the conditional distribution of w_t given $w_{t-1}, \ldots, w_t$. Under normality of a_t, this conditional distribution is normal with conditional mean $\hat{w}_{t|t-1} = E[w_t|w_{t-1}, \ldots \ldots, w_1]$ and conditional variance $\sigma_a^2 v_t = E[(w_t - \hat{w}_{t|t-1})^2]$. Hence, the joint distribution of $\mathbf{w}$ can be conveniently expressed as

$$p(\mathbf{w}|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = \prod_{t=1}^{n} (2\pi\sigma_a^2 v_t)^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2} \sum_{t=1}^{n} \frac{(w_t - \hat{w}_{t|t-1})^2}{v_t}\right] \tag{7.4.6}$$

where the quantities $\hat{w}_{t|t-1}$ and $\sigma_a^2 v_t$ are easily determined recursively from the Kalman filter procedure. The initial values needed to start the Kalman filter recursions are given by $\hat{\boldsymbol{Y}}_{0|0} = \mathbf{0}$, an r-dimensional vector of zeros, and $\mathbf{V}_{0|0} = \text{cov}[\boldsymbol{Y}_0]$. The elements of $\mathbf{V}_{0|0}$ can readily be determined as a function of the autocovariances γ_k and the weights ψ_k of the ARMA(p, q) process w_t, making use of the relation $w_{t+j} = \hat{w}_t(j) + \sum_{k=0}^{j-1} \psi_k a_{t+j-k}$ from Chapter 5. See Jones (1980) for further details. For example, in the case of an ARMA(1, 1) model for w_t, we have $\boldsymbol{Y}_t' = (w_t, \hat{w}_t(1))$, so

$$\mathbf{V}_{0|0} = \text{cov}[\boldsymbol{Y}_0] = \begin{bmatrix} \gamma_0 & \gamma_1 \\ \gamma_1 & \gamma_0 - \sigma_a^2 \end{bmatrix} = \sigma_a^2 \begin{bmatrix} \sigma_a^{-2}\gamma_0 & \sigma_a^{-2}\gamma_1 \\ \sigma_a^{-2}\gamma_1 & \sigma_a^{-2}\gamma_0 - 1 \end{bmatrix}$$

It also is generally the case that the one-step-ahead forecasts $\hat{w}_{t|t-1}$ and the corresponding error variances $\sigma_a^2 v_t$ rather quickly approach their steady-state forms, in which case the Kalman filter calculations at some stage (beyond time t_0, say) could be switched to the simpler form $\hat{w}_{t|t-1} = \sum_{i=1}^{p} \phi_i w_{t-i} - \sum_{i=1}^{q} \theta_i a_{t-i|t-i-1}$, and $\sigma_a^2 v_t = \text{var}[a_{t|t-1}] = \sigma_a^2$, for $t > t_0$, where $a_{t|t-1} = w_t - \hat{w}_{t|t-1}$. For example, refer to Gardner et al. (1980) for further details. On comparison of (7.4.6) with expressions given earlier in (7.1.5) and (7.1.6), and also (A7.3.11) and (A7.3.13), the unconditional sum-of-squares function can be represented in two equivalent forms as

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} [a_t]^2 + \hat{\mathbf{e}}_*' \boldsymbol{\Omega}^{-1} \hat{\mathbf{e}}_* = \sum_{t=1}^{n} \frac{a_{t|t-1}^2}{v_t}$$

where $a_{t|t-1} = w_t - \hat{w}_{t|t-1}$, and also $|\mathbf{M}_n^{(p,q)}|^{-1} = |\boldsymbol{\Omega}||\mathbf{D}| = \prod_{t=1}^{n} v_t$.

Innovations Method. The likelihood function expressed in the form of (7.4.6) is generally referred to as the *innovations form*, and the quantities $a_{t|t-1} = w_t - \hat{w}_{t|t-1}, t = 1, \ldots, n$, are the (finite-sample) *innovations*. Calculation of the likelihood function in this form, based on the state-space representation of the ARMA process and associated Kalman filtering algorithms, has been proposed by many authors including Gardner et al. (1980), Harvey and Phillips (1979), and Jones (1980). The innovations form of the likelihood can also be obtained without directly using the state-space representation through the use of an ''innovations algorithm'' (e.g., see Ansley, 1979; Brockwell and Davis, 1991). This method essentially involves a Cholesky decomposition of an $n \times n$ band covariance matrix of the derived MA(q) process:

$$w_t' = w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

More specifically, using the notation of Appendix A7.3, we write the ARMA model relations for n observations as $\mathbf{L}_\phi \mathbf{w} = \mathbf{L}_\theta \mathbf{a} + \mathbf{F}\mathbf{e}_*$, where $\mathbf{a}' = (a_1, a_2, \ldots, a_n)$ and $\mathbf{e}_*' =$

$(w_{1-p}, \ldots, w_0, a_{1-q}, \ldots, a_0)$ is the $(p+q)$-dimensional vector of pre-sample values. Then, the covariance matrix of the vector of derived variables $\mathbf{L}_\phi \mathbf{w}$ is

$$\mathbf{\Gamma}_{w'} = \text{cov}[\mathbf{L}_\phi \mathbf{w}] = \text{cov}[\mathbf{L}_\theta \mathbf{a} + \mathbf{F}\mathbf{e}_*] = \sigma_a^2(\mathbf{L}_\theta \mathbf{L}_\theta' + \mathbf{F}\mathbf{\Omega}\mathbf{F}') \tag{7.4.7}$$

which is a *band matrix*. That is, $\mathbf{\Gamma}_{w'}$ is a matrix with nonzero elements only in a band about the main diagonal of maximum bandwidth $m = \max(p, q)$, and of bandwidth q after the first m rows since $\text{cov}[w_t', w_{t+j}'] = 0$ for $j > q$. The innovations algorithm obtains the (square-root-free) Cholesky decomposition of the band matrix $\mathbf{L}_\theta \mathbf{L}_\theta' + \mathbf{F}\mathbf{\Omega}\mathbf{F}'$ as $\mathbf{GDG}'$, where $\mathbf{G}$ is a lower triangular band matrix with bandwidth corresponding to that of $\mathbf{\Gamma}_{w'}$ and with ones on the diagonal, and $\mathbf{D}$ is a diagonal matrix with positive diagonal elements $v_t, t = 1, \ldots, n$. Hence, $\text{cov}[\mathbf{w}] = \sigma_a^2 \mathbf{L}_\phi^{-1} \mathbf{GDG}' \mathbf{L}_\phi'^{-1}$ and the quadratic form in the exponent of the likelihood function (7.4.6) is

$$\begin{aligned} \mathbf{w}'\{\text{cov}[\mathbf{w}]\}^{-1}\mathbf{w} &= \frac{1}{\sigma_a^2}\mathbf{w}'(\mathbf{L}_\phi^{-1}\mathbf{GDG}'\mathbf{L}_\phi'^{-1})^{-1}\mathbf{w} \\ &= \frac{1}{\sigma_a^2}\mathbf{e}'\mathbf{D}^{-1}\mathbf{e} = \frac{1}{\sigma_a^2}\sum_{t=1}^{n}\frac{a_{t|t-1}^2}{v_t} \end{aligned} \tag{7.4.8}$$

where $\mathbf{e} = \mathbf{G}^{-1}\mathbf{L}_\phi \mathbf{w} = (a_{1|0}, a_{2|1}, \ldots, a_{n|n-1})'$ is the vector of innovations, which are computed recursively from $\mathbf{Ge} = \mathbf{L}_\phi \mathbf{w}$. Thus, the innovations can be obtained recursively as $a_{1|0} = w_1, a_{2|1} = w_2 - \phi_1 w_1 + \theta_{1,1} a_{1|0}, \ldots, a_{m|m-1} = w_m - \sum_{i=1}^{m-1} \phi_i w_{m-i} + \sum_{i=1}^{m-1} \theta_{i,m-1} a_{m-i|m-i-1}$, and

$$a_{t|t-1} = w_t - \sum_{i=1}^{p} \phi_i w_{t-i} + \sum_{i=1}^{q} \theta_{i,t-1} a_{t-i|t-i-1} \tag{7.4.9}$$

for $t > m$, where the tth row of the matrix $\mathbf{G}$ has the form

$$[0, \ldots, 0, -\theta_{q,t-1}, \ldots, -\theta_{1,t-1}, 1, 0, \ldots, 0]$$

with the 1 in the tth (i.e., diagonal) position. In addition, the coefficients $\theta_{i,t-1}$ in (7.4.9) and the diagonal (variance) elements v_t are obtained recursively through the Cholesky decomposition procedure. In particular, the v_t are given by the recursion

$$v_t = \frac{\gamma_0(w')}{\sigma_a^2} - \sum_{j=1}^{q} \theta_{j,t-1}^2 v_{t-j} \qquad \text{for} \quad t > m \tag{7.4.10}$$

where $\gamma_0(w')/\sigma_a^2 = \text{var}[w_t']/\sigma_a^2 = 1 + \sum_{j=1}^{q} \theta_j^2$.

The ''innovations'' state-space approach to evaluating the exact likelihood function has also been shown to be quite useful in dealing with estimation problems for ARMA models when the series has missing values; see, for example, Jones (1980), Harvey and Pierse (1984), and Wincek and Reinsel (1986).

The exact likelihood function calculated using the Kalman filtering approach can be maximized using numerical optimization algorithms. These typically require the first partial derivatives of the log-likelihood with respect to the unknown parameters, and it is often beneficial to use analytical derivatives. From the form of the likelihood in (7.4.6), it is seen that this involves obtaining partial derivatives of the one-step predictions $\hat{w}_{t|t-1}$ and

of the error variances $\sigma_a^2 v_t$ for each $t = 1, \ldots, n$. Wincek and Reinsel (1986) show how the exact derivatives of $a_{t|t-1} = w_t - \hat{w}_{t|t-1}$ and $\sigma_a^2 v_t = \text{var}[a_{t|t-1}]$ with respect to the model parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ_a^2 can be obtained recursively through differentiation of the updating and prediction equations. This in turn leads to an explicit form of iterative calculations for the maximum likelihood estimation associated with the likelihood (7.4.6), similar to the nonlinear least-squares procedures detailed in Section 7.2.

7.5 ESTIMATION USING BAYES' THEOREM

7.5.1 Bayes' Theorem

In this section, we again use the symbol $\boldsymbol{\xi}$ to represent a general vector of parameters. Bayes' theorem tells us that if $p(\boldsymbol{\xi})$ is the probability distribution for $\boldsymbol{\xi}$ prior to the collection of the data, then $p(\boldsymbol{\xi}|\mathbf{z})$, the distribution of $\boldsymbol{\xi}$ posterior to the data $\mathbf{z}$, is obtained by combining the prior distribution $p(\boldsymbol{\xi})$ and the likelihood $L(\boldsymbol{\xi}|\mathbf{z})$ in the following way:

$$p(\boldsymbol{\xi}|\mathbf{z}) = \frac{p(\boldsymbol{\xi})L(\boldsymbol{\xi}|\mathbf{z})}{\int p(\boldsymbol{\xi})L(\boldsymbol{\xi}|\mathbf{z})d\boldsymbol{\xi}} \tag{7.5.1}$$

The denominator merely ensures that $p(\boldsymbol{\xi}|\mathbf{z})$ integrates to 1. The important part of the expression is the numerator, from which we see that the posterior distribution is proportional to the prior distribution multiplied by the likelihood. Savage (1962) showed that prior and posterior probabilities can be interpreted as subjective probabilities. In particular, often before the data are available, we have very little knowledge about $\boldsymbol{\xi}$, and we would be prepared to agree that over the relevant region, it would have appeared a priori just as likely that $\boldsymbol{\xi}$ had one value as another. In this case, $p(\boldsymbol{\xi})$ could be taken as *locally* uniform, and hence $p(\boldsymbol{\xi}|\mathbf{z})$ would be proportional to the likelihood.

It should be noted that for this argument to hold, it is not necessary for the prior density of $\boldsymbol{\xi}$ to be uniform over its entire range (which for some parameters could be infinite). By requiring that it be *locally uniform*, we mean that it be approximately uniform in the region in which the likelihood is appreciable and that it does not take an overwhelmingly large value outside that region.

Thus, if ξ were the weight of a chair, we could certainly say a priori that it weighed more than an ounce and less than a ton. It is also likely that when we obtained an observation z by weighing the chair on a weighing machine, which had an error standard deviation σ, we could honestly say that we would have been equally happy with a priori values in the range $z \pm 3\sigma$. The exception would be if the weighing machine said that an apparently heavy chair weighed, say, 10 ounces. In this case, the likelihood and the prior would be incompatible, and we should not, of course, use Bayes' theorem to combine them but would check the weighing machine and, if this turned out to be accurate, inspect the chair more closely.

There is, of course, some arbitrariness in this idea. Suppose that we assumed the prior distribution of ξ to be locally uniform. This then implies that the distribution of any linear function of ξ is also locally uniform. However, the prior distribution of some nonlinear transformation $\alpha = \alpha(\xi)$ (such as $\alpha = \log \xi$) could *not* be exactly locally uniform. This arbitrariness will usually have very little effect if we are able to obtain fairly precise

estimates of ξ. We will then be considering ξ only over a small range, and over such a range the transformation from ξ to, say, $\log \xi$ would often be very nearly linear.

Jeffreys (1961) has argued that it is best to choose the metric $\alpha(\xi)$ so that Fisher's measure of information $I_\alpha = -E[\partial^2 l/\partial\alpha^2]$ is independent of the value of α, and hence of ξ. This is equivalent to choosing $\alpha(\xi)$ so that the limiting variance of its maximum likelihood estimate is independent of ξ and is achieved by choosing the prior distribution of ξ to be proportional to $\sqrt{I_\xi}$.

Jeffreys justified this choice of prior on the basis of its invariance to the parameterization employed. Specifically, with this choice, the posterior distributions for $\alpha(\xi)$ and for ξ, where $\alpha(\xi)$ and ξ are connected by a one-to-one transformation, are such that $p(\xi|\mathbf{z}) = p(\alpha|\mathbf{z})\, d\alpha/d\xi$. The same result may be obtained (Box and Tiao, 1973) by the following argument. If for large samples, the expected likelihood function for $\alpha(\xi)$ approaches a normal curve, then the mean and variance of the curve summarize the information to be expected from the data. Suppose, now, that a transformation $\alpha(\xi)$ can be found in which the approximating normal curve has nearly constant variance whatever the true values of the parameter. Then, in this parameterization, the *only* information in prospect from the data is conveyed by the *location* of the expected likelihood function. To say that we know essentially nothing a priori relative to this prospective observational information is to say that we regard different *locations* of α as equally likely a priori. Equivalently, we say that α should be taken as locally uniform.

The generalization of Jeffreys' rule to deal with several parameters is that the joint prior distribution of parameters $\boldsymbol{\xi}$ be taken proportional to

$$|\mathbf{I}_\xi|^{1/2} = \left| -E\left[\frac{\partial^2 l}{\partial\xi_i \partial\xi_j}\right]\right|^{1/2} \tag{7.5.2}$$

It has been urged (e.g., Jenkins, 1964) that the likelihood itself is best considered and plotted in that metric α for which I_α is independent of α. If this is done, it will be noted that the likelihood function and the posterior density function with uniform prior are proportional.

7.5.2 Bayesian Estimation of Parameters

We now consider the estimation of the parameters in an ARIMA model from a Bayesian point of view. It is shown in Appendix A7.3 that the exact likelihood of a time series $\mathbf{z}$ of length $N = n + d$ from an ARIMA(p, d, q) process is of the form

$$L(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{z}) = \sigma_a^{-n} f(\boldsymbol{\phi}, \boldsymbol{\theta}) \exp\left[-\frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2}\right] \tag{7.5.3}$$

where

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} [a_t|\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\theta}]^2 + [\mathbf{e}_*]'\boldsymbol{\Omega}^{-1}[\mathbf{e}_*] \tag{7.5.4}$$

If we have no prior information about σ_a, $\boldsymbol{\phi}$, or $\boldsymbol{\theta}$, and since information about σ_a would supply no information about $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, it is sensible, following Jeffreys, to employ a prior distribution for $\boldsymbol{\phi}$, $\boldsymbol{\theta}$, and σ_a of the form

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a) \propto |\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta})|^{1/2} \sigma_a^{-1}$$

It follows that the posterior distribution is

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a|\mathbf{z}) \propto \sigma_a^{-(n+1)}|\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta})|^{1/2} f(\boldsymbol{\phi}, \boldsymbol{\theta}) \exp\left[-\frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2}\right] \tag{7.5.5}$$

If we now integrate (7.5.5) from zero to infinity with respect to σ_a, we obtain the exact joint posterior distribution of the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ as

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}|\mathbf{z}) \propto |\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta})|^{1/2} f(\boldsymbol{\phi}, \boldsymbol{\theta})\{S(\boldsymbol{\phi}, \boldsymbol{\theta})\}^{-n/2} \tag{7.5.6}$$

7.5.3 Autoregressive Processes

If z_t follows an ARIMA$(p, d, 0)$ process, then $w_t = \nabla^d z_t$ follows a pure AR(p) process. It is shown in Appendix A7.4 that for such a process, the factors $|\mathbf{I}(\boldsymbol{\phi})|^{1/2}$ and $f(\boldsymbol{\phi})$, which in any case are dominated by the term in $S(\boldsymbol{\phi})$, essentially cancel. This yields the remarkably simple result that given the assumptions, the parameters $\boldsymbol{\phi}$ of the AR(p) process in w_t have the posterior distribution

$$p(\boldsymbol{\phi}|\mathbf{z}) \propto \{S(\boldsymbol{\phi})\}^{-n/2} \tag{7.5.7}$$

By this argument, then, the sum-of-squares contours, which are approximate likelihood contours, are, when nothing is known a priori, also contours of posterior probability.

Joint Distribution of the Autoregressive Parameters. It is shown in Appendix A7.4 that for the pure AR process, the least-squares estimates of the ϕ's that minimize $S(\boldsymbol{\phi}) = \boldsymbol{\phi}_u' \mathbf{D} \boldsymbol{\phi}_u$ are given by

$$\hat{\boldsymbol{\phi}} = \mathbf{D}_p^{-1}\mathbf{d} \tag{7.5.8}$$

where $\boldsymbol{\phi}_u' = (1, \boldsymbol{\phi}')$,

$$\mathbf{d} = \begin{bmatrix} D_{12} \\ D_{13} \\ \vdots \\ D_{1,p+1} \end{bmatrix} \qquad \mathbf{D}_p = \begin{bmatrix} D_{22} & D_{23} & \cdots & D_{2,p+1} \\ D_{23} & D_{33} & \cdots & D_{3,p+1} \\ \vdots & \vdots & \vdots & \vdots \\ D_{2,p+1} & D_{3,p+1} & \cdots & D_{p+1,p+1} \end{bmatrix}$$

$$\mathbf{D} = \left[\begin{array}{c|c} D_{11} & -\mathbf{d}' \\ \hline -\mathbf{d} & \mathbf{D}_p \end{array}\right] \tag{7.5.9}$$

and

$$D_{ij} = D_{ji} = \tilde{w}_i\tilde{w}_j + \tilde{w}_{i+1}\tilde{w}_{j+1} + \cdots + \tilde{w}_{n+1-j}\tilde{w}_{n+1-i} \tag{7.5.10}$$

It follows that

$$S(\boldsymbol{\phi}) = \nu s_a^2 + (\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})' \mathbf{D}_p (\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}) \tag{7.5.11}$$

where

$$s_a^2 = \frac{S(\hat{\boldsymbol{\phi}})}{\nu} \qquad \nu = n - p \tag{7.5.12}$$

and

$$S(\hat{\boldsymbol{\phi}}) = \hat{\boldsymbol{\phi}}_u' \mathbf{D} \hat{\boldsymbol{\phi}}_u = D_{11} - \hat{\boldsymbol{\phi}}' \mathbf{D}_p \hat{\boldsymbol{\phi}} = D_{11} - \mathbf{d}' \mathbf{D}_p^{-1} \mathbf{d} \tag{7.5.13}$$

Thus, we can write

$$p(\boldsymbol{\phi}|\mathbf{z}) \propto \left[1 + \frac{(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})' \mathbf{D}_p (\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})}{\upsilon s_a^2} \right]^{-n/2} \tag{7.5.14}$$

Equivalently,

$$p(\boldsymbol{\phi}|\mathbf{z}) \propto \left[1 + \frac{\frac{1}{2} \sum_i \sum_j S_{ij} (\phi_i - \hat{\phi}_i)(\phi_j - \hat{\phi}_j)}{\upsilon s_a^2} \right]^{-n/2} \tag{7.5.15}$$

where

$$S_{ij} = \frac{\partial^2 S(\boldsymbol{\phi})}{\partial \phi_i \partial \phi_j} = 2 D_{i+1,j+1}$$

It follows that, a posteriori, the parameters of an autoregressive process have a multiple t distribution (A7.1.13), with $\upsilon = n - p$ degrees of freedom.

In particular, for the special case $p = 1$, $(\phi - \hat{\phi})/s_{\hat{\phi}}$ is distributed *exactly* in a Student t distribution with $n - 1$ degrees of freedom where, using the general results given above, $\hat{\phi}$ and $s_{\hat{\phi}}$ are given by

$$\hat{\phi} = \frac{D_{12}}{D_{22}} \qquad s_{\hat{\phi}} = \left[\frac{1}{n-1} \frac{D_{11}}{D_{22}} \left(1 - \frac{D_{12}^2}{D_{11} D_{22}} \right) \right]^{1/2} \tag{7.5.16}$$

The quantity $s_{\hat{\phi}}$, for large samples, tends to $[(1 - \phi^2)/n]^{1/2}$ and in the sampling theory framework is identical with the large-sample ''standard error'' for $\hat{\phi}$. However, when using this and similar expressions within the Bayesian framework, it is to be remembered that it is the parameters (ϕ in this case) that are random variables. Quantities such as $\hat{\phi}$ and $s_{\hat{\phi}}$, which are functions of data that have already occurred, are regarded as fixed.

Normal Approximation. For samples of size $n > 50$, in which we are usually interested, the normal approximation to the t distribution is adequate. Thus, very nearly, $\boldsymbol{\phi}$ has a joint p-variate normal distribution $N\{\hat{\boldsymbol{\phi}}, \mathbf{D}_p^{-1} s_a^2\}$ having mean vector $\hat{\boldsymbol{\phi}}$ and variance–covariance matrix $\mathbf{D}_p^{-1} s_a^2$.

Bayesian Regions of Highest Probability Density. In summarizing what the posterior distribution has to tell us about the probability of various $\boldsymbol{\phi}$ values, it is useful to indicate a region of *highest probability density*, called for short an HPD region (Box and Tiao, 1965). A Bayesian $1 - \varepsilon$ HPD region has the following properties:

1. Any parameter point inside the region has higher probability density than any point outside.
2. The total posterior probability mass within the region is $1 - \varepsilon$.

Since $\boldsymbol{\phi}$ has a multiple t distribution, it follows, using the result (A7.1.4), that

$$\Pr\{(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}')\mathbf{D}_p(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}) < p s_a^2 F_\varepsilon(p, \upsilon)\} = 1 - \varepsilon \tag{7.5.17}$$

defines the *exact* $1 - \varepsilon$ HPD region for $\boldsymbol{\phi}$. Now, for $\upsilon = n - p > 100$,

$$pF_\varepsilon(p, \upsilon) \simeq \chi_\varepsilon^2(p)$$

Also,

$$(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})'\mathbf{D}_p(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}) = \frac{1}{2}\sum_i \sum_j S_{ij}(\phi_i - \hat{\phi}_i)(\phi_j - \hat{\phi}_j)$$

Thus, approximately, the HPD region defined in (7.5.17) is such that

$$\sum_i \sum_j S_{ij}(\phi_i - \hat{\phi}_i)(\phi_j - \hat{\phi}_j) < 2 s_a^2 \chi_\varepsilon^2(p) \tag{7.5.18}$$

which if we set $\hat{\sigma}_a^2 = s_a^2$ is identical with the confidence region defined by (7.1.25).

Although these approximate regions are identical, it will be remembered that their interpretation is different. From a sampling theory viewpoint, we say that if a confidence region is computed according to (7.1.25), then for each of a set of repeated samples, a proportion $1 - \varepsilon$ of these regions will include the true parameter point. From the Bayesian viewpoint, we are concerned only with the single sample $\mathbf{z}$, which has actually been observed. Assuming the relevance of the noninformative prior distribution that we have taken, the HPD region includes that proportion $1 - \varepsilon$ of the resulting probability distribution of $\boldsymbol{\phi}$, given $\mathbf{z}$, which has the highest density. In other words, the probability that the value of $\boldsymbol{\phi}$, which gave rise to the data $\mathbf{z}$, lies in the HPD region is $1 - \varepsilon$.

Using (7.5.11), (7.5.12), and (7.5.18), for large samples the approximate $1 - \varepsilon$ Bayesian HPD region is bounded by a contour for which

$$S(\boldsymbol{\phi}) = S(\hat{\boldsymbol{\phi}})\left[1 + \frac{\chi_\varepsilon^2(p)}{n}\right] \tag{7.5.19}$$

which corresponds exactly with the confidence region defined by (7.1.27).

7.5.4 Moving Average Processes

If z_t follows an ARIMA$(0, d, q)$ process, then $w_t = \nabla^d z_t$ follows a pure MA(q) process. Because of the duality in estimation results and in the information matrices, in particular, between the autoregressive model and the moving average model, it follows that in the moving average case the factors $|\mathbf{I}(\boldsymbol{\theta})|^{1/2}$ and $f(\boldsymbol{\theta})$ in (7.5.6), which in any case are dominated by $S(\boldsymbol{\theta})$, also cancel for large samples. Thus, corresponding to (7.5.7), we find that the parameters $\boldsymbol{\theta}$ of the MA(q) process in w_t have the posterior distribution

$$p(\boldsymbol{\theta}|\mathbf{z}) \propto [S(\boldsymbol{\theta})]^{-n/2} \tag{7.5.20}$$

Again the sum-of-squares contours are, for moderate samples, essentially *exact* contours of posterior density. However, because $[a_t]$ is not a linear function of the θ's, $S(\boldsymbol{\theta})$ will not be exactly quadratic in $\boldsymbol{\theta}$, though for large samples it will often be nearly so within the relevant ranges. In that case, we have approximately

$$S(\boldsymbol{\theta}) = \nu s_a^2 + \frac{1}{2}\sum_i \sum_j S_{ij}(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)$$

where $\upsilon s_a^2 = S(\hat{\boldsymbol{\theta}})$ and $\upsilon = n - q$. It follows, after substituting for $S(\boldsymbol{\theta})$ in (7.5.20) and using the exponential approximation, that the following holds:

1. For large samples, $\boldsymbol{\theta}$ is *approximately* distributed in a multivariate normal distribution $N\{\hat{\boldsymbol{\theta}}, 2\{S_{ij}\}^{-1}s_a^2\}$.
2. An approximate HPD region is defined by (7.5.18) or (7.5.19), with q replacing p, and $\boldsymbol{\theta}$ replacing $\boldsymbol{\phi}$.

Example: Posterior Distribution of $\lambda = 1 - \theta$ for an IMA(0, 1, 1) Process. To illustrate, Figure 7.8 shows the approximate posterior density distribution $p(\lambda|\mathbf{z})$ from the data of Series B. It is seen to be approximately normal with its mode at $\hat{\lambda} = 1.09$ and having a standard deviation of about 0.05. A 95% Bayesian HPD interval covers essentially the same range, $0.98 < \lambda < 1.19$, as did the 95% confidence interval. Note that the density has been normalized to have unit area under the curve.

7.5.5 Mixed Processes

If z_t follows an ARIMA(p, d, q) process, then $w_t = \nabla^d z_t$ follows an ARMA (p, q) process

$$\phi(B)\tilde{w}_t = \theta(B)a_t$$

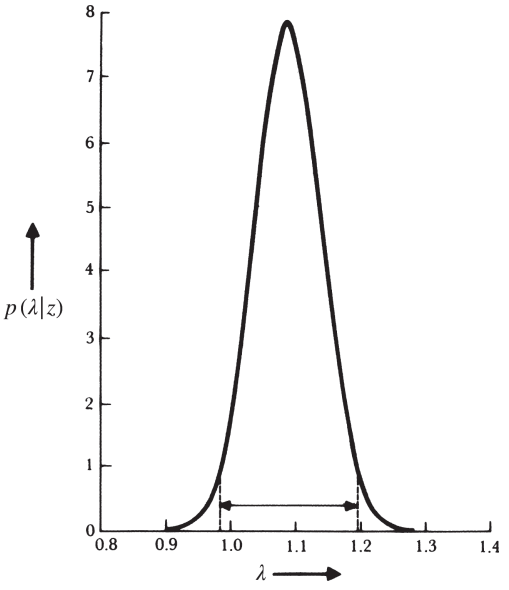


FIGURE 7.8 Posterior density $p(\lambda|\mathbf{z})$ for Series B.

It can be shown that for such a process the factors $|\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta})|^{1/2}$ and $f(\boldsymbol{\phi}, \boldsymbol{\theta})$ in (7.5.5) do not exactly cancel. Instead we can show, based on (7.2.24), that

$$|\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta})|^{1/2} f(\boldsymbol{\phi}, \boldsymbol{\theta}) = J(\boldsymbol{\phi}^* | \boldsymbol{\phi}, \boldsymbol{\theta}) \tag{7.5.21}$$

In (7.5.21), the ϕ^*'s are the $p + q$ parameters obtained by multiplying the autoregressive and moving average operators:

$$(1 - \phi_1^* B - \phi_2^* B^2 - \cdots - \phi_{p+q}^* B^{p+q}) = (1 - \phi_1 B - \cdots - \phi_p B^p) \times (1 - \theta_1 B - \cdots - \theta_q B^q)$$

and J is the Jacobian of the transformation from $\boldsymbol{\phi}^*$ to $(\boldsymbol{\phi}, \boldsymbol{\theta})$, that is,

$$p(\boldsymbol{\phi}, \boldsymbol{\theta} | \mathbf{z}) \propto J(\boldsymbol{\phi}^* | \boldsymbol{\phi}, \boldsymbol{\theta}) [S(\boldsymbol{\phi}, \boldsymbol{\theta})]^{-n/2} \tag{7.5.22}$$

In particular, for the ARMA(1, 1) process, $\phi_1^* = \phi + \theta, \phi_2^* = -\phi\theta, J = |\phi - \theta|$, and

$$p(\phi, \theta | \mathbf{z}) \propto |\phi - \theta| [S(\phi, \theta)]^{-n/2} \tag{7.5.23}$$

In this case, we see that the Jacobian will dominate in a region close to the line $\phi = \theta$ and will produce zero density on the line. This is sensible because the sum of squares $S(\phi, \theta)$ will take the finite value $\sum_{t=1}^{n} \tilde{w}_t^2$ for any $\phi = \theta$ and corresponds to our entertaining the possibility that $\tilde{w}_t$ is white noise. However, in our derivation, we have not constrained the range of the parameters. The possibility that $\phi = \theta$ is thus associated with unlimited ranges for the (equal) parameters. The effect of limiting the parameter space by, for example, introducing the requirements for stationarity and invertibility $(-1 < \phi < 1, -1 < \theta < 1)$ would be to produce a small positive value for the density, but this refinement seems scarcely worthwhile.

The Bayesian analysis reinforces the point made in Section 7.3.5 that estimation difficulties will be encountered with the mixed model and, in particular, with iterative solutions, when there is near redundancy in the parameters (i.e., near common factors between the AR and MA parts). We have already seen that the use of preliminary identification will usually ensure that these situations are avoided.

APPENDIX A7.1 REVIEW OF NORMAL DISTRIBUTION THEORY

A7.1.1 Partitioning of a Positive-Definite Quadratic Form

Consider the positive-definite quadratic form $Q_p = \mathbf{x}' \Sigma^{-1} \mathbf{x}$. Suppose that the $p \times 1$ vector $\mathbf{x}$ is partitioned after the p_1th element, so that $\mathbf{x}' = (\mathbf{x}_1' : \mathbf{x}_2') = (x_1, x_2, \ldots, x_{p_1} : x_{p_1+1}, \ldots, x_p)$, and suppose that the $p \times p$ matrix $\boldsymbol{\Sigma}$ is also partitioned after the p_1th row and column, so that

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} \end{array}\right]$$

It is readily verified that $\boldsymbol{\Sigma}^{-1}$ can be represented as

$$\boldsymbol{\Sigma}^{-1} = \left[\begin{array}{c|c} \mathbf{I} & -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right] \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \end{array}\right] \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline -\boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1} & \mathbf{I} \end{array}\right]$$

Then, since

$$\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = (\mathbf{x}_1' \,:\, \mathbf{x}_2' - \mathbf{x}_1'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}) \times \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \end{array}\right] \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1 \end{pmatrix}$$

$Q_p = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ can always be written as a sum of two quadratic forms Q_{p_1} and Q_{p_2}, containing p_1 and p_2 elements, respectively, where

$$\begin{aligned} Q_p &= Q_{p_1} + Q_{p_2} \\ Q_{p_1} &= \mathbf{x}_1'\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1 \\ Q_{p_2} &= (\mathbf{x}_2 - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1)'(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}(\mathbf{x}_2 - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1) \end{aligned} \tag{A7.1.1}$$

We may also write for the determinant of $\boldsymbol{\Sigma}$

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{11}||\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}| \tag{A7.1.2}$$

A7.1.2 Two Useful Integrals

Let $\mathbf{z}'\mathbf{C}\mathbf{z}$ be a positive-definite quadratic form in $\mathbf{z}$, which has q elements, so that $\mathbf{z}' = (z_1, z_2, \ldots, z_q)$, where $-\infty < z_i < \infty, i = 1, 2, \ldots, q$, and let a, b, and m be positive real numbers. Then, it may be shown that

$$\int_R \left(a + \frac{\mathbf{z}'\mathbf{C}\mathbf{z}}{b}\right)^{-(m+q)/2} d\mathbf{z} = \frac{(b\pi)^{q/2}\Gamma(m/2)}{a^{m/2}|\mathbf{C}|^{1/2}\Gamma[(m+q)/2]} \tag{A7.1.3}$$

where the q-fold integral extends over the entire $\mathbf{z}$ space R, and

$$\frac{\int_{\mathbf{z}'\mathbf{C}\mathbf{z} > qF_0} (1 + \mathbf{z}'\mathbf{C}\mathbf{z}/m)^{-(m+q)/2} d\mathbf{z}}{\int_R (1 + \mathbf{z}'\mathbf{C}\mathbf{z}/m)^{-(m+q)/2} d\mathbf{z}} = \int_{F_0}^{\infty} p(F|q, m)\, dF \tag{A7.1.4}$$

where the function $p(F|q, m)$ is the probability density of the F distribution with q and m degrees of freedom and is defined by

$$p(F|q, m) = \frac{(q/m)^{q/2}\Gamma[(m+q)/2]}{\Gamma(q/2)\Gamma(m/2)} F^{(q-2)/2} \left(1 + \frac{q}{m}F\right)^{-(m+q)/2} \qquad F > 0 \tag{A7.1.5}$$

If m tends to infinity, then

$$\left(1 + \frac{\mathbf{z}'\mathbf{C}\mathbf{z}}{m}\right)^{-(m+q)/2} \qquad \text{tends to} \qquad e^{-(\mathbf{z}'\mathbf{C}\mathbf{z})/2}$$

and writing $qF = \chi^2$, we obtain from (A7.1.4) that

$$\frac{\int_{\mathbf{z}'\mathbf{C}\mathbf{z}>\chi_0^2} e^{-(\mathbf{z}'\mathbf{C}\mathbf{z})/2} d\mathbf{z}}{\int_R e^{-(\mathbf{z}'\mathbf{C}\mathbf{z})/2} d\mathbf{z}} = \int_{\chi_0^2}^{\infty} p(\chi^2|q) d\chi^2 \tag{A7.1.6}$$

where the function $p(\chi^2|q)$ is the probability density of the χ^2 distribution with q degrees of freedom, and is defined by

$$p(\chi^2|q) = \frac{1}{2^{q/2}\Gamma(q/2)} (\chi^2)^{(q-2)/2} e^{-\chi^2/2} \qquad \chi^2 > 0 \tag{A7.1.7}$$

Here and elsewhere $p(x)$ is used as a *general* notation to denote a probability density function for a random variable x.

A7.1.3 Normal Distribution

The random variable x is normally distributed with mean μ and standard deviation σ, or $N(\mu, \sigma^2)$, if its probability density is

$$p(x) = (2\pi)^{-1/2} (\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2} \tag{A7.1.8}$$

Thus, the unit normal variate $u = (x - \mu)/\sigma$ has a distribution $N(0, 1)$. Table E in Part Five shows ordinates $p(u_\varepsilon)$ and values u_ε such that $\Pr\{u > u_\varepsilon\} = \varepsilon$ for chosen values of ε.

Multinormal Distribution. The vector $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ of random variables has a joint p-variate normal distribution $N\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ if its probability density function is

$$p(\mathbf{x}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} e^{-(\mathbf{x}-\mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)/2} \tag{A7.1.9}$$

The multinormal variate $\mathbf{x}$ has mean vector $\boldsymbol{\mu} = E[\mathbf{x}]$ and variance–covariance matrix $\boldsymbol{\Sigma} = \text{cov}[\mathbf{x}]$. The probability density contours are ellipsoids defined by $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \text{constant}$. For illustration, the elliptical contours for a bivariate ($p = 2$) normal distribution are shown in Figure A7.1.

At the point $\mathbf{x} = \boldsymbol{\mu}$, the multivariate normal distribution has its maximum density

$$\max p(\mathbf{x}) = p(\boldsymbol{\mu}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2}$$

The χ^2 Distribution as the Probability Mass Outside a Density Contour of the Multivariate Normal. For the p-variate normal distribution, (A7.1.9), the probability mass outside the density contour defined by

$$(\mathbf{x} - \mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu) = \chi_0^2$$

is given by the χ^2 integral with p degrees of freedom:

$$\int_{\chi_0^2}^{\infty} p(\chi^2|p) d\chi^2$$

where the χ^2 density function is defined as in (A7.1.7). Table F in Part Five shows values of $\chi_\varepsilon^2(p)$, such that $\Pr\{\chi^2 > \chi_\varepsilon^2(p)\} = \varepsilon$ for chosen values of ε.

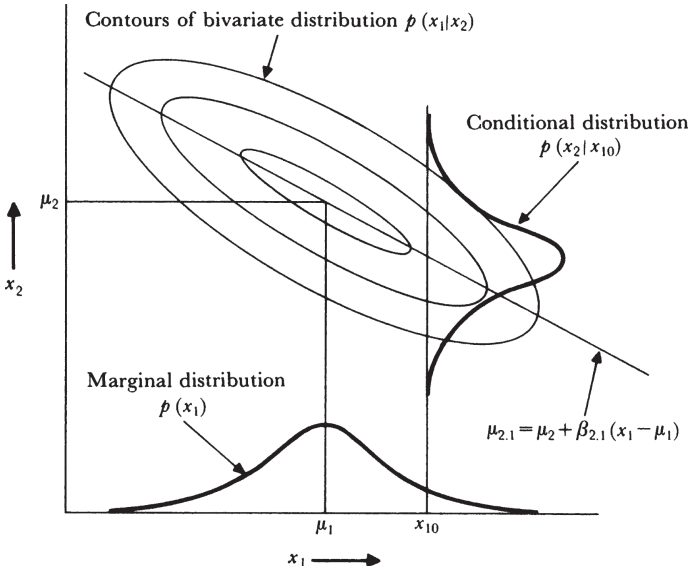


FIGURE A7.1 Contours of a bivariate normal distribution showing the marginal distribution $p(x_1)$ and the conditional distribution $p(x_2|x_{10})$ at $x_1 = x_{10}$.

Marginal and Conditional Distributions for the Multivariate Normal Distribution. Suppose that the vector of $p = p_1 + p_2$ random variables is partitioned after the first p_1 elements, so that

$$\mathbf{x}' = (\mathbf{x}_1' : \mathbf{x}_2') = (x_1, x_2, \ldots, x_{p_1} : x_{p_1+1}, \ldots, x_{p_1+p_2})$$

and that the variance–covariance matrix is

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} \end{array}\right]$$

Then using (A7.1.1) and (A7.1.2), we can write the multivariate normal distribution for the $p = p_1 + p_2$ variates as the *marginal* distribution of $\mathbf{x}_1$ multiplied by the *conditional* distribution of $\mathbf{x}_2$ given $\mathbf{x}_1$, that is,

$$\begin{aligned} p(\mathbf{x}) &= p(\mathbf{x}_1, \mathbf{x}_2) = p(\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1) \\ &= (2\pi)^{-p_1/2}|\boldsymbol{\Sigma}_{11}|^{-1/2} \exp\left[-\frac{(\mathbf{x}_1 - \mu_1)'\boldsymbol{\Sigma}^{-1}{}_{11}(\mathbf{x}_1 - \mu_1)}{2}\right] \\ &\quad \times (2\pi)^{-p_2/2}|\boldsymbol{\Sigma}_{22.11}|^{-1/2} \exp\left[-\frac{(\mathbf{x}_2 - \mu_{2.1})'\boldsymbol{\Sigma}_{22.11}^{-1}(\mathbf{x}_2 - \mu_{2.1})}{2}\right] \end{aligned} \tag{A7.1.10}$$

where

$$\boldsymbol{\Sigma}_{22.11} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}'\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \tag{A7.1.11}$$

and $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \boldsymbol{\beta}_{2.1}(\mathbf{x}_1 - \boldsymbol{\mu}_1) = E[\mathbf{x}_2|\mathbf{x}_1]$ define regression hyperplanes in $(p_1 + p_2)$-dimensional space, tracing the loci of the (conditional) means of the p_2 elements of $\mathbf{x}_2$ as the p_1 elements of $\mathbf{x}_1$ vary. The $p_2 \times p_1$ matrix of regression coefficients is given by $\boldsymbol{\beta}_{2.1} = \boldsymbol{\Sigma}'_{12}\boldsymbol{\Sigma}_{11}^{-1}$.

Both marginal and conditional distributions for the multivariate normal are therefore multivariate normal distributions. It is seen that for the *multivariate normal distribution*, the conditional distribution $p(\mathbf{x}_2|\mathbf{x}_1)$ is, *except for location* (i.e., mean value), identical whatever the value of $\mathbf{x}_1$ (i.e., multivariate normal with identical variance–covariance matrix $\boldsymbol{\Sigma}_{22.11}$).

Univariate Marginals. In particular, the marginal density for a single element x_i $(i = 1, 2, \ldots, p)$ is $N(\mu_i, \sigma_i^2)$, a univariate normal with mean μ_i equal to the ith element of $\boldsymbol{\mu}$ and variance σ_i^2 equal to the ith diagonal element of $\boldsymbol{\Sigma}$.

Bivariate Normal. For illustration, the marginal and conditional distributions for a bivariate normal are shown in Figure A7.1. In this case, the marginal distribution of x_1 is $N(\mu_1, \sigma_1^2)$, while the conditional distribution of x_2 given x_1 is

$$N\left\{\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1 - \rho^2)\right\}$$

where $\rho = (\sigma_1/\sigma_2)\beta_{2.1}$ is the correlation coefficient between x_1 and x_2 and $\beta_{2.1} = \sigma_{12}/\sigma_1^2$ is the regression coefficient of x_2 on x_1.

A7.1.4 Student's *t* Distribution

The random variable x is distributed as a scaled t distribution with mean μ and scale parameter s and with ν degrees of freedom, denoted as $t(\mu, s^2, \nu)$, if its probability density is

$$p(x) = (2\pi)^{-1/2}(s^2)^{-1/2}\left(\frac{\nu}{2}\right)^{-1/2}\Gamma\left(\frac{\nu+1}{2}\right)\Gamma^{-1}\left(\frac{\nu}{2}\right)\left[1 + \frac{(x-\mu)^2}{\nu s^2}\right]^{-(\nu+1)/2} \quad \text{(A7.1.12)}$$

Thus, the standardized t variate $t = (x - \mu)/s$ has distribution $t(0, 1, \nu)$. Table G in Part Five shows values t_ε such that $\Pr\{t > t_\varepsilon\} = \varepsilon$ for chosen values of ε.

Approach to Normal Distribution. For large ν, the product

$$\left(\frac{\nu}{2}\right)^{-1/2}\Gamma\left(\frac{\nu+1}{2}\right)\Gamma^{-1}\left(\frac{\nu}{2}\right)$$

tends to unity, while the right-hand bracket in (A7.1.12) tends to $e^{-(1/2s^2)(x-\mu)^2}$. Thus, if for large ν we write $s^2 = \sigma^2$, the t distribution tends to the normal distribution (A7.1.8).

Multiple t Distribution. Let $\boldsymbol{\mu}' = (\mu_1, \mu_2, \ldots, \mu_p)$ be a $p \times 1$ vector and $\mathbf{S}$ a $p \times p$ positive-definite matrix. Then, the vector random variable $\mathbf{x}$ has a scaled multivariate t distribution $t(\boldsymbol{\mu}, \mathbf{S}, \nu)$, with mean vector $\boldsymbol{\mu}$, scaling matrix $\mathbf{S}$, and ν degrees of freedom if its probability

density is

$$p(\mathbf{x}) = (2\pi)^{-p/2}|\mathbf{S}|^{-1/2}\left(\frac{\nu}{2}\right)^{-p/2}\Gamma\left(\frac{\nu+p}{2}\right)\Gamma^{-1}\left(\frac{\nu}{2}\right) \times \left[1 + \frac{(\mathbf{x}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right]^{-(\nu+p)/2} \tag{A7.1.13}$$

The probability contours of the multiple t distribution are ellipsoids defined by $(\mathbf{x} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \text{constant}$.

Approach to the Multinormal Form. For large ν, the product

$$\left(\frac{\nu}{2}\right)^{-p/2}\Gamma\left(\frac{\nu+p}{2}\right)\Gamma^{-1}\left(\frac{\nu}{2}\right)$$

tends to unity; also, the right-hand bracket in (A7.1.13) tends to $e^{-(\mathbf{x}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}$. Thus, if for large υ we write $\mathbf{S} = \boldsymbol{\Sigma}$, the multiple t tends to the multivariate normal distribution (A7.1.9).

APPENDIX A7.2 REVIEW OF LINEAR LEAST-SQUARES THEORY

A7.2.1 Normal Equations and Least Squares

The linear regression model is assumed to be

$$w_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i \tag{A7.2.1}$$

where the $w_i\,(i = 1, 2, \ldots, n)$ are observations on a response or dependent variable obtained from an experiment in which the independent variables $x_{i1}, x_{i2}, \ldots, x_{ik}$ take on known *fixed* values, the β_i are unknown parameters to be estimated from the data, and the e_i are uncorrelated random errors having zero means and the same common variance σ^2.

The relations (A7.2.1) may be expressed in matrix form as

$$\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

or

$$\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{A7.2.2}$$

where the $n \times k$ matrix $\mathbf{X}$ is assumed to be of full rank k. Gauss's theorem of least-squares may be stated (Barnard, 1963) in the following form: The estimates $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ of the parameters $\boldsymbol{\beta}$, which are linear in the observations and unbiased for $\boldsymbol{\beta}$ and which minimize the mean square error among all such estimates of any linear function $\lambda_1\beta_1 + \lambda_2\beta_2 + \cdots + \lambda_k\beta_k$ of the parameters, are obtained by minimizing the sum of squares

$$S(\boldsymbol{\beta}) = \mathbf{e}'\mathbf{e} = (\mathbf{w} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{w} - \mathbf{X}\boldsymbol{\beta}) \tag{A7.2.3}$$

To establish the minimum of $S(\boldsymbol{\beta})$, we note that the vector $\mathbf{w} - \mathbf{X}\boldsymbol{\beta}$ may be decomposed into two vectors $\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ according to

$$\mathbf{w} - \mathbf{X}\boldsymbol{\beta} = \mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{A7.2.4}$$

Hence, provided that we choose $\hat{\boldsymbol{\beta}}$ so that $\mathbf{X}'(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$, that is,

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{w} \tag{A7.2.5}$$

it follows that

$$S(\boldsymbol{\beta}) = S(\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{A7.2.6}$$

and the vectors $\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ are orthogonal. Since the second term on the right-hand side of (A7.2.6) is a positive-definite quadratic form, it follows that the minimum is attained when $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}$$

is the *least-squares* estimate of $\boldsymbol{\beta}$ given by the solution to the *normal equation* (A7.2.5).

A7.2.2 Estimation of Error Variance

Using (A7.2.3) and (A7.2.5), the sum of squares at the minimum is

$$S(\hat{\boldsymbol{\beta}}) = (\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{w}'\mathbf{w} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \tag{A7.2.7}$$

Furthermore, if we define

$$s^2 = \frac{S(\hat{\boldsymbol{\beta}})}{n - k} \tag{A7.2.8}$$

it may be shown that $E[s^2] = \sigma^2$, and hence s^2 provides an unbiased estimate of the error variance σ^2.

A7.2.3 Covariance Matrix of Least-Squares Estimates

The covariance matrix of the least-squares estimates $\hat{\boldsymbol{\beta}}$ is defined by

$$\begin{aligned}\mathbf{V}(\hat{\boldsymbol{\beta}}) &= \text{cov}[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}'] \\ &= \text{cov}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}, \mathbf{w}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}[\mathbf{w}, \mathbf{w}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\sigma^2\end{aligned} \tag{A7.2.9}$$

since $\text{cov}[\mathbf{w}, \mathbf{w}'] = \mathbf{I}\sigma^2$.

A7.2.4 Confidence Regions

Assuming normality, the quadratic forms $S(\hat{\boldsymbol{\beta}})$ and $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ in (A7.2.6) are independently distributed as σ^2 times chi-squared random variables with $n - k$ and k

degrees of freedom, respectively. Hence,

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{S(\hat{\boldsymbol{\beta}})}\frac{n-k}{k}$$

is distributed as $F(k, n-k)$. Using (A7.2.8), it follows that

$$(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \leq ks^2F_\varepsilon(k, n-k) \tag{A7.2.10}$$

defines a $1-\varepsilon$ confidence region for $\boldsymbol{\beta}$.

A7.2.5 Correlated Errors

Suppose that the errors $\mathbf{e}$ in (A7.2.2) have a *known* covariance matrix $\mathbf{V}$, and let $\mathbf{P}$ be an $n \times n$ nonsingular matrix such that $\mathbf{V}^{-1} = \mathbf{PP}'/\sigma^2$, so that $\mathbf{P}'\mathbf{VP} = \mathbf{I}\sigma^2$. Then, (A7.2.2) may be transformed into

$$\mathbf{P}'\mathbf{w} = \mathbf{P}'\mathbf{X}\boldsymbol{\beta} + \mathbf{P}'\mathbf{e}$$

or

$$\mathbf{w}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{e}^* \tag{A7.2.11}$$

where $\mathbf{w}^* = \mathbf{P}'\mathbf{w}$ and $\mathbf{X}^* = \mathbf{P}'\mathbf{X}$. The covariance matrix of $\mathbf{e}^* = \mathbf{P}'\mathbf{e}$ in (A7.2.11) is

$$\text{cov}[\mathbf{P}'\mathbf{e}, \mathbf{e}'\mathbf{P}] = \mathbf{P}'\text{cov}[\mathbf{e}, \mathbf{e}']\mathbf{P} = \mathbf{P}'\mathbf{VP} = \mathbf{I}\sigma^2$$

Hence, we may apply ordinary least-squares theory with $\mathbf{V} = \mathbf{I}\sigma^2$ to the *transformed* model (A7.2.11), in which $\mathbf{w}$ is replaced by $\mathbf{w}^* = \mathbf{P}'\mathbf{w}$ and $\mathbf{X}$ by $\mathbf{X}^* = \mathbf{P}'\mathbf{X}$. Thus, we obtain the estimates

$$\hat{\boldsymbol{\beta}}_G = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{w}^*$$

with $\mathbf{V}(\hat{\boldsymbol{\beta}}_G) = \text{cov}[\hat{\boldsymbol{\beta}}_\mathbf{G}] = \sigma^2(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}$. In terms of the *original* variables $\mathbf{X}$ and $\mathbf{w}$ of the regression model, since $\mathbf{PP}' = \sigma^2\mathbf{V}^{-1}$, the estimate is

$$\hat{\boldsymbol{\beta}}_G = (\mathbf{X}'\mathbf{PP}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{PP}'\mathbf{w} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{w} \tag{A7.2.12}$$

with

$$\mathbf{V}(\hat{\boldsymbol{\beta}}_G) = \text{cov}[\hat{\boldsymbol{\beta}}_G] = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

The estimator $\hat{\boldsymbol{\beta}}_G$ in (A7.2.12) is generally referred to as the *generalized least-squares* (GLS) estimator, and it follows that this is the estimate of $\boldsymbol{\beta}$ obtained by minimizing the *generalized* sum of squares function

$$S(\boldsymbol{\beta}|\mathbf{V}) = (\mathbf{w}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{w}-\mathbf{X}\boldsymbol{\beta})$$

APPENDIX A7.3 EXACT LIKELIHOOD FUNCTION FOR MOVING AVERAGE AND MIXED PROCESSES

To obtain the required likelihood function for an MA(q) model, we have to derive the probability density function for a series $\mathbf{w}' = (w_1, w_2, \ldots, w_n)$ assumed to be generated by an invertible moving average model of order q:

$$\tilde{w}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \tag{A7.3.1}$$

where $\tilde{w}_t = w_t - \mu$, with $\mu = E[w_t]$. Under the assumption that the a_t's and the $\tilde{w}_t$'s are normally distributed, the joint density may be written as

$$p(\mathbf{w}|\boldsymbol{\theta}, \sigma_a^2, \mu) = (2\pi\sigma_a^2)^{-n/2} |\mathbf{M}_n^{(0,q)}|^{1/2} \exp\left[\frac{-\tilde{\mathbf{w}}'\mathbf{M}_n^{(0,q)}\tilde{\mathbf{w}}}{2\sigma_a^2}\right] \tag{A7.3.2}$$

where $(\mathbf{M}_n^{(p,q)})^{-1}\sigma_a^2$ denotes the $n \times n$ covariance matrix of the w_t's for an ARMA(p, q) process. We now consider a convenient way of evaluating $\tilde{\mathbf{w}}'\mathbf{M}_n^{(0,q)}\tilde{\mathbf{w}}$, and for simplicity, we suppose that $\mu = 0$, so that $w_t = \tilde{w}_t$.

Using the model (A7.3.1), we can write down the n equations:

$$w_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \qquad (t = 1, 2, \ldots, n)$$

These n equations can be conveniently expressed in matrix form in terms of the n-dimensional vectors $\mathbf{w}' = (w_1, w_2, \ldots, w_n)$ and $\mathbf{a}' = (a_1, a_2, \ldots, a_n)$, and the q-dimensional vector of preliminary values $\mathbf{a}_*' = (a_{1-q}, a_{2-q}, \ldots, a_0)$ as

$$\mathbf{w} = \mathbf{L}_\theta \mathbf{a} + \mathbf{F}\mathbf{a}_*$$

where $\mathbf{L}_\theta$ is an $n \times n$ lower triangular matrix with l's on the leading diagonal, $-\theta_1$ on the first subdiagonal, $-\theta_2$ on the second subdiagonal, and so on, with $\theta_i = 0$ for $i > q$. Further, $\mathbf{F}$ is an $n \times q$ matrix with the form $\mathbf{F} = (\mathbf{B}_q', \mathbf{0}')'$ where $\mathbf{B}_q$ is $q \times q$ equal to

$$\mathbf{B}_q = -\begin{bmatrix} \theta_q & \theta_{q-1} & \ldots & \theta_1 \\ 0 & \theta_q & \ldots & \theta_2 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \theta_q \end{bmatrix}$$

Now the joint distribution of the $n + q$ values, which are the elements of $(\mathbf{a}', \mathbf{a}_*')$, is

$$p(\mathbf{a}, \mathbf{a}_*|\sigma_a^2) = (2\pi\sigma_a^2)^{-(n+q)/2} \exp\left[-\frac{1}{2\sigma_a^2}(\mathbf{a}'\mathbf{a} + \mathbf{a}_*'\mathbf{a}_*)\right]$$

Noting that the transformation from $(\mathbf{a}, \mathbf{a}_*)$ to $(\mathbf{w}, \mathbf{a}_*)$ has unit Jacobian and $\mathbf{a} = \mathbf{L}_\theta^{-1}(\mathbf{w} - \mathbf{F}\mathbf{a}_*)$, the joint distribution of $\mathbf{w} = \mathbf{L}_\theta\mathbf{a} + \mathbf{F}\mathbf{a}_*$ and $\mathbf{a}_*$ is

$$p(\mathbf{w}, \mathbf{a}_*|\boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-(n+q)/2} \exp\left[-\frac{1}{2\sigma_a^2} S(\boldsymbol{\theta}, \mathbf{a}_*)\right]$$

where

$$S(\boldsymbol{\theta}, \mathbf{a}_*) = (\mathbf{w} - \mathbf{F}\mathbf{a}_*)' \mathbf{L}_\theta'^{-1} \mathbf{L}_\theta^{-1} (\mathbf{w} - \mathbf{F}\mathbf{a}_*) + \mathbf{a}_*' \mathbf{a}_* \tag{A7.3.3}$$

Now, let $\hat{\mathbf{a}}_*$ be the vector of values that minimize $S(\boldsymbol{\theta}, \mathbf{a}_*)$, which from generalized least-squares theory can be shown to equal $\hat{\mathbf{a}}_* = \mathbf{D}^{-1}\mathbf{F}'\mathbf{L}_\theta'^{-1}\mathbf{L}_\theta^{-1}\mathbf{w}$, where $\mathbf{D} = \mathbf{I}_q + \mathbf{F}'\mathbf{L}_\theta'^{-1}\mathbf{L}_\theta^{-1}\mathbf{F}$. Then, using the result (A7.2.6), we have

$$S(\boldsymbol{\theta}, \mathbf{a}_*) = S(\boldsymbol{\theta}) + (\mathbf{a}_* - \hat{\mathbf{a}}_*)' \mathbf{D} (\mathbf{a}_* - \hat{\mathbf{a}}_*)$$

where

$$S(\boldsymbol{\theta}) = S(\boldsymbol{\theta}, \hat{\mathbf{a}}_*) = (\mathbf{w} - \mathbf{F}\hat{\mathbf{a}}_*)' \mathbf{L}_\theta'^{-1} \mathbf{L}_\theta^{-1} (\mathbf{w} - \mathbf{F}\hat{\mathbf{a}}_*) + \hat{\mathbf{a}}_*' \hat{\mathbf{a}}_* \tag{A7.3.4}$$

is a function of the observations $\mathbf{w}$ but not of the preliminary values $\mathbf{a}_*$. Thus,

$$p(\mathbf{w}, \mathbf{a}_* | \boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-(n+q)/2} \exp \left\{ -\frac{1}{2\sigma_a^2} [S(\boldsymbol{\theta}) + (\mathbf{a}_* - \hat{\mathbf{a}}_*)' \mathbf{D} (\mathbf{a}_* - \hat{\mathbf{a}}_*)] \right\}$$

However, since the joint distribution of $\mathbf{w}$ and $\mathbf{a}_*$ can be factored as

$$p(\mathbf{w}, \mathbf{a}_* | \boldsymbol{\theta}, \sigma_a^2) = p(\mathbf{w} | \boldsymbol{\theta}, \sigma_a^2) p(\mathbf{a}_* | \mathbf{w}, \boldsymbol{\theta}, \sigma_a^2)$$

it follows, similar to (A7.1.10), that

$$p(\mathbf{a}_* | \mathbf{w}, \boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-q/2} \mid \mathbf{D} \mid^{1/2} \exp \left[-\frac{1}{2\sigma_a^2} (\mathbf{a}_* - \hat{\mathbf{a}}_*)' \mathbf{D} (\mathbf{a}_* - \hat{\mathbf{a}}_*) \right] \tag{A7.3.5}$$

$$p(\mathbf{w} | \boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-n/2} \mid \mathbf{D} \mid^{-1/2} \exp \left[-\frac{1}{2\sigma_a^2} S(\boldsymbol{\theta}) \right] \tag{A7.3.6}$$

We can now deduce the following:

1. From (A7.3.5), we see that $\hat{\mathbf{a}}_*$ is the conditional expectation of $\mathbf{a}_*$ given $\mathbf{w}$ and $\boldsymbol{\theta}$. Thus, using the notation introduced in Section 7.1.4, we obtain

$$\hat{\mathbf{a}}_* = [\mathbf{a}_* | \mathbf{w}, \boldsymbol{\theta}] = [\mathbf{a}_*]$$

where $[\mathbf{a}] = \mathbf{L}_\theta^{-1}(\mathbf{w} - \mathbf{F}[\mathbf{a}_*])$ is the conditional expectation of $\mathbf{a}$ given $\mathbf{w}$ and $\boldsymbol{\theta}$, and using (A7.3.4):

$$S(\boldsymbol{\theta}) = [\mathbf{a}]'[\mathbf{a}] + [\mathbf{a}_*]'[\mathbf{a}_*] = \sum_{t=1-q}^{n} [a_t]^2 \tag{A7.3.7}$$

To compute $S(\boldsymbol{\theta})$, the quantities $[a_t] = [a_t | \mathbf{w}, \boldsymbol{\theta}]$ may be obtained by using the estimates $[\mathbf{a}_*]' = ([a_{1-q}], [a_{2-q}], \ldots, [a_0])$ obtained as above by back-forecasting for preliminary values, and computing the elements $[a_1], [a_2], \ldots, [a_n]$ of $[\mathbf{a}]$ recursively from the relation $\mathbf{L}_\theta[\mathbf{a}] = \mathbf{w} - \mathbf{F}[\mathbf{a}_*]$ as

$$[a_t] = w_t + \theta_1 [a_{t-1}] + \ldots + \theta_q [a_{t-q}] \qquad (t = 1, 2, \ldots, n)$$

Note that if the expression for $\hat{\mathbf{a}}_*$ is utilized in (A7.3.4), after rearranging we obtain

$$S(\boldsymbol{\theta}) = \mathbf{w}'\mathbf{L}_\theta'^{-1}(\mathbf{I}_n - \mathbf{L}_\theta^{-1}\mathbf{F}\mathbf{D}^{-1}\mathbf{F}'\mathbf{L}_\theta'^{-1})\mathbf{L}_\theta^{-1}\mathbf{w} = \mathbf{a}^{0\prime}\mathbf{a}^0 - \hat{\mathbf{a}}_*'\mathbf{D}\hat{\mathbf{a}}_*$$

where $\mathbf{a}^0 = \mathbf{L}_\theta^{-1}\mathbf{w}$ denotes the vector whose elements a_t^0 can be calculated recursively from $a_t^0 = w_t + \theta_1 a_{t-1}^0 + \cdots + \theta_q a_{t-q}^0$, $t = 1, 2, \ldots, n$, by setting the initial values $\mathbf{a}_*$ equal to zero. Hence, the first term described above, $S_*(\boldsymbol{\theta}) = \mathbf{a}^{0\prime}\mathbf{a}^0 = \sum_{t=1}^n (a_t^0)^2$, is the conditional sum-of-squares function, given $\mathbf{a}_* = \mathbf{0}$, as discussed in Section 7.1.2.

2. In addition, we find that

$$\mathbf{M}_n^{(0,q)} = \mathbf{L}_\theta'^{-1}(\mathbf{I}_n - \mathbf{L}_\theta^{-1}\mathbf{F}\mathbf{D}^{-1}\mathbf{F}'\mathbf{L}_\theta'^{-1})\mathbf{L}_\theta^{-1}$$

and $S(\boldsymbol{\theta}) = \mathbf{w}'\mathbf{M}_n^{(0,q)}\mathbf{w}$. Also, by comparing (A7.3.6) and (A7.3.2), we have

$$|\mathbf{D}|^{-1} = |\mathbf{M}_n^{(0,q)}|$$

3. The back-forecasts $\hat{\mathbf{a}}_* = [\mathbf{a}_*]$ can be calculated most conveniently from $\hat{\mathbf{a}}_* = \mathbf{D}^{-1}\mathbf{F}'\mathbf{u}$ (i.e., by solving $\mathbf{D}\hat{\mathbf{a}}_* = \mathbf{F}'\mathbf{u}$), where $\mathbf{u} = \mathbf{L}_\theta'^{-1}\mathbf{L}_\theta^{-1}\mathbf{w} = \mathbf{L}_\theta'^{-1}\mathbf{a}^0 = (u_1, u_2, \ldots, u_n)'$. Note that the elements u_t of $\mathbf{u}$ are calculated through a backward recursion as

$$u_t = a_t^0 + \theta_1 u_{t+1} + \ldots + \theta_q u_{t+q}$$

from $t = n$ down to $t = 1$, using zero starting values $u_{n+1} = \ldots = u_{n+q} = 0$, where the a_t^0 denote the estimates of the a_t conditional on the zero starting values $\mathbf{a}_* = \mathbf{0}$. Also, the vector $\mathbf{h} = \mathbf{F}'\mathbf{u}$ consists of the elements $h_j = -\sum_{i=1}^j \theta_{q-j+i}u_i$, $j = 1, \ldots, q$.

4. Finally, using (A7.3.6) and (A7.3.7), the unconditional likelihood is given exactly by

$$L(\boldsymbol{\theta}, \sigma_a^2|\mathbf{w}) = (\sigma_a^2)^{-n/2}|\mathbf{D}|^{-1/2}\exp\left\{-\frac{1}{2\sigma_a^2}\sum_{t=1-q}^{n}[a_t]^2\right\} \tag{A7.3.8}$$

For example, in the MA(1) model with $q = 1$, we have $\mathbf{F}' = -(\theta, 0, \ldots, 0)$, an n-dimensional vector, and $\mathbf{L}_\theta$ is such that $\mathbf{L}_\theta^{-1}$ has first column equal to $(1, \theta, \theta^2, \ldots, \theta^{n-1})'$, so that

$$\mathbf{D} = 1 + \mathbf{F}'\mathbf{L}_\theta'^{-1}\mathbf{L}_\theta^{-1}\mathbf{F} = 1 + \theta^2 + \theta^4 + \ldots + \theta^{2n} = \frac{1 - \theta^{2(n+1)}}{1 - \theta^2}$$

In addition, the conditional values a_t^0 are computed recursively as $a_t^0 = w_t + \theta a_{t-1}^0$, $t = 1, 2, \ldots, n$, using the zero initial value $a_0^0 = 0$, and the values of the vector $\mathbf{u} = \mathbf{L}_\theta'^{-1}\mathbf{a}^0$ are computed in the backward recursion as $u_t = a_t^0 + \theta u_{t+1}$, from $t = n$ to $t = 1$, with $u_{n+1} = 0$. Then,

$$\hat{\mathbf{a}}_* = [a_0] = -\mathbf{D}^{-1}\theta u_1 = -\mathbf{D}^{-1}u_0 = -\frac{u_0(1 - \theta^2)}{1 - \theta^{2(n+1)}}$$

where $u_0 = a_0^0 + \theta u_1 = \theta u_1$, and the exact likelihood for the MA(1) process is

$$L(\theta, \sigma_a^2|\mathbf{w}) = (\sigma_a^2)^{-n/2} \frac{(1-\theta^2)^{1/2}}{(1-\theta^{2(n+1)})^{1/2}} \exp\left\{ -\frac{1}{2\sigma_a^2} \sum_{t=0}^{n} [a_t]^2 \right\} \tag{A7.3.9}$$

Extension to the Autoregressive and Mixed Processes. The method outlined above may be readily extended to provide the unconditional likelihood for the general mixed model

$$\phi(B)\tilde{w}_t = \theta(B)a_t \tag{A7.3.10}$$

which, with $w_t = \nabla^d z_t$, defines the general ARIMA process. Details of the derivation have been presented by Newbold (1974) and Ljung and Box (1979), while an alternative approach to obtain the exact likelihood that uses the Cholesky decomposition of a band covariance matrix (i.e., the innovations method as discussed in Section 7.4) was given by Ansley (1979). First, assuming a zero mean for the process, the relations for the ARMA model may be written in matrix form, similar to before, as

$$\mathbf{L}_\phi \mathbf{w} = \mathbf{L}_\theta \mathbf{a} + \mathbf{F}\mathbf{e}_*$$

where $\mathbf{L}_\phi$ is an $n \times n$ matrix of the same form as $\mathbf{L}_\theta$ but with ϕ_i's in place of θ_i's, $\mathbf{e}_*' = (\mathbf{w}_*', \mathbf{a}_*') = (w_{1-p}, \ldots, w_0, a_{1-q}, \ldots, a_0)$ is the $(p+q)$-dimensional vector of initial values, and

$$\mathbf{F} = \begin{bmatrix} \mathbf{A}_p & \mathbf{B}_q \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

with

$$\mathbf{A}_p = \begin{bmatrix} \phi_p & \phi_{p-1} & \cdots & \phi_1 \\ 0 & \phi_p & \cdots & \phi_2 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \phi_p \end{bmatrix} \quad \text{and} \quad \mathbf{B}_q = -\begin{bmatrix} \theta_q & \theta_{q-1} & \cdots & \theta_1 \\ 0 & \theta_q & \cdots & \theta_2 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \theta_q \end{bmatrix}$$

Let $\mathbf{\Omega}\sigma_a^2 = E[\mathbf{e}_*\mathbf{e}_*']$ denote the covariance matrix of $\mathbf{e}_*$. This matrix has the form

$$\mathbf{\Omega}\sigma_a^2 = \begin{bmatrix} \sigma_a^{-2}\mathbf{\Gamma}_p & \mathbf{C}' \\ \mathbf{C} & \mathbf{I}_q \end{bmatrix} \sigma_a^2$$

where $\mathbf{\Gamma}_p = E[\mathbf{w}_*\mathbf{w}_*']$ is a $p \times p$ matrix with (i,j)th element γ_{i-j}, and $\sigma_a^2\mathbf{C} = E[\mathbf{a}_*\mathbf{w}_*']$ has elements defined by $E[a_{i-q}w_{j-p}] = \sigma_a^2\psi_{j-i-p+q}$ for $j - i - p + q \geq 0$ and 0 otherwise. The ψ_k are the coefficients in the infinite MA operator $\psi(B) = \phi^{-1}(B)\theta(B) = \sum_{k=0}^{\infty} \psi_k B^k, \psi_0 = 1$, and are easily determined recursively through equations in Section 3.4. The autocovariances γ_k in $\mathbf{\Gamma}_p$ can directly be determined in terms of the coefficients ϕ_i, θ_i, and σ_a^2, through use of the first $(p+1)$ equations (3.4.2) (see, e.g., Ljung and Box, 1979).

Similar to the result in (A7.3.3), since $\mathbf{a} = \mathbf{L}_\theta^{-1}(\mathbf{L}_\phi \mathbf{w} - \mathbf{F}\mathbf{e}_*)$ and $\mathbf{e}_*$ are independent, the joint distribution of $\mathbf{w}$ and $\mathbf{e}_*$ is

$$p(\mathbf{w}, \mathbf{e}_*|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-(n+p+q)/2} \, |\boldsymbol{\Omega}|^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2} S(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{e}_*)\right]$$

where

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{e}_*) = (\mathbf{L}_\phi \mathbf{w} - \mathbf{F}\mathbf{e}_*)' \mathbf{L}_\theta'^{-1} \mathbf{L}_\theta^{-1} (\mathbf{L}_\phi \mathbf{w} - \mathbf{F}\mathbf{e}_*) + \mathbf{e}_*' \boldsymbol{\Omega}^{-1} \mathbf{e}_*$$

Again, by generalized least-squares theory, we can show that

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{e}_*) = S(\boldsymbol{\phi}, \boldsymbol{\theta}) + (\mathbf{e}_* - \hat{\mathbf{e}}_*)' \mathbf{D} (\mathbf{e}_* - \hat{\mathbf{e}}_*)$$

where

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = S(\boldsymbol{\phi}, \boldsymbol{\theta}, \hat{\mathbf{e}}_*) = \hat{\mathbf{a}}' \hat{\mathbf{a}} + \hat{\mathbf{e}}_*' \boldsymbol{\Omega}^{-1} \hat{\mathbf{e}}_* \tag{A7.3.11}$$

is the unconditional sum-of-squares function and

$$\hat{\boldsymbol{e}}_* = E[\mathbf{e}_*|\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\theta}] = [\mathbf{e}_*] = \mathbf{D}^{-1} \mathbf{F}' \mathbf{L}_\theta'^{-1} \mathbf{L}_\theta^{-1} \mathbf{L}_\phi \mathbf{w} \tag{A7.3.12}$$

represents the conditional expectation of the preliminary values $\mathbf{e}_*$, with $\mathbf{D} = \boldsymbol{\Omega}^{-1} + \mathbf{F}' \mathbf{L}_\theta'^{-1} \mathbf{L}_\theta^{-1} \mathbf{F}$, and $\hat{\mathbf{a}} = [\mathbf{a}] = \mathbf{L}_\theta^{-1}(\mathbf{L}_\phi \mathbf{w} - \mathbf{F}\hat{\mathbf{e}}_*)$. By factorization of the joint distribution of $\mathbf{w}$ and $\mathbf{e}_*$, we can obtain

$$p(\mathbf{w}|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-n/2} |\boldsymbol{\Omega}|^{-1/2} |\mathbf{D}|^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2} S(\boldsymbol{\phi}, \boldsymbol{\theta})\right] \tag{A7.3.13}$$

as the unconditional likelihood. It follows immediately from (A7.3.13) that the maximum likelihood estimate for σ_a^2 is given by $\hat{\sigma}_a^2 = S(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})/n$, where $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$ denote maximum likelihood estimates.

Again, we note that $S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} [a_t]^2 + \hat{\mathbf{e}}_*' \boldsymbol{\Omega}^{-1} \hat{\mathbf{e}}_*$, and the elements $[a_1], [a_2], \ldots, [a_n]$ of $\hat{\mathbf{a}} = [\mathbf{a}]$ are computed recursively from the relation $\mathbf{L}_\theta[\mathbf{a}] = \mathbf{L}_\phi \mathbf{w} - \mathbf{F}[\mathbf{e}_*]$ as

$$[a_t] = w_t - \phi_1 [w_{t-1}] - \ldots - \phi_p [w_{t-p}] + \theta_1 [a_{t-1}] + \ldots + \theta_q [a_{t-q}]$$

for $t = 1, 2, \ldots, n$, using the back-forecasted values $[\mathbf{e}_*]$ for the preliminary values, with $[w_t] = w_t$ for $t = 1, 2, \ldots, n$. In addition, the back-forecasts $\hat{\mathbf{e}}_* = [\mathbf{e}_*]$ can be calculated from $\hat{\mathbf{e}}_* = \mathbf{D}^{-1} \mathbf{F}' \mathbf{u}$, where $\mathbf{u} = \mathbf{L}_\theta'^{-1} \mathbf{L}_\theta^{-1} \mathbf{L}_\phi \mathbf{w} = \mathbf{L}_\theta'^{-1} \mathbf{a}^0$, and the elements u_t of $\mathbf{u}$ are calculated through the backward recursion as

$$u_t = a_t^0 + \theta_1 u_{t+1} + \ldots + \theta_q u_{t+q}$$

with starting values $u_{n+1} = \ldots = u_{n+q} = 0$, and the a_t^0 are the elements of $\mathbf{a}^0 = \mathbf{L}_\theta^{-1} \mathbf{L}_\phi \mathbf{w}$ and denote the estimates of the a_t conditional on zero starting values $\mathbf{e}_* = \mathbf{0}$. Also, the

vector $\mathbf{h} = \mathbf{F}'\mathbf{u}$ consists of the $p + q$ elements:

$$h_j = \begin{cases} \sum_{i=1}^{j} \phi_{p-j+i} u_i & j = 1, \ldots, p \\ -\sum_{i=1}^{j-p} \theta_{q-j+p+i} u_i & j = p+1, \ldots, p+q \end{cases}$$

Finally, using (A7.1.1) and (A7.1.2), in $S(\boldsymbol{\phi}, \boldsymbol{\theta})$ we may write $\hat{\mathbf{e}}_*' \boldsymbol{\Omega}^{-1} \hat{\mathbf{e}}_* = \hat{\mathbf{a}}_*' \hat{\mathbf{a}}_* + (\hat{\mathbf{w}}_* - \mathbf{C}'\hat{\mathbf{a}}_*)' \mathbf{K}^{-1} (\hat{\mathbf{w}}_* - \mathbf{C}'\hat{\mathbf{a}}_*)$, so that we have

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1-q}^{n} [a_t]^2 + (\hat{\mathbf{w}}_* - \mathbf{C}'\hat{\mathbf{a}}_*)' \mathbf{K}^{-1} (\hat{\mathbf{w}}_* - \mathbf{C}'\hat{\mathbf{a}}_*) \tag{A7.3.14}$$

where $\mathbf{K} = \sigma_a^{-2} \Gamma_p - \mathbf{C}'\mathbf{C}$, as well as $|\boldsymbol{\Omega}| = |\mathbf{K}|$.

Therefore, in general, the likelihood associated with a series $\mathbf{z}$ of $n + d$ values generated by any ARIMA process is given by

$$L(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2 | \mathbf{z}) = (2\pi\sigma_a^2)^{-n/2} \, |\mathbf{M}_n^{(p,q)}|^{1/2} \exp\left[-\frac{S(\boldsymbol{\phi}, \boldsymbol{\theta})}{2\sigma_a^2} \right] \tag{A7.3.15}$$

where

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} [a_t]^2 + \hat{\mathbf{e}}_*' \boldsymbol{\Omega}^{-1} \hat{\mathbf{e}}_*$$

and $|\mathbf{M}_n^{(p,q)}| = |\boldsymbol{\Omega}|^{-1} |\mathbf{D}|^{-1} = |\mathbf{K}|^{-1} |\mathbf{D}|^{-1}$. Also, by expressing the mixed ARMA model as an infinite moving average $\tilde{w}_t = (1 + \psi_1 B + \psi_2 B^2 + \ldots) a_t$, and referring to results for the pure MA model, it follows that in the unconditional sum-of-squares function for the mixed model, we have the relation that $\hat{\mathbf{e}}_*' \boldsymbol{\Omega}^{-1} \hat{\mathbf{e}}_* = \sum_{t=-\infty}^{0} [a_t]^2$. Hence, we also have the representation $S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=-\infty}^{n} [a_t]^2$, and in practice the values $[a_t]$ may be computed recursively with the summation proceeding from some point $t = 1 - Q$, beyond which the $[a_t]$'s are negligible.

Special Case: AR(p). In the special case of a pure AR(p) model, the results described above simplify somewhat. We then have $\mathbf{e}_* = \mathbf{w}_*$, $\boldsymbol{\Omega} = \sigma_a^{-2} \boldsymbol{\Gamma}_p$, $\mathbf{L}_\theta = \mathbf{I}_n$, $\mathbf{D} = \sigma_a^2 \boldsymbol{\Gamma}_p^{-1} + \mathbf{F}'\mathbf{F} = \sigma_a^2 \boldsymbol{\Gamma}_p^{-1} + \mathbf{A}_p' \mathbf{A}_p$, and $\hat{\mathbf{w}}_* = \mathbf{D}^{-1} \mathbf{F}' \mathbf{L}_\phi \mathbf{w} = \mathbf{D}^{-1} \mathbf{A}_p' \mathbf{L}_{11} \mathbf{w}_p$, where $\mathbf{w}_p' = (w_1, w_2, \ldots, w_p)$ and $\mathbf{L}_{11}$ is the $p \times p$ upper left submatrix of $\mathbf{L}_\phi$. It can then be shown that the back-forecasts $\hat{w}_t$ are determined from the relations $\hat{w}_t = \phi_1 \hat{w}_{t+1} + \cdots + \phi_p \hat{w}_{t+p}$, $t = 0, -1, \ldots, 1 - p$, with $\hat{w}_t = w_t$ for $1 \le t \le n$, and hence these are the same as values obtained from the use of the backward model approach, as discussed in Section 7.1.4, for the special case of the AR model. Thus, we obtain the exact sum of squares as $S(\boldsymbol{\phi}) = \sum_{t=1}^{n} [a_t]^2 + \sigma_a^2 \hat{\mathbf{w}}_*' \boldsymbol{\Gamma}_p^{-1} \hat{\mathbf{w}}_*$.

To illustrate, consider the first-order autoregressive process in w_t,

$$w_t - \phi w_{t-1} = a_t \tag{A7.3.16}$$

where w_t might be the dth difference $\nabla^d z_t$ of the actual observations and a series $\mathbf{z}$ of length $n + d$ observations is available. To compute the likelihood (A7.3.15), we require

$$\begin{aligned} S(\phi) &= \sum_{t=1}^{n} [a_t]^2 + (1-\phi^2)\hat{w}_0^2 \\ &= \sum_{t=2}^{n} (w_t - \phi w_{t-1})^2 + (w_1 - \phi\hat{w}_0)^2 + (1-\phi^2)\hat{w}_0^2 \end{aligned}$$

since $\Gamma_1 = \gamma_0 = \sigma_a^2(1-\phi^2)^{-1}$. Now, because $\mathbf{D} = \sigma_a^2\Gamma_1^{-1} + \mathbf{A}_1'\mathbf{A}_1 = \sigma_a^2\gamma_0^{-1} + \phi^2 = 1$, and hence $\hat{w}_0 = \phi w_1$, substituting this into the last two terms of $S(\phi)$ above, it reduces to

$$S(\phi) = \sum_{t=2}^{n} (w_t - \phi w_{t-1})^2 + (1-\phi^2)w_1^2 \tag{A7.3.17}$$

as a result that may be obtained more directly by methods discussed in Appendix A7.4.

Special case: ARMA(1,1). As an example for the mixed model, consider the ARMA(1, 1) model

$$w_t - \phi w_{t-1} = a_t - \theta a_{t-1} \tag{A7.3.18}$$

Then, we have $\mathbf{e}_*' = (w_0, a_0)$, $\mathbf{A}_1 = \phi$, $B_1 = -\theta$, and

$$\sigma_a^2\mathbf{\Omega} = \sigma_a^2 \begin{bmatrix} \sigma_a^{-2}\gamma_0 & 1 \\ 1 & 1 \end{bmatrix}$$

with $\sigma_a^{-2}\gamma_0 = (1 + \theta^2 - 2\phi\theta)/(1-\phi^2)$. Thus, we have

$$\begin{aligned} \mathbf{D} = \mathbf{\Omega}^{-1} + \mathbf{F}'\mathbf{L}_\theta'^{-1}\mathbf{L}_\theta^{-1}\mathbf{F} &= \frac{1}{\sigma_a^{-2}\gamma_0 - 1} \begin{bmatrix} 1 & -1 \\ -1 & \sigma_a^{-2}\gamma_0 \end{bmatrix} \\ &+ \frac{1-\theta^{2n}}{1-\theta^2} \begin{bmatrix} \phi^2 & -\phi\theta \\ -\phi\theta & \theta^2 \end{bmatrix} \end{aligned}$$

and the estimates of the initial values are obtained as $\hat{\boldsymbol{e}}_* = \mathbf{D}^{-1}\mathbf{h}$, where $\mathbf{h}' = (h_1, h_2) = (\phi, -\theta)u_1$, the u_t are obtained from the backward recursion $u_t = a_t^0 + \theta u_{t+1}$, $u_{n+1} = 0$, and $a_t^0 = w_t - \phi w_{t-1}^0 + \theta a_{t-1}^0, t = 1, 2, \ldots, n$, are obtained using the zero initial values $w_0^0 = a_0^0 = 0$, with $w_t^0 = w_t$ for $1 \le t \le n$. Thus, the exact sum of squares is obtained as

$$S(\phi, \theta) = \sum_{t=0}^{n} [a_t]^2 + \frac{(\hat{w}_0 - \hat{a}_0)^2}{\sigma_a^{-2}\gamma_0 - 1} \tag{A7.3.19}$$

with $[a_t] = w_t - \phi[w_{t-1}] + \theta[a_{t-1}], t = 1, 2, \ldots, n$, and $\sigma_a^{-2}\gamma_0 - 1 = \mathbf{K} = (\phi - \theta)^2/(1 - \phi^2)$. In addition, we have $|\mathbf{M}_n^{(1,1)}| = \{|\mathbf{K}||\mathbf{D}|\}^{-1}$, with

$$|\mathbf{K}||\mathbf{D}| = 1 + \frac{1-\theta^{2n}}{1-\theta^2}\frac{(\phi-\theta)^2}{1-\phi^2}$$

APPENDIX A7.4 EXACT LIKELIHOOD FUNCTION FOR AN AUTOREGRESSIVE PROCESS

We now suppose that a given series $\mathbf{w}' = (w_1, w_2, \ldots, w_n)$ is generated by the pth-order stationary autoregressive model:

$$w_t - \phi_1 w_{t-1} - \phi_2 w_{t-2} - \cdots - \phi_p w_{t-p} = a_t$$

where, temporarily, the w_t's are assumed to have mean $\mu = 0$, but as before, the argument can be extended to the case where $\mu \neq 0$. Assuming normality for the a_t's and hence for the w_t's, the joint probability density function of the w_t's is

$$p(\mathbf{w}|\boldsymbol{\phi}, \sigma_a^2) = (2\pi\sigma_a^2)^{-n/2}|\mathbf{M}_n^{(p,0)}|^{1/2}\exp\left[-\frac{\mathbf{w}'\mathbf{M}_n^{(p,0)}\mathbf{w}}{2\sigma_a^2}\right] \tag{A7.4.1}$$

and because of the reversible character of the general process, the $n \times n$ matrix $\mathbf{M}_n^{(p,0)}$ is symmetric about *both* of its principal diagonals. Such a matrix is said to be *doubly* symmetric. Now,

$$p(\mathbf{w}|\boldsymbol{\phi}, \sigma_a^2) = p(w_{p+1}, w_{p+2}, \ldots, w_n|\mathbf{w}_p, \boldsymbol{\phi}, \sigma_a^2)p(\mathbf{w}_p, |\boldsymbol{\phi}, \sigma_a^2)$$

where $\mathbf{w}_p' = (w_1, w_2, \ldots, w_p)$. The first factor on the right may be obtained by making use of the distribution

$$p(a_{p+1}, \ldots, a_n) = (2\pi\sigma_a^2)^{-(n-p)/2}\exp\left[-\frac{1}{2\sigma_a^2}\sum_{t=p+1}^{n} a_t^2\right] \tag{A7.4.2a}$$

For fixed $\mathbf{w}_p$, $(a_{p+1}, \ldots, a_n)$ and $(w_{p+1}, \ldots, w_n)$ are related by the transformation

$$\begin{aligned} a_{p+1} &= w_{p+1} - \phi_1 w_p - \cdots - \phi_p w_1 \\ &\vdots \\ a_n &= w_n - \phi_1 w_{n-1} - \cdots - \phi_p w_{n-p} \end{aligned}$$

which has unit Jacobian. Thus, we obtain

$$\begin{aligned} &p(w_{p+1}, \ldots, w_n \mid \mathbf{w}_p, \phi, \sigma_a^2) \\ &\quad = (2\pi\sigma_a^2)^{-(n-p)/2}\exp\left[-\frac{1}{2\sigma_a^2}\sum_{t=p+1}^{n}(w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p})^2\right] \end{aligned} \tag{A7.4.2b}$$

Also,

$$p(\mathbf{w}_p, |\boldsymbol{\phi}, \sigma_a^2) = (2\pi\sigma_a^2)^{-p/2}|\mathbf{M}_p^{(p,0)}|^{1/2}\exp\left[-\frac{1}{2\sigma_a^2}\mathbf{w}_p'\mathbf{M}_p^{(p,0)}\mathbf{w}_p\right]$$

Thus,

$$p(\mathbf{w}|\boldsymbol{\phi}, \sigma_a^2) = (2\pi\sigma_a^2)^{-n/2}|\mathbf{M}_p^{(p,0)}|^{1/2}\exp\left[\frac{-S(\boldsymbol{\phi})}{2\sigma_a^2}\right] \tag{A7.4.3}$$

where

$$S(\boldsymbol{\phi}) = \sum_{i=1}^{p}\sum_{j=1}^{p} m_{ij}^{(p)} w_i w_j + \sum_{t=p+1}^{n} (w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p})^2 \tag{A7.4.4}$$

Also,

$$\mathbf{M}_p^{(p,0)} = \{m_{ij}^{(p)}\} = \{\gamma_{|i-j|}\}^{-1}\sigma_a^2 = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{bmatrix}^{-1} \sigma_a^2 \tag{A7.4.5}$$

where $\gamma_0, \gamma_1, \ldots, \gamma_{p-1}$ are the theoretical autocovariances of the process, and $|\mathbf{M}_p^{(p,0)}| = |\mathbf{M}_n^{(p,0)}|$.

Now, let $n = p + 1$, so that

$$\mathbf{w}'_{p+1}\mathbf{M}_{p+1}^{(p,0)}\mathbf{w}_{p+1} = \sum_{i=1}^{p}\sum_{j=1}^{p} m_{ij}^{(p)} w_i w_j + \left(w_{p+1} - \phi_1 w_p - \phi_2 w_{p-1} - \cdots - \phi_p w_1\right)^2$$

Then,

$$\mathbf{M}_{p+1}^{(p)} = \left[\begin{array}{ccc|c} & \mathbf{M}_p^{(p)} & & \begin{matrix}0\\0\\\vdots\\\vdots\end{matrix} \\ \hline 0 & 0 & \cdots & 0 \end{array}\right] + \left[\begin{array}{ccc|c} \phi_p^2 & \phi_p\phi_{p-1} & \cdots & -\phi_p \\ \phi_p\phi_{p-1} & \phi_{p-1}^2 & \cdots & -\phi_{p-1} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & -\phi_1 \\ \hline -\phi_p & -\phi_{p-1} & \cdots & 1 \end{array}\right]$$

and the elements of $\mathbf{M}_p^{(p)} = \mathbf{M}_p^{(p,0)}$ can now be deduced from the consideration that both $\mathbf{M}_p^{(p)}$ and $\mathbf{M}_{p+1}^{(p)}$ are doubly symmetric. Thus, for example,

$$\mathbf{M}_2^{(1)} = \begin{bmatrix} m_{11}^{(1)} + \phi_1^2 & -\phi_1 \\ -\phi_1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -\phi_1 \\ -\phi_1 & m_{11}^{(1)} + \phi_1^2 \end{bmatrix}$$

and after equating elements in the two matrices, we have

$$\mathbf{M}_1^{(1)} = m_{11}^{(1)} = 1 - \phi_1^2$$

Proceeding in this way, we find for processes of orders 1 and 2:

$$\mathbf{M}_1^{(1)} = 1 - \phi_1^2 \qquad |\mathbf{M}_1^{(1)}| = 1 - \phi_1^2$$

$$\mathbf{M}_2^{(2)} = \begin{bmatrix} 1 - \phi_2^2 & -\phi_1(1 + \phi_2) \\ -\phi_1(1 + \phi_2) & 1 - \phi_2^2 \end{bmatrix}$$

$$|\mathbf{M}_2^{(2)}| = (1 + \phi_2)^2[(1 - \phi_2)^2 - \phi_1^2]$$

For example, when $p = 1$,

$$p(\mathbf{w}|\phi, \sigma_a^2) = (2\pi\sigma_a^2)^{-n/2}(1-\phi^2)^{1/2}\exp\left\{-\frac{1}{2\sigma_a^2}\left[(1-\phi^2)w_1^2 + \sum_{t=2}^{n}(w_t - \phi w_{t-1})^2\right]\right\}$$

which checks with the result obtained in (A7.3.17). The process of generation must lead to matrices $\mathbf{M}_p^{(p)}$, whose elements are *quadratic* in the ϕ's.

Thus, it is clear from (A7.4.4) that not only is $S(\boldsymbol{\phi}) = \mathbf{w}'\mathbf{M}_n^{(p)}\mathbf{w}$ a quadratic form in the w_t's, but it is also quadratic in the parameters $\boldsymbol{\phi}$. Writing $\boldsymbol{\phi}_u' = (1, \phi_1, \phi_2, \ldots, \phi_p)$, it is clearly true that for some $(p+1)\times(p+1)$ matrix $\mathbf{D}$ whose elements are quadratic functions of the w_t's,

$$\mathbf{w}'\mathbf{M}_n^{(p)}\mathbf{w} = \boldsymbol{\phi}_u'\mathbf{D}\boldsymbol{\phi}_u$$

Now, write

$$\mathbf{D} = \begin{bmatrix} D_{11} & -D_{12} & -D_{13} & \cdots & -D_{1,p+1} \\ -D_{12} & D_{22} & D_{23} & \cdots & D_{2,p+1} \\ \vdots & \vdots & \vdots & & \vdots \\ -D_{1,p+1} & D_{2,p+1} & D_{3,p+1} & \cdots & D_{p+1,p+1} \end{bmatrix} \tag{A7.4.6}$$

Inspection of (A7.4.4) shows that the elements D_{ij} are ''symmetric'' sums of squares and lagged products, defined by

$$D_{ij} = D_{ji} = w_i w_j + w_{i+1}w_{j+1} + \cdots + w_{n+1-j}w_{n+1-i} \tag{A7.4.7}$$

where the sum D_{ij} contains $n - (i-1) - (j-1)$ terms.

Finally, we can write the *exact* probability density, and hence the exact likelihood, as

$$p(\mathbf{w}|\boldsymbol{\phi}, \sigma_a^2) = L(\boldsymbol{\phi}, \sigma_a^2|\mathbf{w}) = (2\pi\sigma_a^2)^{-n/2}|\mathbf{M}_p^{(p)}|^{1/2}\exp\left[\frac{-S(\boldsymbol{\phi})}{2\sigma_a^2}\right] \tag{A7.4.8}$$

where

$$S(\boldsymbol{\phi}) = \mathbf{w}_p'\mathbf{M}_p^{(p)}\mathbf{w}_p + \sum_{t=p+1}^{n}(w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p})^2 = \boldsymbol{\phi}_u'\mathbf{D}\boldsymbol{\phi}_u \tag{A7.4.9}$$

and the log-likelihood is

$$l(\boldsymbol{\phi}, \sigma_a^2|\mathbf{w}) = -\frac{n}{2}\ln(\sigma_a^2) + \frac{1}{2}\ln|\mathbf{M}_p^{(p)}| - \frac{S(\boldsymbol{\phi})}{2\sigma_a^2} \tag{A7.4.10}$$

For example, when $p = 1$, we have

$$\begin{aligned} S(\phi) &= (1-\phi^2)w_1^2 + \sum_{t=2}^{n}(w_t - \phi w_{t-1})^2 \\ &= \sum_{t=1}^{n} w_t^2 - 2\phi\sum_{t=2}^{n} w_{t-1}w_t + \phi^2\sum_{t=2}^{n-1} w_t^2 \equiv D_{11} - 2\phi D_{12} + \phi^2 D_{22} \end{aligned}$$

Maximum Likelihood Estimates. Differentiating with respect to σ_a^2 and each of the ϕ's in (A7.4.10), we obtain

$$\frac{\partial l}{\partial \sigma_a^2} = -\frac{n}{2\sigma_a^2} + \frac{S(\boldsymbol{\phi})}{2(\sigma_a^2)^2} \tag{A7.4.11}$$

$$\frac{\partial l}{\partial \phi_j} = M_j + \sigma_a^{-2}(D_{1,j+1} - \phi_1 D_{2,j+1} - \cdots - \phi_p D_{p+1,j+1})$$

$$j = 1, 2, \ldots, p \tag{A7.4.12}$$

where

$$M_j = \frac{\partial(\frac{1}{2}\ln|\mathbf{M}_p^{(p)}|)}{\partial \phi_j}$$

Hence, maximum likelihood estimates may be obtained by equating these expressions to zero and solving the resultant equations.

We have at once from (A7.4.11)

$$\hat{\sigma}_a^2 = \frac{S(\hat{\boldsymbol{\phi}})}{n} \tag{A7.4.13}$$

Estimates of $\boldsymbol{\phi}$. A difficulty occurs in dealing with equation (A7.4.12) since, in general, the quantities M_j $(j = 1, 2, \ldots, p)$ are complicated functions of the ϕ's. We consider briefly four alternative approximations.

1. *Least-Squares Estimates*. Since the expected value of $S(\boldsymbol{\phi})$ is proportional to n, while the value of $|\mathbf{M}_p^{(p)}|$ is independent of n, (A7.4.8) is for moderate or large sample sizes dominated by the term in $S(\boldsymbol{\phi})$ and the term in $|\mathbf{M}_p^{(p)}|$ is, by comparison, small.

If we ignore the influence of this term, then

$$l(\boldsymbol{\phi}, \sigma_a^2|\mathbf{w}) \simeq -\frac{n}{2}\ln(\sigma_a^2) - \frac{S(\boldsymbol{\phi})}{2\sigma_a^2} \tag{A7.4.14}$$

and the estimates $\hat{\boldsymbol{\phi}}$ of $\boldsymbol{\phi}$ obtained by maximization of (A7.4.14) are the least-squares estimates obtained by minimizing $S(\boldsymbol{\phi})$. Now, from (A7.4.9), $S(\boldsymbol{\phi}) = \boldsymbol{\phi}_u' \mathbf{D} \boldsymbol{\phi}_u$, where $\mathbf{D}$ is a $(p+1) \times (p+1)$ matrix of symmetric sums of squares and products, defined in (A7.4.7). Thus, on differentiating, the minimizing values are

$$\begin{aligned}
D_{12} &= \hat{\phi}_1 D_{22} + \hat{\phi}_2 D_{23} + \cdots + \hat{\phi}_p D_{2,p+1} \\
D_{13} &= \hat{\phi}_1 D_{23} + \hat{\phi}_2 D_{33} + \cdots + \hat{\phi}_p D_{3,p+1} \\
&\vdots \\
D_{1,p+1} &= \hat{\phi}_1 D_{2,p+1} + \hat{\phi}_2 D_{3,p+1} + \cdots + \hat{\phi}_p D_{p+1,p+1}
\end{aligned} \tag{A7.4.15}$$

which, in an obvious matrix notation, can be written as

$$\mathbf{d} = \mathbf{D}_p \hat{\boldsymbol{\phi}}$$

so that

$$\hat{\boldsymbol{\phi}} = \mathbf{D}_p^{-1}\mathbf{d}$$

These least-squares estimates also maximize the posterior density (7.5.15).

2. *Approximate Maximum Likelihood Estimates*. We now recall an earlier result (3.2.3), which may be written as

$$\gamma_j - \phi_1\gamma_{j-1} - \phi_2\gamma_{j-2} - \cdots - \phi_p\gamma_{j-p} = 0 \qquad j > 0 \tag{A7.4.16}$$

Also, on taking expectations in (A7.4.12) and using the fact that $E[\partial l/\partial\phi_j] = 0$, we obtain

$$M_j\sigma_a^2 + (n-j)\gamma_j - (n-j-1)\phi_1\gamma_{j-1} - (n-j-2)\phi_2\gamma_{j-2} - \cdots - (n-j-p)\phi_p\gamma_{j-p} = 0 \tag{A7.4.17}$$

After multiplying (A7.4.16) by n and subtracting the result from (A7.4.17), we obtain

$$M_j\sigma_a^2 = j\gamma_j - (j+1)\phi_1\gamma_{j-1} - \cdots - (j+p)\phi_p\gamma_{j-p}$$

Therefore, on using $D_{i+1,j+1}/(n-j-i)$ as an estimate of $\gamma_{|j-i|}$, a natural estimate of $M_j\sigma_a^2$ is

$$j\frac{D_{1,j+1}}{n-j} - (j+1)\phi_1\frac{D_{2,j+1}}{n-j-1} - \cdots - (j+p)\phi_p\frac{D_{p+1,j+1}}{n-j-p}$$

Substituting this estimate in (A7.4.12) yields

$$\frac{\partial l}{\partial\phi_j} \simeq n\sigma_a^{-2}\left(\frac{D_{1,j+1}}{n-j} - \phi_1\frac{D_{2,j+1}}{n-j-1} - \cdots - \phi_p\frac{D_{p+1,j+1}}{n-j-p}\right) \qquad j = 1, 2, \ldots, p \tag{A7.4.18}$$

leading to a set of linear equations of the form (A7.4.15), but now with

$$D_{ij}^* = \frac{nD_{ij}}{n-(i-1)-(j-1)}$$

replacing D_{ij}.

3. *Conditional Least-Squares Estimates*. For moderate and relatively large n, we might also consider the conditional sum-of-squares function, obtained by adopting the procedure in Section 7.1.3. This yields the sum of squares given in the exponent of the expression in (A7.4.2),

$$S_*(\boldsymbol{\phi}) = \sum_{t=p+1}^{n}(w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p})^2$$

and is the sum of squares associated with the conditional distribution of $w_{p+1}, \ldots, w_n$, given $\mathbf{w}_p' = (w_1, w_2, \ldots, w_p)$. Conditional least-squares estimates are obtained by minimizing $S_*(\boldsymbol{\phi})$, which is a standard linear least-squares regression problem

associated with the linear model $w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \cdots + \phi_p w_{t-p} + a_t, t = p+1, \ldots, n$. This results in the familiar least-squares estimates $\hat{\boldsymbol{\phi}} = \tilde{\mathbf{D}}_p^{-1}\tilde{\mathbf{d}}$, as in (A7.2.5), where $\tilde{\mathbf{D}}_p$ has (i, j) th element $\tilde{D}_{ij} = \sum_{t=p+1}^{n} w_{t-i} w_{t-j}$ and $\tilde{\mathbf{d}}$ has ith element $\tilde{d}_i = \sum_{t=p+1}^{n} w_{t-i} w_t$.

4. *Yule–Walker Estimates*. Finally, if n is moderate or large, as an approximation, we may replace the symmetric sums of squares and products in (A7.4.15) by n times the appropriate autocovariance estimate. For example, D_{ij}, where $|i - j| = k$, would be replaced by $nc_k = \sum_{t=1}^{n-k} \tilde{w}_t \tilde{w}_{t+k}$. On dividing by nc_0 throughout in the resultant equations, we obtain the following relations expressed in terms of the estimated autocorrelations $r_k = c_k/c_0$:

$$\begin{aligned} r_1 &= \hat{\phi}_1 + \hat{\phi}_2 r_1 + \cdots + \hat{\phi}_p r_{p-1} \\ r_2 &= \hat{\phi}_1 r_1 + \hat{\phi}_2 + \cdots + \hat{\phi}_p r_{p-2} \\ &\vdots \\ r_p &= \hat{\phi}_1 r_{p-1} + \hat{\phi}_2 r_{p-2} + \cdots + \hat{\phi}_p \end{aligned}$$

These are the well-known Yule–Walker equations.

In the matrix notation (7.3.1), they can be written $\mathbf{r} = \mathbf{R}\hat{\boldsymbol{\phi}}$, so that

$$\hat{\boldsymbol{\phi}} = \mathbf{R}^{-1}\mathbf{r} \tag{A7.4.19}$$

which corresponds to equations (3.2.7), with $\mathbf{r}$ substituted for $\boldsymbol{\rho}_p$ and $\mathbf{R}$ for $\mathbf{P}_p$.

To illustrate the differences among the four estimates, take the case $p = 1$. Then, $M_1\sigma_a^2 = -\gamma_1$ and, corresponding to (A7.4.12), the exact maximum likelihood estimate of ϕ is the solution of

$$-\gamma_1 + D_{12} - \phi D_{22} \equiv -\gamma_1 + \sum_{t=2}^{n} w_t w_{t-1} - \phi \sum_{t=2}^{n-1} w_t^2 = 0$$

Note that $\gamma_1 = \sigma_a^2 \phi/(1 - \phi^2)$ and the maximum likelihood solution for $\sigma_a^2, \hat{\sigma}_a^2 = S(\phi)/n$ from (A7.4.13), can be substituted in the expression for γ_1 in the likelihood equation above, where $S(\phi) = D_{11} - 2\phi D_{12} + \phi^2 D_{22}$ as in (A7.4.9). This results in a *cubic* equation in ϕ, whose solution yields the maximum likelihood estimate of ϕ. Upon rearranging, the cubic equation for $\hat{\phi}$ can be written as

$$(n-1)D_{22}\hat{\phi}^3 - (n-2)D_{12}\hat{\phi}^2 - (nD_{22} + D_{11})\hat{\phi} + nD_{12} = 0 \tag{A7.4.20}$$

and there is a single *unique* solution to this cubic equation such that $-1 < \hat{\phi} < 1$ (e.g., Anderson, 1971, p. 354).

Approximation 1 corresponds to ignoring the term γ_1 altogether, yielding

$$\hat{\phi} = \frac{\sum_{t=2}^{n} w_t w_{t-1}}{\sum_{t=2}^{n-1} w_1^2} = \frac{D_{12}}{D_{22}}$$

Approximation 2 corresponds to substituting the estimate $\sum_{t=2}^{n} w_t w_{t-1}/(n-1)$ for γ_1, yielding

$$\hat{\phi} = \frac{\sum_{t=2}^{n} w_t w_{t-1}/(n-1)}{\sum_{t=2}^{n-1} w_t^2/(n-2)} = \frac{n-2}{n-1}\frac{D_{12}}{D_{22}}$$

Approximation 3 corresponds to the standard linear model least-squares estimate obtained by regression of w_t on w_{t-1} for $t = 2, 3, \ldots, n$, so that

$$\hat{\phi} = \frac{\sum_{t=2}^{n} w_t w_{t-1}}{\sum_{t=2}^{n} w_{t-1}^2} = \frac{D_{12}}{D_{22} + w_1^2}$$

In effect, this can be viewed as obtained by substituting ϕw_1^2 for γ_1 in the likelihood equation above for ϕ.

Approximation 4 replaces the numerator and denominator by standard autocovariance estimates (2.1.12), yielding

$$\hat{\phi} = \frac{\sum_{t=2}^{n} w_t w_{t-1}}{\sum_{t=1}^{n} w_1^2} = \frac{c_1}{c_0} = r_1 = \frac{D_{12}}{D_{11}}$$

Usually, as in this example, for moderate and large samples, the differences between the estimates given by the various approximations will be small. We have often employed the least-squares estimates given by approximation 1 which can be computed directly from (A7.4.15). However, for computer calculations, it is often simplest, even when the fitted model is autoregressive, to use the general iterative algorithm described in Section 7.2.1, which computes least-squares estimates for any ARMA process.

Estimate of σ_a^2. Using approximation 4 with (A7.4.9) and (A7.4.13),

$$\hat{\sigma}_a^2 = \frac{S(\hat{\boldsymbol{\phi}})}{n} = c_0 \left[1 : \hat{\boldsymbol{\phi}}'\right] \left[\begin{array}{c|c} 1 & -\mathbf{r}' \\ \hline -\mathbf{r} & \mathbf{R} \end{array}\right] \left[\begin{array}{c} 1 \\ \hline \hat{\boldsymbol{\phi}} \end{array}\right]$$

On multiplying out the right-hand side and recalling that $\mathbf{r} - \mathbf{R}\hat{\boldsymbol{\phi}} = \mathbf{0}$, we find that

$$\hat{\sigma}_a^2 = c_0(1 - \mathbf{r}'\hat{\boldsymbol{\phi}}) = c_0(1 - \mathbf{r}'\mathbf{R}^{-1}\mathbf{r}) = c_0(1 - \hat{\boldsymbol{\phi}}'\mathbf{R}\hat{\boldsymbol{\phi}}) \tag{A7.4.21a}$$

It is readily shown that σ_a^2 can be similarly written in terms of the *theoretical* autocorrelations:

$$\sigma_a^2 = \gamma_0(1 - \boldsymbol{\rho}'\boldsymbol{\phi}) = \gamma_0(1 - \boldsymbol{\rho}'\mathbf{P}_p^{-1}\boldsymbol{\rho}) = \gamma_0(1 - \boldsymbol{\phi}'\mathbf{P}_p\boldsymbol{\phi}) \tag{A7.4.21b}$$

agreeing with the result (3.2.8).

Parallel expressions for $\hat{\sigma}_a^2$ may be obtained for approximations 1, 2, and 3.

Information Matrix. Differentiating for a second time in (A7.4.11) and (A7.4.18), we obtain

$$-\frac{\partial^2 l}{\partial(\sigma_a^2)^2} = -\frac{n}{2(\sigma_a^2)^2} + \frac{S(\boldsymbol{\phi})}{(\sigma_a^2)^3} \tag{A7.4.22a}$$

$$\frac{\partial^2 l}{\partial(\sigma_a^2)\partial\phi_j} \simeq -\sigma_a^{-2}\frac{\partial l}{\partial\phi_j} \tag{A7.4.22b}$$

$$-\frac{\partial^2 l}{\partial\phi_i\partial\phi_j} \simeq \frac{n}{\sigma_a^2}\frac{D_{i+1,j+1}}{n-i-j} \tag{A7.4.22c}$$

Now, since

$$E\left[\frac{\partial l}{\partial\phi_j}\right] = 0$$

it follows that for moderate or large samples,

$$E\left[-\frac{\partial^2 l}{\partial(\sigma_a^2)\partial\phi_j}\right] \simeq 0$$

and

$$|\mathbf{I}(\boldsymbol{\phi}, \sigma_a^2)| \simeq |\mathbf{I}(\boldsymbol{\phi})| I(\sigma_a^2)$$

where

$$I(\sigma_a^2) = E\left[-\frac{\partial^2 l}{\partial(\sigma_a^2)^2}\right] = \frac{n}{2(\sigma_a^2)^2}$$

Now, using (A7.4.22c), we have

$$\mathbf{I}(\boldsymbol{\phi}) = -E\left[\frac{\partial^2 l}{\partial\phi_i\partial\phi_j}\right] \simeq \frac{n}{\sigma_a^2}\boldsymbol{\Gamma}_p = \frac{n\boldsymbol{\gamma}_0}{\sigma_a^2}\mathbf{P}_p = n(\mathbf{M}_p^{(p)})^{-1} \tag{A7.4.23}$$

Hence,

$$|\mathbf{I}(\boldsymbol{\phi}, \sigma_a^2)| \simeq \frac{n^{p+1}}{2(\sigma_a^2)^2}|\mathbf{M}_p^{(p,0)}|^{-1}$$

Variances and Covariances of Estimates of Autoregressive Parameters. Now, in circumstances fully discussed by Whittle (1953), the inverse of the information matrix supplies the asymptotic variance–covariance matrix of the maximum likelihood (ML) estimates. Moreover, if the log-likelihood is approximately quadratic and the maximum is not close to a boundary, even if the sample size is only moderate, the elements of this matrix will normally provide adequate approximations to the variances and covariances of the estimates.

Thus, using (A7.4.23) and (A7.4.21b) gives

$$\begin{aligned}\mathbf{V}(\hat{\boldsymbol{\phi}}) = \mathbf{I}^{-1}(\boldsymbol{\phi}) &\simeq n^{-1}\mathbf{M}_p^{(p)} = n^{-1}\sigma_a^2\boldsymbol{\Gamma}_p^{-1} \\ &= n^{-1}(1-\boldsymbol{\rho}'\mathbf{P}_p^{-1}\boldsymbol{\rho})\mathbf{P}_p^{-1} \\ &= n^{-1}(1-\boldsymbol{\phi}'\mathbf{P}_p\boldsymbol{\phi})\mathbf{P}_p^{-1} = n^{-1}(1-\boldsymbol{\rho}'\boldsymbol{\phi})\mathbf{P}_p^{-1}\end{aligned} \tag{A7.4.24}$$

In particular, for autoregressive process of first and second order,

$$V(\hat{\phi}) \simeq n^{-1}(1-\phi^2)$$
$$V(\hat{\phi}_1, \hat{\phi}_2) \simeq n^{-1} \begin{bmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{bmatrix} \tag{A7.4.25}$$

Estimates of the variances and covariances may be obtained by substituting estimates for the parameters in (A7.4.25). For example, we may substitute r_j's for ρ_j's and $\hat{\boldsymbol{\phi}}$ for $\boldsymbol{\phi}$ in (A7.4.24) to obtain

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\phi}}) = n^{-1}(1 - \mathbf{r}'\hat{\boldsymbol{\phi}})\mathbf{R}^{-1} \tag{A7.4.26}$$

APPENDIX A7.5 ASYMPTOTIC DISTRIBUTION OF ESTIMATORS FOR AUTOREGRESSIVE MODELS

We provide details on the asymptotic distribution of least-squares estimator of the parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ for a *stationary* AR(p) model [i.e., all roots of $\phi(B) = 0$ lie outside the unit circle],

$$w_t = \sum_{i=1}^{p} \phi_i w_{t-i} + a_t$$

based on a sample of n observations, where the w_t are assumed to have mean $\mu = 0$ for simplicity, and the a_t are assumed to be independent random variates, with zero means, variances σ_a^2, and finite fourth moments. It is then established that

$$n^{1/2}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{\mathcal{D}} N\{\mathbf{0}, \sigma_a^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi})\} \tag{A7.5.1}$$

as $n \to \infty$, where $\boldsymbol{\Gamma}_p(\boldsymbol{\phi})$ is the $p \times p$ autocovariance matrix of p successive values from the AR(p) process. Hence, for large n the distribution of $\hat{\boldsymbol{\phi}}$ is approximately normal with mean vector $\boldsymbol{\phi}$ and covariance matrix $\mathbf{V}(\hat{\boldsymbol{\phi}}) \simeq n^{-1}\sigma_a^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi})$, that is, $N\{\boldsymbol{\phi}, n^{-1}\sigma_a^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi})\}$.

We can write the AR(p) model as

$$w_t = \mathbf{w}_{t-1}'\boldsymbol{\phi} + a_t \tag{A7.5.2}$$

where $\mathbf{w}_{t-1}' = (w_{t-1}, \ldots, w_{t-p})$. For convenience, assume that observations $w_{1-p}, \ldots, w_0$ are available in addition to $w_1, \ldots, w_n$, so that the (conditional) least-squares estimator of $\boldsymbol{\phi}$ is obtained by minimizing the sum of squares:

$$S(\boldsymbol{\phi}) = \sum_{t=1}^{n} (w_t - \mathbf{w}_{t-1}'\boldsymbol{\phi})^2$$

As $n \to \infty$, the treatment of the p initial observations becomes negligible, so that conditional and unconditional LS estimators are asymptotically equivalent. From the standard results on LS estimates for regression models, we know that the LS estimate of $\boldsymbol{\phi}$ in the AR(p)

model (A7.5.2) is then given by

$$\hat{\boldsymbol{\phi}} = \left(\sum_{t=1}^{n} \mathbf{w}_{t-1} \mathbf{w}'_{t-1} \right)^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} w_t \tag{A7.5.3}$$

Substituting the expression for w_t from (A7.5.2) in (A7.5.3), we see that

$$\hat{\boldsymbol{\phi}} = \boldsymbol{\phi} + \left(\sum_{t=1}^{n} \mathbf{w}_{t-1} \mathbf{w}'_{t-1} \right)^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} a_t$$

so that

$$n^{1/2}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) = \left(n^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} \mathbf{w}'_{t-1} \right)^{-1} n^{-1/2} \sum_{t=1}^{n} \mathbf{w}_{t-1} a_t \tag{A7.5.4}$$

Notice that the information matrix for this model situation is simply

$$\mathbf{I}(\boldsymbol{\phi}) = -\frac{1}{2\sigma_a^2} E\left[\frac{\partial^2 S(\boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'} \right] = \frac{1}{\sigma_a^2} \sum_{t=1}^{n} E[\mathbf{w}_{t-1} \mathbf{w}'_{t-1}] = \frac{n}{\sigma_a^2} \boldsymbol{\Gamma}_p(\boldsymbol{\phi})$$

so that $n\mathbf{I}^{-1}(\boldsymbol{\phi}) \equiv \mathbf{I}_*^{-1}(\boldsymbol{\phi}) = \sigma_a^2 \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi})$ as appears in (A7.5.1).

We let $U_t = \mathbf{w}_{t-1} a_t$ and argue that these terms have zero mean, covariance matrix $\sigma_a^2 \boldsymbol{\Gamma}_p(\boldsymbol{\phi})$, and are mutually uncorrelated. That is, noting that $\mathbf{w}_{t-1}$ and a_t are independent (e.g., elements of $\mathbf{w}_{t-1}$ are functions of $a_{t-1}, a_{t-2}, \ldots$, independent of a_t), we have $E[\mathbf{w}_{t-1} a_t] = E[\mathbf{w}_{t-1}] E[a_t] = 0$, and again by independence of the terms a_t^2 and $\mathbf{w}_{t-1} \mathbf{w}'_{t-1}$,

$$\operatorname{cov}[\mathbf{w}_{t-1} a_t] = E[\mathbf{w}_{t-1} a_t a_t \mathbf{w}'_{t-1}] = E[a_t^2] E[\mathbf{w}_{t-1} \mathbf{w}'_{t-1}] = \sigma_a^2 \boldsymbol{\Gamma}_p(\boldsymbol{\phi})$$

In addition, for any $l > 0$,

$$\begin{aligned} \operatorname{cov}[\mathbf{w}_{t-1} a_t, \mathbf{w}_{t+l-1} a_{t+l}] &= E[\mathbf{w}_{t-1} a_t a_{t+l} \mathbf{w}'_{t+l-1}] \\ &= E[a_t \mathbf{w}_{t-1} \mathbf{w}'_{t+l-1}] E[a_{t+l}] = 0 \end{aligned}$$

because a_{t+l} is independent of the other terms. By similar reasoning,

$$\operatorname{cov}[\mathbf{w}_{t-1} a_t, \mathbf{w}_{t+l-1} a_{t+l}] = 0$$

for any $l < 0$. Hence, the quantity $\sum_{t=1}^{n} \mathbf{w}_{t-1} a_t$ in (A7.5.4) is the sum of n uncorrelated terms each with zero mean and covariance matrix $\sigma_a^2 \boldsymbol{\Gamma}_p(\boldsymbol{\phi})$.

Now, in fact, the partial sums

$$\mathrm{S}_n = \sum_{t=1}^{n} \mathrm{U}_t \equiv \sum_{t=1}^{n} \mathbf{w}_{t-1} a_t \qquad n = 1, 2, \ldots$$

form a *martingale* sequence (with respect to the σ fields generated by the collection of random variables $\{a_n, a_{n-1}, \ldots\}$), characterized by the property that $E[\mathrm{S}_{n+1} | a_n, a_{n-1}, \ldots] = \mathrm{S}_n$. This clearly holds since $\mathrm{S}_{n+1} = \mathbf{w}_n a_{n+1} + \mathrm{S}_n$,

$$E[\mathbf{w}_n a_{n+1} | a_n, a_{n-1}, \ldots] = \mathbf{w}_n E[a_{n+1} | a_n, a_{n-1}, \ldots] = \mathbf{w}_n E[a_{n+1}] = 0$$

and $S_n = \sum_{t=1}^{n} \mathbf{w}_{t-1} a_t$ is a function of $a_n, a_{n-1}, \ldots$ so that $E[S_n | a_n, a_{n-1}, \ldots] = S_n$. In this context, the terms $U_t = \mathbf{w}_{t-1} a_t$ are referred to as *a martingale difference* sequence. Then, by a martingale central limit theorem (e.g., Billingsley, 1999),

$$n^{-1/2} \mathbf{c}' S_n \xrightarrow{\mathcal{D}} N\{\mathbf{0}, \sigma_a^2 \mathbf{c}' \boldsymbol{\Gamma}_p(\boldsymbol{\phi}) \mathbf{c}\}$$

for any vector or constants $\mathbf{c}' = (c_1, \ldots, c_p)$, and by use of the Cramer–Wold device, it follows that

$$n^{-1/2} S_n \equiv n^{-1/2} \sum_{t=1}^{n} \mathbf{w}_{t-1} a_t \xrightarrow{\mathcal{D}} N\{\mathbf{0}, \sigma_a^2 \boldsymbol{\Gamma}_p(\boldsymbol{\phi})\} \tag{A7.5.5}$$

as $n \to \infty$. Also, we know that the matrix $n^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} \mathbf{w}'_{t-1} \xrightarrow{\mathcal{P}} \boldsymbol{\Gamma}_p(\boldsymbol{\phi})$, as an $n \to \infty$, by a weak law of large numbers, since the (i, j) th element of the matrix is $\hat{\gamma}(i-j) = n^{-1} \sum_{t=1}^{n} w_{t-1} w_{t-j}$, which converges in probability to $\gamma(i-j)$ by consistency of sample autocovariances $\hat{\gamma}(i-j)$. Hence, it follows by continuity that

$$\left(n^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} \mathbf{w}'_{t-1} \right)^{-1} \xrightarrow{\mathcal{P}} \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi}) \tag{A7.5.6}$$

Therefore, by a standard limit theory result, applying (A7.5.5) and (A7.5.6) in (A7.5.4), we obtain that

$$n^{1/2}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{\mathcal{D}} \boldsymbol{\Gamma}_p^{-1}(\boldsymbol{\phi}) N\{\mathbf{0}, \sigma_a^2 \boldsymbol{\Gamma}_p(\boldsymbol{\phi})\} \tag{A7.5.7}$$

which leads to the result (A7.5.1).

In addition, it is easily shown that the Yule–Walker (YW) estimator $\tilde{\boldsymbol{\phi}} = \mathbf{R}^{-1} \mathbf{r}$, discussed in Section 7.3.1, is asymptotically equivalent to the LS estimator considered here, in the sense that

$$n^{1/2}(\hat{\boldsymbol{\phi}} - \tilde{\boldsymbol{\phi}}) \xrightarrow{\mathcal{P}} \mathbf{0}$$

as $n \to \infty$. For instance, we can write the YW estimate as $\tilde{\boldsymbol{\phi}} = \tilde{\boldsymbol{\Gamma}}_p^{-1} \tilde{\boldsymbol{\gamma}}_p$ where $\tilde{\boldsymbol{\Gamma}}_p = \hat{\gamma}_0 \mathbf{R}$ and $\tilde{\boldsymbol{\gamma}}_p = \hat{\gamma}_0 \mathbf{r}$. For notational convenience, we write the LS estimate in (A7.5.3) as $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p$ where we denote $\hat{\boldsymbol{\Gamma}}_p = n^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} \mathbf{w}'_{t-1}$ and $\hat{\boldsymbol{\gamma}}_p = n^{-1} \sum_{t=1}^{n} \mathbf{w}_{t-1} w_t$. Then, we have

$$\begin{aligned} n^{1/2}(\hat{\boldsymbol{\phi}} - \tilde{\boldsymbol{\phi}}) &= n^{1/2}(\hat{\boldsymbol{\Gamma}}_p^{-1} \hat{\boldsymbol{\gamma}}_p - \tilde{\boldsymbol{\Gamma}}_p^{-1} \tilde{\boldsymbol{\gamma}}_p) \\ &= n^{1/2} \hat{\boldsymbol{\Gamma}}_p^{-1} (\hat{\boldsymbol{\gamma}}_p - \tilde{\boldsymbol{\gamma}}_p) + n^{1/2} (\hat{\boldsymbol{\Gamma}}_p^{-1} - \tilde{\boldsymbol{\Gamma}}_p^{-1}) \tilde{\boldsymbol{\gamma}}_p \end{aligned} \tag{A7.5.8}$$

and we can readily determine that both $n^{1/2}(\hat{\boldsymbol{\gamma}}_p - \tilde{\boldsymbol{\gamma}}_p) \xrightarrow{\mathcal{P}} \mathbf{0}$ and $n^{1/2}(\hat{\boldsymbol{\Gamma}}_p - \tilde{\boldsymbol{\Gamma}}_p) \xrightarrow{\mathcal{P}} \mathbf{0}$ as $n \to \infty$, and consequently also

$$n^{1/2}(\hat{\boldsymbol{\Gamma}}_p^{-1} - \tilde{\boldsymbol{\Gamma}}_p^{-1}) = \tilde{\boldsymbol{\Gamma}}_p^{-1} n^{1/2} (\tilde{\boldsymbol{\Gamma}}_p - \hat{\boldsymbol{\Gamma}}_p) \hat{\boldsymbol{\Gamma}}_p^{-1} \xrightarrow{\mathcal{P}} \mathbf{0}$$

Therefore, $n^{1/2}(\hat{\boldsymbol{\phi}} - \tilde{\boldsymbol{\phi}}) \xrightarrow{\mathcal{P}} \mathbf{0}$ follows directly from (A7.5.8).

APPENDIX A7.6 EXAMPLES OF THE EFFECT OF PARAMETER ESTIMATION ERRORS ON VARIANCES OF FORECAST ERRORS AND PROBABILITY LIMITS FOR FORECASTS

The variances and probability limits for the forecasts given in Section 5.2.4 are based on the assumption that the parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$ in the ARIMA model are known exactly. In practice, it is necessary to replace these by their estimates $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$. To gain some insight into the effect of estimation errors on the variance of the forecast errors, we consider the special cases of the nonstationary IMA(0, 1, 1) and the stationary first-order autoregressive processes. It is shown that for these processes and for parameter estimates based on series of moderate length, the effect of such estimation errors is small.

IMA(0, 1, 1) Processes. Writing the model $\nabla z_t = a_t - \theta a_{t-1}$ for $t + l, t + l - 1, \ldots, t + 1$, and summing, we obtain

$$z_{t+l} - z_t = a_{t+l} + (1 - \theta)(a_{t+l-1} + \cdots + a_{t+1}) - \theta a_t$$

Denote by $\hat{z}_t(l|\theta)$ the lead l forecast when the parameter θ is known exactly. On taking conditional expectations at time t, for $l = 1, 2, \ldots$, we obtain

$$\begin{aligned} \hat{z}_t(1|\theta) &= z_t - \theta a_t \\ \hat{z}_t(l|\theta) &= \hat{z}_t(1|\theta) \qquad l \geq 2 \end{aligned}$$

Hence, the lead l forecast error is

$$\begin{aligned} e_t(l|\theta) &= z_{t+l} - \hat{z}_t(l|\theta) \\ &= a_{t+l} + (1 - \theta)(a_{t+l-1} + \cdots + a_{t+1}) \end{aligned}$$

and the variance of the forecast error at lead time l is

$$V(l) = E_t[e_t^2(l|\theta)] = \sigma_a^2[1 + (l - 1)\lambda^2] \tag{A7.6.1}$$

where $\lambda = 1 - \theta$.

However, if θ is replaced by its estimate $\hat{\theta}$, obtained from a time series consisting of n values of $w_t = \nabla z_t$, then,

$$\begin{aligned} \hat{z}_t(1|\hat{\theta}) &= z_t - \hat{\theta}\hat{a}_t \\ \hat{z}_t(l|\hat{\theta}) &= \hat{z}_t(1|\hat{\theta}) \qquad l \geq 2 \end{aligned}$$

where $\hat{a}_t = z_t - \hat{z}_{t-1}(1|\hat{\theta})$. Hence, the lead l forecast error using $\hat{\theta}$ is

$$\begin{aligned} e_t(l|\hat{\theta}) &= z_{t+l} - \hat{z}(l|\hat{\theta}) \\ &= z_{t+l} - z_t + \hat{\theta}\hat{a}_t \\ &= e_t(l|\theta) - (\theta a_t - \hat{\theta}\hat{a}_t) \end{aligned} \tag{A7.6.2}$$

Since $\nabla z_t = (1 - \theta B)a_t = (1 - \hat{\theta}B)\hat{a}_t$, it follows that

$$\hat{a}_t = \left(\frac{1 - \theta B}{1 - \hat{\theta}B}\right) a_t$$

and on eliminating $\hat{a}_t$ from (A7.6.2), we obtain

$$e_t(l|\hat{\theta}) = e_t(l|\theta) - \frac{\theta - \hat{\theta}}{1 - \hat{\theta}B} a_t$$

Now,

$$\begin{aligned}
\frac{\theta - \hat{\theta}}{1 - \hat{\theta}B} a_t &= \frac{\theta - \hat{\theta}}{1 - \theta B} \left[1 + \frac{(\theta - \hat{\theta})B}{1 - \theta B} \right]^{-1} a_t \\
&\simeq \frac{\theta - \hat{\theta}}{1 - \theta B} \left[1 - \frac{(\theta - \hat{\theta})B}{1 - \theta B} \right] a_t \\
&= (\theta - \hat{\theta})(a_t + \theta a_{t-1} + \theta^2 a_{t-2} + \cdots) \\
&\quad - (\theta - \hat{\theta})^2 (a_{t-1} + 2\theta a_{t-2} + 3\theta^2 a_{t-3} + \cdots)
\end{aligned} \tag{A7.6.3}$$

On the assumption that the forecast and the estimate $\hat{\theta}$ are based on essentially nonoverlapping data, $\hat{\theta}$ and $a_t, a_{t-1}, \ldots$ are independent. Also, $\hat{\theta}$ will be approximately normally distributed about θ with variance $(1 - \theta^2)/n$, for moderate-sized samples. On these assumptions the variance of the expression in (A7.6.3) may be shown to be

$$\frac{\sigma_a^2}{n} \left(1 + \frac{3}{n} \frac{1 + \theta^2}{1 - \theta^2} \right)$$

Thus, provided that $|\theta|$ is not close to unity,

$$\text{var}[e_t(l|\hat{\theta})] \simeq \sigma_a^2 [1 + (l - 1)\lambda^2] + \frac{\sigma_a^2}{n} \tag{A7.6.4}$$

Clearly, the proportional change in the variance will be greatest for $l = 1$, when the exact forecast error variance reduces to σ_a^2. In this case, for parameter estimates based on a series of moderate length, the probability limits will be increased by a factor $(n + 1)/n$.

First-Order Autoregressive Processes. Writing the AR(1) model $\tilde{z}_t = \phi \tilde{z}_{t-1} + a_t$ at time $t + l$ and taking conditional expectations at time t, the lead l forecast, given the true value of the parameter ϕ, is

$$\hat{\tilde{z}}_t(l|\phi) = \phi \hat{\tilde{z}}_t(l - 1|\phi) = \phi^l \tilde{z}_t$$

Similarly,

$$\hat{\tilde{z}}_t(l|\hat{\phi}) = \hat{\phi} \hat{\tilde{z}}_t(l - 1|\hat{\phi}) = \hat{\phi}^l \tilde{z}_t$$

and hence

$$e_t(l|\hat{\phi}) = \tilde{z}_{t+l} - \hat{\tilde{z}}_t(l|\hat{\phi}) = e_t(l|\phi) + (\phi^l - \hat{\phi}^l)\tilde{z}_t \tag{A7.6.5}$$

Because $e_t(l|\phi) = \tilde{z}_{t+1} - \hat{\tilde{z}}_t(l|\phi) = a_{t+1} + \phi a_{t+l-1} + \cdots + \phi^{l-1} a_{t+1}$ is independent of $\hat{\phi}$ and $\tilde{z}_t$, it follows from (A7.6.5) that

$$E[e_t^2(l|\hat{\phi})] = E[e_t^2(l|\phi)] + E[\tilde{z}_t^2(\phi^l - \hat{\phi}^l)^2]$$

Again, as in the MA(1) case, the estimate $\hat{\phi}$ is assumed to be essentially independent of $\tilde{z}_t$, and for sufficiently large n, $\hat{\phi}$ will be approximately normally distributed about a mean ϕ with variance $(1-\phi^2)/n$. So using (5.4.16) and $E[\tilde{z}_t^2(\phi^l - \hat{\phi}^l)^2] \simeq E[\tilde{z}_t^2]E[(\phi^l - \hat{\phi}^l)^2]$, with $E[\tilde{z}_t^2] = \gamma_0 = \sigma_a^2/(1-\phi^2)$, *on the average*

$$\text{var}[e_t(l|\hat{\phi})] \simeq \sigma_a^2 \frac{1-\phi^{2l}}{1-\phi^2} + \sigma_a^2 \frac{E[(\phi^l - \hat{\phi}^l)^2]}{1-\phi^2} \tag{A7.6.6}$$

When $l = 1$, using $E[(\phi - \hat{\phi})^2] \simeq (1-\phi^2)/n$,

$$\begin{aligned} \text{var}[e_t(1|\hat{\phi})] &\simeq \sigma_a^2 + \frac{\sigma_a^2}{1-\phi^2}\frac{1-\phi^2}{n} \\ &= \sigma_a^2\left(1 + \frac{1}{n}\right) \end{aligned} \tag{A7.6.7}$$

For $l > 1$, we have

$$\phi^l - \hat{\phi}^l = \phi^l - \{\phi - (\phi - \hat{\phi})\}^l \simeq \phi^l - \{\phi^l - l\phi^{l-1}(\phi - \hat{\phi})\} = l\phi^{l-1}(\phi - \hat{\phi})$$

since the remaining terms involving $(\phi - \hat{\phi})^j$ for $j = 2, \ldots, l$ are of smaller order. Thus, on the average, from (A7.6.6) we obtain

$$\begin{aligned} \text{var}[e_t(l|\hat{\phi})] &\simeq \text{var}[e_t(l|\phi)] + \frac{\sigma_a^2}{1-\phi^2} E[l^2\phi^{2(l-1)}(\phi - \hat{\phi})^2] \\ &= \text{var}[e_t(l|\phi)] + \frac{l^2\phi^{2(l-1)}}{n}\sigma_a^2 \end{aligned}$$

and the discrepancy is again of order n^{-1}.

General-Order Autoregressive Processes. Related approximation results for the effect of parameter estimation errors on forecast error variances have been given by Yamamoto (1976) for the general AR(p) model. In particular, we briefly consider the approximation for one-step-ahead forecasts in the AR(p) case. Write the model at time $t+1$ as

$$\tilde{z}_{t+1} = \phi_1\tilde{z}_t + \phi_2\tilde{z}_{t-1} + \cdots + \phi_p\tilde{z}_{t+1-p} + a_{t+1} = \tilde{\mathbf{z}}_t'\boldsymbol{\phi} + a_{t+1}$$

where $\tilde{\mathbf{z}}_t' = (\tilde{z}_t, \tilde{z}_{t-1}, \ldots, \tilde{z}_{t+1-p})$ and $\boldsymbol{\phi}' = (\phi_1, \phi_2, \ldots, \phi_p)$. Then,

$$\hat{\tilde{z}}_t(1|\boldsymbol{\phi}) = \phi_1\tilde{z}_t + \phi_2\tilde{z}_{t-1} + \cdots + \phi_p\tilde{z}_{t+1-p} = \tilde{\mathbf{z}}_t'\boldsymbol{\phi}$$

and similarly, $\hat{\tilde{z}}_t(1|\hat{\boldsymbol{\phi}}) = \tilde{\mathbf{z}}_t'\hat{\boldsymbol{\phi}}$, where $\hat{\boldsymbol{\phi}}$ is the ML estimate of $\boldsymbol{\phi}$ based on n observations. Hence,

$$e_t(1|\hat{\boldsymbol{\phi}}) = e_t(1|\boldsymbol{\phi}) + \tilde{\mathbf{z}}_t'(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}) \tag{A7.6.8}$$

Using similar independence properties as above, as well as $\text{cov}[\tilde{\mathbf{z}}_t] = \boldsymbol{\Gamma}_p$ and the asymptotic distribution approximation for $\hat{\boldsymbol{\phi}}$ (see, e.g., [7.2.19] and [A7.4.23]) that $\text{cov}[\hat{\boldsymbol{\phi}}] \simeq n^{-1}\sigma_a^2\boldsymbol{\Gamma}_p^{-1}$,

it follows that

$$\begin{aligned} E[e_t^2(1|\hat{\boldsymbol{\phi}})] &= E[e_t^2(1|\boldsymbol{\phi})] + E[\{\tilde{\mathbf{z}}_t'(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})\}^2] \\ &= \sigma_a^2 + \text{tr}\{E[\tilde{\mathbf{z}}_t\tilde{\mathbf{z}}_t']E[(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})']\} \\ &= \sigma_a^2 + \text{tr}\{\boldsymbol{\Gamma}_p n^{-1}\sigma_a^2\boldsymbol{\Gamma}_p^{-1}\} \end{aligned}$$

Thus, the approximation for one-step-ahead forecast error variance,

$$\text{var}[e_t(1|\hat{\boldsymbol{\phi}})] \simeq \sigma_a^2\left(1 + \frac{p}{n}\right) \tag{A7.6.9}$$

is readily obtained for the AR model of order p.

APPENDIX A7.7 SPECIAL NOTE ON ESTIMATION OF MOVING AVERAGE PARAMETERS

If the least-squares iteration that involves moving average parameters is allowed to stray outside the invertibility region, parameter values can readily be found that apparently provide sums of squares smaller than the true minimum. However, these do not provide appropriate estimates and are quite meaningless. To illustrate, suppose that a series has been generated by the first-order moving average model $w_t = (1 - \theta B)a_t$ with $-1 < \theta < 1$. Then, the series could equally well have been generated by the corresponding backward process $w_t = (1 - \theta F)e_t$ with $\sigma_e^2 = \sigma_a^2$. Now, the latter process can also be written as $w_t = (1 - \theta^{-1}B)\alpha_t$, where now θ^{-1} is *outside* the invertibility region. However, in this representation $\sigma_\alpha^2 = \sigma_a^2\theta^2$ and is itself a function of θ. Therefore, a valid estimate of θ^{-1} will not be provided by minimizing $\sum_t \alpha_t^2 = \theta^2 \sum_t a_t^2$. Indeed, this has its minimum at $\theta^{-1} = \infty$.

The difficulty may be avoided:

1. By using as starting values rough preliminary estimates within the invertibility region obtained at the identification stage.
2. By checking that all moving average estimates, obtained after convergence has apparently occurred, lie within the invertibility region.

It is also possible to write least-squares programs such that estimates are constrained to lie within the invertibility region, and to check that moving average estimates lie within the invertibility region after each step of the iterative least-squares estimation procedure.

EXERCISES

7.1. The following table shows calculations for an (unrealistically short) series z_t for which the (0, 1, 1) model $w_t = \nabla z_t = (1 - \theta B)a_t$ is being considered with $\theta = -0.5$ and with an unknown starting value a_0.

t	z_t	$w_t = \nabla z_t$	$a_t = w_t - 0.5a_{t-1}$
0	40		a_0
1	42	2	$2 - 0.50a_0$
2	47	5	$4 + 0.25a_0$
3	47	0	$-2 - 0.13a_0$
4	52	5	$6 + 0.06a_0$
5	51	-1	$-4 - 0.03a_0$
6	57	6	$8 + 0.02a_0$
7	59	2	$-2 - 0.01a_0$

(a) Confirm the entries in the table.

(b) Show that the conditional sum of squares is

$$\sum_{t=1}^{7}(a_t| - 0.5, a_0 = 0)^2 = S_*(-0.5|0) = 144.00$$

7.2. Using the data in Exercise 7.1:

(a) Show (using least-squares) that the value $\hat{a}_0$ of a_0 that minimizes $S_*(-0.5|0)$ is

$$\hat{a}_0 = \frac{(2)(0.50) + (4)(-0.25) + \cdots + (-2)(0.0078)}{1^2 + 0.5^2 + \cdots + 0.0078^2} = \frac{-\sum_{t=0}^{n}\theta^t a_t^0}{\sum_{t=0}^{n}\theta^{2t}}$$

where $a_t^0 = (a_t|\theta, a_0 = 0)$ are the conditional values. Compare this expression for $\hat{a}_0$ with that for the exact back-forecast $[a_0]$ in the MA(1) model, where the expression for $[a_0]$ is given preceding the equation (A7.3.9) in Appendix A7.3, and verify that the two expressions are identical.

(b) By first writing this model in the backward form $w_t = (1 - \theta F)e_t$ and recursively computing the e's, show that the value of a_0 obtained in (a) is the same as that obtained by the back-forecasting method.

7.3. Using the value of $\hat{a}_0$ calculated in Exercise 7.2:

(a) Show that the unconditional sum of squares $S(-0.5)$ is 143.4.

(b) Show that for the (0, 1, 1) model, for large n,

$$S(\theta) = S_*(\theta|0) - \frac{\hat{a}_0^2}{1 - \theta^2}$$

7.4. For the process $w_t = \mu_w + (1 - \theta B)a_t$ show that for long series the variance–covariance matrix of the maximum likelihood estimates $\hat{\mu}_w, \hat{\theta}$ is approximately

$$n^{-1}\begin{bmatrix} (1-\theta)^2\sigma_a^2 & 0 \\ 0 & 1 - \theta^2 \end{bmatrix}$$

7.5. **(a)** Problems were experienced in obtaining a satisfactory fit to a series, the last 16 values of which were recorded as follows:

129, 135, 130, 130, 127, 126, 131, 152,
123, 124, 131, 132, 129, 127, 126, 124

Plot the series and suggest where the difficulty might lie.

(b) In fitting a model of the form $(1 - \phi_1 B - \phi_2 B^2) z_t = (1 - \theta B) a_t$ to a set of data, convergence was slow and the coefficient estimates in successive iterations oscillated wildly. Final estimates having large standard errors were obtained as follows: $\hat{\phi}_1 = 1.19$, $\hat{\phi}_2 = -0.34$, $\hat{\theta} = 0.52$. Can you suggest an explanation for the unstable behavior of the model? Why should preliminary identification have eliminated the problem?

(c) In fitting the model $\nabla^2 z_t = (1 - \theta_1 B - \theta_2 B^2) a_t$ convergence was not obtained. The last iteration yielded the values $\hat{\theta}_1 = 1.81$, $\hat{\theta}_2 = 0.52$. Can you explain the difficulty?

7.6. For the ARIMA(1, 1, 1) model $(1 - \phi B) w_t = (1 - \theta B) a_t$, where $w_t = \nabla z_t$:

(a) Write down the linearized form of the model.

(b) Set out how you would start off the calculation of the conditional nonlinear least-squares algorithm with start values $\phi = 0.5$ and $\theta = 0.4$ for a series whose first nine values are shown below.

t	z_t	t	z_t
0	149	5	150
1	145	6	147
2	152	7	142
3	144	8	146
4	150		

7.7. (a) Show that the second-order autoregressive model $\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + a_t$ may be written in orthogonal form as

$$\tilde{z}_t = \frac{\phi_1}{1 - \phi_2} \tilde{z}_{t-1} + \phi_2 \left(\tilde{z}_{t-2} - \frac{\phi_1}{1 - \phi_2} \tilde{z}_{t-1} \right) + a_t$$

suggesting that the approximate estimates

$$r_1 \quad \text{of} \quad \frac{\phi_1}{1 - \phi_2} \quad \text{and} \quad \hat{\phi}_2 = \frac{r_2 - r_1^2}{1 - r_1^2} \quad \text{of} \quad \phi_2$$

are uncorrelated for long series.

(b) Starting from the variance–covariance matrix of $\hat{\phi}_1$ and $\hat{\phi}_2$ or otherwise, show that the variance–covariance matrix of r_1 and $\hat{\phi}_2$ for long series is given approximately by

$$n^{-1} \begin{bmatrix} (1 - \phi_2^2)(1 - \rho_1^2) & 0 \\ 0 & 1 - \phi_2^2 \end{bmatrix}$$

7.8. The preliminary model identification performed in Chapter 6 suggested that either an ARIMA(1, 1, 0) or an ARIMA(0, 2, 2) model might be appropriate for the chemical process temperature readings in Series C. The series is available for download from http://pages.stat.wisc.edu/ reinsel/bjr-data/.

(a) Estimate the parameters of the ARIMA(1, 1, 0) for this series using R.

(b) Estimate the parameters of the ARIMA(0, 2, 2) model and compare the results with those in part (a).

7.9. Repeat the analysis in Exercise 7.8 by fitting (a) an AR(1) and (b) an ARMA(0, 1, 1) model to the chemical process viscosity readings in Series D.

7.10. Daily air quality measurements in New York, from May to September 1973, are available in a file called 'airquality' in the R datasets package. The file provides data on four air quality variables: mean ozone levels at Roosevelt Island, solar radiation at Central Park, maximum daily temperature at La Guardia Airport, and average wind speeds at La Guardia Airport.

(a) Identify suitable models for the daily temperature and wind speed series.

(b) Estimate the parameters of selected models and comment.

7.11. Consider the solar radiation series that is part of the New York airquality data file described in Problem 7.10. This series has a few missing values.

(a) Impute suitable estimates of the missing values. (Note: A formal procedure for estimating missing values is described in Chapter 13, but is not needed here).

(b) Identify a model for the resulting series.

(c) Estimate the parameters of selected model and comment.

7.12. Refer to the annual river flow measurements in the time series 'Nile' analyzed in Exercise 6.7. Estimate the parameters of the model or models identified for this time series and comment.

8

MODEL DIAGNOSTIC CHECKING

The model having been identified and the parameters estimated, *diagnostic checks* are then applied to the fitted model. One useful method of checking a model is to *overfit*, that is, to estimate the parameters in a model somewhat more general than that which we believe to be true. This method assumes that we can guess the direction in which the model is likely to be inadequate. Therefore, it is necessary to supplement this approach by less specific checks applied to the residuals from the fitted model. These allow the data themselves to suggest modifications to the model. In this chapter, we describe two such checks that employ (1) the autocorrelation function of the residuals and (2) the cumulative periodogram of the residuals. Some alternative diagnostic procedures are also discussed. Numerical examples are included to demonstrate the results.

8.1 CHECKING THE STOCHASTIC MODEL

8.1.1 General Philosophy

Suppose that using a particular time series, the model has been identified and the parameters estimated using the methods described in Chapters 6 and 7. The question remains of deciding whether this model is adequate. If there is evidence of serious inadequacy, we need to know how the model should be modified in the next iterative cycle. What we are doing is described only partially by the words ''testing goodness of fit.'' We need to discover *in what way* a model is inadequate, so as to suggest appropriate modification. To illustrate, by reference to familiar procedures outside time series analysis, the scrutiny of residuals for the analysis of variance, described by Anscombe (1961) and Anscombe and Tukey (1963), and the

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

criticism of factorial experiments, leading to normal plotting and other methods, described by Daniel (1959), would be called *diagnostic checks*.

All models are approximations and no model form can ever represent the truth absolutely. Given sufficient data, statistical tests can discredit models that could nevertheless be entirely adequate for the purpose at hand. Alternatively, tests can fail to indicate serious departures from assumptions because of small sample sizes or because these tests are insensitive to the types of discrepancies that occur. The best policy is to devise the most sensitive statistical procedures possible but be prepared to employ models that exhibit slight lack of fit. If diagnostic checks, which have been thoughtfully devised, are applied to a model fitted to a reasonably large body of data and fail to show serious discrepancies, then we should feel comfortable using that model.

8.1.2 Overfitting

One technique that can be used for diagnostic checking is *overfitting*. Having identified what is believed to be a correct model, we actually fit a more elaborate one. This puts the identified model in jeopardy because the more elaborate model contains additional parameters covering feared directions of discrepancy. Careful thought should be given to the question of how the model should be augmented. In particular, in accordance with the discussion on model redundancy in Section 7.3.5, it would not make sense to add factors *simultaneously* to both sides of the ARMA model. Moreover, if the analysis fails to show that the additions are needed, we have, of course, not proved that our model is correct. A model is only capable of being ''proved'' in the biblical sense of being put to the test. As was recommended by Saint Paul in his first epistle to the Thessalonians, what we can do is to ''Prove all things; hold fast to that which is good.''

Example of Overfitting. As an example, we consider again some IBM stock price data. For this analysis, data were employed that are listed as Series B′ in the Collection of Time Series in Part Five of this book. This series consists of IBM stock prices for the period[1] June 29, 1959–June 30, 1960. The (0, 1, 1) model

$$\nabla z_t = (1 - \theta B)a_t$$

with $\hat{\lambda}_0 = 1 - \hat{\theta} = 0.90$, was identified and fitted to the 255 available observations.

The (0, 1, 1) model can equally well be expressed in the form

$$\nabla z_t = \lambda_0 a_{t-1} + \nabla a_t$$

The extended model that was considered in the overfitting procedure was the (0, 3, 3) process

$$\nabla^3 z_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3)a_t$$

or using (4.3.21), in the form

$$\nabla^3 z_t = (\lambda_0 \nabla^2 + \lambda_1 \nabla + \lambda_2)a_{t-1} + \nabla^3 a_t$$

[1]The IBM stock data previously considered, referred to as Series B, cover a different period, May 17, 1961–November 2, 1962.

While this model may seem overly elaborate, the immediate motivation for extending the model in this particular way was to test a suggestion made by Brown (1962) that the series should be forecasted by an adaptive *quadratic* forecast function. Now, it was shown in Chapter 5 that an IMA(0, q, q) process has for its optimal forecasting function an adaptive polynomial of degree $q - 1$. Thus, for the extended (0, 3, 3) model above, the optimal lead l forecast function is the quadratic polynomial in l:

$$\hat{z}_t(l) = b_0^{(t)} + b_1^{(t)} l + b_2^{(t)} l^2$$

where the coefficients $b_0^{(t)}$, $b_1^{(t)}$, and $b_2^{(t)}$ are adjusted as each new piece of data becomes available.

By comparison, the model we have identified is an IMA(0, 1, 1) process, which yields a forecast function

$$\hat{z}_t(l) = b_0^{(t)} \tag{8.1.1}$$

This is a ''polynomial in l'' of degree zero. Hence, the model implies that the forecast $\hat{z}_t(l)$ is independent of l, that is, the forecast at any particular time t is the same for one step ahead, two steps ahead, and so on. In other words, the series contains information only on the future *level* of the series, and nothing about slope or curvature. At first sight, this is somewhat surprising because, using hindsight, quite definite linear and curvilinear trends appear to be present in the series. Therefore, it is worthwhile to check whether nonzero values of λ_1 and λ_2, which would produce predictable trends, actually occur. Sum-of-squares grids for $S(\lambda_1, \lambda_2 | \lambda_0)$ similar to those shown in Figure 7.2 were produced for $\lambda_0 = 0.7$, 0.9, and 1.1, which showed a minimum close to $\hat{\lambda}_0 = 0.9$, $\hat{\lambda}_1 = 0$, and $\hat{\lambda}_2 = 0$. It was clear that values of $\lambda_1 > 0$ and $\lambda_2 > 0$ lead to higher sum of squares, and do not support augmenting the identified IMA(0, 1, 1) model in these directions. This implies, in particular, that a quadratic forecast function would give worse instead of better forecasts than those obtained from (8.1.1), as was indeed shown to be the case in Section A5.3.3.

Computations in R. Estimation of the parameters in the more elaborate IMA(0, 3, 3) models for the IBM series using R also shows that the model can be simplified. The relevant commands along with a partial model output are provided below:

```
>library(astsa}
>ibm2=read.table("ibm2.txt",header=TRUE}
>ibm.ts=ts(ibm2)
>sarima(ibm.ts,0,3,3)

 Coefficients:
          ma1     ma2      ma3
      -2.0215  1.0686  -0.0469
s.e.   0.0705  0.1370   0.0692  sigma^2 estimated as 25.5

> polyroot(c(1,-2.0215,1.0686,-0.0469))
1.013484+0.005832i 1.013484-0.005832i 20.757680+0.000000i

>sarima(ibm.ts,0,1,1)

 Coefficients:
```

```
          ma1  constant
      -0.0848    0.3028
s.e.   0.0634    0.2878         sigma^2 estimated as 25.1
```

We note that the parameter estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ in the IMA(0, 3, 3) model are highly significant. However, the large estimates are introduced as compensation for overdifferencing by setting $d = 3$ in this model. This is confirmed by finding the roots of the moving average polynomial using the command polyroot() in R. The results, which are included above, show that two of the roots are very close to one. Hence, cancellation is possible, reducing the IMA(0, 3, 3) model to a IMA(0, 1, 1) model. The IMA(0, 1, 1) model also provides a slightly better fit to the data as can be seen from the smaller value of $\hat{\sigma}^2$ in the R output for this model.

8.2 DIAGNOSTIC CHECKS APPLIED TO RESIDUALS

The method of overfitting, by extending the model in a particular direction, assumes that we know what kind of discrepancies are to be feared. Procedures less dependent upon such knowledge are based on the analysis of *residuals*. It cannot be too strongly emphasized that *visual inspection of a plot of the residuals themselves* is an indispensable first step in the checking process.

8.2.1 Autocorrelation Check

Suppose that a model $\phi(B)\tilde{w}_t = \theta(B)a_t$ has been fitted to the observed time series with ML estimates $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ obtained for the parameters. The quantities

$$\hat{a}_t = \hat{\theta}^{-1}(B)\hat{\phi}(B)\tilde{w}_t \tag{8.2.1}$$

are then referred to as the *residuals*. The residuals are computed recursively from $\hat{\theta}(B)\hat{a}_t = \hat{\phi}(B)\tilde{w}_t$ as

$$\hat{a}_t = \tilde{w}_t - \sum_{j=1}^{p} \hat{\phi}_j \tilde{w}_{t-j} + \sum_{j=1}^{q} \hat{\theta}_j \hat{a}_{t-j} \qquad t = 1, 2, \ldots, n$$

using either zero initial values (conditional method) or back-forecasted initial values (exact method) for the initial $\hat{a}_t$'s and $\tilde{w}_t$'s. Now, it is possible to show that, if the model is adequate,

$$\hat{a}_t = a_t + O\left(\frac{1}{\sqrt{n}}\right)$$

As the series length increases, the $\hat{a}_t$'s become close to the white noise a_t's. Therefore, one might expect that study of the $\hat{a}_t$'s could indicate the existence and nature of model inadequacy. In particular, recognizable patterns in the estimated autocorrelation function of the $\hat{a}_t$'s could point to appropriate modifications in the model. This point is discussed further in Section 8.3.

Now, suppose that the form of the model was correct and that we *knew* the true parameter values $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Then, using (2.1.13) and a result of Anderson (1942), the estimated

autocorrelations $r_k(a)$, of the a_t's, would be uncorrelated and distributed approximately normally about zero with variance n^{-1}, and hence with a standard error of $n^{-1/2}$. We could use these facts to assess approximately the statistical significance of apparent departures of these autocorrelations from zero.

Now, in practice, we do not know the *true* parameter values. We have only the estimates $(\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$, from which, using (8.2.1), we can calculate not the a_t's but the $\hat{a}_t$'s. The autocorrelations $r_k(\hat{a})$ of the $\hat{a}_t$'s can yield valuable evidence concerning lack of fit and the possible nature of model inadequacy. However, it was pointed out by Durbin (1970) that it might be dangerous to assess the statistical significance of apparent discrepancies of these autocorrelations $r_k(\hat{a})$ from their theoretical zero values on the basis of a standard error $n^{-1/2}$, appropriate to the $r_k(a)$'s. Durbin was able to show, for example, that for the AR(1) process with parameter ϕ, the variance of $r_1(\hat{a})$ is $\phi^2 n^{-1}$, which can be substantially *smaller* than n^{-1}. The large-sample variances and covariances for all the autocorrelations of the $\hat{a}_t$'s from any ARMA process were subsequently derived by Box and Pierce (1970). They showed that while in all cases, a reduction in variance can occur for low lags, and that at these low lags the $r_k(\hat{a})$'s can be highly correlated, these effects usually disappear rather quickly at high lags. Thus, the use of $n^{-1/2}$ as the standard error for $r_k(\hat{a})$ would underestimate the statistical significance of apparent departures from zero of the autocorrelations at low lags but could usually be employed for moderate or high lags.

For illustration, the large-sample one- and two-standard-error limits of the residual autocorrelations $r_k(\hat{a})$'s, for two AR(1) processes and two AR(2) processes, are shown in Figure 8.1. These also supply the corresponding approximate standard errors for moving average processes with the same parameters as indicated in the figure. It is evident that, except at moderately high lags, $n^{-1/2}$ provides an upper bound for the standard errors of the $r_k(\hat{a})$'s rather than the standard errors themselves. If for low lags we use the standard

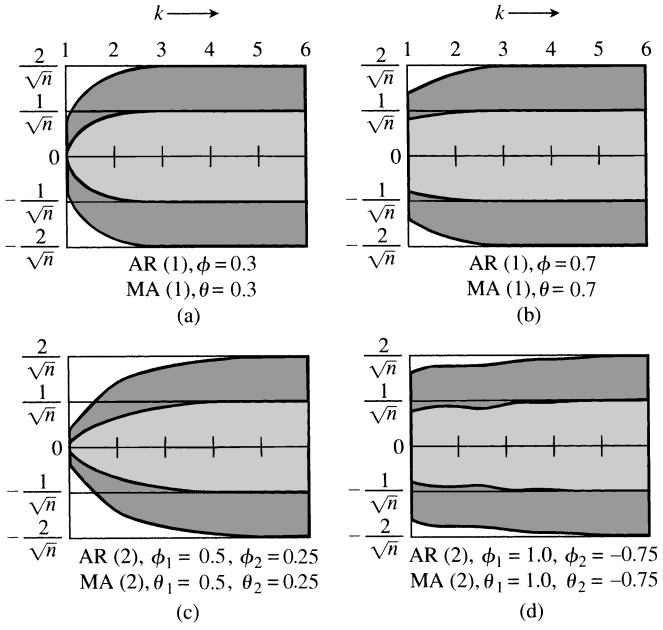


FIGURE 8.1 Standard-error limits for residual autocorrelations $r_k(\hat{a})$.

error $n^{-1/2}$ for the $r_k(\hat{a})$'s, we may seriously *underestimate* the significance of apparent discrepancies.

8.2.2 Portmanteau Lack-of-Fit Test

In addition to considering the $r_k(\hat{a})$'s individually, an indication is often needed of whether, say, the first 10–20 autocorrelations of the $\hat{a}_t$'s *taken as a whole* indicate inadequacy of the model. Suppose that we have the first K autocorrelations[2] $r_k(\hat{a})$ $(k = 1, 2, \ldots, K)$ from any ARIMA(p, d, q) model, then it is possible to show (Box and Pierce, 1970) that if the fitted model is appropriate,

$$Q = n \sum_{k=1}^{K} r_k^2(\hat{a}) \tag{8.2.2}$$

is approximately distributed as $\chi^2(K - p - q)$, where $n = N - d$ is the number of w's used to fit the model. On the other hand, if the model is inappropriate, the average values of Q will be inflated. Therefore, an approximate ''portmanteau'' test of the hypothesis of model adequacy, designed to take account of the difficulties discussed above, may be made by referring an observed value of Q to the percentage points of this χ^2 distribution.

However, Ljung and Box (1978) later showed that, for sample sizes common in practice, the chi-squared distribution may not provide an adequate approximation to the distribution of the statistic Q under the null hypothesis, with the values of Q tending to be somewhat smaller than what is expected under the chi-squared distribution. Empirical evidence to support this was also presented by Davies et al. (1977). Ljung and Box (1978) proposed a modified form of the statistic,

$$\tilde{Q} = n(n + 2) \sum_{k=1}^{K} (n - k)^{-1} r_k^2(\hat{a}) \tag{8.2.3}$$

such that the modified statistic has, approximately, the mean $E[\tilde{Q}] \approx K - p - q$ of the $\chi^2(K - p - q)$ distribution. The motivation for (8.2.3) is that a more accurate value for the variance of $r_k(a)$ from a white noise series is $(n - k)/n(n + 2)$, rather than $1/n$ used in (8.2.2). This modified form of the portmanteau test statistic has been recommended for use as having a null distribution that is much closer to the $\chi^2(K - p - q)$ distribution for typical sample sizes n. Because of its computationally convenient form, this statistics has been implemented in many software packages and has become widely used in applied work. We emphasize, however, that this statistic should not be used as a substitute for careful examination of the residuals and their individual autocorrelation coefficients, and for other diagnostic checks on the fitted model.

Remark. Diagnostic checks based on the residuals and their autocorrelation coefficients are conveniently performed using R. Having fitted a model `m1` to the observed series, the command `tsdiag(m1$residuals, gof.lag=20)` provides a plot of the standardized residuals, a plot of the first 20 residual autocorrelation coefficients, and a plot of the p-values for the

[2]It is assumed here that K is taken sufficiently large so that the weights ψ_j in the model, written in the form $\tilde{w}_t = \phi^{-1}(B)\theta(B)a_t = \psi(B)a_t$ will be negligibly small after $j = K$.

portmanteau statistic $\tilde{Q}$ for increasing values of K. However, while these diagnostics are useful, it appears that the command tsdiag(), at present, determines p-values for $\tilde{Q}$ using a chi-square distribution with K rather than $K - p - q$ degrees of freedom. An alternative is to use diagnostic tools in the R package astsa, where this problem does not appear. An illustration of the use of this package is provided below.

An Empirical Example. In Chapter 7, we examined two potential models for a time series of chemical temperature readings referred to as Series C. The two models were (1) the IMA(0, 2, 2) model $\nabla^2 z_t = (1 - 0.13B - 0.12B^2)a_t$ and (2) the ARIMA(1, 1, 0) model $(1 - 0.82B)\nabla z_t = a_t$. It was decided that the second model gave a preferable representation of the series. Model diagnostics for the IMA(0, 2, 2) model generated using R are provided in Figure 8.2. These include graphs of the standardized residuals, the residual autocorrelation coefficients $r(\hat{a}_k)$, for lags $k = 1, \ldots, 25$, a normal Q–Q plot of the standardized residuals, and a plot of the p-values for the portmanteau statistic $\tilde{Q}$ in (8.2.3) determined for increasing values of K. The graph of the standardized residuals reveals some large residuals around $t = 60$, but apart from that there are no issues. The Q–Q plot confirms the presence of three large residuals but indicates that the normal approximation is adequate otherwise.

Approximate two-standard-error upper bounds on the residual autocorrelation coefficients are included in the graph of the autocorrelation function. Since there are $n = 224$ observations after differencing the series, the approximate upper bound for the standard

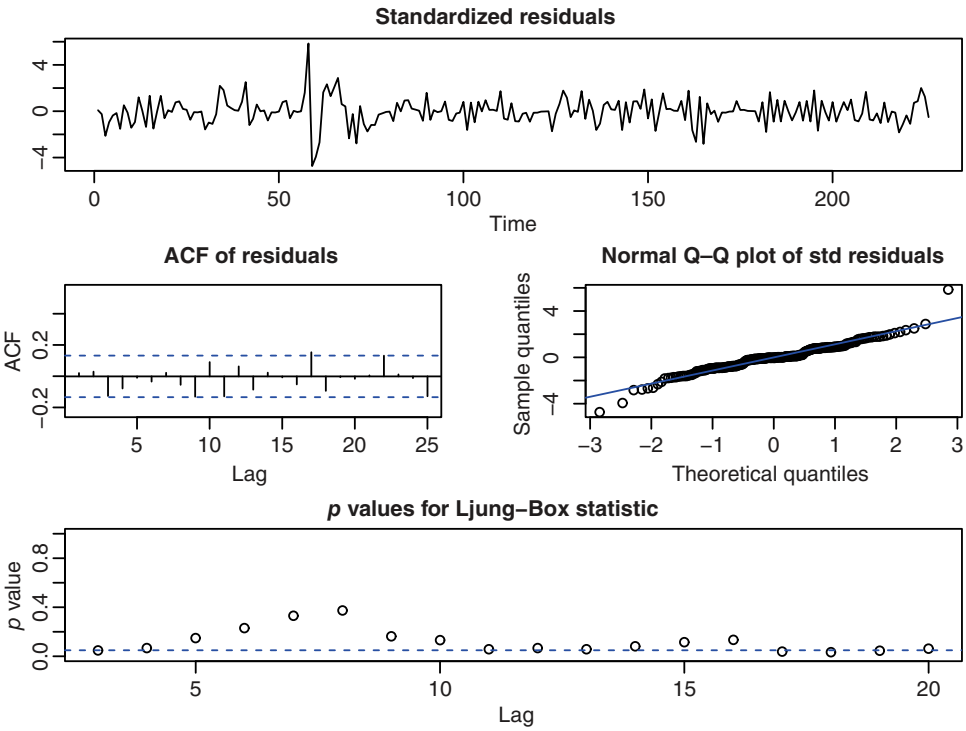


FIGURE 8.2 Model diagnostics for the ARIMA(0, 2, 2) model fitted to the temperature readings in Series C.

error of a single autocorrelation is $1/\sqrt{224} \approx 0.07$. While most of the individual autocorrelations fall within the two-standard-error bounds, several values including $r_3(\hat{a})$, $r_9(\hat{a})$, $r_{11}(\hat{a})$, $r_{17}(\hat{a})$, $r_{22}(\hat{a})$, and $r_{25}(\hat{a})$ are close to these bounds. Of course, occasional large deviations occur even in random series, but taking these results as a whole, there is a suspicion of some lack of fit. This is confirmed by examining the p-values of the portmanteau statistic shown in the bottom graph of Figure 8.2. We note that most of the p-values are at or near the 5% level indicating some lack of fit. This is especially the case for the larger values of K, where the chi-squared distribution is expected to provide a valid approximation.

Model diagnostics for the ARIMA(1, 1, 0) model $(1 - 0.82B)\nabla z_t = a_t$ fitted to the same time series are displayed in Figure 8.3. The graph of the residual autocorrelation function shows fewer large values for this model. This is also reflected in the p-values of the portmanteau statistic shown at the bottom of the graph. These diagnostic checks show a clear improvement over the IMA(0, 2, 2) model examined in Figure 8.2. The graph of the standardized residuals and the normal Q–Q plot reveal that outliers are still present, however. Methods for outlier detection and adjustments will be discussed in Section 13.2, where the ARIMA(1, 1, 0) model for Series C is refitted allowing the outliers at $t = 58$, 59, and 60. Allowing these outliers in the parameter estimation changes the estimate $\hat{\phi}$ only slightly from 0.82 to 0.85. However, a larger change occurs in the estimate of the residual variance, which is reduced by about 26% when the outliers are accounted for in the model.

Before proceeding, we note that Figures 8.2 and 8.3 can be reproduced in R using the following commands:

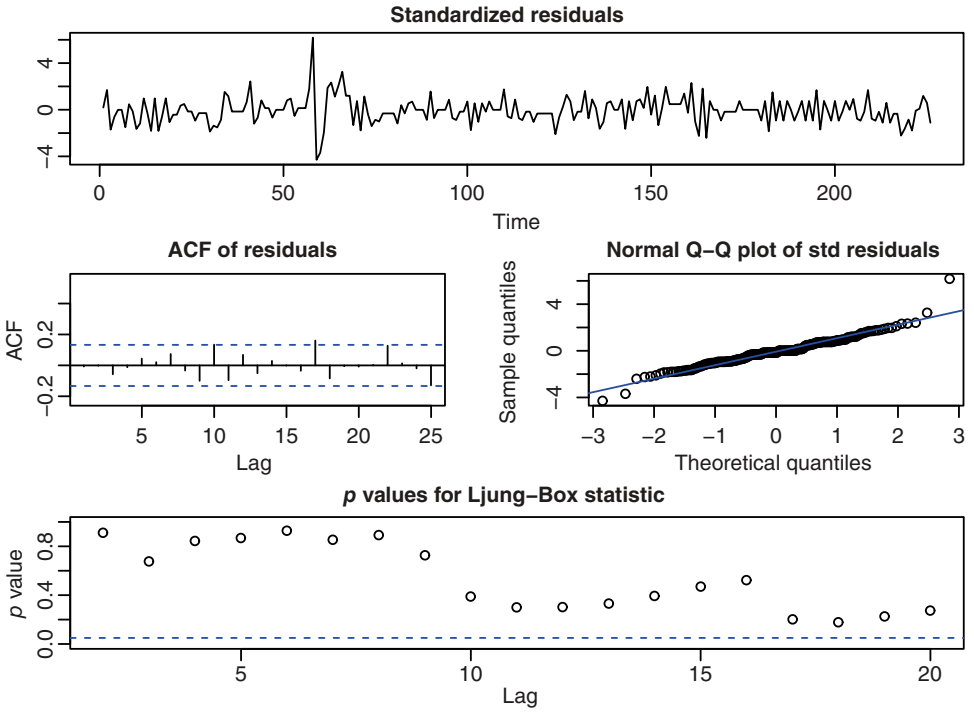


FIGURE 8.3 Model diagnostics for the ARIMA(1, 1, 0) model fitted to the temperature readings in Series C.

```
>library(astsa)
>seriesC=read.table("seriesC.txt",header=T)
>sarima(seriesC,0,2,2,no.constant=TRUE) % Figure 8.2
>sarima(seriesC,1,1,0,no.constant=TRUE) % Figure 8.3
```

Portmanteau Tests for Series A–F. Table 8.1 summarizes the values of the criterion $\tilde{Q}$ in (8.2.3) based on $K = 25$ residual autocorrelations for the models fitted to Series A–F in Table 7.11. However, in regards to the choice of K, a somewhat smaller value would be recommended for use in practice, especially for shorter series such as Series E and F, since the asymptotic theory involved in the distribution of the statistic $\tilde{Q}$ relies on K growing (but only slowly, such that $K/n \to 0$) as the series length n increases. In addition, as noted by Ljung (1986), smaller values of K also have advantages in terms of increased power. This is particularly true for nonseasonal series, where the lack of fit is expected to be most evident in residual autocorrelations at the first few lags.

Inspection of Table 8.1 shows that only two suspiciously large values of $\tilde{Q}$ occur. One is the value $\tilde{Q} = 36.2$ obtained after fitting the IMA(0, 2, 2) model to Series C, which we have discussed already. The other is the value $\tilde{Q} = 38.8$ obtained after fitting an IMA(0, 1, 1) model to Series B. This suggests some model inadequacy since the 5 and 2.5% points for χ^2 with 24 degrees of freedom are 36.4 and 39.3, respectively. The nature of possible model inadequacy for Series B will be examined further in Section 8.2.3.

Other Portmanteau Statistics to Test Model Adequacy. Instead of a portmanteau statistic based on residual autocorrelations, as in (8.2.3), one could alternatively consider a test for model adequacy based on residual *partial* autocorrelations. If the model fitted is adequate, the associated error process a_t is white noise and one should expect the residual partial autocorrelation at any lag k, which we denote as $\hat{\phi}_{kk}(\hat{a})$, not to be significantly different from zero. Therefore, a test for model adequacy can be based on the statistic

$$Q^* = n(n+2)\sum_{k=1}^{K}(n-k)^{-1}\hat{\phi}_{kk}^2(\hat{a}) \tag{8.2.4}$$

TABLE 8.1 Summary of Results of Portmanteau Test Applied to Residuals of Various Models Fitted to Series A–F

Series	$n = N - d$	Fitted Model	$\tilde{Q}$	Degrees of Freedom
A	197	$z_t - 0.92z_{t-1} = 1.45 + a_t - 0.58a_{t-1}$	28.4	23
	196	$\nabla z_t = a_t - 0.70a_{t-1}$	31.9	24
B	368	$\nabla z_t = a_t + 0.09a_{t-1}$	38.8	24
C	225	$\nabla z_t - 0.82\nabla z_{t-1} = a_t$	31.3	24
	224	$\nabla^2 z_t = a_t - 0.13a_{t-1} - 0.12a_{t-2}$	36.2	23
D	310	$z_t - 0.87z_{t-1} = 1.17 + a_t$	11.5	24
	309	$\nabla z_t = a_t - 0.06a_{t-1}$	18.8	24
E	100	$z_t - 1.42z_{t-1} + 0.73z_{t-2} = 14.35 + a_t$	26.8	23
	100	$z_t - 1.57z_{t-1} + 1.02z_{t-2} - 0.21z_{t-3} = 11.31 + a_t$	20.0	22
F	70	$z_t + 0.34z_{t-1} - 0.19z_{t-2} = 58.87 + a_t$	14.7	23

Under the hypothesis of model adequacy, Monti (1994) argued that the statistic Q^* in (8.2.4) is asymptotically distributed as $\chi^2(K-p-q)$, analogous to the asymptotic distribution of the statistic $\tilde{Q}$ in (8.2.3). Hence, a test of model adequacy can be based on referring the value of Q^* to the upper critical value determined from this distribution. The test based on Q^* has been found to be typically at least as powerful as $\tilde{Q}$ in detecting departures from model adequacy, and it seems to be particularly sensitive when the alternative model includes a higher order moving average term. In practice, since residual partial autocorrelations are routinely available, we could consider using both the statistic $\tilde{Q}$ in (8.2.3) and Q^* in (8.2.4) simultaneously in standard model checking procedures.

Another portmanteau goodness-of-fit test statistic based on a general measure of multivariate dependence was proposed by Peña and Rodríguez (2002). Denote the correlation matrix up to order (lag) K of the residuals $\hat{a}_t$ from the fitted ARIMA(p,d,q) model by

$$\hat{\mathbf{P}}_K(\hat{a}) = \begin{bmatrix} 1 & r_1(\hat{a}) & r_2(\hat{a}) & \dots & r_K(\hat{a}) \\ r_1(\hat{a}) & 1 & r_1(\hat{a}) & \dots & r_{K-1}(\hat{a}) \\ r_2(\hat{a}) & r_1(\hat{a}) & 1 & \dots & r_{K-2}(\hat{a}) \\ \vdots & \vdots & \vdots & \dots & \vdots \\ r_K(\hat{a}) & r_{K-1}(\hat{a}) & r_{K-2}(\hat{a}) & \dots & 1 \end{bmatrix}$$

The proposed statistic is based on the determinant of this correlation matrix, a general measure of dependence in multivariate analysis, and is given by

$$\hat{D}_K = n(1 - |\hat{\mathbf{P}}_K(\hat{a})|^{1/K}) \tag{8.2.5}$$

An alternate interpretation for the statistic is obtained from the following relation given by Peña and Rodríquez (2002)

$$|\hat{\mathbf{P}}_K(\hat{a})|^{1/K} = \prod_{k=1}^{K} [1 - \hat{\phi}_{kk}^2(\hat{a})]^{(K+1-k)/K}$$

where the $\hat{\phi}_{kk}(\hat{a})$ are the residual partial autocorrelations as in (8.2.4). This expression shows that $|\hat{\mathbf{P}}_K(\hat{a})|^{1/K}$ is also a weighted function of the first K partial autocorrelations of the residuals. However, in comparison to the statistics (8.2.3) and (8.2.4), relatively more weight is given to the lower lag residual correlations in the statistic (8.2.5). The asymptotic distribution of $\hat{D}_K$ is shown to be a linear combination of K-independent $\chi^2(1)$ random variates, which can be approximated by a gamma distribution (see Peña and Rodríguez, 2002). The authors also proposed and recommended a modification of the statistic $\hat{D}_K$, here denoted as $\tilde{D}_K$, in which the residual autocorrelations $r_k(\hat{a})$ used to form $\hat{\mathbf{P}}_K(\hat{a})$ are replaced by the modified values $\sqrt{(n+2)/(n-k)}r_k(\hat{a})$, similar to the modifications used in the $\tilde{Q}$ and Q^* statistics. Simulation evidence indicates that the statistic $\tilde{D}_K$ may provide considerable increase in power over the statistics $\tilde{Q}$ and Q^* in many cases, due to its greater sensitivity to the lower lag residual correlations. Application of this procedure to detection of several types of nonlinearity, by using sample autocorrelations of squared residuals $\hat{a}_t^2$, was also explored in Peña and Rodríguez (2002). (For discussion of nonlinearities, see Sections 10.2 and 10.3).

Peña and Rodríguez (2006) proposed a modification of their earlier test that has the same asymptotic distribution as $\hat{D}_K$ but better performance in finite sam-

ples. The modified test statistics has the form $D_K^* = -n\sum_{k=1}^{K} w_k \ln[1 - \hat{\phi}_{kk}^2(\hat{a})]$, where $w_k = (K + 1 - k)/(K + 1)$. The statistic is thus proportional to a weighted average of the squared partial autocorrelation coefficients with larger weights given to low-order coefficients and smaller weights to high-order coefficients. The authors considered two approximations to the asymptotic distribution of this statistic, and demonstrated using simulation that the test performs well. Several other authors have extended the work of Peña and Rodríguez (2002) and proposed portmanteau statistics that are asymptotically similar to their statistics; for a discussion and references, see Fisher and Gallagher (2012). See also Li (2004) for a more detailed discussion of diagnostic testing.

8.2.3 Model Inadequacy Arising from Changes in Parameter Values

Another form of model inadequacy occurs when the *form* of the model remains the same but the parameters change over a prolonged period of time. In fact, it appears that this can explain the possible inadequacy of the (0, 1, 1) model fitted to the IBM data.

Table 8.2 shows the results obtained by fitting (0, 1, 1) models separately to the first and second halves of Series B as well as to the complete series. Denoting the estimates of $\lambda = 1 - \theta$ obtained from the two halves by $\hat{\lambda}_1$ and $\hat{\lambda}_2$, we find that the standard error of $\hat{\lambda}_1 - \hat{\lambda}_2$ is $\sqrt{(0.070)^2 + (0.074)^2} = 0.102$. Since the difference $\hat{\lambda}_1 - \hat{\lambda}_2 = 0.26$ is 2.6 times its standard error, it is likely that a real change in λ has occurred. Inspection of the $\tilde{Q}$ values suggests that the (0, 1, 1) model, with parameters appropriately modified for different time periods, might explain the series more exactly. The estimation results for the residual variances $\hat{\sigma}_a^2$ also strongly indicate that a real *change in variability* has occurred between the two halves of the series.

This is confirmed by Figure 8.4 that shows the standardized residuals and other model diagnostics for the IMA(0, 1, 1) model fitted to Series B. An increase in the standardized residuals around time $t = 236$ indicates a change in the characteristics of the series around that time. In fact, fitting the IMA(0, 1, 1) model separately to the first 235 observations and to the remaining 134 observations yields the estimates $\hat{\theta}_1 = -0.26, \hat{\sigma}_{a_1}^2 = 24.55$, and $\hat{\theta}_2 = -0.02, \hat{\sigma}_{a_2}^2 = 99.49$, respectively. Hence, a substantial increase in variability during the latter portion of the series is clearly indicated. Additional approaches to explain and account for inadequacy in the overall IMA(0, 1, 1) model for Series B, which include allowance for conditional heteroscedasticity in the noise, nonlinearity, and mixture transition distributions, have been discussed by Tong (1990) and Le et al. (1996), among others. Some of these modeling approaches will be surveyed in general in Chapter 10.

TABLE 8.2 Comparison of IMA(0, 1, 1) Models Fitted to First and Second Halves of Series B

	n	$\hat{\theta}$	$\hat{\lambda} = 1 - \hat{\theta}$	$\hat{\sigma}(\hat{\lambda}) = [\frac{\hat{\lambda}(2-\hat{\lambda})}{n}]^{1/2}$	Residual Variance $\hat{\sigma}_a^2$	$\tilde{Q}$	Degrees of Freedom
First half	184	−0.29	1.29	±0.070	26.3	24.6	24
Second half	183	−0.03	1.03	±0.074	77.3	37.1	24
Complete	368	−0.09	1.09	±0.052	52.2	38.8	24

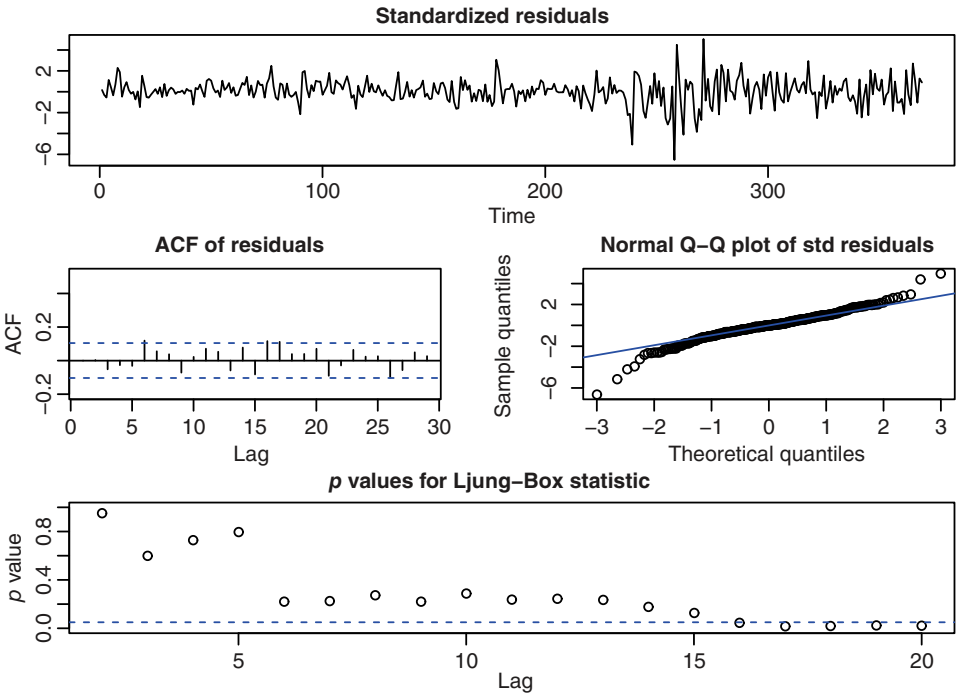


FIGURE 8.4 Model diagnostics for the IMA(0, 1, 1) model fitted to the IBM daily closing stock prices in Series B.

8.2.4 Score Tests for Model Checking

An alternative to the direct use of overfitting in model checking is provided by the Lagrange multiplier or score test procedure, which is also closely related to the portmanteau test procedure. The general score test procedure was presented by Silvey (1959), and its use in diagnostic checking for ARIMA models was discussed initially by Godfrey (1979) and Poskitt and Tremayne (1980). A computational advantage of the score test procedure is that it requires maximum likelihood estimation of parameters only under the null model under test, but it yields tests asymptotically equivalent to the corresponding likelihood ratio tests obtained by directly overfitting the model. Furthermore, the score test statistic is easily computed in the form of the sample size n times a coefficient of determination from a particular ''auxiliary'' regression.

Hence, we assume that an ARMA(p, q) model has been fitted by the maximum likelihood method to the observations $\tilde{w}_t$, and we want to assess the adequacy of the model by testing this null model against the alternative of an ARMA$(p + r, q)$ model or of an ARMA$(p, q + r)$ model. That is, for the ARMA$(p + r, q)$ alternative, we test H_0: $\phi_{p+1} = \cdots = \phi_{p+r} = 0$, while for the ARMA$(p, q + r)$ alternative, we test H_0: $\theta_{q+1} = \cdots = \theta_{q+r} = 0$. The score test procedure is based on the first partial derivatives, or scores, of the log-likelihood function with respect to the model parameters of the alternative model, but evaluated at the ML estimates obtained under the null model. The log-likelihood function is essentially given by $l = -(n/2)\ln(\sigma_a^2) - (\frac{1}{2}\sigma_a^{-2})\sum_{t=1}^{n} a_t^2$. So, the partial derivatives of l with respect

to the parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$ are

$$\frac{\partial l}{\partial \phi_j} = -\frac{1}{\sigma_a^2} \sum_{t=1}^{n} \frac{\partial a_t}{\partial \phi_j} a_t$$
$$\frac{\partial l}{\partial \theta_j} = -\frac{1}{\sigma_a^2} \sum_{t=1}^{n} \frac{\partial a_t}{\partial \theta_j} a_t$$

As in (7.2.9) and (7.2.10), we have

$$-\frac{\partial a_t}{\partial \phi_j} = u_{t-j} \qquad -\frac{\partial a_t}{\partial \theta_j} = v_{t-j}$$

where $u_t = \theta^{-1}(B)\tilde{w}_t = \phi^{-1}(B)a_t$, and $v_t = -\theta^{-1}(B)a_t$. Given residuals $\hat{a}_t$, obtained from ML fitting of the null model, as

$$\hat{a}_t = \tilde{w}_t - \sum_{j=1}^{p} \hat{\phi}_j \tilde{w}_{t-j} + \sum_{j=1}^{q} \hat{\theta}_j \hat{a}_{t-j} \qquad t = 1, 2, \ldots, n$$

the u_t's and v_t's evaluated under the ML estimates of the null model can be calculated recursively, starting with initial values set equal to zero, for example, as

$$u_t = \tilde{w}_t + \hat{\theta}_1 u_{t-1} + \cdots + \hat{\theta}_q u_{t-q}$$
$$v_t = -\hat{a}_t + \hat{\theta}_1 v_{t-1} + \cdots + \hat{\theta}_q v_{t-q}$$

The score vector of first partial derivatives with respect to all the model parameters $\boldsymbol{\beta}$ can be expressed as

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma_a^2} \mathbf{X}'\mathbf{a} \tag{8.2.6}$$

where $\mathbf{a} = (a_1, \ldots, a_n)'$ and $\mathbf{X}$ denotes the $n \times (p + q + r)$ matrix whose tth row consists of $(u_{t-1}, \ldots, u_{t-p-r}, v_{t-1}, \ldots, v_{t-q})$ in the case of the ARMA$(p + r, q)$ alternative model and $(u_{t-1}, \ldots, u_{t-p}, v_{t-1}, \ldots, v_{t-q-r})$ in the case of the ARMA$(p, q + r)$ alternative model. Then, similar to (7.2.17), since the large-sample information matrix for $\boldsymbol{\beta}$ can be consistently estimated by $\hat{\sigma}_a^{-2}\mathbf{X}'\mathbf{X}$, where $\hat{\sigma}_a^2 = n^{-1}\sum_{t=1}^{n} \hat{a}_t^2 = n^{-1}\hat{\mathbf{a}}'\hat{\mathbf{a}}$, it follows that the score test statistic for testing that the additional r parameters are equal to zero is

$$\Lambda = \frac{\hat{\mathbf{a}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{a}}}{\hat{\sigma}_a^2} \tag{8.2.7}$$

Godfrey (1979) noted that the computation of the test statistic in (8.2.7) can be given the interpretation as being equal to n times the coefficient of determination in an auxiliary regression equation. That is, if the alternative model is ARMA$(p + r, q)$, we consider the auxiliary regression equation

$$\hat{a}_t = \alpha_1 u_{t-1} + \cdots + \alpha_{p+r} u_{t-p-r} + \beta_1 v_{t-1} + \cdots + \beta_q v_{t-q} + \varepsilon_t$$

while if the alternative model is ARMA$(p, q + r)$, we consider the regression equation

$$\hat{a}_t = \alpha_1 u_{t-1} + \cdots + \alpha_p u_{t-p} + \beta_1 v_{t-1} + \cdots + \beta_{q+r} v_{t-q-r} + \varepsilon_t$$

Let $\hat{\varepsilon}_t$ denote the residuals from the ordinary least-squares estimation of this regression equation. Then from (8.2.7), it is seen that Λ can be expressed, essentially, as

$$\Lambda = \frac{n(\sum_{t=1}^{n} \hat{a}_t^2 - \sum_{t=1}^{n} \hat{\varepsilon}_t^2)}{\sum_{t=1}^{n} \hat{a}_t^2} = n\left(1 - \frac{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}{\sum_{t=1}^{n} \hat{a}_t^2}\right)$$

which is n times the coefficient of determination of the regression of the $\hat{a}_t$'s on the u_{t-j}'s and the v_{t-j}'s. Under the null hypothesis that the fitted ARMA(p, q) model is correct, the statistic Λ has an asymptotic χ^2 distribution with r degrees of freedom, and the null model is rejected as inadequate for large values of Λ.

As argued by Godfrey (1979) and others, rejection of the null model by the score test procedure should not be taken as evidence to adopt the specific alternative model involved, but simply as evidence against the adequacy of the fitted model. Similarly, the score test is expected to have reasonable power even when the alternative model is not correctly specified. Poskitt and Tremayne (1980) showed, for example, that the score test against an ARMA$(p + r, q)$ model alternative is asymptotically identical to a test against an ARMA$(p, q + r)$ alternative. Hence, the score test procedure may not be sensitive to the particular model specified under the alternative, but its performance will, of course, depend on the choice of the number r of additional parameters specified.

We also note an alternative form for the score statistic Λ. By the ML estimation procedure, it follows that the first partial derivatives, $\partial l/\partial \phi_j$, $j = 1, \ldots, p$, and $\partial l/\partial \theta_j$, $j = 1, \ldots, q$, will be identically equal to zero when evaluated at the ML estimates. Hence, the score vector, $\partial l/\partial \boldsymbol{\beta}$, will contain only r nonzero elements when evaluated at the ML estimates from the null model, these being the partial derivatives with respect to the additional r parameters of the alternative model. Thus, the score statistic in (8.2.7) can also be viewed as a quadratic form in these r nonzero values, whose matrix in the quadratic form is a consistent estimate of the inverse of the covariance matrix of these r score values when evaluated at the ML estimates obtained under the null model. Since these r score values are asymptotically normal with zero means under the null model, the validity of the asymptotic $\chi^2(r)$ distribution under the null hypothesis is easily seen.

Newbold (1980) noted that a score test against the alternative of r additional parameters is closely related to an appropriate test statistic based on the first r residual autocorrelations $r_k(\hat{a})$ from the fitted model. The test statistic is essentially a quadratic form in these first r residual autocorrelations, but of a more complex form than the portmanteau statistic in (8.2.2). As a direct illustration, suppose that the fitted or null model is a pure AR(p) model, and the alternative is an ARMA(p, r) model. Then, it follows from above that the variables v_{t-j} are identical to $-\hat{a}_{t-j}$, since $\theta(B) \equiv 1$ under the null model. Hence, the nonzero elements of the score vector in (8.2.6) are equal to $-n$ times the first r residual autocorrelations, $r_1(\hat{a}), \ldots, r_r(\hat{a})$ from the fitted model, and the score test is thus directly seen to be a quadratic form in these first r residual autocorrelations.

8.2.5 Cumulative Periodogram Check

In some situations, particularly in the fitting of seasonal time series, which are discussed in Chapter 9, it may be feared that we have not adequately taken into account the *periodic* characteristics of the series. Therefore, we are on the lookout for periodicities in the residuals. The autocorrelation function will not be a sensitive indicator of such departures from randomness because periodic effects will typically dilute themselves among several

autocorrelations. The periodogram, on the other hand, is specifically designed for the detection of periodic patterns in a background of white noise.

The periodogram of a time series a_t, $t = 1, 2, \ldots, n$, as defined in Section 2.2.1, is

$$I(f_i) = \frac{2}{n}\left[\left(\sum_{t=1}^{n} a_t \cos(2\pi f_i t)\right)^2 + \left(\sum_{t=1}^{n} a_t \sin(2\pi f_i t)\right)^2\right] \tag{8.2.8}$$

where $f_i = i/n$ is the frequency. Thus, it is a device for correlating the a_t's with sine and cosine waves of different frequencies. A pattern with given frequency f_i in the residuals is reinforced when correlated with a sine or cosine wave at that same frequency, and so produces a large value of $I(f_i)$.

Cumulative Periodogram. Bartlett (1955) and other authors have shown that the *cumulative periodogram* provides an effective means for the detection of periodic nonrandomness.

The power spectrum $p(f)$ for white noise has a constant value $2\sigma_a^2$ over the frequency domain 0–0.5 cycle. Consequently, the cumulative spectrum for white noise

$$P(f) = \int_0^f p(g)\, dg \tag{8.2.9}$$

plotted against f is a straight-line running from (0, 0) to $(0.5, \sigma_a^2)$, that is, $P(f)/\sigma_a^2$ is a straight-line running from (0, 0) to (0.5, 1).

The periodogram $I(f)$ provides an estimate of the power spectrum at frequency f. In fact, for white noise, $E[I(f)] = 2\sigma_a^2$, and hence the estimate is unbiased. It follows that $(1/n)\sum_{i=1}^{j} I(f_i)$ provides an unbiased estimate of the integrated spectrum $P(f_j)$, and

$$C(f_j) = \frac{\sum_{i=1}^{j} I(f_i)}{ns^2} \tag{8.2.10}$$

an estimate of $P(f_j)/\sigma_a^2$, where s^2 is an estimate of σ_a^2. We will refer to $C(f_j)$ as the *normalized cumulative periodogram.*

Now, if the model was adequate and the parameters known *exactly*, the a_t's could be computed from the data and would yield a white noise series. For a white noise series, the plot of $C(f_j)$ against f_j would be scattered about a straight-line joining the points (0, 0) and (0.5, 1). On the other hand, model inadequacies would produce nonrandom a_t's, whose cumulative periodogram could show systematic deviations from this line. In particular, periodicities in the a_t's would tend to produce a series of neighboring values of $I(f_j)$ that were large. These large ordinates would reinforce each other in $C(f_j)$ and form a bump on the expected straight line.

In practice, we do not know the exact values of the parameters, but only their estimated values. Hence, we do not have the a_t's, but only the estimated residuals $\hat{a}_t$'s. However, for large samples, the periodogram for the $\hat{a}_t$'s will have similar properties to that for the a_t's. Thus, careful inspection of the periodogram of the $\hat{a}_t$'s can provide a useful additional diagnostic check, particularly for indicating periodicities taken account of inadequately.

Example: Series C. We have seen that Series C is well fitted by the (1, 1, 0) model:

$$(1 - 0.82B)\nabla z_t = a_t$$

and somewhat less well by the IMA(0, 2, 2) model:

$$\nabla^2 z_t = (1 - 0.13B - 0.12B^2)a_t$$

which is rather similar to it. We illustrate the cumulative periodogram test by showing what happens when we analyze the residual a's after fitting to the series an inadequate IMA(0, 1, 1) model:

$$\nabla z_t = (1 - \theta B)a_t$$

where the least squares estimate of θ is found to be -0.65. The normalized cumulative periodogram plot of the residuals from this model is shown in Figure 8.5(a). We see immediately that there are marked departures from linearity in the cumulative periodogram. These departures are very pronounced at low frequencies, as might be expected, for example, if the degree of differencing is insufficient. Figure 8.5(b) shows the corresponding plot for the best-fitting IMA(0, 2, 2) model. The points of the cumulative periodogram now cluster more closely about the expected line, although, as we have seen in Table 8.1 and Figure 8.2, other evidence points to the inadequacy of this model.

It is wise to indicate on the diagram the period as well as the frequency. This makes for easy identification of the bumps that occur when residuals contain periodicities. For example, in monthly sales data, bumps near periods 12, 24, 36, and so on might indicate that seasonal effects were accounted for inadequately.

The probability relationship between the cumulative periodogram and the integrated spectrum is precisely the same as that between the empirical cumulative frequency function and the cumulative distribution function. For this reason we can assess deviations of the periodogram from that expected if the $\hat{a}_t$'s were white noise, by use of the Kolmogorov–Smirnov test. Using this test, we can place limit lines about the theoretical line. The limit lines are such that if the $\hat{a}_t$ series were white noise, the cumulative periodogram would deviate from the straight line sufficiently to cross these limits only with the stated probability. Now, because the $\hat{a}_t$'s are fitted values and not the true $\hat{a}_t$'s, we know that even when the model is correct, they will not precisely follow a white noise process. Thus, as a test for model inadequacy, application of the Kolmogorov–Smirnov limits will indicate only approximate probabilities. However, it is worthwhile to show these limits on the cumulative periodogram to provide a rough guide as to what deviations to regard with skepticism and what to take more note of.

The limit lines are such that for a truly random or white noise series, they would be crossed a proportion ε of the time. They are drawn at distances $\pm K_\varepsilon/\sqrt{q}$ above and below the theoretical line, where $q = (n - 2)/2$ for n even and $(n - 1)/2$ for n odd. Approximate values for K_ε are given in Table 8.3.

TABLE 8.3 Coefficients for Calculating Approximate Probability Limits for Cumulative Periodogram Test

ε	0.01	0.05	0.10	0.25
K_ε	1.63	1.36	1.22	1.02

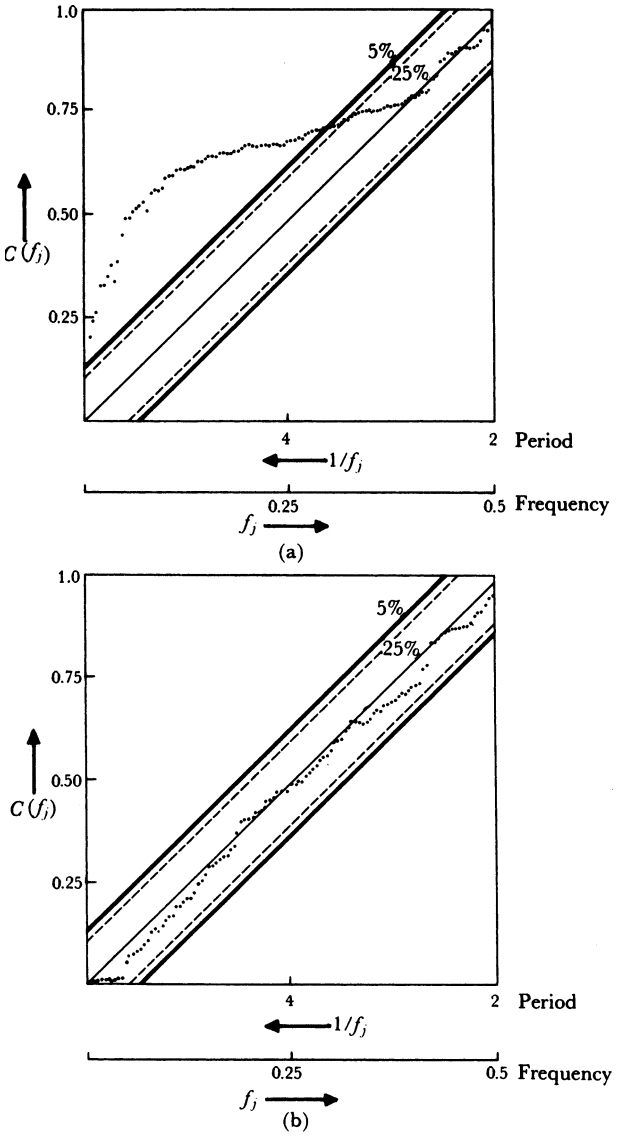


FIGURE 8.5 Series C: cumulative periodograms of residuals from best-fitting models (a) of order (0, 1, 1) and (b) of order (0, 2, 2).

For Series C, $q = (224 - 2)/2 = 111$, and the 5% limit lines inserted on Figure 8.5 deviate from the theoretical line by amounts $\pm 1.36/\sqrt{111} = \pm 0.13$. Similarly, the 25% limit lines deviate by $\pm 1.02/\sqrt{111} = \pm 0.10$.

Conclusions. Each of the model checking procedures described above has essential advantages and disadvantages. Checks based on the study of the estimated autocorrelation function and the cumulative periodogram, although they can point out *unsuspected* peculiarities of the series, may not be particularly sensitive. Tests for specific departures by

overfitting are more sensitive but may fail to warn of trouble other than that specifically anticipated. Portmanteau tests based on the residual autocorrelation and partial autocorrelations, while not always sensitive, provide convenient summary measures that are easy to use. As a result, they are now available in many software packages.

8.3 USE OF RESIDUALS TO MODIFY THE MODEL

8.3.1 Nature of the Correlations in the Residuals When an Incorrect Model Is Used

When the autocorrelation function of the residuals from a fitted model indicates that the model is inadequate, it is necessary to consider in what way the model should be modified. In Section 8.3.2, we show how the autocorrelations of the residuals can be used to suggest such modifications. As an introduction, we consider the effect of fitting an incorrect model on the autocorrelation function of the residuals.

Suppose that the correct model is

$$\phi(B)i\tilde{w}_t = \theta(B)a_t$$

but that an incorrect model

$$\phi_0(B)i\tilde{w}_t = \theta_0(B)b_t$$

is used. Then the residuals b_t, in the incorrect model, will be correlated and since

$$b_t = \theta_0^{-1}(B)\theta(B)\phi_0(B)\phi^{-1}(B)a_t \tag{8.3.1}$$

the autocovariance generating function of the b_t's will be

$$\sigma_a^2[\theta_0^{-1}(B)\theta_0^{-1}(F)\theta(B)\theta(F)\phi_0(B)\phi_0(F)\phi^{-1}(B)\phi^{-1}(F)] \tag{8.3.2}$$

For example, suppose that in an IMA(0, 1, 1) process, instead of the correct value θ, we use some other value θ_0. Then the residuals b_t would follow the mixed process of order (1, 0, 1):

$$(1 - \theta_0 B)b_t = (1 - \theta B)a_t$$

and using (3.4.8), we have

$$\rho_1 = \frac{(1 - \theta\theta_0)(\theta_0 - \theta)}{1 + \theta^2 - 2\theta\theta_0}$$
$$\rho_j = \rho_1\theta_0{}^{j-1} \qquad j = 2, 3, \ldots$$

For example, suppose that in the IMA(0, 1, 1) process,

$$\nabla z_t = (1 - \theta B)a_t$$

we took $\theta_0 = 0.8$ when the correct value was $\theta = 0$. Then

$$\theta_0 = 0.8 \qquad \theta = 0.0$$
$$\rho_1 = 0.8 \qquad \rho_j = 0.8^j$$

Thus, the b_t's would be highly autocorrelated and, since $(1 - 0.8B)b_t = \nabla z_t = a_t$, b_t would follow the autoregressive process

$$(1 - 0.8B)b_t = a_t$$

8.3.2 Use of Residuals to Modify the Model

Suppose that the residuals b_t from the model

$$\phi_0(B)\nabla^{d_0} z_t = \theta_0(B)b_t \tag{8.3.3}$$

appear to be nonrandom, that is, to deviate from white noise behavior. Using the autocorrelation function of b_t, the methods of Chapter 6 may now be applied to identify a model:

$$\phi_1(B)\nabla^{d_1} b_t = \theta_1(B)a_t \tag{8.3.4}$$

for the b_t series.On eliminating b_t between (8.3.3) and (8.3.4), we arrive at a new model:

$$\phi_0(B)\phi_1(B)\nabla^{d_0}\nabla^{d_1} z_t = \theta_0(B)\theta_1(B)a_t \tag{8.3.5}$$

which can now be fitted and diagnostically checked.

For example, suppose that a series had been wrongly identified as an IMA(0, 1, 1) process and fitted to give the model:

$$\nabla z_t = (1 + 0.6B)b_t \tag{8.3.6}$$

Also, suppose that a model

$$\nabla b_t = (1 + 0.8B)a_t \tag{8.3.7}$$

was identified for this residual series. Then on eliminating b_t between (8.3.6) and (8.3.7), we would obtain

$$\begin{aligned}\nabla^2 z_t &= (1 + 0.6B)\nabla b_t \\ &= (1 + 0.6B)(1 - 0.8B)a_t \\ &= (1 - 0.2B - 0.48B^2)a_t\end{aligned}$$

which would suggest that an IMA(0, 2, 2) process should now be entertained.

EXERCISES

8.1. The following are the first 30 residuals obtained when a tentative model was fitted to a time series:

t	Residuals					
1–6	0.78	0.91	0.45	−0.78	−1.90	−2.10
7–12	−0.54	−1.05	0.68	−3.77	−1.40	−1.77
13–18	1.18	0.02	1.29	−1.30	−6.20	−1.89
19–24	0.95	1.49	1.08	0.80	2.02	1.25
25–30	0.52	2.31	1.64	0.78	1.99	1.36

Plot the values and state any reservations you have concerning the adequacy of the model.

8.2. The residuals from a model $\nabla z_t = (1 - 0.6B)a_t$ fitted to a series of $N = 82$ observations yielded the following residual autocorrelations:

k	$r_k(\hat{a})$	k	$r_k(\hat{a})$
1	0.39	6	−0.13
2	0.20	7	−0.05
3	0.09	8	0.06
4	0.04	9	0.11
5	0.09	10	0.02

(a) Plot the residual ACF and determine whether there are any abnormal values relative to white noise behavior.

(b) Calculate the chi-square statistic $\tilde{Q}$ for lags up to $K = 10$ and check whether the residual autocorrelation function as a whole is indicative of model inadequacy.

(c) What modified model would you now tentatively entertain, fit, and check?

8.3. A long series containing $N = 326$ observations was split into two halves and a (1, 1, 0) model $(1 - \phi B)\nabla z_t = a_t$ identified, fitted, and checked for each half. If the estimates of the parameter ϕ for the two halves are $\hat{\phi}^{(1)} = 0.5$ and $\hat{\phi}^{(2)} = 0.7$, is there any evidence that the parameter ϕ has changed?

8.4. **(a)** Show that the variance of the sample mean $\overline{z}$ of n observations from a stationary AR(1) process $(1 - \phi B)\tilde{z}_t = a_t$ is given by

$$\text{var}[\overline{z}] \simeq \frac{\sigma_a^2}{n(1-\phi)^2}$$

(b) The yields from consecutive batches of a chemical process obtained under fairly uniform conditions of process control were shown to follow a stationary AR(1) process $(1 + 0.5B)\tilde{z}_t = a_t$. A technical innovation is made at a given point in time leading to 85 data points with mean $\overline{z}_1 = 41.0$ and residual variance $s_{a1}^2 = 0.1012$ before the innovation is made and 60 data points with $\overline{z}_2 = 43.5$ and $s_{a2}^2 = 0.0895$

after the innovation. Is there any evidence that the innovation has improved (increased) the yield?

8.5. Suppose that a (0, 1, 1) model $\nabla z_t = (1 - \theta B)e_t$, corresponding to the use of an exponentially weighted moving average forecast, with θ arbitrarily chosen to be equal to 0.5, was used to forecast a series that was, in fact, well represented by the (0, 1, 2) model $\nabla z_t = (1 - 0.9B + 0.2B^2)a_t$.

(a) Calculate the autocorrelation function of the lead 1 forecast errors e_t obtained from the (0, 1, 1) model.

(b) Show how this ACF could be used to identify a model for the e_t series, leading to the identification of a (0, 1, 2) model for the z_t series.

8.6. Two time series models, AR(2) and AR(3), were fitted to the yearly time series of sunspot numbers for the period 1770–1869 in Chapter 7. The sunspot data are available for the slightly longer time period 1700–1988 as series 'sunspot.year' in the datasets package in R; type help(sunspot.year) for details. Perform diagnostic checking to determine the adequacy of the AR(2) and AR(3) models for this longer time period. Are there alternative models that you would consider for this series? Would you recommend that a data transformation be used in this case?

8.7. Monthly sales, $\{Y_t\}$, of a company over a period of 150 months are provided as part of Series M in Part 5 of this book. This series is also available as series $BJsales$ along with a related series $BJsales.lead$ in the datasets package in R.

(a) Plot the data and comment.

(b) Perform a statistical analysis to determine a suitable model for this series. Estimate the parameters using the maximum likelihood method.

(c) Repeat the analysis for the series of leading indicator $BJsales.lead$ that is part of the same dataset.

(d) Perform diagnostic checking to determine if there is any lack of fit in the models selected for the two series?

8.8 Global mean surface temperature deviations (from the 1951–1980 average level) are available for the period 1880–2009 as series 'gtemp2' in the astsa package in R.

(a) Plot the data and comment. Are there any unusual features worth noting?

(b) Perform a statistical analysis to determine a suitable model for this series. Estimate the parameters using the maximum likelihood method.

(c) Is there evidences of any lack of fit in the models selected for this series?

(d) Can you suggest an alternative way to analyze this time series? How might an analysis of model generated forecasts impact your choice of model?

8.9 Refer to the daily air quality measurements for New York, May to September 1973, analyzed in Problem 7.10 of Chapter 7. Perform diagnostic checks to determine the adequacy of the models fitted to average daily temperature and wind speed series.

8.10 Repeat the analysis in Problem 8.9 by performing diagnostic checks on the model, or models, considered for the solar radiation series in Problem 7.11.

9

ANALYSIS OF SEASONAL TIME SERIES

In Chapters 3–8, we have considered the properties of a class of linear stochastic models, which are of value in representing stationary and nonstationary time series, and we have seen how these models may be used for forecasting. We then considered the practical problems of identification, fitting, and diagnostic checking that arise when relating these models to actual data. In this chapter, we apply these methods to analyzing and forecasting seasonal time series. A key focus is on seasonal multiplicative time series models that account for time series dependence across seasons as well as between adjacent values in the series. These models are extensions of the ARIMA models discussed in earlier chapters. The methodology is illustrated using a time series commonly referred to as the airline data in the time series literature. We also describe an alternate structural component model approach to representing stochastic seasonal and trend behavior that includes the possibility of the components being deterministic. The chapter concludes with a brief discussion of regression models with autocorrelated errors. These models could include deterministic sine or cosine terms to describe the seasonal behavior of the series.

9.1 PARSIMONIOUS MODELS FOR SEASONAL TIME SERIES

Figure 9.1 shows monthly totals of international airline passengers for the 12-year period from January 1949 to December 1960. This series was discussed by Brown (1962) and is listed as Series G in Part Five of this book. The series is also included as series ''AirPassengers'' in the R datasets package and is conveniently downloaded from there. The series shows a marked seasonal pattern since travel is at its highest in the late summer months, while a secondary peak occurs in the spring. Many other series, particularly sales data, show similar seasonal characteristics.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

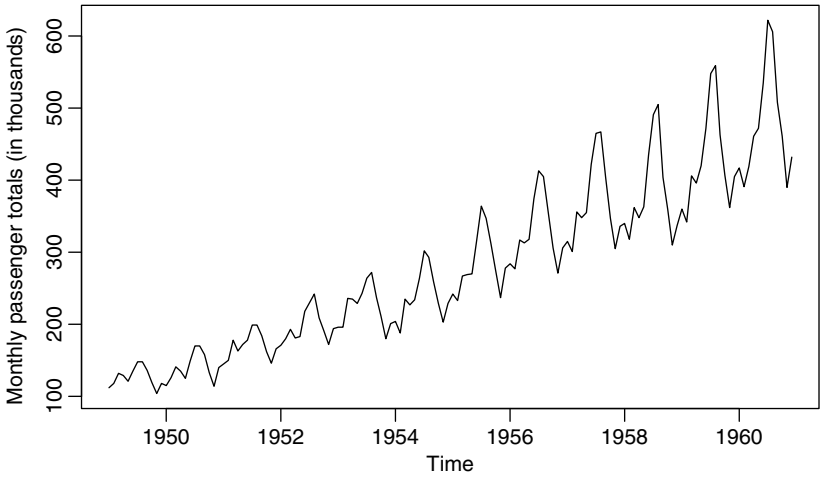


FIGURE 9.1 Totals of international airline passengers in thousands (Series G).

In general, we say that a series exhibits periodic behavior with period s, when similarities in the series occur after s basic time intervals. In the above example, the basic time interval is 1 month and the period is $s = 12$ months. However, examples occur when s can take on other values. For example, $s = 4$ for quarterly data showing seasonal effects within years. It sometimes happens that there is more than one period. Thus, because bills tend to be paid monthly, we would expect weekly business done by a bank to show a periodicity of about 4 within months, while monthly business shows a periodicity of 12.

9.1.1 Fitting Versus Forecasting

A common method of analyzing seasonal time series in the past was to decompose the series arbitrarily into three components: a *trend*, a *seasonal component*, and a *random component*. The trend might be fitted by a polynomial and the seasonal component by a Fourier series. A forecast was then made by projecting these fitted functions. However, such methods could give misleading results if applied indiscriminately. For example, we have seen that the behavior of IBM stock prices in Series B is closely approximated by the random walk model $\nabla z_t = a_t$, that is,

$$z_t = z_0 + \sum_{j=0}^{t-1} a_{t-j} \tag{9.1.1}$$

This implies that $\hat{z}_t(l) = z_t$. In other words, the best forecast of future values of the stock is very nearly today's price. While it is true that short segments of Series B look as if they might be fitted by quadratic curves, this simply reflects the fact that a sum of random deviates can sometimes have this appearance. There is no basis for the use of a quadratic forecast function, which would produce very poor forecasts for this particular series. Similarly, while deterministic trend and seasonal components can provide a good fit to the data, they are often too rigid when it comes to forecasting. In this section, we introduce a seasonal

time series model that requires very few parameters and avoids the assumption of a trend and seasonal component that remains fixed over time.

9.1.2 Seasonal Models Involving Adaptive Sines and Cosines

The general linear model

$$\tilde{z}_t = \sum_{j=1}^{\infty} \pi_j \tilde{z}_{t-j} + a_t = \sum_{j=1}^{\infty} \psi_j a_{t-j} + a_t \tag{9.1.2}$$

with suitable values for the coefficients π_j and ψ_j can be used to describe many seasonal time series. The problem is to choose a suitable *parsimonious parameterization* for such models. We have seen that for nonseasonal series, it is usually possible to obtain a useful and parsimonious representation in the form

$$\varphi(B)\tilde{z}_t = \theta(B)a_t \tag{9.1.3}$$

Moreover, the generalized autoregressive operator $\varphi(B)$ determines the eventual forecast function, which is the solution of the difference equation

$$\varphi(B)\hat{z}_t(l) = 0$$

where B is understood to operate on l. In representing seasonal behavior, we want the forecast function to trace out a periodic pattern. A first thought might be that $\varphi(B)$ should produce a forecast function consisting of a mixture of sines and cosines, and possibly mixed with polynomial terms, to allow changes in the level of the series and changes in the seasonal pattern. Such a forecast function could arise naturally within the structure of the general model (9.1.3). For example, with monthly data, a forecast function that is a sine wave with a 12-month period, adaptive in phase and amplitude, will satisfy the difference equation

$$(1 - \sqrt{3}B + B^2)\hat{z}_t(l) = 0$$

where B is understood to operate on l. However, periodic behavior may not be *economically* represented by mixtures of sines and cosines. Many sine–cosine components would, for example, be needed to represent sales data affected by Christmas, Easter, and other seasonal buying. To take an extreme case, sales of fireworks in Britain are largely confined to the weeks immediately before November 5, when the abortive attempt of Guy Fawkes to blow up the Houses of Parliament is celebrated. An attempt to represent the ''single spike'' of fireworks sales data directly by sines and cosines might be unprofitable. It is clear that a more careful consideration of the problem is needed.

Now, in our previous analysis, we have not necessarily estimated *all* the components of $\varphi(B)$. Where differencing d times was needed to induce stationarity, we have written $\varphi(B) = \phi(B)(1 - B)^d$, which is equivalent to setting d roots of the equation $\varphi(B) = 0$ equal to unity. When such a representation proved adequate, we could proceed with the simpler analysis of $w_t = \nabla^d z_t$. Thus, we have used $\nabla = 1 - B$ as a simplifying operator. In other problems, different types of simplifying operators might be appropriate. For example, the consumption of fuel oil for heat is highly dependent on ambient temperature, which, because the Earth rotates around the sun, is known to follow approximately a sine wave with

period of 12 months. In analyzing sales of fuel oil, it might then make sense to introduce $1 - \sqrt{3}B + B^2$ as a simplifying operator, constituting one of the contributing components of the generalized autoregressive operator $\varphi(B)$. If such a representation proved useful, we could then proceed with the simpler analysis of $w_t = (1 - \sqrt{3}B + B^2)z_t$. This operator is of the homogeneous nonstationary variety, having zeros $e^{\pm(i2\pi/12)}$ on the unit circle.

9.1.3 General Multiplicative Seasonal Model

Simplifying Operator $\mathbf{1-}B^s$**.** The fundamental fact about seasonal time series with period s is that observations that are s intervals apart are similar. Therefore, one can expect that the operation $B^s z_t = z_{t-s}$ will play a particularly important role in the analysis of seasonal series. Furthermore, since nonstationarity is to be expected in the series $z_t, z_{t-s}, z_{t-2s}, \ldots$, the simplifying operation

$$\nabla_s z_t = (1 - B^s)z_t = z_t - z_{t-s}$$

should be useful. This nonstationary operator $1 - B^s$ has s zeros $e^{i(2\pi k/s)}(k = 0, 1, \ldots, s - 1)$ evenly spaced on the unit circle. Moreover, the eventual forecast function satisfies $(1 - B^s)\hat{z}_t(l) = 0$ and so may (but need not) be represented by a full complement of sines and cosines:

$$\hat{z}_t(l) = b_0^{(t)} + \sum_{j=1}^{[s/2]} \left[b_{1j}^{(t)} \cos\left(\frac{2\pi jl}{s}\right) + b_{2j}^{(t)} \sin\left(\frac{2\pi jl}{s}\right) \right]$$

where the b's are adaptive coefficients, and where $[s/2] = \frac{1}{2}s$ if s is even and $[s/2] = \frac{1}{2}(s - 1)$ if s is odd.

Multiplicative Model. When a series exhibits seasonal behavior with known periodicity s, it is useful to display the data in the form of a table containing s columns, such as Table 9.1, which shows the logarithms of the airline data. For seasonal data, special care is needed in selecting an appropriate transformation. In this example, data analysis supports the use of the logarithm (see Section 9.3.5).

The arrangement of Table 9.1 emphasizes the fact that, in periodic data, there are not one but two time intervals of importance. For this example, these intervals correspond to months and years. Specifically, we expect relationships to occur (a) between the observations for successive months in a particular year and (b) between the observations for the same month in successive years. The situation is somewhat like that in a two-way analysis of variance model, where similarities can be expected between observations in the same column and between observations in the same row.

For the airline data, the seasonal effect implies that an observation for a particular month, say April, is related to the observations for previous Aprils. Suppose that the t-th observation z_t is for the month of April. We might be able to link this observation z_t to observations in previous Aprils by a model of the form

$$\Phi(B^s)\nabla_s^D z_t = \Theta(B^s)\alpha_t \tag{9.1.4}$$

TABLE 9.1 Natural Logarithms of Monthly Passenger Totals (Measured in Thousands) in International Air Travel (Series G)

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1949	4.718	4.771	4.883	4.860	4.796	4.905	4.997	4.997	4.913	4.779	4.644	4.771
1950	4.745	4.836	4.949	4.905	4.828	5.004	5.136	5.136	5.063	4.890	4.736	4.942
1951	4.977	5.011	5.182	5.094	5.147	5.182	5.293	5.293	5.215	5.088	4.984	5.112
1952	5.142	5.193	5.263	5.199	5.209	5.384	5.438	5.489	5.342	5.252	5.147	5.268
1953	5.278	5.278	5.464	5.460	5.434	5.493	5.576	5.606	5.468	5.352	5.193	5.303
1954	5.318	5.236	5.460	5.245	5.455	5.576	5.710	5.680	5.557	5.434	5.313	5.434
1955	5.489	5.451	5.587	5.595	5.598	5.753	5.897	5.849	5.743	5.613	5 648	5.628
1956	5.649	5.624	5.759	5.746	5.762	5.924	6.023	6.004	5.872	5.724	5.602	5.724
1957	5.753	5.707	5.875	5.852	5.872	6.045	6.142	6.146	6.001	5.849	5.720	5.817
1958	5.829	5.762	5.892	5.852	5.894	6.075	6.196	6.225	6.001	5.883	5.737	5.820
1959	5.886	5.835	6.006	5.981	6.040	6.157	6.306	6.326	6.138	6.009	5.892	6.004
1960	6.033	5.969	6.038	6.133	6.157	6.282	6.433	6.407	6.230	6.133	5.966	6.068

where $s = 12$, $\nabla_s = 1 - B^s$, and $\Phi(B^s)$, $\Theta(B^s)$ are polynomials in B^s of degrees P and Q, respectively, and satisfying stationarity and invertibility conditions. Similarly, a model

$$\Phi(B^s)\nabla_s^D z_{t-1} = \Theta(B^s)\alpha_{t-1} \tag{9.1.5}$$

might be used to link the current behavior for March with previous March observations, and so on, for each of the 12 months. Moreover, it is usually reasonable to assume that the parameters $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ contained in these monthly models would be approximately the same for each month.

Now the error components, $\alpha_t, \alpha_{t-1}, \ldots$, in these models would not in general be uncorrelated. For example, the total of airline passengers in April 1960, while related to previous April totals, would also be related to totals in March 1960, February 1960, January 1960, and so on. Thus, we would expect that α_t in (9.1.4) would be related to α_{t-1} in (9.1.5) and to α_{t-2}, and so on. Therefore, to account for such relationships, we introduce a second model

$$\phi(B)\nabla^d \alpha_t = \theta(B)a_t \tag{9.1.6}$$

where now a_t is a white noise process and $\phi(B)$ and $\theta(B)$ are polynomials in B of degrees p and q, respectively, and satisfying stationarity and invertibility conditions, and $\nabla = \nabla_1 = 1 - B$.

Substituting (9.1.6) in (9.1.4), we obtain a general multiplicative model

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D z_t = \theta_q(B)\Theta_Q(B^s)a_t \tag{9.1.7}$$

where, for this particular example, $s = 12$. Also, the subscripts p, P, q, and Q have been added to indicate the orders of the various operators. The resulting multiplicative process will be said to be of order $(p, d, q) \times (P, D, Q)_s$. A similar argument can be used to obtain models with three or more periodic components to take care of multiple seasonalities.

In the next two sections, we examine some basic forms of the seasonal model introduced above and demonstrate their potential for forecasting. We also consider the problems of identification, estimation, and diagnostic checking that arise in relating such models to data. No new principles are needed to do this, merely an application of the procedures and ideas already discussed in Chapters 6–8. This is illustrated in the next section where a seasonal ARIMA model of order $(0, 1, 1) \times (0, 1, 1)_{12}$ is used to represent the airline data.

9.2 REPRESENTATION OF THE AIRLINE DATA BY A MULTIPLICATIVE $(0, 1, 1) \times (0, 1, 1)_{12}$ MODEL

9.2.1 Multiplicative $(0, 1, 1) \times (0, 1, 1)_{12}$ Model

We have seen that a simple and widely applicable stochastic model for the analysis of nonstationary time series, which contains no seasonal component, is the IMA(0, 1, 1) process. Suppose, following the argument presented above, that we have a seasonal time series and employ the model

$$\nabla_{12} z_t = (1 - \Theta B^{12})\alpha_t$$

for linking z's 1-year apart. Suppose further that we employ a similar model

$$\nabla \alpha_t = (1 - \theta B) a_t$$

for linking α's 1-month apart, where in general θ and Θ will have different values. Then, on combining these expressions, we obtain the seasonal multiplicative model

$$\nabla \nabla_{12} z_t = (1 - \theta B)(1 - \Theta B^{12}) a_t \tag{9.2.1}$$

of order $(0, 1, 1) \times (0, 1, 1)_{12}$. The model written explicitly is

$$z_t - z_{t-1} - z_{t-12} + z_{t-13} = a_t - \theta a_{t-1} - \Theta a_{t-12} + \theta \Theta a_{t-13} \tag{9.2.2}$$

The invertibility region for this model, required by the condition that the roots of $(1 - \theta B)(1 - \Theta B^{12}) = 0$ lie outside the unit circle, is defined by the inequalities $-1 < \theta < 1$ and $-1 < \Theta < 1$. Note that the moving average operator $(1 - \theta B)(1 - \Theta B^{12}) = 1 - \theta B - \Theta B^{12} + \theta \Theta B^{13}$, on the right-hand side of (9.2.1), is of order $q + sQ = 1 + 12(1) = 13$.

We will show below that the logged airline data are well represented by a model of this form, where to a sufficient approximation, $\hat{\theta} = 0.4$, $\hat{\Theta} = 0.6$, and $\hat{\sigma}_a^2 = 1.34 \times 10^{-3}$. However, as a preliminary, we first consider how this model and with these parameter values inserted can be used to forecast future values of the series.

9.2.2 Forecasting

In Chapter 4, we saw that there are three basically different ways of considering the general model, each giving rise to a different way of viewing the forecast in Chapter 5. We consider now these three approaches for the forecasting of the seasonal model introduced above.

Difference Equation Approach. Forecasts are best *computed* directly from the difference equation itself. Thus, since

$$z_{t+l} = z_{t+l-1} + z_{t+l-12} - z_{t+l-13} + a_{t+l} - \theta a_{t+l-1} - \Theta a_{t+l-12} + \theta \Theta a_{t+l-13} \tag{9.2.3}$$

after setting $\theta = 0.4$, $\Theta = 0.6$, the minimum mean square error forecast at lead time l and origin t is given immediately by

$$\hat{z}_t(l) = [z_{t+l-1} + z_{t+l-12} - z_{t+l-13} + a_{t+l} - 0.4 a_{t+l-1} - 0.6 a_{t+l-12} + 0.24 a_{t+l-13}] \tag{9.2.4}$$

where

$$[z_{t+l}] = E[z_{t+l} | z_t, z_{t-1}, \cdots; \theta, \Theta]$$

is the conditional expectation of z_{t+l} taken at origin t. In this expression, the parameters are assumed to be known, and knowledge of the series $z_t, z_{t-1}, \ldots$ is assumed to extend into the remote past.

Practical application depends upon the following facts:

1. Invertible models fitted to actual data usually yield forecasts that depend appreciably only on recent values of the series.

2. The forecasts are insensitive to small changes in parameter values such as are introduced by estimation errors.

Now

$$[z_{t+j}] = \begin{cases} z_{t+j} & j \le 0 \\ \hat{z}_t(j) & j > 0 \end{cases} \tag{9.2.5}$$

$$[a_{t+j}] = \begin{cases} a_{t+j} & j \le 0 \\ 0 & j > 0 \end{cases} \tag{9.2.6}$$

Thus, to obtain the forecasts, we simply replace unknown z's by forecasts and unknown a's by zeros. The known a's are, of course, the one-step-ahead forecast errors already computed, that is, $a_t = z_t - \hat{z}_{t-1}(1)$.

For example, to obtain the 3-months-ahead forecast, we have

$$z_{t+3} = z_{t+2} + z_{t-9} - z_{t-10} + a_{t+3} - 0.4a_{t+2} - 0.6a_{t-9} + 0.24a_{t-10}$$

Taking conditional expectations at the origin t gives

$$\hat{z}_t(3) = \hat{z}_t(2) + z_{t-9} - z_{t-10} - 0.6a_{t-9} + 0.24a_{t-10}$$

Substituting $a_{t-9} = z_{t-9} - \hat{z}_{t-10}(1)$ and $a_{t-10} = z_{t-10} - \hat{z}_{t-11}(1)$ on the right-hand side also yields

$$\hat{z}_t(3) = \hat{z}_t(2) + 0.4z_{t-9} - 0.76z_{t-10} + 0.6\hat{z}_{t-10}(1) - 0.24\hat{z}_{t-11}(1) \tag{9.2.7}$$

which expresses the forecast in terms of previous z's and previous forecasts of z's.

Figure 9.2 shows the forecasts for lead times up to 36 months, all made at the arbitrarily selected origin, July 1957. We see that the simple model, containing only two parameters, faithfully reproduces the seasonal pattern and supplies excellent forecasts. It is to be remembered, of course, that like all predictions obtained from the general linear stochastic model, the forecast function is adaptive. When changes occur in the seasonal pattern, these will be appropriately projected into the forecast. It will be noticed that when the 1-month-ahead forecast is too high, there is a tendency for all future forecasts from the point to be high. This is to be expected because, as has been noted in Appendix A5.1, forecast errors from the same origin, but for different lead times, are highly correlated. Of course, a forecast for a long lead time, such as 36 months, may necessarily contain a fairly large error. However, in practice, an initially remote forecast will be updated continually, and as the lead shortens, greater accuracy will be possible.

The preceding forecasting procedure is robust to moderate changes in the parameter values. Thus, if we used $\theta = 0.5$ and $\Theta = 0.5$, instead of $\theta = 0.4$ and $\Theta = 0.6$, the forecasts would not be greatly affected. This is true even for forecasts made several steps ahead (e.g., 12 months). The approximate effect on the one-step-ahead forecasts of modifying the values of the parameters can be seen by studying the sum-of-squares surface. Thus, we know that the approximate confidence region for the k parameters $\boldsymbol{\beta}$ is bounded, in general, by the contour $S(\boldsymbol{\beta}) = S(\hat{\boldsymbol{\beta}})[1 + \chi^2_\varepsilon(k)/n]$, which includes the true parameter point with probability $1 - \varepsilon$. Therefore, we know that, had the *true* parameter values been employed,

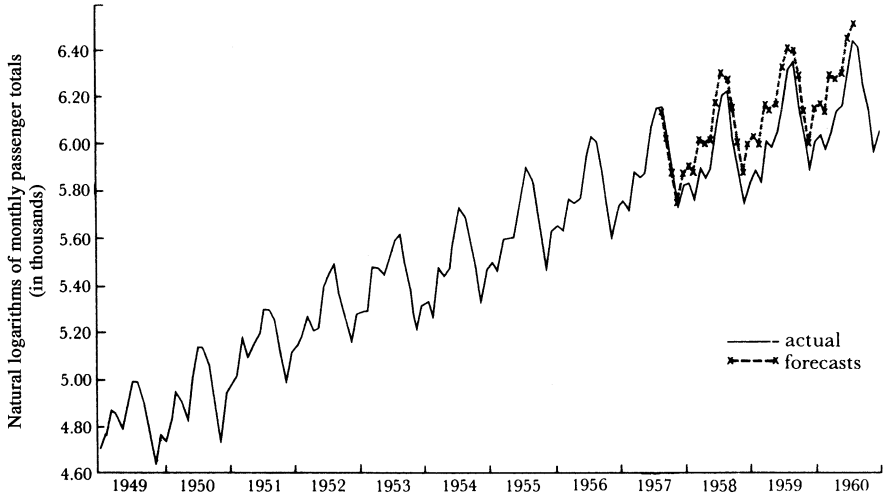


FIGURE 9.2 Airline data with forecasts for 1, 2, 3, …, 36 months ahead, all made from an arbitrary selected origin, July 1957.

with this same probability the mean square of the one-step-ahead forecast errors could not have been increased by a factor greater than $1 + \chi^2_\varepsilon(k)/n$.

Forecast Function, Its Updating, and the Forecast Error Variance. In practice, the difference equation procedure is by far the simplest and most convenient way for actually *computing* forecasts and updating them. However, the difference equation itself does not reveal very much about the *nature* of the forecasts and their updating. To cast light on these aspects, we now consider the forecasts from other points of view.

Forecast Function. Using (5.1.12) yields $z_{t+l} = \hat{z}_t(l) + e_t(l)$, where

$$e_t(l) = a_{t+l} + \psi_1 a_{t+l-1} + \cdots + \psi_{l-1} a_{t+1} \tag{9.2.8}$$

Now, the moving average operator on the right-hand side of (9.2.1) is of order 13. Hence, for $l > 13$, the forecasts satisfy the difference equation

$$(1 - B)(1 - B^{12})\hat{z}_t(l) = 0 \qquad l > 13 \tag{9.2.9}$$

where, in this equation, B operates on the lead time l.

We now write $l = (r, m) = 12r + m, r = 0, 1, 2, \ldots$ and $m = 1, 2, \ldots, 12$, to represent a lead time of r years and m months, so that, for example, $l = 15 = (1, 3)$. Then, the forecast function, which is the solution of (9.2.9), with starting conditions given by the first 13 forecasts, is of the form

$$\hat{z}_t(l) = \hat{z}_t(r, m) = b^{(t)}_{0,m} + r b^{(t)}_1 \qquad l > 0 \tag{9.2.10}$$

This forecast function contains 13 adjustable coefficients $b^{(t)}_{0,1}, b^{(t)}_{0,2}, \ldots, b^{(t)}_{0,12}, b^{(t)}_1$. These represent 12 monthly contributions and 1 yearly contribution and are determined by the

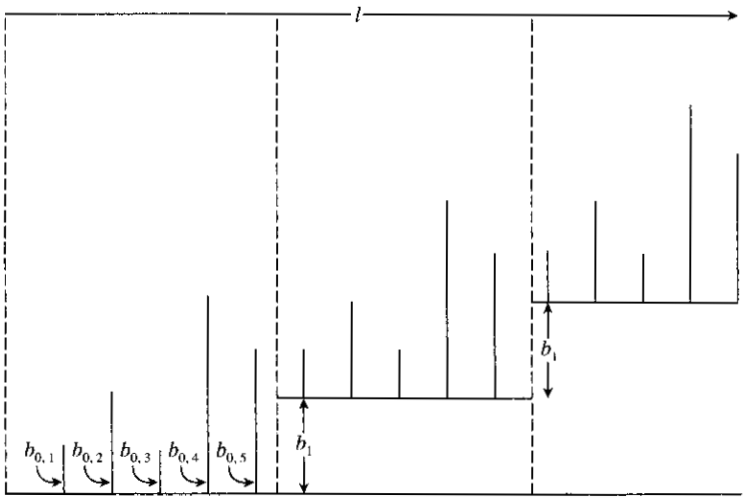


FIGURE 9.3 Seasonal forecast function generated by the model $\nabla\nabla_{12}z_t = (1 - \theta B)(1 - \Theta B^s)a_t$, with $s = 5$.

first 13 forecasts. The nature of this function is more clearly understood from Figure 9.3, which shows a forecast function of this kind, but with period $s = 5$, so that there are six adjustable coefficients $b_{0,1}^{(t)}, b_{0,2}^{(t)}, \ldots, b_{0,5}^{(t)}, b_1^{(t)}$.

Equivalently, since $\hat{z}_t(l)$ satisfies (9.2.9) and the roots of $(1 - B)(1 - B^{12}) = 0$ are $1, 1, -1, e^{\pm(i2\pi k/12)}, k = 1, \ldots, 5$, on the unit circle, the forecast function, as in (5.3.3), can be represented as

$$\hat{z}_t(l) = \sum_{j=1}^{5} \left[b_{1j}^{(t)} \cos\left(\frac{2\pi jl}{12}\right) + b_{2j}^{(t)} \sin\left(\frac{2\pi jl}{12}\right) \right] + b_{16}^{(t)}(-1)^l + b_0^{(t)} + b_1^{*(t)} l$$

This shows that $\hat{z}_t(l)$ consists of a mixture of sinusoids at the seasonal frequencies $2\pi j/12, j = 1, \ldots, 6$, plus a linear trend with slope $b_1^{*(t)}$. The coefficients $b_{1j}^{(t)}, b_{2j}^{(t)}, b_0^{(t)}$, and $b_1^{*(t)}$ in the expression above are all adaptive with regard to the forecast origin t, being determined by the first 13 forecasts. In comparison to (9.2.10), it is clear, for example, that $b_1^{(t)} = 12b_1^{*(t)}$, and represents the *annual* rate of change in the forecasts $\hat{z}_t(l)$, whereas $b_1^{*(t)}$ is the *monthly* rate of change.

The ψ Weights. To determine updating formulas and to obtain the variance of the forecast error $e_t(l)$ in (9.2.8), we need the ψ weights in the form $z_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}$ of the model. We can write the moving average operator in (9.2.1) in the form

$$(1 - \theta B)(1 - \Theta B^{12}) = (\nabla + \lambda B)(\nabla_{12} + \Lambda B^{12})$$

where $\lambda = 1 - \theta, \Lambda = 1 - \Theta, \nabla_{12} = 1 - B^{12}$. Hence, the model may be written as

$$\nabla\nabla_{12}z_t = (\nabla + \lambda B)(\nabla_{12} + \Lambda B^{12})a_t$$

By equating coefficients in $\nabla\nabla_{12}\psi(B) = (\nabla + \lambda B)(\nabla_{12} + \Lambda B^{12})$, it can be seen that the ψ weights satisfy $\psi_0 = 1, \psi_1 - \psi_0 = \lambda - 1, \psi_{12} - \psi_{11} - \psi_0 = \Lambda - 1, \psi_{13} - \psi_{12} - \psi_1 + \psi_0 = (\lambda - 1)(\Lambda - 1)$, and $\psi_j - \psi_{j-1} - \psi_{j-12} + \psi_{j-13} = 0$ otherwise. Thus, the ψ weights for this process are

$$\begin{aligned} \psi_1 = \psi_2 = \cdots = \psi_{11} &= \lambda & \psi_{12} &= \lambda + \Lambda \\ \psi_{13} = \psi_{14} = \cdots = \psi_{23} &= \lambda(1 + \Lambda) & \psi_{24} &= \lambda(1 + \Lambda) + \Lambda \\ \psi_{25} = \psi_{26} = \cdots = \psi_{35} &= \lambda(1 + 2\Lambda) & \psi_{36} &= \lambda(1 + 2\Lambda) + \Lambda \end{aligned}$$

and so on. Writing ψ_j as $\psi_{r,m} = \psi_{12r+m}$, where $r = 0, 1, 2, \ldots$ and $m = 1, 2, \ldots, 12$, refer, respectively, to years and months, we obtain

$$\psi_{r,m} = \lambda(1 + r\Lambda) + \delta\Lambda \tag{9.2.11}$$

where

$$\delta = \begin{cases} 1 \text{ when } m = 12 \\ 0 \text{ when } m \neq 12 \end{cases}$$

Updating. The general updating formula (5.2.5) is

$$\hat{z}_{t+1}(l) = \hat{z}_t(l+1) + \psi_l a_{t+1}$$

Thus, if $m \neq s = 12$,

$$b_{0,m}^{(t+1)} + r b_1^{(t+1)} = b_{0,m+1}^{(t)} + r b_1^{(t)} + (\lambda + r\lambda\Lambda) a_{t+1}$$

and on equating coefficients of r, the updating formulas are

$$\begin{aligned} b_{0,m}^{(t+1)} &= b_{0,m+1}^{(t)} + \lambda a_{t+1} \\ b_1^{(t+1)} &= b_1^{(t)} + \lambda\Lambda a_{t+1} \end{aligned} \tag{9.2.12}$$

Alternatively, if $m = s = 12$,

$$b_{0,12}^{(t+1)} + r b_1^{(t+1)} = b_{0,1}^{(t)} + (r+1) b_1^{(t)} + (\lambda + \Lambda + r\lambda\Lambda) a_{t+1}$$

and in this case,

$$\begin{aligned} b_{0,12}^{(t+1)} &= b_{0,1}^{(t)} + b_1^{(t)} + (\lambda + \Lambda) a_{t+1} \\ b_1^{(t+1)} &= b_1^{(t)} + \lambda\Lambda a_{t+1} \end{aligned} \tag{9.2.13}$$

In studying these relations, it should be remembered that $b_{0,m}^{(t+1)}$ will be the updated version of $b_{0,m+1}^{(t)}$. Thus, if the origin t was January of a particular year, $b_{0,2}^{(t)}$ would be the coefficient for March. After a month had elapsed, we should move the forecast origin to February and the updated version for the March coefficient would now be $b_{0,1}^{(t+1)}$.

Forecast Error Variance. Knowledge of the ψ weights enables us to calculate the variance of the forecast errors at any lead time l, using the result (5.1.16), namely

$$V(l) = (1 + \psi_1^2 + \cdots + \psi_{l-1}^2)\sigma_a^2 \tag{9.2.14}$$

Thus, setting $\lambda = 0.6, \Lambda = 0.4, \sigma_a^2 = 1.34 \times 10^{-3}$ in (9.2.11) and (9.2.14), the estimated standard deviations $\hat{\sigma}(l)$ of the forecast errors of the log airline data are readily calculated for different lead times.

Forecasts as a Weighted Average of Previous Observations. If we write the model in the form

$$z_t = \sum_{j=1}^{\infty} \pi_j z_{t-j} + a_t$$

the one-step-ahead forecast is

$$\hat{z}_t(1) = \sum_{j=1}^{\infty} \pi_j z_{t+1-j}$$

The π weights may be obtained by equating coefficients in

$$(1 - B)(1 - B^{12}) = (1 - \theta B)(1 - \Theta B^{12})(1 - \pi_1 B - \pi_2 B^2 - \cdots)$$

Thus,

$$\begin{aligned} \pi_j &= \theta^{j-1}(1 - \theta) \qquad j = 1, 2, \ldots, 11 \\ \pi_{12} &= \theta^{11}(1 - \theta) + (1 - \Theta) \\ \pi_{13} &= \theta^{12}(1 - \theta) - (1 - \theta)(1 - \Theta) \\ \pi_j &- \theta\pi_{j-1} - \Theta\pi_{j-12} + \theta\Theta\pi_{j-13} \qquad j \geq 14 \end{aligned} \tag{9.2.15}$$

These weights are plotted in Figure 9.4 for the parameter values $\theta = 0.4$ and $\Theta = 0.6$.

The reason that the weight function takes the particular form shown in the figure may be understood as follows: the process (9.2.1) may be written as

$$a_{t+1} = \left(1 - \frac{\lambda B}{1 - \theta B}\right)\left(1 - \frac{\Lambda B^{12}}{1 - \Theta B^{12}}\right) z_{t+1} \tag{9.2.16}$$

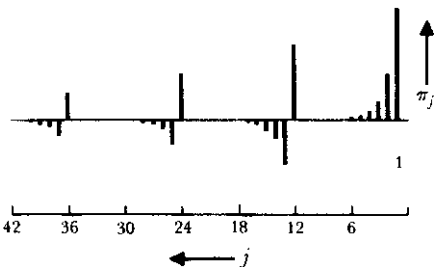


FIGURE 9.4 The π weights for $(0, 1, 1) \times (0, 1, 1)_{12}$ process fitted to the airline data ($\theta = 0.4, \Theta = 0.6$).

We now use the notation $\text{EWMA}_\lambda(z_t)$ to mean an exponentially weighted moving average, with parameter $\lambda = 1 - \theta$ of values $z_t, z_{t-1}, z_{t-2}, \ldots$, so that

$$\text{EWMA}_\lambda(z_t) = \frac{\lambda}{1 - \theta B} z_t = \lambda z_t + \lambda\theta z_{t-1} + \lambda\theta^2 z_{t-2} + \cdots$$

Similarly, we use $\text{EWMA}_\Lambda(z_t)$ to mean an exponentially weighted moving average, with parameter $\Lambda = 1 - \Theta$, of values $z_t, z_{t-12}, z_{t-24}, \ldots$, so that

$$\text{EWMA}_\Lambda(z_t) = \frac{\Lambda}{1 - \Theta B^{12}} z_t = \Lambda z_t + \Lambda\Theta z_{t-12} + \Lambda\Theta^2 z_{t-24} + \cdots$$

Substituting $\hat{z}_t(1) = z_{t+1} - a_{t+1}$, in (9.2.16), we obtain

$$\hat{z}_t(1) = \text{EWMA}_\lambda(z_t) + \text{EWMA}_\Lambda(z_{t-11} - \text{EWMA}_\lambda(z_{t-12})) \tag{9.2.17}$$

Thus, the forecast is an EWMA taken over previous months, modified by a second EWMA of discrepancies found between similar monthly EWMAs and actual performance in previous years. As a particular case, if $\theta = 0$ ($\lambda = 1$), (9.2.17) would reduce to

$$\begin{aligned}\hat{z}_t(1) &= z_t + \text{EWMA}_\Lambda(z_{t-11} - z_{t-12}) \\ &= z_t + \Lambda[(z_{t-11} - z_{t-12}) + \Theta(z_{t-23} - z_{t-24}) + \cdots]\end{aligned}$$

which shows that first differences are forecast as the seasonal EWMA of first differences for similar months from previous years.

For example, suppose that we were attempting to predict December sales for a department store. These sales would include a heavy component from Christmas buying. The first term on the right-hand side of (9.2.17) would be an EWMA taken over previous months up to November. However, we know this will be an underestimate, so we correct it by taking a second EWMA over previous years of the *discrepancies* between actual December sales and the corresponding monthly EWMAs taken over previous months in those years.

The forecasts for lead times $l > 1$ can be generated from the π weights by substituting forecasts of shorter lead time for unknown values, as displayed in the general expression (5.3.6) of Section 5.3.3. Alternatively, explicit values for the weights applied directly to $z_t, z_{t-1}, z_{t-2}, \ldots$ may be computed, for example, from (5.3.9) or from (A5.2.3).

Calculation of Forecasts in R. Forecasts of future values of a time series that follows a multiplicative seasonal model can be calculated using R. A convenient option available in R is the command sarima.for() in the astsa package. For a series z_t that follows a multiplicative model with period s, the command is sarima.for(z,n.ahead,p,d,q,P,D,Q,s), where n.ahead is the lead time. Thus, to generate forecasts up to 24 steps ahead for the logged airline series using the model $\nabla\nabla_{12} z_t = (1 - \theta B)(1 - \Theta B^{12}) a_t$, the commands are

```
> library(astsa)
> ap=ts(seriesG,start=c(1949,1),frequency=12)
> log.AP=log(ap)
> m1=sarima.for(log.AP,24,0,1,1,0,1,1,12)
> m1   % retrieves output from a file
```

The output includes the forecasts (''pred'') and the prediction errors (''se'') of the forecasts. A graph of the forecasts with ± 2 prediction error limits attached is provided as part of the

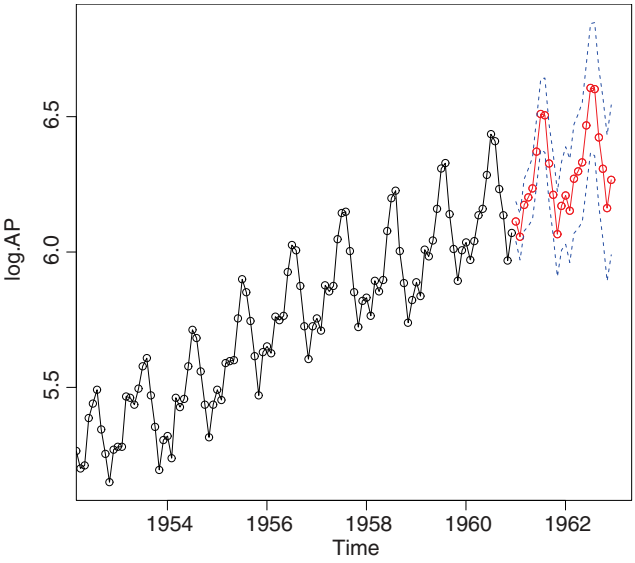


FIGURE 9.5 Forecasts along with ±2 prediction error limits for the logarithm of the airline data generated from the model $\nabla\nabla_{12} z_t = (1 - \theta B)(1 - \Theta B^{12})a_t$.

output. Figure 9.5 shows the forecasts generated for the logged airline data using these commands.

9.2.3 Model Identification

The identification of the nonseasonal IMA(0, 1, 1) process depends upon the fact that, after taking first differences, the autocorrelations for all lags beyond the first are zero. For the multiplicative $(0, 1, 1) \times (0, 1, 1)_{12}$ process (9.2.1), the only nonzero autocorrelations of $\nabla\nabla_{12} z_t$ are those at lags 1, 11, 12, and 13. In fact, from (9.2.2) the model is viewed as

$$w_t = a_t - \theta a_{t-1} - \Theta a_{t-12} + \theta\Theta a_{t-13}$$

which is an MA model of order 13 for $w_t = \nabla\nabla_{12} z_t$. The autocovariances of w_t are thus given by

$$\begin{aligned}
\gamma_0 &= [1 + \theta^2 + \Theta^2 + (\theta\Theta)^2]\sigma_a^2 = (1 + \theta^2)(1 + \Theta^2)\sigma_a^2 \\
\gamma_1 &= [-\theta - \Theta(\theta\Theta)]\sigma_a^2 = -\theta(1 + \Theta^2)\sigma_a^2 \\
\gamma_{11} &= \theta\Theta\sigma_a^2 \\
\gamma_{12} &= [-\Theta - \theta(\theta\Theta)]\sigma_a^2 = -\Theta(1 + \theta^2)\sigma_a^2 \\
\gamma_{13} &= \theta\Theta\sigma_a^2
\end{aligned} \tag{9.2.18}$$

In particular, these expressions imply that

$$\rho_1 = \frac{-\theta}{1 + \theta^2} \qquad \text{and} \qquad \rho_{12} = \frac{-\Theta}{1 + \Theta^2}$$

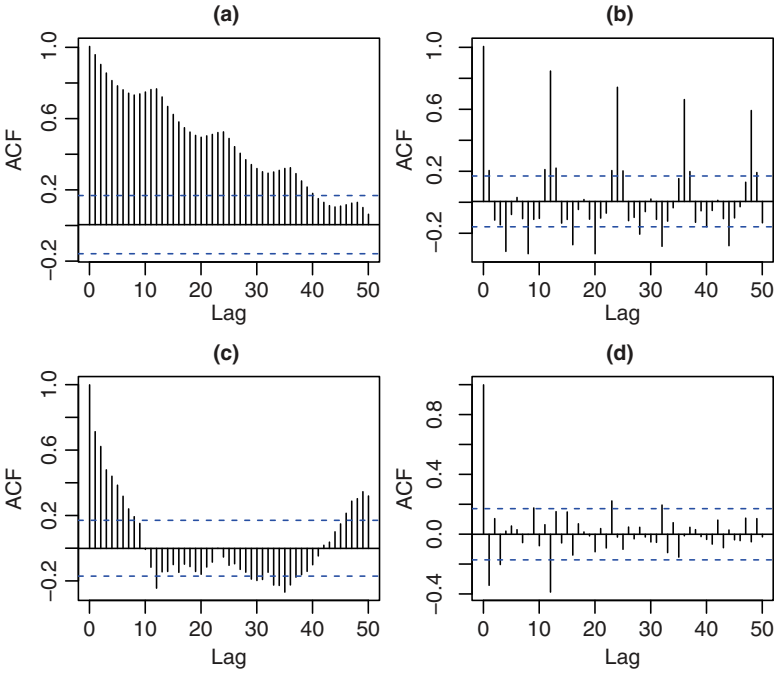


FIGURE 9.6 Estimated autocorrelation function of logged airline data: (a) undifferenced series, (b) first differenced series, (c) seasonally differenced series, and (d) series with regular and seasonal differencing.

so that the value ρ_1 is unaffected by the presence of the seasonal MA factor $(1 - \Theta B^{12})$ in the model (9.2.1), while the value of ρ_{12} is unaffected by the nonseasonal or regular MA factor $(1 - \theta B)$.

Figure 9.6 shows the estimated autocorrelations of the airline data for (a) the logged series, z_t, (b) the logged series differenced with respect to months only, ∇z_t, (c) the logged series differenced with respect to years only, $\nabla_{12} z_t$, and (d) the logged series differenced with respect to months and years, $\nabla\nabla_{12} z_t$. The autocorrelations for z_t are large and fail to die out at higher lags. While simple differencing reduces the correlations in general, a very heavy periodic component remains. This is evidenced particularly by very large correlations at lags 12, 24, 36, and 48. Simple differencing with respect to period 12 results in correlations which are first persistently positive and then persistently negative. By contrast, the differencing $\nabla\nabla_{12}$ markedly reduces correlations throughout.

The autocorrelations of $\nabla\nabla_{12} z_t$ exhibit spikes at lags 1 and 12, compatible with the theoretical autocovariances in (9.2.18) for model (9.2.1). As an alternative, however, the autocorrelations for $\nabla_{12} z_t$ might be viewed as dying out at a slow exponential rate beginning from lag one. Hence, there is also the possibility that $\nabla_{12} z_t$ may follow a nonseasonal ARMA(1, 1) model with ϕ relatively close to one, rather than a nonstationary IMA(0, 1, 1) model as in (9.2.1). However, in practice, the distinction between these two models may not be substantial and the latter model will not be explored further here. The choice between the nonstationary and stationary AR(1) factor could, in fact, be tested using unit root procedures similar to those described in Section 10.1 of the next chapter.

The autocorrelation functions shown in Figure 9.6 was generated in R using the following commands:

```
> library(astsa)
> log.AP=log(ts(seriesG))
> par(mfrow=c(2,2))
> acf(log.AP,50,main='(a)')
> acf(diff(log.AP),50,main='(b)')
> acf(diff(log.AP,12),50,main='(c)')
> acf(diff(diff(log.AP,12)),50,main='(d)')
```

On the assumption that the model is of the form (9.2.1), the variances for the estimated higher lag autocorrelations are approximated by Bartlett's formula (2.1.15), which in this case becomes

$$\text{var}[r_k] \simeq \frac{1 + 2(\rho_1^2 + \rho_{11}^2 + \rho_{12}^2 + \rho_{13}^2)}{n} \qquad k > 13 \tag{9.2.19}$$

Substituting estimated correlations for the ρ's and setting $n = 144 - 13 = 131$ in (9.2.19), where $n = 131$ is the number of differences $\nabla\nabla_{12}z_t$, we obtain a standard error $\hat{\sigma}(r) \simeq 0.11$. The dashed lines shown in Figure 9.6 are approximate two-standard-error limits computed under the assumption that there is no autocorrelation in the series so that var$[r_k]$ = 1/n.

Preliminary Estimates. As with the nonseasonal model, by equating appropriate observed sample correlations to their expected values, approximate values can be obtained for the parameters θ and Θ. On substituting the sample estimates $r_1 = -0.34$ and $r_{12} = -0.39$ in the expressions

$$\rho_1 = \frac{-\theta}{1 + \theta^2} \qquad \rho_{12} = \frac{-\Theta}{1 + \Theta^2}$$

we obtain rough estimates $\hat{\theta} \simeq 0.39$ and $\hat{\Theta} \simeq 0.48$. A table summarizing the behavior of the autocorrelation function for some specimen seasonal models, useful in identification and in obtaining preliminary estimates of the parameters, is given in Appendix A9.1.

9.2.4 Parameter Estimation

Contours of the sum-of-squares function $S(\theta, \Theta)$ for the model (9.2.1) fitted to the airline data are shown in Figure 9.7, together with the appropriate 95% confidence region. The least-squares estimates (LE) are seen to be very nearly $\hat{\theta} = 0.4$ and $\hat{\Theta} = 0.6$. The grid of values for $S(\theta, \Theta)$ was computed using the technique described in Chapter 7. It was shown there that given n observations $\mathbf{w}$ from a linear process defined by

$$\phi(B)w_t = \theta(B)a_t$$

the quadratic form $\mathbf{w}'\mathbf{M}_n\mathbf{w}$, which appears in the exponent of the likelihood, can always be expressed in terms of a sum of squares of the conditional expectation of a's and a quadratic function of the conditional expectation of the $p + q$ initial values

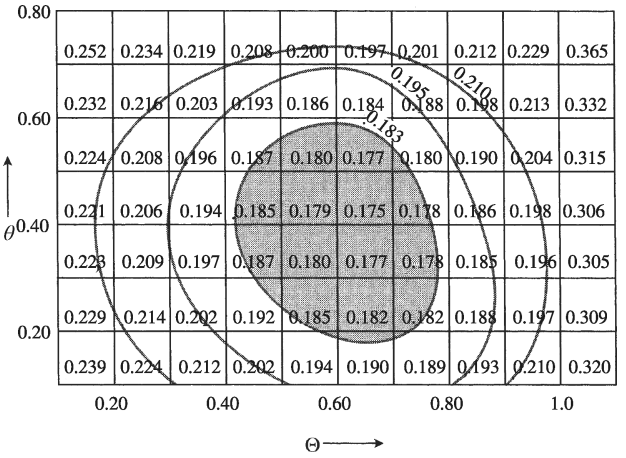


FIGURE 9.7 Contours of $S(\theta, \Theta)$ with shaded 95% confidence region for the model $\nabla\nabla_{12} z_t = (1 - \theta B)(1 - \Theta B^{12}) a_t$ fitted to the airline data.

$\mathbf{e}_* = (w_{1-p}, \ldots, w_0, a_{1-q}, \ldots, a_0)'$, that is,

$$\mathbf{w}'\mathbf{M}_n\mathbf{w} = S(\boldsymbol{\phi}, \theta) = \sum_{t=-\infty}^{n} [a_t]^2 = \sum_{t=1}^{n} [a_t]^2 + [\mathbf{e}_*]'\boldsymbol{\Omega}^{-1}[\mathbf{e}_*]$$

where $[a_t] = [a_t|\mathbf{w}, \boldsymbol{\phi}, \theta]$, $[\mathbf{e}_*] = [\mathbf{e}_*|\mathbf{w}, \boldsymbol{\phi}, \theta]$, and $\text{cov}[\mathbf{e}_*] = \sigma_a^2\boldsymbol{\Omega}$. Furthermore, $S(\boldsymbol{\phi}, \theta)$ plays a central role in the estimation of the parameters $\boldsymbol{\phi}$ and θ from both a sampling theory and a likelihood or Bayesian point of view.

The computation for seasonal models follows precisely the course described in Section 7.1.5 for nonseasonal models. The airline series has $N = 144$ observations. This reduces to n=131 observations after the differencing $w_t = \nabla\nabla_{12} z_t$. The $[a_t]$ in $S(\theta, \Theta)$ can be calculated recursively using an approximate approach that iterates between the forward and backward versions of the $(0, 1, 1) \times (0, 1, 1)_{12}$ model. Alternatively, an exact method discussed in Appendix A7.3 and also used in Section 7.1.5 can be employed. For the present model, this involves first computing the conditional estimates of the a_t, using zero initial values $a_{-12}^0 = a_{-11}^0 = \cdots = a_0^0 = 0$, through a recursive calculation as

$$a_t^0 = w_t + \theta a_{t-1}^0 + \Theta a_{t-12}^0 - \theta\Theta a_{t-13}^0 \qquad t = 1, \ldots, n \tag{9.2.20}$$

Then a backward recursion is used to obtain a series u_t as

$$u_t = a_t^0 + \theta u_{t+1} + \Theta u_{t+12} - \theta\Theta u_{t+13} \qquad t = n, \ldots, 1$$

using zero initial values $u_{n+1} = \cdots = u_{n+13} = 0$. Finally, the exact estimate for the vector of initial values $\mathbf{a}_*' = (a_{-12}, \ldots, a_o)$ is obtained by solving the equations $\mathbf{D}[\mathbf{a}_*] = \mathbf{F}'\mathbf{u}$, as described in (A7.3.12) of Appendix A7.3. Letting $\mathbf{h} = \mathbf{F}'\mathbf{u} = (h_{-12}, h_{-11}, \ldots, h_0)'$, the values h_{-j} are computed as

$$h_{-j} = -(\theta u_{-j+1} + \Theta u_{-j+12} - \theta\Theta u_{-j+13})$$

with $u_{-j} = 0, j \geq 0$. Once the initial values are estimated, the remaining $[a_t]$ values for $t = 1, 2, \ldots, n$ are calculated recursively as in (9.2.20), and hence the exact sum of squares $S(\theta, \Theta) = \sum_{t=-12}^{131} [a_t]^2$ is obtained.

Iterative Calculation of Least-Squares Estimates $\hat{\theta}$, $\hat{\Theta}$. While it is essential to plot sums-of-squares surfaces in a new situation, or whenever difficulties arise, an iterative linearization technique may be used in straightforward situations to supply the least-squares estimates and their approximate standard errors. The procedure has been set out in Section 7.2.1, and no new difficulties arise in estimating the parameters of seasonal models.

For the present example, we can write approximately

$$a_{t,0} = (\theta - \theta_0)x_{t,1} + (\Theta - \Theta_0)x_{t,2} + a_t$$

where

$$x_{t,1} = -\frac{\partial a_t}{\partial \theta}\bigg|_{\theta_0,\Theta_0} \qquad x_{t,2} = -\frac{\partial a_t}{\partial \Theta}\bigg|_{\theta_0,\Theta_0}$$

and where θ_0 and Θ_0 are guessed values and $a_{t,0} = [a_t|\theta_0, \Theta_0]$. As explained and illustrated in Section 7.2.2, the derivatives are most easily computed numerically. Alternatively, the derivatives could be obtained to any degree of accuracy by recursive calculation.

Proceeding this way and using as starting values, the preliminary estimates $\hat{\theta} = 0.39, \hat{\Theta} = 0.48$ obtained above, parameter estimates correct to two decimals are available in three iterations. The estimated variance of the residuals is $\hat{\sigma}_a^2 = 1.34 \times 10^{-3}$. From the inverse of the matrix of sums of squares and products of the x's on the last iteration, the standard errors of the estimates may now be calculated. The least-squares estimates followed by their standard errors are then

$$\hat{\theta} = 0.40 \pm 0.08$$
$$\hat{\Theta} = 0.61 \pm 0.07$$

agreeing closely with the values obtained from the sum-of-squares plot.

Large-Sample Variances and Covariances for the Estimates. As in Section 7.2.6, large-sample formulas for the variances and covariances of the parameter estimates may be obtained. In this case, from the model equation $w_t = a_t - \theta a_{t-1} - \Theta a_{t-12} + \theta\Theta a_{t-13}$, the derivatives $x_{t,1} = -\partial a_t/\partial\theta$ are seen to satisfy

$$x_{t,1} - \theta x_{t-1,1} - \Theta x_{t-12,1} + \theta\Theta x_{t-13,1} + a_{t-1} - \Theta a_{t-13} = 0$$

hence $(1 - \theta B)(1 - \Theta B^{12})x_{t,1} = -(1 - \Theta B^{12})a_{t-1}$, or simply $(1 - \theta B)x_{t,1} = -a_{t-1}$. Thus, using a similar derivation for $x_{t,2} = -\partial a_t/\partial\Theta$, we obtain that

$$x_{t,1} \simeq -(1 - \theta B)^{-1} a_{t-1} = -\sum_{j=0}^{\infty} \theta^j B^j a_{t-1}$$

$$x_{t,2} \simeq -(1 - \Theta B^{12})^{-1} a_{t-12} = -\sum_{i=0}^{\infty} \Theta^i B^{12i} a_{t-12}$$

Therefore, for large samples, the information matrix is

$$\mathbf{I}(\theta, \Theta) = n \begin{bmatrix} (1-\theta^2)^{-1} & \theta^{11}(1-\theta^{12}\Theta)^{-1} \\ \theta^{11}(1-\theta^{12}\Theta)^{-1} & (1-\Theta^2)^{-1} \end{bmatrix}$$

Provided that $|\theta|$ is not close to unity, the off-diagonal term is negligible, and approximate values for the variances and covariances of $\hat{\theta}$ and $\hat{\Theta}$ are

$$V(\hat{\theta}) \simeq n^{-1}(1-\theta^2) \qquad V(\hat{\Theta}) \simeq n^{-1}(1-\Theta^2)$$
$$\operatorname{cov}[\hat{\theta}, \hat{\Theta}] \simeq 0 \tag{9.2.21}$$

In the present example, substituting the values $\hat{\theta} = 0.40$, $\hat{\Theta} = 0.61$, and $n = 131$, we obtain

$$V(\hat{\theta}) \simeq 0.0064 \qquad V(\hat{\Theta}) \simeq 0.0048$$

and

$$\sigma(\hat{\theta}) \simeq 0.08 \qquad \sigma(\hat{\Theta}) \simeq 0.07$$

which, to this accuracy, are identical with the values obtained directly from the iteration. It is also interesting to note that the parameter estimates $\hat{\theta}$ and $\hat{\Theta}$, associated with months and years, respectively, are virtually uncorrelated.

Parameter Estimation in R. The parameters of the model

$$\nabla\nabla_{12} z_t = w_t = (1-\theta B)(1-\Theta B^{12})a_t$$

can be estimated in R using the command sarima(log.AP,p,d,q,P,D,Q,S=12) in the astsa package as demonstrated below. The resulting estimates of the two parameters θ and Θ are 0.40 and 0.56, respectively, with corresponding standard errors of 0.09 and 0.07. The full likelihood function, including the determinant, is used for parameter estimation, which accounts for the difference between the parameter estimates derived above and those obtained in R. Also, in viewing the output, it should be noted that R defines the moving average operators with positive signs, in contrast to the negative signs used in this text.

```
  > library(astsa)
  > log.AP=log(ts(seriesG))
  > m1.AP=sarima(log.AP, 0,1,1,0,1,1,S=12)
  > m1.AP % Retrieves output from file

OUTPUT:
Call:
stats:arima(x=xdata,order=c(p,d,q),seasonal= list(order=c(P,D,Q),
period=S),optim.control=list(trace=trc,REPORT=1,reltol=tol))

Coefficients:
          ma1     sma1
      -0.4018  -0.5569
s.e.   0.0896   0.0731

sigma^2 estimated as 0.001348: log likelihood=244.7,  aic=-483.4
```

9.2.5 Diagnostic Checking

Before proceeding further, we check the adequacy of fit of the model by examining the residuals from the fitted model.

Autocorrelation Checks. The standardized residuals calculated from the fitted model and the estimated autocorrelations of the residuals are shown in Figure 9.8. The figure is generated as part of the output from the estimation command ''sarima'' in R. The residual autocorrelations do not present evidence of any lack of fit, since none of the values fall outside the approximate two-standard-error limits of 0.18. This conclusion is also supported by the p values of the portmanteau statistics $\tilde{Q} = n(n+2)\sum_{k=1}^{K} r_k^2(\hat{a})/(n-k)$ which are shown for different values of K in the last part of the graph.

Periodogram Check. The cumulative periodogram (see Section 8.2.5) for the residuals is shown in Figure 9.9. The Kolmogorov–Smirnov 5 and 25% probability limits, which as we have seen in Section 8.2.5 supply a very rough guide to the significance of apparent deviations, fail in this instance to indicate any significant departure from the assumed model.

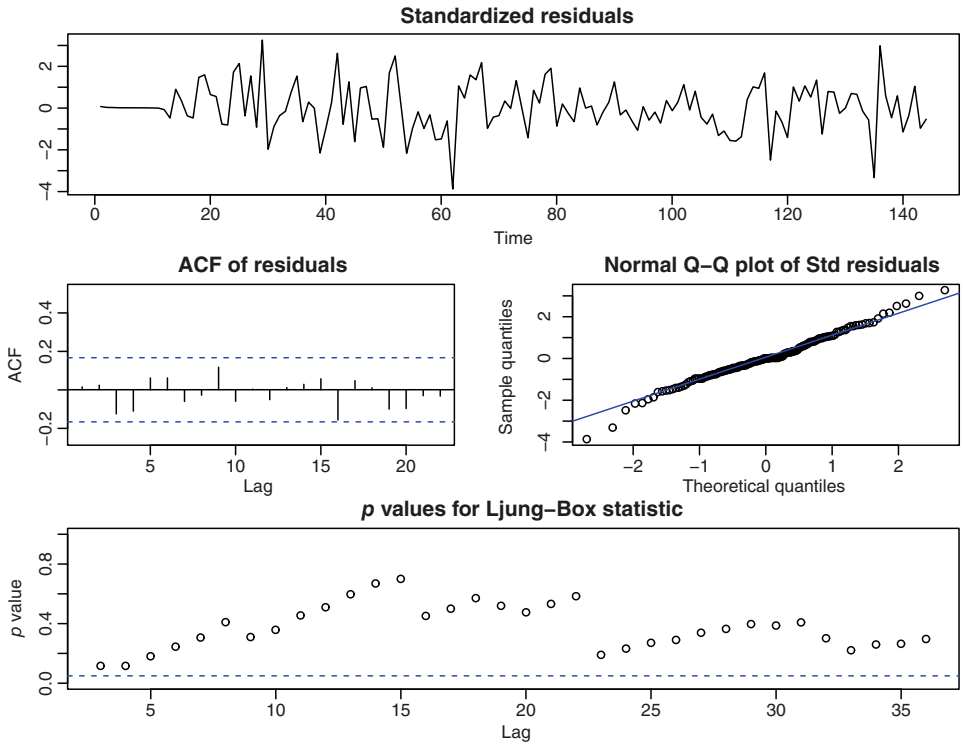


FIGURE 9.8 Diagnostic checks on the residuals from the fitted model.

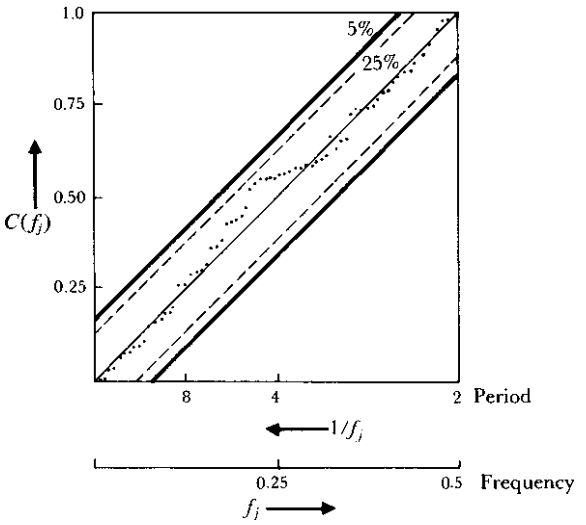


FIGURE 9.9 Cumulative periodogram check on residuals from the model $\nabla\nabla_{12}z_t = (1 - 0.40B)(1 - 0.61B^{12})a_t$, fitted to the airline data.

9.3 SOME ASPECTS OF MORE GENERAL SEASONAL ARIMA MODELS

9.3.1 Multiplicative and Nonmultiplicative Models

In previous sections, we discussed methods of dealing with seasonal time series, and in particular, we examined an example of a multiplicative model. We have seen how this model can provide a useful representation with remarkably few parameters. It now remains to study other seasonal models of this kind, and insofar as new considerations arise, the associated processes of identification, estimation, diagnostic checking, and forecasting.

Suppose, in general, that we have a seasonal effect associated with period s. Then, the general class of multiplicative models may be typified in the manner shown in Figure 9.10. In the multiplicative model, it is assumed that the ''between periods'' development of the series is represented by some model

$$\Phi_P(B^s)\nabla_s^D z_{r,m} = \Theta_Q(B^s)\alpha_{r,m}$$

while ''within periods'' the α's are related by

$$\phi_p(B)\nabla^d \alpha_{r,m} = \theta_q(B)a_{r,m}$$

Obviously, we could change the order in which we considered the two types of models and in either case obtain the general multiplicative model

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D z_{r,m} = \theta_q(B)\Theta_Q(B^s)a_{r,m} \tag{9.3.1}$$

where $a_{r,m}$ is a white noise process with zero mean. In practice, the usefulness of models such as (9.3.1) depends on how far it is possible to parameterize actual time series parsimoniously in these terms. In fact, experience has shown that this is possible for a variety of seasonal time series coming from widely different sources. While the multiplicative model (9.2.1)

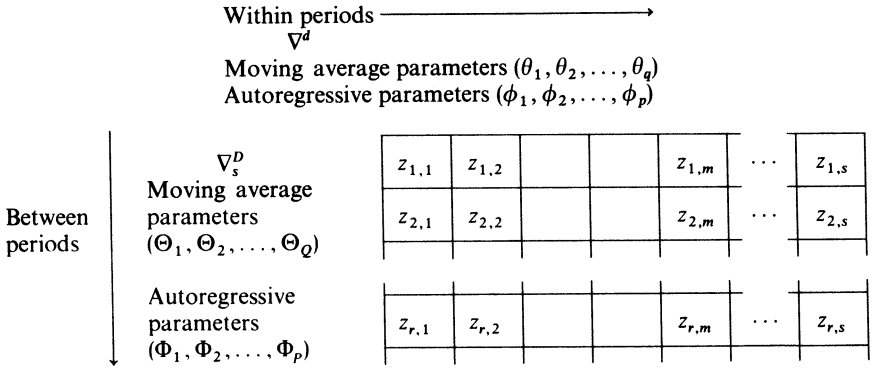


FIGURE 9.10 Two-way table for multiplicative seasonal model.

has been found to fit many time series, other models of the form (9.3.1) have also been found to be useful in practise.

It is not possible to obtain a completely adequate fit with multiplicative models for all series. One modification that is sometimes useful allows the mixed moving average operator to be nonmultiplicative. By this is meant that we replace the operator $\theta_q(B)\Theta_Q(B^s)$ on the right-hand side of (9.3.1) by a more general moving average operator $\theta^*_{q^*}(B)$. Alternatively, or in addition, it may be necessary to replace the autoregressive operator $\phi_p(B)\Phi_P(B^s)$ on the left by a more general autoregressive operator $\phi^*_{p^*}(B)$. Some examples of nonmultiplicative models are given in Appendix A9.1. These are numbered 4, 4a, 5, and 5a.

In those cases where a nonmultiplicative model is found necessary, experience suggests that the best-fitting multiplicative model can provide a good starting point from which to construct a better nonmultiplicative model. The situation is reminiscent of the problems encountered in analyzing two-way analysis of variance tables, where additivity of row and column constants may or may not be an adequate assumption, but may provide a good point of departure.

Our general strategy for relating multiplicative or nonmultiplicative models to data is that which we have already discussed and illustrated in some detail in Section 9.2. Using the autocorrelation function for guidance:

1. The series is differenced with respect to ∇ and/or ∇_s, so as to produce stationarity.
2. By inspection of the autocorrelation function of the suitably differenced series, a tentative model is selected.
3. From the values of appropriate autocorrelations of the differenced series, preliminary estimates of the parameters are obtained. These can be used as starting values in the search for the least-squares or maximum likelihood estimates.
4. After fitting, the diagnostic checking process applied to the residuals either may lead to the acceptance of the tentative model or, alternatively, may suggest ways in which it can be improved, leading to refitting and repetition of the diagnostic checks.

As a few practical guidelines for model specification, we note that for seasonal series the order of seasonal differencing D needed would almost never be greater than one, and especially for monthly series with $s = 12$, the orders P and Q of the seasonal AR and MA

operators $\Phi(B^s)$ and $\Theta(B^s)$ would rarely need to be greater than 1. This is particularly so when the series length of available data is not sufficient to warrant a more complicated form of model with $P > 1$ or $Q > 1$.

9.3.2 Model Identification

A useful aid in model identification is the list in Appendix A9.1 that gives the autocovariance structure of $w_t = \nabla^d \nabla_s^D z_t$ for a number of simple seasonal models. This list makes no claim to be comprehensive. However, it does include some frequently encountered models, and the reader should have no difficulty in discovering the characteristics of others that may seem useful. It should be emphasized that rather simple models, such as models 1 and 2 in the appendix, have provided adequate representations for many seasonal series.

Since the multiplicative seasonal ARMA models for the differences $w_t = \nabla\nabla_s z_t$ may be viewed as special forms of ARMA models with orders $p + sP$ and $q + sQ$, their autocovariances can be derived from the principles of Chapter 3, as was done in the previous section for the MA model $w_t = a_t - \theta a_{t-1} - \Theta a_{t-12} + \theta\Theta a_{t-13}$. For further illustration, consider the model

$$(1 - \phi B)w_t = (1 - \Theta B^s)a_t$$

which is a special form of ARMA model with AR order 1 and MA order s. First, since the ψ weights for this model for w_t satisfy $\psi_j - \phi\psi_{j-1} = 0, j = 1, \ldots, s-1$, we have $\psi_j = \phi^j, j = 1, \ldots, s-1$, as well as $\psi_s = \phi^s - \Theta$ and $\psi_j = \phi\psi_{j-1}, j > s$. It is then easy to see that the autocovariances for w_t will satisfy

$$\begin{aligned} \gamma_0 &= \phi\gamma_1 + \sigma_a^2(1 - \Theta\psi_s) \\ \gamma_j &= \phi\gamma_{j-1} - \sigma_a^2\Theta\psi_{s-j} \qquad j = 1, \ldots, s \\ \gamma_j &= \phi\gamma_{j-1} \qquad j > s \end{aligned} \tag{9.3.2}$$

Solving the first two equations for γ_0 and γ_1, we obtain

$$\gamma_0 = \sigma_a^2 \frac{1 - \Theta(\phi^s - \Theta) - \phi^s\Theta}{1 - \phi^2} = \sigma_a^2 \frac{1 + \Theta^2 - 2\phi^s\Theta}{1 - \phi^2}$$

$$\gamma_1 = \sigma_a^2 \frac{\phi[1 - \Theta(\phi^s - \Theta)] - \phi^{s-1}\Theta}{1 - \phi^2} = \sigma_a^2 \frac{\phi(1 + \Theta^2 - \phi^s\Theta) - \phi^{s-1}\Theta}{1 - \phi^2}$$

with $\gamma_j = \phi\gamma_{j-1} - \sigma_a^2\Theta\phi^{s-j} = \phi^j\gamma_0 - \sigma_a^2\Theta\phi^{s-j}(1 - \phi^{2j})/(1 - \phi^2), j = 1, \ldots, s$ and $\gamma_j = \phi\gamma_{j-1} = \phi^{j-s}\gamma_s, j > s$. Hence, in particular, for monthly data with $s = 12$ and $|\phi|$ not too close to one, the autocorrelation function ρ_j for this process will behave, for low lags, similarly to that of a regular AR(1) process, $\rho_j \simeq \phi^j$ for small j, while the value of ρ_{12} will be close to $-\Theta/(1 + \Theta^2)$.

A fact of considerable utility in deriving autocovariances of a multiplicative process is that for such a process, the autocovariance generating function (3.1.11) is the product or the generating functions of the components. Thus, in (9.3.1) if the component models for $\nabla^d z_t$ and $\nabla_s^D \alpha_t$,

$$\phi_p(B)\nabla^d z_t = \theta_q(B)\alpha_t \qquad \Phi_P(B^s)\nabla_s^D \alpha_t = \Theta_Q(B)a_t$$

have autocovariance generating function $\gamma(B)$ and $\Gamma(B^s)$, the autocovariance generating function for $w_t = \nabla^d \nabla_s^D z_t$ in (9.3.1) is

$$\gamma(B)\Gamma(B^s)$$

Another point to be remembered is that it may be useful to parameterize more general models in terms of their departures from related multiplicative forms in a manner now illustrated.

The three-parameter nonmultiplicative operator

$$1 - \theta_1 B - \theta_{12} B^{12} - \theta_{13} B^{13} \tag{9.3.3}$$

employed in models 4 and 5 in the appendix may be written as

$$(1 - \theta_1 B)(1 - \theta_{12} B^{12}) - k B^{13}$$

where

$$k = \theta_1 \theta_{12} - (-\theta_{13})$$

An estimate of k that was large compared with its standard error would indicate the need for a nonmultiplicative model in which the value of θ_{13} is not tied to the values of θ_1 and θ_{12}. On the other hand, if k is small, then on writing $\theta_1 = \theta$, $\theta_{12} = \Theta$, the model approximates the multiplicative $(0, 1, 1) \times (0, 1, 1)_{12}$ model.

9.3.3 Parameter Estimation

No new problems arise in the estimation of the parameters of general seasonal models. The unconditional sum of squares is computed quite generally by the methods set out fully in Section 7.1.5 and illustrated further in Section 9.2.4. As always, contour plotting can illuminate difficult situations. In well-behaved situations, iterative least-squares with numerical determination of derivatives yield rapid convergence to the least-squares estimates, together with approximate variances and covariances of the estimates. Recursive procedures can be derived in each case, which allow direct calculation of derivatives, if desired.

Large-Sample Variances and Covariances of the Estimates. The large-sample information matrix $\mathbf{I}(\boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\Phi}, \boldsymbol{\Theta})$ is given by evaluating $E[\mathbf{X}'\mathbf{X}]$, where, as in Section 7.2.6, $\mathbf{X}$ is the $n \times (p + q + P + Q)$ matrix of derivatives with reversed signs. Thus, for the general multiplicative model

$$a_t = \theta^{-1}(B)\Theta^{-1}(B^s)\phi(B)\Phi(B^s)w_t$$

where $w_t = \nabla^d \nabla_s^D z_t$, the required derivatives are

$$\frac{\partial a_t}{\partial \theta_i} = \theta^{-1}(B)B^i a_t \qquad \frac{\partial a_t}{\partial \Theta_i} = \Theta^{-1}(B^s)B^{si} a_t$$

$$\frac{\partial a_t}{\partial \phi_j} = -\phi^{-1}(B)B^j a_t \qquad \frac{\partial a_t}{\partial \Phi_j} = -\Phi^{-1}(B^s)B^{sj} a_t$$

Approximate variances and covariances of the estimates are obtained as before, by inverting the matrix $\mathbf{I}(\boldsymbol{\phi}, \theta, \boldsymbol{\Phi}, \boldsymbol{\Theta})$.

9.3.4 Eventual Forecast Functions for Various Seasonal Models

We now consider the characteristics of the eventual forecast functions for a number of seasonal models. For a seasonal model with single periodicity s, the eventual forecast function at origin t for lead time l is the solution of the difference equation

$$\phi(B)\Phi(B^s)\nabla^d \nabla_s^D \hat{z}_t(l) = 0$$

Table 9.2 shows this solution for various choices of the difference equation; also shown is the number of initial values on which the behavior of the forecast function depends.

In Figure 9.11, the behavior of each forecast function is illustrated for $s = 4$. It will be convenient to regard the lead time $l = rs + m$ as referring to a forecast r years and m quarters ahead. In the diagram, an appropriate number of initial values (required to start the forecast off and indicated by bold dots) has been set arbitrarily and the course of the forecast

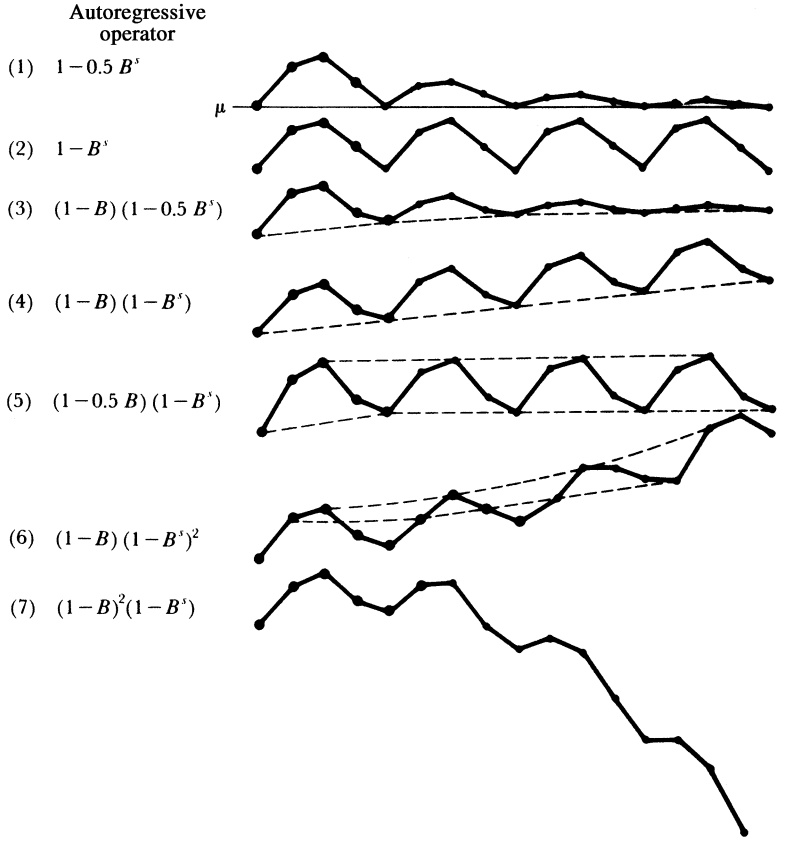


FIGURE 9.11 Behavior of the seasonal forecast function for various choices of the general seasonal autoregressive operator.

TABLE 9.2 Eventual Forecast Functions for Various Generalized Autoregressive Operators

Generalized Autoregressive Operator	Eventual Forecast Function $\hat{z}(r, m)^a$	Number of Initial Values on which Forecast Function Depends
(1) $1 - \Phi B^s$	$\mu + (b_{0,m} - \mu)\Phi^r$	s
(2) $1 - B^s$	$b_{0,m}$	s
(3) $(1 - B)(1 - \Phi B^s)$	$b_0 + (b_{0,m} - b_0)\Phi^r + b_1 \left\{ \frac{1 - \Phi^r}{1 - \Phi} \right\}$	$s + 1$
(4) $(1 - B)(1 - B^s)$	$b_{0,m} + b_1 r$	$s + 1$
(5) $(1 - \phi B)(1 - B^s)$	$b_{0,m} + b_1 \phi^{m-1} \left\{ \frac{1 - \phi^{sr}}{1 - \phi^s} \right\}$	$s + 1$
(6) $(1 - B)(1 - B^s)^2$	$b_{0,m} + b_{1,m} r + \frac{1}{2} b_2 r(r - 1)$	$2s + 1$
(7) $(1 - B)^2(1 - B^s)$	$b_{0,m} + [b_1 + (m - 1)b_2]r + \frac{1}{2} b_2 sr(r - 1)$	$s + 2$

[a]Coefficients b are all adaptive and depend upon forecast origin t.

function traced to the end of the fourth period. When the difference equation involves an autoregressive parameter, its value has been set equal to 0.5.

The constants $b_{0,m}$, b_1, and so on, appearing in the solutions in Table 9.2, should strictly be indicated by $b_{0,m}^{(t)}$, $b_1^{(t)}$, and so on, since each one depends on the origin t of the forecast, and these constants are adaptively modified each time the origin changes. The superscript t has been omitted temporarily to simplify notation.

The operator labeled (1) in Table 9.2 is stationary, with the model containing a fixed mean μ. It is autoregressive in the seasonal pattern, and the forecast function decays with each period, approaching closer and closer to the mean.

Operator (2) in Table 9.2 is nonstationary in the seasonal component. The forecasts for a particular quarter are linked from year to year by a polynomial of degree 0. Thus, the basic forecast of the seasonal component is exactly reproduced in forecasts of future years.

Operator (3) in Table 9.2 is nonstationary with respect to the basic time interval but stationary in the seasonal component. Operator (3) in Figure 9.11 shows the general level of the forecast approaching asymptotically the new level

$$b_0 + \frac{b_1}{1 - \Phi}$$

where, at the same time, the superimposed predictable component of the stationary seasonal effect dies out exponentially.

In Table 9.2, operator (4) is the limiting case of the operator (3) as Φ approaches unity. The operator is nonstationary with respect to both the basic time interval and the periodic component. The basic initial forecast pattern is reproduced, as is the incremental yearly increase. This is the type of forecast function given by the multiplicative $(0, 1, 1) \times (0, 1, 1)_{12}$ process fitted to the airline data.

Operator (5) is nonstationary in the seasonal pattern but stationary with respect to the basic time interval. The pattern approaches exponentially an asymptotic basic pattern

$$\hat{z}_t(\infty, m) = b_{0,m} + \frac{b_1 \phi^{m-1}}{1 - \phi^s}$$

Operator (6) is nonstationary in both the basic time interval and the seasonal component. An overall quadratic trend occurs over years, and a particular kind of modification occurs in the seasonal pattern. Individual quarters not only have their own level $b_{0,m}$ but also their own rate of change of level $b_{1,m}$. Therefore, when this kind of forecast function is appropriate, we can have a situation where, for example, as the lead time is increased, the difference in summer over spring sales can be forecast to increase from one year to the next, while at the same time, the difference in autumn over summer sales can be forecast to decrease.

In Table 9.2, operator (7) is again nonstationary in both the basic time interval and in the seasonal component, and there is again a quadratic tendency over years with the incremental changes in the forecasts from one quarter to the next changing linearly. However, in this case, they are restricted to have a common *rate* of change.

9.3.5 Choice of Transformation

It is particularly true for seasonal models that the weighted averages of previous data values, which comprise the forecasts, may extend far back into the series. Care is therefore needed in choosing a transformation in terms or which a parsimonious linear model will closely apply over a sufficient stretch of the series. Simple graphical analysis can often suggest such a transformation. Thus, an appropriate transformation may be suggested by determining in what metric the amplitude of the seasonal component is roughly independent of the level of the series. To illustrate how a data-based transformation may be chosen more exactly, denote the *untransformed* airline data by x, and let us assume that some power transformation [$z = x^\lambda$ for $\lambda \neq 0$, $z = \ln(x)$ for $\lambda = 0$] may be needed to make the model (9.2.1) appropriate. Then, as suggested in Section 4.1.3, the approach of Box and Cox (1964) may be followed, and the maximum likelihood value obtained by fitting the model to $x^{(\lambda)} = (x^\lambda - 1)/\lambda \dot{x}^{\lambda-1}$ for various values of λ, and choosing the value of λ that results in the smallest residual sum of squares s_λ. In this expression, $\dot{x}$ is the geometric mean of the series x, and it is easily shown that $x^{(0)} = \dot{x} \ln(x)$. For the airline data, we find

λ	S_λ	λ	S_λ	λ	S_λ
−0.4	13,825.5	−0.1	11,627.2	0.2	11,784.3
−0.3	12,794.6	0.0	11,458.1	0.3	12,180.0
−0.2	12,046.0	0.1	11,554.3	0.4	12,633.2

The maximum likelihood value is thus close to $\lambda = 0$, confirming the appropriateness of the logarithmic transformation for the airline series.

9.4 STRUCTURAL COMPONENT MODELS AND DETERMINISTIC SEASONAL COMPONENTS

A traditional method to represent a seasonal time series has been to decompose the series into trend, seasonal, and noise components, as $z_t = T_t + S_t + N_t$, where the trend T_t and seasonal component S_t are represented as deterministic functions of time using polynomial and sinusoidal functions, respectively. However, as noted in Section 9.1.1, the deterministic nature of the trend and seasonal components limits the applicability of these models.

Subsequently, models that permit random variation in the trend and seasonal components, referred to as *structural component* models, have become increasingly popular for time series modeling (e.g., Harvey, 1989; Harvey and Todd, 1983; Gersch and Kitagawa, 1983; Kitagawa and Gersch, 1984; Hillmer and Tiao, 1982; and Durbin and Koopman, 2012). We discuss these models briefly in the following sections.

9.4.1 Structural Component Time Series Models

In general, a univariate structural component time series model is one in which an observed series z_t is formulated as the sum of unobservable component or "signal" time series. Although the components are unobservable and cannot be uniquely specified, they will usually have direct meaningful interpretation, such as representing the seasonal behavior or the long-term trend of an economic time series or a physical signal that is corrupted by measurement noise in the engineering setting. Thus, the models attempt to describe the main features of the series as well as provide a basis for forecasting, signal extraction, seasonal adjustments, and other applications. For a monthly time series, the trend T_t might be assumed to follow a simple random walk model or some extension such as the ARIMA(0, 1, 1) model $(1 - B)T_t = (1 - \theta B)a_t$, or the AIRMA(0, 2, 2) model $(1 - B)^2 T_t = (1 - \theta_1 B - \theta_2 B^2)a_t$, while the seasonal component might be specified as a ''seasonal random walk'' $(1 - B^{12})S_t = b_t$, where a_t and b_t are independent white noise processes.

An appeal of this structural modeling approach, especially for seasonal adjustments and signal extraction, is that Kalman filtering and smoothing methods based on state-space formulations of the model, as discussed in Section 5.5, can be employed. The exact likelihood function can be constructed based on the state-space model form, as described in Section 7.4, and used for parameter estimation. The Kalman filtering and smoothing procedures can then be used to obtain estimates of the unobservable component series such as the trend $\{T_t\}$ and seasonal $\{S_t\}$ components, which are now included as elements within the state vector Y_t in the general state-space model (5.5.4) and (5.5.5).

Basic Structural Model. As a specific illustration, consider the *basic structural model* (BSM) for seasonal time series with period s as formulated by Harvey (1989). The model is defined by $z_t = T_t + S_t + \varepsilon_t$, where T_t follows the ''local linear trend model'' defined by

$$T_t = T_{t-1} + \beta_{t-1} + \eta_t \quad \beta_t = \beta_{t-1} + \xi_t \tag{9.4.1}$$

and S_t follows the ''dummy variable seasonal component model'' defined by

$$(1 + B + B^2 + \ldots + B^{s-1})S_t = \omega_t \tag{9.4.2}$$

where η_t, ξ_t, ω_t, and ε_t are mutually uncorrelated white noise processes with zero means and variances σ_η^2, σ_ξ^2, σ_ω^2, and σ_ε^2, respectively.

This local linear trend model is a stochastic generalization of the deterministic linear trend $T_t = \alpha + \beta t$, where α and β are constants. In (9.4.1), the effect of the random disturbance η_t is to allow the level of the trend to shift up and down, while ξ_t allows the slope to change. As special limiting cases, if $\sigma_\xi^2 = 0$, then $\beta_t = \beta_{t-1}$ and so β_t is a fixed constant β for all t and the trend follows the random walk with drift $(1 - B)T_t = \beta + \eta_t$. If $\sigma_\eta^2 = 0$ in addition, then (9.4.1) collapses to the deterministic model $T_t = T_{t-1} + \beta$ or $T_t = \alpha + \beta t$. The seasonal component model (9.4.2) requires the seasonal effects S_t to sum to zero over

s consecutive values of a seasonal period, subject to a random disturbance with mean zero which allows the seasonal effects to change gradually over time. Again, a special limiting case of deterministic seasonal components with a fixed seasonal pattern about an average of zero, $S_t = S_{t-s}$ with $S_t + S_{t-1} + \ldots + S_{t-s+1} = 0$, occurs when $\sigma_\omega^2 = 0$. Thus, one attraction of a model such as (9.4.1) and (9.4.2) is that it generalizes a regression-type in which the trend is represented by a fixed straight line and the seasonality by fixed seasonal effects using indicator variables, by allowing the trend and seasonality to vary over time, and still yields the deterministic components as special limiting cases.

We illustrate the state-space representation of the model (9.4.1) and (9.4.2) for the case of quarterly time series with $s = 4$. For this, we define the state vector as

$$Y_t = (T_t, \beta_t, S_t, S_{t-1}, S_{t-2})'$$

and let $\mathbf{a}_t = (\eta_t, \xi_t, \omega_t)'$. Then we have the transition equation

$$Y_t = \begin{bmatrix} T_t \\ \beta_t \\ S_t \\ S_{t-1} \\ S_{t-2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} T_{t-1} \\ \beta_{t-1} \\ S_{t-1} \\ S_{t-2} \\ S_{t-3} \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \eta_t \\ \xi_t \\ \omega_t \end{bmatrix} \tag{9.4.3}$$

or $\boldsymbol{Y}_t = \boldsymbol{\Phi} Y_{t-1} + \Psi \mathbf{a}_t$, together with the observation equation $z_t = T_t + S_t + \varepsilon_t \equiv [1 \quad 0 \quad 1 \quad 0 \quad 0] Y_t + \varepsilon_t = \mathbf{H} Y_t + \varepsilon_t$. Hence, the variance component parameters of the structural model can be estimated by maximum likelihood methods using the state-space representation and innovations form of the likelihood function, as discussed in Sections 5.5 and 7.4. Once these estimates are obtained, the desired optimal smoothed estimates $\hat{T}_{t|n} = E[T_t | z_1, \ldots, z_n]$ and $\hat{S}_{t|n} = E[S_t | z_1, \ldots, z_n]$ of the trend and seasonal components based on the observed series $z_1, \ldots, z_n$ can readily be obtained by applying the Kalman filtering and smoothing techniques to the state-space representation.

Relation to ARIMA Model. It should be noted from general results of Appendix A4.3 that structural models such as the BSM have an equivalent ARIMA model representation, which is sometimes referred to as its *reduced form* in this context. For instance, the process T_t defined by the local linear trend model (9.4.1) satisfies

$$(1 - B)^2 T_t = (1 - B)\beta_{t-1} + (1 - B)\eta_t = \xi_{t-1} + (1 - B)\eta_t$$

It follows from Appendix A4.3.1 that $\xi_{t-1} + (1 - B)\eta_t$ can be represented as an MA(1) process $(1 - \theta B)a_t$, so that $(1 - B)^2 T_t = (1 - \theta B)a_t$ and T_t has the ARIMA(0, 2, 1) model as a reduced form. For another illustration, consider $z_t = T_t + S_t + N_t$, where it is assumed that

$$(1 - B)T_t = (1 - \theta_T B)a_t \quad (1 - B^{12})S_t = (1 - \Theta_s B^{12})b_t$$

and $N_t = c_t$ is white noise. Then, we have

$$\begin{aligned} &(1 - B)(1 - B^{12})z_t \\ &\quad = (1 - B^{12})(1 - \theta_T B)a_t + (1 - B)(1 - \Theta_s B^{12})b_t + (1 - B)(1 - B^{12})c_t \end{aligned}$$

and according to the developments in Appendix A4.3, the right-hand-side expression above can be represented as the MA model $(1 - \theta_1 B - \theta_{12} B^{12} - \theta_{13} B^{13})\epsilon_t$, where ϵ_t is white noise, since the right-hand side will have nonzero autocovariances only at the lags 0, 1, 11, 12, and 13. Under additional structure, the MA operator could have the multiplicative form, but in general we see that the foregoing structural model, $z_t = T_t + S_t + N_t$, has an equivalent ARIMA model representation as

$$(1 - B)(1 - B^{12})z_t = (1 - \theta_1 B - \theta_{12} B^{12} - \theta_{13} B^{13})\epsilon_t$$

Example: Airline Data. Harvey (1989, Sec. 4.5) reported results of maximum likelihood estimation of the BSM defined by (9.4.1) and (9.4.2) for the logged monthly airline passenger data, using the data period from 1949 to 1958. The ML estimates were such that $\hat{\sigma}_\xi^2 = 0$ and $\hat{\sigma}_\omega^2$ was very small relative to $\hat{\sigma}_\eta^2$ and $\hat{\sigma}_\epsilon^2$. The zero estimate $\hat{\sigma}_\xi^2 = 0$ implies that the model (9.4.1) for the trend T_t reduces to the random walk with *constant* drift, $(1 - B)T_t = \beta + \eta_t$, while the seasonal component model is $(1 + B + \ldots + B^{11})S_t = \omega_t$. Differencing the series z_t thus implies that

$$\begin{aligned} w_t = (1 - B)(1 - B^{12})z_t &= (1 - B)(1 - B^{12})T_t + (1 - B)(1 - B^{12})S_t \\ &\quad + (1 - B)(1 - B^{12})\epsilon_t \\ &= (1 - B^{12})\eta_t + (1 - B)^2\omega_t + (1 - B)(1 - B^{12})\epsilon_t \end{aligned}$$

It readily follows that the autocovariances of the differenced series $w_t = \nabla\nabla_{12} z_t$ for this model are

$$\begin{aligned} \gamma_0 &= 2\sigma_\eta^2 + 6\sigma_\omega^2 + 4\sigma_\epsilon^2 \\ \gamma_1 &= -4\sigma_\omega^2 - 2\sigma_\epsilon^2 \\ \gamma_2 &= \sigma_\omega^2 \\ \gamma_{11} &= \sigma_\epsilon^2 = \gamma_{13} \\ \gamma_{12} &= -\sigma_\eta^2 - 2\sigma_\epsilon^2 \end{aligned} \tag{9.4.4}$$

and $\gamma_j = 0$ otherwise. In particular, these give the autocorrelations

$$\rho_1 = -\frac{\sigma_\epsilon^2 + 2\sigma_\omega^2}{2\sigma_\epsilon^2 + \sigma_\eta^2 + 3\sigma_\omega^2}$$

$$\rho_{12} = \frac{2\sigma_\epsilon^2 + \sigma_\eta^2}{2(2\sigma_\eta^2 + \sigma_\eta^2 + 3\sigma_\omega^2)}$$

and $\rho_{11} = \rho_{13} = \sigma_\epsilon^2/[2(2\sigma_\epsilon^2 + \sigma_\eta^2 + 3\sigma_\omega^2)]$.

The autocorrelations calculated using estimates of the variance components given in Table 4.5.3 of Harvey (1989) are shown in Table 9.3 for the logged airline data. Also shown in Table 9.3 are the autocorrelations for the differenced series $w_t = \nabla\nabla_{12} z_t$ in the seasonal $(0, 1, 1) \times (0, 1, 1)_{12}$ model. These were calculated from (9.2.18) using the parameter estimates $\hat{\theta} = 0.396$, $\hat{\Theta} = 0.614$, and $\hat{\sigma}_a^2 = 1.34 \times 10^{-3}$ reported in Section 9.2.4. Table 9.3 shows a close agreement between the two sets of autocorrelations. Hence, for the logged airline data, both modeling approaches provide very similar representations of the basic trend and seasonality in the series.

TABLE 9.3 Comparison of the Autocorrelations of $w_t = \nabla\nabla_{12} z_t$ for the Basic Structural Model and the Seasonal ARIMA Model $(0,1,1) \times (0,1,1)_{12}$ for Logged Airline Data

Model	ρ_1	ρ_2	ρ_{11}	ρ_{12}	ρ_{13}
Basic structural model	−0.26	0.00	0.12	−0.49	0.12
ARIMA $(0,1,1) \times (0,1,1)_{12}$	−0.34	0.00	0.15	−0.45	0.15

9.4.2 Deterministic Seasonal and Trend Components and Common Factors

Now in some applications, particularly in the physical sciences, a seasonal or trend component could be nearly deterministic. For example, suppose the seasonal component can be approximated as

$$S_t = \beta_0 + \sum_{j=1}^{6} \left[\beta_{1j} \cos\left(\frac{2\pi jt}{12}\right) + \beta_{2j} \sin\left(\frac{2\pi jt}{12}\right) \right]$$

where the β coefficients are constants. We note that this can be viewed as a special case of the previous examples, since S_t satisfies $(1 + B + B^2 + \ldots + B^{11})S_t = 12\beta_0$ or $(1 - B^{12})S_t = 0$. Now, ignoring the trend component for the present and assuming that $z_t = S_t + N_t$, where $(1 - B^{12})S_t = 0$ and $N_t = (1 - \theta_N B)a_t$, say, we find that z_t follows the seasonal ARIMA model

$$(1 - B^{12})z_t = (1 - \theta_N B)(1 - B^{12})a_t$$

However, we now notice the presence of a *common factor* of $1 - B^{12}$ in both the generalized AR operator and the MA operator of this model; equivalently, we might say that $\Theta = 1$ for the seasonal MA operator $\Theta(B^{12}) = (1 - \Theta B^{12})$. This is caused by and, in fact, is indicative of the presence of the deterministic seasonal component S_t in the original form of the model.

In general, the presence of deterministic seasonal or trend components in the structure of a time series z_t is characterized by common factors of $(1 - B^s)$ or $(1 - B)$ in the generalized AR operator and the MA operator of the model. We can state the result more formally as follows. Suppose that z_t follows the model $\varphi(B)z_t = \theta_0 + \theta(B)a_t$, and the operators $\varphi(B)$ and $\theta(B)$ contain a common factor $G(B)$, so that $\varphi(B) = G(B)\varphi_1(B)$ and $\theta(B) = G(B)\theta_1(B)$. Hence, the model is

$$G(B)\varphi_1(B)z_t = \theta_0 + G(B)\theta_1(B)a_t \tag{9.4.5}$$

Let $G(B) = 1 - g_1 B - \cdots - g_r B^r$ and suppose that this polynomial has roots $G_1^{-1}, \ldots, G_r^{-1}$ which are distinct. Then, the common factor $G(B)$ can be canceled from both sides of the above model, but a term of the form $\sum_{i=1}^{r} c_i G_i^t$ needs to be added. Thus, the model (9.4.5) can be expressed in the equivalent form as

$$\varphi_1(B)z_t = c_{0t} + \sum_{i=1}^{r} c_i G_i^t + \theta_1(B)a_t \tag{9.4.6}$$

where the c_i are constants, and c_{0t} is a term that satisfies $G(B)c_{0t} = \theta_0$. Modifications of the result for the case where some of the roots G_i^{-1} are repeated are straightforward.

Thus, it is seen that an equivalent representation for the above model is

$$\varphi_1(B)z_t = x_t + \theta_1(B)a_t$$

where x_t is a deterministic function of t that satisfies $G(B)x_t = \theta_0$. Note that roots in $G(B)$ corresponding to ''stationary factors,'' such that $|G_i| < 1$, will make a contribution to the component x_t that is only transient and so negligible, and hence these terms may be ignored. Thus, only those factors whose roots correspond to nonstationary ''differencing'' and other ''simplifying'' operators, such as $(1 - B)$ and $(1 - B^s)$, with roots $|G_i| = 1$ need to be included in the deterministic component x_t. These common factors will, of course, give rise to deterministic functions in x_t that are of the form of polynomials, sine and cosine functions, and products of these, depending on the roots of the common factor $G(B)$.

Examples. For a few simple examples, the model $(1 - B)z_t = \theta_0 + (1 - B)\theta_1(B)a_t$ has an equivalent form $z_t = c_1 + \theta_0 t + \theta_1(B)a_t$, which occurs upon cancellation of the common factor $(1 - B)$, while the model $(1 - \sqrt{3}B + B^2)z_t = \theta_0 + (1 - \sqrt{3}B + B^2)\theta_1(B)a_t$ has an equivalent model form as $z_t = c_0 + c_1 \cos(2\pi t/12) + c_2 \sin(2\pi t/12) + \theta_1(B)a_t$, where $(1 - \sqrt{3} + 1)c_0 = \theta_0$.

Detection of a deterministic component such as x_t above in a time series z_t may occur after an ARIMA model is estimated and common or near-common factors are identified. Hence, the ARIMA time series methodology, in a sense, can indicate when a time series may contain deterministic seasonal or trend components. The presence of a deterministic component is characterized by a factor in the MA operator with roots on, or very near to the unit circle, which correspond to a differencing factor that has been applied to the original series in the formulation of the ARIMA model. When this situation occurs, the series is sometimes said to be ''over-differenced''. Formal tests for the presence of a unit root in the MA operator implying the presence of a deterministic component, have been developed by Saikkonen and Luukkonen (1993), Leybourne and McCabe (1994), and Tam and Reinsel (1997, 1998), among others. These tests can also be viewed as tests for unit roots in the generalized AR operator $\varphi(B)$ in the sense that if one performs the differencing and then concludes that the MA operator does not have a unit root, then the unit root in the AR operator is supported.

Deterministic components implied by the cancellation of factors could be estimated directly by a combination of regression models and ARIMA time series methods, as will be discussed in Section 9.5. An additional consequence of the presence of deterministic factors for forecasting is that at least some of the coefficients $b_j^{(t)}$ in the general forecast function $\hat{z}_t(l)$ for z_{t+l} in (5.3.3) will not be adaptive but will be deterministic (fixed) constants. Results such as those described above concerning the relationship between common factors in the generalized AR and the MA operators of ARIMA models and the presence of deterministic polynomial and sinusoidal components have been discussed by Abraham and Box (1978), Harvey (1981), and Bell (1987).

9.4.3 Estimation of Unobserved Components in Structural Models

A common problem of interest for the structural model is the estimation of the unobservable series S_t from values of the observed series z_t. We suppose that S_t and z_t are stationary processes with zero means and autocovariance functions $\gamma_s(l) = E[S_t S_{t+1}]$ and $\gamma_z(l) = E[z_t z_{t+1}]$, and cross-covariance function $\gamma_{sz}(l) = E[S_t z_{t+1}]$. Then, specifically, suppose

we observe the values $z_t, t \le \tau$, and want to determine the linear filter

$$\hat{S}_t = \sum_{u=0}^{\infty} v_u^{(\tau)} z_{\tau-u} \equiv v^{(\tau)}(B) z_\tau \tag{9.4.7}$$

of $\{z_t\}$ such that the value $\hat{S}_t$ is close to S_t in the mean square error sense, that is, $E[(S_t - \hat{S}_t)^2]$ is a minimum among all possible linear filters. A typical model for which this problem arises is the ''signal extraction'' model, in which there is a signal S_t of interest, but what is observed is a noise-corrupted version of the signal so that

$$z_t = S_t + N_t$$

where N_t is a noise component. The problem then is to estimate values of the signal series S_t given values on the observed series z_t. Often, the filtering and smoothing algorithms for the state-space model, as discussed in Section 5.5.3, can be applied to this situation. However, while these algorithms are computationally attractive in practice, explicit expressions for the coefficients $v_u^{(\tau)}$ in (9.4.7) cannot usually be obtained directly from the state-space algorithms. These expressions can be derived more readily in the ''classical'' approach, which assumes that an infinite extent of observations is available for filtering or smoothing. This section provides a brief overview of some classical filtering and smoothing results that can be used to study the coefficients in (9.4.7). Typically, from a practical point of view, the classical results provide a good approximation to exact filtering and smoothing results that are based on a finite sample of observations $z_1, \ldots, z_n$.

Smoothing and Filtering for Time Series. We suppose that $\{z_t\}$ has the infinite MA representation

$$z_t = \psi(B) a_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}$$

where the a_t are white noise with variance σ_a^2. Also, let $g_{zs}(B) = \sum_{j=-\infty}^{\infty} \gamma_{zs}(j) B^j$ be the cross-covariance generating function between z_t and S_t. Then, it can be derived (e.g., Whittle, 1963, Chapters 5 and 6; Priestley, 1981, Chapter 10) that the optimal linear filter for the estimate $\hat{S}_t = \sum_{u=0}^{\infty} v_u^{(\tau)} z_{\tau-u} = v^{(\tau)}(B) z_\tau$, where $v^{(\tau)}(B) = \sum_{u=0}^{\infty} v_u^{(\tau)} B^u$, is given by

$$v^{(\tau)}(B) = \frac{1}{\sigma_a^2 \psi(B)} \left[\frac{B^{\tau-t} g_{zs}(B)}{\psi(B^{-1})} \right]_+ \tag{9.4.8}$$

Here, for a general operator $v(B) = \sum_{j=-\infty}^{\infty} v_j B^j$, the notation $[v(B)]_+$ is used to denote $\sum_{j=0}^{\infty} v_j B^j$.

To derive the result (9.4.8) for the optimal linear filter, note that, since $z_t = \psi(B) a_t$, the linear filter can be expressed as

$$\hat{S}_t = v^{(\tau)}(B) z_\tau = v^{(\tau)}(B) \psi(B) a_\tau = h^{(\tau)}(B) a_\tau$$

where $h^{(\tau)}(B) = v^{(\tau)}(B)\psi(B) = \sum_{j=0}^{\infty} h_j^{(\tau)} B^j$. Then, we can determine the coefficients $h_j^{(\tau)}$ to minimize the mean squared error $E[(S_t - \hat{S}_t)^2] = E[(S_t - \sum_{j=0}^{\infty} h_j^{(\tau)} a_{\tau-j})^2]$. Since the $\{a_t\}$ are mutually uncorrelated, by standard linear least-squares arguments the values of

the coefficients that minimize this mean squared error are

$$h_j^{(\tau)} = \frac{\text{cov}[a_{\tau-j}, S_t]}{\text{var}[a_{\tau-j}]} = \frac{\gamma_{as}(j+t-\tau)}{\sigma_a^2} \qquad j \geq 0$$

Hence, the optimal linear filter is

$$h^{(\tau)}(B) = \frac{1}{\sigma_a^2} \sum_{j=0}^{\infty} \gamma_{as}(j+t-\tau) B^j = \frac{1}{\sigma_a^2} [B^{\tau-t} g_{as}(B)]_+ \tag{9.4.9}$$

where $g_{as}(B)$ denotes the cross-covariance generating function between a_t and S_t. Also, note that $\gamma_{zs}(j) = \text{cov}[\sum_{i=0}^{\infty} \psi_i a_{t-i}, S_{t+j}] = \sum_{i=0}^{\infty} \psi_i \gamma_{as}(i+j)$, so it follows that $g_{zs}(B) = \psi(B^{-1}) g_{as}(B)$. Therefore, the optimal linear filter in (9.4.9) is $h^{(\tau)}(B) = (1/\sigma_a^2)[B^{\tau-t} g_{zs}(B)/\psi(B^{-1})]_+$, and, hence, the optimal filter in terms of $\hat{S}_t = \upsilon^{(\tau)}(B) z_\tau$ is $\upsilon^{(\tau)}(B) = h^{(\tau)}(B)/\psi(B)$, which yields the result (9.4.8). The mean squared error of the optimal filter, since $\hat{S}_t = \sum_{j=0}^{\infty} h_j^{(\tau)} a_{\tau-j}$, is easily seen from the above derivation to be

$$E[(S_t - \hat{S}_t)^2] = E[S_t^2] - E[\hat{S}_t^2] = \text{var}[S_t] - \sigma_a^2 \sum_{j=0}^{\infty} \{h_j^{(\tau)}\}^2$$

In the smoothing case where $\tau = +\infty$, that is, we estimate S_t based on the infinite record of observations z_u, $-\infty < u < \infty$, by a linear filter $\hat{S}_t = \sum_{u=-\infty}^{\infty} \upsilon_u z_{t-u} = \upsilon(B) z_t$, the result (9.4.8) for the optimal filter reduces to

$$\upsilon(B) = \frac{g_{zs}(B)}{g_{zz}(B)} = \frac{g_{zs}(B)}{\sigma_a^2 \psi(B) \psi(B^{-1})} \tag{9.4.10}$$

For the signal extraction problem, we have $z_t = S_t + N_t$, where it is usually assumed that the signal $\{S_t\}$ and the noise process $\{N_t\}$ are independent. Thus, in this case we have $g_{zs}(B) = g_{ss}(B)$, and so in the smoothing case $\tau = +\infty$, we have $\upsilon(B) = g_{ss}(B)/g_{zz}(B)$ or $\upsilon(B) = g_{ss}(B)/[g_{ss}(B) + g_{nn}(B)]$.

Smoothing Relations for the Signal Plus Noise or Structural Components Model. The preceding results can be applied specifically to the model $z_t = S_t + N_t$, where we assume that the signal process $\{S_t\}$ and the noise process $\{N_t\}$ are independent and satisfy ARMA models, $\phi_s(B) S_t = \theta_s(B) b_t$ and $\phi_n(B) N_t = \theta_n(B) c_t$, where b_t and c_t are independent white noise processes with variances σ_b^2 and σ_c^2. It follows from Appendix A4.3 that the observed process z_t also satisfies an ARMA model $\phi(B) z_t = \theta(B) a_t$, where $\phi(B) = \phi_s(B) \phi_n(B)$, assuming no common factors in the AR operators. It then follows that the optimal linear ''smoother'' $\hat{S}_t = \sum_{u=-\infty}^{\infty} \upsilon_u z_{t-u} = \upsilon(B) z_t$ of S_t, based on the infinite set of values z_u, $-\infty < u < \infty$, has a filter given by

$$\upsilon(B) = \frac{g_{ss}(B)}{g_{zz}(B)} = \frac{\sigma_b^2 \phi(B) \phi(B^{-1}) \theta_s(B) \theta_s(B^{-1})}{\sigma_a^2 \theta(B) \theta(B^{-1}) \phi_s(B) \phi_s(B^{-1})} \tag{9.4.11}$$

In practice, since the series S_t and N_t are not observable, the models for S_t and N_t would usually not be known. Thus, the optimal filter would not be known in practice. However, by

developing a model for the observed series z_t and placing certain restrictions on the form of the models for S_t and N_t beyond those implied by the model for z_t, e.g., by assuming N_t is white noise with the largest possible variance, one may obtain reasonable approximations to the optimal filter $v(B)$. While optimal smoothing results, such as (9.4.10), have been derived for the case where S_t and N_t are stationary processes, Bell (1984) showed that the results extend to the nonstationary case under reasonable assumptions for the nonstationary signal S_t and noise N_t processes.

As noted earlier, an alternative to the classical filtering approach in the structural components models is to express the model in state-space form and use Kalman filtering and smoothing techniques to estimate the components, as illustrated, for example, by Kitagawa and Gersch (1984). For further discussion of this approach, see also Harvey (1989) and Durbin and Koopman (2012).

Seasonal Adjustments. The filtering and smoothing methods described above have applications to seasonal adjustments of economic and business time series (i.e., estimating and removing the seasonal component from the series). Approaches of the type discussed were used by Hillmer and Tiao (1982) to decompose a time series uniquely into mutually independent seasonal, trend, and irregular components. A model-based approach to seasonal adjustments was also considered by Cleveland and Tiao (1976). Seasonal adjustments are commonly performed by statistical agencies in the U.S. and elsewhere, and the methods used have received considerable attention in the literature. For an overview and further discussion, see, for example, Ghysels and Osborn (2001, Chapter 4), Bell and Sotiris (2010), Chu, Tiao, and Bell (2012), and Bell, Chu, and Tiao (2012).

9.5 REGRESSION MODELS WITH TIME SERIES ERROR TERMS

The previous discussion of deterministic components in Section 9.4.2 motivates consideration of time series models that include regression terms such as deterministic sine and cosine functions to represent seasonal behavior or stochastic predictor variables, in addition to a serially correlated ''noise'' or error term. We will assume that the noise series N_t follows a stationary ARMA process; otherwise, differencing may be need to be considered. Thus, letting w_t be a ''response'' series of interest, we wish to represent w_t in terms of its linear dependence on k explanatory or predictor time series variables $x_{t1}, \ldots, x_{tk}$ as follows:

$$w_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + N_t \qquad t = 1, \ldots, n \tag{9.5.1}$$

where the errors N_t follow a zero-mean ARMA(p, q) model, $\phi(B)N_t = \theta(B)a_t$. The traditional linear regression model was reviewed briefly in Appendix A7.2. Using similar notations with $\mathbf{w} = (w_1, \ldots, w_n)'$, $\boldsymbol{N} = (N_1, \ldots, N_n)'$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$, the model (9.5.1) may be written in matrix form as $\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{N}$, and with covariance matrix $\mathbf{V} = \text{cov}[\boldsymbol{N}]$. In the standard regression model, the errors N_t are assumed to be uncorrelated with common variance σ_N^2, so that $\mathbf{V} = \sigma_N^2\mathbf{I}$, and the ordinary least squares (LS) estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{w}$ has well-known properties such as $\text{cov}[\hat{\boldsymbol{\beta}}] = \sigma_N^2(\mathbf{X}'\mathbf{X})^{-1}$. However, in the case of autocorrelated errors, this property no longer holds and the ordinary least-squares estimator has covariance matrix

$$\text{cov}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Moreover, standard inference procedures based on the t and F distributions are no longer valid due to the lack of independence.

When $\text{cov}[\boldsymbol{N}] = \mathbf{V} \neq \sigma_N^2 \mathbf{I}$, the best linear unbiased estimator of $\boldsymbol{\beta}$ is the *generalized least-squares* (GLS) estimator given by

$$\hat{\boldsymbol{\beta}}_G = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{w} \tag{9.5.2}$$

which has $\text{cov}[\hat{\boldsymbol{\beta}}_G] = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$. The estimator $\hat{\boldsymbol{\beta}}_G$ is the best linear unbiased estimator in the sense that $\text{var}[\boldsymbol{c}'\hat{\boldsymbol{\beta}}_G]$ is a minimum among all possible linear unbiased estimators of $\boldsymbol{\beta}$, for every arbitrary k-dimensional vector of constants $c' = (c_1, \ldots, c_k)$; in particular, $\text{var}[c'\hat{\boldsymbol{\beta}}_G] \leq \text{var}[c'\hat{\boldsymbol{\beta}}]$ holds relative to the ordinary LS estimator $\hat{\boldsymbol{\beta}}$. It follows that $\hat{\boldsymbol{\beta}}_G$ in (9.5.2) is the estimate of $\boldsymbol{\beta}$ obtained by minimizing the generalized sum of squares $S(\boldsymbol{\beta}; \mathbf{V}) = (\mathbf{w} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})$ with $\mathbf{V}$ given. This estimator also corresponds to the maximum likelihood estimator under the assumption of normality of the errors when the covariance matrix $\mathbf{V}$ is known. Of course, a practical limitation to use of the GLS estimate $\hat{\boldsymbol{\beta}}_G$ is that the ARMA noise model and its parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ needed to determine $\mathbf{V}$ must be known, which is typically not true in practice. This motivates an iterative model building and estimation procedure discussed below.

9.5.1 Model Building, Estimation, and Forecasting Procedures for Regression Models

When a regression model is fitted to time series data, one should always consider the possibility that the errors are autocorrelated. Often, a reasonable approach to identify an appropriate model for the error N_t is first to obtain the least-squares estimate $\hat{\boldsymbol{\beta}}$, and then compute the corresponding regression model residuals

$$\hat{N}_t = w_t - \hat{\beta}_1 x_{t1} - \hat{\beta}_2 x_{t2} - \cdots - \hat{\beta}_k x_{tk} \tag{9.5.3}$$

This residual series can be examined by the usual time series methods, such as inspection of its sample ACF and PACF, to identify an appropriate ARMA model for N_t. This would typically be adequate to specify a tentative model for the error term N_t, especially when the explanatory variables x_{ti} are deterministic functions such as sine and cosine functions, or polynomial terms. In such cases, it is known (e.g., Anderson, 1971, Section 10.2) that the least-squares estimator for $\boldsymbol{\beta}$ is an asymptotically efficient estimator relative to the best linear estimator. In addition, it is known that the sample autocorrelations and partial autocorrelations calculated using the residuals from the preliminary least-squares fit are asymptotically equivalent to those obtained from the actual noise series N_t (e.g., Anderson, 1971, Section 10.3; Fuller, 1996, Section 9.3).

Hence, the complete model that we consider is

$$w_t = \boldsymbol{x}_t'\boldsymbol{\beta} + N_t \qquad \phi(B)(1-B)^d N_t = \theta(B)a_t \qquad t = 1, \ldots, n \tag{9.5.4}$$

where $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tk})'$. Estimates of all parameters can be obtained by maximum likelihood methods. The resulting estimate for $\boldsymbol{\beta}$ has the GLS form

$$\hat{\boldsymbol{\beta}}_G = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{w}$$

but where V is replaced by the estimate $\hat{\mathbf{V}}$ obtained from the MLEs $\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q$ of the ARMA parameters for N_t. Also, $\text{cov}[\hat{\boldsymbol{\beta}}_G] \simeq (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$. The estimation can be

performed iteratively, alternating between calculation of $\hat{\boldsymbol{\beta}}_G$ for given estimates $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$, and reestimation of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, given $\hat{\boldsymbol{\beta}}_G$ and the estimated noise series $\hat{N}_t = w_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_G$.

Transformed Model. With the ARMA model specified for N_t, the computation of the generalized least-squares estimator of $\boldsymbol{\beta}$ can be carried out in a computationally convenient manner as follows. Let $\mathbf{P}'$ be a lower triangular matrix, such that $\mathbf{P}'\mathbf{V}\mathbf{P} = \sigma_a^2\mathbf{I}$, that is, $\mathbf{V}^{-1} = \mathbf{P}\mathbf{P}'/\sigma_a^2$. Then, as in Appendix A7.2.5, the GLS estimator can be obtained from the transformed regression model

$$\mathbf{P}'\mathbf{w} = \mathbf{P}'\mathbf{X}\boldsymbol{\beta} + \mathbf{P}'\boldsymbol{N} \tag{9.5.5}$$

or $\mathbf{w}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{a}$, where the transformed variables are $\mathbf{w}^* = \mathbf{P}'\mathbf{w}$, $\mathbf{X}^* = \mathbf{P}'\mathbf{X}$, and $\boldsymbol{a} = \mathbf{P}'\boldsymbol{N}$. Since the covariance matrix of the error vector $\boldsymbol{a} = \mathbf{P}'\boldsymbol{N}$ in the transformed model is

$$\text{cov}[\boldsymbol{a}] = \mathbf{P}'\text{cov}[\boldsymbol{N}]\mathbf{P} = \mathbf{P}'\mathbf{V}\mathbf{P} = \sigma_a^2\mathbf{I}$$

we can now use ordinary least-squares to estimate $\boldsymbol{\beta}$ in the transformed model. That is, the GLS estimator of $\boldsymbol{\beta}$ is obtained as the LS estimator in terms of the transformed variables $\mathbf{w}^*$ and $\mathbf{X}^*$ as

$$\hat{\boldsymbol{\beta}}_G = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{w}^* \qquad \text{with} \qquad \text{cov}[\hat{\boldsymbol{\beta}}_G] = \sigma_a^2(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} \tag{9.5.6}$$

However, since the ARMA parameters for N_t are not known in practice, one must still iterate between the computation of $\hat{\boldsymbol{\beta}}_G$ using the current estimates of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ to form the transformation matrix $\hat{\mathbf{P}}'$, and estimation of the ARMA parameters based on $\hat{N}_t = w_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_G$ constructed from the current estimate of $\boldsymbol{\beta}$. The computational procedure used to determine the exact sum-of-squares function for the specified ARMA model will also essentially determine the nature of the transformation matrix $\mathbf{P}'$. For instance, the innovations algorithm described in Section 7.4 gives the sum of squares for an ARMA model as $S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sigma_a^2\mathbf{w}'\mathbf{V}^{-1}\mathbf{w} = \mathbf{e}'\mathbf{D}^{-1}\mathbf{e}$, where $\mathbf{e} = \mathbf{G}^{-1}\mathbf{L}_\phi\mathbf{w}$ and $\mathbf{D} = \text{diag}(v_1, \ldots, v_n)$, and $\mathbf{G}$ and $\mathbf{L}_\phi$ are specific lower triangular matrices. Hence, the innovations algorithm can be viewed as providing the transformation matrix $\mathbf{P}' = \mathbf{D}^{-1/2}\mathbf{G}^{-1}\mathbf{L}_\phi$ such that $\mathbf{w}^* = \mathbf{D}^{-1/2}\mathbf{G}^{-1}\mathbf{L}_\phi\mathbf{w} \equiv \mathbf{P}'\mathbf{w}$ has covariance matrix of the ''standard'' form

$$\text{cov}[\mathbf{w}^*] = \mathbf{P}'\text{cov}[\mathbf{w}]\mathbf{P} = \mathbf{D}^{-1/2}\mathbf{G}^{-1}\mathbf{L}_\phi\,\text{cov}[\mathbf{w}]\mathbf{L}_\phi'\mathbf{G}'^{-1}\mathbf{D}^{-1/2} = \sigma_a^2\mathbf{I}$$

Therefore, the required transformed variables $\mathbf{w}^* = \mathbf{P}'\mathbf{w}$ and $\mathbf{X}^* = \mathbf{P}'\mathbf{X}$ in (9.5.6) can be obtained by applying the innovations algorithm recursive calculations (e.g.,(7.4.9)) to the series $\mathbf{w} = (w_1, \ldots, w_n)'$ and to each column, $\mathbf{x}_i' = (x_{1i}, \ldots, x_{ni})'$, $i = 1, \ldots, k$, of the matrix $\mathbf{X}$.

Example. We take the simple example of an AR(1) model, $(1 - \phi B)N_t = a_t$, for the noise N_t, for illustration. Then the covariance matrix $\mathbf{V}$ of $\boldsymbol{N}$ has (i, j)th element given by $\gamma_{i-j} = \sigma_a^2\phi^{|i-j|}/(1 - \phi^2)$. The $n \times n$ matrix $\mathbf{P}'$ such that $\mathbf{P}'\mathbf{V}\mathbf{P} = \sigma_a^2\mathbf{I}$ has its (1, 1) element equal to $(1 - \phi^2)^{1/2}$, its remaining diagonal elements equal to 1, its first subdiagonal elements equal to $-\phi$, and all remaining elements equal to zero. Hence, the transformed variables are $w_1^* = (1 - \phi^2)^{1/2}w_1$ and $w_t^* = w_t - \phi w_{t-1}, t = 2, 3, \ldots, n$, and similarly for the transformed explanatory variables x_{ti}^*. In effect, with AR(1) errors, the original model

(9.5.1) has been transformed by applying the AR(1) operator $(1 - \phi B)$ throughout the equation to obtain

$$w_t - \phi w_{t-1} = \beta_1(x_{t1} - \phi x_{t-1,1}) + \beta_2(x_{t2} - \phi x_{t-1,2}) + \cdots + \beta_k(x_{tk} - \phi x_{t-1,k}) + a_t \tag{9.5.7}$$

or, equivalently, $w_t^* = \beta_1 x_{t1}^* + \beta_2 x_{t2}^* + \cdots + \beta_k x_{tk}^* + a_t$, where the errors a_t now are uncorrelated. Thus, ordinary least-squares applies to the transformed regression model, and the resulting estimator is the same as the GLS estimator in the original regression model.

Generalization of the transformation procedure to higher order AR models is straightforward. Apart from special treatment for the initial p observations, the transformed variables are $w_t^* = \phi(B)w_t = w_t - \phi_1 w_{t-1} - \cdots - \phi_p w_{t-p}$ and $x_{ti}^* = \phi(B)x_{ti} = x_{ti} - \phi_1 x_{t-1,i} - \cdots - \phi_p x_{t-p,i}$, $i = 1, \ldots, k$. The exact form of the transformation in the case of mixed ARMA models will be more complicated [an approximate form is $w_t^* \simeq \theta^{-1}(B)\phi(B)w_t$, and so on] but can be determined through the same procedure as is used to construct the exact sum-of-squares function for the ARMA model.

Forecasting. Forecasting for regression models with time series errors is straightforward when future values $x_{t+l,i}$ of the explanatory variables are known, as would be the case for deterministic functions such as sine and cosine functions, for example. Then, based on forecast origin t, the lead l forecast of

$$w_{t+l} = \beta_1 x_{t+l,1} + \cdots + \beta_k x_{t+l,k} + N_{t+l}$$

based on past values through time t, is

$$\hat{w}_t(l) = \beta_1 x_{t+l,1} + \beta_2 x_{t+l,2} + \cdots + \beta_k x_{t+l,k} + \hat{N}_t(l) \tag{9.5.8}$$

where $\hat{N}_t(l)$ is the usual l-step-ahead forecast of N_{t+l} from the ARMA(p, q) model, $\phi(B)N_t = \theta(B)a_t$, based on the past values of the noise series N_t. The forecast error is

$$e_t(l) = w_{t+l} - \hat{w}_t(l) = N_{t+l} - \hat{N}_t(l) = \sum_{i=0}^{l-1} \psi_i a_{t+l-i} \tag{9.5.9}$$

with $V(l) = \text{var}[e_t(l)] = \sigma_a^2 \sum_{i=0}^{l-1} \psi_i^2$, just the forecast error and its variance from the ARMA model for the noise series N_t, where the ψ_i are the coefficients in $\psi(B) = \phi^{-1}(B)\theta(B)$ for the noise model.

Example. For the model

$$w_t = \beta_0 + \beta_1 \cos\left(\frac{2\pi t}{12}\right) + \beta_2 \sin\left(\frac{2\pi t}{12}\right) + N_t$$

where $(1 - \phi B)N_t = a_t$, the forecasts are

$$\hat{w}_t(l) = \beta_0 + \beta_1 \cos\left[\frac{2\pi(t+l)}{12}\right] + \beta_2 \sin\left[\frac{2\pi(t+l)}{12}\right] + \hat{N}_t(l)$$

with $\hat{N}_t(l) = \phi^l N_t$. Note that these forecasts are similar in functional form to those that would be obtained in an ARMA(1, 3) model (with zero constant term) for the series

$(1 - B)(1 - \sqrt{3}B + B^2)w_t$, except that the β coefficients in the forecast function for the regression model case are deterministic, not adaptive, as was noted at the end of Section 9.4.2.

In practice, estimates of $\boldsymbol{\beta}$ and the time series model parameters would be used to obtain the estimated noise series $\hat{N}_t$ from which forecasts of future values would be made. The effect of parameter estimation errors on the variance of the corresponding forecast error was investigated by Baillie (1979) for regression models with autoregressive errors, generalizing a similar study by Yamamoto (1976) conducted for pure autoregressive models.

More detailed discussions of regression analysis with time series errors are given by Harvey and Phillips (1979) and by Wincek and Reinsel (1986), who also consider the possibility of missing data. A state-space approach with associated Kalman filtering calculations, as discussed in Section 7.4, can be employed for the regression model with time series errors, and this corresponds to one particular choice for the transformation matrix $\mathbf{P}'$ in the above discussion. A specific application of the use of regression models with time series errors to model calendar effects in seasonal time series was given by Bell and Hillmer (1983), while Reinsel and Tiao (1987) used regression models with time series errors to model atmospheric ozone data for estimation of trends.

One common application of regression models for seasonal time series is where seasonality can be modeled as a *deterministic seasonal mean* model. Then, for monthly seasonal data, for example, we might consider a model of the form

$$z_t = \beta_0 + \sum_{j=1}^{6} \left[\beta_{1j} \cos\left(\frac{2\pi jt}{12}\right) + \beta_{2j} \sin\left(\frac{2\pi jt}{12}\right) \right] + N_t \tag{9.5.10}$$

where N_t is modeled as an ARIMA process. As an example, Reinsel and Tiao (1987) consider the time series z_t of monthly averages of atmospheric total column ozone measured at the station Aspendale, Australia, for the period from 1958 to 1984. This series is highly seasonal, and so in terms of ARIMA modeling, the seasonal differences $w_t = (1 - B^{12})z_t$, were considered. Based on the sample ACF and PACF of w_t, the following model was specified and estimated,

$$(1 - 0.48B - 0.22B^2)(1 - B^{12})z_t = (1 - 0.99\ B^{12})a_t$$

and the model was found to be adequate. We see that this model contains a near-common seasonal difference factor $(1 - B^{12})$, and consequently, it is equivalent to the model that contains a deterministic seasonal component, $z_t = S_t + N_t$, of exactly the form given in (9.5.10), and where N_t follows the AR(2) model, $(1 - 0.48B - 0.22B^2)N_t = a_t$. This model was estimated using regression methods similar to those discussed above.

Sometimes, the effects of a predictor variable $\{x_t\}$ on z_t are not confined to a single time period t, but the effects are more dynamic over time and are ''distributed'' over several time periods. With a single predictor variable, this would lead to models of the form

$$z_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \beta_3 x_{t-2} + \cdots + N_t$$

where N_t might be an ARIMA process. For parsimonious modeling, the regression coefficients β_i can be formulated as specific functions of a small number or unknown parameters. Such models are referred to as *transfer function* models or *dynamic regression* models, and will be considered in detail in Chapters 11 and 12.

Remark. Note that regression models with autocorrelated errors can be fitted to data using the arima() function in R with an argument xreg added to account for regression terms; type help(arima) for details. For further discussion, see also Venables and Ripley (2002). An alternative available in the MTS package of R is the function tfm1() that can be used to fit a regression model with a single input variable X_t. We demonstrate the use of this function to develop a dynamic regression model in Chapter 12. A similar function which allows for two input series is also available in the MTS package of R.

9.5.2 Restricted Maximum Likelihood Estimation for Regression Models

A detracting feature of the maximum likelihood estimator (MLE) of the ARMA parameters in the linear regression model (9.5.1) is that the MLE can produce a nonnegligible bias for small to moderate sample sizes. This bias could have significant impact on inferences of the regression parameters $\boldsymbol{\beta}$ based on the GLS estimation, through the approximation $\text{cov}[\hat{\boldsymbol{\beta}}_G] \simeq (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$, where $\hat{\mathbf{V}}$ involves the ML estimates of the ARMA parameters. One ''preventive'' approach for reducing the bias is to use the *restricted maximum likelihood* (REML) estimation procedure, also known as the *residual maximum likelihood* estimation procedure, for the ARMA model parameters.

The REML method has been popular and commonly used in the estimation of variance components in mixed-effects linear models. For ARMA models, this procedure has been used by Cooper and Thompson (1977) and Tunnicliffe Wilson (1989), among others. Cheang and Reinsel (2000, 2003) compared the ML and REML estimation methods, and bias characteristics in particular, for time series regression models with AR and ARMA noise (as well as fractional ARIMA noise, see Section 10.4). They established approximate bias characteristics for these estimators, and confirmed empirically that REML typically reduces the bias substantially over ML estimation. Consequently, the REML approach leads to more accurate inferences about the regression parameters.

The REML estimation of the parameters in the ARMA noise models differs from the ML estimation in that it explicitly takes into account the fact that the regression parameters $\boldsymbol{\beta}$ are unknown and must be estimated (i.e., estimation of ARMA parameters relies on the residuals $\hat{N}_t = w_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_G$ rather than the ''true'' noise $N_t = w_t - \boldsymbol{x}_t'\boldsymbol{\beta}$). In the REML estimation method, the estimates of $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ_a^2 are determined so as to maximize the restricted likelihood function. This is the likelihood function based on observation of the ''residual vector'' of error contrasts $\mathbf{u} = \mathbf{H}'\mathbf{w}$ only, whose distribution is free of the regression parameters $\boldsymbol{\beta}$, rather than the likelihood based on the 'full' vector of observations $\mathbf{w}$. Here, $\mathbf{H}'$ is any $(n-k) \times n$ full rank matrix such that $\mathbf{H}'\mathbf{X} = \mathbf{0}$, so the regression effects are eliminated in $\mathbf{u} = \mathbf{H}'\mathbf{w}$ and its distribution is free of the parameters $\boldsymbol{\beta}$.

Assuming normality, the distribution of $\mathbf{w}$ is normal with mean vector $E(\mathbf{w}) = \mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\text{cov}[\mathbf{w}] = \mathbf{V}$, which we write as $\mathbf{V} = \sigma_a^2\mathbf{V}_*$ for convenience of notation. Then, $\mathbf{u} = \mathbf{H}'\mathbf{w}$ has normal distribution with zero mean vector and covariance matrix $\text{cov}[\mathbf{u}] = \sigma_a^2\mathbf{H}'\mathbf{V}_*\mathbf{H}$. Thus, the likelihood of $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ_a^2 based on $\mathbf{u}$, that is, the density of $\mathbf{u}$, is

$$p(\mathbf{u}|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = (2\pi\sigma_a^2)^{-(n-k)/2}|\mathbf{H}'\mathbf{V}_*\mathbf{H}|^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2}\mathbf{u}'(\mathbf{H}'\mathbf{V}_*\mathbf{H})^{-1}\mathbf{u}\right]$$

It has been established (e.g., Harville, 1974, 1977), however, that this likelihood (i.e., density) can be expressed in an equivalent form that does not involve the particular choice of error contrast matrix $\mathbf{H}'$ as

$$\begin{aligned} L_*(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) &\equiv p(\mathbf{u}|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) \\ &= (2\pi\sigma_a^2)^{-(n-k)/2} |\mathbf{X}'\mathbf{X}|^{1/2} |\mathbf{V}_*|^{-1/2} \\ &\quad \times |\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X}|^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2} S(\hat{\boldsymbol{\beta}}_G, \boldsymbol{\phi}, \boldsymbol{\theta})\right] \end{aligned} \tag{9.5.11}$$

where

$$\begin{aligned} S(\hat{\boldsymbol{\beta}}_G, \boldsymbol{\phi}, \boldsymbol{\theta}) &= (\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}_G)'\mathbf{V}_*^{-1}(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}}_G) \\ &\equiv \mathbf{w}'(\mathbf{V}_*^{-1} - \mathbf{V}_*^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_\star^{-1})\mathbf{w} \end{aligned}$$

and $\hat{\boldsymbol{\beta}}_G = (\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{w}$. Evaluation of the restricted likelihood (9.5.11) requires little additional computational effort beyond that of the ''full'' likelihood, only the additional factor $|\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X}|$. Therefore, numerical determination of the REML estimates of $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ_a^2 is very similar to methods for ML estimation of the ARMA model parameters. However, one difference is that the REML estimate of σ_a^2 takes into account the loss in degrees of freedom that results from estimating the regression parameters and is given by $\hat{\sigma}_a^2 = S(\hat{\boldsymbol{\beta}}_G, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})/(n-k)$ as opposed to $S(\hat{\boldsymbol{\beta}}_G, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})/n$ for the ML estimate, although arguments can be put forth for use of the divisor $n-k-p-q$ rather than $n-k$ in the REML estimate $\hat{\sigma}_a^2$. For further discussion and details related to REML estimation, see Tunnicliffe Wilson (1989) and Cheang and Reinsel (2000, 2003).

APPENDIX A9.1 AUTOCOVARIANCES FOR SOME SEASONAL MODELS

See the following Table A9.1:

TABLE A9.1 Autocovariances for Some Seasonal Models

Model	(Autocovariances of $w_t)/\sigma_a^2$	Special Characteristics
(1) $w_t = (1 - \theta B)(1 - \Theta B^s)a_t$ $w_t = a_t - \theta a_{t-1} - \Theta a_{t-s} + \theta\Theta a_{t-s-1}$ $s \geq 3$	$\gamma_0 = (1 + \theta^2)(1 + \Theta^2)$ $\gamma_1 = -(1 + \Theta^2)$ $\gamma_{s-1} = \theta\Theta$ $\gamma_s = -\Theta(1 + \theta^2)$ $\gamma_{s+1} = \gamma_{s-1}$ All other autocovariances are zero.	(a) $\gamma_{s-1} = \gamma_{s+1}$ (b) $\rho_{s-1} = \rho_{s+1} = \rho_1\rho_s$
(2) $(1 - \Phi B^s)w_t = (l - \theta B)(1 - \Theta B^s)a_t$ $w_t - \Phi w_{t-s} = a_t - \theta a_{t-l} - \Theta a_{t-s} + \theta\Theta a_{t-s-1}$ $s \geq 3$	$\gamma_0 = (1 + \theta^2)\left[1 + (\Theta - \Phi)^2 \times (1 - \Phi^2)^{-1}\right]$ $\gamma_1 = -\theta\left[1 + (\Theta - \Phi)^2 \times (1 - \Phi^2)^{-1}\right]$ $\gamma_{s-1} = \theta\left[\Theta - \Phi - \Phi(\Theta - \Phi)^2 \times (1 - \Phi^2)^{-1}\right]$ $\gamma_s = -(1 + \theta^2)\left[\Theta - \Phi - \Phi(\Theta - \Phi)^2 \times (1 - \Phi^2)^{-1}\right]$ $\gamma_{s+1} = \gamma_{s-1}$ $\gamma_j = \Phi\gamma_{j-s}\ j \geq s + 2$ For $s \geq 4, \gamma_2, \gamma_3, \ldots, \gamma_{s-2}$ are all zero.	(a) $\gamma_{s-1} = \gamma_{s+1}$ (b) $\gamma_j = \Phi\gamma_{j-s}\ \ j \geq s + 2$
(3) $w_t = (1 - \theta_1 B - \theta_2 B^2)$ $\times (1 - \Theta_1 B^s - \Theta_2 B^{2s})a_t$ $w_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \Theta_1 a_{t-s}$ $+\theta_1\Theta_1 a_{t-s-1} + \theta_2\Theta_1 a_{t-s-2}$ $-\Theta_2 a_{t-2s} + \theta_1\Theta_2 a_{t-2s-1}$ $+\theta_2\Theta_2 a_{t-2s-2}$ $s \geq 5$	$\gamma_0 = (1 + \theta_1^2 + \theta_2^2)(1 + \Theta_1^2 + \Theta_2^2)$ $\gamma_1 = -\theta_1(1 - \theta_2)(1 + \Theta_1^2 + \Theta_2^2)$ $\gamma_2 = -\theta_2(1 + \Theta_1^2 + \Theta_2^2)$ $\gamma_{s-2} = \theta_2\Theta_1(1 - \Theta_2)$ $\gamma_{s-1} = \theta_1\Theta_1(1 - \theta_2)(1 - \Theta_2)$ $\gamma_s = -\Theta_1(1 - \Theta_2)(1 + \theta_1^2 + \theta_2^2)$ $\gamma_{s+1} = \gamma_{s-1}$ $\gamma_{s+2} = \gamma_{s-2}$ $\gamma_{2s-2} = \theta_2\Theta_2$ $\gamma_{2s-1} = \theta_1(1 - \theta_2)\Theta_2$ $\gamma_{2s} = -\Theta_2(1 + \theta_1^2 + \theta_2^2)$ $\gamma_{2s+1} = \gamma_{2s-1}$	(*a*) $\gamma_{s-2} = \gamma_{s+2}$ (*b*) $\gamma_{s-1} = \gamma_{s+1}$ (*c*) $\gamma_{s-2} = \gamma_{2s+2}$ (*d*) $\gamma_{2s-1} = \gamma_{2s+1}$

TABLE A9.1 (*continued*)

Model	(Autocovariances of $w_t)/\sigma_a^2$	Special Characteristics
	$\gamma_{2s+2} = \gamma_{2s-2}$ All other autocovariances are zero.	
(3a) *Special case of model 3* $w_t = (1 - \theta_1 B - \theta_2 B^2)(1 - \Theta B^s)a_t$ $w_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \Theta a_{t-s}$ $+\theta_1 \Theta a_{t-s-1} + \theta_2 \Theta a_{t-s-2}$ $s \geq 5$	$\gamma_0 = (1 + \theta_1^2 + \theta_2^2)(1 + \Theta^2)$ $\gamma_1 = -\theta_1(1 - \theta_2)(1 + \Theta^2)$ $\gamma_2 = -\theta_2(1 + \Theta^2)$ $\gamma_{s-2} = \theta_2 \Theta$ $\gamma_{s-1} = \theta_1(1 - \theta_2)\Theta$ $\gamma_s = -\Theta(1 + \theta_1^2 + \theta_2^2)$ $\gamma_{s+1} = \gamma_{s-1}$ $\gamma_{s+2} = \gamma_{s-2}$ All other autocovariances are zero.	(a) $\gamma_{s-2} = \gamma_{s+2}$ (b) $\gamma_{s-1} = \gamma_{s+1}$
(3b) *Special case of model 3* $w_t = (1 - \theta B)(1 - \Theta_1 B^s - \Theta_2 B^{2s})a_t$ $w_t = a_t - \theta a_{t-1} - \Theta_1 a_{t-s} + \theta \Theta_1 a_{t-s-1}$ $-\Theta_2 a_{t-2s} + \theta \Theta_2 a_{t-2s-1}$ $s \geq 3$	$\gamma_0 = (1 + \theta^2)(1 + \Theta_1^2 + \Theta_2^2)$ $\gamma_1 = -\theta(1 + \Theta_1^2 + \Theta_2^2)$ $\gamma_{s-1} = \theta \Theta_1(1 - \Theta_2)$ $\gamma_s = -\Theta_1(1 - \Theta_2)(1 + \theta^2)$ $\gamma_{s+1} = \gamma_{s-1}$ $\gamma_{2s-1} = \theta \Theta_2$ $\gamma_{2s} = -\Theta_2(1 + \theta^2)$ $\gamma_{2s+1} = \gamma_{2s-1}$ All other autocovariances are zero.	(a) $\gamma_{s-1} = \gamma_{s+1}$ (b) $\gamma_{2s-1} = \gamma_{2s+1}$
(4) $w_t = (1 - \theta_1 B - \theta_s B^s - \theta_{s+1} B^{s+1})a_t$ $w_t = a_t - \theta_1 a_{t-1} - \theta_s a_{t-s} - \theta_{s+1} a_{t-s-1}$ $s \geq 3$	$\gamma_0 = 1 + \theta_1^2 + \theta_s^2 + \theta_{s+1}^2$ $\gamma_1 = -\theta_1 + \theta_s \theta_{s+1}$ $\gamma_{s-1} = \theta_1 \theta_s$ $\gamma_s = -\theta_s + \theta_1 \theta_{s+1}$ $\gamma_{s+1} = -\theta_{s+1}$ All other autocovariances are zero	(a) In general, $\gamma_{s-1} \neq \gamma_{s+1}$ $\gamma_1 \gamma_s \neq \gamma_{s+1}$

(*continued*)

TABLE A9.1 (*continued*)

Model	(Autocovariances of $w_t)/\sigma_a^2$	Special Characteristics
(4a) *Special case of model 4* $w_t = (1 - \theta_1 B - \theta_s B^s)a_t$ $w_t = a_t - \theta_1 a_{t-1} - \theta_s a_{t-s}$ $s \geq 3$	$\gamma_0 = 1 + \theta_1^2 + \theta_s^2$ $\gamma_1 = -\theta_1$ $\gamma_{s-1} = \theta_1\theta_s$ $\gamma_s = -\theta_s$ All other autocovariances are zero.	(a) Unlike model 4, $\gamma_{s+1} = 0$
(5) $(1 - \Phi B^s)w_t = (1 - \theta_1 B - \theta_s B^s$ $-\theta_{s+1}B^{s+1})a_t$ $w_t - \Phi w_{t-s} = a_t - \theta_1 a_{t-1} - \theta_s a_{t-s}$ $-\theta_{s+1}a_{t-s-1}$ $s \geq 3$	$\gamma_0 = 1 + \theta_1^2 + \dfrac{(\theta_s - \Phi)^2}{1 - \Phi^2} + \dfrac{(\theta_{s+1} + \theta_1\Phi)^2}{1 - \Phi^2}$ $\gamma_1 = -\theta_1 + \dfrac{(\theta_s - \Phi)(\theta_{s+1} + \theta_1\Phi)}{1 - \Phi^2}$ $\gamma_{s-1} = (\theta_s - \Phi)\left[\theta_1 + \Phi\dfrac{\theta_{s+1} + \theta_1\Phi}{1 - \Phi^2}\right]$ $\gamma_s = -(\theta_s - \Phi)\left[1 - \Phi\dfrac{\theta_s - \Phi}{1 - \Phi^2}\right]$ $+(\theta_{s+1} + \theta_1\Phi)\left[\theta_1 + \Phi\dfrac{\theta_{s+1} - \theta_1\Phi}{1 - \Phi^2}\right]$ $\gamma_{s+1} = -(\theta_{s+1} + \theta_1\Phi)\left[1 - \Phi\dfrac{\theta_s - \Phi}{1 - \Phi^2}\right]$ $\gamma_j = \Phi\gamma_{j-s}\ j \geq s + 2$ For $s \geq 4, \gamma_2, \ldots, \gamma_{s-2}$ are all zero.	(a) $\gamma_{s-1} \neq \gamma_{s+1}$ (b) $\gamma_j = \Phi\gamma_{j-s}\ j \geq s + 2$
(5a) *Special case of model 5* $(1 - \Phi B^s)w_t = (1 - \theta_1 B - \theta_s B^s)a_t$ $w_t - \Phi w_{t-s} = a_t - \theta_1 a_{t-1} - \theta_s a_{t-s}$ $s \geq 3$	$\gamma_0 = 1 + \dfrac{\theta_1^2 + (\theta_s - \Phi)^2}{1 - \Phi^2}$ $\gamma_1 = -\theta_1\left[1 - \Phi\dfrac{\theta_s - \Phi}{1 - \Phi^2}\right]$ $\gamma_{s-1} = \dfrac{\theta_1(\theta_s - \Phi)}{1 - \Phi^2}$ $\gamma_s = \dfrac{\Phi\theta_1^2 - (\theta_s - \Phi)(1 - \Phi\theta_s)}{1 - \Phi^2}$ $\gamma_j = \Phi\gamma_{j-s}\ j \geq s + 1$ For $s \geq 4, \gamma_2, \ldots, \gamma_{s-2}$ are all zero.	(a) Unlike model 5, $\gamma_{s+1} = \Phi\gamma_1$

EXERCISES

9.1. Show that the seasonal difference operator $1 - B^{12}$, often useful in the analysis of monthly data, may be factorized as follows:

$$(1 - B^{12}) = (1 + B)(1 - \sqrt{3}B + B^2)(1 - B + B^2)(1 + B^2)(1 + B + B^2) \times (1 + \sqrt{3}B + B^2)(1 - B)$$

Plot the zeros of this expression in the unit circle and show by actual numerical calculation and plotting of the results that the factors in the order given above correspond to sinusoids with frequencies (in cycles per year) of 6, 5, 4, 3, 2, 1, together with a constant term. [For example, the difference equation $(1 - B + B^2)x_t = 0$ with arbitrary starting values $x_1 = 0, x_2 = 1$ yields $x_3 = 1, x_4 = 0, x_5 = -1$, and so on, generating a sine wave of frequency 2 cycles per year.]

9.2. A method that has sometimes been used for ''deseasonalizing'' monthly time series employs an equally weighted 12-month moving average:

$$\overline{z}_t = \frac{1}{12}(z_t + z_{t-1} + \cdots + z_{t-11})$$

(a) Using the decomposition $(1 - B^{12}) = (1 - B)(1 + B + B^2 + \cdots + B^{11})$, show that $12(\overline{z}_t - \overline{z}_{t-1}) = (1 - B^{12})z_t$.

(b) The exceedance for a given month over the previous moving average may be computed as $z_t - \overline{z}_{t-1}$. A quantity u_t may then be calculated that compares the current exceedance with the average of similar monthly exceedances experienced over the last k years. Show that u_t may be written as

$$u_t = \left(1 - \frac{B}{12}\frac{1 - B^{12}}{1 - B}\right)\left(1 - \frac{B^{12}}{k}\frac{1 - B^{12k}}{1 - B^{12}}\right) z_t$$

9.3. It has been shown (Tiao et al., 1975) that monthly averages for the (smog-producing) oxidant level in Azusa, California, may be represented by the model

$$(1 - B^{12})z_t = (1 + 0.2B)(1 - 0.9B^{12})a_t \qquad \sigma_a^2 = 1.0$$

(a) Compute and plot the ψ_j weights of this model.

(b) Compute and plot the π_j weights of this model.

(c) Calculate the standard deviations of the forecast errors 3 months and 12 months ahead.

(d) Obtain the eventual forecast function.

9.4. The monthly oxidant averages in parts per hundred million in Azusa from January 1969 to December 1972 were as follows:

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1969	2.1	2.6	4.1	3.9	6.7	5.1	7.8	9.3	7.5	4.1	2.9	2.6
1970	2.0	3.2	3.7	4.5	6.1	6.5	8.7	9.1	8.1	4.9	3.6	2.0
1971	2.4	3.3	3.3	4.0	3.6	6.2	7.7	6.8	5.8	4.1	3.0	1.6
1972	1.9	3.0	4.5	4.2	4.8	5.7	7.1	4.8	4.2	2.3	2.1	1.6

Using the model of Exercise 9.3, compute the forecasts for the next 24 months. (Approximate unknown a's by zeros.)

9.5. Thompson and Tiao (1971) have shown that the outward station movements of telephones (logged data) in Wisconsin are well represented by the model

$$(1 - 0.5B^3)(1 - B^{12})z_t = (1 - 0.2B^9 - 0.3B^{12} - 0.2B^{13})a_t$$

Obtain and plot the autocorrelation function of $w_t = (1 - B^{12})z_t$ for lags $1, 2, \ldots, 24$.

9.6. Consider the airline series analyzed earlier in this chapter. We have seen that the logarithm of the series is well represented by the multiplicative model $w_t = (1 - \theta B)(1 - \Theta_{12}B^{12})a_t$

(a) Compute and plot the 36-step-ahead forecasts and associated ± 2 forecast error limits for the logged series.

(b) Use the results in part (a) to obtain 12-step-ahead forecasts and associated forecast error limits for the original series. Plot the results.

9.7. Quarterly earnings per share of the U.S. company Johnson & Johnson are available for the period 1960–1980 as series 'JohnsonJohnson' in the R datasets package.

(a) Plot the time series using the graphics capabilities in R.

(b) Determine a variance stabilizing transformation for the series.

(c) Plot the autocorrelation functions and identify a suitable model (or models) for the series.

(d) Estimate the parameters of the model (or models) identified in part (**c**) and assess the statistical significance of the estimated parameters.

(e) Perform diagnostic checks to determine the adequacy of the fitted model.

(f) Compute and plot the l-step-ahead forecasts and associated two-standard-error prediction limits, $l = 1, \ldots, 4$, for this series.

9.8. Monthly Mauna Loa atmospheric CO_2 concentration readings for the period 1959–1997 are available as series 'co_2' in the R datasets package.

(a) Plot the time series and comment on the pattern in the data.

(b) Examine the autocorrelation structure and develop a suitable time series model for this series.

(c) Compute and plot the 12-step-ahead forecasts and associated two-standard-error prediction limits.

9.9. A time series representing the total monthly electricity generated in the United States (in millions of kilowatt-hours) for the period January 1970 to December 2005 is available as series 'electricity' in the R TSA package.

(a) Plot the series and comment. Is a variance stabilizing transformation needed for this case?

(b) Determine a suitable model for the series following the iterative three-stage procedure of model identification, parameter estimation, and diagnostics checking.

(c) Is there evidence of a deterministic seasonal pattern in this series? If so, how would this impact your choice of model for this series?

9.10. Consider the time series model $w_t = \beta_0 + N_t$ where N_t follows the AR(1) model $N_t = \phi N_{t-1} + a_t$. Assume that a series of length n is available for analysis.

(a) Assuming that the parameter ϕ is known, derive the generalized least-squares estimator of the constant β_0 in this model.

(b) Repeat the derivation in part (**a**) assuming that N_t follows the seasonal AR model $N_t = \phi_4 N_{t-4} + a_t$.

9.11. Suppose the quarterly seasonal process $\{z_t\}$ is represented as $z_t = S_t + a_{2t}$, where S_t follows a ''seasonal random walk'' model $(1 - B^4)S_t = \theta_0 + a_{1t}$, and a_{1t} and a_{2t} are independent white noise processes with variances $\sigma^2_{a_1}$ and $\sigma^2_{a_2}$, respectively. Show that z_t follows the seasonal ARIMA model $(1 - B^4)z_t = \theta_0 + (1 - \Theta B^4)a_t$, and determine expressions for Θ and σ^2_a in terms of the variance parameters of the other two processes. Discuss the implication if the resulting value of Θ is equal (or very close) to one, with regard to deterministic seasonal components.

9.12. Monthly averages of hourly ozone readings in downtown Los Angeles for the period from January 1955 to December 1972 are included as Series R in Part 5 of this book; see also http://pages.stat.wisc.edu/reinsel/bjr-data/.

(a) Plot the time series and comment.

(b) Develop a suitable time model for this time series. Discuss the adequacy of the selected model.

10

ADDITIONAL TOPICS AND EXTENSIONS

In previous chapters, the properties of *linear* autoregressive–moving average models have been examined extensively and it has been shown how these models can be used to represent stationary and nonstationary time series that arise in practice. This chapter will discuss additional topics that either supplement or extend the material presented in earlier chapters. We begin by discussing unit root tests that can be used as a supplementary tool to determine whether a time series is unit root nonstationary and can be transformed to a stationary series through differencing. This topic is discussed in Section 10.1. Unit root testing has received considerable attention in the econometrics literature, in particular, since it appears to be a common starting point for applied research in macroeconomics. For example, unit root tests are an integral part of the methodology used to detect long-term equilibrium relationships among nonstationary economic time series, commonly referred to as cointegration. In Section 10.2, we consider models for conditional heteroscedastic time series, which exhibit periods of differing degrees of volatility or variability depending on the past history of the series. Such behavior is common in many economic and financial time series, in particular. In Section 10.3, we introduce several classes of nonlinear time series models, which are capable of capturing some distinctive features in the behavior of processes that deviate from linear Gaussian time series. Finally, Section 10.4 looks at models for long memory processes, which are characterized by the much slower convergence to zero of their autocorrelation function ρ_k as $k \to \infty$ compared with the dependence structure of ARMA processes.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

10.1 TESTS FOR UNIT ROOTS IN ARIMA MODELS

As discussed in earlier chapters, the initial decision concerning the need for differencing is based, informally, on characteristics of the time series plot of z_t and of its sample autocorrelation function. In particular, a failure of the autocorrelations r_k to dampen out sufficiently quickly would indicate that the time series is nonstationary and needs to be differenced. This can be evaluated further using formal tests for unit roots in the autoregressive operator of the model. Testing for unit roots has received considerable attention in the time series literature motivated by econometric applications, in particular. Early contributions to this area include work by Dickey and Fuller (1979, 1981). These authors proposed tests based on the conditional least-squares estimator for an autoregressive process and the corresponding ''t-statistic.'' While the underlying concepts are fairly straightforward, a number of challenges arise in practice. In particular, the distribution theory for parameter estimates and associated test statistics developed for stationary time series do not apply when a unit root is present in the model. The asymptotic distributions are functions of standard Brownian motions and do not have convenient closed-form expressions. As a result, the percentiles of the distributions needed to perform the tests have to be evaluated using numerical approximations or by simulation. Moreover, the form of the test statistics and their asymptotic distributions are impacted by the presence of deterministic terms such as constants or time trends in the model. The size and power characteristics of unit root tests can also be a concern for shorter time series. This section provides a brief description of the tests proposed by Dickey and Fuller and summarizes some of the subsequent developments. For a more detailed discussion of unit root testing, see, for example, Hamilton (1994) and Fuller (1996). Reviews of unit root tests and their applications are provided by Dickey et al. (1986), Pantula et al. (1994), Phillips and Xiao (1998), and Haldrup et al. (2013), among others.

10.1.1 Tests for Unit Roots in AR Models

Simple AR(1) Model. To introduce unit root testing, we first examine the simple AR(1) model $z_t = \phi z_{t-1} + a_t, t = 1, 2, \ldots, n$, with $z_0 = 0$ and no constant term. We are interested in testing the hypothesis that $\phi = 1$ so that the series follows a random walk. The conditional least-squares (CLS) estimator of ϕ is given by

$$\hat{\phi} = \frac{\sum_{t=2}^{n} z_{t-1} z_t}{\sum_{t=2}^{n} z_{t-1}^2} = \phi + \frac{\sum_{t=2}^{n} z_{t-1} a_t}{\sum_{t=2}^{n} z_{t-1}^2}$$

In the stationary case with $|\phi| < 1$, the statistic $n^{1/2}(\hat{\phi} - \phi)$ has an approximate normal distribution with zero mean and variance $(1 - \phi^2)$. However, when $\phi = 1$, so that $z_t = \sum_{j=0}^{t-1} a_{t-j} + z_0$ in the integrated form, it can be shown that

$$n(\hat{\phi} - 1) = \frac{n^{-1} \sum_{t=2}^{n} z_{t-1} a_t}{n^{-2} \sum_{t=2}^{n} z_{t-1}^2} = O_p(1)$$

bounded in probability as $n \to \infty$, with both the numerator and denominator possessing nondegenerate and nonnormal limiting distributions. Hence, in the nonstationary case the estimator $\hat{\phi}$ approaches its true value $\phi = 1$ with increasing sample size n at a faster rate than in the stationary case.

The limiting distribution of $n(\hat{\phi}-1)$ was studied by Dickey and Fuller (1979) who showed that under the null hypothesis $\phi = 1$

$$n(\hat{\phi}-1) \xrightarrow{\mathcal{D}} \frac{\frac{1}{2}(\Lambda^2-1)}{\Gamma} \tag{10.1.1}$$

where $(\Gamma, \Lambda) = (\sum_{i=1}^{\infty} \gamma_i^2 Z_i^2, \sum_{i=1}^{\infty} 2^{1/2}\gamma_i Z_i)$, with $\gamma_i = 2(-1)^{i+1}/[(2i-1)\pi]$, and the Z_i are iid $N(0, 1)$ distributed random variables. An equivalent representation for the distribution is given by

$$\begin{aligned} n(\hat{\phi}-1) &\xrightarrow{\mathcal{D}} \frac{\int_0^1 B(u)\,dB(u)}{\int_0^1 B(u)^2\,du} \\ &= \frac{\frac{1}{2}(B(1)^2-1)}{\int_0^1 B(u)^2\,du} \end{aligned} \tag{10.1.2}$$

where $B(u)$ is a (continuous-parameter) standard Brownian motion process on [0, 1]; see Chan and Wei (1988). Such a process is characterized by the properties that $B(0) = 0$, increments over nonoverlapping intervals are independent, and $B(u+s) - B(s)$ is distributed as normal $N(0, u)$. Basically, $B(u)$ is the limit as $n \to \infty$ of the process

$$\frac{n^{-1/2}}{\sigma_a} z_{[nu]} = \frac{n^{-1/2}}{\sigma_a} \sum_{t=1}^{[nu]} a_t$$

where $[nu]$ denotes the largest integer part of $nu, 0 < u < 1$.

By the functional central limit theorem (Billingsley, 1999; Hall and Heyde, 1980, Section 4.2), $n^{-1/2} z_{[nu]}/\sigma_a$ converges in law as $n \to \infty$ to the standard Brownian motion process $\{B(u), 0 < u < 1\}$. The random walk model $z_t = z_{t-1} + a_t$ with $z_0 = 0$ implies that $z_{t-1}a_t = \frac{1}{2}(z_t^2 - z_{t-1}^2 - a_t^2)$, so that

$$n^{-1}\sum_{t=2}^{n} z_{t-1}a_t = \frac{1}{2}\left[n^{-1}z_n^2 - n^{-1}\sum_{t=1}^{n} a_t^2\right] \xrightarrow{\mathcal{D}} \frac{\sigma_a^2}{2}[B(1)^2 - 1] \tag{10.1.3}$$

since $n^{-1}z_n^2 = \sigma_a^2(n^{-1/2}z_n/\sigma_a)^2 \xrightarrow{\mathcal{D}} \sigma_a^2 B(1)^2$ while $n^{-1}\sum_{t=1}^{n} a_t^2 \xrightarrow{\mathcal{P}} \sigma_a^2$ by the law of large numbers. In addition,

$$n^{-2}\sum_{t=2}^{n} z_{t-1}^2 = \sigma_a^2 \int_0^1 \left(\frac{n^{-1/2}z_{[nu]}}{\sigma_a}\right)^2 du + o_p(1) \xrightarrow{\mathcal{D}} \sigma_a^2 \int_0^1 B(u)^2\,du \tag{10.1.4}$$

by the continuous mapping theorem (Billingsley, 1999; Hall and Heyde, 1980, p. 276). Hence, these last two results establish the representation (10.1.2).

The limiting distribution of $n(\hat{\phi}-1)$ described above does not have a closed-form representation but it can be evaluated numerically using simulation. Tables for the percentiles of the limiting distribution are given by Fuller (1996, Appendix 10.A). Fuller also provides

tables for the limiting distribution of the ''Studentized'' statistic

$$\hat{\tau} = \frac{\hat{\phi} - 1}{s_a(\sum_{t=2}^{n} z_{t-1}^2)^{-1/2}} \tag{10.1.5}$$

where $s_a^2 = (n-2)^{-1}(\sum_{t=2}^{n} z_t^2 - \hat{\phi} \sum_{t=2}^{n} z_{t-1} z_t)$ is the residual mean square. These results can be used to test the random walk hypothesis that $\phi = 1$. Since the alternative hypothesis of stationarity is one-sided, the test rejects $\phi = 1$ when $\hat{\tau}$ is sufficiently negative. The test based on $\hat{\tau}$ is commonly referred to as the Dickey–Fuller (DF) test in the literature.

Higher Order AR Models. To extend the results to higher order models, we consider a generalized AR($p+1$) process $z_t = \sum_{j=1}^{p+1} \varphi_j z_{t-j} + a_t$, or $\varphi(B) z_t = a_t$, where $\varphi(B)$ contains a single unit root so that $\varphi(B) = \phi(B)(1-B)$ and $\phi(B) = 1 - \sum_{j=1}^{p} \phi_j B^j$ is a *stationary* AR operator of order p. Hence,

$$\varphi(B) z_t = \phi(B)(1-B) z_t = z_t - z_{t-1} - \sum_{j=1}^{p} \phi_j (z_{t-j} - z_{t-j-1}) + a_t$$

Testing for a unit root in $\varphi(B)$ is then equivalent to testing $\rho = 1$ in the model

$$z_t = \rho z_{t-1} + \sum_{j=1}^{p} \phi_j (z_{t-j} - z_{t-j-1}) + a_t$$

or equivalently testing $\rho - 1 = 0$ in the model

$$(z_t - z_{t-1}) = (\rho - 1) z_{t-1} + \sum_{j=1}^{p} \phi_j (z_{t-j} - z_{t-j-1}) + a_t$$

In fact, for any generalized AR($p+1$) model $z_t = \sum_{j=1}^{p+1} \varphi_j z_{t-j} + a_t$, it is seen that the model can be written in an equivalent form as

$$w_t = (\rho - 1) z_{t-1} + \sum_{j=1}^{p} \phi_j w_{t-j} + a_t \tag{10.1.6}$$

where $w_t = z_t - z_{t-1}$, $\rho - 1 = -\varphi(1) = \sum_{j=1}^{p+1} \varphi_j - 1$, and $\phi_j = \sum_{i=1}^{p} \varphi_i - 1$. Hence, the existence of a unit root in the AR operator $\varphi(B)$ is equivalent to $\rho = \sum_{j=1}^{p+1} \varphi_j = 1$.

Based on this last form of the model, let $(\hat{\rho} - 1, \hat{\phi}_1, \ldots, \hat{\phi}_p)$ denote the usual conditional least-squares estimates of the parameters in (10.1.6) obtained by regressing w_t on $z_{t-1}, w_{t-1}, \ldots, w_{t-p}$. Then, under the unit root model where $\rho = 1$ and $\phi(B)$ is stationary, it follows from Fuller (1996, Theorem 10.1.2 and Corollary 10.1.2.1) that

$$(\hat{\rho} - 1) \Big/ \left\{ s_a \left(\sum_{t=p+2}^{n} z_{t-1}^2 \right)^{-1/2} \right\}$$

has the same limiting distribution as the Studentized statistic $\hat{\tau}$ in (10.1.5) for the AR(1) model, while $(n-p-1)(\hat{\rho}-1)c$, where $c=\sum_{j=0}^{\infty}\psi_j$ with $\psi(B)=\phi^{-1}(B)$, has approximately the same distribution as the statistic $n(\hat{\phi}-1)$ for the AR(1) model. Also, it follows that the statistic, denoted as $\hat{\tau}$, formed by dividing $(\hat{\rho}-1)$ by its estimated standard error from the least-squares regression will be asymptotically equivalent to the statistic $(\hat{\rho}-1)/\{s_a(\sum_{t=p+2}^{n}z_{t-1}^2)^{-1/2}\}$, and hence will have the same limiting distribution as the statistic $\hat{\tau}$ for the AR(1) case; see Said and Dickey (1984).

The test statistic $\hat{\tau}$ formed from the regression of w_t on $z_{t-1}, w_{t-1}, \ldots, w_{t-p}$ as described above can thus be used to test for a unit root in the AR$(p+1)$ model $\varphi(B)z_t=a_t$. This is the well-known augmented Dickey–Fuller (ADF) test. Furthermore, as shown by Fuller (1996, Theorem 10.1.2), the limiting distribution of the least-squares estimates $(\hat{\phi}_1,\ldots,\hat{\phi}_p)$ for the parameters of the stationary operator $\phi(B)$ in the model is the same as the standard asymptotic distribution for least-squares estimates obtained by regressing the stationary differenced series w_t on $w_{t-1},\ldots,w_{t-p}$. The estimation results for the stationary AR model discussed earlier in Section 7.2.6 are therefore valid in this case.

Inclusion of a Constant Term. The results described above extend with suitable modifications to the more practical case where a constant term θ_0 is included in the least-squares regression. Under stationarity, the constant is related to the mean of the process and equals $\theta_0=(1-\varphi_1-\cdots-\varphi_{p+1})\mu=(1-\rho)\mu$. The least-squares regression yields a test statistic analogous to $\hat{\tau}$ above denoted by $\hat{\tau}_\mu$, although the limiting distribution of this test statistic is derived under the assumption that $\theta_0=0$ under the null hypothesis $\phi=1$. For example, for the AR(1) model $z_t=\phi z_{t-1}+\theta_0+a_t$ with $\theta_0=(1-\phi)\mu$, the least-squares estimator for ϕ is

$$\hat{\phi}_\mu=\frac{\sum_{t=2}^{n}(z_{t-1}-\bar{z}_{(1)})(z_t-\bar{z}_{(0)})}{\sum_{t=2}^{n}(z_{t-1}-\bar{z}_{(1)})^2} \tag{10.1.7}$$

where $\bar{z}_{(i)}=(n-1)^{-1}\sum_{t=2}^{n}z_{t-i}, i=0,1,$ so that $\hat{\phi}_\mu=\phi+\sum_{t=2}^{n}(z_{t-1}-\bar{z}_{(1)})a_t/\sum_{t=2}^{n}(z_{t-1}-\bar{z}_{(1)})^2$. When $\phi=1$, the representation for the limiting distribution of $n(\hat{\phi}_\mu-1)$ analogous to (10.1.2) is given by

$$n(\hat{\phi}_\mu-1)\xrightarrow{\mathcal{D}}\frac{\int_0^1 B(u)dB(u)-\xi B(1)}{\int_0^1 B(u)^2du-\xi^2} \tag{10.1.8}$$

where $\xi=\int_0^1 B(u)du$, and it is assumed that $\theta_0=(1-\phi)\mu=0$ when $\phi=1$. The corresponding Studentized test statistic for $\phi=1$ in the AR(1) case is

$$\hat{\tau}_\mu=\frac{\hat{\phi}_\mu-1}{s_a[\sum_{t=2}^{n}(z_{t-1}-\bar{z}_{(1)})^2]^{-1/2}} \tag{10.1.9}$$

The limiting distribution of $\hat{\tau}_\mu$ readily follows from the result in (10.1.8). Tables of percentiles of the distribution of $\hat{\tau}_\mu$ when $\phi=1$ are provided by Fuller (1996, p. 642). Note that under $\phi=1$, since $z_t=\sum_{j=0}^{t-1}a_{t-j}+z_0$ in the truncated random shock or integrated form, the terms $z_t-\bar{z}_{(0)}$ and $z_{t-1}-\bar{z}_{(1)}$ do not involve the initial value z_0. Therefore, the distribution theory for the least-squares estimator $\hat{\phi}_\mu$ does not depend on any assumption

concerning z_0. Also, the results for the first-order AR(1) model with a constant term extend to higher order autoregressive models in much the same way as it does when the constant term θ_0 is absent from the model. The tables developed for the percentiles of the limiting distribution of statistic $\hat{\tau}_\mu$ can thus be used for higher order AR models as well.

The procedures described above are based on conditional LS estimation or equivalently on the conditional likelihood assuming that the noise term a_t follows a normal distribution. Pantula et al. (1994) studied *unconditional* likelihood estimation for the AR model with a unit root. They showed that the limiting distributions of estimators and test statistics for unit root based on the unconditional likelihood are different from those based on the conditional approach. For example, in the simple AR(l) model $z_t = \phi z_{t-1} + a_t$ with no constant term included in the estimation, the unconditional log-likelihood is

$$l(\phi, \sigma_a^2) = -\frac{n}{2}\ln(\sigma_a^2) + \frac{1}{2}\ln(1-\phi^2) - \frac{1}{2\sigma_a^2}\left[\sum_{t=2}^{n}(z_t - \phi z_{t-1})^2 + (1-\phi^2)z_1^2\right]$$

as shown in Appendix A7.4. The unconditional ML estimator $\hat{\phi}$, which maximizes $l(\phi, \sigma_a^2)$, is a root of the cubic equation in $\hat{\phi}$ given by (A7.4.20). Pantula et al. (1994) derived the asymptotic distribution of $n(\hat{\phi}_1 - 1)$ and concluded, using Monte Carlo studies, that tests for unit root in AR models based on the unconditional maximum likelihood estimator are more powerful than those based on the conditional maximum likelihood estimator for moderate values of n.

Processes with Deterministic Linear Trend. The asymptotic distribution theory related to the least-squares estimator $\hat{\phi}_\mu$ in (10.1.7) depends heavily on the condition that the constant term θ_0 is zero under the null hypothesis $\phi = 1$, since the behavior of the process $z_t = z_{t-1} + \theta_0 + a_t$ differs fundamentally between the cases $\theta_0 = 0$ and $\theta_0 \neq 0$. When $\theta_0 = 0$, the process is a random walk with zero drift. When $\theta_0 \neq 0$, the model can be written as $z_t = \theta_0 t + z_0 + u_t$, where $u_t = u_{t-1} + a_t$. The process $\{z_t\}$ is now a random walk with drift and its long-term behavior in many respects is dominated by the deterministic linear trend term $\theta_0 t$ contained in z_t. If θ_0 has a nonzero value under the hypothesis $\phi = 1$, then $n^{3/2}(\hat{\phi}_\mu - 1)$ converges in distribution to $N(0, 12\sigma_a^2/\theta_0^2)$ as $n \to \infty$. Thus, when $\theta_0 \neq 0$ the asymptotic normal distribution theory applies to the least-squares estimator $\hat{\phi}_\mu$ and to the corresponding test statistic $\hat{\tau}_\mu$. For details, see Fuller (1996, Section 10.1.2) and Hamilton (1994, Section 17.4).

For a time series that exhibits a persistent trend, it is often of interest to determine whether the trend arises from the drift term of a random walk or it is due to a deterministic trend added to a stationary AR(1) model, for example. The previous formulation of the AR(1) model with nonzero constant $z_t = \phi z_{t-1} + \theta_0 + a_t$ does not allow this, since when $|\phi| < 1$ this model implies a process with constant mean $\mu = E[z_t] = \theta_0/(1-\phi)$, independent of time. An alternate formulation of the AR(1) model that allows for a deterministic linear time trend that is not linked to ϕ is

$$z_t = \alpha + \theta_0 t + u_t \qquad \text{where} \qquad u_t = \phi u_{t-1} + a_t \qquad t = 1, \ldots, n \tag{10.1.10}$$

This model has a linear trend with slope $\theta_0 \neq 0$ regardless of whether $\phi = 1$ or $\phi \neq 1$. It is of interest to note the relation between parameters in this form relative to the previous

form. Applying the operator $(1 - \phi B)$ to (10.1.10), the model can be expressed as

$$z_t = \phi z_{t-1} + \alpha_0 + \delta_0 t + a_t \tag{10.1.11}$$

where $\alpha_0 = \alpha(1 - \phi) + \phi\theta_0$ and $\delta_0 = \theta_0(1 - \phi)$. Hence, in this form $\alpha_0 = \theta_0$ and $\delta_0 = 0$ are obtained under $\phi = 1$, so that $z_t = z_{t-1} + \theta_0 + a_t$. The presence of the linear time trend in (10.1.10) thus leads to a model with a nonzero constant but a zero coefficient for the time trend under the null hypothesis $\phi = 1$. The constant θ_0 is referred to as a drift term and measures the expected change in the series when the time increases by one unit.

A common procedure to test for a unit root in this model is to perform least-squares estimation with the linear trend term t in addition to the constant included in the regression. The resulting estimator of ϕ, denoted as $\hat{\phi}_\tau$, is such that the limiting distribution of $n(\hat{\phi}_\tau - 1)$, under $\phi = 1$, does not depend on the value of the constant $\alpha_0 = \theta_0$ but still requires the coefficient δ_0 of the time variable t to be zero under the null hypothesis. Hence, this estimator $\hat{\phi}_\tau$ can be used as the basis of a valid test of $\phi = 1$ regardless of the value of the constant θ_0. Tables of percentiles of the null distribution of $n(\hat{\phi}_\tau - 1)$ and of the corresponding Studentized statistic $\hat{\tau}_\tau$ are available in Fuller (1996, p. 642).

Alternative procedures to test $\phi = 1$ in the presence of a possible deterministic linear trend, which are valid regardless of the value of the constant term, have been proposed by several authors. Bhargava (1986) developed a locally most powerful invariant test for unit roots. Schmidt and Phillips (1992) used a score (or Lagrange multiplier (LM)) test for the model (10.1.10), and Ahn (1993) extended this approach to allow for a more general ARMA model for the noise process u_t. Elliott et al. (1996) used a point optimal testing approach with maximum power against a local alternative for the same model. The power gains were obtained by a preliminary generalized least-squares (GLS) detrending procedure using a local alternative to $\phi = 1$, followed by use of the least-squares estimate $\hat{\phi}$ and corresponding test statistic $\hat{\tau}$ obtained from the detrended series. Subsequent contributions to this area include work by Ng and Perron (2001), Perron and Qu (2007), and Harvey et al. (2009), among others.

10.1.2 Extensions of Unit Root Testing to Mixed ARIMA Models

The test procedures described above and other similar ones have been extended to testing for unit roots in mixed ARIMA$(p, 1, q)$ models (e.g., see Said and Dickey (1984, 1985) and Solo (1984b)), as well as models with higher order differencing (e.g., see Dickey and Pantula (1987)). Said and Dickey (1984) showed that the Dickey–Fuller procedure, which was originally developed for autoregressive models of known order p, remains valid asymptotically for an ARIMA$(p, 1, q)$ model where p and q are unknown. The authors approximated the mixed model by an autoregressive model of sufficiently high order and applied the ADF test to the resulting AR model. The approximation assumes that the lag length of the autoregression increases with the length of the series, n, at a controlled rate less than $n^{1/3}$. Phillips (1987) and Phillips and Perron (1988) proposed a number of unit root tests that have become popular in the econometrics literature. These tests differ from the ADF tests in how they deal with serial correlation and heteroscedasticity in the error process. Thus, while the ADF tests approximate the ARMA structure by a high-order autoregression, the Phillips and Perron tests deal with serial correlation by directly modifying the test statistics to account for serial correlation. Likelihood ratio type of unit root tests have also been considered for the mixed ARIMA model based on both conditional and unconditional normal distribution likelihoods by Yap and Reinsel (1995) and Shin and

Fuller (1998), among others. Simulation studies suggest that these tests often perform better than $\hat{\tau}$-type test statistics for mixed ARIMA models.

Motivated by problems in macroeconomics and related fields, the literature has continued to grow and many other extensions have been developed. These include the use of bootstrap methods for statistical inference as discussed, for example, by Palm et al. (2008). The use of Bayesian methods for unit root models has also been considered. The problem of distinguishing unit root nonstationary series from series with structural breaks such as level shifts or trend changes has been considered by many researchers. The methodology has also been extended and modified to deal with more complex series involving nonlinearities, time-varying volatility, and fractionally integrated processes with long-range dependence. Tests with a null hypothesis of stationarity, rather than unit root nonstationarity, have also been proposed in the literature. For further discussion and references, see, for example, Phillips and Xiao (1998) and Haldrup et al. (2013).

Example: Series C. To illustrate unit root testing, consider the series of temperature readings referred to as Series C. Two potential models identified for this series in Chapter 6 were the ARIMA(1, 1, 0) and the ARIMA(0, 2, 0). Since there is some doubt about the need for the second differencing in the ARIMA(0, 2, 0) model, with the alternative model being a stationary AR(1) for the first differences, we investigate this more formally. The AR(1) model $\nabla z_t = \phi \nabla z_{t-1} + a_t$ for the first differences can be written as $\nabla^2 z_t = (\phi - 1)\nabla z_{t-1} + a_t$, and in this form the conditional least-squares regression estimate $\hat{\phi} - 1 = -0.187$ is obtained, with an estimated standard error of 0.038, and $\hat{\sigma}_a^2 = 0.018$. Note that this implies $\hat{\phi} = 0.813$ similar to results in Tables 6.5 and 7.6. The Studentized statistic to test $\phi = 1$ is $\hat{\tau} = -4.87$, which is far more negative than the lower one percentage point of -2.58 for the distribution of $\hat{\tau}$ in the tables of Fuller (1996). Also, $\hat{\tau}_\mu = -4.96$ was obtained when a constant term is included in the AR(1) model for ∇z_t. Hence, these estimation results do not support the need for second differencing and point to a preference for the ARIMA(1, 1, 0) model.

Implementation in R. Tests for unit roots can be performed using the package fUnitRoots available in the FinTS package in R. If z represents the time series of interest, the command used to perform the augmented Dickey–Fuller test is

```
> adfTest(z,lags,type=c("nc","c","ct")
```

where lags denotes the number of lags in the autoregressive model and type indicates whether or not a constant or trend should be included in the fitted model. The argument ''nc'' specifies that no constant should be included in the model, ''c'' is used for constant only, and ''ct'' specifies a trend plus a constant. For lags equal to 0, the test is the original Dickey–Fuller test. Otherwise, lags represents the order of the stationary autoregressive polynomial in (10.1.6). For a mixed ARMA model, it represents the order of the autoregressive approximation to this model.

The calculations for Series C described above can be performed in R as follows:

```
> library(fUnitRoots)
> adfTest(diff(ts(seriesC)),0,type=c("nc"))

Title: Augmented Dickey-Fuller Test
Test Results:
```

```
  PARAMETER:
  Lag Order: 0
  STATISTIC:  Dickey-Fuller: -4.8655
  P VALUE:  0.01

> adfTest(diff(ts(seriesC)),0,type=c("c"))

  Title: Augmented Dickey-Fuller Test
  Test Results:
  PARAMETER:
  Lag Order: 0
  STATISTIC:  Dickey-Fuller: -4.962
  P VALUE:  0.01
```

The values of the test statistics agree in both cases with those quoted in the example. Note that the output shows the p value but does not give the critical value for the test. If the critical values are needed, they can be obtained in R using the command

```
>adfTable(trend=c("nc","c","ct"), statistic=c("nc","c","ct"))
```

Example: Series A. For further illustration, consider Series A that represents concentration readings of a chemical process at 2-hour intervals and has $n = 197$ observations. In Chapters 6 and 7, two possible ARMA/ARIMA models were proposed for this series. One is the nearly nonstationary ARMA(1, 1) model, $(1 - \phi B)z_t = \theta_0 + (1 - \theta B)a_t$, with estimates $\hat{\phi} = 0.92$, $\hat{\theta} = 0.58$, $\hat{\theta}_0 = 1.45$, and $\hat{\sigma}_a^2 = 0.0974$. The second is the nonstationary ARIMA(0, 1, 1) model, $(1 - B)z_t = (1 - \theta B)a_t$, with estimates $\hat{\theta} = 0.71$ and $\hat{\sigma}_a^2 = 0.1004$. Below we use the ADF test to test the hypothesis that differencing is needed so that the series follows the ARIMA(0, 1, 1) model. To determine the order k of the autoregressive approximation to this model, we first use the R command ar(z) to select a suitable value for k based on the AIC criterion. The output suggests an AR(6) model, which is then used for the test. A slightly different choice of k does not alter the conclusion.

```
  > library(fUnitRoots)
  > ar(diff(ts(seriesA)),aic=TRUE)

  Call: ar(x = diff(ts(seriesA)), aic = TRUE)
  Coefficients:
  1 2 3 4 5 6
  -0.6098 -0.3984 -0.3585 -0.3175 -0.3142 -0.2139
  Order selected 6 sigma^2 estimated as 0.09941

  > adfTest(ts(seriesA),6,type=c("nc"))

  Title: Augmented Dickey-Fuller Test
  Test Results:
  PARAMETER:
  Lag Order: 6
  STATISTIC:  Dickey-Fuller: 0.6271
  P VALUE:  0.8151
```

The p values are large and the test does not reject the null hypothesis that the series needs to be differenced, suggesting that ARIMA(0, 1, 1) is the preferred model. A similar conclusion was reached by Solo (1984b) who used a Lagrange multiplier test to determine the need for differencing.

10.2 CONDITIONAL HETEROSCEDASTIC MODELS

This section presents an overview of some models that have been developed to describe time-varying variability or volatility in a time series. To first introduce some notation, we note that the ARMA(p, q) process $\phi(B)z_t = \theta_0 + \theta(B)a_t$ can be written as the sum of a predictable part and a prediction error as

$$z_t = E[z_t|F_{t-1}] + a_t$$

where F_{t-1} represents the past information available at time $t-1$ and a_t represents the prediction error. For the ARMA model, F_{t-1} is a function of past observations and past error terms, but could more generally include external regression variables X_t. The assumption made thus far is that the prediction errors a_t are *independent* random variables with a constant variance $\text{Var}[a_t] = \sigma_a^2$ that is independent of the past. However, this assumption appears inconsistent with the heteroscedasticity often seen for time series in business and economics, in particular. For example, financial time series such as stock returns often exhibit periods when the volatility is high and periods when it is lower. This characteristic feature, or *stylized fact*, is commonly referred to as volatility clustering. For illustration, Figure 10.1(a) shows the weekly S&P 500 Index over the period January 3, 2000 to May 27, 2014 for a total of 751 observations. The log returns calculated as $\ln(p_t/p_{t-1}) = \ln(p_t) - \ln(p_{t-1})$, where p_t represents the original time series, are shown in Figure 10.1(b). We note that while the original time series is nonstationary, the returns fluctuate around a stable mean level. However, the variability around the mean changes and volatility clusters

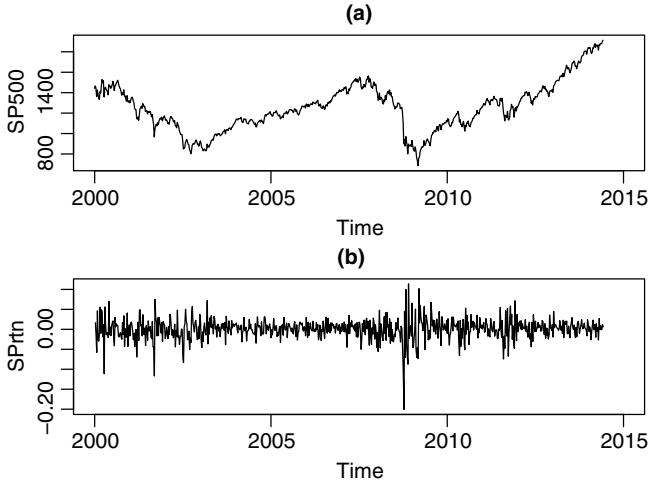


FIGURE 10.1 (a) Time plot of the weekly S&P 500 Index from January 3, 2000 to May 27, 2014, and (b) the weekly log returns on the S&P 500 Index.

are clearly visible. Note the high volatility during and following the 2008 financial crisis, in particular. Another common feature of financial time series is that the marginal distributions are leptokurtic and tend to have heavier tails than those of a normal distribution. A number of other stylized facts have been documented and investigated for financial data (for discussion and references, see, for example, Teräsvirta et al., 2010, Chapter 8).

The autoregressive conditional heteroscedastic (ARCH) model was introduced by Engle (1982) to describe time-varying variability in a series of inflation rates. An extension of this model called the generalized conditional heteroscedastic (GARCH) model was proposed by Bollerslev (1986). These models are capable of describing not only volatility clustering but also features such as heavy-tailed behavior that is common in many economic and financial time series. Still, there are other features related to volatility that are not captured by the basic ARCH and GARCH models. This has led to a number of extensions and alternative formulations aimed at addressing these issues. This section presents a brief description of the ARCH and GARCH models along with some extensions proposed in the literature. The literature in this area is extensive and only a select number of developments will be discussed. A more complete coverage can be found in survey papers by Bollerslev et al. (1992, 1994), Bera and Higgins (1993), Li et al. (2003), and Teräsvirta (2009), among others. Volatility modeling is also discussed in several time series texts, including Franses and van Dijk (2000), Mills and Markellos (2008), Teräsvirta et al. (2010), and Tsay (2010). Textbooks devoted to volatility modeling include Francq and Zakoïan (2010) and Xekalaki and Degiannakis (2010).

10.2.1 The ARCH Model

For a stationary ARMA process, the unconditional mean of the series is constant over time while the conditional mean $E[z_t|F_{t-1}]$ varies as a function of past observations. Parallel to this, the ARCH model assumes that the unconditional variance of the error process is constant over time but allows the conditional variance of a_t to vary as a function of past squared errors. Letting $\sigma_t^2 = \text{var}[a_t|F_{t-1}]$ denote the conditional variance of a_t, given the past F_{t-1}, the basic ARCH(s) model can be formulated as

$$a_t = \sigma_t e_t \tag{10.2.1}$$

where $\{e_t\}$ is a sequence of iid random variables with mean zero and variance 1, and

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_s a_{t-s}^2 \tag{10.2.2}$$

with $\alpha_0 > 0$, $\alpha_i \geq 0$, for $i = 1, \ldots, s-1$, and $\alpha_s > 0$. The parameter constraints are imposed to ensure that the conditional variance σ_t^2 is positive. The additional constraint $\sum_{i=1}^{s} \alpha_i < 1$ ensures that the a_t are covariance stationary with finite unconditional variance σ_a^2. For some time series, such as stock returns, the original observations are typically serially uncorrelated and the a_t are observed directly. Alternatively, the a_t can be the noise sequence associated with an ARMA or regression-type model. For modeling purposes, the e_t in (10.2.1) are usually assumed to follow a standard normal or a Student t-distribution.

The ARCH model was used by Engle (1982) to study the variance of UK inflation rates and by Engle (1983) to describe the variance of U.S. inflation rates. The ARCH model and its later extensions by Bollerslev (1986) and others quickly found other applications. For example, Diebold and Nerlove (1989) showed that the ARCH model may be used to generate statistically and economically meaningful measures of exchange rate volatility.

Bollerslev (1987) used the GARCH extension of the ARCH model to analyze the conditional volatility of financial returns observed at a monthly or higher frequency. In Weiss (1984), ARMA models with ARCH errors were used to model the time series behavior of 13 different U.S. macroeconomic time series. Bollerslev et al. (1992) describe a large number of other applications in their review of volatility models. While a majority of applications have been in finance and economics, the models have also been used in other fields. For example, Campbell and Diebold (2005) used volatility models in their analysis of the daily average temperatures for four U.S. cities. The models have also been used for variables such as wind speeds, air quality measurements, earthquake series, and in the analysis of speech signals. For selected references, see Francq and Zekoïan (2010, p. 12).

Some Properties of the ARCH Model. To establish some properties of the ARCH model, we first examine the ARCH(1) model where

$$\sigma_t^2 = \text{var}[a_t|F_{t-1}] = E[a_t^2 \mid F_{t-1}] = \alpha_0 + \alpha_1 a_{t-1}^2 \tag{10.2.3}$$

with $\alpha_0 > 0$ and $\alpha_1 > 0$. The form of the model shows that the conditional variance σ_t^2 will be large if a_{t-1} was large in absolute value and vice versa. A large (small) value of σ_t^2 will in turn tend to generate a large (small) value of a_t, thus giving rise to volatility clustering.

It follows from (10.2.1) that $E\,[a_t \mid F_{t-1}] = 0$. The *unconditional* mean of a_t is also zero since

$$E[a_t] = E\left[E[a_t \mid F_{t-1}]\right] = 0$$

Furthermore, the a_t are serially uncorrelated since for $j > 0$,

$$E[a_t a_{t-j}] = E\left[E[a_t a_{t-j} \mid F_{t-1}]\right] = E\left[a_{t-j} E[a_t \mid F_{t-1}]\right] = 0$$

But the a_t are not mutually independent since they are interrelated through their conditional variances. The lack of serial correlation is an important property that makes the ARCH model suitable for modeling asset returns that are expected to be uncorrelated by the efficient market hypothesis.

We also assume that the a_t have equal *unconditional* variances, $\text{var}[a_t] = E[a_t^2] = \sigma_a^2$, for all t, so that the process is weakly stationary. If $\alpha_1 < 1$, the unconditional variance exists and equals

$$\sigma_a^2 = \text{var}[a_t] = \frac{\alpha_0}{1 - \alpha_1} \tag{10.2.4}$$

This follows since

$$\sigma_a^2 = E[a_t^2] = E\left[E[a_t^2 \mid F_{t-1}]\right] = E[\alpha_0 + \alpha_1 a_{t-1}^2] = \alpha_0 + \alpha_1 \sigma_a^2$$

Further substituting $\alpha_0 = \sigma_a^2(1 - \alpha_1)$ from (10.2.4) into (10.2.3), we see that

$$\sigma_t^2 = \sigma_a^2 + \alpha_1(a_{t-1}^2 - \sigma_a^2) \tag{10.2.5}$$

or, equivalently, $\sigma_t^2 - \sigma_a^2 = \alpha_1(a_{t-1}^2 - \sigma_a^2)$. Hence, the conditional variance of a_t will be above the unconditional variance whenever a_{t-1}^2 is larger than the unconditional variance σ_a^2.

To study the tail behavior of a_t, we examine the fourth moment $\mu_4 = E[a_t^4]$. If a_t is normally distributed, conditional on the past, then

$$E[a_t^4 \mid F_{t-1}] = 3\sigma_t^4 = 3(\alpha_0 + \alpha_1 a_{t-1}^2)^2$$

Therefore, the fourth unconditional moment of a_t satisfies

$$E[a_t^4] = E\left[E[a_t^4 \mid F_{t-1}]\right] = 3\left[\alpha_0^2 + 2\alpha_0\alpha_1 E[a_{t-1}^2] + \alpha_1^2 E[a_{t-1}^4]\right]$$

Thus, if $\{a_t\}$ is fourth-order stationary so that $\mu_4 = E[a_t^4] = E[a_{t-1}^4]$, then

$$\mu_4 = \frac{3(\alpha_0^2 + 2\alpha_0\alpha_1\sigma_a^2)}{1 - 3\alpha_1^2} \equiv \frac{3\alpha_0^2(1-\alpha_1^2)}{(1-\alpha_1)^2(1-3\alpha_1^2)} \tag{10.2.6}$$

Since $\mu_4 = E[a_t^4] > 0$, this expression shows that α_1 must satisfy $0 < \alpha_1 < 1/\sqrt{3}$ in order for a_t to have finite fourth moment. Further, if κ denotes the unconditional kurtosis of a_t, then

$$\kappa = \frac{E[a_t^4]}{\left(E[a_t^2]\right)^2} = \frac{3(1-\alpha_1^2)}{1-3\alpha_1^2}$$

This value exceeds 3, the kurtosis of the normal distribution. Hence, the marginal distribution of a_t has heavier tails than those of the normal distribution. This is an additional feature of the ARCH model that makes it useful for modeling financial asset returns where heavy-tailed behavior is the norm.

To derive an alternative form of the ARCH process, we let $v_t = a_t^2 - \sigma_t^2$, so that $a_t^2 = \sigma_t^2 + v_t$. The random variables v_t then have zero mean and they are serially uncorrelated since

$$\begin{aligned} E[(a_t^2 - \sigma_t^2)(a_{t-j}^2 - \sigma_{t-j}^2)] &= E[E\{(a_t^2 - \sigma_t^2)(a_{t-j}^2 - \sigma_{t-j}^2) \mid F_{t-1}\}] \\ &= E[(a_{t-j}^2 - \sigma_{t-j}^2)E\{(a_t^2 - \sigma_t^2) \mid F_{t-1}\}] = 0 \end{aligned}$$

Further, since $\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2$, we find that the ARCH(1) model can be written as

$$a_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + v_t \tag{10.2.7}$$

This form reveals that the process of squared errors a_t^2 can be viewed as an AR(1) model with uncorrelated innovations v_t. The innovations are heteroscedastic and also non-Gaussian in this case, however.

For the ARCH(s) model in (10.2.2), we similarly have

$$a_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_s a_{t-s}^2 + v_t$$

so that the a_t^2 has the form of an AR(s) process. Other results related to the moments and the kurtosis of the ARCH(1) model also extend to higher order ARCH models. In particular, if $\sum_{i=1}^{s} \alpha_i < 1$, then the unconditional variance is

$$\sigma_a^2 = \frac{\alpha_0}{1 - \sum_{i=1}^{s} \alpha_i}$$

as shown by Engle (1982). Necessary and sufficient conditions for the existence of higher order even moments of the ARCH(s) process were given by Milhøj (1985).

Forecast Errors for the ARCH Model. Forecasts of a future value z_{t+l} generated from ARMA models with iid errors a_t have forecast errors that depend on the lead time l but are independent of the time origin t from which the forecasts are made. Baillie and Bollerslev (1992) showed that the minimum mean square error forecasts of z_{t+l} are the same irrespective of whether the shocks a_t are heteroscedastic or not. For an ARMA process with ARCH errors, this implies, in particular, that the one-step-ahead forecast error equals a_{t+1} while the l-step-ahead forecast error can be written as $e_t(l) = \sum_{j=0}^{l-1} \psi_j a_{t+l-j}$ with $\psi_0 = 1$. The presence of conditional heteroscedasticity will, however, impact the variance of the forecast errors.

For an ARCH(1) process, the conditional variance of the one-step-ahead forecast error a_{t+1} is given by (10.2.5) as

$$E[e_t^2(1)|F_t] = \sigma_{t+1}^2 = \sigma_a^2 + \alpha_1(a_t^2 - \sigma_a^2) \tag{10.2.8}$$

The conditional variance of the one-step-ahead forecast error can thus be smaller or larger than the unconditional variance depending on the difference between the last squared error a_t^2 and σ_a^2.

Conditional variances of multistep-ahead forecast errors $e_t(l)$ can also be shown to depend on the past squared errors based on

$$E[e_t^2(l) \mid F_t] = \sum_{j=0}^{l-1} \psi_j^2 E[a_{t+l-j}^2 \mid F_t]$$

where for the ARCH(1) model

$$\begin{aligned} E[a_{t+h}^2|F_t] &= E[E(a_{t+h}^2 \mid F_t)] \\ &= \alpha_0 + \alpha_1 E[a_{t+h-1}^2 \mid F_t] \\ &= \alpha_0(1 + \alpha_1 + \cdots + \alpha_1^{h-1}) + \alpha_1^h a_t^2 \qquad \text{for} \quad h > 0 \end{aligned}$$

From this and using (10.2.4) it can be verified that

$$E[e_t^2(l) \mid F_t] = \sigma_a^2 \sum_{j=0}^{l-1} \psi_j^2 + \sum_{j=0}^{l-1} \psi_j^2 \alpha_1^{l-j}(a_t^2 - \sigma_a^2) \tag{10.2.9}$$

which simplifies to (10.2.8), for $l = 1$. The first term on the right-hand side of this expression is the conventional prediction error variance assuming that the errors a_t are homoscedastic while the second term reflects the impact of the ARCH effects. This term varies over time and can again be positive or negative depending on the difference $a_t^2 - \sigma_a^2$. The variance of the predicted values thus varies over time and can be larger or smaller than that under homoscedasticity. For the general ARCH(s) model, the second term on the right-hand side will be a function of s past values $a_t^2, \ldots, a_{t-s+1}^2$.

If the time series z_t follows an AR(1) model, the ψ weights are given by $\psi_j = \phi^{j-1}$. If ϕ equals zero, so that the mean of the series is a constant independent of the past, expression (10.2.9) simplifies to $\sigma_a^2 + \alpha_1^l(a_t^2 - \sigma_a^2)$. We note that this is the conditional l-step-ahead forecast of the conditional variance σ_{t+l}^2 for the ARCH(1) model. This forecast

could be calculated more directly as $E[\sigma^2_{t+l}|F_t] = \alpha_0 + \alpha_1 E[a^2_{t+l-1}|F_t]$, where $E[a^2_{t+l-1}|F_t]$ can be generated recursively from the AR model for a^2_t. The result follows by setting $\alpha_0 = \sigma^2_a(1-\alpha_1)$.

10.2.2 The GARCH Model

The ARCH model has a disadvantage in that it often requires a high lag order s to adequately describe the evolution of volatility over time. An extension of the ARCH model called the *generalized* ARCH, or GARCH, model was introduced by Bollerslev (1986) to overcome this issue. The GARCH(s, r) model assumes that $a_t = \sigma_t e_t$, where the $\{e_t\}$ again are iid random variables with mean zero and variance 1, and where σ_t is given by

$$\sigma^2_t = \alpha_0 + \sum_{i=1}^{s} \alpha_i a^2_{t-i} + \sum_{j=1}^{r} \beta_j \sigma^2_{t-j} \tag{10.2.10}$$

with $\alpha_0 > 0$, $\alpha_i \geq 0$, $i = 1, \ldots, s-1$, $\alpha_s > 0$, $\beta_j \geq 0$, $j = 1, \ldots, r-1$, and $\beta_r > 0$. These parameter constraints are sufficient for the conditional variance σ^2_t to be positive. Nelson and Cao (1992) showed that these constraints can be relaxed slightly to allow some of the parameters to be negative while the conditional variance still remains positive. The additional constraint $\sum_{i=1}^{m}(\alpha_i + \beta_i) < 1$, where $m = \max(s, r)$ with $\alpha_i = 0$, for $i > s$, and $\beta_j = 0$, for $j > r$, ensures that the unconditional variance σ^2_a is finite.

The simplest and most widely used model in this class is the GARCH(1, 1) model where

$$\sigma^2_t = E[a^2_t \mid F_{t-1}] = \alpha_0 + \alpha_1 a^2_{t-1} + \beta_1 \sigma^2_{t-1}$$

Since the constants α_1 and β_1 are positive, we see that a large value of a^2_{t-1} or σ^2_{t-1} results in a large value of σ^2_t. As for the ARCH process, this model therefore accounts for volatility clustering.

Assuming that $\alpha_1 + \beta_1 < 1$, the unconditional variance of a_t is

$$\sigma^2_a = \text{var}[a_t] = \alpha_0/[1 - (\alpha_1 + \beta_1)]$$

Also, assuming that the conditional distributions are normal, the fourth unconditional moment of a_t is finite provided that $(\alpha_1 + \beta_1)^2 + 2\alpha^2_1 < 1$ (Bollerslev, 1986). In addition, the kurtosis of the marginal distribution of a_t equals

$$\kappa = \frac{E(a^4_t)}{[E(a^2_t)]^2} = \frac{3[1 - (\alpha_1 + \beta_1)^2]}{1 - (\alpha_1 + \beta_1)^2 - 2\alpha^2_1} > 3$$

As in the ARCH case, the unconditional distribution of a_t thus has heavier tails than the normal distribution and is expected to give rise to a higher frequency of extreme observations or ''outliers'' than would be the case under normality.

Now let $v_t = a^2_t - \sigma^2_t$ so that $\sigma^2_t = a^2_t - v_t$, where the v_t have zero mean and are serially uncorrelated. We then see that the GARCH(1, 1) model can be rearranged as $a^2_t - v_t = \alpha_0 + \alpha_1 a^2_{t-1} + \beta_1(a^2_{t-1} - v_{t-1})$, or

$$a^2_t = \alpha_0 + (\alpha_1 + \beta_1)a^2_{t-1} + v_t - \beta_1 v_{t-1} \tag{10.2.11}$$

The process of *squared errors* thus has the form of an ARMA(1, 1) model with uncorrelated innovations v_t. The v_t are in general heteroscedastic, however. In the special case of $\beta_1 = 0$, the model reduces to $a_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + v_t$, which is the AR(1) form of the ARCH(1) model. For the general GARCH(s, r) process, expression (10.2.11) generalizes to

$$a_t^2 = \alpha_0 + \sum_{i=1}^{m}(\alpha_i + \beta_i)a_{t-i}^2 + v_t - \sum_{i=1}^{s}\beta_i v_{t-i}$$

which has the form of an ARMA process for a_t^2 with AR order equal to $m = \max(r, s)$. The autocorrelation structure of a_t^2 also mimics that of the ARMA process provided that fourth unconditional moment of a_t is finite (Bollerslev, 1988).

The necessary and sufficient condition for second-order stationarity of the GARCH(s, r) process is

$$\sum_{i=1}^{s}\alpha_i + \sum_{i=1}^{r}\beta_i = \sum_{i=1}^{m}(\alpha_i + \beta_i) < 1$$

When this condition is met, the unconditional variance is

$$\sigma_a^2 = \text{var}[a_t] = \alpha_0 \Big/ \left[1 - \sum_{i=1}^{m}(\alpha_i + \beta_i)\right]$$

This was shown by Bollerslev (1986) who also gave necessary and sufficient conditions for the existence of all higher order moments for the GARCH(1, 1) model and the fourth-order moments for GARCH(1, 2) and GARCH(2, 1) models. Extensions of these results have been given by He and Teräsvirta (1999) and Ling and McAleer (2002), among others. The expressions for the higher order moments and the constraints on the parameters needed to ensure their existence become more complex for the higher order models. The model specification also becomes more difficult. On the other hand, numerous studies have shown that low-order models such as the GARCH(1, 1), GARCH(2, 1), and GARCH(1, 2) models are often adequate in practice, with the GARCH(1, 1) model being the most popular.

10.2.3 Model Building and Parameter Estimation

Testing for ARCH/GARCH Effects. The preceding results motivate the use of the ACF and PACF of the squares a_t^2 for model specification and for basic preliminary checking for the presence of ARCH/GARCH effects in the errors a_t. For an ARMA model with heteroscedastic errors, a starting point for the analysis is an examination of the sample ACF and PACF of the squared residuals $\hat{a}_t^2$ obtained from fitting an ARMA model to the observed series. In particular, let $r_k(\hat{a}^2)$ denote the sample autocorrelations of the squared residuals $\hat{a}_t^2$ so that

$$r_k(\hat{a}^2) = \sum_{t=1}^{n-k}(\hat{a}_t^2 - \hat{\sigma}_a^2)(\hat{a}_{t+k}^2 - \hat{\sigma}_a^2) \Big/ \sum_{t=1}^{n}(\hat{a}_t^2 - \hat{\sigma}_a^2)^2$$

where $\hat{\sigma}_a^2 = n^{-1}\sum_{t=1}^{n}\hat{a}_t^2$ is the residual variance estimate. Analogous to the modified portmanteau statistic described in Section 8.2.2, McLeod and Li (1983) proposed the

portmanteau statistic

$$\tilde{Q}(\hat{a}^2) = n(n+2)\sum_{k=1}^{K} r_k^2(\hat{a}^2)/(n-k) \tag{10.2.12}$$

to detect departures from the ARMA assumptions. As a portmanteau test, this test does not assume a specific alternative, but the type of departures for which $\tilde{Q}(\hat{a}^2)$ can be useful includes conditional heteroscedasticity in the form of ARCH/GARCH effects, and bilinear type of nonlinearity in the conditional mean of the process (see Section 10.3 for discussion of bilinear models). McLeod and Li (1983) showed that the statistic $\tilde{Q}(\hat{a}^2)$ has approximately the χ^2 distribution with K degrees of freedom under the assumption that the ARMA model alone is adequate. The distribution is similar to that of the usual portmanteau statistic $\tilde{Q}$ based on the residuals $\hat{a}_t$, with the exception that the degrees of freedom in the case of (10.2.12) are *not affected* by the fact that $p+q$ ARMA parameters have been estimated. The potentially more powerful portmanteau statistics by Peña and Rodríguez (2002, 2006) discussed in Section 8.2 could also be applied to the squared residuals $\hat{a}_t^2$.

An alternative test for ARCH effects is the score or Lagrange multiplier test proposed by Engle (1982). The score statistic Λ for testing the null hypothesis H_0: $\alpha_i = 0$, $i = 1, \ldots, s$, has a convenient form and can be expressed as n times the coefficient of determination in the least-squares fitting of the auxiliary regression equation

$$\hat{a}_t^2 = \alpha_0 + \alpha_1\hat{a}_{t-1}^2 + \alpha_2\hat{a}_{t-2}^2 + \cdots + \alpha_s\hat{a}_{t-s}^2 + \varepsilon_t$$

Assuming normality of the a_t's, the score statistic Λ has an asymptotic χ^2 distribution with s degrees of freedom under the null model of no ARCH effects. The test procedure is thus to fit a time series model to the observed series, save the residuals $\hat{a}_t$, and regress the squared residuals on a constant and s lagged values of the $\hat{a}_t^2$. The resulting value of nR^2 is then referred to a χ^2 distribution with s degrees of freedom. Even though this test was derived for the ARCH(s) model, it has been shown to be useful for detecting other forms of conditional heteroscedasticity as well. Also, the test is asymptotically equivalent to the McLeod–Li portmanteau test based on the autocorrelations of the squared residuals (see Luukkonen et al., 1988b). Thus, although the latter was derived as a pure significance test, it is also a LM test against ARCH effects.

Parameter Estimation. The parameter estimation for models with ARCH or GARCH errors is typically performed using the conditional maximum likelihood method. For estimation of an ARMA model $\phi(B)z_t = \theta_0 + \theta(B)a_t$ with ARCH or GARCH errors a_t, we assume that a_t is conditionally normally distributed as $N(0, \sigma_t^2)$. The z_t are then conditionally normal, given $z_{t-1}, z_{t-2}, \ldots$, and from the joint density function $p(\boldsymbol{z}) = \prod_{t=1}^{n} p(z_t \mid z_{t-1}, \ldots, z_1)$ we obtain the log-likelihood function

$$l = \log(L) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\log(\sigma_t^2) - \frac{1}{2}\sum_{t=1}^{n} a_t^2/\sigma_t^2 \tag{10.2.13}$$

where $a_t = z_t - \sum_{i=1}^{p}\phi_i z_{t-i} - \theta_0 + \sum_{i=1}^{q}\theta_i a_{t-i}$ and σ_t^2 is given by (10.2.2) or (10.2.10). A discussion of the iterative maximization of the likelihood function along with other results related to the parameter estimation can be found, for example, in Engle (1982), Weiss (1984, 1986), and Bollerslev (1986). When an ARMA model with ARCH or GARCH errors

is fitted to the series, the information matrix of the log-likelihood is block diagonal with respect to the conditional mean and variance parameters, so that iterations can be carried out separately with respect to the two sets of parameters. The so-called BHHH algorithm by Berndt, Hall, Hall, and Hausman (1974) provides a convenient method to perform the calculations. This algorithm has the advantage that only first-order derivatives are needed for the optimization. These derivatives can be evaluated numerically or analytically. Use of analytical first derivatives is often recommended as they improve the precision of the parameter estimates. Provided that the fourth-order moment of the process is finite, the resulting estimates of the ARMA–ARCH parameters are consistent and asymptotically normal as shown by Weiss (1986).

The normal distribution was originally proposed by Engle (1982) to model the conditional distribution of the disturbances a_t. As discussed earlier, the conditional normal distribution results in a leptokurtic unconditional distribution. Nevertheless, in financial applications the normal distribution sometimes fails to capture the excess kurtosis that is present in stock returns and other variables. To overcome this drawback, Bollerslev (1987) suggested using a standardized Student t-distribution with $\nu > 2$ degrees of freedom for the estimation. The density function of the t-distribution is

$$f(x|\nu) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\pi(\nu-2)}}\left(1+\frac{x^2}{(\nu-2)}\right)^{-(\nu+1)/2}$$

where $\Gamma(\nu) = \int_0^\infty e^{-x}x^{\nu-1}dx$ is the Gamma function and ν measures the tail thickness. As is well known, the distribution is symmetric around zero and approaches a normal distribution as $\nu \to \infty$. For $\nu > 4$, the fourth moment exists and the conditional kurtosis equals $3(\nu-2)/(\nu-4)$. Since this value exceeds 3, the tails are heavier than those of the normal distribution. The log-likelihood function based on the t-distribution is given by

$$\begin{aligned} l = \log(L) = n\left[\log\Gamma\left(\frac{\nu+1)}{2}\right) - \log\left(\frac{\nu}{2}\right) - \frac{1}{2}\log(\pi(\nu-2))\right] \\ -\frac{1}{2}\sum_{t=1}^{n}\left[\log(\sigma_t^2) + (1+\nu)\log\left(1+\frac{a_t^2}{(\nu-2)\sigma_t^2}\right)\right] \end{aligned}$$

Here, ν is either prespecified or estimated jointly with other parameters. If ν is specified in advance, values between 5 and 8 are often used; see Tsay (2010). With ν prespecified, the conditional likelihood function is maximized by minimizing the second term of the likelihood function given above.

Nelson (1991) suggested using the generalized error distribution (GED) for the estimation. The density function of a GED random variable normalized to have mean zero and variance one is given by

$$f(x|\eta) = \frac{\eta\exp(-0.5|x/\lambda|^\eta)}{\lambda 2^{(1+1/\eta)}\Gamma(1/\eta)}$$

where $\lambda = [2^{(-2/\eta)}\Gamma(1/\eta)/\Gamma(3/\eta)]^{1/2}$. For the tail thickness parameter $\eta = 2$, the distribution equals the normal distribution used in (10.2.13). For $\eta < 2$, the distribution has thicker tails than the normal distribution. The reverse is true for $\eta > 2$. Box and Tiao (1973) call the GED distribution an exponential power distribution.

In addition to having excess kurtosis, the distribution of a_t may also be skewed. A discussion of potential sources for skewness can be found in He et al. (2008). To allow for skewness as well as heavy tails, the likelihood calculations can be based on skewed versions for the Student t-distribution and the GED distributions available in software packages such as R. Other forms of skewed distributions have also been considered.

In practice, it is often difficult to know whether the specified probability distribution is the correct one. An alternative approach is to continue to base the parameter estimation on the normal likelihood function in (10.2.13). This method is commonly referred to as the quasi-maximum likelihood (QML) estimation. The asymptotic properties of the resulting QML estimator for the ARCH, GARCH, and ARMA–GARCH models have been studied by many authors with early contributions provided by Weiss (1986) and Bollerslev and Wooldridge (1992). For further discussion and references, see, for example, Francq and Zakoïan (2009, 2010).

Diagnostic Checking. Methods for model checking include informal graphical checks using time series plots and Q–Q plots of the residuals along with a study of their dependence structure. The assumption underlying the ARCH and GARCH models is that the standardized innovations a_t/σ_t are independent and identically distributed. Having estimated the parameters of model, the adequacy of the mean value function can be checked by examining the autocorrelation and partial autocorrelation functions of the standardized residuals $\hat{a}_t/\hat{\sigma}_t$. Similar checks on the autocorrelation and partial autocorrelations of the squared standardized residuals are useful for examining the adequacy of the volatility model. These checks are often supplemented by the portmanteau test proposed by McLeod and Li (1983) or the score test proposed by Engle (1982). However, while these statistics can provide useful indications of lack of fit, their asymptotic distributions are impacted by the estimation of the ARCH or GARCH parameters. Li and Mak (1994) derived an alternative portmanteau statistic that asymptotically follows the correct χ^2_K distribution. This statistic is a quadratic form in the first m autocorrelations of the squared standardized residuals but has a more complex form than the $\tilde{Q}$ statistic in (10.2.12). Analogous modifications of Engle's score test based on ARCH residuals were discussed by Lundbergh and Teräsvirta (2002). More recent contributions to model checking include work by Wong and Ling (2005), Ling and Tong (2011), Fisher and Gallagher (2012), and many others.

10.2.4 An Illustrative Example: Weekly S&P 500 Log Returns

To demonstrate the model building process, we consider the weekly log returns on the S&P 500 Index displayed in Figure 10.1(b) for the period January 3, 2000 to May 27, 2014. Figure 10.2 shows the ACF of the returns along with the ACF of the squared returns. We note that there is little, if any, serial correlation in the returns themselves. The mean value function μ_t will thus be taken as a constant. However, the squared returns are clearly correlated and show a pattern consistent with that of an ARCH or a GARCH model. The PACF of the squared returns (not shown) has a pattern that persists over several lags suggesting that a GARCH may be appropriate for the volatility.

The parameters can be estimated in R using the function garchFit() in the fGarch package. The normal distribution is the default error distribution for the ARCH or GARCH models. Other options include the Student t-distribution and the GED distributions along with skewed versions of these distributions. For demonstration, we will fit a GARCH(1, 1)

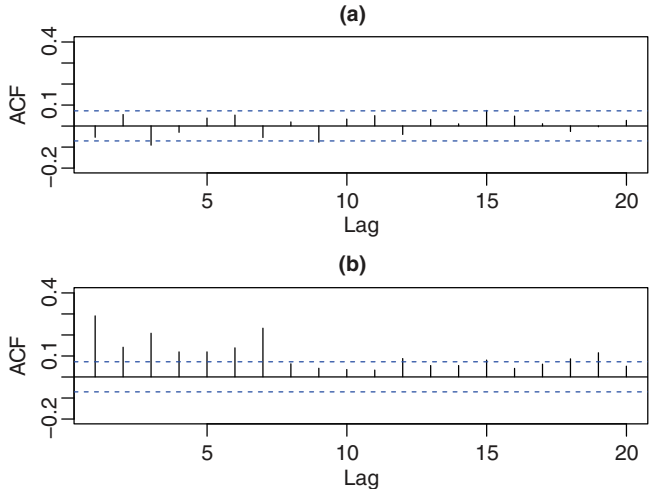


FIGURE 10.2 Autocorrelation functions for (a) the S&P 500 weekly log returns and (b) the squared weekly log returns.

model with normal errors to the returns. The R commands and a partial model output are provided below, where the log returns are denoted by SPrtn:

```
  >library(fGarch)
  >m1=garchFit(~garch(1,1),data=SPrtn,trace=F)
  >summary(m1) % Retrieve model output

Title: GARCH Modelling
Call: garchFit(formula=~garch(1,1),data=SPrtn, trace=F)

Mean and Variance Equation:  data ~ garch(1,1)
Conditional Distribution:  norm

Coefficient(s):
     mu       omega      alpha1       beta1
 2.1875e-03  3.5266e-05  2.1680e-01  7.3889e-01

 Error Analysis:
        Estimate   Std. Error   t value   Pr(>|t|)
 mu     2.187e-03   6.875e-04    3.182    0.00146 **
 omega  3.527e-05   1.153e-05    3.058    0.00223 **
 alpha1 2.168e-01   4.189e-02    5.176   2.27e-07 ***
 beta1  7.389e-01   4.553e-02   16.230    < 2e-16 ***

 Standardised Residuals Tests:
                                Statistic    p-Value
 Jarque-Bera Test      Chi^2    77.92548       0
 Shapiro-Wilk Test  R    W       0.9815283 3.990011e-08
 Ljung-Box Test     R    Q(10)   6.910052  0.7339084
 Ljung-Box Test     R    Q(20)   16.43491  0.689303
```

```
Ljung-Box Test      R^2  Q(10)    12.64346  0.244295
Ljung-Box Test      R^2  Q(20)    18.15442  0.5772367
LM Arch Test        R    TR^2     14.05565  0.297169

Information Criterion Statistics:
  AIC         BIC         SIC       HQIC
-4.751772 -4.727132 -4.751829 -4.742278
```

Letting w_t denote the log returns, the fitted model is

$$w_t = 0.002187 + a_t, \quad \sigma_t^2 = 0.000035 + 0.2168a_{t-1}^2 + 0.7389\sigma_{t-1}^2$$

where all the parameter estimates are statistically significant. The portmanteau tests for serial correlation in the standardized residuals and in their squared values indicate no lack of fit. However, the Jarque–Bera and Shapiro–Wilk tests for normality suggest that the model is not fully adequate. To examine this issue, the Student t-distribution and its skewed version were tested by adding the argument cond.dist="std" and cond.dist ="sstd", respectively, to the garchFit command. The GED distribution and its skewed version were also tested. Although these modifications improved the fit, the results are for simplicity not shown here.

The standardized residuals from the fitted model and the ACF of the squared standardized residuals are shown in Figure 10.3. A normal Q–Q plot is also included in this graph. Visual inspection of the standardized residuals and the Q–Q plot confirms the results of the normality tests discussed above. The ACF of the squared residuals indicates no lack of fit although a marginally significant correlation is present at lag 1. This value would be reduced by fitting a GARCH(1, 2) model to the data. But this potential refinement is not pursued here. Finally, estimates of the conditional standard deviation σ_t are displayed in Figure 10.4(a). Figure 10.4(b) displays the volatility shown earlier in Figure 10.1(b) with two standard deviation limits now superimposed around the series. A variety of other graphs can be generated using the R command plot(m1), where m1 refers to the fitted model. In addition, l-step-ahead forecasts of future volatility based on the conditional standard deviations shown in Figure 10.4 can be generated using the R command predict(m1,l).

10.2.5 Extensions of the ARCH and GARCH Models

While the ARCH and GARCH models allow for volatility clustering and capture thick-tailed behavior of the underlying unconditional distributions, they do not account for certain other features that are commonly observed in financial data. For example, so-called leverage effects are often observed in stock returns, where a negative innovation tends to increase the volatility more than a positive innovation of the same magnitude. In symmetric ARCH and GARCH models, on the other hand, the variance depends on the magnitude of the innovations but not their signs. Another limitation of the basic ARCH and GARCH models is the assumption that the conditional mean of the process is unaffected by the volatility. This assumption ignores the so-called risk premium that relates to the fact that investors expect to receive higher returns as compensation for taking on riskier assets. The presence of this feature would generate a positive relationship between expected return and volatility. Below we describe some extensions and modifications of the ARCH and GARCH models that have been proposed to address such issues.

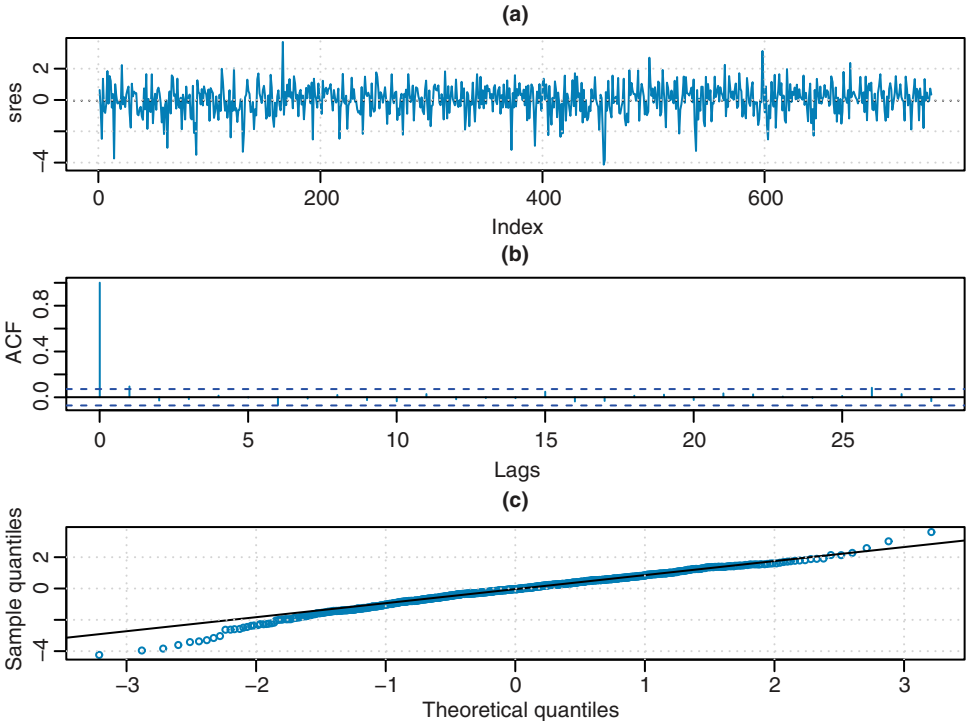


FIGURE 10.3 Model diagnostics for the GARCH(1, 1) model fitted to the S&P 500 weekly log returns: (a) standardized residuals, (b) autocorrelation function of the squared standardized residuals, and (c) a normal Q–Q plot of the standardized residuals.

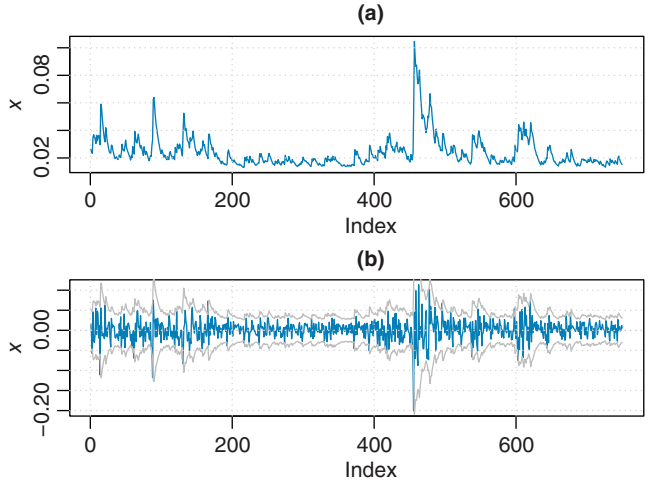


FIGURE 10.4 Conditional standard deviations for the S&P 500 weekly log returns (a) and the weekly log returns with two standard deviation limits imposed (b).

Exponential GARCH Models. The earliest model that allows for an asymmetric response due to leverage effects is the exponential GARCH, or EGARCH, model introduced by Nelson (1991). The EGARCH(1, 1) model is defined as $a_t = \sigma_t e_t$, where

$$\ln(\sigma_t^2) = \alpha_0 + g(e_{t-1}) + \beta_1 \ln(\sigma_{t-1}^2)$$

The function $g(e_{t-1})$ determines the asymmetry and is defined as the weighted innovation

$$g(e_{t-1}) = \alpha_1 e_{t-1} + \gamma_1[|e_{t-1}| - E(|e_{t-1}|)]$$

where α_1 and γ_1 are real constants. The model then becomes

$$\ln(\sigma_t^2) = \alpha_0 + \alpha_1 e_{t-1} + \gamma_1 |e_{t-1}| - \gamma_1 E(|e_{t-1}|) + \beta_1 \ln(\sigma_{t-1}^2)$$

From here it is easy to see that a positive shock has the effect $(\alpha_1 + \gamma_1)e_{t-1}$ while a negative shock has the effect $(\alpha_1 - \gamma_1)e_{t-1}$. The use of $g(e_{t-1})$ thus allows the model to respond asymmetrically to ''good news'' and ''bad news.'' Since bad news typically has a larger impact on volatility than good news, the value of α_1 is expected to be negative when leverage effects are present. Note that since the EGARCH model describes the relation between the logarithm of the conditional variance σ_t^2 and past information, the model does not require any restrictions on the parameters to ensure that σ_t^2 is nonnegative. The general EGARCH(s, r) model has the form

$$\ln(\sigma_t^2) = \alpha_0 + \sum_{i=1}^{s} g_i(e_{t-i}) + \sum_{j=1}^{r} \beta_j \ln(\sigma_{t-j}^2)$$

with

$$g_i(e_{t-i}) = \alpha_i e_{t-i} + \gamma_i(|e_{t-i}| - E(|e_{t-i}|)$$

However, as in the GARCH case, the first-order model is the most popular in practice.

Nelson (1991) specified the likelihood function assuming that the errors follow a generalized error distribution that includes the normal distribution as a special case. Properties of the QML estimator based on the normality assumption for the EGARCH(1, 1) model were studied by Straumann and Mikosch (2006) who verified the conditions for consistency of this estimator. Further properties and details related to the model building process can be found in Tsay (2010) and Teräsvirta et al. (2010), for example.

The GJR and Threshold GARCH Models. The so-called GJR-GARCH model of Glosten, Jagannathan, and Runkle (1993) and the threshold GARCH model of Zakoïan (1994) provide an alternative way to allow for asymmetric effects of positive and negative volatility shocks. Starting from the GARCH(1, 1) model, the GJR model assumes that the parameter associated with a_{t-1}^2 depends on the sign of the shock so that

$$\sigma_t^2 = \alpha_0 + (\alpha_1 + \gamma_1 I_{t-1})a_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

where the indicator variable I_{t-1} assumes the value 1 if a_{t-1} is negative and zero if it is positive. The constraints on the parameters needed to ensure that the conditional variance σ_t^2 is nonnegative are readily derived from those of the GARCH(1, 1) process. Using this formulation, the noise term a_{t-1} has a coefficient $\alpha_1 + \gamma_1$ when it is negative, and α_1 when it is positive. This allows negative shocks to have a larger impact on the volatility. The

GJR model is relatively simple and empirical studies have shown that the model performs well in practice. For general GARCH(s, r), the model generalizes to

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{s}(\alpha_i + \gamma_i I_{t-i})a_{t-i}^2 + \sum_{j=1}^{r}\beta_j\sigma_{t-j}^2$$

although applications with r and s greater than 1 seem to be very rare. Zakoïan (1994) introduced a model with the same functional form as the GJR model, but instead of modeling the conditional variance, Zakoïan models the conditional standard deviation. Since the coefficient associated with a_{t-1} changes its value as a_{t-1} crosses the *threshold* zero, Zakoïan referred to this model as a threshold GARCH, or TGARCH, model.

Nonlinear Smooth Transition Models. For the threshold model described above, the impact of past shocks changes abruptly as a_{t-i} crosses the zero threshold. Attempts have been made in the literature to develop nonlinear extensions of ARCH and GARCH models that allow for more flexibility and a smoother transition as a lagged value a_{t-i} crosses a specified threshold. These extensions include the logistic smooth transition GARCH model proposed by Hagerud (1997), and a similar model proposed independently by González-Rivera (1998). This model assumes that the model parameters α_i in the ARCH or GARCH model are not constant but functions of the lagged a_{t-i} so that $\alpha_i = \alpha_{1i} + \alpha_{2i}F(a_{t-i})$, $i = 1, \ldots, s$, where $F(\cdot)$ is a transition function. Hagerud considered two transition functions, the logistic and the exponential. The GARCH(s, r) model with a logistic transition function has the form

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{s}[\alpha_{1i} + \alpha_{2i}F(a_{t-i})]a_{t-i}^2 + \sum_{j=1}^{r}\beta_j\sigma_{t-j}^2$$

where

$$F(a_{t-i}) = \frac{1}{1 + \exp(-\theta a_{t-i})} - \frac{1}{2}$$

with $\theta > 0$. In contrast to the GJR model that follows one process when the innovations are positive and another process when the innovations are negative, the transition between the two states is smooth in the present model. Hagerud provided conditions for stationarity and nonnegativity of the conditional variances.

Lanne and Saikkonen (2005) proposed a smooth transition GARCH process that uses the lagged conditional variance σ_{t-1}^2 as the transition variable, and is suitable for describing high persistence in the conditional variance. The first-order version of this model can be written as

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \delta_1 G_1(\theta; \sigma_{t-1}^2) + \beta_1\sigma_{t-1}^2$$

where the transition function $G_1(\theta; \sigma_{t-1}^2)$ is a continuous, monotonically increasing bounded function of σ_{t-1}^2. Lanne and Saikkonen used the cumulative distribution function of the gamma distribution as the transition function. The original purpose for introducing this model was to remedy a tendency of GARCH models to exaggerate the persistence in volatility as evidenced by $\Sigma(\alpha_i + \beta_i)$ often being very close to one. Using empirical examples involving exchange rates, the authors showed that this formulation alleviates the problem

of exaggerated persistence. For further discussion of these and related models, see, for example, Mills and Markellos (2008) and Teräsvirta (2009).

GARCH-M Models. Many theories in finance postulate a direct relationship between the expected return on an investment and its risk. To account for this, the GARCH-in-mean, or GARCH-M, model, allows the conditional mean of a GARCH process to depend on the conditional variance σ_t^2. This model originates from the ARCH-M model proposed by Engle et al. (1987). The mean value function is specified as

$$\mu_t = \beta_0 + \beta_1 g(\sigma_t^2)$$

where $g(\sigma_t^2)$ is a positive-valued function and β_1 is a positive constant called the risk premium parameter. An increase or decrease in the conditional mean is here associated with the sign of the partial derivative of the function $g(\sigma_t^2)$ with respect to σ_t^2. In many applications, $g(\sigma_t^2)$ is taken to be the identity function or the square root function so that $g(\sigma_t^2) = \sigma_t^2$ or $g(\sigma_t^2) = \sigma_t$. The parameters of the GARCH-M model can be estimated using the maximum likelihood method. However, because of the dependence of the conditional mean on the conditional variance, the information matrix is no longer block diagonal with respect to the conditional mean and variance parameters. This makes joint maximization of the likelihood function with respect to the two sets of parameters necessary. Also, consistent estimation of the parameters in the GARCH-M models requires the full model be correctly specified. Applications of the GARCH-M model to stock returns, exchange rates, and interest rates were discussed by Bollerslev et al. (1992).

IGARCH and FIGARCH Models. As noted earlier, the GARCH(1, 1) model is weakly stationary assuming that $(\alpha_1 + \beta_1) < 1$. When the GARCH model is applied to high-frequency financial data, it is often found that $\alpha_1 + \beta_1$ is close to or equal to1. Engle and Bollerslev (1986) refer to a model with $\alpha_1 + \beta_1 = 1$ as an integrated GARCH, or IGARCH, model. The motivation is that this implies a unit root in the autoregressive part of the ARMA(1, 1) representation of the GARCH(1, 1) model for a_t^2 in (10.2.11). With $\alpha_1 + \beta_1 = 1$, the model becomes $(1 - B)a_t^2 = \alpha_0 + \nu_t - \beta_1 \nu_{t-1}$. Similar to a random walk process, this process is not mean reverting since the unconditional variance of the process is not finite. Also, the impact of a large shock on the forecasts of future values will not diminish for increasing lead times. But while the GARCH(1,1) process is not weakly stationary, Nelson (1990) showed that the process has time-invariant probability distributions and is thus strictly stationary. A necessary condition for strict stationarity is $E[ln(\alpha_1 a_{t-1}^2 + \beta_1)] < 0$. For further discussion of this model, see, for example, Teräsvirta (2009).

Fractionally integrated GARCH, or FIGARCH, models have also been proposed in the literature. These differ from the IGARCH model in that the degree of differencing d is allowed to be a fraction rather than a constant. The FIGARCH(1, 1) model, in particular, is of the form $(1 - B)^d a_t^2 = \alpha_0 + \nu_t - \beta_1 \nu_{t-1}$, where d is a constant such that $0 < d < 0.5$. For the FIGARCH model, the empirical autocorrelations of a_t^2 need not be very large but they decay very slowly as the lag k increases. This is indicative of so-called *long memory* behavior in the series. Models involving fractional differencing will be discussed further in Section 10.4 in relation to long-range dependence in the conditional mean μ_t.

Other Models. Numerous other models have been proposed to account for conditional heteroscedasticity. For example, a natural extension of the ARCH(s) model specified in

(10.2.1) is to let $\sigma_t^2 = \alpha_0 + \boldsymbol{a}_{t-1}'\boldsymbol{\Omega}\boldsymbol{a}_{t-1}$, where $\boldsymbol{a}_{t-1} = (a_{t-1}, \ldots, a_{t-s})'$ and $\boldsymbol{\Omega}$ is a $s \times s$ nonnegative definite matrix. The ARCH(s) model is then a special case that requires that $\boldsymbol{\Omega}$ be diagonal. One way that the above form can arise is through the conditional heteroscedastic ARMA (CHARMA) model specification discussed by Tsay (1987). Other approaches to volatility modeling include the random coefficient autoregressive model of Nicholls and Quinn (1982) and the stochastic volatility models of Melino and Turnbull (1990), Jacquier et al. (1994), and Harvey et al. (1994). A brief description of the stochastic volatility models is provided below.

10.2.6 Stochastic Volatility Models

Stochastic volatility models are similar to GARCH models but introduce a stochastic innovation term to the equation that describes the evolution of the conditional variance σ_t^2. To ensure positiveness of the conditional variances, stochastic volatility models are defined in terms of $\ln(\sigma_t^2)$ instead of σ_t^2. A basic version of a stochastic volatility model is defined by $a_t = \sigma_t e_t$ as in (10.2.1) with $\ln(\sigma_t^2)$ satisfying

$$\ln(\sigma_t^2) = \alpha_0 + \beta_1 \ln(\sigma_{t-1}^2) + \cdots + \beta_r \ln(\sigma_{t-r}^2) + \upsilon_t \tag{10.2.14}$$

where e_t are iid normal $N(0, 1)$, υ_t are iid normal $N(0, \sigma_\upsilon^2)$, $\{e_t\}$ and $\{\upsilon_t\}$ are independent processes, and the roots of the characteristic equation $1 - \sum_{j=1}^{r} \beta_i B^j = 0$ are outside the unit circle. Note, for example, the stochastic volatility model equation for $r = 1$ is $\ln(\sigma_t^2) = \alpha_0 + \beta_1 \ln(\sigma_{t-1}^2) + \upsilon_t$, which is somewhat analogous to the GARCH(1, 1) model equation, $\sigma_t^2 = \alpha_0 + \beta_1\sigma_{t-1}^2 + \alpha_1 a_{t-1}^2$. Alternatively, replacing $g(e_{t-1})$ by υ_t in the EGARCH(1, 1) model, we obtain (10.2.14) with $r = 1$. Some properties of the stochastic volatility model for $r = 1$ are provided by Jacquier et al. (1994). Also note that we may write $a_t^2 = \sigma_t^2 e_t^2$ so that $\ln(a_t^2) = \ln(\sigma_t^2) + \ln(e_t^2)$. This allows the stochastic volatility model to be viewed as a state-space model, with the last relation representing the observation equation and the transition equation being developed from (10.2.14). Difficulty in parameter estimation is increased for stochastic volatility models, however, since likelihoods based on the state-space model are non-Gaussian. Quasi-likelihood methods may thus be needed. Jacquier et al. (1994) give a good summary of estimation techniques, including quasi-likelihood methods with Kalman filtering and the expectation maximization (EM) algorithm and Markov chain Monte Carlo (MCMC) methods. They also provide a comparison of estimation results between the different methods.

A discussion and examples of the use of Markov chain Monte Carlo methods for parameter estimation can also be found in Tsay (2010, Chapter 12). A general overview of the stochastic volatility literature is given by a collection of articles in the books edited by Shephard (2005) and Andersen et al. (2009).

10.3 NONLINEAR TIME SERIES MODELS

Many processes occurring in the natural sciences, engineering, finance, and economics exhibit some form of nonlinear behavior. This includes features that can not be modeled using Gaussian linear processes such as lack of time reversibility evidenced, for example, by pseudocyclical patterns where the values slowly rise to a peak and then quickly

decline to a trough. Time series that exhibit occasional bursts of outlying values are also unlikely under the linear Gaussian assumption. The prevalence of such series has led to an interest in developing nonlinear time series models that can account for such behavior. Nonlinear models proposed in the literature include bilinear models, threshold autoregressive (TAR) models, exponential autoregressive (EXPAR) models, and stochastic or random coefficient models. These models describe nonlinearities in the conditional mean as opposed to nonlinearities in the conditional variance as discussed in Section 10.2. When nonlinearities are present, model identification and estimation become more complicated, including the fundamental problem of which type of nonlinear model might be useful for a particular time series. This section presents a brief description of some nonlinear models that have been proposed in the literature. More comprehensive discussions are available in texts such as Tong (1983, 1990), Priestley (1988), Franses and van Dijk (2000), Fan and Yao (2003), Tsay (2010, Chapter 4), and Teräsvirta et al. (2010).

10.3.1 Classes of Nonlinear Models

Many nonlinear ARMA models can be viewed as special cases of the following general form:

$$\begin{aligned} z_t - \phi_1(\boldsymbol{Y}_{t-1})z_{t-1} - \cdots - \phi_p(\boldsymbol{Y}_{t-1})z_{t-p} \\ = \theta_0(\boldsymbol{Y}_{t-1}) + a_t - \theta_1(\boldsymbol{Y}_{t-1})a_{t-1} - \cdots - \theta_q(\boldsymbol{Y}_{t-1})a_{t-q} \end{aligned} \tag{10.3.1}$$

where

$$\boldsymbol{Y}_{t-1} = (z_{t-1}, \ldots, z_{t-p}, a_{t-1}, \ldots, a_{t-q})'$$

and $\phi_i(\boldsymbol{Y}_{t-1})$ and $\theta_i(\boldsymbol{Y}_{t-1})$ are functions of the ''state vector'' $\boldsymbol{Y}_{t-1}$ at time $t-1$. For specific cases, we mention the following models.

1. *Bilinear Models.* Let the ϕ_i be constants, and set $\theta_j(\boldsymbol{Y}_{t-1}) = b_j + \Sigma_{i=1}^{k} b_{ij} z_{t-i}$. Then we have the model

$$z_t - \phi_1 z_{t-1} - \cdots - \phi_p z_{t-p} = \theta_0 + a_t - \sum_{j=1}^{q} b_j a_{t-j} - \sum_{i=1}^{k}\sum_{j=1}^{q} b_{ij} z_{t-i} a_{t-j} \tag{10.3.2}$$

Equivalently, with the notations $p^* = \max(p, k)$, $\phi_i = 0$, $i > p$, $b_{ij} = 0$, $i > k$, and $\alpha_i(t) = \Sigma_{j=1}^{q} b_{ij} a_{t-j}$, (10.3.2) can be expressed in the form

$$z_t - \sum_{i=1}^{p^*} [\phi_i - \alpha_i(t)] z_{t-i} = \theta_0 + a_t - \sum_{j=1}^{q} b_j a_{t-j}$$

and be viewed in the form of an ARMA model with random coefficients for the AR parameters, which are linear functions of past values of the innovations process a_t. The statistical properties of bilinear models were studied extensively by Granger and Anderson (1978). Methods for analysis and parameter estimation were also studied by Subba Rao (1981) and Subba Rao and Gabr (1984), and various special cases of these models have been examined by subsequent authors.

Conditions for stationarity and other properties have been studied for the general bilinear model by Tuan (1985, 1986) and Liu and Brockwell (1988), in particular. For example, consider the simple first-order bilinear model $z_t - \phi_1 z_{t-1} = a_t - b_{11} z_{t-1} a_{t-1}$. It is established that a condition for second-order stationarity of such a process $\{z_t\}$ is $\phi_1^2 + \sigma_a^2 b_{11}^2 < 1$, and that the autocovariances of z_t under stationarity will satisfy $\gamma_j = \phi_1 \gamma_{j-1}$ for $j > 1$. Thus, this process will have essentially the same autocovariance structure as an ARMA(1, 1) process. This example highlights the fact that moments higher than the second order are typically needed in order to distinguish between linear and nonlinear models.

2. *Amplitude-Dependent Exponential AR Models.* Let $\theta_i = 0$, and set $\phi_i(\boldsymbol{Y}_{t-1}) = b_i + \pi_i e^{-c z_{t-1}^2}$, where $c > 0$ is a constant. Then we have

$$z_t - \sum_{i=1}^{p} (b_i + \pi_i e^{-c z_{t-1}^2}) z_{t-i} = a_t \tag{10.3.3}$$

This class of models was introduced by Haggan and Ozaki (1981), with an aim to construct models that reproduce features of nonlinear random vibration theory.

3. *Threshold AR, or TAR, Models.* Let $\theta_i = 0, i \geq 1$, and for some integer time lag d and some ''threshold'' constant c, let

$$\phi_i(\boldsymbol{Y}_{t-1}) = \begin{cases} \phi_i^{(1)} & \text{if } z_{t-d} \leq c \\ \phi_i^{(2)} & \text{if } z_{t-d} > c \end{cases}$$

$$\theta_0(\boldsymbol{Y}_{t-1}) = \begin{cases} \theta_0^{(1)} & \text{if } z_{t-d} \leq c \\ \phi_0^{(2)} & \text{if } z_{t-d} > c \end{cases}$$

Then we have the model

$$z_t = \begin{cases} \theta_0^{(1)} + \sum_{i=1}^{p} \phi_i^{(1)} z_{t-i} + a_t^{(1)} & \text{if } z_{t-d} \leq c \\ \theta_0^{(2)} + \sum_{i=1}^{p} \phi_i^{(2)} z_{t-i} + a_t^{(2)} & \text{if } z_{t-d} > c \end{cases} \tag{10.3.4}$$

where $\{a_t^{(1)}\}$ and $\{a_t^{(2)}\}$ are each white noise processes with variances σ_1^2 and σ_2^2, respectively (e.g., we can take $a_t^{(j)} = \sigma_j a_t$). The value c is called the threshold parameter and d is the delay parameter. A special case arises when the parameter c is replaced by a lagged value of the series itself, resulting in a model called the self-exciting TAR (SETAR) model.

The model (10.3.4) readily extends to an ''l-threshold'' model of the form

$$z_t = \theta_0^{(j)} + \sum_{i=1}^{p} \phi_i^{(j)} z_{t-i} + a_t^{(j)} \qquad \text{if} \qquad c_{j-1} < z_{t-d} \leq c_j \qquad j = 1, \ldots, l$$

with threshold parameters $c_1 < c_2 < \cdots < c_{l-1}$ (and $c_0 = -\infty, c_l = +\infty$), which define a partition of the real line into l subintervals. The first-order threshold model,

$$z_t = \theta_0^{(j)} + \phi^{(j)} z_{t-1} + a_t^{(j)} \qquad \text{if} \qquad c_{j-1} < z_{t-1} \le c_j$$

for example, may thus be regarded as a piecewise linear approximation to a general nonlinear first-order model $z_t = g(z_{t-1}) + a_t$, where $g(\cdot)$ is some general nonlinear function.

The TAR models were introduced by Tong (1978) and Tong and Lim (1980) and discussed in detail by Tong (1983, 1990). Tong (2007) gives a brief discussion of their origin. The basic threshold AR model can be seen as a piecewise linear AR model, with a somewhat abrupt change from one equation or ''regime'' to another dependent on whether or not a threshold value c_j is exceeded by z_{t-d}. A generalization that allows for less abrupt transition from one regime to another has been developed as a class of models known as smooth transition AR (STAR) models; see, for example, Teräsvirta (1994) and Teräsvirta et al. (2010). For the case of a single threshold $l = 1$, the basic form of a STAR model is

$$z_t = \theta_0^{(1)} + \sum_{i=1}^{p} \phi_i^{(1)} z_{t-i} + \left(\theta_0^{(2)} + \sum_{i=1}^{p} \phi_i^{(2)} z_{t-i} \right) F(z_{t-d}) + a_t$$

where $F(z) = 1/[1 + \exp\{-\gamma(z - c)\}]$ in the case of a logistic STAR model and in the normal STAR model $F(z) = \Phi(\gamma(z - c))$, with $\Phi(\cdot)$ equal to the cumulative distribution function of the standard normal distribution. By letting $\gamma \to \infty$, we see that $F(z)$ tends to the indicator function, and the usual two-regime TAR model (10.3.4) is obtained as a special case. The TAR model and its extensions have been used to model nonlinear series in many diverse areas such as finance and economics, the environmental sciences, hydrology, neural science, population dynamics, and physics; for selected references, see Fan and Yao (2003, p. 126).

Other types of nonlinear models include the stochastic or random coefficient models. For example, in the simple AR(1) model we consider $z_t = \phi_t z_{t-1} + a_t$, where ϕ_t is not a constant but is a stochastic parameter. Possible assumptions on the mechanism generating the ϕ_t include (i) the ϕ_t are iid random variables with mean ϕ and variance σ_ϕ^2, independent of the process $\{a_t\}$, and (ii) the ϕ_t follow an AR(1) process themselves,

$$\phi_t - \phi = \alpha(\phi_{t-1} - \phi) + e_t$$

where ϕ is the mean of the ϕ_t process and the e_t are iid random variables with mean 0 and variance σ_e^2, independent of a_t. Estimation for the first case was considered in detail by Nicholls and Quinn (1982), while the second case may in principle be estimated using state-space methods (e.g., Ledolter, 1981).

Additional classes of nonlinear models include the general state-dependent model form (10.3.1) examined extensively by Priestley (1980, 1988), or more general nonparametric autoregressive model forms such as nonlinear additive autoregressive models considered by Chen and Tsay (1993), and adaptive spline threshold autoregressive models used by Lewis and Stevens (1991). Nonparametric and semiparametric methods such as kernel regression and artificial neural networks have also been used to model nonlinearity. A review of nonlinear time series models with special emphasis on nonparametric methods was provided by Tjøstheim (1994). More recent discussions of the developments in this

area can be found in Fan and Yao (2003), Gao (2007), and Teräsvirta et al. (2010). A discussion of nonlinear models with applications to finance is provided by Tsay (2010, Chapter 4).

10.3.2 Detection of Nonlinearity

Many methods have been proposed to detect nonlinearity of a time series. In addition to informal graphical methods and inspection of higher order moments, such as third- and fourth-order moments, these include more formal test procedures by Hinich (1982), Subba Rao and Gabr (1980), McLeod and Li (1983), Keenan (1985), Tsay (1986a), Petruccelli and Davies (1986), Luukkonen et al. (1988a), and others. Some of these tests exploit the nonlinear dependence structure that is reflected in the higher order moments, and many of the tests are developed as portmanteau tests based on a linear model, with an alternative not explicitly specified. Other tests are Lagrange multiplier or score-type procedures against specified alternative models. For example, the tests of Luukkonen et al. (1988a) are score-type tests against STAR alternatives. The tests of Subba Rao and Gabr (1980) and Hinich (1982) are nonparametric tests that use a bispectral approach, while the test of Petruccelli and Davies (1986) is based on cumulative sums of standardized residuals from autoregressive fitting to the data. The portmanteau test statistic (10.2.12) of McLeod and Li (1983) is based on sample autocorrelations of squared residuals $\hat{a}_t^2$ from a fitted linear ARMA model. This test was introduced as a test for nonlinearity, although simulations suggest that it may be more powerful against ARCH alternatives. A modest gain in power may be possible by basing the nonlinearity checks on the portmanteau statistics proposed by Peña and Rodríguez (2002, 2006).

Keenan (1985) proposed an F-test for nonlinearity using an analogue of Tukey's single-degree-of-freedom test for nonadditivity. The test is also similar to the regression specification error test (RESET) proposed by Ramsey (1969) for linear regression models. The test can be implemented by first fitting an AR(m) model to the observed series z_t, where m is a suitably selected order. The fitted values are retained and their squares are added as a predictor variable to the AR(m) model. This model is then refitted and the coefficient associated with the predictor variable is tested for significance. This procedure thus amounts to determining whether inclusion of the squared predicted values helps improve the prediction.

Tsay (1986a) proposed an extension based on testing whether second-order terms have additional predictive ability. The procedure can be carried out as follows: First fit a linear AR(m) model and obtain the residuals $\hat{a}_t$ from this fit. Then consider the $M = \frac{1}{2}m(m+1)$ component vector

$$\boldsymbol{Z}_t = (z_{t-1}^2, \ldots, z_{t-m}^2, z_{t-1}z_{t-2}, \ldots, z_{t-m+1}z_{t-m})'$$

consisting of all squares and distinct cross-products of the lagged values $z_{t-1}, \ldots, z_{t-m}$. Now perform a multivariate least-squares regression of the elements of $\boldsymbol{Z}_t$ on the set of regressors $\{1, z_{t-1}, \ldots, z_{t-m}\}$ and obtain the multivariate residual vectors $\hat{\boldsymbol{U}}_t$, for $t = m+1, \ldots, n$. Finally, perform a least-squares regression $\hat{a}_t = \hat{\boldsymbol{U}}_t'\boldsymbol{\beta} + e_t$ of the AR(m) model residuals $\hat{a}_t$ on the M-dimensional vectors $\hat{\boldsymbol{U}}_t$ as regressor variables, and let $\hat{F}$ be the F ratio of the

regression mean square to the error mean square from that regression, so that

$$\hat{F} = \frac{(\sum_t \hat{a}_t \hat{U}_t')(\sum_t \hat{U}_t \hat{U}_t')^{-1}(\sum_t \hat{U}_t \hat{a}_t)/M}{\sum_{t=m+1}^{n} \hat{e}_t^2/(n-m-M-1)} \tag{10.3.5}$$

Under the assumption of linearity, $\hat{F}$ has, for large n, an approximate F distribution with M and $n-m-M-1$ degrees of freedom, and the null hypothesis of linearity is rejected for large values of $\hat{F}$. Extension to a procedure for residuals $\hat{a}_t$ from a fitted ARMA(p, q) model was also mentioned by Tsay (1986a).

If one aggregates or condenses the information in the M-dimensional vector $\boldsymbol{Z}_t$ into a single variable $\hat{z}_t^2 = (\hat{\theta}_0 + \Sigma_{i=1}^{m} \hat{\phi}_i z_{t-i})^2$, which is the square of the fitted value from the AR(m) model, and performs the remaining steps outlined above, one obtains the earlier test by Keenan (1985). The associated test statistic is

$$\hat{F} = \frac{(\sum_t \hat{u}_t \hat{a}_t)^2/(\sum_t \hat{u}_t^2)}{\sum_{t=m+1}^{n} \hat{e}_t^2/(n-2m-2)}$$

with 1 and $n-2m-2$ degrees of freedom. Luukkonen et al. (1988b) and Tong (1990, Section 5.3) noted a score test interpretation of the procedures proposed by Keenan (1985) and Tsay (1986a). Both tests are available in the TSA package of R and can be implemented using the commands Keenan.test(z) and Tsay.test(z). For further discussion, see Tsay (2010, Chapter 4).

10.3.3 An Empirical Example

For illustration, we consider modeling of the Canadian lynx dataset, consisting of annual numbers of Canadian lynx trapped in the MacKenzie River district for the period 1821 to 1934. The series is available in the R datasets package. For several reasons, the $\log_{10}$ transformation of the data is used in the analysis, denoted as z_t, $t = 1, \ldots, n$, with $n = 114$. Examination of the time series plot of z_t in Figure 10.5 shows a very strong cyclical behavior, with period around 10 years. It also shows an asymmetry or lack of time reversibility in that the sample values rise to their peak or maximum values more slowly than they fall away to their minimum values (typically, about 6-year segments of rising and 4-year segments of falling). This is a feature exhibited by many nonlinear processes. There are biological/population reasons that would also support a nonlinear process, especially one involving a threshold mechanism; see, for example, Tong (1990).

The sample ACF and PACF of the series $\{z_t\}$ are shown in Figure 10.6. The ACF exhibits the cyclic feature clearly, and based on features of the sample PACF a linear AR(4) model is initially fitted to the series, with $\hat{\sigma}_a^2 = 0.0519$. The presence of some moderate autocorrelation at higher lags, around lags 10 and 12, in the residuals from the fitted AR(4) model suggested the following more refined model that was estimated by conditional LS:

$$\begin{aligned} z_t = {} & 1.149 + 1.038 z_{t-1} - 0.413 z_{t-2} + 0.252 z_{t-3} - 0.229 z_{t-4} \\ & + 0.188 z_{t-9} - 0.232 z_{t-12} + a_t \end{aligned} \tag{10.3.6}$$

with residual variance estimate $\hat{\sigma}_a^2 = 0.0380$.

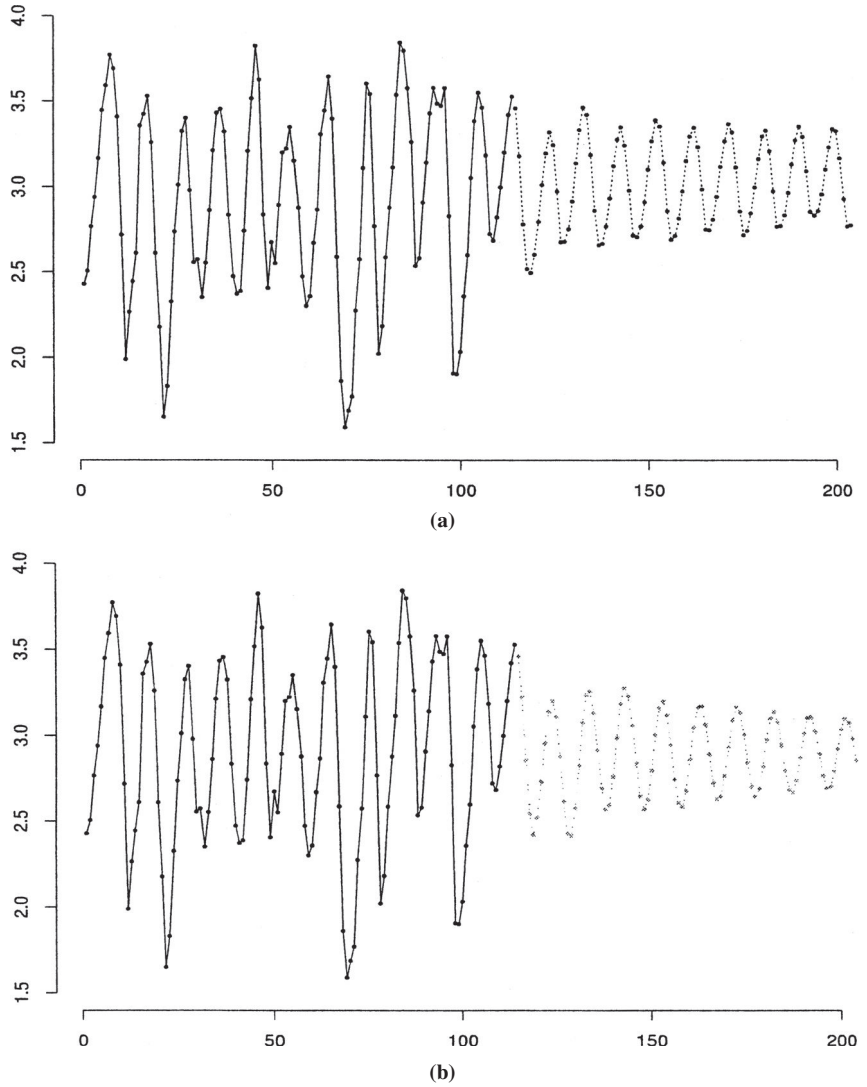


FIGURE 10.5 Logarithms (base 10) of the Canadian lynx time series for 1821–1934, with forecasts for 90 periods ahead from (a) the TAR model and (b) the linear subset AR(12) model.

Some diagnostics of this fitted model suggest possible nonlinearity. Specifically, there is strong autocorrelation in the squared residuals $\hat{a}_t^2$ at lag 2, with $r_2(\hat{a}^2) = 0.401$, and nonlinear features exist in scatter plots of the ''fitted values'' $\hat{z}_t \equiv \hat{z}_{t-1}(1)$ and residuals $\hat{a}_t = z_t - \hat{z}_{t-1}(1)$ versus lagged values z_{t-j}, for lags $j = 2, 3, 4$. But the tests by Keenan (1985) and Tsay (1986a), implemented in the TSA package of R, are inconclusive in that the Keenan test rejects linearity whereas the Tsay test does not (see the output below). However, it appears that the failure of the Tsay test to detect the nonlinearity may be due to the way the package computes the Tsay statistic. This computation uses 77 parameters and results in an observation/parameter ratio of $114/77 < 2$, which is too small for valid inference.

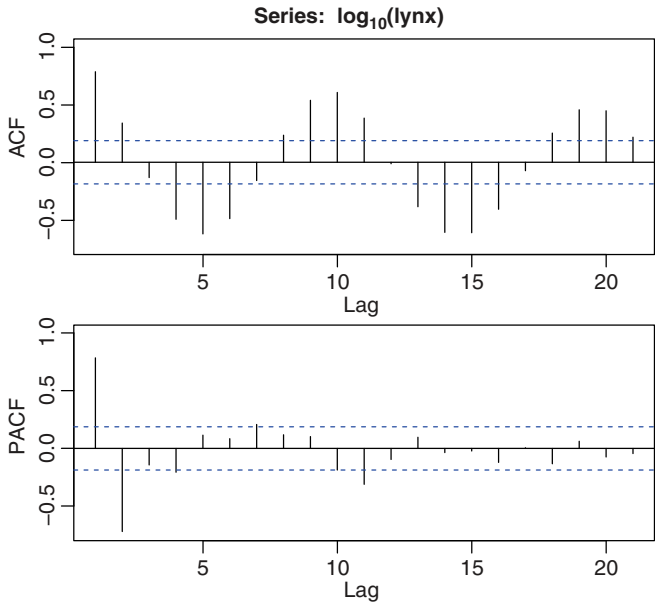


FIGURE 10.6 Autocorrelation and partial autocorrelation functions for the logarithm of the Canadian lynx series.

```
> library(TSA)
> data(lynx)
> z=log10(lynx)
> Keenan.test(z)
    $test.stat: 11.66997
    $p.value:  0.000955
    $order: 11
> Tsay.test(z)
    $test.stat: 1.316
    $p.value:  0.2256
    $order: 11
```

Tong (1990) specified a TAR model, with time delay of $d = 2$ and threshold value of about $c \approx 3.10$ for this series. A threshold version of the AR model in (10.3.6), with two phases and terms at lags 1, 2, 3, 4, 9, and 12, was estimated by conditional LS. After eliminating nonsignificant parameter estimates, we arrived at the following estimated threshold AR model:

$$\begin{aligned} z_t &= 1.3206 + 0.9427 z_{t-1} - 0.2161 z_{t-4} \\ &\quad - 0.1411 z_{t-12} + a_t^{(1)} \quad \text{if} \quad z_{t-2} \le 3.10 \\ &= 1.8259 + 1.1971 z_{t-1} - 0.7266 z_{t-2} + 0.1667 z_{t-9} \\ &\quad - 0.2229 z_{t-12} + a_t^{(2)} \qquad \text{if} \quad z_{t-2} > 3.10 \end{aligned}$$

with residual variance estimates $\hat{\sigma}_1^2 = 0.0249$ and $\hat{\sigma}_2^2 = 0.0386$ (pooled $\hat{\sigma}^2 = 0.0328$).

The approximate ''eventual'' forecast function from this model will lead to periodic limit cycle behavior with an approximate period of 9 years (see Tong (1990) for discussion of limit cycles). Although exact minimum MSE forecasts $\hat{z}_n(l)$ for lead times $l > 2$ are not easily computed for the fitted threshold AR model, approximate forecasts for larger l can be obtained by projecting series values forward with future white noise terms $a_t^{(i)}$ set to 0 (see Teräsvirta et al. (2010, Chapter 14) for other options). Values obtained in this way for the eventual forecast function from the TAR model are depicted for 90 years, $l = 1, \ldots, 90$, in Figure 10.5(a). These values exhibit a limit cycle with a period of essentially 9 years (in fact, the period is 28 years with 3 ''subcycles''), and the asymmetric feature of slower rise to peak values and faster fall to minimum values is visible. In contrast, the stationary linear AR model will give a forecast function in the form of very slowly damped sinusoidal oscillations that will eventually decay to the mean value of the process, 2.90. This forecast function is shown in Figure 10.5(b).

Other nonlinear models have been considered for the Canadian lynx data. For examples, Subba Rao and Gabr (1984) have estimated a bilinear model for these data, an AR(2) model with random coefficients was fitted by Nicholls and Quinn (1982), and an amplitude-dependent exponential AR model of order 11 was fitted to the mean-adjusted log lynx data by Haggan and Ozaki (1981).

10.4 LONG MEMORY TIME SERIES PROCESSES

The autocorrelation function ρ_k of a stationary ARMA(p, q) process decreases rapidly as $k \to \infty$, since the autocorrelation function is geometrically bounded so that

$$|\rho_k| \leq CR^k, \qquad k = 1, 2, \ldots$$

where $C > 0$ and $0 < R < 1$. Processes with this property are often referred to as *short memory* processes. Stationary processes with much more slowly decreasing autocorrelation function, known as *long memory* processes, have

$$\rho_k \sim Ck^{2d-1} \qquad \text{as} \qquad k \to \infty \tag{10.4.1}$$

where $C > 0$ and $-0.5 < d < 0.5$. Empirical evidence suggests that long memory processes are common in fields as diverse as hydrology (e.g., Hurst, 1951; McLeod and Hipel, 1978), geophysics, and financial economics. The sample autocorrelations of such processes are not necessarily large, but tend to persist over a long period. The latter could suggest a need for differencing to achieve stationarity, although taking a first difference may be too extreme. This motivates the notion of fractional differencing and consideration of the class of fractionally integrated processes.

10.4.1 Fractionally Integrated Processes

A notable class of stationary long memory processes z_t is the *fractionally integrated ARMA*, or ARFIMA, processes defined for $-0.5 < d < 0.5$ by the relation

$$\phi(B)(1 - B)^d z_t = \theta(B)a_t \tag{10.4.2}$$

where $\{a_t\}$ is a white noise sequence with zero mean and variance σ_a^2, and $\phi(B) = 0$ and $\theta(B) = 0$ have all roots greater than one in absolute value. The class of models in (10.4.2) was initially proposed and studied by Granger and Joyeux (1980) and Hosking (1981) as an intermediate compromise between fully integrated ARIMA processes and short memory ARMA processes. More comprehensive treatments of these models can be found in texts by Beran (1994), Robinson (2003), and Palma (2007).

For $d > -1$, the operator $(1 - B)^d$ in (10.4.2) is defined by the binomial expansion

$$(1 - B)^d = \sum_{j=0}^{\infty} \pi_j B^j \tag{10.4.3}$$

where $\pi_0 = 1$ and

$$\pi_j = \frac{\Gamma(j - d)}{\Gamma(j + 1)\Gamma(-d)} = \prod_{0<k\leq j} \frac{k - 1 - d}{k} \qquad j = 1, 2, \ldots \tag{10.4.4}$$

and $\Gamma(x)$ is the gamma function. Hence, the π_j follow the simple recursion

$$\pi_j = \left(\frac{j - 1 - d}{j}\right) \pi_{j-1}$$

A particular special case is the *fractionally integrated white noise* process w_t, defined by

$$(1 - B)^d w_t = a_t$$

For $-0.5 < d < 0.5$, since the power series expansion of $\psi(B) = (1 - B)^{-d} \equiv \Sigma_{j=0}^{\infty} \psi_j B^j$ converges for $|B| \leq 1$, it follows that such a process $\{w_t\}$ is stationary and has the infinite MA representation

$$w_t = (1 - B)^{-d} a_t = \sum_{j=0}^{\infty} \psi_j a_{t-j} \tag{10.4.5}$$

where

$$\psi_j = \frac{\Gamma(j + d)}{\Gamma(j + 1)\Gamma(d)} = \prod_{0<k\leq j} \frac{k - 1 + d}{k} \sim \frac{1}{\Gamma(d)} j^{d-1} \qquad \text{as} \qquad j \to \infty \tag{10.4.6}$$

It can also be shown (Hosking, 1981; Brockwell and Davis, 1991, Chapter 12) that the fractionally integrated white noise process has variance

$$\gamma_0(w) = \text{var}[w_t] = \frac{\sigma^2{}_a \Gamma(1 - 2d)}{[\Gamma(1 - d)]^2}$$

and ACF

$$\rho_h(w) = \frac{\Gamma(h + d)\Gamma(1 - d)}{\Gamma(h - d + 1)\Gamma(d)} = \prod_{0<k\leq h} \frac{k - 1 + d}{k - d} \qquad h = 1, 2, \ldots \tag{10.4.7}$$

In particular, we have $\rho_1(w) = d/(1 - d)$, and $\rho_h(w) = [(h - 1 + d)/(h - d)]\rho_{h-1}(w)$. It follows, using Stirling's formula $\Gamma(x) \sim \sqrt{2\pi} e^{-x+1}(x - 1)^{x-1/2}$ as $x \to \infty$, that the ACF

behaves like

$$\rho_h(w) \sim h^{2d-1}\frac{\Gamma(1-d)}{\Gamma(d)} \qquad \text{as} \qquad h \to \infty$$

the characteristic feature of the ACF of a long memory process. In addition, by use of the Levinson–Durbin recursion algorithm described in Appendix A3.2, values for the partial autocorrelations of the fractionally integrated white noise process can be determined by induction and shown to be $\phi_{kk} = d/(k-d), k = 1, \ldots$.

The fractionally integrated white noise process itself may be of limited use in modeling long memory behavior since the single parameter d can allow for only a restrictive class of autocorrelation function forms. This process can be useful, however, in building of the more general class of long memory processes. In fact, we can see from the above definition that a fractionally integrated ARMA(p, d, q) process, $\phi(B)(1-B)^d z_t = \theta(B)a_t$, can be interpreted as an ''ARMA(p, q) process driven by fractionally integrated white noise,'' that is, $\{z_t\}$ satisfies $\phi(B)z_t = \theta(B)w_t$, with $(1-B)^d w_t = a_t$. From general results on linear filtering, we see that the exact autocovariance function of $\{z_t\}$ can be expressed in terms of the autocovariance function of the fractionally integrated white noise process $\{w_t\}$ as

$$\gamma_h(z) = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\psi_j\psi_k\gamma_{h+j-k}(w) \tag{10.4.8}$$

where the ψ_j are the coefficients in $\psi(B) = \phi(B)^{-1}\theta(B) = \sum_{j=0}^{\infty}\psi_j B^j$ and

$$\begin{aligned}\gamma_h(w) = \gamma_0(w)\rho_h(w) &= \sigma_a^2\frac{\Gamma(1-2d)\Gamma(h+d)}{\Gamma(h-d+1)\Gamma(d)\Gamma(1-d)} \\ &\equiv \sigma_a^2\frac{(-1)^h\Gamma(1-2d)}{\Gamma(h-d+1)\Gamma(1-h-d)}\end{aligned}$$

is the autocovariance function of the fractionally integrated white noise process $\{w_t\}$.

In terms of the spectrum, from (3.1.12) the spectrum of a fractionally integrated ARIMA (p, d, q) process $\{z_t\}$ is

$$p_z(f) = 2\sigma_a^2|1-e^{-i2\pi f}|^{-2d}\frac{|\theta(e^{-i2\pi f})|^2}{|\phi(e^{-i2\pi f})|^2} \qquad 0 \le f \le \frac{1}{2} \tag{10.4.9}$$

where $p_w(f) = 2\sigma_a^2|1-e^{-i2\pi f}|^{-2d} \equiv 2\sigma_a^2[2\sin(\pi f)]^{-2d}$ is the spectrum of the fractionally integrated white noise process. In particular, we see that $p_z(f)$ does not remain finite as $f \to 0$ for $0 < d < \frac{1}{2}$. Since $\sin(x) \sim x$ as $x \to 0$, we have the behavior that

$$p_z(f) \sim 2\sigma_a^2\left[\frac{|\theta(1)|^2}{|\phi(1)|^2}\right](2\pi f)^{-2d} \equiv C^* f^{-2d} \qquad \text{as} \qquad f \to 0$$

which is a distinguishing feature of the spectrum of long memory processes, for $0 < d < \frac{1}{2}$.

Two Simple Special Cases. In practice, ARIMA(p, d, q) models are likely to be most useful for small values of p and q. So, we mention a few specific details given by Hosking (1981) about characteristics of two of the simplest such models. First, consider the fractional ARIMA$(1, d, 0)$ model, $(1-\phi B)(1-B)^d z_t = a_t$, with AR parameter $-1 < \phi < 1$. Then

$(1-\phi B)z_t = w_t$ or $z_t = (1-\phi B)^{-1}w_t = \sum_{j=0}^{\infty}\phi^j w_{t-j}$, so using (10.4.7) and (10.4.8) with $\psi_j = \phi^j$ it follows that the autocorrelation function of $\{z_t\}$ is

$$\rho_l(z) = \frac{\rho_l(w)}{1-\phi}\frac{F(d+l,1;1-d+l;\phi)+F(d-l,1;1-d-l;\phi)-1}{F(1+d,1;1-d;\phi)}$$

where $F(a,b;c;x)$ is the hypergeometric function defined by

$$\begin{aligned}F(a,b;c;x) &= 1+\frac{ab}{c\cdot 1}x+\frac{a(a+1)b(b+1)}{c(c+1)\cdot 1\cdot 2}x^2+\cdots\\ &= \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)}\sum_{k=0}^{\infty}\frac{\Gamma(a+k)\Gamma(b+k)}{\Gamma(c+k)k!}x^k\end{aligned}$$

and

$$\begin{aligned}\gamma_0(z) &= \gamma_0(w)\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\phi^{j+k}\rho_{j-k}(w)\\ &= \frac{\gamma_0(w)}{1-\phi^2}[2F(d,1;1-d;\phi)-1] = \frac{\sigma_a^2\Gamma(1-2d)}{\Gamma(1-d)^2}\frac{F(1+d,1;1-d;\phi)}{1+\phi}\end{aligned}$$

Given ϕ and d, values of $F(d+l,1;1-d+l;\phi)$ required in computing the $\gamma_l(z) = \gamma_0(z)\rho_l(z)$ may be obtained more conveniently using the recurrence relation

$$F(d+l-1,1;1-d+l-1;\phi) = \frac{d+l-1}{1-d+l-1}\phi F(d+l,1;1-d+l;\phi)+1$$

Second, for the fractional ARIMA$(0,d,1)$ model, $(1-B)^d z_t = (1-\theta B)a_t$, with $-1<\theta<1$, we have $z_t = (1-\theta B)w_t$. So again using (10.4.7) and (10.4.8), now with $\psi_0 = 1, \psi_1 = -\theta$, and $\psi_j = 0$ for $j>1$, we find that

$$\gamma_l(z) = \gamma_0(w)[(1+\theta^2)\rho_l(w)-\theta\rho_{l+1}(w)-\theta\rho_{l-1}(w)]$$

and the ACF of $\{z_t\}$ is

$$\rho_l(z) = \rho_l(w)\frac{al^2-(1-d)^2}{l^2-(1-d)^2}$$

where

$$a = (1-\theta)^2\left[1+\theta^2-\frac{2\theta d}{1-d}\right]^{-1}$$

with

$$\gamma_0(z) = \gamma_0(w)[1+\theta^2-2\theta\rho_1(w)] = \left[\frac{\sigma_a^2\Gamma(1-2d)}{\Gamma(1-d)^2}\right]\left[1+\theta^2-\frac{2\theta d}{1-d}\right]$$

10.4.2 Estimation of Parameters

We first briefly mention the sampling properties of the sample mean

$$\bar{z} = \left(\frac{1}{n}\right)\sum_{t=1}^{n} z_t$$

for estimation of the mean $\mu = E[z_t]$ from a fractionally integrated ARMA process. From the general result that $\text{var}[\bar{z}] = (\gamma_0(z)/n)[1 + 2\sum_{h=1}^{n-1}\{(n-h)/n\}\rho_h(z)]$ and the property that $\rho_h(z) \sim Ch^{2d-1}$ as $h \to \infty$, it follows that

$$n^{1-2d}\text{var}[\bar{z}] \to C^*$$

for $-0.5 < d < 0.5$, where $C^* > 0$ is a certain constant. Hence, we see that $\text{var}[\bar{z}] \simeq C^*/n^{1-2d}$, whereas for short memory processes ($d = 0$), the variance of the sample mean behaves like $\text{var}[\bar{z}] \simeq C^*/n$. Thus, for $0 < d < 0.5$, the process mean μ can be much less accurately estimated by the sample mean. Equivalently, a much longer series length n is required for accurate estimation of μ for long memory processes. Hosking (1996) derived asymptotic distribution results for sample autocorrelations $\hat{\rho}_l(z)$ of long memory processes.

Estimation of the parameters $d, \boldsymbol{\phi}, \boldsymbol{\theta}$, and σ_a^2 in a fractionally integrated ARIMA (p, d, q) process can be performed by maximum likelihood (e.g., Sowell, 1992). However, direct evaluation of the exact likelihood function is rather slow due partly to the complicated nature of the autocovariance function of the process. Therefore, approximate ML estimation methods have been considered by Beran (1994, 1995) and others. Another convenient approach is to obtain an estimate of the parameter d initially by certain methods (e.g., using a frequency-domain nonparametric approach; see Geweke and Porter-Hudak (1983)), and then estimate $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ_a^2 by relatively standard ML methods for the given estimate of d. Asymptotic normality and the form of limiting covariance matrix of (approximate) ML estimators have been established by Beran (1995) and argued by Li and McLeod (1986). Notice that for $d \geq 0.5$, the fractionally integrated ARMA process is nonstationary. For such cases, in practice the typical procedure is to first difference the nonstationary process in the usual way, thus reducing it to a fractionally integrated process with a parameter d in the ''stationary'' range $-0.5 \leq d < 0.5$.

One approximate maximum likelihood estimation method is suggested by expressing the general fractional ARIMA process z_t in (10.4.2) in the infinite AR form as

$$z_t - \sum_{j=1}^{\infty} \pi_j^* z_{t-j} = a_t \tag{10.4.10}$$

where

$$\pi^*(B) = 1 - \sum_{j=1}^{\infty} \pi_j^* B^j = \theta^{-1}(B)\phi(B)(1-B)^d$$

The π_j^* coefficients can be obtained recursively based on the relation $\theta(B)\pi^*(B) = \phi(B)(1-B)^d \equiv \varphi(B)$, similar to Section 4.2.3, as

$$\pi_j^* - \theta_1\pi_{j-1}^* - \cdots - \theta_q\pi_{j-q}^* = \varphi_j \qquad j = 1, 2, \ldots$$

where $\varphi(B) = \phi(B)(1-B)^d = 1 - \sum_{j=1}^{\infty} \varphi_j B^j$. For example, in an ARIMA$(1, d, 1)$ model, the π_j^* satisfy $\pi_j^* - \theta_1 \pi_{j-1}^* = \varphi_j$, with $\varphi_j = \pi_j - \phi_1 \pi_{j-1}$ for $j \geq 1$, where the π_j are the coefficients in (10.4.3) and (10.4.4). In the approximate maximum likelihood or least-squares method, the truncated errors

$$\varepsilon_t(\boldsymbol{\beta}) = z_t - \sum_{j=1}^{t-1} \pi_j^* z_{t-j} \qquad t = 1, \ldots, n \tag{10.4.11}$$

are considered as a function of $\boldsymbol{\beta} = (\boldsymbol{\phi}', \boldsymbol{\theta}', d)'$, and the estimate $\hat{\boldsymbol{\beta}}$ is determined by minimizing the sum of squares $S(\boldsymbol{\beta}) = \sum_{t=1}^{n} \varepsilon_t^2(\boldsymbol{\beta})$. The corresponding approximate ML estimate of σ_a^2 is then taken as $\hat{\sigma}_a^2 = S(\hat{\boldsymbol{\beta}})/n$. For very long time series, it might be advisable to discard the first several $\varepsilon_t^2(\boldsymbol{\beta})$ terms in the sum-of-squares function to be minimized (e.g., the first 10–20 values), to avoid the effects of the inaccuracy in the approximation (10.4.11) for small values of t.

For practical implementation of the approximate maximum likelihood method, we might consider the following modification suggested because the series $(1-B)^d z_t$ follows the ARMA(p, q) model. Construct the series of truncated values of $(1-B)^d z_t \equiv \pi(B) z_t$ as

$$\tilde{z}_t(d) = z_t + \sum_{j=1}^{t-1} \pi_j(d) z_{t-j} \qquad t = 1, \ldots, n$$

for each d a grid of values within $-0.5 \leq d < 0.5$, where the $\pi_j(d)$ are the coefficients in (10.4.3) and (10.4.4). Then for *each* (fixed) value of d on the grid, obtain ML estimates of the ARMA parameters $\boldsymbol{\phi}, \boldsymbol{\theta}$, and σ_a^2, for the time series $\tilde{z}_t(d), \ldots, \bar{z}_n(d)$, by the usual likelihood and sum-of-squares methods of Chapter 7. The estimate $\hat{d}$ is then taken as the value of d that gives the minimum $\hat{\sigma}_a^2$ or the maximum of the likelihood, and the estimates $\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}$ associated with this value of d are the corresponding approximate ML estimates.

Estimation procedures directly extend to the more practical case of the fractional ARIMA model with an unknown nonzero mean μ,

$$\phi(B)(1-B)^d (z_t - \mu) = \theta(B) a_t$$

Although asymptotic theory is established to show that estimation of the additional unknown mean parameter μ does not affect the limiting distribution of the ARIMA parameter estimates $\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \hat{d}$, empirical simulation evidence (e.g., Hauser, 1999; Cheang and Reinsel, 2003) suggests that sampling properties of these estimates can be adversely affected even for moderately large sample lengths. This behavior may be related to previous discussion concerning the lower accuracy in estimation of the mean μ of a fractional ARIMA process. A possible remedy to obtain improved estimates of the ARIMA model parameters in the case of an unknown mean μ, or in situations of more general regression models with fractional ARIMA noise, is use of the restricted maximum likelihood estimation method as discussed in Section 9.5.2.

Forecasting. As with parameter estimation, forecasting for fractionally integrated ARMA processes (10.4.2) is not as convenient as for ARIMA processes with nonnegative *integer* value of d, because of the higher complexity of the differencing operator $(1-B)^d$ in the fractional case. Unlike the standard ARIMA model, forecasts cannot be obtained

conveniently directly from a *finite-order* difference equation form. For the fractional ARIMA model, it is simpler to consider forecasts based on the infinite AR form (10.4.10). Then, similar to (5.3.5) and (5.3.6), from this form we obtain that the l-step-ahead forecast of z_{t+l} based on the infinite past observations through origin t, z_t, z_{t-1}, …, is

$$\hat{z}_t(l) = \sum_{j=1}^{\infty} \pi_j^* \hat{z}_t(l-j) \tag{10.4.12}$$

where $\hat{z}_t(l-j) = z_{t+l-j}$ for $j \geq l$ as usual. For practical use, with forecasts based on a finite series of n available observations $z_1, \ldots, z_n$ and n sufficiently large, the sum in (10.4.12) must be truncated as $\hat{z}_n(l) = \sum_{j=1}^{n+l-1} \pi_j^* \hat{z}_n(l-j)$.

Conversely, the process z_t has the infinite MA form

$$z_t = \psi(B)a_t = \sum_{j=0}^{\infty} \psi_j a_{t-j}$$

where $\psi(B) = \Sigma_{j=0}^{\infty} \psi_j B^j = \phi^{-1}(B)(1-B)^{-d}\theta(B) \equiv \varphi^{-1}(B)\theta(B)$. From the same reasoning as in Chapter 5, we also have the equivalent representation of the lead-l forecast in (10.4.12) as

$$\hat{z}_t(l) = \sum_{j=l}^{\infty} \psi_j a_{t+l-j} \tag{10.4.13}$$

So the forecast error is $e_t(l) = z_{t+l} - \hat{z}_t(l) = \sum_{j=0}^{l-1} \psi_j a_{t+l-j}$, with variance

$$\sigma^2(l) = \text{var}[e_t(l)] = \sigma_a^2 \sum_{j=0}^{l-1} \psi_j^2$$

Example: Series A. Consider again Series A, which is a time series of chemical process concentration readings with $n = 197$ observations. Two possible models were proposed for this series in Chapters 6 and 7. One was the ''nearly nonstationary'' ARMA(1, 1) model, $(1-\phi B)z_t = \theta_0 + (1-\theta B)a_t$, with estimates $\hat{\phi} = 0.92$, $\hat{\theta} = 0.58$, $\hat{\theta}_0 = 1.45$, and $\hat{\sigma}_a^2 = 0.0974$. The second was the nonstationary IMA(0, 1, 1) model, $(1-B)z_t = (1-\theta B)a_t$, with estimates $\hat{\theta} = 0.71$ and $\hat{\sigma}_a^2 = 0.1004$. The unit root test performed in Section 10.1 suggests that the nonstationary IMA(0, 1, 1) model may be more appropriate. Beran (1995) also examined these data and found that an ARIMA$(0, d, 0)$ model, that is, a fractionally integrated white noise model, $(1-B)^d(z_t - \mu) = a_t$, fits the series well, with estimates $\hat{d} = 0.41$ and $\hat{\sigma}_a^2 = 0.0978$. Notice that the estimate of d is less than, but close to, the nonstationary boundary of $d < 0.5$ for an ARIMA$(0, d, 0)$ process, giving further support to the notion that it is very difficult to determine whether this process is stationary or not based on the series length of only $n = 197$ observations. In certain respects, especially in terms of long memory characteristics, the fractional ARIMA$(0, d, 0)$ model of Beran (1995) may be viewed as intermediate between the two models suggested earlier. For comparison, in Table 10.1 we display the first 30 ψ_j coefficients of the ''infinite'' MA representation for each of the three models considered. Notice that while the ψ_j, for ≥ 2, are initially smaller for the ARIMA$(0, d, 0)$ model than for the ARMA(1, 1) model, they decay relatively more slowly and become larger than those of the ARMA(1, 1) for all lags $j \geq 18$. In contrast,

TABLE 10.1 Coefficients ψ_j of the "Infinite" MA Representations for Three ARIMA Models Fitted to the Chemical Process Concentration Readings in Series A.

j	ARMA (1, 1)	IMA (0, 1, 1)	ARMA $(0, d, 0)$	j	ARMA (1, 1)	IMA (0, 1, 1)	ARMA $(0, d, 0)$
1	0.34000	0.290	0.41000	16	0.09734	0.290	0.08938
2	0.31280	0.290	0.28905	17	0.08955	0.290	0.08628
3	0.28778	0.290	0.23220	18	0.08239	0.290	0.08345
4	0.26475	0.290	0.19795	19	0.07580	0.290	0.08086
5	0.24357	0.290	0.17460	20	0.06974	0.290	0.07848
6	0.22409	0.290	0.15743	21	0.06416	0.290	0.07627
7	0.20616	0.290	0.14416	22	0.05902	0.290	0.07423
8	0.18967	0.290	0.13353	23	0.05430	0.290	0.07232
9	0.17449	0.290	0.12477	24	0.04996	0.290	0.07054
10	0.16054	0.290	0.11741	25	0.04596	0.290	0.06888
11	0.14769	0.290	0.11111	26	0.04228	0.290	0.06732
12	0.13588	0.290	0.10565	27	0.03890	0.290	0.06585
13	0.12501	0.290	0.10086	28	0.03579	0.290	0.06446
14	0.11501	0.290	0.09661	29	0.03293	0.290	0.06315
15	0.10581	0.290	0.09281	30	0.03029	0.290	0.06190

for the IMA(0, 1, 1) model we know that the $\psi_j = 1 - \theta$, for all $j > 1$, do not decay, which may not be an appropriate feature of a model for this process.

Remark. The parameters of the ARIMA$(0, d, 0)$ model can be estimated using the fracdiff package in R as shown below. From the partial output included, we see that the estimates $\hat{d} = 0.40$ and $\hat{\sigma}_a^2 = (0.3123734)^2 = 0.0976$ are close to the values quoted above.

```
> library(fracdiff)
> fracdiff(seriesA, nar=0, nma=0, M=30)
  Call:  fracdiff(x = numA, nar=0, nma=0, M=30)
  Coefficients:  d = 0.4001903
  sigma[eps] = 0.3123734
```

EXERCISES

10.1 Download from the Internet the daily stock prices of a company of your choosing.

(a) Plot the data using the graphics capabilities in R. Are there any unusual features worth noting? Perform a statistical test to determine the presence of a unit root in the series.

(b) Compute and plot the series of daily log returns. Does the graph show evidence of volatility clustering? Perform a statistical analysis to determine whether an AR–ARCH model would be appropriate for your series. If so, fit the model to the returns.

10.2 Daily closing prices of four major European stock indices are available for the period 1991–1998 in the file ''EuStockMarkets'' in the R datasets package; see help(EuStockMarkets) for details.

(a) Select two series and plot the data using R. Are there any unusual features worth noting? Perform a statistical test to determine the presence of a unit root in these series.

(b) Compute and plot the series of daily log returns. Do the graphs show evidence of volatility clustering? Perform a statistical analysis to determine whether AR–ARCH models would be appropriate for your series. State the final models selected.

10.3 Consider the ARCH(1) process $\{a_t\}$ defined by $a_t = \sigma_t e_t$, with $\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2$, where the e_t are independent, identically distributed variates with mean 0 and variance 1, and assume that $0 < \alpha_1 < 1$.

(a) Verify that $a_t^2 = \alpha_0 \sum_{j=0}^{\infty} \alpha_1^j e_t^2 e_{t-1}^2 \cdots e_{t-j}^2 = e_t^2 \alpha_0 \left(1 + \sum_{j=1}^{\infty} \alpha_1^j e_{t-1}^2 \cdots e_{t-j}^2\right)$ or

$$a_t = e_t \left\{ \alpha_0 \left(1 + \sum_{j=1}^{\infty} \alpha_1^j e_{t-1}^2 \cdots e_{t-j}^2 \right) \right\}^{1/2}$$

provides a causal (strictly) stationary representation (solution) of the ARCH model equations, that is, such that $\sigma_t^2 = \alpha_0 \left(1 + \sum_{j=1}^{\infty} \alpha_1^j e_{t-1}^2 \cdots e_{t-j}^2\right)$ satisfies $\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 \equiv \alpha_0 + \alpha_1 e_{t-1}^2 \sigma_{t-1}^2$.

(b) Use the representation for a_t in (a) to show that $E[a_t] = 0$, $E[a_t^2] = \text{var}[a_t] = \alpha_0/(1 - \alpha_1)$, and $E[a_t a_{t-k}] = \text{cov}[a_t, a_{t-k}] = 0$ for $k \neq 0$.

(c) Define $X_t = a_t^2$ and assume, in addition, that $\alpha_1^2 < \frac{1}{3}$, so that $E[a_t^4] < \infty$, that is, $E[X_t^2] < \infty$. Show that the process $\{X_t\}$ satisfies the relation $X_t = e_t^2(\alpha_0 + \alpha_1 X_{t-1})$, and deduce from this that the autocovariances of $\{X_t\}$ satisfy $\text{cov}[X_t, X_{t-k}] = \alpha_1 \text{cov}[X_{t-1}, X_{t-k}]$ for $k \geq 1$. Hence, conclude that $\{X_t\}$ has the same autocorrelation function as an AR(1) process with AR parameter $\phi = \alpha_1$.

10.4 Consider the GARCH(1, 1) model $a_t = \sigma_t e_t$, where the e_t are iid random variables with mean 0 and variance 1, and $\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2$. Show that the unconditional variance of a_t equals $\text{var}[a_t] = \alpha_0/[1 - (\alpha_1 + \beta_1)]$.

10.5 Derive the five-step-ahead forecast of the conditional variance σ_t^2 from a time origin h for the GARCH(1, 1) process. Repeat the derivation for a GARCH(2, 1) process.

10.6 Suppose that a time series of stock returns $\{r_t\}$ can be represented using an ARCH(1)-M process $r_t = \delta \sigma_t^2 + a_t$, $a_t = \sigma_t e_t$, and $\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2$, where the e_t are iid Normal(0, 1).

(a) Derive the conditional and unconditional mean of the series.

(b) Show that the ARCH-in-mean effect makes the $\{r_t\}$ serially correlated and calculate the ACF $\rho_k, k = 1, 2, \ldots$.

10.7 Assume that $\{z_t\}$ is a stationary, zero mean, Gaussian process with autocovariance function $\gamma_k(z)$ and autocorrelation function $\rho_k(z)$. Use the property that for zero

mean Gaussian variates,

$$E[z_t z_{t+i} z_{t+j} z_{t+k}] = E[z_t z_{t+i}]E[z_{t+j} z_{t+k}] + E[z_t z_{t+j}]E[z_{t+i} z_{t+k}] + E[z_t z_{t+k}]\, E[z_{t+i} z_{t+j}]$$

to show that $\text{cov}[z_t^2, z_{t+k}^2] = 2\gamma_k^2(z)$ and hence that the autocorrelation function of the process of squared values $X_t = z_t^2$ is $\rho_k(X) = \rho_k^2(z)$.

10.8 Consider the first-order bilinear model $z_t = \phi z_{t-1} + a_t - b z_{t-1} a_{t-1}$, where the a_t are independent variates with mean 0 and variance σ_a^2. Assume the process $\{z_t\}$ is stationary, which involves the condition that $\phi^2 + \sigma_a^2 b^2 < 1$, and assume that $\{z_t\}$ has a causal stationary representation of the form $z_t = a_t + f(a_{t-1}, a_{t-2}, \ldots)$.

(a) Verify that $E[z_t a_t] = \sigma_a^2$, and so also that $\mu_z = E[z_t]$ satisfies $(1 - \phi)\mu_z = -b\sigma_a^2$.

(b) Establish that the autocovariances γ_k of $\{z_t\}$ satisfy $\gamma_k = \phi\gamma_{k-1}$ for $k > 1$, so that the process has the same autocovariance structure as an ARMA(1, 1) process.

10.9 Consider the annual sunspot series referred to as Series E in this text. The series is also available for a slightly longer time period as series ''sunspot.year'' in the R datasets package.

(a) Plot the time series and fit an AR(3) model to the series.

(b) Use the procedure described by McLeod and Li (1983) to test for nonlinearity in the series.

(c) Repeat part (b) using the Keenan and Tsay tests for nonlinearity.

(d) Describe how you might fit a nonlinear time series model to this series.

10.10 Measurements of the annual flow of the river Nile at Aswan from 1871 to 1970 are provided as series ''Nile'' in the R datasets package; type help(Nile) for details.

(a) Plot the data along with the ACF and PACF of the series. Fit an appropriate ARIMA model to this series and comment.

(b) Perform a statistical analysis to determine whether there is evidence of long memory dependence in this series.

(c) If the answer in (b) is affirmative, develop a fractionallly integrated ARMA (i.e. ARFIMA) model for the series.

PART THREE

TRANSFER FUNCTION AND MULTIVARIATE MODEL BUILDING

Suppose that X measures the level of an *input* to a dynamic system. For example, X might be the concentration of some constituent in the feed to a chemical process. Suppose that the level of X influences the level of a system *output* Y. For example, Y might be the yield of product from the chemical process. It will usually be the case that because of the inertia of the system, a change in X from one level to another will have no immediate effect on the output but, instead, will produce delayed response with Y eventually coming to equilibrium at a new level. We refer to such a change as a *dynamic* response. A model that describes this dynamic response is called a *transfer function model*. We shall suppose that observations of input and output are made at equispaced intervals of time. The associated transfer function model will then be called a *discrete* transfer function model.

Models of this kind can describe not only the behavior of industrial processes but also that of economic and business systems. Transfer function model building is important because it is only when the dynamic characteristics of a system are understood that intelligent direction, manipulation, and control of the system is possible.

Even under carefully controlled conditions, influences other than X will affect Y. We refer to the combined effect on Y of such influences as the *disturbance* or the *noise*. Such model that can be related to real data must take account of not only the dynamic relationship associating X and Y but also the noise infecting the system. Such joint models are obtained by combining a deterministic transfer function model with a stochastic noise model.

In Chapter 11 we introduce a class of linear transfer function models capable of representing many of the dynamic relationships commonly met in practice. In Chapter 12 we show how, taking account of corrupting noise, they may be related to data. Given the observed series X and Y, the development of the combined transfer function and noise model is accomplished by procedures of *identification, estimation*, and *diagnostic checking,*

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

which closely parallel those already described for univariate time series. In Chapter 13 we describe how simple pulse and step indicator variables can be used as inputs in transfer function models to represent and assess the effects of unusual *intervention* events on the behavior of a time series Y. In Chapter 14 the concepts and methods of bivariate time series analysis and transfer function modeling are extended to the general study of dynamic relationships among *several* time series through development of statistical models and methods of *multivariate* time series analysis.

11

TRANSFER FUNCTION MODELS

In this chapter, we introduce a class of discrete linear transfer function models. These models take advantage of the dynamic relationship between two time series for prediction, control, and other applications. The models considered can be used to represent commonly occurring dynamic situations and are parsimonious in their use of parameters.

11.1 LINEAR TRANSFER FUNCTION MODELS

We assume that pairs of observations (X_t, Y_t) are available at equispaced intervals of time of an input X and an output Y from some dynamic system, as illustrated in Figure 11.1. In some situations, both X and Y are essentially continuous but are observed only at discrete times. It then makes sense to consider not only what the data has to tell us about the model representing transfer from one discrete series to another, but also what the discrete model might be able to tell us about the corresponding continuous model. In other examples, the discrete series are all that exist, and there is no underlying continuous process. Where we relate continuous and discrete systems, we shall use the basic sampling interval as the unit of time. That is, periods of time will be measured by the number of sampling intervals they occupy. Also, a discrete observation X_t will be deemed to have occurred ''at time t.''

When we consider the value of a continuous variable, say Y at time t, we denote it by $Y(t)$. If t happens to be a time at which a discrete variable Y is observed, its value is denoted by Y_t. When we wish to emphasize the dependence of a discrete output Y, not only on time but also on the level of the input X, we write $Y_t(X)$.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

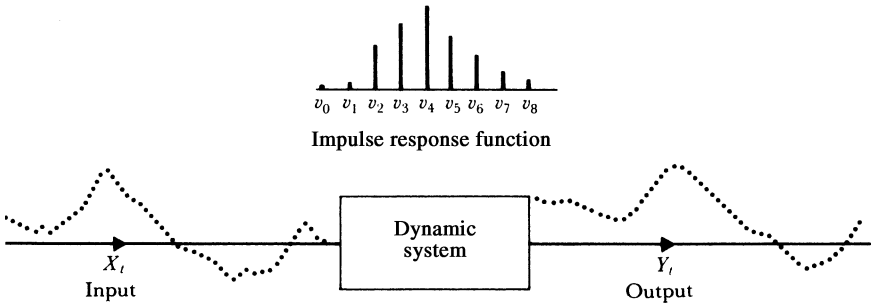


FIGURE 11.1 Input to, and output from, a dynamic system.

11.1.1 Discrete Transfer Function

With suitable inputs and outputs, which are left to the imagination of the reader, the dynamic system of Figure 11.1 might represent an industrial process, the economy of a country, or the behavior of a particular corporation or government agency.

From time to time, we refer to the *steady-state* level of the output obtained when the input is held at some fixed value. By this, we mean that the value $Y_\infty(X)$ at which the discrete output from a stable system *eventually* comes to equilibrium when the input is held at the fixed level X. Very often, over the range of interest, the relationship between $Y_\infty(X)$ and X will be approximately linear. Hence, if we use Y and X to denote *deviations* from convenient origins situated on the line, we can write the steady-state relationship as

$$Y_\infty = gX \tag{11.1.1}$$

where g is called the *steady-state gain*, and it is understood that Y_∞ is a function of X.

Now, suppose the level of the input is being varied and that X_t and Y_t represent *deviations* at time t from equilibrium. Then, it frequently happens that to an adequate approximation, the inertia of the system can be represented by a *linear filter* of the form

$$\begin{aligned} Y_t &= v_0 X_t + v_1 X_{t-1} + v_2 X_{t-2} + \cdots \\ &= (v_0 + v_1 B + v_2 B^2 + \cdots) X_t \\ &= v(B) X_t \end{aligned} \tag{11.1.2}$$

in which the output deviation at some time t is represented as a linear aggregate of input deviations at times $t, t-1, \ldots$. The operator $v(B)$ is called the *transfer function* of the filter.

Impulse Response Function. The weights $v_0, v_1, v_2, \ldots$ in (11.1.2) are called the *impulse response function* of the system. This is because the v_j may be regarded as the output or *response* at times $j \geq 0$ to a unit *pulse* input at time 0, that is, an input X_t such that $X_t = 1$ if $t = 0$, $X_t = 0$ otherwise. The impulse response function is shown in Figure 11.1 in the form of a bar chart. When there is no immediate response, one or more of the initial v's, say $v_0, v_1, \ldots, v_{b-1}$, will be equal to zero.

According to (11.1.2), the output deviation can be regarded as a linear aggregate of a series of superimposed impulse response functions scaled by the deviations X_t. This is illustrated in Figure 11.2, which shows a hypothetical impulse response function and the transfer it induces from the input to the output. In the situation illustrated, the input and

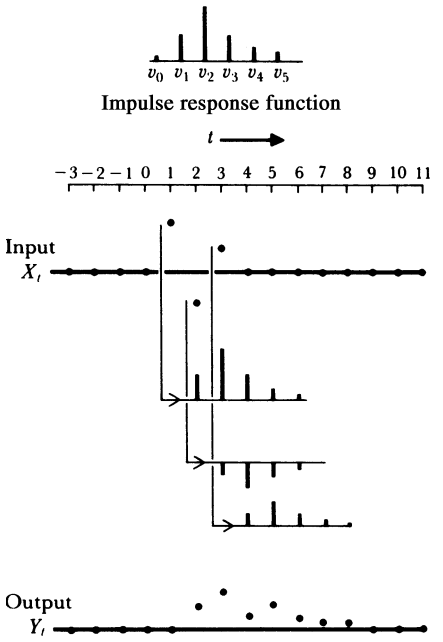


FIGURE 11.2 Linear transfer from input X_t to output Y_t.

output are initially in equilibrium. The deviations that occur in the input at times $t = 1$, $t = 2$, and $t = 3$ produce impulse response patterns of deviations in the output, which add together to produce the overall output response.

Relation Between the Incremental Changes. Denote by

$$y_t = Y_t - Y_{t-1} = \nabla Y_t$$

and by

$$x_t = X_t - X_{t-1} = \nabla X_t$$

the *incremental changes* in Y and X. We often wish to relate such changes. On differencing (11.1.2), we obtain

$$y_t = v(B)x_t$$

Thus, we see that the incremental changes y_t and x_t satisfy the same transfer function model as do Y_t and X_t.

Stability. If the infinite series $v_0 + v_1 B + v_2 B^2 + \cdots$ converges for $|B| \leq 1$, or equivalently, if the v_j are absolutely summable, so that $\sum_{j=0}^{\infty} |v_j| < \infty$, then the system is said to be *stable*. We shall be concerned here only with stable systems and consequently, impose this condition on the models we study. The stability condition implies that a finite incremental change in the input results in a finite incremental change in the output.

Now, suppose that X is held indefinitely at the value +1. Then, according to (11.1.1), Y will adjust and maintain itself at the value g. On substituting in (11.1.2) the values $Y_t = g$, $1 = X_t = X_{t-1} = X_{t-2} = \ldots$, we obtain

$$\sum_{j=0}^{\infty} v_j = g \tag{11.1.3}$$

Thus, for a stable system the sum of the impulse response weights converges and is equal to the steady-state gain of the system.

Parsimony. It would often be unsatisfactory to parameterize the system in terms of the v_j's of (11.1.2). The use of that many parameters could, at the estimation stage, lead to inaccurate and unstable estimation of the transfer function. Furthermore, it is usually inappropriate to estimate the weights v_j directly because for many real situations the v_j's would be functionally related, as we now see.

11.1.2 Continuous Dynamic Models Represented by Differential Equations

First-Order Dynamic System. Consider Figure 11.3. Suppose that at time t, $X(t)$ is the *volume* of water in tank A and $Y_1(t)$ the volume of water in tank B, which is connected to A by a pipe. For the time being we ignore tank C, shown by dashed lines. Now suppose that water can be forced in or out of A through pipe P and that mechanical devices are available that make it possible to force the level and hence the volume X in A to follow any desired pattern *irrespective* of what happens in B.

Now if the volume X in the first tank is held at some *fixed* level, water will flow from one tank to the other until the levels are equal. If we now reset the volume X to some other value, again a flow between the tanks will occur until equilibrium is reached. The volume in B at equilibrium as a function of the fixed volume in A yields the steady-state

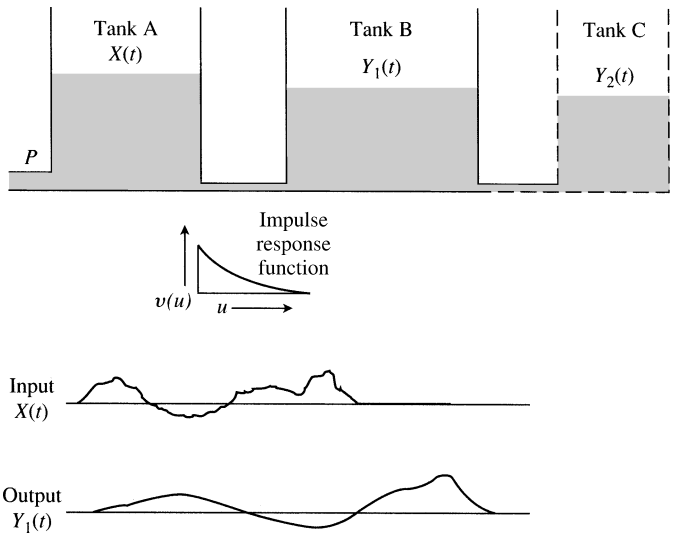


FIGURE 11.3 Representation of a simple dynamic system.

relationship

$$Y_{1\infty} = g_1 X \tag{11.1.4}$$

In this case the steady-state gain g_1 physically represents the ratio of the cross-sectional areas of the two tanks. If the levels are not in equilibrium at some time t, it is to be noted that the difference in the water level between the tanks is proportional to $g_1 X(t) - Y_1(t)$.

Suppose now that by forcing liquid in and out of pipe P, the volume $X(t)$ is made to follow a pattern like that labeled ''Input $X(t)$'' in Figure 11.3. Then, the volume $Y_1(t)$ in B will correspondingly change in some pattern such as that labeled on the figure as ''Output $Y_1(t)$.'' In general, the function $X(t)$ that is responsible for driving the system is called the *forcing function*.

To relate output to input, we note that to a close approximation, the rate of flow through the pipe will be proportional to the difference in head. That is,

$$\frac{dY_1}{dt} = \frac{1}{T_1}[g_1 X(t) - Y_1(t)] \tag{11.1.5}$$

where T_1 is a constant. The differential equation (11.1.5) may be rewritten in the form

$$(1 + T_1 D)Y_1(t) = g_1 X(t) \tag{11.1.6}$$

where $D = d/dt$. The dynamic system so represented by a first-order differential equation is often referred to as a first-order dynamic system. The constant T_1 is called the *time constant* of the system. The same first-order model can approximately represent the behavior of many simple systems. For example, $Y_1(t)$ might be the outlet temperature of water from a water heater, and $X(t)$ the flow rate of water into the heater.

It is possible to show (see, e.g., Jenkins and Watts, 1968) that the solution of a linear differential equation such as (11.1.6) can be written in the form

$$Y_1(t) = \int_0^\infty v(u)X(t-u)du \tag{11.1.7}$$

where in general $v(u)$ is the (continuous) impulse response function. We see that $Y_1(t)$ is generated from $X(t)$ as a continuously weighted aggregate, just as Y_t is generated from X_t as a discretely weighted aggregate in (11.1.2). Furthermore, we see that the role of weight function played by $v(u)$ in the continuous case is precisely parallel to that played by v_j in the discrete situation. For the particular first-order system defined by (11.1.6),

$$v(u) = g_1 T_1^{-1} e^{-u/T_1}$$

Thus, the impulse response in this case undergoes simple exponential decay, as indicated in Figure 11.3.

In the continuous case, determination of the output for a completely arbitrary forcing function, such as shown in Figure 11.3, is normally accomplished by simulation on an analog computer, or by using numerical procedures on a digital machine. Solutions are available analytically only for special forcing functions. Suppose, for example, that with the hydraulic system empty, $X(t)$ was suddenly raised to a level $X(t) = 1$ and maintained at that value. Then, we shall refer to the forcing function, which was at a steady level of zero and changed instantaneously to a steady level of unity, as a (unit) *step function*. The response of the system to such a function, called the *step response* to the system, is derived

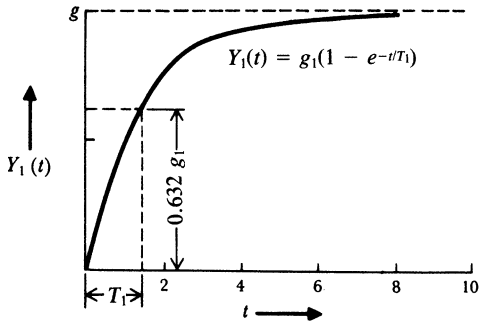


FIGURE 11.4 Response of a first-order system to a unit step change.

by solving the differential equation (11.1.6) with a unit step input, to obtain

$$Y_1(t) = g_1(1 - e^{-t/T_1}) \tag{11.1.8}$$

Thus, the level in tank B rises exponentially in the manner shown in Figure 11.4. Now, when $t = T_1, Y_1(t) = g_1(1 - e^{-1}) = 0.632g_1$. Thus, the time constant T_1 is the time required after the initiation of a step input for the first-order system (11.1.6) to reach 63.2% of its final equilibrium level.

Sometimes there is an initial period of pure *delay* or *dead time* before the response to a given input change begins to take effect. For example, if there were a long length of pipe between A and B in Figure 11.3, a sudden change in level in A could not begin to take effect until liquid had flowed down the pipe. Suppose that the delay thus introduced occupies τ units of time. Then, the response of the delayed system would be represented by a differential equation like (11.1.6), but with $t - \tau$ replacing t on the right-hand side, so that

$$(1 + T_1 D)Y_1(t) = g_1 X(t - \tau) \tag{11.1.9}$$

The corresponding impulse and step response functions for this system would be of precisely the same shape as for the undelayed system, but the functions would be translated along the horizontal axis a distance τ.

Second-Order Dynamic System. Consider Figure 11.3 once more. Imagine a three-tank system in which a pipe leads from tank B to a third tank C, the volume of liquid in which is denoted by $Y_2(t)$. Let T_2 be the time constant for the additional system and g_2 its steady-state gain. Then, $Y_2(t)$ and $Y_1(t)$ are related by the differential equation

$$(1 + T_2 D)Y_2(t) = g_2 Y_1(t)$$

After substitution in (11.1.6), we obtain a *second-order* differential equation linking the output from the third tank and the input to the first:

$$[1 + (T_1 + T_2)D + T_1 T_2 D^2]Y_2(t) = gX(t) \tag{11.1.10}$$

where $g = g_1 g_2$. For such a system, the impulse response function is a mixture of two exponentials

$$v(u) = \frac{g(e^{-u/T_1} - e^{-u/T_2})}{T_1 - T_2} \tag{11.1.11}$$

and the response to a unit step is given by

$$Y_2(t) = g\left(1 - \frac{T_1 e^{-t/T_1} - T_2 e^{-t/T_2}}{T_1 - T_2}\right) \tag{11.1.12}$$

The continuous curve R in Figure 11.5 shows the response to a unit step for the system

$$(1 + 3D + 2D^2)Y_2(t) = 5X(t)$$

for which $T_1 = 1$, $T_2 = 2$, $g = 5$. Note that unlike the first-order system, the second-order system has a step response that has zero slope initially.

A more general second-order system is defined by

$$(1 + \Xi_1 D + \Xi_2 D^2)Y(t) = gX(t) \tag{11.1.13}$$

where

$$\Xi_1 = T_1 + T_2 \qquad \Xi_2 = T_1 T_2 \tag{11.1.14}$$

and the constants T_1 and T_2 may be complex. If we write

$$T_1 = \frac{1}{\zeta} e^{i\lambda} \qquad T_2 = \frac{1}{\zeta} e^{-i\lambda} \tag{11.1.15}$$

then (11.1.13) becomes

$$\left(1 + \frac{2\cos\lambda}{\zeta} D + \frac{1}{\zeta^2} D^2\right) Y(t) = gX(t) \tag{11.1.16}$$

The impulse response function (11.1.11) then reduces to

$$v(u) = g \frac{\zeta e^{-\zeta u \cos\lambda} \sin(\zeta u \sin\lambda)}{\sin\lambda} \tag{11.1.17}$$

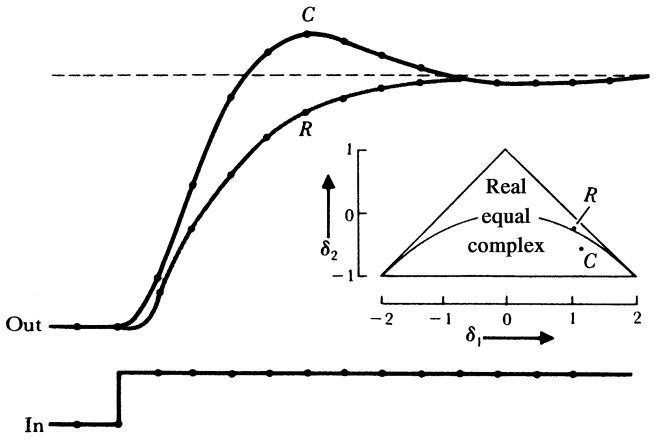


FIGURE 11.5 Step responses of coincident, discrete, and continuous second-order systems having characteristic equations with real roots (curve ***R***) and complex roots (curve ***C***).

and the response (11.1.12) to a unit step, to

$$Y(t) = g\left[1 - \frac{e^{-\zeta t \cos\lambda}\sin(\zeta t \sin\lambda + \lambda)}{\sin\lambda}\right] \tag{11.1.18}$$

The continuous curve C in Figure 11.5 shows the response to a unit step for the system

$$(1 + \sqrt{2}D + 2D^2)Y(t) = 5X(t)$$

for which $\lambda = \pi/3$ and $\zeta = \sqrt{2}/2$. It will be noticed that the response overshoots the value $g = 5$ and then comes to equilibrium as a damped sine wave. This behavior is typical of underdamped systems, as they are called. In general, a second-order system is said to be *overdamped, critically damped*, or *underdamped*, depending on whether the constants T_1 and T_2 are real, real and equal, or complex. The overdamped system has a step response that is a mixture of two exponentials, given by (11.1.12), and will always remain below the asymptote $Y(\infty) = g$. As with the first-order system, the response can be made subject to a period of dead time by replacing t on the right-hand side of (11.1.13) by $t - \tau$. Many quite complicated dynamic systems can be closely approximated by such second-order systems with delay.

More elaborate linear dynamic systems can be represented by allowing not only the level of the forcing function $X(t)$ but also its rate of change dX/dt and higher derivatives to influence the behavior of the system. Thus, a general model for representing (continuous) dynamic systems is the linear differential equation

$$(1 + \Xi_1 D + \cdots + \Xi_R D^R)Y(t) = g(1 + H_1 D + \cdots + H_S D^S)X(t - \tau) \tag{11.1.19}$$

11.2 DISCRETE DYNAMIC MODELS REPRESENTED BY DIFFERENCE EQUATIONS

11.2.1 General Form of the Difference Equation

Corresponding to the continuous representation (11.1.19), discrete dynamic systems are often parsimoniously represented by the general linear *difference* equation

$$(1 + \xi_1\nabla + \cdots + \xi_r\nabla^r)Y_t = g(1 + \eta_1\nabla + \cdots + \eta_s\nabla^s)X_{t-b} \tag{11.2.1}$$

which we refer to as a transfer function model of order (r, s). The difference equation (11.2.1) may also be written in terms of the backward shift operator B, with $\nabla = 1 - B$, as

$$(1 - \delta_1 B - \cdots - \delta_r B^r)Y_t = (\omega_0 - \omega_1 B - \cdots - \omega_s B^s)X_{t-b} \tag{11.2.2}$$

or as

$$\delta(B)Y_t = \omega(B)X_{t-b}$$

Equivalently, writing $\Omega(B) = \omega(B)B^b$, the model becomes

$$\delta(B)Y_t = \Omega(B)X_t \tag{11.2.3}$$

Comparing (11.2.3) with (11.1.2) we see that the transfer function for this model is

$$v(B) = \delta^{-1}(B)\Omega(B) \tag{11.2.4}$$

Thus, the transfer function is represented by the ratio of two polynomial operators in B.

Dynamics of ARIMA Stochastic Models. The ARIMA model

$$\varphi(B)z_t = \theta(B)a_t$$

used for the representation of a time series $\{z_t\}$ relates z_t and a_t by the linear filtering operation

$$z_t = \varphi^{-1}(B)\theta(B)a_t$$

where a_t is white noise. Thus, the ARIMA model postulates that a time series can be usefully represented as an output from a dynamic system to which the input is white noise and for which the transfer function can be parsimoniously expressed as the ratio of two polynomial operators in B.

Stability of the Discrete Models. The requirement of stability for the discrete transfer function models exactly parallels that of stationarity for the ARMA stochastic models. In general, for stability we require that the roots of the characteristic equation

$$\delta(B) = 0$$

with B regarded as a variable, lie outside the unit circle. In particular, this implies that for the first-order model with $\delta(B) = 1 - \delta_1 B$, the parameter δ_1 satisfies

$$-1 < \delta_1 < 1$$

and for the second-order model (see, e.g., Fig. 11.5), the parameters δ_1, δ_2 satisfy

$$\delta_2 + \delta_1 < 1$$

$$\delta_2 - \delta_1 < 1$$

$$-1 < \delta_2 < 1$$

On writing (11.2.2) in full as

$$Y_t = \delta_1 Y_{t-1} + \cdots + \delta_r Y_{t-r} + \omega_0 X_{t-b} - \omega_1 X_{t-b-1} - \cdots - \omega_s X_{t-b-s}$$

we see that if X_t is held indefinitely at a value $+1$, Y_t will eventually reach the value

$$g = \frac{\omega_0 - \omega_1 - \cdots - \omega_s}{1 - \delta_1 - \cdots - \delta_r} \tag{11.2.5}$$

which expresses the steady-state gain in terms of the parameters of the model.

11.2.2 Nature of the Transfer Function

If we employ a transfer function model defined by the difference equation (11.2.2), then substituting

$$Y_t = v(B)X_t \tag{11.2.6}$$

in (11.2.2), we obtain the identity

$$\begin{aligned}(1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r)(v_0 + v_1 B + v_2 B^2 + \cdots) \\ = (\omega_0 - \omega_1 B - \cdots - \omega_s B^s)B^b\end{aligned} \tag{11.2.7}$$

On equating coefficients of B, we find

$$v_j = \begin{cases} 0 & j < b \\ \delta_1 v_{j-1} + \delta_2 v_{j-2} + \cdots + \delta_r v_{j-r} + \omega_0 & j = b \\ \delta_1 v_{j-1} + \delta_2 v_{j-2} + \cdots + \delta_r v_{j-r} - \omega_{j-b} & j = b+1, b+2, \ldots, b+s \\ \delta_1 v_{j-1} + \delta_2 v_{j-2} + \cdots + \delta_r v_{j-r} & j > b+s \end{cases} \tag{11.2.8}$$

The weights $v_{b+s}, v_{b+s-1}, \ldots, v_{b+s-r+1}$ supply r starting values for the homogeneous difference equation

$$\delta(B)v_j = 0 \qquad j > b + s$$

The solution $v_j = f(\boldsymbol{\delta}, \boldsymbol{\omega}, j)$ of this difference equation applies to all values v_j for which $j \geq b + s - r + 1$.

Thus, in general, the impulse response weights v_j consist of:

1. b zero values $v_0, v_1, \ldots, v_{b-1}$.
2. A further $s - r + 1$ values $v_b, v_{b+1}, \ldots, v_{b+s-r}$ following no fixed pattern (no such values occur if $s < r$).
3. Values v_j with $j \geq b + s - r + 1$ following the pattern dictated by the rth-order difference equation, which has r starting values $v_{b+s}, v_{b+s-1}, \ldots, v_{b+s-r+1}$. Starting values v_j for $j < b$ will, of course, be zero.

Step Response. We now write $V(B)$ for the generating function of the step response weights V_j, which represent the *response* at times $j \geq 0$ to a unit *step* at time 0, $X_t = 1$ if $t \geq 0$, $X_t = 0$ if $t < 0$, so that $V_j = \sum_{i=0}^{j} v_i$ for $j \geq 0$. Thus,

$$\begin{aligned}V(B) &= V_0 + V_1 B + V_2 B^2 + \cdots \\ &= v_0 + (v_0 + v_1)B + (v_0 + v_1 + v_2)B^2 + \cdots\end{aligned} \tag{11.2.9}$$

and

$$v(B) = (1 - B)V(B) \tag{11.2.10}$$

Substitution of (11.2.10) in (11.2.7) yields the identity

$$(1 - \delta_1^* B - \delta_2^* B^2 - \cdots - \delta_{r+1}^* B^{r+1})(V_0 + V_1 B + V_2 B^2 + \cdots) = (\omega_0 - \omega_1 B - \cdots - \omega_s B^s) B^b \tag{11.2.11}$$

with

$$(1 - \delta_1^* B - \delta_2^* B^2 - \cdots - \delta_{r+1}^* B^{r+1}) = (1 - B)(1 - \delta_1 B - \cdots - \delta_r B^r) \tag{11.2.12}$$

The identity (11.2.11) for the step response weights V_j exactly parallels the identity (11.2.7) for the impulse response weights, except that the left-hand operator $\delta^*(B)$ is of order $r + 1$ instead of r.

Using the results (11.2.8), it follows that the step response function is defined by:

1. b zero values $V_0, V_1, \ldots, V_{b-1}$.
2. A further $s - r$ values $V_b, V_{b+1}, \ldots, V_{b+s-r-1}$ following no fixed pattern (no such values occur if $s < r + 1$).
3. Values V_j, with $j \geq b + s - r$, which follow the pattern dictated by the $(r + 1)$th-order difference equation $\delta^*(B)V_j = 0$, which has $r + 1$ starting values V_{b+s}, $V_{b+s-1}, \ldots, V_{b+s-r}$. Starting values V_j for $j < b$ will, of course, be zero.

11.2.3 First- and Second-Order Discrete Transfer Function Models

Details of transfer function models for all combinations of $r = 0, 1, 2$ and $s = 0, 1, 2$ are shown in Table 11.1. Specific examples of the models, with bar charts showing step response and impulse response, are given in Figure 11.6. The equations at the end of Table 11.1 allow the parameters ξ, g, η of the ∇ form of the model to be expressed in terms of the parameters $\boldsymbol{\delta}, \boldsymbol{\omega}$ of the B form. These equations are given for the most general of the models considered, namely that for which $r = 2$ and $s = 2$. All the other models are special cases of this one, and the corresponding equations for these are obtained by setting appropriate parameters to zero. For example, if $r = 1$ and $s = 1, \xi_2 = \eta_2 = \delta_2 = \omega_2 = 0$, then

$$\delta_1 = \frac{\xi_1}{1 + \xi_1} \qquad \omega_0 = \frac{g(1 + \eta_1)}{1 + \xi_1} \qquad \omega_1 = \frac{g\eta_1}{1 + \xi_1}$$

In Figure 11.6, the starting values for the difference equations satisfied by the impulse and step responses, respectively, are indicated by circles on the bar charts.

Discussion of the Models in Table 11.1. The models, whose properties are summarized in Table 11.1 and Figure 11.6, will require careful study, since they are useful in representing many commonly met dynamic systems. In all the models the operator B^b on the right ensures that the first nonzero term in the impulse response function is v_b. In the examples in Figure 11.6, the value of g is assumed to equal 1, and b is assumed to equal 3.

Models with $r = 0$. With r and s both equal to zero, the impulse response consists of a single nonzero value $v_b = \omega_0 = g$. The output is proportional to the input but is displaced by b time intervals. More generally, if we have an operator of order s on the right, the instantaneous input will be delayed b intervals and will be spread over $s + 1$ values in proportion to $v_b = \omega_0, v_{b+1} = -\omega_1, \ldots, v_{b+s} = -\omega_s$. The step response is obtained by summing the impulse

TABLE 11.1 Impulse Response Functions for Transfer Function Models of the Form $\delta_r(B)Y_t = \omega_s(B)B^bX_t$

rsb	∇ Form	B Form	Impulse Response V_j	
00b	$Y_t = gX_{t-b}$	$Y_t = \omega_0 B^b X_t$	0	$j < b$
			ω_0	$j = b$
			0	$j > b$
01b	$Y_t = g(1 + \eta_1\nabla)X_{t-b}$	$Y_t = (\omega_0 - \omega_1 B)B^b X_t$	0	$j < b$
			ω_0	$j = b$
			$-\omega_1$	$j = b + 1$
			0	$j > b + 1$
02b	$Y_t = g(1 + \eta_1\nabla + \eta_2\nabla^2)X_{t-b}$	$Y_t = (\omega_0 - \omega_1 B - \omega_2 B^2)B^b X_t$	0	$j < b$
			ω_0	$j = b$
			$-\omega_1$	$j = b + 1$
			$-\omega_2$	$j = b + 2$
			0	$j > b + 2$
10b	$(1 + \xi_1\nabla)Y_t = gX_{t-b}$	$(1 - \delta_1 B)Y_t = \omega_0 B^b X_t$	0	$j < b$
			ω_0	$j = b$
			$\delta_1 v_{j-1}$	$j > b$
11b	$(1 + \xi_1\nabla)Y_t = g(1 + \eta_1\nabla)X_{t-b}$	$(1 - \delta_1 B)Y_t = (\omega_0 - \omega_1 B)B^b X_t$	0	$j < b$
			ω_0	$j = b$
			$\delta_1\omega_0 - \omega_1$	$j = b + 1$
			$\delta_1 v_{j-1}$	$j > b + 1$
12b	$(1 + \xi_1\nabla)Y_t = g(1 + \eta_1\nabla + \eta_2\nabla^2)X_{t-b}$	$(1 - \delta_1 B)Y_t = (\omega_0 - \omega_1 B - \omega_2 B^2)B^b X_t$	0	$j < b$
			ω_0	$j = b$
			$\delta_1\omega_0 - \omega_1$	$j = b + 1$
			$\delta_1^2\omega_0 - \delta_1\omega_1 - \omega_2$	$j = b + 2$
			$\delta_1 v_{j-1}$	$j > b + 2$

(*continued*)

TABLE 11.1 Impulse Response Functions for Transfer Function Models of the Form $\delta_r(B)Y_t = \omega_s(B)B^bX_t$ (*continued*)

rsb	∇ Form	B Form	Impulse Response V_j	
20*b*	$(1+\xi_1\nabla+\xi_2\nabla^2)Y_t = gX_{t-b}$	$(1-\delta_1B-\delta_2B^2)Y_t = \omega_0B^bX_t$	0	$j<b$
			ω_0	$j=b$
			$\delta_1v_{j-1}+\delta_1v_{j-2}$	$j>b$
21*b*	$(1+\xi_1\nabla+\xi_2\nabla^2)Y_t = g(1+\eta_1\nabla)X_{t-b}$	$(1-\delta_1B-\delta_2B^2)Y_t = (\omega_0-\omega_1B)B^bX_t$	0	$j<b$
			ω_0	$j=b$
			$\delta_1\omega_0-\omega_1$	$j=b+1$
			$\delta_1v_{j-1}+\delta_2v_{j-2}$	$j>b+1$
22*b*	$(1+\xi_1\nabla+\xi_2\nabla^2)Y_t = g(1+\eta_1\nabla+\eta_2\nabla^2)X_{t-b}$	$(1-\delta_1B-\delta_2B^2)Y_t = (\omega_0-\omega_1B-\omega_2B^2)B^bX_t$	0	$j<b$
			ω_0	$j=b$
			$\delta_1\omega_0-\omega_1$	$j=b+1$
			$(\delta_1^2+\delta_2)\omega_0-\delta_1\omega_1-\omega_2$	$j=b+2$
			$\delta_1v_{j-1}+\delta_2v_{j-2}$	$j=b+2$

$$\xi_1 = \frac{\delta_1+2\delta_2}{1-\delta_1-\delta_2}, \quad \xi_2 = \frac{-\delta_2}{1-\delta_1-\delta_2} \qquad \delta_1 = \frac{\xi_1+2\xi_2}{1+\xi_1+\xi_2}, \quad \delta_2 = \frac{-\xi_2}{1+\xi_1+\xi_2}$$

$$g = \frac{\omega_0-\omega_1-\omega_2}{1-\delta_1-\delta_2} \qquad \omega_0 = \frac{g(1+\eta_1+\eta_2)}{1+\xi_1+\xi_2}$$

$$\eta_1 = \frac{\omega_1+2\omega_2}{\omega_0-\omega_1-\omega_2} \qquad \omega_1 = \frac{g(\eta_1+2\eta_2)}{1+\xi_1+\xi_2}$$

$$\eta_2 = \frac{-\omega_2}{\omega_0-\omega_1-\omega_2} \qquad \omega_2 = \frac{-g\eta_2}{1+\xi_1+\xi_2}$$

$$1-\delta_1-\delta_2 = (1+\xi_1+\xi_2)^{-1}$$

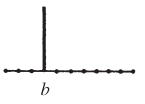
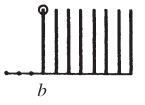
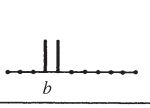
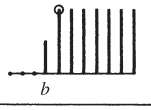
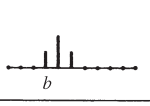
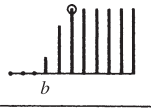
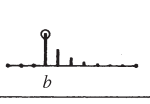
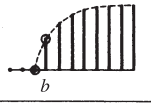
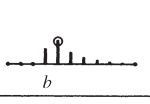
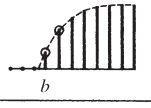
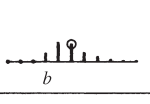
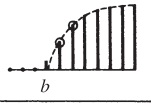
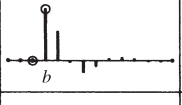
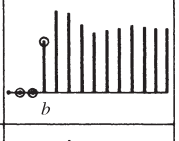
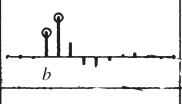
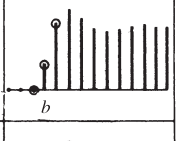
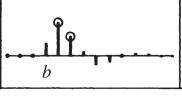
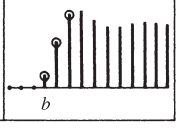
r, s, b	∇ Form	B Form	Impulse Response v_j	Step Response $V_j = \sum_{i=0}^{j} v_i$
003	$Y_t = X_{t-3}$	$Y_t = B^3 X_t$		
013	$Y_t = (1 - 0.5\nabla)\, X_{t-3}$	$Y_t = (0.5 + 0.5B)\, B^3 X_t$		
023	$Y_t = (1 - \nabla + 0.25\nabla^2)\, X_{t-3}$	$Y_t = (0.25 + 0.50B + 0.25B^2)\, B^3 X_t$		
103	$(1+\nabla)\, Y_t = X_{t-3}$	$(1 - 0.5B)\, Y_t = 0.5B^3 X_t$		
113	$(1+\nabla)\, Y_t = (1 - 0.5\nabla)\, X_{t-3}$	$(1 - 0.5B)\, Y_t = (0.25 + 0.25B)\, B^3 X_t$		
123	$(1+\nabla)\, Y_t = (1 - \nabla + 0.25\nabla^2)\, X_{t-3}$	$(1 - 0.5B)\, Y_t = (0.125 + 0.25B + 0.125B^2)\, B^3 X_t$		
203	$(1 - 0.25\nabla + 0.5\nabla^2)\, Y_t = X_{t-3}$	$(1 - 0.6B + 0.4B^2)\, Y_t = 0.8B^3 X_t$		
213	$(1 - 0.25\nabla + 0.5\nabla^2)\, Y_t = (1 - 0.5\nabla)\, X_{t-3}$	$(1 - 0.6B + 0.4B^2)\, Y_t = (0.4 + 0.4B)\, B^3 X$		
223	$(1 - 0.25\nabla + 0.5\nabla^2)\, Y_t = (1 - \nabla + 0.25\nabla^2)\, X_{t-3}$	$(1 - 0.6B + 0.4B^2)\, Y_t = (0.2 + 0.4B + 0.2B^2)\, B^3 X_t$		

FIGURE 11.6 Examples of impulse and step response functions with gain $g = 1$.

response and eventually satisfies the difference equation $(1 - B)V_j = 0$ with starting values $V_{b+s} = g = \omega_0 - \omega_1 - \cdots - \omega_s$.

Models with $r = 1$. With $s = 0$, the impulse response tails off exponentially (geometrically) from the initial starting value $v_b = \omega_0 = g/(1 + \xi_1) = g(1 - \delta_1)$. The step re-

sponse increases exponentially until it attains the value $g = 1$. If the exponential step response is extrapolated backwards as indicated by the dashed line, it cuts the time axis at time $b - 1$. This corresponds to the fact that $V_{b-1} = 0$ as well as $V_b = v_b$ are starting values for the appropriate difference equation $(1 - \delta_1 B)(1 - B)V_j = 0$.

With $s = 1$, there is an initial value $v_b = \omega_0 = g(1 + \eta_1)/(1 + \xi_1)$ of the impulse response, which does not follow a pattern. The exponential pattern induced by the difference equation $v_j = \delta_1 v_{j-1}$ associated with the left-hand operator begins with the starting value $v_{b+1} = (\delta_1 \omega_0 - \omega_1) = g(\xi_1 - \eta_1)/(1 + \xi_1)^2$. The step response function follows an exponential curve, determined by the difference equation $(1 - \delta_1 B)(1 - B)V_j = 0$, which approaches g asymptotically from the starting value $V_b = v_b$ and $V_{b+1} = v_b + v_{b+1}$. An exponential curve projected by the dashed line backwards through the points will, in general, cut the time axis at some intermediate point in the time interval. We show in Section 11.3 that certain discrete models, which approximate continuous first-order systems having *fractional* periods of delay, may in fact be represented by a first-order difference equation with an operator of order $s = 1$ on the right.

With $s = 2$, there are two values v_b and v_{b+1} for the impulse response that do not follow a pattern, followed by exponential fall off beginning with v_{b+2}. Correspondingly, there is a single preliminary value V_b in the step response that does not coincide with the exponential curve projected by the dashed line. This curve is, as before, determined by the difference equation $(1 - \delta_1 B)(1 - B)V_j = 0$ but with starting values V_{b+1} and V_{b+2}.

Models with r = 2. The flexibility of the model with $s = 0$ is limited because the first starting value of the impulse response is fixed to be zero. More useful models are obtained for $s = 1$ and $s = 2$. The use of these models in approximating continuous second-order systems is discussed in Section 11.3 and in Appendix A11.1.

The behavior of the dynamic weights v_j, which eventually satisfy

$$v_j - \delta_1 v_{j-1} - \delta_2 v_{j-2} = 0 \qquad j > b + s \tag{11.2.13}$$

depends on the nature of the roots S_1^{-1} and S_2^{-1}, of the *characteristic equation*

$$1 - \delta_1 B - \delta_2 B^2 = (1 - S_1 B)(1 - S_2 B) = 0$$

This dependence is shown in Table 11.2. As in the continuous case, the model may be overdamped, critically damped, or underdamped, depending on the nature of the roots of the characteristic equation.

When the roots are complex, the solution of (11.2.13) will follow a damped sine wave, as in the examples of second-order systems in Figure 11.6. When the roots are real, the solution will be the sum of two exponentials. As in the continuous case considered in

TABLE 11.2 Dependence of Nature of Second-Order System on the Roots of $1 - \delta_1 B - \delta_2 B^2 = 0$

Roots (S_1^{-1}, S_2^{-1})	Condition	Damping
Real	$\delta_1^2 + 4\delta_2 > 0$	Overdamped
Real and equal	$\delta_1^2 + 4\delta_2 = 0$	Critically damped
Complex	$\delta_1^2 + 4\delta_2 < 0$	Underdamped

Section 11.1.2, the system can then be thought of as equivalent to two discrete first-order systems arranged in series and having parameters S_1 and S_2.

The weights V_j for the step response eventually satisfy a difference equation

$$(V_j - g) - \delta_1(V_{j-1} - g) - \delta_2(V_{j-2} - g) = 0$$

which is of the same form as (11.2.13). Thus, the behavior of the step response V_j about its asymptotic value g parallels the behavior of the impulse response about the time axis. In the situation where there are complex roots, the step response ''overshoots'' the value g and then oscillates about this value until it reaches equilibrium. When the roots are real and positive, the step response, which is the sum of two exponential terms, approaches its asymptote g without crossing it. However, if there are negative real roots, the step response may overshoot and oscillate as it settles down to its equilibrium value.

In Figure 11.5, the dots indicate two discrete step responses, labeled R and C, respectively, in relation to a discrete step input indicated by dots at the bottom of the figure. The difference equation models[1] corresponding to R and C are

$$\text{R}: \quad (1 - 0.97B + 0.22B^2)Y_t = 5(0.15 + 0.09B)X_{t-1}$$

$$\text{C}: \quad (1 - 1.15B + 0.49B^2)Y_t = 5(0.19 + 0.15B)X_{t-1}$$

Also shown in Figure 11.5 is a diagram of the stability region with the parameter points (δ_1, δ_2) marked for each of the two models. Note that the system described by model R, which has real positive roots, has no overshoot while that for model C, which has complex roots, does have overshoot.

11.2.4 Recursive Computation of Output for Any Input

It would be extremely tedious if it were necessary to use the impulse response form (11.1.2) of the model to compute the output for a given input. Fortunately, this is not necessary. Instead, we may employ the difference equation model directly. In this way it is a simple matter to compute the output recursively for any input. For example, consider the model with $r = 1, s = 0, b = 1$, and with $\xi = 1$ and $g = 5$. Thus,

$$(1 + \nabla)Y_t = 5X_{t-1}$$

or equivalently,

$$(1 - 0.5B)Y_t = 2.5X_{t-1} \tag{11.2.14}$$

Table 11.3 shows the calculation of Y_t when the input X_t is (a) a unit pulse input, (b) a unit step input, and (c) a ''general'' input. In all cases, it is assumed that the output has the initial value $Y_0 = 0$. To perform the recursive calculation, the difference equation is written out with Y_t on the left. Thus,

$$Y_t = 0.5Y_{t-1} + 2.5X_{t-1}$$

[1]The parameters in these models were in fact selected, in a manner to be discussed in Section 11.3.2, so that at the discrete points, the step responses exactly matched those of the continuous systems introduced in Section 11.1.2.

TABLE 11.3 Calculation of Output from Discrete First-Order System for Impulse, Step, and General Input

	(a) Impulse Input		(b) Step Input		(c) General Input	
t	Input X_t	Output Y_t	Input X_t	Output Y_t	Input X_t	Output Y_t
0	0	0	0	0	0	0
1	1	0	1	0	1.5	0
2	0	2.50	1	2.50	0.5	3.75
3	0	1.25	1	3.75	2.0	3.12
4	0	0.62	1	4.38	1.0	6.56
5	0	0.31	1	4.69	−2.5	5.78
6	0	0.16	1	4.84	0.5	−3.36

and, for example, in the case of the ''general'' input

$$Y_1 = 0.5 \times 0 + 2.5 \times 0 = 0$$
$$Y_2 = 0.5 \times 0 + 2.5 \times 1.5 = 3.75$$
$$Y_3 = 0.5 \times 3.75 + 2.5 \times 0.5 = 3.125$$

and so on. These inputs and outputs are plotted in Figure 11.7(a), (b), and (c).

In general, we see that having written the transfer function model in the form

$$Y_t = \delta_1 Y_{t-1} + \cdots + \delta_r Y_{t-r} + \omega_0 X_{t-b} - \omega_1 X_{t-b-1} - \cdots - \omega_s X_{t-b-s}$$

it is an easy matter to compute the discrete output for any discrete input. To start off the recursion, we need to know certain initial values. This need is not, of course, a shortcoming of the method of calculation but comes about because with a transfer function model, the initial values of Y *will* depend on values of X that occurred before observation was begun. In practice, when the necessary initial values are not known, we can substitute mean values for unknown Y's and X's (zeros if these quantities are considered as deviations from their means). The early calculated values will then depend upon this choice of the starting values. However, for a stable system, the effect of this choice will be negligible after a period sufficient for the impulse response to become negligible. If this period is p_o time intervals, an alternative procedure is to compute $Y_{p_o}, Y_{p_o+1}, \ldots$ directly from the impulse response until enough values are available to set the recursion going.

11.2.5 Transfer Function Models with Added Noise

In practice, the output Y could not be expected to follow exactly the pattern determined by the transfer function model, even if that model were entirely adequate. Disturbances of various kinds other than X normally corrupt the system. A disturbance might originate at any point in the system, but it is often convenient to consider it in terms of its net effect on the output Y, as indicated in Figure 1.5. If we assume that the disturbance, or noise N_t, is independent of the level of X and is additive with respect to the influence of X,

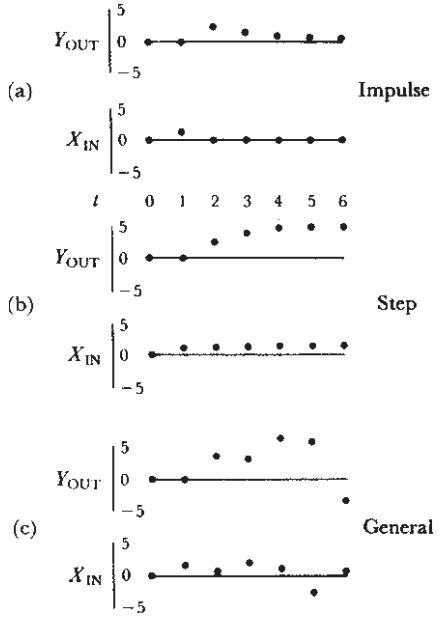


FIGURE 11.7 Response of a first-order system to (a) an impulse, (b) a step, and (c) a ''general'' input.

we can write

$$Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + N_t \tag{11.2.15}$$

If the noise process N_t can be represented by an ARIMA(p, d, q) model

$$N_t = \varphi^{-1}(B)\theta(B)a_t$$

where a_t is white noise, the model (11.2.15) can be written finally as

$$Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + \varphi^{-1}(B)\theta(B)a_t \tag{11.2.16}$$

In Chapter 12, we describe methods for identifying, fitting, and checking combined transfer function–noise models of the form (11.2.16).

11.3 RELATION BETWEEN DISCRETE AND CONTINUOUS MODELS

The discrete dynamic model, defined by a linear difference equation, is of importance in its own right. It provides a sensible class of transfer functions and needs no other justification. In many examples, no question will arise of attempting to relate the discrete model to a supposed underlying continuous model because no underlying continuous series properly exists. However, in some cases, for example, where instantaneous observations are taken periodically on a chemical reactor, the discrete record can be used to tell us something about the continuous system. In particular, control engineers are used to thinking in terms

of the time constants and dead times of continuous systems and may best understand the results of the discrete model analysis when so expressed.

As before, we denote a continuous output and input at time t by $Y(t)$ and $X(t)$, respectively. Suppose that the output and input are related by the linear filtering operation

$$Y(t) = \int_0^\infty v(u)X(t-u)du$$

Suppose now that only discrete observations $(X_t, Y_t), (X_{t-1}, Y_{t-1}), \ldots$ of output and input are available at equispaced intervals of time $t, t-1, \ldots$ and that the discrete output and input are related by the discrete linear filter

$$Y_t = \sum_{j=0}^{\infty} v_j X_{t-j}$$

Then, for certain special cases, and with appropriate assumptions, useful relationships may be established between the discrete and continuous models.

11.3.1 Response to a Pulsed Input

A special case, which is of importance in the design of the discrete control schemes discussed in Part Four, arises when the opportunity for adjustment of the process occurs immediately after observation of the output, so that the input variable is allowed to remain at the same level between observations. The typical appearance of the resulting square wave, or *pulsed input* as we shall call it, is shown in Figure 11.8. We denote the fixed level at which the input is held during the period $t-1 < \tau < t$ by X_{t-1+}.

Consider a continuous linear system that has b whole periods of delay plus a fractional period c of further delay. Thus, in terms of previous notation, $b + c = \tau$. Then, we can represent the output from the system as

$$Y(t) = \int_0^\infty v(u)X(t-u)du$$

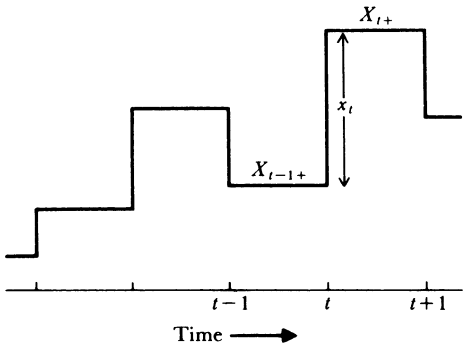


FIGURE 11.8 Example of a pulsed input.

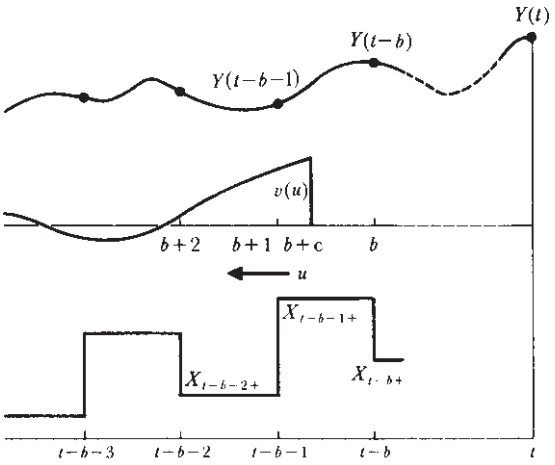


FIGURE 11.9 Transfer to output from a pulsed input.

where the impulse response function $v(u)$ is zero for $u < b + c$. Now for a pulsed input, as shown in Figure 11.9, the output at time t will be given exactly by

$$Y(t) = \left[\int_{b+c}^{b+1} v(u)du\right] X_{t-b-1+} + \left[\int_{b+1}^{b+2} v(u)du\right] X_{t-b-2+} + \cdots$$

Thus,

$$Y(t) = Y_t = v_b X_{t-b-1+} + v_{b+1} X_{t-b-2+} + \cdots$$

Therefore, for a *pulsed input*, there exists a discrete linear filter that is such that at times $t, t-1, t-2, \ldots$, the continuous output $Y(t)$ *exactly* equals the discrete output.

Given a pulsed input, consider the output Y_t from a discrete model

$$\xi(\nabla)Y_t = \eta(\nabla)X_{t-b-1+} \tag{11.3.1}$$

of order (r, r) in relation to the continuous output from the Rth-order model

$$(1 + \Xi_1 D + \Xi_2 D^2 + \cdots + \Xi_R D^R)Y(t) = X(t - b - c) \tag{11.3.2}$$

subject to the same input. It is shown in Appendix A11.1 that for suitably chosen values of the parameters $(\boldsymbol{\Xi}, c)$, the outputs will coincide exactly if $R = r$. Furthermore, if $c = 0$, the output from the continuous model (11.3.2) will be identical at the discrete times with that of a discrete model (11.3.1) of order $(r, r-1)$. We refer to the related continuous and discrete models as *discretely coincident* systems. If, then, a discrete model of the form (11.3.1) of order (r, r) has been obtained, then on the assumption *that the continuous model would be represented by the rth-order differential equation (11.3.2)*, the parameters, and in particular the time constants for the discretely coincident continuous system, may be written explicitly in terms of the parameters of the discrete model.

The parameter relationships for a delayed second-order system have been derived in Appendix A11.1. From these, the corresponding relationships for simpler systems may be obtained by setting appropriate constants equal to zero, as we shall now discuss.

11.3.2 Relationships for First- and Second-Order Coincident Systems

Undelayed First-Order System.

B Form. The continuous system satisfying

$$(1 + TD)Y(t) = gX(t) \tag{11.3.3}$$

is, for a pulsed input, discretely coincident with the discrete system satisfying

$$(1 - \delta B)Y_t = \omega_0 X_{t-1+} \tag{11.3.4}$$

where

$$\delta = e^{-1/T} \qquad T = (-\ln \delta)^{-1} \qquad \omega_0 = g(1 - \delta) \tag{11.3.5}$$

∇ ***Form***. Alternatively, the difference equation may be written

$$(1 + \xi\nabla)Y_t = gX_{t-1+} \tag{11.3.6}$$

where

$$\xi = \frac{\delta}{1 - \delta} \tag{11.3.7}$$

To illustrate, we reconsider the example of Section 11.2.4 for the ''general'' input. The output for this case is calculated in Table 11.3(c) and plotted in Figure 11.7(c). Suppose that, in fact, we had a continuous system:

$$(1 + 1.44D)Y(t) = 5X(t)$$

Then this would be discretely coincident with the discrete model (11.2.14) actually considered, namely,

$$(1 - 0.5B)Y_t = 2.5X_{t-1+}$$

If the input and output were continuous and the input were pulsed, the actual course of the response would be that shown by the continuous lines in Figure 11.10. The output would in fact follow a series of exponential curves. Each dashed line shows the further course that the response would take if no further change in the input were made. The curves correspond exactly at the discrete sample points with the discrete output already calculated in Table 11.3(c) and plotted in Figure 11.7(c).

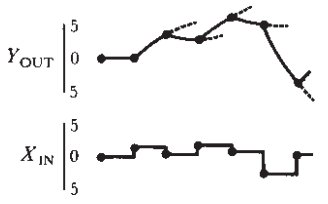


FIGURE 11.10 Continuous response of the system $(1 + 1.44D)Y(t) = 5X(t)$ to a pulsed input.

Delayed First-Order System.

B Form. The continuous system satisfying

$$(1 + TD)Y(t) = gX(t - b - c) \tag{11.3.8}$$

is, for a pulsed input, discretely coincident with the discrete system satisfying

$$(1 - \delta B)Y_t = (\omega_0 - \omega_1 B)X_{t-b-1+} \tag{11.3.9}$$

where

$$\delta = e^{-1/T} \qquad \omega_0 = g(1 - \delta^{1-c}) \qquad \omega_1 = g(\delta - \delta^{1-c}) \tag{11.3.10}$$

∇ ***Form***. Alternatively, the difference equation may be written

$$(1 + \xi\nabla)Y_t = g(1 + \eta\nabla)X_{t-b-1+} \tag{11.3.11}$$

where

$$\xi = \frac{\delta}{1 - \delta} \qquad -\eta = \frac{\delta(\delta^{-c} - 1)}{1 - \delta} \tag{11.3.12}$$

Now

$$(1 + \eta\nabla)X_{t-b-1+} = (1 + \eta)X_{t-b-1+} - \eta X_{t-b-2+} \tag{11.3.13}$$

can be regarded as an interpolation at an increment $(-\eta)$ between X_{t-b-1+} and X_{t-b-2+}. Table 11.4 allows the corresponding parameters $(\xi, -\eta)$ and (T, c) of the discrete and continuous models to be determined for a range of alternatives.

Undelayed Second-Order System.

B Form. The continuous system satisfying

$$(1 + T_1 D)(1 + T_2 D)Y(t) = gX(t) \tag{11.3.14}$$

TABLE 11.4 Values of $-\eta$ for Various Values of T and c for a First-Order System with Delay; Corresponding Values of ξ and δ

			$-\eta$ for				
δ	ξ	T	$c = 0.9$	$c = 0.7$	$c = 0.5$	$c = 0.3$	$c = 0.1$
0.9	9.00	9.49	0.90	0.69	0.49	0.29	0.10
0.8	4.00	4.48	0.89	0.68	0.47	0.28	0.09
0.7	2.33	2.80	0.88	0.66	0.46	0.26	0.09
0.6	1.50	1.95	0.88	0.64	0.44	0.25	0.08
0.5	1.00	1.44	0.87	0.62	0.41	0.23	0.07
0.4	0.67	1.09	0.85	0.60	0.39	0.21	0.06
0.3	0.43	0.83	0.84	0.57	0.35	0.19	0.05
0.2	0.25	0.62	0.82	0.52	0.31	0.15	0.04
01	0.11	0.43	0.77	0.45	0.24	0.11	0.03

is, for a pulsed input, discretely coincident with the system

$$(1 - \delta_1 B - \delta_2 B^2)Y_t = (\omega_0 - \omega_1 B)X_{t-1+} \tag{11.3.15}$$

or equivalently, with the system

$$(1 - S_1 B)(1 - S_2 B)Y_t = (\omega_0 - \omega_1 B)X_{t-1+} \tag{11.3.16}$$

where

$$\begin{aligned} S_1 &= e^{-1/T_1} \qquad S_2 = e^{-1/T_2} \\ \omega_0 &= g(T_1 - T_2)^{-1}[T_1(1 - S_1) - T_2(1 - S_2)] \\ \omega_1 &= g(T_1 - T_2)^{-1}[T_1 S_2(1 - S_1) - T_2 S_1(1 - S_2)] \end{aligned} \tag{11.3.17}$$

∇ *Form*. Alternatively, the difference equation may be written

$$(1 + \xi_1 \nabla + \xi_2 \nabla^2)Y_t = g(1 + \eta_1 \nabla)X_{t-1+} \tag{11.3.18}$$

where

$$-\eta_1 = (1 - S_1)^{-1}(1 - S_2)^{-1}(T_1 - T_2)^{-1}[T_2 S_1(1 - S_2) - T_1 S_2(1 - S_1)] \tag{11.3.19}$$

may be regarded as the increment of an interpolation between X_{t-1+} and X_{t-2+}. Values for ξ_1 and ξ_2 in terms of the δ's can be obtained directly using the results given in Table 11.1.

As a specific example, Figure 11.5 shows the step response for two discrete systems we have considered before, together with the corresponding continuous responses from the discretely coincident systems.

The pair of models are, for curve C,

$$\begin{aligned} &\text{Continuous}: \quad (1 + 1.41D + 2D^2)Y(t) = 5X(t) \\ &\text{Discrete}: \quad (1 - 1.15B + 0.49B^2)Y_t = 5(0.19 + 0.15B)X_{t-1+} \end{aligned}$$

and for curve R,

$$\begin{aligned} &\text{Continuous}: \quad (1 + 2D)(1 + D)Y(t) = 5X(t) \\ &\text{Discrete}: \quad (1 - 0.97B + 0.22B^2)Y_t = 5(0.15 + 0.09B)X_{t-1+} \end{aligned}$$

The continuous curves were drawn using (11.1.18) and (11.1.12), which give the continuous step responses for second-order systems having, respectively, complex and real roots.

The discrete representation of the response of a second-order continuous system with delay to a pulsed input is given in Appendix A11.1.

11.3.3 Approximating General Continuous Models by Discrete Models

Perhaps we should emphasize once more that the discrete transfer function models do not need to be justified in terms of, or related to, continuous systems. They are of importance in their own right in allowing a discrete output to be calculated from a discrete input. However, in some instances, such relationships are of interest.

For continuous systems, the pulsed input arises of itself in control problems when the convenient way to operate is to make an observation on the output Y and then immediately

to make any adjustment that may be needed on the input variable X. Thus, the input variable stays at a fixed level between observations, and we have a pulsed input. The relationships established in the previous sections may then be applied immediately. In particular, these relationships indicate that with the notation we have used, the *undelayed* discrete system is represented by

$$\xi(\nabla)Y_t = \eta(\nabla)X_{t-1+}$$

in which the subscript $t-1+$ on X is one step behind the subscript t on Y.

Use of Discrete Models When Continuous Records Are Available. Even though we have a continuous record of input and output, it may be convenient to determine the dynamic characteristics of the system by discrete methods, as we describe in Chapter 12. Thus, if pairs of values are read off with a sufficiently short sampling interval, very little is lost by replacing the continuous record by the discrete one.

One way in which the discrete results may then be used to approximate the continuous transfer function is to treat the input as though it were pulsed, that is, to treat the input record as if the discrete input observed at time j extended from just after $j-\frac{1}{2}$ to $j+\frac{1}{2}$. Thus, $X(t) = X_j(j-\frac{1}{2} < t \le j+\frac{1}{2})$. We can then relate the discrete result to that of the continuous record by using the pulsed input equations with X_t replacing X_{t+} and with $b+c-\frac{1}{2}$ replacing $b+c$, that is, with one half a time period subtracted from the delay. The continuous record will normally be read at a sufficiently small sampling interval so that sudden changes do not occur between the sampled points. In this case, the approximation will be very close.

APPENDIX A11.1 CONTINUOUS MODELS WITH PULSED INPUTS

We showed in Section 11.3.1 (see also Fig. 11.9) that for a pulsed input, the output from any delayed continuous linear system

$$Y(t) = \int_0^\infty v(u)X(t-u)du$$

where $v(u) = 0, u < b+c$, exactly given at the discrete times $t, t-1, t-2, \ldots$ by the discrete linear filter

$$Y_t = v(B)X_{t-1+}$$

where the weights $v_0, v_1, \ldots, v_{b-1}$ are zero and the weights $v_b, v_{b+1}, \ldots$ are given by

$$v_b = \int_{b+c}^{b+1} v(u)du \tag{A11.1.1}$$

$$v_{b+j} = \int_{b+j}^{b+j+1} v(u)du \qquad j \ge 1 \tag{A11.1.2}$$

Now suppose that the dynamics of the continuous system is represented by the Rth-order linear differential equation

$$\Xi(D)Y(t) = gX(t-b-c) \tag{A11.1.3}$$

which may be written in the form

$$\prod_{h=1}^{R}(1+T_hD)Y(t) = gX(t-b-c)$$

where $T_1, T_2, \ldots, T_R$ may be real or complex. We now show that for a pulsed input, the output from this continuous system is discretely coincident with that from a discrete difference equation model of order (r, r), or of order $(r, r-1)$ if $c = 0$. Now $v(u)$ is zero for $u < b + c$ and for $u \geq b + c$ is in general nonzero and satisfies the differential equation

$$\prod_{h=1}^{R}(1+T_hD)v(u-b-c) = 0 \qquad u \geq b+c$$

Thus,

$$v(u) = 0 \qquad u < b+c$$
$$v(u) = \alpha_1 e^{-(u-b-c)/T_1} + \alpha_2 e^{-(u-b-c)/T_2} + \cdots + \alpha_R e^{-(u-b-c)/T_R} \qquad u \geq b+c$$

Hence, using (A11.1.1) and (A11.1.2),

$$v_b = \sum_{h=1}^{R} \alpha_h T_h [1 - e^{-(1-c)/T_h}] \tag{A11.1.4}$$

$$v_{b+j} = \sum_{h=1}^{R} \alpha_h T_h (1 - e^{-1/T_h}) e^{c/T_h} e^{-j/T_h} \qquad j \geq 1 \tag{A11.1.5}$$

It will be noted that in the particular case when $c = 0$, the weights v_{b+j} are given by (A11.1.2) for $j = 0$ as well as for $j > 0$.

Now consider the difference equation model of order (r, s),

$$\delta(B)Y_t = \omega(B)B^b X_{t-1+} \tag{A11.1.6}$$

If we write

$$\Omega(B) = \omega(B)B^b$$

the discrete transfer function $v(B)$ for this model satisfies

$$\delta(B)v(B) = \Omega(B) \tag{A11.1.7}$$

As we have observed in (11.2.8), by equating coefficients in (A11.1.7) we obtain b zero weights $v_0, v_1, \ldots, v_{b-1}$, and if $s \geq r$, a further $s - r + 1$ values $v_b, v_{b+1}, \ldots, v_{b+s-r}$ which do not follow a pattern. The weights v_j eventually satisfy

$$\delta(B)v_j = 0 \qquad j > b+s \tag{A11.1.8}$$

with $v_{b+s}, v_{b+s-1}, \ldots, v_{b+s-r+1}$ supplying the required r starting values. Now write

$$\delta(B) = \prod_{h=1}^{r}(1 - S_h B)$$

where $S_1^{-1}, S_2^{-1}, \ldots, S_r^{-1}$ are the roots of the equation $\delta(B) = 0$. Then, the solution of (A11.1.8) is of the form

$$v_j = A_1(\boldsymbol{\omega})S_1^j + A_2(\boldsymbol{\omega})S_2^j + \cdots + A_r(\boldsymbol{\omega})S_r^j \qquad j > b + s - r \tag{A11.1.9}$$

where the coefficients $A_h(\boldsymbol{\omega})$ are suitably chosen so that the solutions of (A11.1.9) for $j = s - r + 1, s - r + 2, \ldots, s$ generate the starting values $v_{b+s-r+1}, \ldots, v_{b+s}$, and the notation $A_h(\boldsymbol{\omega})$ is used as a reminder that the A_h's are functions of $\omega_0, \omega_1, \ldots, \omega_s$. Thus, if we set $s = r$, for given parameters $(\boldsymbol{\omega}, \boldsymbol{\delta})$ in (A11.1.6), and hence for given parameters $(\boldsymbol{\omega}, \mathbf{S})$, there will be a corresponding set of values $A_h(\boldsymbol{\omega})$ $(h = 1, 2, \ldots, r)$ that produce the appropriate r starting values $v_{b+1}, v_{b+2}, \ldots, v_{b+r}$. Furthermore, we know that $v_b = \omega_0$. Thus,

$$v_b = \omega_0 \tag{A11.1.10}$$

$$v_{b+j} = \sum_{h=1}^{r} A_h(\boldsymbol{\omega})S_h^j \tag{A11.1.11}$$

and we can equate the values of the weights in (A11.1.4) and (A11.1.5), which come from the differential equation, to those in (A11.1.10) and (A11.1.11), which come from the difference equation. To do this, we must set

$$R = r \qquad S_h = e^{-1/T_h}$$

and the remaining $r + 1$ equations

$$\omega_0 = \sum_{h=1}^{r} \alpha_h T_h (1 - S_h^{1-c})$$

$$A_h(\boldsymbol{\omega}) = \alpha_h T_h (1 - S_h) S_h^{-c}$$

determine $c, \alpha_1, \alpha_2, \ldots, \alpha_r$ in terms of the S_h's and ω_j's.

When $c = 0$, we set $s = r - 1$, and for given parameters $(\boldsymbol{\omega}, \mathbf{S})$ in the difference equation, there will then be a set of r values $A_h(\boldsymbol{\omega})$ that are functions of $\omega_0, \omega_1, \ldots, \omega_{r-1}$, which produce the r starting values $v_b, v_{b+1}, \ldots, v_{b+r-1}$ and which can be equated to the values given by (A11.1.5) for $j = 0, 1, \ldots, r - 1$. To do this, we set

$$R = r \qquad S_h = e^{-1/T_h}$$

and the remaining r equations

$$A_h(\boldsymbol{\omega}) = \alpha_h T_h (1 - S_h)$$

determine $\alpha_1, \alpha_2, \ldots, \alpha_r$, in terms of the S_h's and ω_j's.

It follows, in general, that for a pulsed input the output at times $t, t - 1, \ldots$ from the continuous rth-order dynamic system defined by

$$\Xi(D)Y(t) = gX(t - b - c) \tag{A11.1.12}$$

is identical to the output from a discrete model

$$\xi(\nabla)Y_t = g\eta(\nabla)X_{t-b-1+} \tag{A11.1.13}$$

of order (r, r) with the parameters suitably chosen. Furthermore, if $c = 0$, the output from the continuous model (A11.1.12) is identical at the discrete times to that of a model (A11.1.13) of order $(r, r-1)$.

We now derive the discrete model corresponding to the second-order system with delay, from which the results given in Section 11.3.2 may be obtained as special cases.

Second-Order System with Delay. Suppose that the differential equation relating input and output for a continuous system is given by

$$(1 + T_1 D)(1 + T_2 D)Y(t) = gX(t - b - c) \tag{A11.1.14}$$

Then, the continuous impulse response function is

$$v(u) = g(T_1 - T_2)^{-1}(e^{-(u-b-c)/T_1} - e^{-(u-b-c)/T_2}) \qquad u > b + c \tag{A11.1.15}$$

For a pulsed input, the output at discrete times $t, t-1, t-2, \ldots$ will be related to the input by the difference equation

$$(1 + \xi_1\nabla + \xi_2\nabla^2)Y_t = g(1 + \eta_1\nabla + \eta_2\nabla^2)X_{t-b-1+} \tag{A11.1.16}$$

with suitably chosen values of the parameters. This difference equation can also be written

$$(1 - \delta_1 B - \delta_2 B^2)Y_t = (\omega_0 - \omega_1 B - \omega_2 B^2)X_{t-b-1+}$$

or

$$(1 - S_1 B)(1 - S_2 B)Y_t = (\omega_0 - \omega_1 B - \omega_2 B^2)X_{t-b-1+} \tag{A11.1.17}$$

Using (A11.1.1) and (A11.1.2) and writing

$$S_1 = e^{-1/T_1} \qquad S_2 = e^{-1/t_2}$$

we obtain

$$v_b = \int_{b+c}^{b+1} v(u)\,du = g(T_1 - T_2)^{-1}[T_1(1 - S_1^{1-c}) - T_2(1 - S_2^{1-c})]$$

$$v_{b+j} = \int_{b+j}^{b+j+1} v(u)\,du = g(T_1 - T_2)^{-1}[T_1 S_1^{-c}(1 - S_1)S_1^j - T_2 S_2^{-c}(1 - S_2)S_2^j] \qquad j \geq 1$$

Thus,

$$\begin{aligned}(T_1 - T_2)v(B) = {} & gB^b T_1[1 - S_1^{1-c} + S_1^{-c}(1 - S_1)(1 - S_1 B)^{-1}S_1 B] \\ & - gB^b T_2[1 - S_2^{1-c} + S_2^{-c}(1 - S_2)(1 - S_2 B)^{-1}S_2 B]\end{aligned}$$

But from (A11.1.17),

$$v(B) = \frac{B^b(\omega_0 - \omega_1 B - \omega_2 B^2)}{(1 - S_1 B)(1 - S_2 B)}$$

Hence, we obtain

$$\begin{aligned}\omega_0 &= g(T_1 - T_2)^{-1}[T_1(1 - S_1^{1-c}) - T_2(1 - S_2^{1-c})] \\ \omega_1 &= g(T_1 - T_2)^{-1}[(S_1 + S_2)(T_1 - T_2) + T_2 S_2^{1-c}(1 + S_1) - T_1 S_1^{1-c}(1 + S_2)] \\ \omega_2 &= g S_1 S_2 (T_1 - T_2)^{-1}[T_2(1 - S_2^{-c}) - T_1(1 - S_1^{-c})]\end{aligned} \tag{A11.1.18}$$

and

$$\delta_1 = S_1 + S_2 = e^{-1/T_1} + e^{-1/T_2} \qquad \delta_2 = -S_1 S_2 = -e^{-(1/T_1)-(1/T_2)} \tag{A11.1.19}$$

Complex Roots. If T_1 and T_2 are complex, corresponding expressions are obtained by substituting

$$T_1 = \zeta^{-1} e^{i\lambda} \qquad T_2 = \zeta^{-1} e^{-i\lambda} \qquad (i^2 = -1)$$

yielding

$$\begin{aligned}\omega_0 &= g\left\{1 - \frac{e^{-\zeta(1-c)\cos\lambda}\sin[\zeta(1-c)\sin\lambda + \lambda]}{\sin\lambda}\right\} \\ \omega_2 &= g\delta_2\left[1 - \frac{e^{\zeta c\cos\lambda}\sin(-\zeta c\sin\lambda + \lambda)}{\sin\lambda}\right] \\ \omega_1 &= \omega_0 - \omega_2 - (1 - \delta_1 - \delta_2)g\end{aligned} \tag{A11.1.20}$$

where

$$\begin{aligned}\delta_1 &= 2e^{-\zeta\cos\lambda}\cos(\zeta\sin\lambda) \\ \delta_2 &= -e^{-2\zeta\cos\lambda}\end{aligned} \tag{A11.1.21}$$

APPENDIX A11.2 NONLINEAR TRANSFER FUNCTIONS AND LINEARIZATION

The linearity (or additivity) of the transfer function models we have considered implies that the overall response to the sum of a number of individual inputs will be the sum of the individual responses to those inputs. Specifically, that if $Y_t^{(1)}$ is the response at time t to an input history $\{X_t^{(1)}\}$ and $\{Y_t^{(2)}\}$ is the response at time t to an input history $\{X_t^{(2)}\}$ the response at time t to an input history $\{X_t^{(1)} + X_t^{(2)}\}$ would be $Y_t^{(1)} + Y_t^{(2)}$, and similarly for continuous inputs and outputs. In particular, if the input level is multiplied by some constant, the output level is multiplied by this same constant. In practice, this assumption is probably never quite true, but it supplies a useful approximation for many practical situations.

Models for nonlinear systems may sometimes be obtained by allowing the parameters to depend upon the level of the input in some prescribed manner. For example, suppose that a system were being studied over a range where Y had a maximum η, and for any X

the steady-state relation could be approximated by the quadratic expression

$$Y_\infty = \eta - \frac{1}{2}k(\mu - X)^2$$

where Y and X are, as before, deviations from a convenient origin. Then,

$$g(X) = \frac{dY_\infty}{dX} = k(\mu - X)$$

and the dynamic behavior of the system might then be capable of representation by the first-order difference equation (11.3.4) but with variable gain proportional to $k(\mu - X)$. Thus,

$$Y_t = \delta Y_{t-1} + k(\mu - X_{t-1+})(1 - \delta)X_{t-1+} \tag{A11.2.1}$$

Dynamics of a Simple Chemical Reactor. It sometimes happens that we can make a theoretical analysis of a physical problem that will yield the appropriate form for the transfer function. In particular, this allows us to see very specifically what is involved in the linearized approximation.

As an example, suppose that a pure chemical A is continuously fed through a stirred tank reactor, and in the presence of a catalyst a certain proportion of it is changed to a product B, with no change of overall volume; hence the material continuously leaving the reactor consists of a mixture of B and unchanged A.

Suppose that initially the system is in equilibrium and that with quantities measured in suitable units:

1. μ is the rate at which A is fed to the reactor (and consequently is also the rate at which the mixture of A and B leaves the reactor).
2. η is the proportion of unchanged A at the outlet, so that $1 - \eta$ is the proportion of the product B at the outlet.
3. V is the volume of the reactor.
4. k is a constant determining the rate at which the product B is formed.

Suppose that the reaction is ''first order'' with respect to A, which means that the rate at which B is formed and A is used up is proportional to the amount of A present. Then, the rate of formation of B is $kV\eta$, but the rate at which B is leaving the outlet is $\mu(1 - \eta)$, and since the system is in equilibrium,

$$\mu(1 - \eta) = kV\eta \tag{A11.2.2}$$

Now, suppose that the equilibrium of the system is disturbed, the rate of feed to the reactor at time t being $\mu + X(t)$ and the corresponding concentration of A in the outlet being $\eta + Y(t)$. Now, the rate of chemical formation of B, which now equals $kV[\eta + Y(t)]$, will in general no longer exactly balance the rate at which B is flowing out of the system, which now equals $[\mu + X(t)][1 - \eta - Y(t)]$. The difference in these two quantities is the rate of increase in the amount of B within the reactor, which equals $-V[dY(t)/dt]$. Thus,

$$-V\frac{dY(t)}{dt} = kV[\eta + Y(t)] - [\mu + X(t)][1 - \eta - Y(t)] \tag{A11.2.3}$$

Using (A11.2.2) and rearranging, (A11.2.3) may be written

$$(kV + \mu + VD)Y(t) = X(t)[1 - \eta - Y(t)]$$

or

$$(1 + TD)Y(t) = (1 - \frac{Y(t)}{1 - \eta})X(t) \tag{A11.2.4}$$

where

$$T = \frac{V}{kV + \mu} \qquad g = \frac{1 - \eta}{kV + \mu} \tag{A11.2.5}$$

Now (A11.2.4) is a nonlinear differential equation, since it contains a term $X(t)$ multiplied by $Y(t)$. However, in some practical circumstances, it could be adequately approximated by a linear differential equation, as we now show.

Processes operate under a wide range of conditions, but certainly a not unusual situation might be one where $100(1 - \eta)$, the percentage conversion of feed A to product B was, say, 80%, and $100Y(t)$, the percentage fluctuation that was of practical interest, was 4%. In this case, the factor $1 - Y(t)/(1 - \eta)$ would vary from 0.95 to 1.05 and, to a good approximation, could be replaced by unity. The nonlinear differential equation (A11.2.4) could then be replaced by the linear first-order differential equation

$$(1 + TD)Y(t) = gX(t)$$

where T and g are as defined in Section 11.1.2. If the system was observed at discrete intervals of time, this equation could be approximated by a linear difference equation.

Situations can obviously occur when nonlinearities are of importance. This is particularly true of optimization studies, where the range of variation for the variables may be large. A device that is sometimes useful when the linear assumption is not adequate is to represent the dynamics by a set of linear models applicable over different ranges of the input variables. This approach could lead to nonlinear transfer function models similar in spirit to the threshold AR stochastic models considered in Section 10.3. However, for discrete systems it is often less clumsy to work directly with a nonlinear difference equation that can be ''solved'' recursively rather than analytically. For example, we might replace the nonlinear differential equation (A11.2.4) by the nonlinear difference equation

$$(1 + \xi_1 \nabla)Y_t = g(1 + \eta_{12} Y_{t-1})X_{t-1}$$

which has a form analogous to a particular case of the bilinear stochastic models discussed in Section 10.3.

EXERCISES

11.1. In the following transfer function models, X_t is the methane gas feed rate to a gas furnace, measured in cubic feet per minute, and Y_t the percent carbon dioxide in the outlet gas:

(1) $Y_t = 10 + \dfrac{25}{1 - 0.7B} X_{t-1}$

(2) $Y_t = 10 + \dfrac{22 - 12.5B}{1 - 0.85B} X_{t-2}$

(3) $Y_t = 10 + \dfrac{20 - 8.5B}{1 - 1.2B + 0.4B^2} X_{t-3}$

(a) Verify that the models are stable.

(b) Calculate the steady-state gain g, expressing it in the appropriate units.

11.2. For each of the models of Exercise 11.1, calculate from the difference equation and plot the responses to:

(a) A unit impulse $(0, 1, 0, 0, 0, 0, \ldots)$ applied at time $t = 0$

(b) A unit step $(0, 1, 1, 1, 1, 1, \ldots)$ applied at time $t = 0$

(c) A ramp input $(0, 1, 2, 3, 4, 5, \ldots)$ applied at time $t = 0$

(d) A periodic input $(0, 1, 0, -1, 0, 1, \ldots)$ applied at time $t = 0$

Estimate the period and damping factor of the step response to model (3).

11.3. Use equation (11.2.8) to obtain the impulse weights v_j for each of the models of Exercise 11.1, and check that they are the same as the impulse response obtained in Exercise 11.2(a).

11.4. Express the models of Exercise 11.1 in ∇ form.

11.5. **(a)** Calculate and plot the response of the two-input system

$$Y_t = 10 + \frac{6}{1 - 0.7B} X_{1,t-1} + \frac{8}{1 - 0.5B} X_{2,t-2}$$

to the orthogonal and randomized input sequences shown below.

t	X_{1t}	X_{2t}	t	X_{1t}	X_{2t}
0	0	0	5	1	−1
1	−1	1	6	1	1
2	1	−1	7	−1	−1
3	−1	−1	8	−1	1
4	1	1			

(b) Calculate the gains g_1 and g_2 of Y with respect to X_1 and X_2, respectively, and express the model in ∇ form.

12

IDENTIFICATION, FITTING, AND CHECKING OF TRANSFER FUNCTION MODELS

In Chapter 11, a parsimonious class of discrete linear transfer function models was introduced:

$$Y_t - \delta_1 Y_{t-1} - \cdots - \delta_r Y_{t-r} = \omega_0 X_{t-b} - \omega_1 X_{t-b-1} - \cdots - \omega_s X_{t-b-s}$$

or

$$Y_t = \delta^{-1}(B)\omega(B)X_{t-b}$$

where X_t and Y_t are deviations from equilibrium of the system input and output. In practice, the system will be infected by disturbances, or noise, whose net effect is to corrupt the output predicted by the transfer function model by an amount N_t. The combined transfer function–noise model may then be written as

$$Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + N_t$$

In this chapter, methods are described for identifying, fitting, and checking transfer function–noise models when simultaneous pairs of observations (X_1, Y_1), $(X_2, Y_2), \ldots, (X_N, Y_N)$ of the input and output are available at discrete equispaced times $1, 2, \ldots, N$.

Engineering methods for estimating transfer functions are usually based on the choice of special inputs to the system, for example, step and sine wave inputs (Young, 1955) and ''pulse'' inputs (Hougen, 1964). These methods have been useful when the system is affected by small amounts of noise but are less satisfactory otherwise. In the presence of

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

appreciable noise, it is necessary to use statistical methods for estimating the transfer function. Two previous approaches that have been tried for this problem are direct estimation of the impulse response in the time domain and direct estimation of the gain and phase characteristics in the frequency domain, as described, for example, by Briggs et al. (1965), Hutchinson and Shelton (1967), and Jenkins and Watts (1968). These methods are often unsatisfactory because they involve the estimation of too many parameters. For example, to determine the gain and phase characteristics, it is necessary to estimate two parameters at each frequency. The approach adopted in this chapter is to estimate the parameters in parsimonious difference equation models. Throughout most of the chapter we assume that the input X_t is itself a stochastic process. Models of the kind discussed are useful in representing and forecasting certain multiple time series.

12.1 CROSS-CORRELATION FUNCTION

In the same way that the autocorrelation function was used to identify stochastic models for univariate time series, the data analysis tool employed for the identification of transfer function models is the *cross-correlation function* between the input and output. In this section, we describe the basic properties of the cross-correlation function and in the next section show how it can be used to identify transfer function models.

12.1.1 Properties of the Cross-Covariance and Cross-Correlation Functions

Bivariate Stochastic Processes. We have seen in Chapter 2 that to analyze a time series, it is useful to regard it as a realization of a hypothetical population of time series called a stochastic process. Now, suppose that we want to describe an input time series X_t and the corresponding output time series Y_t from some physical system. For example, Figure 12.1 shows continuous data representing the (coded) input gas feed rate and corresponding output CO_2 concentration from a gas furnace. Then we can regard this pair of time series as realizations of a hypothetical population of pairs of time series, called *a bivariate stochastic process* (X_t, Y_t). We will assume that the data are read off at equispaced times yielding a pair of discrete time series, generated by a discrete bivariate process, and that values of the time series at times $t_0 + h, t_0 + 2h, \ldots, t_0 + Nh$ are denoted by $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$.

In this chapter, we will use the gas furnace data read at intervals of 9 seconds for illustration. The resulting time series (X_t, Y_t) consist of 296 observations and are listed as Series J in the Collection of Time Series section in Part Five. Further details about the data will be given in Section 12.2.2.

Cross-Covariance and Cross-Correlation Functions. We have seen in Chapter 2 that a stationary Gaussian stochastic process can be described by its mean μ and autocovariance function γ_k, or, equivalently, by its mean μ, variance σ^2, and autocorrelation function ρ_k. Moreover, since $\gamma_k = \gamma_{-k}$ and $\rho_k = \rho_{-k}$, the autocovariance and autocorrelation functions need to be considered only for nonnegative values of the lag $k = 0, 1, 2, \ldots$.

In general, a bivariate stochastic process (X_t, Y_t) need not be stationary. However, as in Chapter 4, we assume that the appropriately differenced process (x_t, y_t), where $x_t = \nabla^{d_x} X_t$ and $y_t = \nabla^{d_y} Y_t$, is stationary. The stationarity assumption implies in particular that the two processes x_t and y_t have constant means μ_x and μ_y and constant variances σ_x^2 and σ_y^2. If, in addition, it is assumed that the bivariate process is Gaussian, or normal, it is uniquely

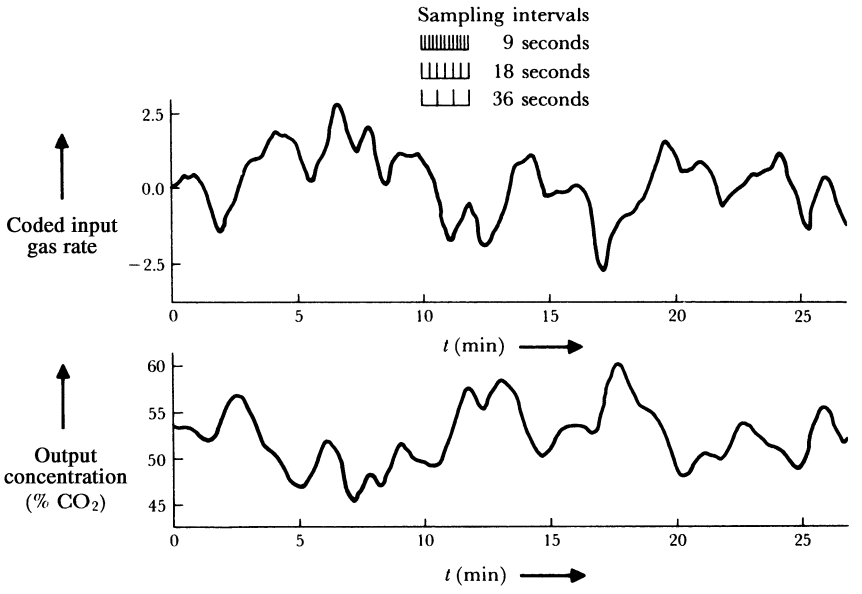


FIGURE 12.1 Input gas rate and output CO_2 concentration from a gas furnace.

characterized by its means μ_x and μ_y and its covariance matrix. Figure 12.2 shows the different kinds of covariances that need to be considered.

The autocovariance coefficients of each of the two series at lag k are defined by the usual formula:

$$\gamma_{xx}(k) = E[(x_t - \mu_x)(x_{t+k} - \mu_x)] = E[(x_t - \mu_x)(x_{t-k} - \mu_x)]$$
$$\gamma_{yy}(k) = E[(y_t - \mu_y)(y_{t+k} - \mu_y)] = E[(y_t - \mu_y)(y_{t-k} - \mu_y)]$$

where we now use the extended notation $\gamma_{xx}(k)$ and $\gamma_{yy}(k)$ for the autocovariances of the x_t and y_t series. The only other covariances that can appear in the covariance matrix are the *cross-covariance* coefficients between x_t and y_t series at lag $+k$:

$$\gamma_{xy}(k) = E[(x_t - \mu_x)(y_{t+k} - \mu_y)] \qquad k = 0, 1, 2, \ldots \tag{12.1.1}$$

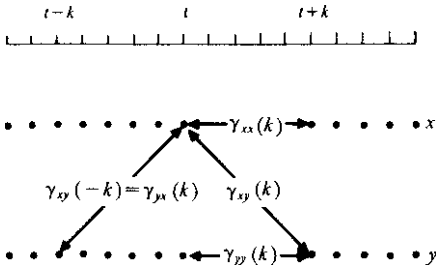


FIGURE 12.2 Autocovariances and cross-covariances of a bivariate stochastic process.

and the cross-covariance coefficients between the y_t and x_t series at lag $+k$:

$$\gamma_{yx}(k) = E[(y_t - \mu_y)(x_{t+k} - \mu_x)] \qquad k = 0, 1, 2, \ldots \tag{12.1.2}$$

Under (bivariate) stationarity, these cross-covariances must be the same for all t and hence are functions only of the lag k.

Note that, in general, $\gamma_{xy}(k)$ will not be the same as $\gamma_{yx}(k)$. However, since

$$\gamma_{xy}(k) = E[(x_{t-k} - \mu_x)(y_t - \mu_y)] = E[(y_t - \mu_y)(x_{t-k} - \mu_x)] = \gamma_{yx}(-k)$$

we need to define only one function $\gamma_{xy}(k)$ for $k = 0, \pm 1, \pm 2, \ldots$. The function $\gamma_{xy}(k) = \text{cov}[x_t, y_{t+k}]$, as defined in (12.1.1) for $k = 0, \pm 1, \pm 2, \ldots$, is called the *cross-covariance function* of the stationary bivariate process. Similarly, the correlation between x_t and y_{t+k}, which is the dimensionless quantity given by

$$\rho_{xy}(k) = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y} \qquad k = 0, \pm 1, \pm 2, \ldots \tag{12.1.3}$$

is called the *cross-correlation* coefficient at lag k, and the function $\rho_{xy}(k)$, defined for $k = 0, \pm 1, \pm 2, \ldots$, the *cross-correlation function* of the stationary bivariate process.

Since $\rho_{xy}(k)$ is not in general equal to $\rho_{xy}(-k)$, the cross-correlation function, in contrast to the autocorrelation function, is not symmetric about $k = 0$. In fact, it will sometimes happen that the cross-correlation function is zero over some range $-\infty$ to i or i to $+\infty$. For example, consider the cross-covariance function between the series a_t and z_t for the ''delayed'' first-order autoregressive process:

$$(1 - \phi B)\tilde{z}_t = a_{t-b} \qquad -1 < \phi < 1 \qquad b > 0$$

where a_t is white noise with zero mean and variance σ_a^2. Then since

$$\tilde{z}_{t+k} = a_{t+k-b} + \phi a_{t+k-b-1} + \phi^2 a_{t+k-b-2} + \cdots$$

the cross-covariance function between the series a_t and z_t is

$$\gamma_{az}(k) = E[a_t \tilde{z}_{t+k}] = \begin{cases} \phi^{k-b}\sigma_a^2 & k \geq b \\ 0 & k < b \end{cases}$$

Hence, for the delayed autoregressive process, the cross-correlation function is

$$\rho_{az}(k) = \begin{cases} \phi^{k-b}\dfrac{\sigma_a}{\sigma_z} = \phi^{k-b}(1 - \phi^2)^{1/2} & k \geq b \\ 0 & k < b \end{cases}$$

Figure 12.3 shows this cross-correlation function when $b = 2$ and $\phi = 0.6$.

12.1.2 Estimation of the Cross-Covariance and Cross-Correlation Functions

We assume that after differencing the original input and output time series d times, there are $n = N - d$ pairs of values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ available for analysis. Then it is shown, for example, in Jenkins and Watts (1968), that an estimate $c_{xy}(k)$ of the

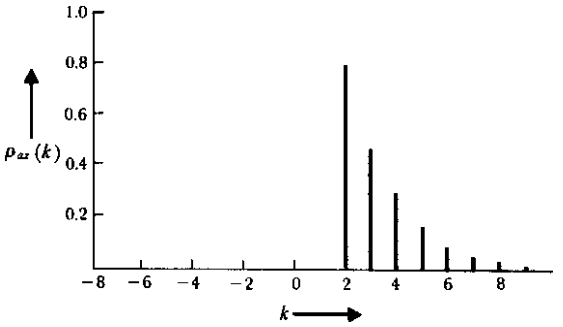


FIGURE 12.3 Cross-correlation function between a_t and z_t for delayed autoregressive process $\tilde{z}_t - 0.6\tilde{z}_{t-1} = a_{t-2}$.

cross-covariance coefficient at lag k is provided by

$$c_{xy}(k) = \begin{cases} \dfrac{1}{n}\displaystyle\sum_{t=1}^{n-k}(x_t - \bar{x})(y_{t+k} - \bar{y}) & k = 0, 1, 2, \ldots \\ \dfrac{1}{n}\displaystyle\sum_{t=1}^{n+k}(y_t - \bar{y})(x_{t-k} - \bar{x}) & k = 0, -1, -2, \ldots \end{cases} \tag{12.1.4}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the x_t series and y_t series, respectively. Similarly, the estimate $r_{xy}(k)$ of the cross-correlation coefficient $\rho_{xy}(k)$ at lag k may be obtained by substituting in (12.1.3) the estimates $c_{xy}(k)$ for $\gamma_{xy}(k)$, $s_x = \sqrt{c_{xx}(0)}$ for σ_x, and $s_y = \sqrt{c_{yy}(0)}$ for σ_y, yielding

$$r_{xy}(k) = \frac{c_{xy}(k)}{s_x s_y} \qquad k = 0, \pm 1, \pm 2, \ldots \tag{12.1.5}$$

The top graph in Figure 12.4 shows the estimated cross-correlation function $r_{xy}(k)$ between the input and output series for the discrete gas furnace data obtained by reading the continuous data of Figure 12.1 at intervals of 9 seconds. Note that the cross-correlation function is not symmetrical about zero and has a well-defined peak at $k = +5$, indicating that the output lags behind the input. The cross-correlations are negative. This is to be expected since an *increase* in the coded input produces a *decrease* in the output as seen from Figure 12.1. The autocorrelation functions of the input and output variables are also included in Figure 12.4. Both variables are highly autocorrelated and the slowly decaying patterns are indicative of an autoregressive dependence structure in these series.

Figure 12.4 can be reproduced in R as follows:

```
> gasfur = read.table('SeriesJ.txt',header=T)
> X = gasfur[,1]
> Y = gasfur[,2]
> CCF=ccf(Y,X)
> ACF.y=acf(Y)
> ACF.x=acf(X)
> par(mfrow=c(3,1))
> plot(CCF,ylab="CCF",main="Cross Correlation Between
```

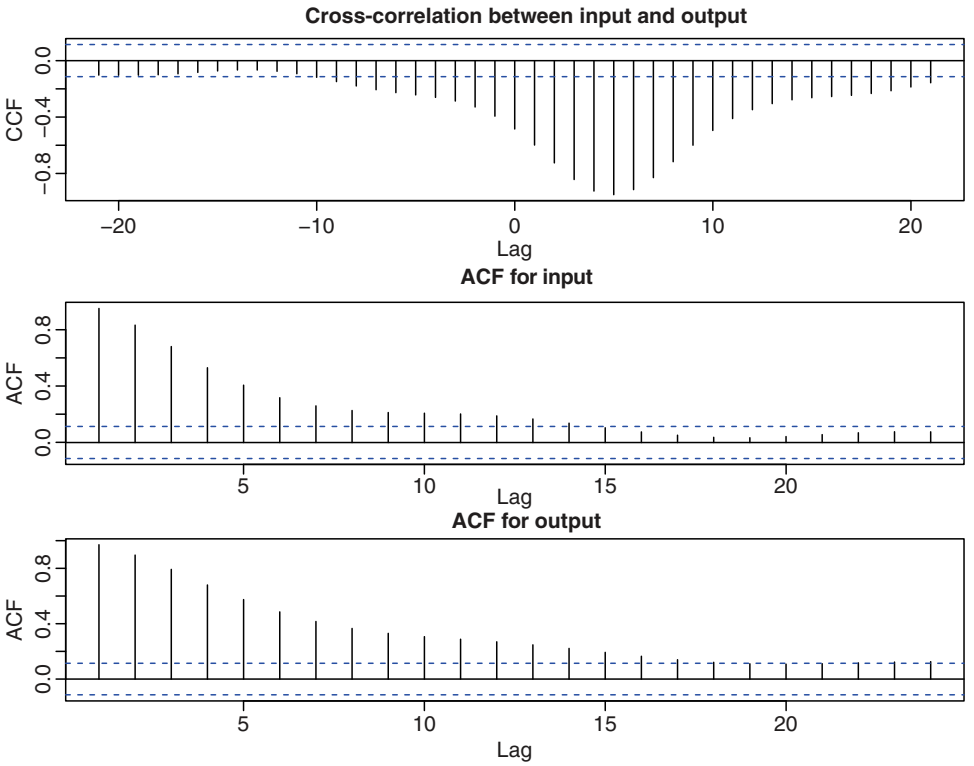


FIGURE 12.4 Estimated cross-correlation function between input and output for coded gas furnace data read at 9-second intervals along with the autocorrelation functions for the individual series.

```
    Input and Output")
> plot(ACF.x,main="ACF for Input")
> plot(ACF.y,main="ACF for Output")
```

12.1.3 Approximate Standard Errors of Cross-Correlation Estimates

A crude check as to whether certain values of the cross-correlation function $\rho_{xy}(k)$ could be effectively zero may be made by comparing the corresponding cross-correlation estimates with their approximate standard errors. Bartlett (1955) showed that the covariance between two cross-correlation estimates $r_{xy}(k)$ and $r_{xy}(k+l)$ is, on the normal assumption, and $k \geq 0$, given by

$$\begin{aligned}
&\operatorname{cov}[r_{xy}(k), r_{xy}(k+l)] \\
&\quad \simeq (n-k)^{-1} \sum_{\upsilon=-\infty}^{\infty} \{\rho_{xx}(\upsilon)\rho_{yy}(\upsilon+l) + \rho_{xy}(-\upsilon)\rho_{xy}(\upsilon+2k+l) \\
&\qquad\qquad + \rho_{xy}(k)\rho_{xy}(k+l)[\rho_{xy}^2(\upsilon) + \tfrac{1}{2}\rho_{xx}^2(\upsilon) + \tfrac{1}{2}\rho_{yy}^2(\upsilon)] \\
&\qquad\qquad - \rho_{xy}(k)[\rho_{xx}(\upsilon)\rho_{xy}(\upsilon+k+l) + \rho_{xy}(-\upsilon)\rho_{yy}(\upsilon+k+l)] \\
&\qquad\qquad - \rho_{xy}(k+l)[\rho_{xx}(\upsilon)\rho_{xy}(\upsilon+k) + \rho_{xy}(-\upsilon)\rho_{yy}(\upsilon+k)]\} \quad (12.1.6)
\end{aligned}$$

In particular, setting $l = 0$,

$$\begin{aligned}\operatorname{var}[r_{xy}(k)] &\simeq (n-k)^{-1}\sum_{v=-\infty}^{\infty}\{\rho_{xx}(v)\rho_{yy}(v)+\rho_{xy}(k+v)\rho_{xy}(k-v)\\ &\quad+\rho_{xy}^2(k)[\rho_{xy}^2(v)+\tfrac{1}{2}\rho_{xx}^2(v)+\tfrac{1}{2}\rho_{yy}^2(v)]\\ &\quad-2\rho_{xy}(k)[\rho_{xx}(v)\rho_{xy}(v+k)+\rho_{xy}(-v)\rho_{yy}(v+k)]\}\end{aligned} \tag{12.1.7}$$

Formulas that apply to important special cases can be derived from these general expressions. For example, if we assume that $x_t \equiv y_t$, it becomes appropriate to set

$$\rho_{xx}(v) = \rho_{yy}(v) = \rho_{xy}(v) = \rho_{xy}(-v)$$

On making this substitution in (12.1.6) and (12.1.7), we obtain an expression for the covariance between two autocorrelation estimates and, more particularly, the expression for the variance of an autocorrelation estimate given earlier in (2.1.13).

It is often the case that two processes are appreciably cross-correlated only over some rather narrow range of lags. Suppose it is postulated that $\rho_{xy}(v)$ is nonzero *only* over some range $Q_1 \le v \le Q_2$. Then,

1. If neither $k, k+l$, nor $k+\frac{1}{2}l$ are included in this range, all terms in (12.1.6) except the first are zero, and

$$\operatorname{cov}[r_{xy}(k), r_{xy}(k+l)] \simeq (n-k)^{-1}\sum_{v=-\infty}^{\infty}\rho_{xx}(v)\rho_{yy}(v+l) \tag{12.1.8}$$

2. If k is not included in this range, then in a similar way (12.1.7) reduces to

$$\operatorname{var}[r_{xy}(k)] \simeq (n-k)^{-1}\sum_{v=-\infty}^{\infty}\rho_{xx}(v)\rho_{yy}(v) \tag{12.1.9}$$

In particular, on the hypothesis that the two processes have *no cross-correlation*, that is, cross-correlations are zero for all lags, it follows that the simple formulas (12.1.8) and (12.1.9) apply for *all* lags k and $k+l$.

Another special case of some interest occurs when two processes are *not cross-correlated and one is white noise*. Suppose that $x_t = a_t$ is generated by a white noise process but y_t is autocorrelated. Then from (12.1.8),

$$\operatorname{cov}[r_{ay}(k), r_{ay}(k+l)] \simeq (n-k)^{-1}\rho_{yy}(l) \tag{12.1.10}$$

$$\operatorname{var}[r_{ay}(k)] \simeq (n-k)^{-1} \tag{12.1.11}$$

Hence, it follows that

$$\rho[r_{ay}(k), r_{ay}(k+l)] \simeq \rho_{yy}(l) \tag{12.1.12}$$

Thus, in this case the cross-correlations have the *same* autocorrelation function as the process generating the output y_t. Thus, even though a_t and y_t are *not* cross-correlated, the sample cross-correlation function can be expected to vary about zero with standard deviation $(n-k)^{-1/2}$ *in a systematic pattern* typical of the behavior of the autocorrelation

function $\rho_{yy}(l)$. Finally, if two processes are *both* white noise and are not cross-correlated, the covariance between cross-correlation estimates at different lags will be zero.

12.2 IDENTIFICATION OF TRANSFER FUNCTION MODELS

We now show how to *identify* a combined transfer function–noise model

$$Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + N_t$$

for a linear system corrupted by noise N_t at the output and assumed to be generated by an ARIMA process that is statistically independent[1] of the input X_t. Specifically, the objective at this stage is to obtain some idea of the orders r and s of the denominator and numerator operators in the transfer function model and to derive initial guesses for the parameters $\boldsymbol{\delta}, \boldsymbol{\omega}$, and the delay parameter b. In addition, we aim to make initial guesses of the orders p, d, q of the ARIMA process describing the noise at the output and to obtain initial estimates of the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ in that model. The tentative transfer function and noise models so obtained can then be used as a starting point for more efficient estimation methods described in Section 12.3.

Outline of the Identification Procedure. Suppose that the transfer function model

$$Y_t = \upsilon(B)X_t + N_t \tag{12.2.1}$$

may be parsimoniously parameterized in the form

$$Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + N_t \tag{12.2.2}$$

where $\delta(B) = 1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r$ and $\omega(B) = \omega_0 - \omega_1 B - \omega_2 B^2 - \cdots - \omega_s B^s$. The identification procedure is as follows:

1. Derive rough estimates $\hat{\upsilon}_j$ of the impulse response weights υ_j in (12.2.1).
2. Use the estimates $\hat{\upsilon}_j$ so obtained to make guesses of the orders r and s of the denominator and numerator operators in (12.2.2) and of the delay parameter b.
3. Substitute the estimates $\hat{\upsilon}_j$ in equations (11.2.8) with values of r, s, and b obtained from step 2 to obtain initial estimates of the parameters $\boldsymbol{\delta}$ and $\boldsymbol{\omega}$ in (12.2.2).

Knowing the $\hat{\upsilon}_j$, values of b, r, and s may be guessed using the following facts established in Section 11.2.2. For a model of the form of (12.2.2), the impulse response weights υ_j consist of:

1. b zero values $\upsilon_0, \upsilon_1, \ldots, \upsilon_{b-1}$.
2. A further $s - r + 1$ values $\upsilon_b, \upsilon_{b+1}, \ldots, \upsilon_{b+s-r}$ following no fixed pattern (no such values occur if $s < r$).

[1]When the input is at our choice, we can guarantee that it is independent of N_t by *generating* X_t according to some random process.

3. Values v_j with $j \geq b + s - r + 1$ that follow the pattern dictated by an rth-order difference equation that has r starting values $v_{b+s}, \ldots, v_{b+s-r+1}$. Starting values v_j for $j < b$ will, of course, be zero.

Differencing of the Input and Output. The basic tool that is employed here in the identification procedure is the cross-correlation function between input and output. When the processes are nonstationary, it is assumed that stationarity can be induced by suitable differencing. Nonstationary behavior is suspected if the estimated auto- and cross-correlation functions of the (X_t, Y_t) series fail to damp out quickly. We assume that a degree of differencing[2] d necessary to induce stationarity has been achieved when the estimated auto- and cross-correlations $r_{xx}(k)$, $r_{yy}(k)$, and $r_{xy}(k)$ of $x_t = \nabla^d X_t$ and $y_t = \nabla^d Y_t$ damp out quickly. In practice, d is usually 0, 1, or 2.

Identification of the Impulse Response Function Without Prewhitening. Suppose that after differencing d times, the model (12.2.1) can be written in the form

$$y_t = v_0 x_t + v_1 x_{t-1} + v_2 x_{t-2} + \cdots + n_t \tag{12.2.3}$$

where $y_t = \nabla^d Y_t$, $x_t = \nabla^d X_t$, and $n_t = \nabla^d N_t$ are stationary processes with zero means. Then, on multiplying throughout in (12.2.3) by x_{t-k} for $k \geq 0$, we obtain

$$x_{t-k} y_t = v_0 x_{t-k} x_t + v_1 x_{t-k} x_{t-1} + \cdots + x_{t-k} n_t \tag{12.2.4}$$

If we make the further assumption that x_{t-k} is uncorrelated with n_t for all k, taking expectations in (12.2.4) yields the set of equations

$$\gamma_{xy}(k) = v_0 \gamma_{xx}(k) + v_1 \gamma_{xx}(k-1) + \cdots \qquad k = 0, 1, 2, \ldots \tag{12.2.5}$$

Suppose that the weights v_j are effectively zero beyond $k = K$. Then the first $K + 1$ of the equations (12.2.5) can be written as

$$\boldsymbol{\gamma}_{xy} = \boldsymbol{\Gamma}_{xx} \mathbf{v} \tag{12.2.6}$$

where

$$\boldsymbol{\gamma}_{xy} = \begin{bmatrix} \gamma_{xy}(0) \\ \gamma_{xy}(1) \\ \vdots \\ \gamma_{xy}(K) \end{bmatrix} \qquad \mathbf{v} = \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_K \end{bmatrix}$$

$$\boldsymbol{\Gamma}_{xx} = \begin{bmatrix} \gamma_{xx}(0) & \gamma_{xx}(1) & \cdots & \gamma_{xx}(K) \\ \gamma_{xx}(1) & \gamma_{xx}(0) & \cdots & \gamma_{xx}(K-1) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{xx}(K) & \gamma_{xx}(K-1) & \cdots & \gamma_{xx}(0) \end{bmatrix}$$

[2]The procedures outlined can equally well be used when different degrees of differencing are employed for input and output.

Substituting estimates $c_{xx}(k)$ of the autocovariance function of the input x_t and estimates $c_{xy}(k)$ of the cross-covariance function between the input x_t and output y_t, (12.2.6) provides $K+1$ linear equations for the first $K+1$ weights. However, these equations, which do not in general provide efficient estimates, are cumbersome to solve for large K and in any case require knowledge of the point K beyond which the v_j are effectively zero. The sample version of equations (12.2.6) represents essentially, apart from ''end effects,'' the least-squares normal equations from linear regression of y_t on $x_t, x_{t-1}, \ldots, x_{t-K}$, in which it is assumed, implicitly, that the noise n_t in (12.2.3) is not autocorrelated. This is one source of the inefficiency in this identification method, which may be called the *regression method*. To improve the efficiency of this method, Liu and Hanssens (1982) (see also Pankratz (1991, Chapter 5)) suggest performing generalized least-squares estimation of the regression equation $y_t = v_0 x_t + v_1 x_{t-1} + \cdots + v_K x_{t-K} + n_t$ assuming the noise n_t follows some autocorrelated time series ARMA model. They also discuss generalization of this method of identification of impulse response functions to the case with multiple input processes $X_{1,t}, X_{2,t}, \ldots, X_{m,t}$ in the model, that is, $Y_t = v_1(B)X_{1,t} + \cdots + v_m(B)X_{m,t} + N_t$.

12.2.1 Identification of Transfer Function Models by Prewhitening the Input

Considerable simplification in the identification process would occur if the input to the system were white noise. Indeed, as discussed in more detail in Section 12.5, when the choice of the input is at our disposal, there is much to recommend such an input. When the original input follows some other stochastic process, simplification is possible by *prewhitening*.

Suppose that the suitably differenced input process x_t is stationary and is capable of representation by some member of the general linear class of autoregressive–moving average models. Then, given a set of data, we can carry out our usual identification and estimation methods to obtain a model for the x_t process:

$$\theta_x^{-1}(B)\phi_x(B)x_t = \alpha_t \tag{12.2.7}$$

which, to a close approximation, transforms the correlated input series x_t to the uncorrelated white noise series α_t. At the same time, we can obtain an estimate s_α^2 of σ_α^2 from the sum of squares of the $\hat{\alpha}_t$'s. If we now apply this *same* transformation to y_t to obtain

$$\beta_t = \theta_x^{-1}(B)\phi_x(B)y_t$$

then the model (12.2.3) may be written as

$$\beta_t = v(B)\alpha_t + \varepsilon_t \tag{12.2.8}$$

where ε_t is the transformed noise series defined by

$$\varepsilon_t = \theta_x^{-1}(B)\phi_x(B)\eta_t \tag{12.2.9}$$

On multiplying (12.2.8) on both sides by α_{t-k} and taking expectations, we obtain

$$\gamma_{\alpha\beta}(k) = v_k\sigma_\alpha^2 \tag{12.2.10}$$

where $\gamma_{\alpha\beta}(k) = E[\alpha_{t-k}\beta_t]$ is the cross-covariance at lag $+k$ between the series α_t and β_t. Thus,

$$v_k = \frac{\gamma_{\alpha\beta}(k)}{\sigma_\alpha^2}$$

or, in terms of the cross-correlations,

$$v_k = \frac{\rho_{\alpha\beta}(k)\sigma_\beta}{\sigma_\alpha} \qquad k = 0, 1, 2, \ldots \tag{12.2.11}$$

Hence, after prewhitening the input, the cross-correlation function between the prewhitened input and correspondingly transformed output is directly proportional to the impulse response function. We note that the effect of prewhitening is to convert the nonorthogonal set of equation (12.2.6) into the orthogonal set (12.2.10).

In practice, we do not know the theoretical cross-correlation function $\rho_{\alpha\beta}(k)$, so we must substitute estimates in (12.2.11) to give

$$\hat{v}_k = \frac{r_{\alpha\beta}(k)s_\beta}{s_\alpha} \qquad k = 0, 1, 2, \ldots \tag{12.2.12}$$

The preliminary estimates $\hat{v}_k$ so obtained are again, in general, statistically inefficient but can provide a rough basis for selecting suitable operators $\delta(B)$ and $\omega(B)$ in the transfer function model. An additional feature of the prewhitening method is that because the prewhitened input series α_t is *white noise*, so that $\rho_{\alpha\alpha}(k) = 0$ for all $k \neq 0$, there are considerable simplifications in formulas (12.1.7) and (12.1.9) for $\text{var}[r_{\alpha\beta}(k)]$. In particular, on the assumption that the series α_t and β_t are not cross correlated, the result (12.1.11) applies to give simply $\text{var}[r_{\alpha\beta}(k)] \simeq (n-k)^{-1}$. We now illustrate this identification and preliminary estimation procedure with an actual example.

12.2.2 Example of the Identification of a Transfer Function Model

In an investigation on adaptive optimization (Kotnour et al., 1966), a gas furnace was employed in which air and methane combined to form a mixture of gases containing CO_2 (carbon dioxide). The air feed was kept constant, but the methane feed rate could be varied in any desired manner, and the resulting CO_2 concentration in the off-gases measured. The continuous data of Figure 12.1 were collected to provide information about the dynamics of the system over a region of interest where it was known that an approximately linear steady-state relationship applied. The continuous stochastic input series $X(t)$ shown in the top half of Figure 12.1 was generated by passing white noise through a linear filter. The process had mean zero and, during the realization that was used for this experiment, varied from -2.5 to $+2.5$. It was desired that the actual methane gas feed rate should cover a range from 0.5 to 0.7 ft^3/min. To ensure this, the input gas feed rate was caused to follow the process:

$$\text{Methane gas input feed} = 0.60 - 0.04X(t)$$

For simplicity, we will work throughout with the ''coded'' input $X(t)$. The final transfer function expressed in terms of the actual feed rate is readily obtained by substitution. Series J in the Collection of Time Series section in Part Five shows 296 successive pairs

TABLE 12.1 Estimated Cross-Correlation Function After Prewhitening and Approximate Impulse Response Function for Gas Furnace Data

k	$r_{\alpha\beta}(k)$	$\hat{\sigma}(r)$	$\hat{v}_k$	$r_{\beta\beta}(k)$	k	$r_{\alpha\beta}(k)$	$\hat{\sigma}(r)$	$\hat{v}_k$	$r_{\beta\beta}(k)$
0	−0.00	0.06	−0.02	1.00	6	−0.27	0.06	−0.52	0.12
1	0.05	0.06	0.10	0.23	7	−0.17	0.06	−0.32	0.05
2	−0.03	0.06	−0.06	0.36	8	−0.03	0.06	−0.06	0.09
3	−0.29	0.05	−0.53	0.13	9	0.03	0.06	0.06	0.01
4	−0.34	0.06	−0.63	0.08	10	−0.06	0.06	−0.10	0.10
5	−0.46	0.05	−0.88	0.01					

of observations (X_t, Y_t) read off from the continuous records at 9-second intervals. In this particular experiment, the nature of the input disturbance was known because it was deliberately induced. However, we proceed as if it were not known. As shown in Figure 12.4, the estimated auto- and cross-correlation functions of X_t and Y_t damp out fairly quickly, confirming that no differencing is necessary. The usual model identification and fitting procedures applied to the input series X_t indicate that it is well described by a third-order autoregressive process

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)X_t = \alpha_t$$

with $\hat{\phi}_1 = 1.97$, $\hat{\phi}_2 = -1.37$, $\hat{\phi}_3 = 0.34$, and $s_\alpha^2 = 0.0353$. Hence, the transformations

$$\alpha_t = (1 - 1.97B + 1.37B^2 - 0.34B^3)X_t$$
$$\beta_t = (1 - 1.97B + 1.37B^2 - 0.34B^3)Y_t$$

are applied to the input and output series to yield the series α_t and β_t with $s_\alpha = 0.188$ and $s_\beta = 0.358$. The estimated cross-correlation function between α_t and β_t is listed in Table 12.1 and plotted in Figure 12.5. Table 12.1 also includes the estimate (12.2.12) of the impulse response function,

$$\hat{v}_k = \frac{0.358}{0.188} r_{\alpha\beta}(k)$$

The approximate standard errors $\hat{\sigma}(r)$ for the estimated cross-correlations $r_{\alpha\beta}(k)$ shown in Table 12.1 are the square roots of the variances obtained from expression (12.1.7):

1. With cross-correlations up to lag +2 and from lag +8 onward assumed equal to zero
2. With autocorrelations $\rho_{\alpha\alpha}(k)$ assumed zero for $k > 0$
3. With autocorrelations $\rho_{\beta\beta}(k)$ assumed zero for $k > 4$
4. With estimated correlations $r_{\alpha\beta}(k)$ and $r_{\beta\beta}(k)$ from Table 12.1 replacing theoretical values.

For this example, the standard errors $\hat{\sigma}(r)$ differ very little from the approximate values $(n-k)^{-1/2}$, or as a further approximation $n^{-1/2} = 0.06$, appropriate under the hypothesis that the series are uncorrelated. The estimated cross-correlations along with the approximate two standard error limits are plotted in Figure 12.5.

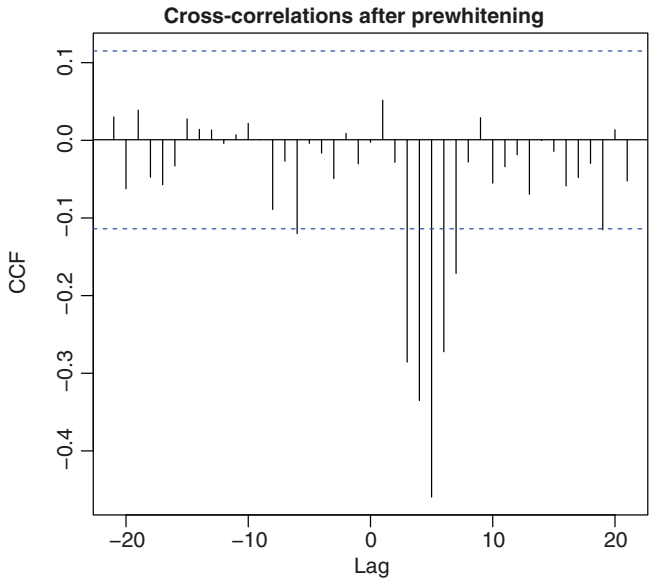


FIGURE 12.5 Estimated cross-correlation function for coded gas furnace data after prewhitening.

The values $\hat{v}_0, \hat{v}_1$, and $\hat{v}_2$ are small compared with their standard errors, suggesting that $b = 3$ (that there are two whole periods of delay). Using the results of Section 12.1.1, the subsequent pattern of the $\hat{v}$'s might be accounted for by a model with (r, s, b) equal to either (1, 2, 3) or (2, 2, 3). The first model would imply that v_3 and v_4 were preliminary values following no fixed pattern and that v_5 provided the starting value for an exponential decay determined by the difference equation $v_j - \delta v_{j-1} = 0, j > 5$. The second model would imply that v_3 was a single preliminary value and that v_4 and v_5 provided the starting values for a pattern of double exponential decay or damped sinusoidal decay determined by the difference equation $v_j - \delta_1 v_{j-1} - \delta_2 v_{j-2} = 0, j > 5$. Thus, the preliminary identification suggests a transfer function model

$$(1 - \delta_1 B - \delta_2 B^2) Y_t = (\omega_0 - \omega_1 B - \omega_2 B^2) X_{t-b} \tag{12.2.13}$$

or some simplification of it, probably with $b = 3$.

Calculations in R. The prewhitening, the calculation of $r_{\alpha\beta}(k)$, $\hat{v}_k$, and $r_{\beta\beta}(k)$ in Table 12.1, and the creation of Figure 12.5 can be performed using the R code provided below. Note, however, that the results from R differ very slightly from those shown in Table 12.1, possibly due to round-off and differences in the treatment of initial values in the series.

```
> mm1=arima(X,order=c(3,0,0))
> mm1  % Prints the AR(3) coefficients for X

   Call:  arima(x = X, order = c(3, 0, 0))
   Coefficients:
         ar1      ar2     ar3  intercept
      1.9691  -1.3651  0.3394    -0.0606
```

```
   s.e.  0.0544   0.0985  0.0543     0.1898
   sigma^2 estimated as 0.0353:log likelihood=72.6,aic=-135.1

> f1=c(1,-mm1$coef[1:3])        % Creates a filter to transform Y
> f1                    ar1         ar2        ar3
        1.000     -1.9691     1.3651    -0.3394

> Yf=filter(Y,f1,method=c("convolution"),sides=1)
> yprev=Yf[4:296]               % transformed Y
> xprev=mm1$residuals[4:296]    % transformed X
> CCF=ccf(yprev,xprev)          % computes the cross-correlations
> CCF                           % retrieves the cross-correlations
> vk=(sd(yprev)/sd(xprev))*CCF$acf    % impulse response function
> ACF=acf(yprev)                % autocorrelations of transformed Y
> plot(CCF,ylab='CCF',main='Cross-correlations after prewhitening')
```

Preliminary Estimates. Assuming the model (12.2.13) with $b = 3$, the equations (11.2.8) for the impulse response function are

$$\begin{aligned} v_j &= 0 \qquad j < 3 \\ v_3 &= \omega_0 \\ v_4 &= \delta_1 v_3 - \omega_1 \\ v_5 &= \delta_1 v_4 + \delta_2 v_3 - \omega_2 \\ v_6 &= \delta_1 v_5 + \delta_2 v_4 \\ v_7 &= \delta_1 v_6 + \delta_2 v_5 \end{aligned} \tag{12.2.14}$$

Substituting the estimates $\hat{v}_k$ from Table 12.1 in the last two of these equations, we obtain

$$\begin{aligned} -0.88\hat{\delta}_1 - 0.63\hat{\delta}_2 &= -0.52 \\ -0.52\hat{\delta}_1 - 0.88\hat{\delta}_2 &= -0.32 \end{aligned}$$

which give preliminary estimates $\hat{\delta}_1 = 0.57$ and $\hat{\delta}_2 = 0.02$. If these values are now substituted in the second, third, and fourth of equations (12.2.14), we obtain

$$\begin{aligned} \hat{\omega}_0 &= \hat{v}_3 = -0.53 \\ \hat{\omega}_1 &= \hat{\delta}_1 \hat{v}_3 - \hat{v}_4 = (0.57)(-0.53) + 0.63 = 0.33 \\ \hat{\omega}_2 &= \hat{\delta}_1 \hat{v}_4 + \hat{\delta}_2 \hat{v}_3 - \hat{v}_5 = (0.57)(-0.63) + (0.02)(-0.53) + 0.88 = 0.51 \end{aligned}$$

Thus, the preliminary identification suggests a tentative transfer function model:

$$(1 - 0.57B - 0.02B^2)Y_t = -(0.53 + 0.33B + 0.51B^2)X_{t-3}$$

The estimates so obtained can be used as starting values for the more efficient iterative estimation methods, which will be described in Section 12.3. Note that the estimate $\hat{\delta}_2$ is very small and suggests that this parameter may be omitted, but we will retain it for the time being.

12.2.3 Identification of the Noise Model

Reverting to the general case, suppose that (where necessary, after suitable differencing) the model could be written as

$$y_t = v(B)x_t + n_t$$

where $n_t = \nabla^d N_t$. Given that a preliminary estimate $\hat{v}(B)$ of the transfer function has been obtained in the manner discussed in Section 12.2.2, an estimate of the noise series is provided by

$$\hat{n}_t = y_t - \hat{v}(B)x_t$$

that is,

$$\hat{n}_t = y_t - \hat{v}_0 x_t - \hat{v}_1 x_{t-1} - \hat{v}_2 x_{t-2} - \cdots$$

Alternatively, $\hat{v}(B)$ may be replaced by the tentative transfer function model estimate $\hat{\delta}^{-1}(B)\hat{\omega}(B)B^b$ determined by preliminary identification. Thus,

$$\hat{n}_t = y_t - \hat{\delta}^{-1}(B)\hat{\omega}(B)x_{t-b}$$

and $\hat{n}_t$ may be computed by first calculating $\hat{y}_t = \hat{\delta}^{-1}(B)\hat{\omega}(B)x_{t-b}$ recursively through $\hat{\delta}(B)\hat{y}_t = \hat{\omega}(B)x_{t-b}$ as

$$\hat{y}_t = \hat{\delta}_1\hat{y}_{t-1} + \cdots + \hat{\delta}_r\hat{y}_{t-r} + \hat{\omega}_0 x_{t-b} - \hat{\omega}_1 x_{t-b-1} - \cdots - \hat{\omega}_s x_{t-b-s} \tag{12.2.15}$$

and then computing the noise series from $\hat{n}_t = y_t - \hat{y}_t$. In either case, study of the estimated autocorrelation function and partial autocorrelation function of $\hat{n}_t$ can lead to identification of the noise model.

It is also possible to identify the noise using the correlation functions for the input and output, after prewhitening, in the following way. Suppose that the input could be exactly prewhitened to give

$$\beta_t = v(B)\alpha_t + \varepsilon_t \tag{12.2.16}$$

where the known relationship

$$\varepsilon_t = \theta_x^{-1}(B)\phi_x(B)n_t \tag{12.2.17}$$

would link ε_t and n_t. If a stochastic model could be found for ε_t, then, using (12.2.17), a model could be deduced for n_t and hence for N_t. If we now write $v(B)\alpha_t = u_t$, so that $\beta_t = u_t + \varepsilon_t$, and provided that our independence assumption concerning x_t and n_t, and hence concerning u_t and ε_t, is justified, we can write

$$\gamma_{\beta\beta}(k) = \gamma_{uu}(k) + \gamma_{\varepsilon\varepsilon}(k) \tag{12.2.18}$$

Since α_t is white noise, $\gamma_{uu}(k)$ may be obtained using the result (3.1.8), which gives the autocorrelation function of a linear process. Thus,

$$\gamma_{uu}(k) = \sigma_\alpha^2 \sum_{j=0}^{\infty} v_j v_{j+k} = \frac{1}{\sigma_\alpha^2} \sum_{j=0}^{\infty} \gamma_{\alpha\beta}(j)\gamma_{\alpha\beta}(j+k)$$

TABLE 12.2 Estimated Autocorrelation and Partial Autocorrelation Functions of the Noise in Gas Furnace Data

k	r_k	$\hat{\phi}_{kk}$	k	r_k	$\hat{\phi}_{kk}$
1	0.89	0.89	7	0.01	−0.02
2	0.71	−0.43	8	−0.03	0.01
3	0.51	−0.13	9	−0.05	−0.01
4	0.32	0.02	10	−0.04	0.08
5	0.17	0.04	11	−0.03	−0.06
6	0.07	−0.02	12	−0.03	−0.10

using (12.2.10). Hence, using (12.2.18), the autocovariances of ε_t may be obtained from $\gamma_{\varepsilon\varepsilon}(k) = \gamma_{\beta\beta}(k) - \gamma_{uu}(k)$, with autocorrelations

$$\rho_{\varepsilon\varepsilon}(k) = \frac{\gamma_{\varepsilon\varepsilon}(k)}{\gamma_{\varepsilon\varepsilon}(0)} = \frac{\rho_{\beta\beta}(k) - \gamma_{uu}(k)/\gamma_{\beta\beta}(0)}{1 - \gamma_{uu}(0)/\gamma_{\beta\beta}(0)}$$
$$= \frac{\rho_{\beta\beta}(k) - \sum_{j=0}^{\infty} \rho_{\alpha\beta}(j)\rho_{\alpha\beta}(j+k)}{1 - \sum_{j=0}^{\infty} \rho_{\alpha\beta}^2(j)}$$

Now, in practice, it is necessary to *estimate* the prewhitening transformation. Having made the approximate prewhitening transformation, rough values for $\rho_{\varepsilon\varepsilon}(k)$ could be obtained by substituting the estimates $r_{\alpha\beta}(j)$ of the cross-correlation function between transformed input and output and $r_{\beta\beta}(j)$ of the autocorrelation function of the transformed output.

Application to the Gas Furnace Example. Table 12.2 shows the first 12 values of the sample autocorrelations and partial autocorrelations of the noise series $\hat{N}_t = Y_t - \hat{y}_t$, where $\hat{y}_t = \hat{\delta}^{-1}(B)\hat{\omega}(B)X_{t-3}$ is computed as in (12.2.15) using the preliminary estimates for the transfer function model obtained previously. That is, the values are computed as

$$\hat{y}_t = 0.57\hat{y}_{t-1} - (0.53X_{t-3} + 0.33X_{t-4} + 0.51X_{t-5})$$

The partial autocorrelations of $\hat{N}_t$ indicate that a second-order autoregressive model might be an adequate representation, and the least-squares estimates obtained from the $\hat{N}_t$ values for the AR(2) model yield

$$(1 - 1.54B + 0.64B^2)N_t = a_t \tag{12.2.19}$$

with $\hat{\sigma}_a^2 = 0.057$.

Thus, the analysis of this section and Section 12.1.2 suggests the identification

$$Y_t = \frac{\omega_0 - \omega_1 B - \omega_2 B^2}{1 - \delta_1 B - \delta_2 B^2} X_{t-3} + \frac{1}{1 - \phi_1 B - \phi_2 B^2} a_t \tag{12.2.20}$$

for the gas furnace model. Furthermore, the initial estimates $\hat{\omega}_0 = -0.53, \hat{\omega}_1 = 0.33, \hat{\omega}_2 = 0.51, \hat{\delta}_1 = 0.57, \hat{\delta}_2 = 0.02, \hat{\phi}_1 = 1.54$, and $\hat{\phi}_2 = -0.64$ can be used as rough starting values for the nonlinear estimation procedures that we describe in Section 12.3.

12.2.4 Some General Considerations in Identifying Transfer Function Models

Some general remarks can now be made concerning the procedure for identifying transfer function and noise models that we have just described.

1. For many practical situations, when the effect of noise is appreciable, a delayed first- or second-order system such as that given by (12.2.13), or some simplification of it, would often provide as elaborate a model as could be justified for the data.
2. Efficient estimation is only possible assuming the model *form* to be known. The estimates $\hat{v}_k$ given by (12.2.12) are in general *necessarily* inefficient therefore. They are employed at the identification stage because they are easily computed and can indicate a form of model worthy to be fitted by more elaborate means.
3. Even if these were efficient estimates, the number of $\hat{v}$'s required to trace out the impulse response function fully would typically be considerably larger than the number of parameters in a transfer function model. In cases where the δ's and ω's in an adequate transfer function model could be estimated accurately, nevertheless, the estimates of the corresponding v's could have large variances and be highly correlated.
4. The variance of

$$r_{\alpha\beta}(k) = \hat{v}_k \frac{s_\alpha}{s_\beta}$$

is of order $1/n$. Thus, we can expect that the estimates $r_{\alpha\beta}(k)$ and hence the $\hat{v}_k$ will be buried in noise unless σ_α is reasonably large compared with the residual noise, or unless n is large. Thus, the identification procedure requires the variation in the input X_t to be reasonably large compared with the variation due to the noise and/or a large volume of data is available. These requirements are satisfied by the gas furnace data for which, as we show in Section 12.3, the initial identification is remarkably good. When these requirements are not satisfied, the identification procedure may fail. Usually, this will mean that only very rough estimates are possible with the available data. However, some kind of rudimentary modeling may be possible by postulating a plausible but simple transfer function/noise model, fitting directly by the least-squares procedures of the next section, and applying diagnostic checks leading to elaboration of the model when this proves necessary.
5. It should, perhaps, be emphasized that the prewhitened series α_t and β_t, and their cross-correlation function, $r_{\alpha\beta}(k)$, in particular, are used only for the purpose of identification of the form of the transfer function model. Once the model form is identified, the original series X_t and Y_t, not the prewhitened series, are used for parameter estimation, forecasting, and so on.
6. An alternative method for identification of the transfer function–noise model was proposed by Haugh and Box (1977), and similar ideas were also discussed by Priestley (1981, Chapter 9). The method, which might be referred to as ''double prewhitening,'' involves prewhitening *both* input and output series. That is, separate univariate ARIMA models are built for both the input and the output processes, and then the cross-correlation structure of the resulting (univariate white noise) residuals from these models is examined. However, while sometimes useful, this procedure can

become overly complicated in terms of the final model specified, due to the use of two sets of prewhitening factors.

7. The above discussion has focused on transfer function models with a single input variable X_t. An alternative method of identifying transfer function models, which readily generalizes to deal with multiple inputs, is given in Appendix A12.1. Transfer function models can also be specified using methods developed for multivariate time series analysis as demonstrated by Tiao and Box (1981). A discussion of such methods is given in Chapter 14.

Lack of Uniqueness of the Model. Suppose that a particular dynamic system is represented by the model

$$Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + \varphi^{-1}(B)\theta(B)a_t \tag{12.2.21}$$

Then it could equally well be represented by

$$L(B)Y_t = L(B)\delta^{-1}(B)\omega(B)X_{t-b} + L(B)\varphi^{-1}(B)\theta(B)a_t \tag{12.2.22}$$

where $L(B)$ could be an arbitrary common factor, and hence would be redundant. Similar to the discussion in Section 7.3.5 on parameter redundancy for ARMA models, for uniqueness of model parameterization in (12.2.21) it is clear that the possibility of common factors in the operators $\delta(B)$ and $\omega(B)$, or in the $\varphi(B)$ and $\theta(B)$ operators, must be avoided. The chance that we may iterate toward a model of unnecessarily complicated form is reduced if we base our strategy on the following considerations:

1. Since rather simple transfer function models of first or second order, with or without delay, are often adequate, iterative model building should begin with a fairly simple model, looking for further simplification if this is possible, and reverting to more complicated models only as the need is demonstrated.
2. One should be always on the look out for the possibility of removing a factor common to two or more of the operators on Y_t, X_t, and a_t. In practice, we will be dealing with estimated coefficients, which may be subject to rather large sampling errors, so that only approximate common factors in the factorizations can be expected. Thus, a very careful analysis may be needed to detect such factors. Of course, having removed what appears to be a common factor, the model can be refitted and checked to show whether the simplification can be justified.
3. When simplification by factorization is possible, but is overlooked, the least-squares estimation procedure may become extremely unstable since the minimum will tend to lie on a line or surface in the parameter space rather than at a point. Conversely, instability in the solution can point to the possibility of simplification of the model. As noted earlier, one reason for carrying out the identification procedure before fitting the model is to avoid redundancy or, conversely, to achieve *parsimony* in parameterization.

Remark. If the operator $L(B)$ in (12.2.22) were set equal to $\varphi(B)\delta(B)$, we would obtain

$$\varphi(B)\delta(B)Y_t = \varphi(B)\omega(B)X_{t-b} + \delta(B)\theta(B)a_t \tag{12.2.23}$$

which can be written as

$$\delta^*(B)Y_t = \omega^*(B)X_{t-b} + \theta^*(B)a_t \tag{12.2.23a}$$

Models of the general form of (12.2.23a) have been referred to as ARMAX models in the econometric literature (e.g., Hannan and Deistler, 1988; Hannan et al., 1979; Reinsel, 1979). As can be seen, care is needed to avoid the occurrence of common factors among the operators in this form.

12.3 FITTING AND CHECKING TRANSFER FUNCTION MODELS

12.3.1 Conditional Sum-of-Squares Function

We now consider the problem of efficiently and simultaneously estimating the parameters b, $\boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\phi}$, and $\boldsymbol{\theta}$ in the tentatively identified model

$$y_t = \delta^{-1}(B)\omega(B)x_{t-b} + n_t \tag{12.3.1}$$

where $y_t = \nabla^d Y_t$, $x_t = \nabla^d X_t$, and $n_t = \nabla^d N_t$ are all stationary processes and

$$n_t = \phi^{-1}(B)\theta(B)a_t \tag{12.3.2}$$

It is assumed that $n = N - d$ pairs of values are available for the analysis and that Y_t and X_t (y_t and x_t if $d > 0$) denote deviations from expected values. These expected values may be estimated along with the other parameters, but for the lengths of time series normally worth analyzing it will usually be sufficient to use the sample means as estimates. When $d > 0$, it will frequently be true that expected values for y_t and x_t are zero.

If starting values $\mathbf{x}_0$, $\mathbf{y}_0$, and $\mathbf{a}_0$ prior to the commencement of the series were available, then given the data, for any choice of the parameters $(b, \boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\phi}, \boldsymbol{\theta})$ and of the starting values $(\mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_0)$ we could calculate, successively, values of

$$a_t = a_t(b, \boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\phi}, \boldsymbol{\theta} | \mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_0)$$

for $t = 1, 2, \ldots, n$. Under the normal assumption for the a_t's, a close approximation to the maximum likelihood estimates of the parameters can be obtained by minimizing the *conditional sum-of-squares function*,

$$S_0(b, \boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} a_t^2(b, \boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\phi}, \boldsymbol{\theta} | \mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_0) \tag{12.3.3}$$

Three-Stage Procedure for Calculating the a's. Given appropriate starting values, the generation of the a_t's *for any particular choice of the parameter values* may be accomplished using the following three-stage procedure.

First, the output y_t from the transfer function model may be computed from

$$y_t = \delta^{-1}(B)\omega(B)x_{t-b}$$

that is, from

$$\delta(B)y_t = \omega(B)x_{t-b}$$

or from

$$y_t - \delta_1 y_{t-1} - \cdots - \delta_r y_{t-r} = \omega_0 x_{t-b} - \omega_1 x_{t-b-1} - \cdots - \omega_s x_{t-b-s} \tag{12.3.4}$$

Having calculated the y_t series, then using (12.3.1), the noise series n_t can be obtained from

$$n_t = y_t - y_t \tag{12.3.5}$$

Finally, the a_t's can be obtained from (12.3.2) written in the form

$$\theta(B)a_t = \phi(B)n_t$$

that is,

$$a_t = \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} + n_t - \phi_1 n_{t-1} - \cdots - \phi_p n_{t-p} \tag{12.3.6}$$

Starting Values. As discussed in Section 7.1.3 for stochastic model estimation, the effect of transients can be minimized if the difference equations are started off from a value of t for which all previous x_t's and y_t's are known. Thus, y_t in (12.3.4) is calculated from $t = u + 1$ onward, where u is the larger of r and $s + b$. This means that n_t will be available from n_{u+1} onward; hence, if unknown a_t's are set equal to their unconditional expected values of zero, the a_t's may be calculated from a_{u+p+1} onward. Thus, the conditional sum-of-squares function is

$$S_0(b, \boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=u+p+1}^{n} a_t^2(b, \boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\phi}, \boldsymbol{\theta} | \mathbf{x}_0, \mathbf{y}_0, \mathbf{a}_0) \tag{12.3.7}$$

Example Using the Gas Furnace Data. For these data, the model (12.2.20), namely

$$Y_t = \frac{\omega_0 - \omega_1 B - \omega_2 B^2}{1 - \delta_1 B - \delta_2 B^2} X_{t-3} + \frac{1}{1 - \phi_1 B - \phi_2 B^2} a_t$$

has been identified. Equations (12.3.4), (12.3.5), and (12.3.6) then become

$$y_t = \delta_1 y_{t-1} + \delta_2 y_{t-2} + \omega_0 X_{t-3} - \omega_1 X_{t-4} - \omega_2 X_{t-5} \tag{12.3.8}$$

$$N_t = Y_t - y_t \tag{12.3.9}$$

$$a_t = N_t - \phi_1 N_{t-1} - \phi_2 N_{t-2} \tag{12.3.10}$$

Thus, (12.3.8) can be used to generate y_t from $t = 6$ onward and (12.3.10) to generate a_t from $t = 8$ onward. The slight loss of information that results will not be important for a sufficiently long length of series. For example, since $N = 296$ for the gas furnace data, the loss of seven values at the beginning of the series is of little practical consequence.

In the example above, we have assumed that $b = 3$. To estimate b, the values of $\boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\phi}$, and $\boldsymbol{\theta}$, which minimize the conditional sum of squares, can be calculated for each value of b in the likely range and the overall minimum with respect to $b, \boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\phi}$, and $\boldsymbol{\theta}$ obtained.

12.3.2 Nonlinear Estimation

A nonlinear least-squares algorithm, analogous to that given for fitting the stochastic model in Section 7.2.4, can be used to obtain the least-squares estimates and their approximate

standard errors. The algorithm will behave well when the sum-of-squares function is roughly quadratic. However, the procedure can sometimes run into trouble, in particular if the parameter estimates are very highly correlated (if, for example, the model approaches singularity due to near-common factors in the factorizations of the operators), or, in some cases, if estimates are near a boundary of the permissible parameter space. In difficult cases, the estimation situation may be clarified by plotting sum-of-squares contours for selected two-dimensional sections of the parameter space.

The nonlinear least-squares algorithm can be implemented as follows: At any stage of the iteration, and for some fixed value of the delay parameter b, let the best guesses available for the remaining parameters be denoted by

$$\boldsymbol{\beta}_0' = (\delta_{1,0}, \ldots, \delta_{r,0};\ \omega_{0,0}, \ldots, \omega_{s,0}; \phi_{1,0}, \ldots, \phi_{p,0}; \theta_{1,0}, \ldots, \theta_{q,0})$$

Now let $a_{t,0}$ denote that value of a_t computed from the model, as in Section 12.3.1, for the guessed parameter values $\boldsymbol{\beta}_0$ and denote the negative of the derivatives of a_t with respect to the parameters as follows:

$$d_{i,t}^{(\delta)} = -\frac{\partial a_t}{\partial \delta_i}\bigg|_{\boldsymbol{\beta}_0} \qquad d_{j,t}^{(\omega)} = -\frac{\partial a_t}{\partial \omega_j}\bigg|_{\boldsymbol{\beta}_0} \qquad d_{g,t}^{(\phi)} = -\frac{\partial a_t}{\partial \phi_g}\bigg|_{\boldsymbol{\beta}_0} \qquad d_{h,t}^{(\theta)} = -\frac{\partial a_t}{\partial \theta_h}\bigg|_{\boldsymbol{\beta}_0} \tag{12.3.11}$$

Then a Taylor series expansion of $a_t = a_t(\boldsymbol{\beta})$ about parameter values $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ can be rearranged in the form

$$\begin{aligned} a_{t,0} \simeq & \sum_{i=1}^{r} (\delta_i - \delta_{i,0}) d_{i,t}^{(\delta)} + \sum_{j=0}^{s} (\omega_j - \omega_{j,0}) d_{j,t}^{(\omega)} \\ & + \sum_{g=1}^{p} (\phi_g - \phi_{g,0}) d_{g,t}^{(\phi)} + \sum_{h=1}^{q} (\theta_h - \theta_{h,0}) d_{h,t}^{(\theta)} + a_t \end{aligned} \tag{12.3.12}$$

We proceed as in Section 7.2 to obtain adjustments $\delta_i - \delta_{i,0}, \omega_j - \omega_{j,0}$, and so on, by fitting this linearized equation by standard linear least-squares. By adding the adjustments to the first guesses $\boldsymbol{\beta}_0$, a set of second guesses can be formed and the procedure repeated until convergence is reached.

The derivatives in (12.3.11) may be computed recursively. However, it seems simplest to work with a standard nonlinear least-squares computer program in which derivatives are determined numerically and an option is available of ''constrained iteration'' to prevent instability. It is then necessary only to program the computation of a_t itself.

The covariance matrix of the estimates may be obtained from the converged value of the matrix

$$(\mathbf{X}_{\hat{\boldsymbol{\beta}}}' \mathbf{X}_{\hat{\boldsymbol{\beta}}})^{-1} \hat{\sigma}_a^2 \simeq \hat{\text{cov}}[\hat{\boldsymbol{\beta}}]$$

as described in Section 7.2.2; in addition, the least-squares estimates $\hat{\boldsymbol{\beta}}$ have been shown to have a multivariate normal asymptotic distribution (e.g., Pierce, 1972a; Reinsel, 1979). If the delay b, which is an integer, needs to be estimated, the iteration may be run to convergence for a series of values of b and the value of b giving the minimum sum of squares selected. One special feature (see, for example, Pierce, 1972a) of the covariance matrix of the least-squares estimates $\hat{\boldsymbol{\beta}}$ is that it will be approximately a block diagonal matrix whose two blocks on the diagonal consist of the covariance matrices of the

parameters $(\hat{\boldsymbol{\delta}}', \hat{\boldsymbol{\omega}}') = (\hat{\delta}_1, \ldots, \hat{\delta}_r, \hat{\omega}_0, \ldots, \hat{\omega}_s)$ and $(\hat{\boldsymbol{\phi}}', \hat{\boldsymbol{\theta}}') = (\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q)$, respectively. Thus, the parameter estimates of the transfer function part of the model are approximately uncorrelated with the estimates of the noise part of the model, which results from the assumed independence between the input X_t and the white noise a_t in the model.

More exact sum-of-squares and exact likelihood function methods could also be employed in the estimation of the transfer function–noise models, as in the case of the ARMA models discussed in Chapter 7 (see, e.g. Newbold, 1973). The state-space model Kalman filtering and innovations algorithm approach to the exact likelihood evaluation discussed in Section 7.4 could also be used. However, for moderate and large n and nonseasonal data, there will generally be little difference between the conditional and exact methods.

Remark. Commercially available software packages such as SAS and SCA include algorithms for estimating the parameters in transfer function–noise models. The software package R can also be used for model fitting. In particular, the newly released package MTS for multivariate time series analysis that we will use in Chapter 14 has a function tfm1() that fits a transfer function–noise model to a dataset with a single input variable X. A demonstration of this package is given in Section 12.4.1. A second function tfm2() fits a model with two input variables to the data.

12.3.3 Use of Residuals for Diagnostic Checking

Serious model inadequacy can usually be detected by examining

1. The autocorrelation function $r_{\hat{a}\hat{a}}(k)$ of the residuals $\hat{a}_t = a_t(\hat{b}, \hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}})$ from the fitted model.
2. Certain cross-correlation functions involving input and residuals: in particular, the cross-correlation function $r_{\alpha\hat{a}}(k)$ between prewhitened input α_t and the residuals $\hat{a}_t$.

Suppose, if necessary after suitable differencing, that the model can be written as

$$\begin{aligned} y_t &= \delta^{-1}(B)\omega(B)x_{t-b} + \phi^{-1}(B)\theta(B)a_t \\ &= \upsilon(B)x_t + \psi(B)a_t \end{aligned} \tag{12.3.13}$$

Now, suppose that we select an incorrect model leading to residuals a_{0t}, where

$$y_t = \upsilon_0(B)x_t + \psi_0(B)a_{0,t}$$

Then

$$a_{0t} = \psi_0^{-1}(B)[\upsilon(B) - \upsilon_0(B)]x_t + \psi_0^{-1}(B)\psi(B)a_t \tag{12.3.14}$$

Thus, it is apparent in general that if a wrong model is selected, the a_{0t}'s will be autocorrelated and the a_{0t}'s will be cross-correlated with the x_t's and hence with the α_t's, which generate the x_t's.

Now consider what happens in two special cases: (1) when the transfer function model is correct but the noise model is incorrect, and (2) when the transfer function model is incorrect.

Transfer Function Model Correct, Noise Model Incorrect. If $v_0(B) = v(B)$ but $\psi_0(B) \neq \psi(B)$, then (12.3.14) becomes

$$a_{0t} = \psi_0^{-1}(B)\psi(B)a_t \tag{12.3.15}$$

Therefore, the a_{0t}'s would *not* be cross-correlated with x_t's or with α_t's. However, the a_{0t} process would be autocorrelated, and the form of the autocorrelation function could indicate appropriate modification of the noise structure, as discussed for univariate ARIMA models in Section 8.3.

Transfer Function Model Incorrect. From (12.3.14) it is apparent that if the transfer function model were incorrect, not only would the a_{0t}'s be cross-correlated with the x_t's (and α_t's), but *also the a_{0t}'s would be autocorrelated*. This would be true even if the noise model were correct, for then (12.3.14) would become

$$a_{0t} = \psi^{-1}(B)[v(B) - v_0(B)]x_t + a_t \tag{12.3.16}$$

Whether or not the noise model was correct, a cross-correlation analysis could indicate the modifications needed in the transfer function model. This aspect is clarified by considering the model after prewhitening. If the output and the input are assumed to be transformed so that the input is white noise, then, as in (12.2.8), we may write the model as

$$\beta_t = v(B)\alpha_t + \varepsilon_t$$

where $\beta_t = \theta_x^{-1}(B)\phi_x(B)y_t$ and $\varepsilon_t = \theta_x^{-1}(B)\phi_x(B)n_t$. Now, consider the quantities

$$\varepsilon_{0t} = \beta_t - v_0(B)\alpha_t$$

Since $\varepsilon_{0t} = [v(B) - v_0(B)]\alpha_t + \varepsilon_t$, arguing as in Section 12.1.1, the cross-correlations between the ε_{0t}'s and the α_t's measure the discrepancy between the correct and incorrect impulse functions. Specifically, as in (12.2.11),

$$v_k - v_{0k} = \frac{\rho_{\alpha\varepsilon_0}(K)\sigma_{\varepsilon_0}}{\sigma_\alpha} \qquad k = 0, 1, 2, \ldots \tag{12.3.17}$$

12.3.4 Specific Checks Applied to the Residuals

In practice, we do not know the process parameters exactly but must apply our checks to the residuals $\hat{a}_t$ computed after least-squares fitting. Even if the functional form of the fitted model were adequate, the parameter estimates would differ somewhat from the true values and the distribution of the autocorrelations of the residuals $\hat{a}_t$'s would also differ to some extent from that of the autocorrelations of the a_t's. Therefore, some caution is necessary in using the results of the previous sections to suggest the behavior of residual correlations. The brief discussion that follows is based in part on a more detailed study by Pierce (1972b).

Autocorrelation Checks. Suppose that a transfer function–noise model having been fitted by least-squares and the residuals $\hat{a}_t$'s calculated by substituting least-squares estimates for the parameters and the estimated autocorrelation function $r_{\hat{a}\hat{a}}(k)$ of these residuals is computed. Then, as we have seen

1. If the autocorrelation function $r_{\hat{a}\hat{a}}(k)$ shows marked correlation patterns, this suggests model inadequacy.
2. If the cross-correlation checks do not indicate inadequacy of the transfer function model, the inadequacy is probably in the fitted noise model $n_t = \psi_0(B)\hat{a}_{0t}$.

In the latter case, identification of a subsidiary model

$$\hat{a}_{0t} = T(B)a_t$$

to represent the correlation of the residuals from the primary model can, in accordance with (12.3.15), indicate roughly the form

$$n_t = \psi_0(B)T(B)a_t$$

to take for the modified noise model. However, in making assessments of whether an apparent discrepancy of estimated autocorrelations from zero is, or is not, likely to point to a nonzero theoretical value, certain facts must be borne in mind analogous to those discussed in Section 8.2.1.

Suppose that after allowing for starting values, $m = n - u - p$ values of the $\hat{a}_t$'s are actually available for this computation. Then if the model was correct in functional form and the *true parameter values were substituted*, the residuals would be white noise and the estimated autocorrelations would be distributed mutually independently about zero with variance $1/m$. When estimates are substituted for the parameter values, the distributional properties of the estimated autocorrelations at low lags are affected. In particular, the variance of these estimated low-lag autocorrelations can be considerably less than $1/m$, and the values can be highly correlated. Thus, with k small, comparison of an estimated autocorrelation $r_{\hat{a}\hat{a}}(k)$ with a ''standard error'' $1/\sqrt{m}$ could greatly underestimate its significance. Also, ripples in the estimated autocorrelation function at low lags can arise simply because of the high induced correlation between these estimates. If the amplitude of such low-lag ripples is small compared with $1/\sqrt{m}$, they could have arisen by chance alone and need not be indicative of some real pattern in the theoretical autocorrelations.

A helpful overall check, which takes account of these distributional effects produced by fitting, is as follows. Consider the first K estimated autocorrelations $r_{\hat{a}\hat{a}}(1), \ldots, r_{\hat{a}\hat{a}}(K)$ and let K be taken sufficiently large so that if the model is written as $y_t = v(B)x_t + \psi(B)a_t$, the weights ψ_j can be expected to be negligible for $j > K$. Then if the functional form of the model is adequate, the quantity

$$Q = m\sum_{k=1}^{K} r_{\hat{a}\hat{a}}^2(k) \tag{12.3.18}$$

is approximately distributed as χ^2 with $K - p - q$ degrees of freedom. Note that the degrees of freedom in χ^2 depend on the number of parameters in the noise model but not on the number of parameters in the transfer function model. By referring Q to a table of percentage points of χ^2, we can obtain an approximate test of the hypothesis of model adequacy. However, in practice, the modified statistic

$$\tilde{Q} = m(m+2)\sum_{k=1}^{K}(m-k)^{-1}r_{\hat{a}\hat{a}}^2(k) \tag{12.3.18a}$$

analogous to (8.2.3) of Section 8.2.2 for the ARIMA model, would be recommended instead of (12.3.18) because $\tilde{Q}$ provides a closer approximation to the chi-squared distribution than Q under the null hypothesis of model adequacy.

Cross-Correlation Check. As we have seen in Section 12.3.3,

1. A pattern of markedly nonzero cross-correlations $r_{x\hat{a}}(k)$ suggests inadequacy of the transfer function model.
2. A somewhat different cross-correlation analysis can suggest the *type* of modification needed in the transfer function model. Specifically, if the fitted transfer function is $\hat{v}_0(B)$ and we consider the cross-correlations between the quantities $\hat{\varepsilon}_{0t} = \beta_t - \hat{v}_0(B)\alpha_t$ and α_t, rough estimates of the discrepancies $v_k - v_{0k}$ are given by

$$\frac{r_{\alpha\hat{\varepsilon}_0}(k)s_{\hat{\varepsilon}_0}}{s_\alpha}$$

Suppose that the model were of the correct functional form and *true* parameter values had been substituted. The residuals would be white noise uncorrelated with the x_t's and, using (12.1.11), the variance of the $r_{xa}(k)$ for an effective length of series m would be approximately $1/m$. However, unlike the autocorrelations $r_{aa}(k)$, these cross-correlations will not be approximately uncorrelated. In general, if the x_t's are autocorrelated, so are the cross-correlations $r_{xa}(k)$. In fact, as has been seen in (12.1.12), on the assumption that the x_t's and the a_t's have no cross-correlation, the correlation coefficient between $r_{xa}(k)$ and $r_{xa}(k+l)$ is

$$\rho[r_{xa}(k), r_{xa}(k+l)] \simeq \rho_{xx}(l) \tag{12.3.19}$$

That is, approximately, the cross-correlations have the *same* autocorrelation function as does the original input series x_t. Thus, when the x_t's are autocorrelated, a perfectly adequate transfer function model will give rise to estimated cross-correlations $r_{x\hat{a}}(k)$, which, although small in magnitude, may show *pronounced patterns*. This effect is eliminated if the check is made by computing cross-correlations $r_{\alpha\hat{a}}(k)$ with the *prewhitened* input α_t.

As with the autocorrelations, when estimates are substituted for parameter values, the distributional properties of the estimated cross-correlations are affected. However, a rough overall test of the hypothesis of model adequacy, similar to the autocorrelation test, can be obtained based on the magnitudes of the estimated cross-correlations. To employ the check, the cross-correlations $r_{\alpha\hat{a}}(k)$ for $k = 0, 1, 2, \ldots, K$ between the input α_t in *prewhitened* form and the residuals $\hat{a}_t$ are estimated, and K is chosen sufficiently large so that the weights v_j and ψ_j in (12.3.13) can be expected to be negligible for $j > K$. The effects resulting from the use of estimated parameters in calculating residuals are, as before, principally confined to cross-correlations of low order whose variances are considerably less than m^{-1} and that may be highly correlated even when the input is white noise.

For an overall test, Pierce (1972b) showed that

$$S = m\sum_{k=0}^{K} r^2_{\alpha\hat{a}}(k) \tag{12.3.20}$$

is approximately distributed as χ^2 with $K + 1 - (r + s + 1)$ degrees of freedom, where $(r + s + 1)$ is the number of parameters fitted in the transfer function model. Note that the number of degrees of freedom is independent of the number of parameters fitted in the noise model. Based on studies of the behavior of the Q statistic discussed in Chapter 8, the modified statistic, $\tilde{S} = m(m+2)\sum_{k=0}^{K}(m-k)^{-1}r_{\alpha\hat{a}}^2(k)$, might be suggested for use in practice because it may more accurately approximate the χ^2 distribution under the null model, although detailed investigations of its performance have not been made (however, see empirical results in Poskitt and Tremayne, 1981).

12.4 SOME EXAMPLES OF FITTING AND CHECKING TRANSFER FUNCTION MODELS

12.4.1 Fitting and Checking of the Gas Furnace Model

We now illustrate the approach described in Section 12.2 to the fitting of the model

$$Y_t = \frac{\omega_0 - \omega_1 B - \omega_2 B^2}{1 - \delta_1 B - \delta_2 B^2} X_{t-3} + \frac{1}{1 - \phi_1 B - \phi_2 B^2} a_t$$

which was identified for the gas furnace data in Sections 12.2.2 and 12.2.3.

Nonlinear Estimation. Using the initial estimates $\hat{\omega}_0 = -0.53, \hat{\omega}_1 = 0.33, \hat{\omega}_2 = 0.51, \hat{\delta}_1 = 0.57, \hat{\delta}_2 = 0.02, \hat{\phi}_1 = 1.54$, and $\hat{\phi}_2 = -0.64$ derived in Sections 12.2.2 and 12.2.3 with the conditional least-squares algorithm described in Section 12.3.2, least-squares values, to two decimals, were achieved in four iterations. However, to test whether the results would converge in much less favorable circumstances, Table 12.3 shows the iterations produced with all starting values taken to be either $+0.1$ or -0.1. The fact that, even then, convergence was achieved in 10 iterations with as many as seven parameters in the model is encouraging.

The last line in Table 12.3 shows the rough preliminary estimates obtained at the identification stage in Sections 12.2.2 and 12.2.3. It is seen that for this example, they are in close agreement with the least-squares estimates given on the previous line. Thus, the final fitted transfer function model is

$$\begin{array}{rcl} (1 - 0.57B - 0.01B^2)Y_t & = & -(0.53 + 0.37B + 0.51B^2)X_{t-3} \\ (\pm 0.21)(\pm 0.14) & & (\pm 0.08)(\pm 0.15)(\pm 0.16) \end{array} \qquad (12.4.1)$$

and the fitted noise model is

$$\begin{array}{c} (1 - 1.53B + 0.63B^2)N_t = a_t \\ (\pm 0.05)(\pm 0.05) \end{array} \qquad (12.4.2)$$

with $\hat{\sigma}_a^2 = 0.0561$, where the limits in parentheses are the ± 1 standard error limits obtained from the nonlinear least-squares estimation procedure.

Diagnostic Checking. Before accepting the model above as an adequate representation of the system, autocorrelation and cross-correlation checks should be applied, as described in Section 12.3.4. The first 36 lags of the residual autocorrelations are given in Table 12.4(a) and plotted in Figure 12.6(a), together with their approximate two standard error limits

TABLE 12.3 Convergence of Nonlinear Least-Squares Fit of Gas Furnace Data

Iteration	ω_0	ω_1	ω_2	δ_1	δ_2	ϕ_1	ϕ_2	Sum of Squares
0	0.10	−0.10	−0.10	0.10	0.10	0.10	0.10	13,601.00
1	−0.46	0.63	0.60	0.14	0.27	1.33	−0.27	273.10
2	−0.52	0.45	0.31	0.40	0.52	1.37	−0.43	92.50
3	−0.63	0.60	0.01	0.12	0.73	1.70	−0.76	31.80
4	−0.54	0.50	0.29	0.24	0.42	1.70	−0.81	19.70
5	−0.50	0.31	0.51	0.63	0.09	1.56	−0.68	16.84
6	−0.53	0.38	0.53	0.54	0.01	1.54	−0.64	16.60
7	−0.53	0.37	0.51	0.56	0.01	1.53	−0.63	16.60
8	−0.53	0.37	0.51	0.56	0.01	1.53	−0.63	16.60
9	−0.53	0.37	0.51	0.57	0.01	1.53	−0.63	16.60
Preliminary estimates	−0.53	0.33	0.51	0.57	0.02	1.54	−0.64	

$\pm 2/\sqrt{m} \simeq 0.12$ ($m = 289$) under the assumption that the model is adequate. There seems to be no evidence of model inadequacy from the behavior of individual autocorrelations. This is confirmed by calculating the $\tilde{Q}$ criterion in (12.3.18a), which is

$$\tilde{Q} = (289)(291)\sum_{k=1}^{36}(289-k)^{-1}r_{\hat{a}\hat{a}}^2(k) = 43.8$$

Comparison of $\tilde{Q}$ with the χ^2 table for $K - p - q = 36 - 2 - 0 = 34$ degrees of freedom provides no grounds for questioning model adequacy.

The first 36 lags of the cross-correlation function $r_{x\hat{a}}(k)$ between the input X_t and the residuals $\hat{a}_t$ are given Table 12.4(b) and shown in Figure 12.6(b), together with their approximate two standard error limits $\pm 2/\sqrt{m}$. It is seen that although the cross-correlations $r_{x\tilde{a}}(k)$ do not exceed their two standard error limits, they are themselves highly autocorrelated. This is to be expected because as indicated by (12.3.19), the estimated cross-correlations follow the same stochastic process as does the input X_t, and as we have already seen, for this example the input was highly autocorrelated.

The corresponding cross-correlations between the prewhitened input α_t and the residuals $\hat{a}_t$ are given in Table 12.4(c) and shown in Figure 12.6(c). The $\tilde{S}$ criterion yields

$$\tilde{S} = (289)(291)\sum_{k=0}^{35}(289-k)^{-1}r_{a\hat{a}}^2(k) = 32.1$$

Comparison of $\tilde{S}$ with the X^2 table for $K + 1 - (r + s + 1) = 36 - 5 = 31$ degrees of freedom again provides no evidence that the model is inadequate.

Parameter Estimation Using R. We will now use the R software to fit the model employed in (12.4.1) and (12.4.2) to the gas furnace data. The parameter estimation can be performed using the function tfm1() in the MTS package developed for multivariate time series analysis. The arguments of this function are tfm1(Y, X,orderX=c(r,s,b),orderN=c(p,d,q)). The function call and the resulting output are shown below:

TABLE 12.4 Estimated Autocorrelation and Cross-Correlation Functions of Residuals from Fitted Gas Furnace Model

Lag k					(a) Autocorrelation $r_{\hat{a}\hat{a}}(k)$								Upper Bound to Standard Error
1–12	0.02	0.06	−0.07	−0.05	−0.05	0.12	0.03	0.03	−0.08	0.05	0.02	0.10	±0.06
13–24	−0.04	0.05	−0.09	−0.01	−0.08	0.00	−0.12	0.00	−0.01	0.08	0.02	−0.01	±0.06
25–36	0.04	−0.02	0.02	0.09	−0.12	0.06	−0.03	−0.06	0.11	0.02	0.03	0.06	±0.06
					(b) $r_{x\hat{a}}(k)$ between the input and the output residuals								
0–11	0.00	0.00	0.00	0.00	0.00	0.00	−0.01	−0.02	−0.03	−0.05	−0.06	−0.05	±0.06
12–23	−0.03	−0.03	−0.03	−0.07	−0.10	−0.12	−0.12	−0.10	−0.04	−0.01	−0.01	−0.02	±0.06
24–35	−0.03	−0.04	−0.04	−0.02	−0.01	0.02	0.04	0.05	0.06	0.07	0.07	0.06	±0.06
					(c) $r_{u\hat{a}}(k)$ between the prewhitened input and the output residuals								
0–11	−0.06	0.03	−0.01	0.00	0.01	0.01	0.01	−0.04	0.02	0.07	−0.03	−0.02	±0.06
12–23	−0.03	−0.11	0.02	0.04	0.04	0.01	0.01	−0.15	−0.03	−0.07	−0.08	0.02	±0.06
24–35	−0.01	0.02	0.05	−0.07	0.00	0.04	−0.15	0.04	0.03	−0.02	0.00	−0.03	±0.06

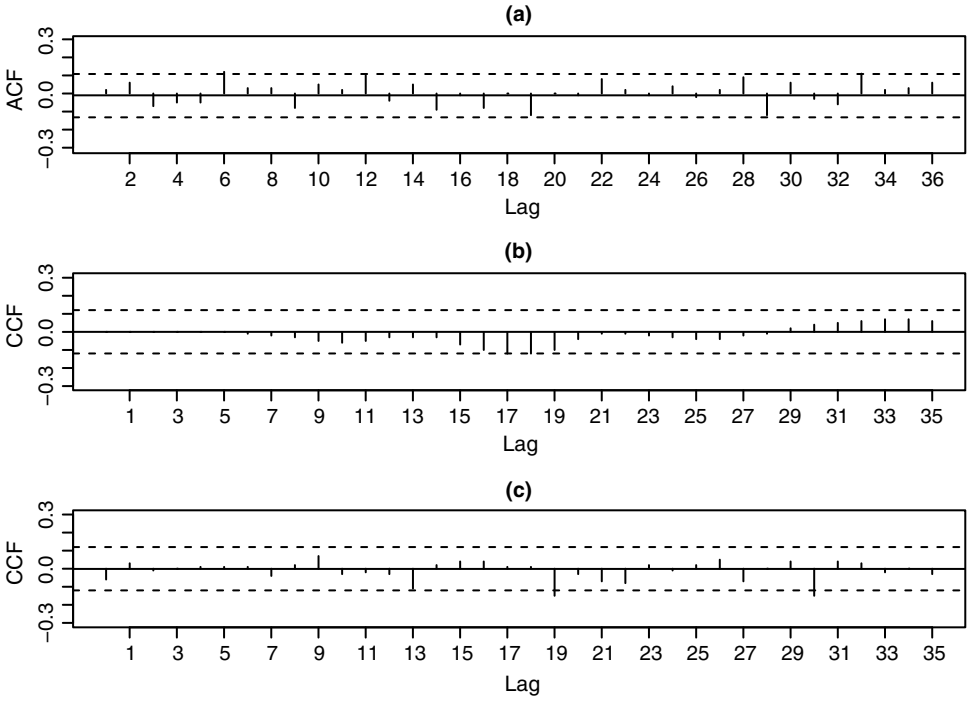


FIGURE 12.6 (a) Estimated autocorrelations of the residuals $r_{\hat{a}\hat{a}}(k)$ from the fitted gas furnace model, (b) estimated cross-correlations $r_{x\hat{a}}(k)$ between the input and the output residuals $r_{x\hat{a}}(k)$, and (c) estimated cross-correlations $r_{\alpha\hat{a}}(k)$ between the prewhitened input and the output residuals.

```
>  library(MTS)
>  m1=tfm1(Y,X,orderX=c(2,2,3),orderN=c(2,0,0))
   Model Output:
     Delay:  3
     Transfer function coefficients & s.e.:
     in the order: constant, omega, and delta: 1 3 2
         [,1]    [,2]   [,3]   [,4]  [,5]    [,6]
  v    53.371 -0.5302 -0.371 -0.511 0.565 -0.0119
  se.v  0.142  0.0745  0.146  0.149 0.200  0.1415
  ARMA order: [1] 2 0 0
  ARMA coefficients & s.e.:
              [,1]    [,2]
  coef.arma 1.5315 -0.6321
  se.arma   0.0472  0.0502

>  names(m1)    % check contents of output
     [1] "estimate"  "sigma2"  "residuals" "varcoef" "Nt"
>  m1$sigma2
     [1] 0.0576   % residual variance
>  acf(m1$residuals)       % acf of the residuals
>  ccf(m1$residuals,X)     % cross-correlation
```

```
            between input series and residuals
>  ccf(m1$residuals,xprev) % cross-correlation
            between prewhitened input and residuals
```

Using the output from R and allowing for sign differences in the definition of $\omega(B)$, the estimated transfer function–noise model is

$$Y_t = \frac{-(0.53 + 0.37B + 0.51B^2)}{1 - 0.57B + 0.01B^2}X_{t-3} + \frac{1}{1 - 1.53B + 0.63B^2}a_t$$

We see that the parameter estimates for the transfer function and noise models are nearly identical to those shown in (12.4.1) and (12.4.2). The estimate of the residual variance is 0.0576, which is also close to the value 0.0561 quoted in the text. In addition, the residual autocorrelations, the cross-correlations between the input X_t and the residuals, and the cross-correlations between the prewhitened input and the residuals (not shown) were small and close to those displayed in Figure 12.6 although some minor differences were seen in the patterns.

Step and Impulse Responses. The estimate $\hat{\delta}_2 = 0.01$ in (12.4.1) is very small compared with its standard error ± 0.14, and the parameter δ_2 can in fact be omitted from the model without affecting the estimates of the remaining parameters to the accuracy considered. The final form of the combined transfer function–noise model for the gas furnace data is

$$Y_t = \frac{-(0.53 + 0.37B + 0.51B^2)}{1 - 0.57B}X_{t-3} + \frac{1}{1 - 1.53B + 0.63B^2}a_t$$

The step and impulse response functions corresponding to the transfer function model

$$(1 - 0.57B)Y_t = -(0.53 + 0.37B + 0.51B^2)X_{t-3}$$

are given in Figure 12.7. Using (11.2.5), the steady-state gain of the coded data is

$$g = \frac{-(0.53 + 0.37 + 0.51)}{1 - 0.57} = -3.3$$

The results agree very closely with those obtained by cross-spectral analysis (Jenkins and Watts, 1968).

Choice of Sampling Interval. When a choice is available, the sampling interval should be taken as fairly short compared with the time constants expected for the system. When in doubt, the analysis can be repeated with several trial sampling intervals. In the choice of sampling interval, it is the noise at the output that is important, and its variance should approach a minimum value as the interval is shortened. Thus, in the gas furnace example that we have used for illustration, a pen recorder was used to provide a continuous record of input and output. The discrete data that we have actually analyzed were obtained by reading off values from this continuous record at points separated by 9-second intervals. This interval was chosen because inspection of the traces shown in Figure 12.1 suggested that it ought to be adequate to allow all the variation (apart from slight pen chatter) that occurred in input and output to be taken account of. The use of this kind of common sense is usually a reliable guide in choosing the interval. The estimated mean square error for the gas furnace data, obtained by dividing $\sum_t (Y_t - \hat{Y}_t)^2$ by the appropriate number of

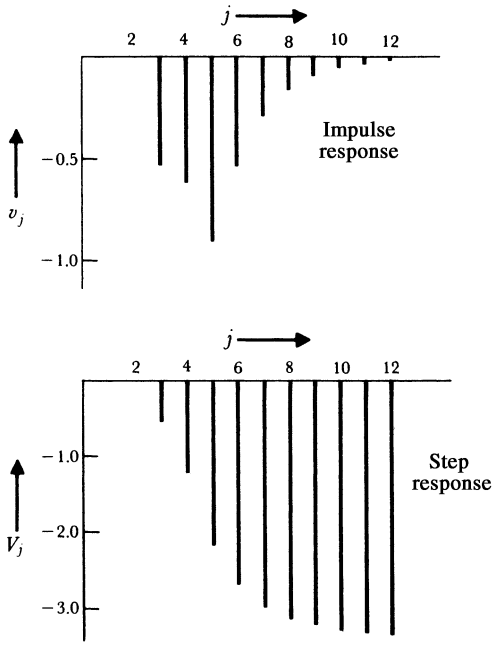


FIGURE 12.7 Impulse and step responses for transfer function model $(1 - 0.57B)Y_t = -(0.53 + 0.37B + 0.51B^2)X_{t-3}$ fitted to coded gas furnace data.

TABLE 12.5 Mean Square Error at the Output for Various Choices of the Sampling Interval for Gas Furnace Data

	Interval (Seconds)						
	9	18	27	36	45	54	72
Number of data points N	296	148	98	74	59	49	37
MS error	0.71	0.78	0.74	0.95	0.97	1.56	7.11

degrees of freedom, is shown for various time intervals in Table 12.5. These values are also plotted in Figure 12.8. Little change in mean square error occurs until the interval is almost 40 seconds, when a very rapid rise occurs. There is little difference in the mean square error, or indeed the plotted step response, for the 9-, 18-, and 27-second intervals, but a considerable change occurs when the 36-second interval is used. It will be seen that the 9-second interval we have used in this example is, in fact, conservative.

12.4.2 Simulated Example with Two Inputs

The fitting of models involving more than one input series involves no difficulty in principle, except for the increase in the number of parameters that has to be handled. For example, for two inputs we can write the model as

$$y_t = \delta_1^{-1}(B)\omega_1(B)x_{1,t-b_1} + \delta_2^{-1}(B)\omega_2(B)x_{2,t-b_2} + n_t$$

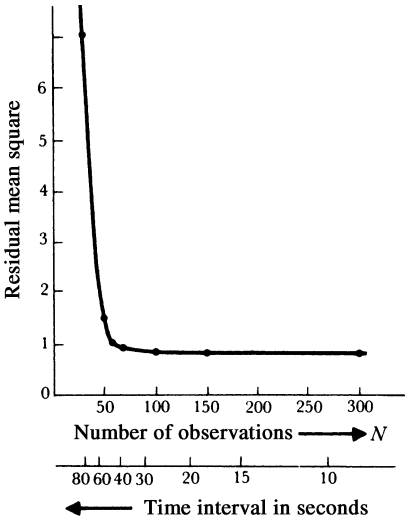


FIGURE 12.8 Mean square error at the output for various choices of sampling interval.

with

$$n_t = \phi^{-1}(B)\theta(B)a_t$$

where $y_t = \nabla^d Y_t$, $x_{1,t} = \nabla^d X_{1,t}$, $x_{2,t} = \nabla^d X_{2,t}$, and $n_t = \nabla^d N_t$ are stationary processes. To compute the a_t's, we first calculate for specific values of the parameters $b_1, \boldsymbol{\delta}_1, \boldsymbol{\omega}_1$,

$$y_{1,t} = \delta_1^{-1}(B)\omega_1(B)x_{1,t-b_1} \tag{12.4.3}$$

and for specific values of $b_2, \boldsymbol{\delta}_2, \boldsymbol{\omega}_2$,

$$y_{2,t} = \delta_2^{-1}(B)\omega_2(B)x_{2,t-b_2} \tag{12.4.4}$$

Then the noise n_t can be calculated from

$$n_t = y_t - y_{1,t} - y_{2,t} \tag{12.4.5}$$

and finally, a_t from

$$a_t = \theta^{-1}(B)\phi(B)n_t \tag{12.4.6}$$

Simulated Example. It is clear that even simple situations can lead to the estimation of a large number of parameters. The example below, with two input variables and delayed first-order models, has eight unknown parameters. To illustrate the behavior of the iterative nonlinear least-squares procedure described in Section 12.3.2 when used to obtain estimates of the parameters in such models, an experiment was performed using manufactured data, details of which are given in Box et al. (1967b). The data were generated from the model written in ∇ form as

$$Y_t = \beta + g_1 \frac{1 + \eta_1 \nabla}{1 + \xi_1 \nabla} X_{1,t-1} + g_2 \frac{1 + \eta_2 \nabla}{1 + \xi_2 \nabla} X_{2,t-1} + \frac{1}{1 - \phi_1 B} a_t \tag{12.4.7}$$

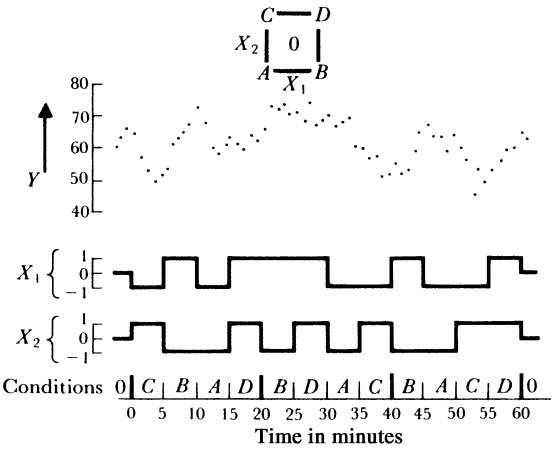


FIGURE 12.9 Data for simulated two-input example (Series K).

with $\beta = 60, g_1 = 13.0, \eta_1 = -0.6, \xi_1 = 4.0, g_2 = -5.5, \eta_2 = -0.6, \xi_2 = 4.0, \phi_1 = 0.5$, and $\sigma_a^2 = 9.0$. The input variables X_1 and X_2 were changed according to a randomized 2^2 factorial design replicated three times. Each input condition was supposed to be held fixed for 5 minutes and output observations taken every minute. The data are plotted in Figure 12.9 and appear as Series K in the Collection of Time Series section in Part Five.

The constrained iterative nonlinear least-squares program, described in Chapter 7, was used to obtain the least-squares estimates, so that it was only necessary to set up the calculation of the a_t's. Thus, for specified values of the parameters $g_1, g_2, \xi_1, \xi_2, \eta_1$, and η_2, the values $y_{1,t}$ and $y_{2,t}$ can be obtained from

$$(1 + \xi_1 \nabla) y_{1,t} = g_1 (1 + \eta_1 \nabla) X_{1,t-1}$$
$$(1 + \xi_2 \nabla) y_{2,t} = g_2 (1 + \eta_2 \nabla) X_{2,t-1}$$

and can be used to calculate

$$N_t = Y_t - y_{1,t} - y_{2,t}$$

Finally, for a specified value of ϕ_1, a_t can be calculated from

$$a_t = N_t - \phi_1 N_{t-1}$$

It was assumed that the process inputs had been maintained at their center conditions for some time before the start of the experiment, so that $y_{1,t}, y_{2,t}$, and N_t may be computed from $t = 0$ onward and a_t from $t = 1$.

Two runs were made of the nonlinear least-squares procedure using two different sets of initial values. In the first, the parameters were chosen as representing what a person reasonably familiar with the process might guess for initial values. In the second, the starting value for β was chosen to be the sample mean $\overline{Y}$ of all observations and all other starting values were set equal to 0.1. Thus, the second run represents a much more extreme situation than would normally arise in practice. Convergence with the first set of initial values occurred after five iterations, while convergence with the second set occurred

after nine iterations. These results suggest that in realistic circumstances, multiple inputs can be handled without serious estimation difficulties.

12.5 FORECASTING WITH TRANSFER FUNCTION MODELS USING LEADING INDICATORS

Frequently, forecasts of a time series $Y_t, Y_{t-1}, \ldots$ may be considerably improved by using information coming from some associated series $X_t, X_{t-1}, \ldots$. This is particularly true if changes in Y tend to be *anticipated* by changes in X, in which case economists call X a ''leading indicator'' for Y.

To obtain an optimal forecast using information from both series Y_t and X_t, we first build a transfer function–noise model connecting the series Y_t and X_t in the manner already outlined. Suppose, using previous notations, that an adequate model is

$$Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + \varphi^{-1}(B)\theta(B)a_t \quad b \geq 0 \tag{12.5.1}$$

In general, the noise component of this model, which is assumed statistically independent of the input X_t, is nonstationary with $\varphi(B) = \phi(B)\nabla^d$, so that if $y_t = \nabla^d Y_t$ and $x_t = \nabla^d X_t$,

$$y_t = \delta^{-1}(B)\omega(B)X_{t-b} + \phi^{-1}(B)\theta(B)a_t$$

Also, we will assume that an adequate stochastic model for the input or leading series X_t is

$$X_t = \varphi_x^{-1}(B)\theta_x(B)\alpha_t \tag{12.5.2}$$

so that with $\varphi_x(B) = \phi_x(B)\nabla^d$,

$$x_t = \phi_x^{-1}(B)\theta_x(B)\alpha_t$$

12.5.1 Minimum Mean Square Error Forecast

Now (12.5.1) may be written as

$$Y_t = \upsilon(B)\alpha_t + \psi(B)a_t \tag{12.5.3}$$

with the a_t's and the α_t's statistically independent white noise, and

$$\upsilon(B) = \delta^{-1}(B)\omega(B)B^b\varphi_x^{-1}(B)\theta_x(B)$$

Arguing as in Section 5.1.1, suppose that the forecast $\hat{Y}_t(l)$ of Y_{t+l} made at origin t is of the form

$$\hat{Y}_t(l) = \sum_{j=0}^{\infty} \nu_{l+j}^0 \alpha_{t-j} + \sum_{j=0}^{\infty} \psi_{l+j}^0 a_{t-j}$$

Then

$$Y_{t+l} - \hat{Y}_t(l) = \sum_{i=0}^{l-1}(\nu_i \alpha_{t+l-i} + \psi_i a_{t+l-i}) + \sum_{j=0}^{\infty}[(\nu_{l+j} - \nu_{l+j}^0)\alpha_{t-j} + (\psi_{l+j} - \psi_{l+j}^0)a_{t-j}]$$

and

$$E[(Y_{t+l} - \hat{Y}_t(l))^2] = (\nu_0^2 + \nu_1^2 + \cdots + \nu_{l-1}^2)\sigma_\alpha^2 + (1 + \psi_1^2 + \cdots + \psi_{l-1}^2)\sigma_a^2 + \sum_{j=0}^{\infty}[(\nu_{l+j} - \nu_{l+j}^0)^2\sigma_\alpha^2 + (\psi_{l+j} - \psi_{l+j}^0)^2\sigma_a^2]$$

which is minimized only if $\upsilon_{l+j}^0 = \upsilon_{l+j}$ and $\psi_{l+j}^0 = \psi_{l+j}$ for $j = 0, 1, 2 \ldots$. Thus, the minimum mean square error forecast $\hat{Y}_t(l)$ of Y_{t+l} at origin t is given by the conditional expectation of Y_{t+l} at time t, based on the past history of information on *both* series Y_t and X_t through time t. Theoretically, this expectation is conditional on knowledge of the series from the infinite past up to the present origin t. As in Chapter 5, such results are of practical use because, usually, the forecasts depend appreciably only on *recent* past values of the series X_t and Y_t.

Computation of the Forecast. Now (12.5.1) may be written as

$$\varphi(B)\delta(B)Y_t = \varphi(B)\omega(B)X_{t-b} + \delta(B)\theta(B)a_t$$

which we will write as

$$\delta^*(B)Y_t = \omega^*(B)X_{t-b} + \theta^*(B)a_t$$

Then, using square brackets to denote conditional expectations at time t, and writing $p^* = p + d$, we have for the lead l forecast

$$\begin{aligned}\hat{Y}_t(l) = [Y_{t+l}] = {} & \delta_1^*[Y_{t+l-1}] + \cdots + \delta_{p^*+r}^*[Y_{t+l-p^*-r}] + \omega_0^*[X_{t+l-b}] \\ & - \cdots - \omega_{p^*+s}^*[X_{t+l-b-p^*-s}] + [a_{t+l}] - \theta_1^*[a_{t+l-1}] \\ & - \cdots - \theta_{q+r}^*[a_{t+l-q-r}]\end{aligned} \tag{12.5.4}$$

where

$$[Y_{t+j}] = \begin{cases} Y_{t+j} & j \le 0 \\ \hat{Y}_t(j) & j > 0 \end{cases}$$

$$[X_{t+j}] = \begin{cases} X_{t+j} & j \le 0 \\ \hat{X}_t(j) & j > 0 \end{cases} \tag{12.5.5}$$

$$[a_{t+j}] = \begin{cases} a_{t+j} & j \le 0 \\ 0 & j > 0 \end{cases}$$

and a_t is calculated from (12.5.1), which if $b \geq 1$ is equivalent to

$$a_t = Y_t - \hat{Y}_{t-1}(1)$$

Thus, by appropriate substitutions, the minimum mean square error forecast is readily computed directly using (12.5.4) and (12.5.5). The forecasts $\hat{X}_t(j)$ are obtained in the usual way (see Section 5.2) utilizing the univariate ARIMA model (12.5.2) for the input series X_t.

It is important to note that the conditional expectations in (12.5.4) and (12.5.5) are taken with respect to values in *both* series Y_t and X_t through time t, but because of the assumed independence between input X_t and noise N_t in (12.5.1), it follows in particular that we will have

$$\hat{X}_t(j) = E[X_{t+j}|X_t, X_{t-1}, \ldots, Y_t, Y_{t-1}, \ldots] = E[X_{t+j}|X_t, X_{t-1}, \ldots]$$

That is, given the past values of the input series X_t, the optimal forecasts of its future values depend only on the past X's and cannot be improved by the additional knowledge of the past Y's; hence, the optimal values $\hat{X}_t(j)$ can be obtained directly from the univariate model (12.5.2).

Variance of the Forecast Error. The v_j weights and the ψ_j weights of (12.5.3) may be obtained explicitly by equating coefficients in

$$\delta(B)\varphi_x(B)v(B) = \omega(B)\theta_x(B)B^b$$

and in

$$\varphi(B)\psi(B) = \theta(B)$$

The variance of the lead l forecast error is then given by

$$V(l) = E[(Y_{t+l} - \hat{Y}_t(l))^2] = \sigma_\alpha^2 \sum_{j=b}^{l-1} v_j^2 + \sigma_a^2 \sum_{j=0}^{l-1} \psi_j^2 \tag{12.5.6}$$

Forecasts as a Weighted Aggregate of Previous Observations. For any given example, it is instructive to consider precisely how the forecasts of future values of the series Y_t utilize the previous values of the X_t and Y_t series. We have seen in Section 5.3.3 how the forecasts may be written as linear aggregates of previous values of the series. Thus, for forecasts of the input or leading indicator, we could write

$$\hat{X}_t(l) = \sum_{j=1}^{\infty} \pi_j^{(l)} X_{t+1-j} \tag{12.5.7}$$

The weights $\pi_j^{(1)} = \pi_j$ arise when the model (12.5.2) is written in the infinite autoregressive form

$$\alpha_t = \theta_x^{-1}(B)\varphi_x(B)X_t = X_t - \pi_1 X_{t-1} - \pi_2 X_{t-2} - \cdots$$

and may thus be obtained by explicitly equating coefficients in

$$\varphi_x(B) = (1 - \pi_1 B - \pi_2 B^2 - \cdots)\theta_x(B)$$

Also, using (5.3.9),

$$\pi_j^{(l)} = \pi_{j+l-1} + \sum_{h=1}^{l-1} \pi_h \pi_j^{(l-h)} \tag{12.5.8}$$

In a similar way, we can write the transfer function model (12.5.1) in the form

$$a_t = Y_t - \sum_{j=1}^{\infty} P_j Y_{t-j} - \sum_{j=0}^{\infty} Q_j X_{t-j} \tag{12.5.9}$$

It should be noted that if the transfer function between the input or leading indicator series X_t and the output Y_t is such that $b > 0$, then $\nu_j = 0$ for $j < b$, and so $Q_0, Q_1, \ldots, Q_{b-1}$ in (12.5.9) will also be zero.

Now (12.5.9) may be written as

$$a_t = \left(1 - \sum_{j=1}^{\infty} P_j B^j\right) Y_t - \left(\sum_{j=0}^{\infty} Q_j B^j\right) X_t$$

Comparison with (12.5.1) shows that the P_j and Q_j weights may be obtained by equating coefficients in the expressions

$$\theta(B)\left(1 - \sum_{j=1}^{\infty} P_j B^j\right) = \varphi(B)$$

$$\theta(B)\delta(B)\left(\sum_{j=0}^{\infty} Q_j B^j\right) = \varphi(B)\omega(B)B^b$$

On substituting $t + l$ for t in (12.5.9), and taking conditional expectations at origin t, we have the lead l forecast in the form

$$\hat{Y}_t(l) = \sum_{j=1}^{\infty} P_j[Y_{t+l-j}] + \sum_{j=0}^{\infty} Q_j[X_{t+l-j}] \tag{12.5.10}$$

Now the lead 1 forecast is $\hat{Y}_t(1) = \Sigma_{j=1}^{\infty} P_j Y_{t+1-j} + Q_0[X_{t+1}] + \Sigma_{j=1}^{\infty} Q_j X_{t+1-j}$, which for $b > 0$ becomes

$$\hat{Y}_t(1) = \sum_{j=1}^{\infty} P_j Y_{t+1-j} + \sum_{j=1}^{\infty} Q_j X_{t+1-j}$$

Also, the quantities in square brackets in (12.5.10) are either known values of the X_t and Y_t series or forecasts that are linear functions of these known values.

Thus, we can write the lead l forecast in terms of the values of the series that have already occurred at time t in the form

$$\hat{Y}_t(l) = \sum_{j=1}^{\infty} P_j^{(l)} Y_{t+1-j} + \sum_{j=1}^{\infty} Q_j^{(l)} X_{t+1-j} \tag{12.5.11}$$

where, for $b > 0$, the coefficients $P_j^{(l)}$ and $Q_j^{(l)}$ may be computed recursively as follows:

$$\begin{aligned} P_j^{(1)} &= P_j \qquad Q_j^{(1)} = Q_j \\ P_j^{(l)} &= P_{j+l-1} + \sum_{h=1}^{l-1} P_h P_j^{(l-h)} \\ Q_j^{(l)} &= Q_{j+l-1} + \sum_{h=1}^{l-1} \left\{ P_h Q_j^{(l-h)} + Q_h \pi_j^{(l-h)} \right\} \end{aligned} \tag{12.5.12}$$

12.5.2 Forecast of CO_2 Output from Gas Furnace

For illustration, consider the gas furnace data shown in Figure 12.1. For this example, the fitted model (see Section 12.4.1) was

$$Y_t = \frac{-(0.53 + 0.37B + 0.51B^2}{1 - 0.57B} X_{t-3} + \frac{1}{1 - 1.53B + 0.63B^2} a_t$$

and $(1 - 1.97B + 1.37B^2 - 0.34B^3)X_t = \alpha_t$. The forecast function, written in the form (12.5.4), is thus

$$\begin{aligned} \hat{Y}_t(l) = [Y_{t+l}] = {} & 2.1[Y_{t+l-1}] - 1.5021[Y_{t+l-2}] + 0.3591[Y_{t+l-3}] \\ & - 0.53[X_{t+l-3}] + 0.4409[X_{t+l-4}] - 0.2778[X_{t+l-5}] \\ & + 0.5472[X_{t+l-6}] - 0.3213[X_{t+l-7}] \\ & + [a_{t+l}] - 0.57[a_{t+l-1}] \end{aligned}$$

Figure 12.10 shows the forecasts for lead times $l = 1, 2, \ldots, 12$ made at origin $t = 206$. The π_j, P_j, and Q_j weights for the model are given in Table 12.6.

Figure 12.10 shows the weights $P_j^{(5)}$ and $Q_j^{(5)}$ appropriate to the lead 5 forecast. The weights ν_i and ψ_i of (12.5.3) are listed in Table 12.7. Using estimates $\hat{\sigma}_\alpha^2 = 0.0353$ and $\hat{\sigma}_a^2 = 0.0561$, obtained in Sections 12.2.2 and 12.4.1, respectively, (12.5.6) may be employed to obtain variances of the forecast errors and the 50 and 95% probability limits shown in Figure 12.10.

To illustrate the advantages of using an input or leading indicator series X_t in forecasting, assume that only the Y_t series is available. The usual identification and fitting procedure

TABLE 12.6 π_j, P_j, and Q_j Weights for Gas Furnace Model

j	π_j	P_j	Q_j	j	π_j	P_j	Q_j
1	1.97	1.53	0	7	0	0	−0.07
2	−1.37	−0.63	0	8	0	0	−0.04
3	0.34	0	−0.53	9	0	0	−0.02
4	0	0	0.14	10	0	0	−0.01
5	0	0	−0.20	11	0	0	−0.01
6	0	0	0.43				

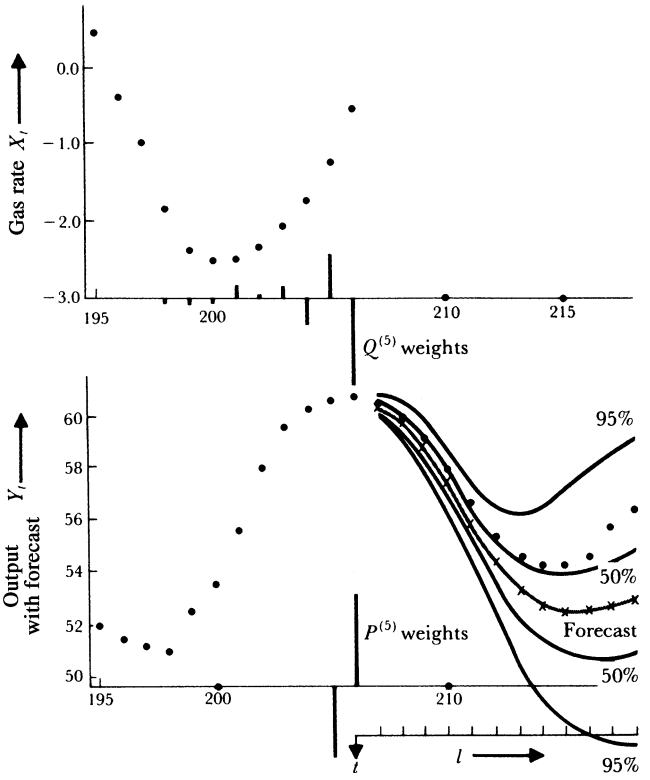


FIGURE 12.10 Forecast of CO_2 output from a gas furnace using input and output series.

TABLE 12.7 ν_i and ψ_i Weights for Gas Furnace Model

i	ν_i	ψ_i	i	ν_i	ψ_i
0	0	1	6	−5.33	0.89
1	0	1.53	7	−6.51	0.62
2	0	1.71	8	−6.89	0.39
3	−0.53	1.65	9	−6.57	0.20
4	−1.72	1.45	10	−5.77	0.06
5	−3.55	1.18	11	−4.73	−0.03

applied to this series indicated that it is well described by an ARMA(4, 2) process,

$$(1 - 2.42B + 2.388B^2 - 1.168B^3 + 0.23B^4)Y_t = (1 - 0.31B + 0.47B^2)\varepsilon_t$$

with $\sigma_\varepsilon^2 = 0.1081$. Table 12.8 shows estimated standard deviations of forecast errors made with and without the leading indicator series X_t. As might be expected, for short lead times use of the leading indicator can produce forecasts of considerably greater accuracy.

Univariate Modeling Check. To further confirm the univariate modeling results for the series Y_t, we can use results from Appendix A4.3 to obtain the nature of the univariate

TABLE 12.8 Estimated Standard Deviations of Forecast Errors Made With and Without the Leading Indicator for Gas Furnace Data

l	With Leading Indicator	Without Leading Indicator	l	With Leading Indicator	Without Leading Indicator
1	0.23	0.33	7	1.52	2.74
2	0.43	0.77	8	1.96	2.86
3	0.59	1.30	9	2.35	2.95
4	0.72	1.82	10	2.65	3.01
5	0.86	2.24	11	2.87	3.05
6	1.12	2.54	12	3.00	3.08

ARIMA model for Y_t that is implied by the transfer function–noise model between Y_t and X_t and the univariate AR(3) model for X_t. These models imply that

$$\begin{aligned}&(1-0.57B)(1-1.53B+0.63B^2)Y_t\\&=-(0.53+0.37B+0.51B^2)(1-1.53B+0.63B^2)X_{t-3}\\&\quad+(1-0.57B)a_t\end{aligned}\tag{12.5.13}$$

But since

$$\varphi_x(B)=1-1.97B+1.37B^2-0.34B^3\simeq(1-1.46B+0.60B^2)(1-0.52B)$$

in the AR(3) model for X_t, the right-hand side of (12.5.13) reduces approximately to $-(0.53+0.37B+0.51B^2)(1-0.52B)^{-1}\alpha_{t-3}+(1-0.57B)a_t$, and hence we obtain

$$\begin{aligned}&(1-0.52B)(1-0.57B)(1-1.53B+0.63B^2)Y_t\\&\quad=-(0.53+0.37B+0.51B^2)\alpha_{t-3}+(1-0.52B)(1-0.57B)a_t\end{aligned}$$

The results of Appendix A4.3 imply that the right-hand side of this last equation has an MA(2) model representation as $(1-\theta_1B-\theta_2B^2)\epsilon_t$, and the nonzero autocovariances of the MA(2) are determined from the right-hand side expression above to be

$$\lambda_0=0.1516\qquad\lambda_1=-0.0657\qquad\lambda_2=0.0262$$

Hence, the implied univariate model for Y_t would be ARMA(4, 2), with approximate AR operator equal to $(1-2.62B+2.59B^2-1.14B^3+0.19B^4)$, and from methods of Appendix A6.2, the MA(2) operator would be $(1-0.44B+0.21B^2)$, with $\sigma_\epsilon^2=0.1220$; that is, the univariate model for Y_t would be

$$(1-2.62B+2.59B^2-1.14B^3+0.19B^4)Y_t=(1-0.44B+0.21B^2)\epsilon_t$$

This model result is in good agreement with the univariate model actually identified and fitted to the series Y_t, which gives an additional check and provides further support to the transfer function–noise model that has been specified for the gas furnace data.

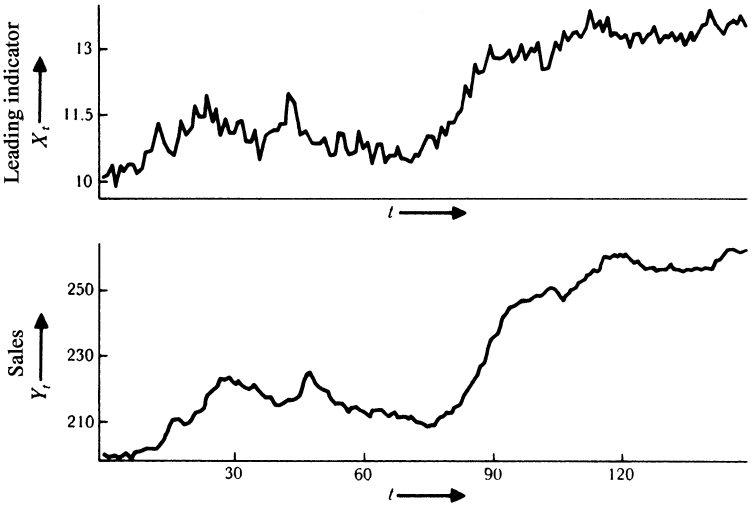


FIGURE 12.11 Sales data with leading indicator.

12.5.3 Forecast of Nonstationary Sales Data Using a Leading Indicator

As a second illustration, consider the data on sales Y_t in relation to a leading indicator X_t, plotted in Figure 12.11 and listed as Series M in the Collection of Time Series section in Part Five. The data are typical of that arising in business forecasting and are well fitted by the nonstationary model[3]

$$y_t = 0.035 + \frac{4.82}{1 - 0.72B} x_{t-3} + (1 - 0.54B)a_t$$
$$x_t = (1 - 0.32B)\alpha_t$$

with y_t and x_t first differences of the series. The forecast function, in the form (12.54), is then

$$\begin{aligned}\hat{Y}_t(l) = [Y_{t+l}] = 1.72[Y_{t+l-1}] - 0.72[Y_{t+l-2}] + 0.0098 + 4.82[X_{t+l-3}] \\ - 4.82[X_{t+l-4}] + [a_{t+l}] - 1.26[a_{t+l-1}] \\ + 0.3888[a_{t+l-2}]\end{aligned}$$

Figure 12.12 shows the forecasts for lead times $l = 1, 2, \ldots, 12$ made at origin $t = 89$. The weights ν_j and ψ_j are given in Table 12.9.

Using the estimates $\hat{\sigma}_\alpha^2 = 0.0676$ and $\hat{\sigma}_a^2 = 0.0484$, obtained in fitting the above model, the variance of the forecast error may be found from (12.5.6). In particular, $V(l) = \sigma_a^2 \Sigma_{j=0}^{l-1} \psi_j^2$ for $l = 1, 2$, and 3 in this specific case (note the delay of $b = 3$ in the transfer function model). The 50 and 95% probability limits are shown in Figure 12.12. It will be seen that in this particular example, the use of the leading indicator allows very accurate forecasts to be obtained for lead times $l = 1, 2$, and 3.

The π_j, P_j, and Q_j weights for this model are given in Table 12.10. The weights $p_j^{(5)}$ and $Q_j^{(5)}$ appropriate to the lead 5 forecast are shown in Figure 12.12.

[3]Using data the latter part of which is listed as Series M.

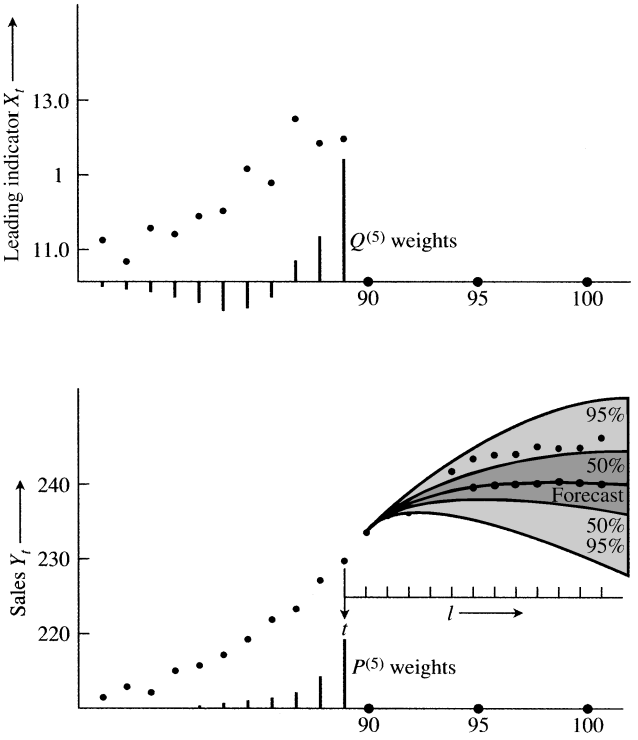


FIGURE 12.12 Forecast of sales at origin $t = 89$ with P and Q weights for lead 5 forecast.

TABLE 12.9 v_j and ψ_j Weights for Nonstationary Model for Sales Data

j	v_j	ψ_j	j	v_j	ψ_j
0	0	1	6	9.14	0.46
1	0	0.46	7	9.86	0.46
2	0	0.46	8	10.37	0.46
3	4.82	0.46	9	10.75	0.46
4	6.75	0.46	10	11.02	0.46
5	8.14	0.46	11	11.21	0.46

12.6 SOME ASPECTS OF THE DESIGN OF EXPERIMENTS TO ESTIMATE TRANSFER FUNCTIONS

In some engineering applications, the form of the input X_t can be deliberately chosen so as to obtain good estimates of the parameters in the transfer function–noise model:

$$Y_t = \delta^{-1}(B)\omega(B)X_{t-b} + N_t$$

The estimation of the transfer function is equivalent to estimation of a dynamic ''regression'' model, and the methods that can be used are very similar to those used in ordinary

TABLE 12.10 π_j, P_j, and Q_j Weights for Nonstationary Model for Sales Data

j	π_j	P_j	Q_j	j	π_j	P_j	Q_j
1	0.68	0.46	0	9	0.00	0.00	−0.74
2	0.22	0.25	0	10	0.00	0.00	−0.59
3	0.07	0.13	4.82	11	0.00	0.00	−0.29
4	0.02	0.07	1.25	12	0.00	0.00	−0.13
5	0.01	0.04	−0.29	13	0.00	0.00	−0.06
6	0.00	0.02	−0.86	14	0.00	0.00	−0.02
7	0.00	0.01	−0.97	15	0.00	0.00	0.00
8	0.00	0.01	−0.89				

nondynamic regression. As might be expected, the same problems (see e.g. Box, 1966) face us.

As with static regression, it is very important to be clear on the objective of the investigation. In some situations, we want to answer the question: If the input X is merely observed (but not interfered with), what can this tell us of the present and future behavior of the output Y under *normal* conditions of process operation? In other situations, the appropriate question is: If the input X is *changed* in some specific way, what *change* will be induced in the present and future behavior of the output Y? The types of data we need to answer these two questions are different.

To answer the first question unambiguously, we must use data obtained by observing, *but not interfering with*, the normal operation of the system. In contrast, the second question can only be answered unambiguously from data in which *deliberate* changes have been induced into the input of the system; that is, the data must be specially generated by *a designed experiment*.

Clearly, if X is to be used as a control variable, that is, a variable that may be used to manipulate the output, we need to answer the second question. To understand how we can design experiments to obtain valid estimates of the parameters of a cause-and-effect relationship, it is necessary to examine the assumptions of the analysis.

A critical assumption is that the X_t's are distributed independently of the N_t's. When this assumption is violated, the following issues arise:

1. The estimates we obtain are, in general, not even consistent. Specifically, as the sample size is made large, the estimates converge not on the true values but on other values differing from the true values by an unknown amount.
2. The violation of this assumption is not detectable by examining the data. Therefore, the possibility that in any particular situation the independence assumption may not be true is a particularly disturbing one. The only way it is possible to guarantee its truth is by deliberately *designing* the experiment rather than using data that have simply ''happened.'' Specifically, we must deliberately generate and feed into the process an input X_t, which we know to be uncorrelated with N_t because we have generated it by some external random process.

The input X_t can, of course, be autocorrelated; it is necessary only that it should not be cross-correlated with N_t. To satisfy this requirement, we could, for example, draw a set of random variates α_t and use them to generate any desired input process $X_t = \psi_X(B)\alpha_t$.

Alternatively, we can choose a fixed ''design,'' for example, the factorial design used in Section 12.4.2, and randomize the order in which the runs are made. Appendix A12.2 contains a preliminary discussion of some elementary design problems, and it is sufficient to expose some of the difficulties in the practical selection of the ''optimal'' stochastic input. In particular, as is true in a wider context: (1) it is difficult to decide what is a sensible criterion for optimality, and (2) the choice of ''optimal'' input depends on the values of the unknown parameters that are to be optimally estimated. In general, a white noise input has distinct advantages in simplifying identification, and if nothing very definite were known about the system under study, it would provide a sensible initial choice of input.

APPENDIX A12.1 USE OF CROSS-SPECTRAL ANALYSIS FOR TRANSFER FUNCTION MODEL IDENTIFICATION

In this appendix, we show that an alternative method for identifying transfer function models, which does not require prewhitening of the input, can be based on spectral analysis. Furthermore, it is easily generalized to multiple inputs.

A12.1.1 Identification of Single-Input Transfer Function Models

Suppose that the transfer function $v(B)$ is *defined* so as to allow the possibility of nonzero impulse response weights v_j for j a negative integer, so that

$$v(B) = \sum_{k=-\infty}^{\infty} v_k B^k$$

Then if, corresponding to (12.2.3), the transfer function–noise model is

$$y_t = v(B)x_t + n_t$$

equations (12.2.5) become

$$\gamma_{xy}(k) = \sum_{j=-\infty}^{\infty} v_j \gamma_{xx}(k-j) \qquad k = 0, \pm 1, \pm 2, \ldots \tag{A12.1.1}$$

We now define a *cross-covariance generating function*

$$\gamma^{xy}(B) = \sum_{k=-\infty}^{\infty} \gamma_{xy}(k) B^k \tag{A12.1.2}$$

which is analogous to the autocovariance generating function (3.1.10). On multiplying throughout in (A12.1.1) by B^k and summing, we obtain

$$\gamma^{xy}(B) = v(B)\gamma^{xx}(B) \tag{A12.1.3}$$

If we now substitute $B = e^{-i2\pi f}$ in (A12.1.2), we obtain the cross-spectrum $p_{xy}(f)$ between input x_t and output y_t. Making the same substitution in (A12.1.3) yields

$$v(e^{-i2\pi f}) = \frac{p_{xy(f)}}{p_{xx}(f)} \qquad -\frac{1}{2} \le f < \frac{1}{2} \tag{A12.1.4}$$

where

$$v(e^{-i2\pi f}) = G(f)e^{i2\pi\phi(f)} = \sum_{k=-\infty}^{\infty} v_k e^{-i2\pi fk} \tag{A12.1.5}$$

is called the *frequency response function* of the system transfer function relationship and is the Fourier transform of the impulse response function. Since $v(e^{-i2\pi f})$ is complex valued, we write it as a product involving a *gain function* $G(f) =\mid v(e^{i2\pi f}) \mid$ and a *phase function* $\phi(f)$. Equation (A12.1.4) shows that the frequency response function is the ratio of the cross-spectrum to the input spectrum. Methods for estimating the frequency response function $v(e^{-i2\pi f})$ are described by Jenkins and Watts (1968). Knowing $v(e^{-i2\pi f})$, the impulse response function v_k can then be obtained from

$$v_k = \int_{-1/2}^{1/2} v(e^{-i2\pi f})e^{i2\pi fk}df \tag{A12.1.6}$$

Using a similar approach, the autocovariance generating function of the noise n_t is

$$\gamma^{nn}(B) = \gamma^{yy}(B) - \frac{\gamma^{xy}(B)\gamma^{xy}(F)}{\gamma^{xx}(B)} \tag{A12.1.7}$$

On substituting $B = e^{-i2\pi f}$ in (A12.1.7), we obtain the expression

$$p_{nn}(f) = p_{yy}(f)[1 - k_{xy}^2(f)] \tag{A12.1.8}$$

for the spectrum of the noise process, where

$$k_{xy}^2(f) = \frac{\mid p_{xy}(f) \mid^2}{p_{xx}(f)p_{yy}(f)}$$

and $k_{xy}(f)$ is the *coherency spectrum* between the series x_t and y_t. The coherency spectrum $k_{xy}(f)$ at each frequency f behaves like a correlation coefficient between the random components at frequency f in the spectral representations of x_t and y_t. Knowing the noise spectrum, the noise autocovariance function $\gamma_{nn}(k)$ may then be obtained from

$$\gamma_{nn}(k) = 2\int_0^{1/2} p_{nn}(f)\cos(2\pi fk)df$$

By substituting estimates of the spectra such as those described in Jenkins and Watts (1968), estimates of the impulse response weights v_k and noise autocorrelation function are obtained. These can be used to identify the transfer function model and noise model as described in Sections 12.2.1 and 6.2.1.

A12.1.2 Identification of Multiple-Input Transfer Function Models

We now generalize the model

$$\begin{aligned} Y_t &= v(B)X_t + N_t \\ &= \delta^{-1}(B)\omega(B)X_{t-b} + N_t \end{aligned}$$

to allow for several inputs $X_{1,t}, X_{2,t}, \ldots, X_{m,t}$. Thus,

$$Y_t = \upsilon_1(B)X_{1,t} + \cdots + \nu_m(B)X_{m,t} + N_t \tag{A12.1.9}$$

$$= \delta_1^{-1}(B)\omega_1(B)X_{1,t-b_1} + \cdots + \delta_m^{-1}(B)\omega_m(B)X_{m,t-b_m} + N_t \tag{A12.1.10}$$

where $\upsilon_j(B) = \delta_j^{-1}(B)\omega_j(B)B^{b_j}$ is the generating function of the impulse response weights relating $X_{j,t}$ to the output Y_t. We assume, as before, that after differencing, (A12.1.9) may be written as

$$y_t = \upsilon_1(B)x_{1,t} + \cdots + \upsilon_m(B)x_{m,t} + n_t$$

where $y_t, x_{1,t}, \ldots, x_{m,t}$, and n_t are all jointly stationary processes. Multiplying throughout by $x_{1,t-k}, x_{2,t-k}, \ldots, x_{m,t-k}$ in turn, taking expectations, and forming the generating functions, we obtain

$$\begin{aligned}
\gamma^{x_1y}(B) &= \upsilon_1(B)\gamma^{x_1x_1}(B) + \upsilon_2(B)\gamma^{x_1x_2}(B) + \cdots + \upsilon_m(B)\gamma^{x_1x_m}(B) \\
\gamma^{x_2y}(B) &= \upsilon_1(B)\gamma^{x_2x_1}(B) + \upsilon_2(B)\gamma^{x_2x_2}(B) + \cdots + \upsilon_m(B)\gamma^{x_2x_m}(B) \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
\gamma^{x_my}(B) &= \upsilon_1(B)\gamma^{x_mx_1}(B) + \upsilon_2(B)\gamma^{x_mx_2}(B) + \cdots + \upsilon_m(B)\gamma^{x_mx_m}(B)
\end{aligned} \tag{A12.1.11}$$

On substituting $B = e^{-i2\pi f}$, the spectral equations are obtained. For example, with $m = 2$,

$$p_{x_1y}(f) = H_1(f)p_{x_1x_1}(f) + H_2(f)p_{x_1x_2}(f)$$
$$p_{x_2y}(f) = H_1(f)p_{x_2x_1}(f) + H_2(f)p_{x_2x_2}(f)$$

and the frequency response functions $H_1(f) = \upsilon_1(e^{-i2\pi f})$ and $H_2(f) = \upsilon_2(e^{-i2\pi f})$ can be calculated as described in Jenkins and Watts (1968). The impulse response weights can then be obtained using the inverse transformation (A12.1.6).

APPENDIX A12.2 CHOICE OF INPUT TO PROVIDE OPTIMAL PARAMETER ESTIMATES

Suppose that the input to a dynamic system can be made to follow an imposed stochastic process that is our choice. For example, it might be an autoregressive process, a moving average process, or white noise. To illustrate the problems involved in the optimal selection of this stochastic process, it is sufficient to consider an elementary example.

A12.2.1 Design of Optimal Inputs for a Simple System

Suppose that a system is under study for which the transfer function–noise model is assumed to be

$$Y_t = \beta_1 Y_{t-1} + \beta_2 X_{t-1} + a_t \qquad |\beta_1| < 1 \tag{A12.2.1}$$

where a_t is white noise. It is also assumed that the input and output processes are stationary and that X_t and Y_t denote deviations of these processes from their respective means. For large samples, and associated with any fixed probability, the approximate area of the Bayesian HPD region for β_1 and β_2, and also of the corresponding confidence region, is

proportional to $\Delta^{-1/2}$, where Δ is the determinant

$$\Delta = \begin{vmatrix} E[Y_t^2] & E[Y_t X_t] \\ E[Y_t X_t] & E[X_t^2] \end{vmatrix}$$

We will proceed by attempting to find the design minimizing the area of the HPD or confidence region and thus maximizing Δ. Now

$$\begin{aligned} E[Y_t^2] &= \sigma_Y^2 = \sigma_X^2 \beta_2^2 \frac{1+2q}{1-\beta_1^2} + \frac{\sigma_a^2}{1-\beta_1^2} \\ E[Y_t X_t] &= \sigma_X^2 \frac{\beta_2}{\beta_1} q \\ E[X_t^2] &= \sigma_X^2 \end{aligned} \tag{A12.2.2}$$

where

$$q = \sum_{i=1}^{\infty} \beta_1^i \rho_i \qquad \sigma_X^2 \rho_i = E[X_t X_{t-i}]$$

The value of the determinant may be written in terms of σ_X^2 as

$$\Delta = \frac{\sigma_X^2 \sigma_a^2}{1-\beta_1^2} + \frac{\beta_2^2 \sigma_X^4}{(1-\beta_1^2)^2} - \frac{\sigma_X^4 \beta_2^2}{\beta_1^2} \left(q - \frac{\beta_1^2}{1-\beta_1^2} \right)^2 \tag{A12.2.3}$$

Thus, as might be expected, the area of the region can be made small by making σ_X^2 large (i.e., by varying the input variable over a wide range). In practice, there may be limits to the amount of variation that can be allowed in X. Let us proceed by first supposing that σ_X^2 is held fixed at some specified value.

Solution with σ_X^2 Fixed. With $(1-\beta_1^2) > 0$ and for any fixed σ_X^2, we see from (A12.2.3) that Δ is maximized by setting

$$q = \frac{\beta_1^2}{1-\beta_1^2}$$

that is,

$$\beta_1 \rho_1 + \beta_1^2 \rho_2 + \beta_1^3 \rho_3 + \cdots = \beta_1^2 + \beta_1^4 + \beta_1^6 + \cdots$$

There are an infinite number of ways in which, for given β_1, this equality could be achieved. One obvious solution is

$$\rho_i = \beta_1^i$$

Thus, one way to maximize Δ for fixed σ_X^2 would be to force the input to follow the autoregressive process

$$(1 - \beta_1 B)X_t = \alpha_t$$

where α_t is a white noise process with variance $\sigma_\alpha^2 = \sigma_X^2(1 - \beta_1^2)$.

Solution with σ_Y^2 Fixed. So far we have supposed that σ_Y^2 is unrestricted. In some cases, we might wish to avoid too great a variation in the output rather than in the input. Suppose that σ_Y^2 is held equal to some fixed acceptable value but that σ_X^2 is unrestricted. Then the value of the determinant Δ can be written in terms of σ_Y^2 as

$$\Delta = \frac{\sigma_Y^4}{\beta_2^2}\left[\frac{\sigma_Y^2 - \sigma_a^2}{\sigma_Y^2} - \frac{\beta_1^2}{s^2}\left(\frac{q+s}{1+2q}\right)^2\right] \tag{A12.2.4}$$

where

$$s = \frac{\beta_1^2 r}{1 + \beta_1^2 r} \tag{A12.2.5}$$

and

$$r = \frac{\sigma_Y^2}{\sigma_Y^2 - \sigma_a^2} \tag{A12.2.6}$$

The maximum is achieved by setting

$$q = -s = \frac{-\beta_1^2 r}{1 + \beta_1^2 r} \tag{A12.2.7}$$

that is,

$$\beta_1\rho_1 + \beta_1^2\rho_2 + \beta_1^3\rho_3 + \cdots = -\beta_1^2 r + \beta_1^4 r^2 - \beta_1^6 r^3 + \cdots$$

There are again infinite ways of satisfying this equality. In particular, one solution is

$$\rho_i = (-\beta_1 r)^i \tag{A12.2.8}$$

which can be obtained by forcing the input to follow the autoregressive process

$$(1 + \beta_1 r B)X_t = \alpha_t \tag{A12.2.9}$$

where α_t is a white noise process with variance $\sigma_\alpha^2 = \sigma_X^2(1 - \beta_1^2 r^2)$. Since r is essentially positive, the sign of the parameter $(-\beta_1 r)$ of this autoregressive process is opposite to that obtained for the optimal input with σ_X^2 fixed.

Solution with $\sigma_Y^2 \times \sigma_X^2$ Fixed. In practice, it might happen that excessive variations in input and output were both to be avoided. If it were true that a given *percentage* decrease in the variance of X was equally as desirable as the same *percentage* decrease in the variance of Y, it would be sensible to maximize Δ subject to a fixed value of the product $\sigma_X^2 \times \sigma_Y^2$.

The determinant is

$$\Delta = \sigma_X^2 \sigma_Y^2 - \frac{\sigma_X^4 \beta_2^2 q^2}{\beta_1^2} \tag{A12.2.10}$$

which is maximized for fixed $\sigma_X^2 \sigma_Y^2$ only if $q = 0$. Once again there are an infinite number of solutions. However, by using a white noise input, Δ is maximized *whatever the value of* β_1. For such an input, using (A12.2.2), σ_X^2 is the positive root of

$$\sigma_X^4 \beta_2^2 + \sigma_X^2 \sigma_a^2 - k(1 - \beta_1^2) = 0 \tag{A12.2.11}$$

where $k = \sigma_X^2 \sigma_Y^2$, which is fixed.

A12.2.2 Numerical Example

Suppose that we were studying the first-order dynamic system (A12.2.1) with $\beta_1 = 0.50$ and $\beta_2 = 1.00$, so that

$$Y_t = 0.50 Y_{t-1} + 1.00 X_{t-1} + \alpha_t$$

where $\sigma_\alpha^2 = 0.2$.

σ_X^2 Fixed, σ_Y^2 Unrestricted. Suppose at first that the design is chosen to maximize Δ with $\sigma_X^2 = 1.0$. Then one optimal choice for the input X_t will be the autoregressive process

$$(1 - 0.5B) X_t = \alpha_t$$

where the white noise process α_t would have variance $\sigma_\alpha^2 = \sigma_X^2(1 - \beta_1^2) = 0.75$. Using (A12.2.2), the variance σ_Y^2 of the output would be 2.49, and the scheme will achieve a Bayesian region for β_1 and β_2 whose area is proportional to $\Delta^{-1/2} = 0.70$.

σ_Y^2 Fixed, σ_X^2 Unrestricted. The above scheme is optimal under the assumption that the input variance is $\sigma_X^2 = 1$ and the output variance is unrestricted. This output variance then turns out to be $\sigma_Y^2 = 2.49$. If, instead, the input variance were unrestricted, then with a *fixed* output variance of 2.49, we could, of course, do considerably better. In fact, using (A12.2.6), $r = 1.087$ and hence $\beta_1 r \simeq 0.54$, so that from (A12.2.9) one optimal choice for the unrestricted input would be the autoregressive process

$$(1 + 0.54B) X_t = \alpha_t$$

where in this case α_t is a white noise process with $\sigma_\alpha^2 = \sigma_X^2(1 - \beta_1^2 r^2)$. Using (A12.2.2) with $\sigma_Y^2 = 2.49$ fixed and $q = -0.214$ from (A12.2.7), the variance σ_X^2 of the input would now be increased to 2.91, so that $\sigma_\alpha^2 = 2.05$, and $\Delta^{-1/2}$, which measures the area of the Bayesian region, would be reduced to $\Delta^{-1/2} = 0.42$.

Product $\sigma_Y^2 \times \sigma_X^2$ Fixed. Finally, we consider a scheme that attempts to control both σ_Y^2 and σ_X^2 by maximizing Δ with $\sigma_Y^2 \times \sigma_X^2$ fixed. In the previous example in which σ_Y^2 was fixed, we found that $\Delta^{-1/2} = 0.42$ with $\sigma_X^2 = 2.91$ and $\sigma_Y^2 = 2.49$, so that the product is

$2.91 \times 2.49 = 7.25$. If our objective had been to minimize $\Delta^{-1/2}$ while keeping this product equal to 7.25, we could have made an optimal choice *without knowledge of* β_1 by choosing a white noise input $X_t = \alpha_t$. Using (A12.2.11), $\sigma_X^2 = \sigma_\alpha^2 = 2.29$, $\sigma_Y^2 = 3.16$, and in this case, as expected, $\Delta^{-1/2} = 0.37$, slightly smaller than that in the previous example.

It is worth considering this example in terms of spectral ideas. To optimize with σ_X^2 fixed, we have used an autoregressive input with ϕ_x positive that has high power at low frequencies. Since the gain of the system is high at low frequencies, this achieves maximum transfer from X to Y and so induces large variations in Y. When σ_Y^2 is fixed, we have introduced an input that is an autoregressive process with ϕ_x negative. This has high power at high frequencies. Since there is minimum transfer from X to Y at high frequencies, the disturbance in X must now be made large at these frequencies. When the product $\sigma_X^2 \times \sigma_Y^2$ is fixed, the ''compromise'' input white noise is indicated and does not require knowledge of β_1. This final maximization of Δ is equivalent to minimizing the (magnitude of the) correlation between the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, and in fact the correlation between these estimates is zero when a white noise input is used.

Conclusions. This investigation shows the following:

1. The optimal choice of design rests heavily on how we define ''optimal.''
2. Both in the case where α_X^2 is held fixed and in the case where α_Y^2 is held fixed, the optimal choices require specific stochastic processes for the input X_t whose parameters are functions of the *unknown* dynamic parameters. Thus, we are in the familiar paradoxical situation where we can do a better job of data gathering only to the extent that we already know something about the answer we seek. A sequential approach, where we improve the design as we find out more about the parameters, is a possibility worth further investigation. In particular, a pilot investigation using a possibly nonoptimal input, say white noise, could be used to generate data from which preliminary estimates of the dynamic parameters could be obtained. These estimates could then be used to specify a further input using one of our previous criteria.
3. The use of white noise is shown, *for the simple case investigated*, to be optimal for a sensible criterion of optimality, and its use as an input requires no prior knowledge of the parameters.

EXERCISES

12.1. Estimate of the cross-correlation function at lags $-1, 0$, and $+1$ for the following series of five pairs of observations:

t	1	2	3	4	5
x_t	11	7	8	12	14
y_t	7	10	6	7	10

12.2. If two series may be represented in ψ-weight form as

$$y_t = \psi_y(B)a_t \qquad x_t = \psi_x(B)a_t$$

(a) Show that their cross-covariance generating function

$$\gamma^{xy}(B) = \sum_{k=-\infty}^{\infty} \gamma_{xy}(k) B^k$$

is given by $\sigma_a^2 \psi_y(B)\psi_x(F)$.

(b) Use the above result to obtain the cross-covariance function between y_t and x_t when

$$y_t = (1 - \theta B)a_t \qquad x_t = (1 - \theta_1' B - \theta_2' B^2)a_t$$

12.3. After estimating a prewhitening transformation $\theta_x^{-1}(B)\phi_x(B)x_t = \alpha_t$ for an input series x_t and then computing the transformed output $\beta_t = \theta_x^{-1}(B)\phi_x(B)y_t$, cross-correlations $r_{\alpha\beta}(k)$ were obtained as follows:

k	$r_{\alpha\beta}(k)$	k	$r_{\alpha\beta}(k)$
0	0.05	5	0.24
1	0.31	6	0.07
2	0.52	7	−0.03
3	0.43	8	0.10
4	0.29	9	0.07

with $\hat{\sigma}_\alpha = 1.26$, $\hat{\sigma}_\beta = 2.73$, and $n = 187$.

(a) Obtain approximate standard errors for the cross-correlations.

(b) Calculate rough estimates for the impulse response weights v_j of a transfer function between y_t and x_t.

(c) Suggest a model form for the transfer function and give rough estimates of its parameters.

12.4. It is frequently the case that the user of an estimated transfer function–noise model $y_t = \delta^{-1}(B)\omega(B)B^b x_t + n_t$ will want to establish whether the steady-state gain $g = \delta^{-1}(1)\omega(1)$ makes good sense.

(a) For the first-order transfer function system

$$y_t = \frac{\omega_0}{1 - \delta B} x_{t-1}$$

show that an approximate standard error $\hat{\sigma}(\hat{g})$ of the estimate $\hat{g} = \hat{\omega}_0/(1 - \hat{\delta})$ is given by

$$\frac{\hat{\sigma}^2(\hat{g})}{\hat{g}^2} \simeq \frac{\text{var}[\hat{\omega}_0]}{\hat{\omega}_0^2} + \frac{\text{var}[\hat{\delta}]}{(1 - \hat{\delta})^2} + \frac{2\text{cov}[\hat{\omega}_0, \hat{\delta}]}{\hat{\omega}_0(1 - \hat{\delta})}$$

(b) Calculate $\hat{g}$ and an approximate value for $\hat{\sigma}(\hat{g})$ when $\hat{\omega}_0 = 5.2$, $\hat{\delta} = 0.65$, $\hat{\sigma}(\hat{\omega}_0) = 0.5$, $\hat{\sigma}(\hat{\delta}) = 0.1$, and $\text{cov}[\hat{\omega}_0, \hat{\delta}] = 0.025$.

12.5. Consider the regression model

$$Y_t = \beta_1 X_{1,t} + \beta_2 X_{2,t} + N_t$$

where N_t is a nonstationary error term following an IMA(0, 1, 1) process $\nabla N_t = a_t - \theta a_{t-1}$. Show that the regression model may be rewritten in the form

$$Y_t - \bar{Y}_{t-1} = \beta_1(X_{1,t} - \bar{X}_{1,t-1}) + \beta_2(X_{2,t} - \bar{X}_{2,t-1}) + a_t$$

where $\bar{Y}_{t-1}$, $\bar{X}_{1,t-1}$, and $\bar{X}_{2,t-1}$ are exponentially weighted moving averages so that, for example,

$$\bar{Y}_{t-1} = (1-\theta)(Y_{t-1} + \theta Y_{t-2} + \theta^2 Y_{t-3} + \cdots)$$

It will be seen that the fitting of this regression model with nonstationary noise by maximum likelihood is equivalent to fitting the *deviations* of the independent and dependent variables from *local updated exponentially weighted moving averages* by ordinary least-squares. (Refer to Section 9.5.1 for related ideas regarding transformation of regression models with autocorrelated noise N_t.)

12.6. Quarterly measurements of unemployment and the gross domestic product (GDP) in the United Kingdom over the period 1955–1969 are included in Series P in Part Five of this book; see also http://pages.stat.wisc.edu/ reinsel/bjr-data/.

(a) Plot the two time series using R.

(b) Calculate and plot the autocorrelation and partial autocorrelation functions of the two series. Repeat the calculations for the first differences of the two series. Would a variance stabilizing transformation be helpful for model development?

(c) Calculate and plot the cross-correlation function between the two series.

12.7. Refer to Exercise 12.6. Build (identify, estimate, and check) a transfer function–noise model that uses the GDP series X_t as input to help explain variations in the logged unemployment series Y_t.

12.8. Consider the transfer function–noise model fitted to the gas furnace data in (12.4.1) and (12.4.2). Note that the estimate of δ_2 is very close to zero. Re-estimate the parameters of this model setting δ_2 equal to zero. Describe the resulting impact on the estimate of the residual variance and other model parameters.

12.9. A bivariate time series consisting of sales data and a leading indicator is listed as Series M in Part Five of this book. The series is also available as ''BJsales'' in the datasets package of R.

(a) Plot the two time series using R.

(b) Calculate and plot the autocorrelation and partial autocorrelation functions of the two series. Find a suitable model for the leading indicator series.

(c) Calculate and plot the cross-correlation function between the two variables.

(d) Calculate and plot the cross-correlation function after prewhitening the series using the time series model developed in part (b).

(e) Estimate the impulse response function v_k for the two series.

12.10. Refer to Exercise 12.9. A bivariate transfer function–noise model was given for these series in Section 12.5.3.

(a) Use the results from Exercise 12.9 to justify the choice of transfer function model. Derive preliminary estimates of the parameters in this model.

(b) Justify the choice of the noise model given in Section 12.5.3.

(c) Estimate the parameters of the combined transfer function–noise model and perform the appropriate diagnostic checks on the fitted model.

13

INTERVENTION ANALYSIS, OUTLIER DETECTION, AND MISSING VALUES

Time series are often affected by special events or conditions such as policy changes, strikes, advertising promotions, environmental regulations, and similar events, which we will refer to as *intervention* events. In Section 13.1, we describe the method of intervention analysis, which can account for the expected effects of these interventions. For this, the transfer function models of the previous chapters are used, but in the intervention analysis model, the input series will be in the form of a simple pulse or step indicator function to signal the presence or absence of the event. The timing of the intervention event is assumed to be known in this analysis. Section 13.2 considers the related problem of detecting outlying or unusual behavior in a time series at an unknown point of time. Depending on how the outlier enters and its likely impact on the time series, two types of outlier models, additive outlier (AO) and innovational outlier (IO) models, are considered. A somewhat related problem of missing values in a time series is discussed in Section 13.3. The key focus of this section is on parameter estimation and evaluation of the likelihood function of an ARMA model for time series with missing values. However, consideration is also given to estimation of the missing values in the series.

13.1 INTERVENTION ANALYSIS METHODS

13.1.1 Models for Intervention Analysis

In the setting of intervention analysis, it is assumed that an intervention event has occurred at a known point in time T of a time series. It is of interest to determine whether there is

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

any evidence of a change or effect, of an expected kind, on the time series Y_t associated with the event. We consider the use of transfer function models to model the nature of and estimate the magnitude of the effects of the intervention, and hence to account for the possible unusual behavior in the time series related to the event. Based on the study by Box and Tiao (1975), the type of model we consider has the form

$$Y_t = \frac{\omega(B)B^b}{\delta(B)}\xi_t + N_t \tag{13.1.1}$$

where the term $\mathcal{Y}_t = \delta^{-1}(B)\omega(B)B^b\xi_t$ represents the effects of the intervention event in terms of the deterministic input series ξ_t, and N_t is the noise series that represents the underlying time series without the intervention effects. It is assumed that N_t follows an ARIMA(p, d, q) model, $\varphi(B)N_t = \theta(B)a_t$, with $\varphi(B) = \phi(B)(1 - B)^d$. Multiplicative seasonal ARIMA models as presented in Chapter 9 can also be included for N_t, but special note of the seasonal models will not be made in this chapter.

There are two common types of deterministic input variables ξ_t that have been found useful to represent the impact of intervention events on a time series. Both of these are indicator variables taking only the values 0 and 1 to denote the nonoccurrence and occurrence of the intervention. One type is a *step function* at time T, given by

$$S_t^{(T)} = \begin{cases} 0 & t < T \\ 1 & t \geq T \end{cases} \tag{13.1.2}$$

which would typically be used to represent the effects of an intervention that are expected to remain permanently after time T to some extent. The other type is a *pulse function* at T, given by

$$P_t^{(T)} = \begin{cases} 0 & t \neq T \\ 1 & t = T \end{cases} \tag{13.1.3}$$

which could represent the effects of an intervention that are temporary or transient and will die out after time T. These indicator input variables are used in many situations where the effects of the intervention cannot be represented as the response to a quantitative variable because such a quantitative variable does not exist or it is impractical or impossible to obtain measurements on such a variable.

Because of the deterministic nature of the indicator input series ξ_t in (13.1.1), unlike the transfer function model situation of Chapter 12, identification of the structure of the intervention model operator $v(B) = \delta^{-1}(B)\omega(B)B^b$ cannot be based on the technique of prewhitening. Instead, it is necessary to postulate the form of the intervention model by considering the mechanisms that might cause the change or effect and the implied form of the change that would be expected. In addition, the identification may be aided by direct inspection of the data to suggest the form of effect due to the known event, and supplementary evidence may sometimes be available from examination of the residuals from a model fitted before the intervention term is introduced.

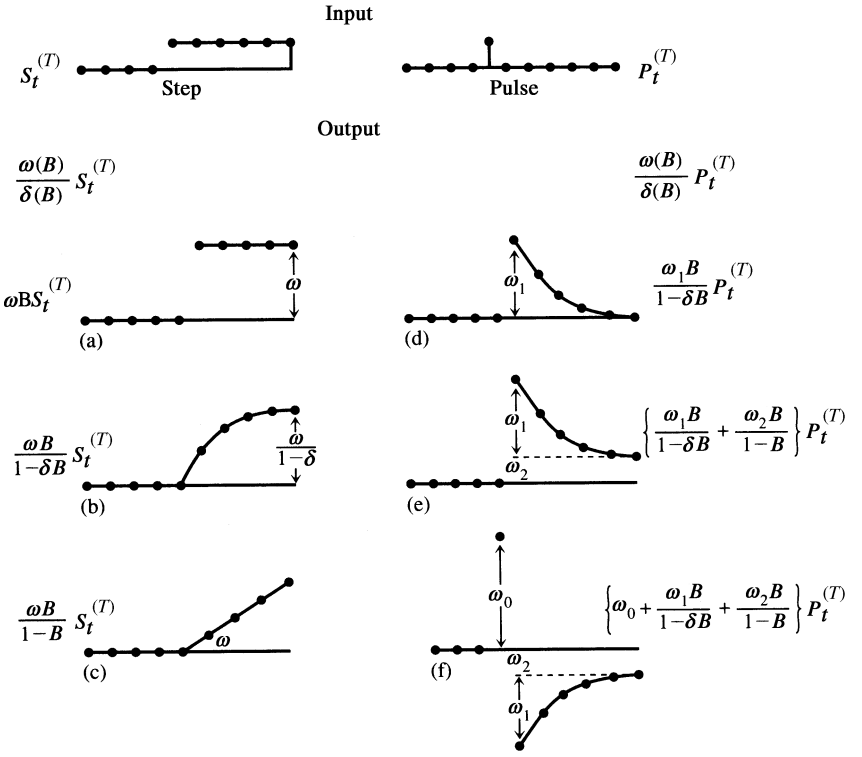


FIGURE 13.1 Responses to a step and a pulse input: (a–c) Response to a step input for various simple transfer function models, and (d–f) Response to a pulse input for some common models of interest.

Response Patterns Useful in Intervention Analysis. Several different response patterns

$$\mathcal{Y}_t = \delta^{-1}(B)\omega(B)B^b\xi_t$$

are possible through different choices of the transfer function. Figure 13.1 shows the responses for various simple transfer functions with both step and pulse indicators as input. For example, the model $\mathcal{Y}_t = \omega B S_t^{(T)}$ in Figure 13.1(a) can be used to represent a permanent step change in level of unknown magnitude ω after time T, while the form

$$\mathcal{Y}_t = \frac{\omega B}{1-\delta B} S_t^{(T)} \qquad 0 < \delta < 1 \tag{13.1.4}$$

in Figure 13.1(b), which implies that $\mathcal{Y}_t = \omega(1-\delta^{t-T})/(1-\delta)$, $t \geq T$, corresponds to a gradual change with rate δ that eventually approaches the long-run change in level equal to $\omega/(1-\delta)$. Similarly, the model

$$\mathcal{Y}_t = \frac{\omega_1 B}{1-\delta B} P_t^{(T)} \quad 0 < \delta < 1 \tag{13.1.5}$$

in Figure 13.1(d), which implies that $\mathcal{Y}_t = \omega_1 \delta^{t-T-1}, t > T$, would represent a sudden ''pulse'' change after time T of unknown magnitude ω_1, followed by a gradual decay of

rate δ back to the original preintervention level with no permanent effect. More complex response patterns can be obtained by various linear combinations of the simpler forms, such as in the case of Figure 13.1(f). It is also noted that since $(1 - B)S_t^{(T)} = P_t^{(T)}$, any of the transfer function models that involve $S_t^{(T)}$ could equally well be represented in terms $P_t^{(T)}$.

The following additional points concerning the intervention models are worthy of note. The function $\mathcal{Y}_t$ represents the additional effect of the intervention event over the noise or ''background'' series N_t. Hence, when possible, the model $N_t = [\theta(B)/\varphi(B)]a_t$ for the noise is identified based on the usual procedures applied to the time series observations available before the date of the intervention, that is, Y_t, $t < T$. Also, it is assumed in model (13.1.1) that only the level of the series is affected by the intervention and, in particular, that the form and the parameters of the time series model for N_t are the same before and after the intervention. One should also recognize that there can be considerable differences in the accuracy with which the intervention model parameters can be estimated depending on whether the noise N_t is stationary or nonstationary, as well as on whether permanent or transitory effects are postulated.

In general, the parameter estimates and their standard errors for the intervention model

$$Y_t = \frac{\omega(B)B^b}{\delta(B)}\xi_t + \frac{\theta(B)}{\varphi(B)}a_t \tag{13.1.6}$$

are obtained by the least-squares method of estimation for transfer function–noise models, as described in Section 12.3. Diagnostic checking based on the residuals $\hat{a}_t$ from the fitted model can also be performed using methods similar to those previously employed to assess the adequacy of a fitted model.

13.1.2 Example of Intervention Analysis

Box and Tiao (1975) considered the monthly time series consisting of the rate of change in the U.S. consumer price index (CPI) for the period July 1953 through December 1972. Beginning in September 1971, phase I economic control went into effect for 3 months, and after that phase II was in effect. The problem was to investigate the possible effect of the phase I and II controls on the rate of change in the CPI.

Inspection of the sample autocorrelation functions of the rate of change of the CPI and its first differences for the 218 monthly observations prior to phase I suggested a noise model of the form

$$(1 - B)N_t = (1 - \theta B)a_t \tag{13.1.7}$$

with maximum likelihood estimates $\hat{\theta} = 0.84$ and $\hat{\sigma}_a = 0.0019$. Examination of the residuals and their autocorrelations reveals no obvious inadequacies in this model.

Then, to address the question of the possible effects of phase I and II controls, it is assumed that phase I and II are expected to produce changes in the level of the rate of change of the CPI, and that the form of the noise model remains the same. Based on these assumptions, the appropriate model to assess the impact of the controls is

$$Y_t = \omega_1\xi_{1t} + \omega_2\xi_{2t} + \frac{1 - \theta B}{1 - B}a_t \tag{13.1.8}$$

where

$$\xi_{1t} = \begin{cases} 1 & t = \text{September, October, or November 1971} \\ 0 & \text{otherwise} \end{cases}$$

$$\xi_{2t} = \begin{cases} 1 & t \geq \text{December 1971} \\ 0 & \text{otherwise} \end{cases}$$

The nonlinear least-squares estimates of the parameters in model (13.1.8) were obtained, with standard errors in parentheses, as

$$\hat{\theta} = 0.85(0.05) \qquad \hat{\omega}_1 = -0.0022(0.0010) \qquad \hat{\omega}_2 = -0.0008(0.0009)$$

Hence, the analysis suggests that a drop in the rate of increase of the CPI is associated with phase I, but the effect of phase II is much less certain.

Many other examples of the use of intervention analysis have appeared in the literature. These include studies of the effects of regulations for engine design changes in new cars on oxidant pollution levels in the Los Angeles area (Box and Tiao, 1975), the effect of a change in policy in relation to debt collection on bad debt collections (Jenkins, 1979), the effectiveness of seat belt legislation on road deaths (Bhattacharyya and Layton, 1979), and the impact of the Arab oil embargo on electricity consumption in the United States (Montgomery and Weatherby, 1980).

13.1.3 Nature of the ML Estimator for a Simple Level Change Model

It is instructive to consider the nature of the maximum likelihood estimator of the intervention parameters, such as those in (13.1.8), for some relatively simple situations. We consider the simple model

$$Y_t = \omega \xi_t + N_t \tag{13.1.9}$$

where $N_t = \phi^{-1}(B)\theta(B)a_t$. This model can be written, formally, as

$$\pi(B)Y_t = \omega \pi(B)\xi_t + a_t \tag{13.1.10}$$

where $\pi(B) = \theta^{-1}(B)\phi(B) = 1 - \sum_{i=1}^{\infty} \pi_i B^i$. Letting $w_t = \pi(B)Y_t$ and $x_t = \pi(B)\xi_t$, we can write (13.1.10) in the form of a simple linear model $w_t = \omega x_t + a_t$, $t = 1, 2, \ldots, n$. Hence, the maximum likelihood estimator of ω is approximately

$$\hat{\omega} = \frac{\sum_{t=1}^{n} x_t w_t}{\sum_{t=1}^{n} x_t^2} \tag{13.1.11}$$

with $\text{var}[\hat{\omega}] = \sigma_a^2 / \sum_{t=1}^{n} x_t^2$.

Example with a Step Change Input and Nonstationary Noise. Let us consider a special case of (13.1.9) where $\xi_t = BS_t^{(T)}$ represents a step change after time T. Then, $x_t = \pi(B)BS_t^{(T)} = 1 - \sum_{i=1}^{t-T-1} \pi_i$, $t > T + 1$, with $x_{T+1} = 1$ and $x_t = 0$ for $t \leq T$. For the discussion that follows, we suppose that n is large, and that a relatively large number of observations are available before and after the intervention time T.

Now suppose that the noise N_t in (13.1.9) is nonstationary with generalized autoregressive operator $\varphi(B) = \phi(B)(1 - B)$ so that

$$\pi(B) = \theta^{-1}(B)\phi(B)(1 - B) = \tilde{\pi}(B)(1 - B)$$

with $\tilde{\pi}(B) = \theta^{-1}(B)\phi(B) = 1 - \sum_{j=1}^{\infty} \tilde{\pi}_j B^j$. Then, $x_t = \tilde{\pi}(B)B(1 - B)S_t^{(T)} = \tilde{\pi}(B)P_{t-1}^{(T)} = \tilde{\pi}_{t-T-1}$, $t \geq T + 1$, and hence

$$\sum_{t=1}^{n} x_t^2 = \sum_{t=T+1}^{n} \tilde{\pi}_{t-T-1}^2 \simeq \sum_{i=0}^{\infty} \tilde{\pi}_i^2 \equiv \eta_0$$

Also, $w_t = \pi(B)Y_t = Y_t - \overline{Y}_{t-1}$, where $\overline{Y}_{t-1} = \sum_{i=1}^{\infty} \pi_i Y_{t-i}$ is a weighted average of values prior to t (since $\sum_{i=1}^{\infty} \pi_i = 1$ when $d = 1$). Following the results in Box and Tiao (1975), it can then be shown that

$$\begin{aligned}\sum_{t=1}^{n} x_t w_t = \sum_{t=T+1}^{n} \tilde{\pi}_{t-T-1} w_t &\simeq \tilde{\pi}(B)\tilde{\pi}(F)(1 - B)Y_{T+1} \\ &= \sum_{s=0}^{\infty} \alpha_s Y_{T+1+s} - \sum_{s=0}^{\infty} \alpha_s Y_{T-s}\end{aligned}$$

where $a_s = \eta_s - \eta_{s+1}$ and the η_s are coefficients in $\tilde{\pi}(B)\tilde{\pi}(F) = \eta_0 + \sum_{s=1}^{\infty} \eta_s(B^s + F^s)$, such that $\sum_{s=0}^{\infty} \alpha_s = \eta_0 \equiv \sum_{i=0}^{\infty} \tilde{\pi}_i^2$. Therefore, in this situation, the maximum likelihood estimator of ω is

$$\hat{\omega} = \frac{\sum_{t=1}^{n} x_t w_t}{\sum_{t=1}^{n} x_t^2} \simeq (\eta_0)^{-1} \left(\sum_{s=0}^{\infty} \alpha_s Y_{T+1+s} - \sum_{s=0}^{\infty} \alpha_s Y_{T-s} \right) \tag{13.1.12}$$

with $\text{var}[\hat{\omega}] \simeq \sigma_a^2(\eta_0)^{-1}$. The estimator $\hat{\omega}$ can thus be interpreted as a contrast between two weighted moving averages, one consisting of the observations after the intervention and the other for the observations before the intervention, where the weights (α_s/η_0) are symmetrical.

For example, consider the case where N_t follows the IMA(0, 1, 1) model, $(1 - B)N_t = (1 - \theta B)a_t$, so that $\tilde{\pi}(B) = (1 - \theta B)^{-1}$ with $\tilde{\pi} = \theta^i$, $i \geq 1$. Then, $\eta_s = \theta^s/(1 - \theta^2)$, $s = 0, 1, \ldots$, and so $\alpha_s = (\theta^s - \theta^{s+1})/(1 - \theta^2) = \theta^s/(1 + \theta)$. Hence, the estimator in (13.1.12) becomes

$$\hat{\omega} \simeq (1 - \theta)^{-1} \left(\sum_{s=0}^{\infty} \theta^s Y_{T+1+s} - \sum_{s=0}^{\infty} \theta^s Y_{T-s} \right) \tag{13.1.12a}$$

with $\text{var}[\hat{\omega}] \simeq \sigma_a^2(1 - \theta^2)$. The estimator $\hat{\omega}$ is thus a contrast between two *exponentially* weighted moving averages, one consisting of the observations after the intervention and the other for the observations before the intervention.

Now, as a second case, suppose that the noise instead follows the ARIMA(1, 1, 0) model, so that $\tilde{\pi}(B) = (1 - \phi B)$ with $\tilde{\pi}_1 = -\phi$ and $\tilde{\pi}_i = 0$ for $i > 1$. Then $\eta_0 = 1 + \phi^2$, $\eta_1 = -\phi$, and $\eta_s = 0$, $s > 1$. Hence, $\sum_{t=1}^{n} x_t^2 = 1 + \phi^2 = \eta_0$, $\alpha_0 = 1 + \phi + \phi^2$, $\alpha_1 = -\phi$, and it

follows that

$$\sum_{t=1}^{n} x_t w_t = (1-\phi B)(1-\phi F)(1-B)Y_{T+1}$$
$$= [(1+\phi+\phi^2)Y_{T+1} - \phi Y_{T+2}] - [(1+\phi+\phi^2)Y_T - \phi Y_{T-1}]$$

Thus, for this case we have

$$\hat{\omega} = \frac{\sum_{t=1}^{n} x_t w_t}{\sum_{t=1}^{n} x_t^2}$$
$$= (1+\phi^2)^{-1}\{[(1+\phi+\phi^2)Y_{T+1} - \phi Y_{T+2}]$$
$$- [(1+\phi+\phi^2)Y_T - \phi Y_{T-1}]\} \tag{13.1.13}$$

with $\text{var}[\hat{w}] = \sigma_a^2/(1+\phi^2)$. Again, the estimator $\hat{w}$ can be viewed as a contrast between two weighted averages of the same form, one of the postintervention observations Y_{T+1} and Y_{T+2} and the other of the preintervention observations Y_T and Y_{T-1}, but the weighted averages are only finite in extent because the noise model contains only an AR factor $(1-\phi B)$ and no MA factor as in the previous case.

Comparison with a Case with Stationary Noise. Finally, we consider a simpler situation of model (13.1.9), in which the noise is *stationary*, for example, an AR(1) model $(1-\phi B)N_t = a_t$. In this situation we obtain $x_t = (1-\phi B)BS_t^{(T)} = 1-\phi$ for $t > T+1$ with $x_{T+1} = 1$ and $w_t = (1-\phi B)Y_t = Y_t - \phi Y_{t-1}$. Then, it readily follows that

$$\hat{\omega} = \frac{\sum_{t=1}^{n} x_t w_t}{\sum_{t=1}^{n} x_t^2}$$
$$\simeq \frac{(1-\phi)\sum_{t=T+1}^{n}(Y_t - \phi Y_{t-1})}{(n-T)(1-\phi)^2} \simeq \overline{Y}_2 \tag{13.1.14}$$

where $\overline{Y}_2 = (n-T)^{-1}\sum_{t=T+1}^{n} Y_t$ denotes an unweighted average of all observations after the intervention, with $\text{var}[\hat{\omega}] = \sigma_a^2/[1+(n-T-1)(1-\phi)^2] \simeq \sigma_a^2/[(n-T)(1-\phi)^2]$. Notice that because of the stationarity of the noise, we have an *unweighted* average of postintervention observations and also that there is no adjustment for the preintervention observations because they are assumed to be *stationary* about a *known mean of zero*. Also note that in the stationary case, the variance of $\hat{\omega}$ decreases proportionally with $1/(n-T)$, whereas in the previous nonstationary noise situations, $\text{var}[\hat{\omega}]$ is essentially a constant not dependent on the sample size. This reflects the differing degrees of accuracy in the estimators of intervention model parameters, such as the level shift parameter ω, that can be expected in large samples between the nonstationary noise and the stationary noise model situations.

Specifically, in the model (13.1.9), with $\xi_t = BS_t^{(T)}$ equal to a step input, suppose that the noise process N_t is nonstationary ARIMA with $d = 1$, so that $\phi(B)(1-B)N_t = \theta(B)a_t$. Then, by applying the differencing operator $(1-B)$, the model

$$Y_t = \omega BS_t^{(T)} + N_t \tag{13.1.15}$$

can also be expressed as

$$y_t = \omega B P_t^{(T)} + n_t \tag{13.1.16}$$

where $y_t = (1 - B)Y_t$ and $n_t = (1 - B)N_t$, and hence n_t is a stationary ARMA(p, q) process. Therefore, the MLE of ω for the original model (13.1.15) with a (permanent) *step* input effect and *nonstationary* noise ($d = 1$) will have features similar to the MLE in the model (13.1.16), which has a (transitory) *pulse* input effect and *stationary* noise.

Of course, the model (13.1.9) can be generalized to allow for an unknown nonzero mean ω_0 before the intervention, $Y_t = \omega_0 + \omega\xi_t + N_t$, with $\xi_t = BS_t^{(T)}$, so that ω represents the *change in mean level* after the intervention. Then, for the stationary AR(1) noise model case, for example, similar to (13.1.14), it can be shown that the MLE of ω is $\hat{\omega} \simeq \overline{Y}_2 - \overline{Y}_1$, where $\overline{Y}_1 = T^{-1}\sum_{t=1}^{T} Y_t$ denotes the sample mean of all preintervention observations.

13.2 OUTLIER ANALYSIS FOR TIME SERIES

Time series observations may sometimes be affected by isolated events, disturbances, or errors that create spurious effects in the series and result in unusual patterns in the observations that are not consistent with the overall behavior of the time series. Such unusual observations may be referred to as *outliers*. They may be the result of unusual external events such as strikes, sudden political or economic changes, unusual weather events, sudden changes in a physical system, and so on, or simply due to recording or gross errors in measurement. The presence of such outliers in a time series can have substantial effects on the behavior of sample autocorrelations, partial autocorrelations, estimates of ARMA model parameters, and forecasting, and can even affect the specification of the model. If the time of occurrence T of an event that results in the outlying behavior is known, the unusual effects can often be accounted for by the use of intervention analysis techniques discussed in Section 13.1. However, since in practice the presence of outliers is often not known at the start of the analysis, additional procedures for detection of outliers and assessment of their possible impacts are important. In this section we discuss some useful models for representing outliers and corresponding methods, similar to the methods of intervention analysis, for detection of outliers. Some relevant references that deal with the topics of outlier detection, influence of outliers, and robust methods of estimation include Bruce and Martin (1989), Chang et al. (1988), Chen and Liu (1993), Martin and Yohai (1986), and Tsay (1986).

13.2.1 Models for Additive and Innovational Outliers

Following the work of Fox (1972), we consider two simple intervention models to represent two different types of outliers that might occur in practice. These are the *additive outlier* (AO) and the *innovational outlier* (IO) models. Let z_t denote the underlying time series process that is free of the impact of outliers, and let Y_t denote the observed time series. We assume that z_t follows the ARIMA(p, d, q) model $\varphi(B)z_t = \theta(B)a_t$. Then, an additive outlier at time T, or ''observational outlier,'' is modeled as

$$Y_t = \omega P_t^{(T)} + z_t = \omega P_t^{(T)} + \frac{\theta(B)}{\varphi(B)}a_t \tag{13.2.1}$$

where $P_t^{(T)} = 1$ if $t = T$, $P_t^{(T)} = 0$ if $t \neq T$, denotes the pulse indicator at time T. An innovational outlier at time T, or ''innovational shock,'' is modeled as

$$Y_t = \frac{\theta(B)}{\varphi(B)}(\omega P_t^{(T)} + a_t) = \omega \frac{\theta(B)}{\varphi(B)} P_t^{(T)} + z_t \tag{13.2.2}$$

Hence, an AO affects the level of the observed time series only at time T, $Y_T = \omega + z_T$, by an unknown additive amount ω, while an IO represents an extraordinary random shock at time T, $a_T + \omega = a_T^*$, which affects all succeeding observations $Y_T, Y_{T+1}, \ldots$ through the dynamics of the system described by $\psi(B) = \theta(B)/\varphi(B)$, such that $Y_t = \omega\psi_i + z_t$ for $t = T + i \geq T$. For a stationary series, the effect of the IO is temporary since ψ_i decay exponentially to 0, but for nonstationary series with $d \geq 1$, there can be permanent effects that approach a level shift or even ramp effect since ψ_i do not decay to 0. More generally, an observed time series Y_t might be affected by outliers of different types at several points of time $T_1, T_2, \ldots, T_k$, and the multiple outlier model of the following general form

$$Y_t = \sum_{j=1}^{k} \omega_j \nu_j(B) P_t^{(T_j)} + z_t \tag{13.2.3}$$

could be considered for use, where $\nu_j(B) = 1$ for an AO at time T_j and $\nu_j(B) = \theta(B)/\varphi(B)$ for an IO at time T_j. Problems of interest associated with these outlier models are to identify the timing and the type of outliers and to estimate the magnitude ω of the outlier effect, so that the analysis of the time series will adjust for these outlier effects.

Tsay (1988), Chen and Tiao (1990), and Chen and Liu (1993), among others, also consider allowance in (13.2.3) for level shift type of outlier effect at unknown time of the form $\omega S_t^{(T)}$. The occurrence of such an effect is often encountered in series where the underlying process z_t that is nonstationary, and such that there is a factor $(1 - B)$ in the AR operator $\varphi(B)$ of the ARIMA model for z_t. Then recall that $(1 - B)S_t^{(T)} = P_t^{(T)}$ so that a level shift type of outlier effect for the nonstationary observed series Y_t is equivalent to an AO effect for the first differenced series $(1 - B)Y_t$.

13.2.2 Estimation of Outlier Effect for Known Timing of the Outlier

We first consider the estimation of the impact ω of an AO in (13.2.1) and that of an IO in (13.2.2), respectively, in the situation where the parameters of the time series model for the underlying process z_t are assumed known. To motivate iterative procedures that have been proposed for the general case, it will also be assumed that the timing T of the outlier is given.

Let $\pi(B) = \theta^{-1}(B)\varphi(B) = 1 - \sum_{i=1}^{\infty} \pi_i B^i$ and define $e_t = \pi(B)Y_t$ for $t = 1, 2, \ldots, n$, in terms of the observed series Y_t. Then we can write the above outlier models, (13.2.2) and (13.2.1), respectively, as

$$\text{IO}: \quad e_t = \omega P_t^{(T)} + a_t \tag{13.2.4a}$$

$$\text{AO}: \quad e_t = \omega\pi(B)P_t^{(T)} + a_t = \omega x_{1t} + a_t \tag{13.2.4b}$$

where for the AO model, $x_{1t} = \pi(B)P_t^{(T)} = -\pi_i$ if $t = T + i \geq T$, $x_{lt} = 0$ if $t < T$, with $\pi_0 = -1$. Thus, we see from (13.2.4) that the information about an IO is contained solely in the ''residual'' e_T at the particular time T, whereas that for an AO is spread over the

stretch of residuals $e_T, e_{T+1}, e_{T+2}, \ldots$, with generally decreasing weights $1, -\pi_1, -\pi_2, \ldots$, because the π_i are absolutely summable due to the invertibility of the MA operator $\theta(B)$. Equivalently, when an AO is present at time T, we can see that the residuals constructed from the observed series Y_t, for $t \geq T$, will be affected as $e_t = \pi(B)Y_t = a_t - \omega\pi_i$ for $t = T + i$. Hence, in the presence of an AO, a relatively high proportion of the constructed residuals could be influenced and distorted relative to the underlying white noise series a_t. Consequently, the presence of AOs that are unaccounted for typically tend to have a much more substantial adverse effect on estimates of the autocorrelations and parameters of the ARMA model for z_t compared to the presence of innovational outliers.

From least-squares principles, the least-squares estimator of the outlier impact ω in the IO model is simply the residual at time T,

$$\text{IO:} \quad \hat{\omega}_{I,T} = e_T \tag{13.2.5a}$$

with $\text{var}[\hat{w}_{I,T}] = \sigma_a^2$, while that in the AO model is the linear combination of $e_T, e_{T+1}, \ldots$,

$$\text{AO:} \quad \hat{\omega}_{A,T} = \frac{e_T - \sum_{i=1}^{n-T} \pi_i e_{T+i}}{\sum_{i=0}^{n-T} \pi_i^2} = \frac{\pi^*(F)e_T}{\tau^2} \tag{13.2.5b}$$

with $\text{var}[\hat{\omega}_{A,T}] = \sigma_a^2/\tau^2$, where $\tau^2 = \sum_{i=0}^{n-T} \pi_i^2$ and $\pi^*(F) = 1 - \pi_1 F - \pi_2 F^2 - \cdots - \pi_{n-T} F^{n-T}$. The notation in (13.2.5) reflects the fact that the estimates depend upon the time T. Note that in an underlying autoregressive model $\varphi(B)z_t = a_t$, since then $\pi^*(B) = \pi(B) = \varphi(B)$ for $T < n - p - d$, and $e_t = \varphi(B)Y_t$, in terms of the observations Y_t, the estimate $\hat{\omega}_{A,T}$ in (13.2.5b) can be written as

$$\hat{\omega}_{A,T} = \frac{\varphi(F)\varphi(B)Y_T}{\tau^2}$$

Since $\tau^2 \geq 1$, it is seen in general that $\text{var}[\hat{\omega}_{A,T}] \leq \text{var}[\hat{\omega}_{I,T}] = \sigma_a^2$, and in some cases $\text{var}[\hat{\omega}_{A,T}]$ can be much smaller than σ_a^2. For example, in an MA(1) model for z_t, the variance of $\hat{\omega}_{A,T}$ would be $\sigma_a^2(1 - \theta^2)/(1 - \theta^{2(n-T+1)}) \simeq \sigma_a^2(1 - \theta^2)$ when $n - T$ is large.

Significance tests for the presence of an outlier of type AO or IO at the given time T can be formulated as a test of $\omega = 0$ in either model (13.2.1) or (13.2.2), against $\omega \neq 0$. The likelihood ratio test criteria can be derived for both situations and essentially take the form of the standardized statistics

$$\lambda_{I,T} = \frac{\hat{\omega}_{I,T}}{\sigma_a} \quad \text{and} \quad \lambda_{A,T} = \frac{\tau\hat{\omega}_{A,T}}{\sigma_a} \tag{13.2.6}$$

respectively, for IO and AO types. Under the null hypothesis that $\omega = 0$, both statistics in (13.2.6) will have the standard normal distribution.

For the level-shift-type outlier model $Y_t = \omega S_t^{(T)} + z_t$, we have $e_t = \omega\pi(B)S_t^{(T)} + a_t$ and

$$\pi(B)S_t^{(T)} = \left[\frac{\pi(B)}{1 - B}\right] P_t^{(T)} \equiv \tilde{\pi}(B)P_t^{(T)}$$

with $\tilde{\pi}(B) = \pi(B)/(1-B) = 1 - \sum_{j=1}^{\infty} \tilde{\pi}_j B^j$. So it follows from the estimation results in (13.2.4b) and (13.2.5b) that the MLE of ω in the level shift model is $\hat{\omega}_{L,T} = \tilde{\pi}^*(F)e_T/\tilde{\tau}^2$ with

$$\tilde{\pi}^*(F) = 1 - \tilde{\pi}_1 F - \tilde{\pi}_2 F^2 - \cdots - \tilde{\pi}_{n-T} F^{n-T}$$

and $\tilde{\tau}^2 = 1 + \tilde{\pi}_1^2 + \cdots + \pi\tilde{\pi}_{n-T}^2$. When $d = 1$ in the ARIMA model, $\varphi(B) = \phi(B)(1-B)$ and $\tilde{\pi}(B) = \theta^{-1}(B)\phi(B)$, and, as discussed earlier, the results for this situation are the same as for the AO in terms of the model for the first differences:

$$(1-B)Y_t = \omega(1-B)S_t^{(T)} + (1-B)z_t = \omega P_t^{(T)} + \frac{\theta(B)}{\phi(B)}a_t$$

13.2.3 Iterative Procedure for Outlier Detection

In practice, the time T of a possible outlier as well as the model parameters are unknown. To address the problem of detection of outliers at unknown times, iterative procedures that are relatively convenient computationally have been proposed by Chang et al. (1988), Tsay (1986), and Chen and Liu (1993) to identify and adjust for the effects of outliers.

At the first stage of this procedure, the ARIMA model is estimated for the observed time series Y_t in the usual way, assuming that the series contains no outliers. The residuals $\hat{e}_t$ from the model are obtained as $\hat{e}_t = \theta^{-1}(B)\hat{\varphi}(B)Y_t = \hat{\pi}(B)Y_t$, and $\hat{\sigma}_a^2 = n^{-1}\sum_{t=1}^{n}\hat{e}_t^2$ is obtained. Then the statistics, as in (13.2.6),

$$\hat{\lambda}_{I,t} = \frac{\hat{\omega}_{I,t}}{\hat{\sigma}_a} \quad \text{and} \quad \hat{\lambda}_{A,t} = \frac{\hat{\tau}\hat{\omega}_{A,t}}{\hat{\sigma}_a}$$

are computed for each time $t = 1, 2, \ldots, n$, as well as

$$\hat{\lambda}_T = \max_t[\max(|\hat{\lambda}_{I,t}|, |\hat{\lambda}_{A,t}|)]$$

where T denotes the time when this maximum occurs. The possibility of an outlier of type IO is identified at time T if $\hat{\lambda} = |\hat{\lambda}_{I,T}| > c$, where c is a prespecified constant with typical values for c of 3.0, 3.5, or 4.0. The effect of this IO can be eliminated from the residuals by defining $\tilde{e}_T = \hat{e}_T - \hat{\omega}_{I,T} = 0$ at T. If $\hat{\lambda}_T = |\hat{\lambda}_{A,T}| > c$, the possibility of an AO is identified at T, and its impact is estimated by $\hat{\omega}_{A,T}$ as in (13.2.5b). The effect of this AO can be removed from the residuals by defining $\hat{e}_t = \hat{e}_t - \hat{\omega}_{A,T}\hat{\pi}(B)P_t^{(T)} = \hat{e}_t + \hat{\omega}_{A,T}\hat{\pi}_{t-T}$ for $t \geq T$. In either case, a new estimate $\tilde{\sigma}_a^2$ is computed from the modified residuals $\tilde{e}_t$.

If any outliers are identified, the modified residuals $\tilde{e}_t$ and modified estimate $\tilde{\sigma}_a^2$, but the same parameters $\hat{\pi}(B) = \hat{\theta}^{-1}(B)\hat{\varphi}(B)$, are used to compute new statistics $\hat{\lambda}_{I,t}$ and $\hat{\lambda}_{A,t}$. The preceding steps are then repeated until all outliers are identified. Suppose that this procedure identifies outliers at k time points $T_1, T_2, \ldots, T_k$. Then the overall outlier model, as in (13.2.3),

$$Y_t = \sum_{j=1}^{k} \omega_j \nu_j(B) P_t^{(T_j)} + \frac{\theta(B)}{\varphi(B)}a_t \tag{13.2.7}$$

is estimated for the observed series Y_t, where $v_j(B) = 1$ for an AO and $v_j(B) = \theta(B)/\varphi(B)$ for an IO at time T_j. A revised set of residuals

$$\hat{e}_t = \hat{\theta}^{-1}(B)\hat{\varphi}(B)\left[Y_t - \sum_{j=1}^{k} \hat{\omega}_j \hat{v}_j(B) P_t^{(T_j)}\right]$$

and a new $\hat{\sigma}_a^2$ are obtained from this fitted model. The previous steps of the procedure can then be repeated with new residuals, until all outliers are identified and a final model of the general form of (13.2.7) is estimated. If desired, a modified time series of observations in which the effects of the outliers have been removed can be constructed as $\tilde{z}_t = Y_t - \sum_{j=1}^{k} \hat{\omega}_j \hat{v}_j(B) P_t^{(T_j)}$.

The procedure above can be implemented, with few modifications, to any existing software capable of estimation of ARIMA and transfer function–noise models. An implementation in the R package will be demonstrated below. The technique can be a useful tool in the identification of potential time series outliers that if undetected could have a negative impact on the effectiveness of modeling and estimation. However, there should be some cautions concerning the systematic use of such ''outlier adjustment'' procedures, particularly with regard to the overall interpretation of results, the appropriateness of a general model specification for ''outliers'' such as (13.2.7), which treats the outliers as deterministic constants, and the possibilities for ''overspecification'' in the number of outliers. Whenever possible, it would always be highly desirable to search for the causes or sources of the outliers that may be identified by the foregoing procedure, so that the outlying behavior can be better understood and properly accounted for in the analysis. Also, although the foregoing procedures should perform well when the series has only a few relatively isolated outliers, there could be difficulties due to ''masking effects'' when the series has multiple outliers that occur in patches, especially when they are in the form of additive outliers and level shift effects. Modifications to the basic procedure to help remedy these difficulties associated with multiple outliers, including *joint* estimation of all identified outlier effects and the model parameters within the iteration stages, were proposed by Chen and Liu (1993).

13.2.4 Examples of Analysis of Outliers

We consider two numerical examples to illustrate the application of the outlier analysis procedures, discussed in the previous sections. For computational convenience, conditional least-squares estimation methods are used throughout in these examples.

Series D. The first example involves Series D, which represents ''uncontrolled" viscosity readings every hour from a chemical process. In Chapter 7, an AR(1) model $(1 - \phi B)z_t = \theta_0 + a_t$ has been suggested and fitted to this series. In the outlier detection procedure, the model is first estimated assuming that no outliers are present, and the results are given in Table 13.1(a). Then the AO and IO statistics as in (13.2.6) are computed for each time point t, using $\hat{\sigma}_a^2 = 0.08949$. Based on a critical value of $c = 3.5$, we lead to identification of an IO of rather large magnitude at time $T = 217$. The effect of this IO is removed by modifying the residual at T, a new estimate $\tilde{\sigma}_a^2 = 0.08414$ is obtained, and new outlier statistics are computed using $\tilde{\sigma}_a$. At this stage, no outliers are identified. Then, the time

TABLE 13.1 Outlier Detection and Parameter Estimation Results for Series C and D Examples

	Parameter[a]						Outlier			
	$\hat{\theta}_0$	$\hat{\phi}$	$\hat{\omega}_1$	$\hat{\omega}_2$	$\hat{\omega}_3$	$\hat{\sigma}_a^2$	Time	$\hat{\omega}$	$\hat{\lambda}$	Type
(a) Series D										
Cycle 1	1.269	0.862				0.0895	217	−1.28	−4.29	IO
	(0.258)	(0.028)								
Final	1.181	0.872	−1.296			0.0841				
	(0.251)	(0.027)	(0.292)							
(b) Series C										
Cycle 1		0.813				0.0179	58	0.76	5.65	IO
		(0.038)					59	−0.51	−4.16	IO
							60	−0.44	−3.74	IO
Final		0.851	0.745	−0.551	−0.455	0.0132				
		(0.035)	(0.116)	(0.120)	(0.116)					

[a] Standard errors of parameter estimates are in parentheses.

series parameters and the outlier parameter ω in model (13.2.2), that is, in the model

$$Y_t = \frac{1}{1-\phi B}[\theta_0 + \omega P_t^{(T)} + a_t]$$

are estimated simultaneously, and the estimates are given in Table 13.1(a). Repeating the outlier detection procedure based on these new parameter estimates and corresponding residuals does not reveal any other outliers. Hence, only one extreme IO is identified, and adjusting for this IO does not result in much change in the estimate $\hat{\phi}$ of the time series model parameter, but gives about a 6% reduction in the estimate of σ_a^2. Several other potential outliers, at times $t = 29, 113, 115, 171, 268$, and 272, were also suggested during the outlier procedure as having values of the test statistics $\hat{\lambda}$ slightly greater than 3.0 in absolute value, but adjustment for such values did not affect the estimates of the model substantially.

Series C. The second example we consider is Series C, the ''uncontrolled'' temperature readings every minute in a chemical process. The model previously identified and fitted to this series is the ARIMA(1, 1, 0) model, $(1 - \phi B)(1 - B)z_t = a_t$. The estimation results for this model obtained assuming there are no outliers are given in Table 13.1(b). Proceeding with the sequence of calculations of the outlier test statistics and using the critical value of $c = 3.5$, we first identify an IO at time 58. The residual at time 58 is modified, we obtain a new estimate $\tilde{\sigma}_a^2 = 0.01521$, and next an IO at time 59 is identified. This residual is modified, a new estimate $\tilde{\sigma}_a^2 = 0.01409$ is obtained, and then another IO at time 60 is indicated. After this, no further outliers are identified. These innovational outliers at times 58, 59, and 60 are rather apparent in Figure 13.2(a), which shows a time series plot of the residuals from the initial model fit before any adjustment for outliers.

Then the time series outlier model

$$(1 - B)Y_t = \frac{1}{1-\phi B}[\omega_1 P_t^{(58)} + \omega_2 P_t^{(59)} + \omega_3 P_t^{(60)} + a_t]$$

is estimated for the series, and the results are presented in Table 13.1(b). The residuals are shown in Figure 13.2(b). No other outliers are detected when the outlier procedure is repeated with the new model parameter estimates. In this example we see that adjustment for the outliers has a little more effect on the estimate $\hat{\phi}$ of the time series parameter than in the previous case, and it reduces the estimate of σ_a^2 substantially by about 26%. Figure 13.2(b) clearly shows the reduction in variability due to the outlier adjustment.

Calculations Using R. The detection and adjustment for outliers in time series can be performed using the TSA package in R. The code needed to do the analysis for Series C and D is as follows:

```
> library(TSA)
> m1.C=arima(seriesC,order=c(1,1,0))
> m1.C
> detectAO(m1.C); detectIO(m1.C)
> m2.C=arimax(seriesC,order=c(1,1,0),io=c(58,59,60))
> m2.C
```

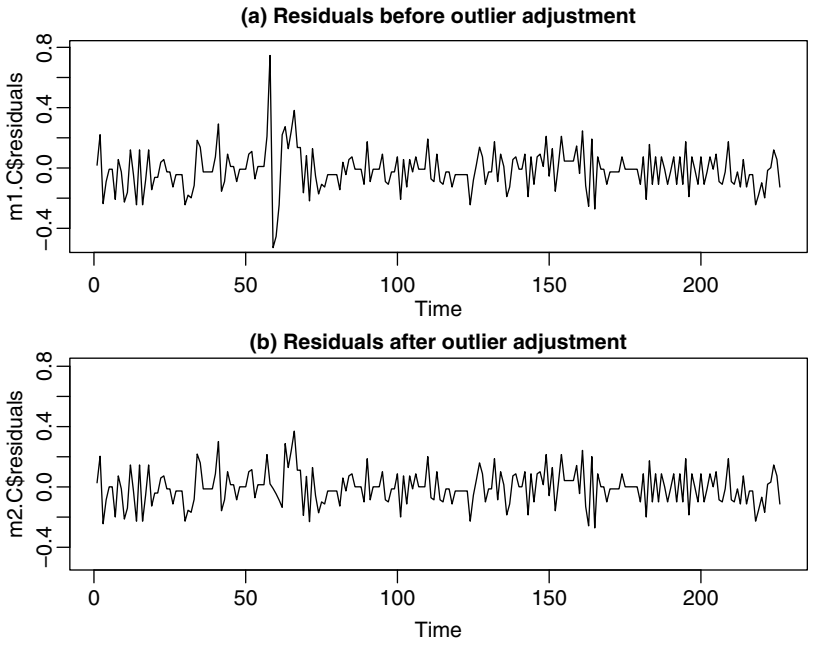


FIGURE 13.2 Residuals from the ARIMA(1, 1, 0) model fitted to Series C before and after adjustment for innovational outliers at $t = 58$, 59, and 60.

```
> m1.D=arima(seriesD,order=c(1,0,0))
> m1.D
> detectAO(m1.D); detectIO(m1.D)
> m2.D=arimax(seriesD,order=c(1,0,0),io=c(217))
> m2.D
```

Figure 13.2 that shows the residuals for Series C before and after the outlier adjustment can be reproduced in R as follows:

```
> par(mfrow=c(2,1))
> plot(m1.C$residuals,ylim=c(-0.5,0.8),
       main='(a) Residuals before outlier adjustment')
> plot(m2.C$residuals,ylim=c(-0.5,0.8),
       main='(b) Residuals after outlier adjustment')
```

13.3 ESTIMATION FOR ARMA MODELS WITH MISSING VALUES

In some situations in practice, the values of a time series z_t may not be observed at equally spaced times because there may be ''missing values'' corresponding to certain time points. In this section we discuss briefly the maximum likelihood estimation of parameters in an ARIMA(p, d, q) model for such situations, through consideration of the calculation of the exact Gaussian likelihood function for the observed data. It is shown that for series with missing observations, the likelihood function can conveniently be constructed using the

state-space form of the model and associated Kalman filtering procedures, as discussed in Sections 5.5 and 7.4, but modified to accommodate the missing data. These methods for evaluation of the likelihood in cases of irregularly spaced observations have been examined by Jones (1980), Harvey and Pierse (1984), Ansley and Kohn (1983, 1985), and Wincek and Reinsel (1986), among others. We also address briefly the related issue of estimation of the missing values in the time series.

13.3.1 State-Space Model and Kalman Filter with Missing Values

We suppose n observations are available at integer times $t_1 < t_2 < \cdots < t_n$, not equally spaced, from an ARIMA(p, d, q) process, which follows the model $\phi(B)(1 - B)^d z_t = \theta(B)a_t$. From Section 5.5.1, the process z_t has the state-space formulation given by

$$Y_t = \mathbf{\Phi} Y_{t-1} + \mathbf{\Psi} a_t \tag{13.3.1}$$

with $z_t = \mathbf{H}Y_t = [1, 0, \ldots, 0]Y_t$, where Y_t is the r-dimensional state vector and $r = \max(p + d, q + 1)$. Let $\Delta_i = t_i - t_{i-1}$ denote the time difference between successive observations $z_{t_{i-1}}$ and z_{t_i}, $i = 2, \ldots, n$. By successive substitutions, Δ_i times, on the right-hand side of (13.3.1), we obtain

$$Y_{t_i} = \mathbf{\Phi}^{\Delta_i} Y_{t_{i-1}} + \sum_{j=0}^{\Delta_i - 1} \mathbf{\Phi}^j \mathbf{\Psi} a_{t_i - j} \equiv \mathbf{\Phi}_i^* Y_{t_{i-1}} + a_{t_i}^* \tag{13.3.2}$$

where $\mathbf{\Phi}_i^* = \mathbf{\Phi}^{\Delta_i}$ and $a_{t_i}^* = \sum_{j=0}^{\Delta_i - 1} \mathbf{\Phi}^j \mathbf{\Psi} a_{t_i - j}$, with

$$\operatorname{cov}\left[a_{t_i}^*\right] = \mathbf{\Sigma}_i = \sigma_a^2 \sum_{j=0}^{\Delta_i - 1} \mathbf{\Phi}^j \mathbf{\Psi}\mathbf{\Psi}' \mathbf{\Phi}'^j$$

Thus, (13.3.2) together with the observation equation $z_{t_i} = \mathbf{H}Y_{t_i}$ constitutes a state-space model form for the *observed* time series data $z_{t_1}, z_{t_2}, \ldots, z_{t_n}$.

Therefore, the Kalman filter recursive equations as in (5.5.6) to (5.5.9) can be directly employed to obtain the state predictors $\hat{Y}_{t_i|t_{i-1}}$ and their error covariance matrices $\mathbf{V}_{t_i|t_{i-1}}$. So we can obtain the predictors

$$\hat{z}_{t_i|t_{i-1}} = E[z_{t_i} | z_{t_{i-1}}, \ldots, z_{t_1}] = \mathbf{H}\hat{Y}_{t_i|t_{i-1}} \tag{13.3.3}$$

for the observations z_{t_i} based on the previous *observed* data and their error variances

$$\sigma_a^2 v_i = \mathbf{H}\mathbf{V}_{t_i|t_{i-1}}\mathbf{H}' = E[(z_{t_i} - \hat{z}_{t_i|t_{i-1}})^2] \tag{13.3.4}$$

readily from the recursive Kalman filtering procedure. More specifically, the updating equations (5.5.6) and (5.5.7) in this missing data setting take the form

$$\hat{Y}_{t_i|t_i} = \hat{Y}_{t_i|t_{i-1}} + \mathbf{K}_i(z_{t_i} - \mathbf{H}\hat{Y}_{t_i|t_{i-1}}) \tag{13.3.5}$$

with

$$\mathbf{K}_i = \mathbf{V}_{t_i|t_{i-1}}\mathbf{H}'[\mathbf{H}\mathbf{V}_{t_i|t_{i-1}}\mathbf{H}']^{-1} \tag{13.3.6}$$

while the prediction equations (5.5.8) are given by

$$\hat{\mathbf{Y}}_{t_i|t_{i-1}} = \mathbf{\Phi}_i^* \hat{\mathbf{Y}}_{t_{i-1}|t_{i-1}} = \mathbf{\Phi}^{\Delta_i} \hat{\mathbf{Y}}_{t_{i-1}|t_{i-1}} \qquad \mathbf{V}_{t_i|t_{i-1}} = \mathbf{\Phi}_i^* \mathbf{V}_{t_{i-1}|t_{i-1}} \mathbf{\Phi}_i^{*\prime} + \mathbf{\Sigma}_i \tag{13.3.7}$$

with

$$\mathbf{V}_{t_t|t_i} = [\mathbf{I} - \mathbf{K}_i \mathbf{H}] \mathbf{V}_{t_i|t_{i-1}} \tag{13.3.8}$$

Notice that the calculation of the prediction equations (13.3.7) can be interpreted as computation of the successive one-step-ahead predictions:

$$\hat{\mathbf{Y}}_{t_{i-1}+j|t_{i-1}} = \mathbf{\Phi} \hat{\mathbf{Y}}_{t_{i-1}+j-1|t_{i-1}}$$
$$\mathbf{V}_{t_{i-1}+j|t_{i-1}} = \mathbf{\Phi} \mathbf{V}_{t_{i-1}+j-1|t_{i-1}} \mathbf{\Phi}' + \sigma_a^2 \mathbf{\Psi} \mathbf{\Psi}'$$

for $j = 1, \ldots, \Delta_i$, without any updating since there are no observations available between the time points t_{i-1} and t_i to provide any additional information for updating.

Exact Likelihood Function with Missing Values. The exact likelihood for the vector of observations $\mathbf{z}' = (z_{t_1}, z_{t_2}, \ldots, z_{t_n})$ is obtained directly from the quantities in (13.3.3) and (13.3.4) because the joint density of $\mathbf{z}$ can be expressed as the product of the conditional densities of the z_{t_i}, given $z_{t_{i-1}}, \ldots, z_{t_1}$, for $i = 2, \ldots, n$, which are Gaussian with conditional means and variances given by (13.3.3) and (13.3.4). Hence, the joint density of the observations $\mathbf{z}$ can be expressed as

$$p(\mathbf{z}|\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2) = \prod_{i=1}^{n} (2\pi\sigma_a^2 v_i)^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2} \sum_{i=1}^{n} \frac{(z_{t_i} - \hat{z}_{t_i|t_{i-1}})^2}{v_i} \right] \tag{13.3.9}$$

In (13.3.9), the quantities $\hat{z}_{t_i|t_{i-1}}$ and $\sigma_a^2 v_i$ are directly determined from the recursive filtering calculations (13.3.5)–(13.3.8). In the case of a stationary ARMA(p, q) model, the initial conditions required to start the filtering procedure can be determined readily (see, for example, Jones, (1980) and Section 5.5.2). However, for the nonstationary ARIMA model situation, some additional assumptions need to be specified concerning the process and the initial conditions. Appropriate methods for such cases have been examined by Ansley and Kohn (1985).

As a simple example to illustrate the missing data methods, consider the stationary AR(1) model $(1 - \phi B) z_t = a_t$. Then, (13.3.2) directly becomes (see, for example, Reinsel and Wincek, 1987)

$$z_{t_i} = \phi^{\Delta_i} z_{t_{i-1}} + \sum_{j=0}^{\Delta_{i-1}} \phi^j a_{t_i - j} \tag{13.3.10}$$

and it is readily determined that

$$\hat{z}_{t_i|t_{i-1}} = \phi^{\Delta_i} z_{t_{i-1}} \text{ and } \sigma_i^2 = \sigma_a^2 v_i = \frac{\sigma_a^2 (1 - \phi^{2\Delta_i})}{1 - \phi^2} \tag{13.3.11}$$

Hence, the likelihood for the observed data in the first-order autoregressive model with missing values is as given in (13.3.9), with these expressions for $\hat{z}_{t_i|t_{i-1}}$ and $\sigma_a^2 v_i$.

13.3.2 Estimation of Missing Values of an ARMA Process

A related problem of interest that often arises in the context of missing values for time series is that of estimating the missing values. Studies based on interpolation of missing values for ARIMA time series from a least-squares viewpoint were performed by Brubacher and Tunnicliffe Wilson (1976), Damsleth (1980), and Abraham (1981). Within the framework of the state-space formulation, estimates of missing values and their corresponding error variances can be derived conveniently through the use of recursive smoothing methods associated with the Kalman filter, which were discussed briefly in Section 5.5.3 and are described in general terms in Anderson and Moore (1979), for example. These methods have been considered more specifically for the ARIMA model with missing values by Harvey and Pierse (1984) and by Kohn and Ansley (1986).

For the special case of a pure autoregressive model, $\phi(B)z_t = a_t$, some rather simple and explicit interpolation results are available. For example, in an AR(p) process with a single missing value at time T surrounded by at least p consecutive observed values both before and after time T, it is well known (see, for example, Brubacher and Tunnicliffe Wilson, 1976) that the optimal interpolation of the missing value z_T is given by

$$\hat{z}_T = -d_0^{-1} \sum_{j=1}^{p} d_j (z_{T-j} + z_{T+j}) \tag{13.3.12}$$

where $d_j = \sum_{i=j}^{p} \phi_i \phi_{i-j}$, $\phi_0 = -1$, and $d_0 = 1 + \sum_{i=1}^{p} \phi_i^2$, with $E[(z_T - \hat{z}_T)^2] = \sigma_a^2 d_0^{-1} = \sigma_0^2 (1 + \sum_{i=1}^{p} \phi_i^2)^{-1}$. Notice that the value in (13.3.12) can be expressed as $\hat{z}_T = z_T - [\phi(F)\phi(B)z_T/d_0]$, with interpolation error equal to

$$\hat{e}_T = z_T - \hat{z}_T = \frac{\phi(F)\phi(B)z_T}{d_0} \tag{13.3.13}$$

As one way to establish the result (13.3.12), for convenience of discussion, suppose that z_T is the only missing value among times $t = 1, \ldots, n$, with $p + 1 \le T \le n - p$. Using a normal distribution assumption, the optimal (minimum MSE) estimate of z_T is $\hat{z}_T = E[z_T | z_1, \ldots, z_{T-1}, z_{T+1}, \ldots, z_n]$, which is also the best linear estimate without the normality assumption. Then, by writing the joint density of $\mathbf{z} = (z_1, \ldots, z_n)'$ in the form

$$p(z_1, \ldots, z_{T-1}, z_{T+1}, \ldots, z_n) p(z_T | z_1, \ldots, z_{T-1}, z_{T+1}, \ldots, z_n)$$

from basic properties of the multivariate normal distribution and its conditional distributions, it is easily deduced that the estimate $\hat{z}_T$, the conditional mean, is identical to the value of z_T that minimizes the ''sum-of-squares'' function in the exponent of the joint multivariate normal density of $\mathbf{z}$. Thus, since z_T occurs only in $p + 1$ terms of the exponent sum of squares, this reduces to finding the value of z_T to minimize $S = \sum_{i=0}^{p} a_{T+i}^2$, where

$a_t = z_t - \sum_{l=1}^{p} \phi_l z_{t-l}$. Now we obtain

$$\begin{aligned}\frac{\partial S}{\partial z_T} &= 2\left[\left(z_T - \sum_{l=1}^{p}\phi_l z_{T-l}\right) - \sum_{i=1}^{p}\phi_i\left(z_{T+i} - \sum_{l=1}^{p}\phi_l z_{T+i-l}\right)\right] \\ &= 2\left[\left(1 + \sum_{i=1}^{p}\phi_i^2\right)z_T + \sum_{i=0}^{p}\phi_i\left\{\sum_{l\neq i}^{p}\phi_l z_{T+i-l}\right\}\right] \\ &= 2\left[\left(1 + \sum_{i=1}^{p}\phi_i^2\right)z_T + \sum_{j=1}^{p}\left(\sum_{i=j}^{p}\phi_i\phi_{i-j}\right)(z_{T-j} + z_{T+j})\right]\end{aligned}$$

where $\phi_0 = -1$. Setting this partial derivative to zero and solving for z_T, we find that the estimate is given by $\hat{z}_T = -d_0^{-1}\sum_{j=1}^{p} d_j(z_{T-j} + z_{T+j})$, where $d_j = \sum_{i=j}^{p}\phi_i\phi_{i-j}$ and $d_0 = 1 + \sum_{i=1}^{p}\phi_i^2$. Notice that the estimate $\hat{z}_T$ can be seen to be determined from the solution for z_T to the relation $\phi(F)\phi(B)z_T = 0$, where $\phi(B) = 1 - \sum_{i=1}^{p}\phi_i B^i$ is the AR(p) operator. It can also be established that the error variance of the missing data estimate is given by $E[(z_T - \hat{z}_T)^2] = \sigma_a^2 d_0^{-1}$.

In the general ARMA model situation, Bruce and Martin (1989) and Ljung (1993), among others, have noted a close connection between the likelihood function construction in the case of missing values and the formulation of the consecutive data model likelihood with AOs specified for each time point that corresponds to a missing value. Hence, in effect, in such a time series AO model for consecutive data, for given values of the ARMA model parameters, the estimate of the outlier effect parameter ω corresponds to the interpolation error in the missing data situation. For example, in the autoregressive model situation, compare the result in (13.3.13) with the result given following (13.2.5b) for the AO model. Specifically, since $\pi(B) = \phi(B)$ in the AR(p) model, $e_T = \phi(B)Y_T$ and the estimate in (13.2.5b) reduces to $\hat{\omega}_{A,T} = [\phi(F)\phi(B)Y_T]/d_0 \equiv Y_T - \hat{Y}_T = \hat{e}_T$, the interpolation error given in (13.3.13). Furthermore, the sum-of-squares function in the likelihood (13.3.9) for the missing data situation is equal to the sum of squares obtained from a complete set of consecutive observations in which an AO has been assumed at each time point where a missing value occurs and for which the likelihood is evaluated at the maximum likelihood estimates for each of the corresponding AO effect parameters ω, for given values of the time series model parameters ϕ and θ. As an illustration, for the simple AR(1) model situation with a single isolated missing value at time T, from (13.3.11) the relevant term in the missing data sum-of-squares function is

$$\begin{aligned}\frac{(z_{T+1} - \phi^2 z_{T-1})^2}{1+\phi^2} &\equiv [(z_T - \hat{\omega}) - \phi z_{T-1}]^2 + [z_{T+1} - \phi(z_T - \hat{\omega})]^2 \\ &= (\hat{z}_T - \phi z_{T-1})^2 + (z_{T+1} - \phi\hat{z}_T)^2 \qquad (13.3.14)\end{aligned}$$

where

$$\hat{\omega} = z_T - \frac{\phi}{1+\phi^2}(z_{T-1} + z_{T+1}) = z_T - \hat{z}_T$$

is the maximum likelihood estimate of the outlier effect ω in the AO model (13.2.1), and the latter expressions in (13.3.14) represent the sum-of-squares terms in the consecutive data situation but with an AO modeled at time T.

Treating missing data as additive outliers does have an impact on estimation of the ARMA model parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, however, and ML estimates of these parameters in the missing data case are not identical to estimates that maximize a complete data likelihood for which an AO has been assumed at each time point where a missing value occurs. In fact, Basu and Reinsel (1996) established that MLEs of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ for the missing data situation are the same as estimates obtained from a model that assumes complete data with an AO at each time point where a missing value occurs when the method of *restricted maximum likelihood* estimation (e.g., as discussed in Section 9.5.2) is employed for this latter model formulation. We provide the following argument to establish this result.

Connection Between Exact Likelihood Function for Missing Data Situation and Restricted Likelihood. Let $\mathbf{z}_n = (z_{t_1}, z_{t_2}, \ldots, z_{t_n})'$ denote the $n \times 1$ vector of observations from the ARMA(p, q) process $\phi(B)z_t = \theta(B)a_t$ with $t_1 \equiv 1$ and $t_n = T$. Let $\mathbf{z}_0$ denote the $T \times 1$ vector consisting of the observations $\mathbf{z}_n$ with 0's inserted for values of times where observations are missing, and for convenience arrange as $\mathbf{z}_0 = (\mathbf{z}_n', 0')'$. Also, let $\mathbf{z} = (\mathbf{z}_n', \mathbf{z}_m')'$ denote the corresponding vector of (complete) values of the process, where $\mathbf{z}_m$ is the $m \times 1$ vector of the ''missing values,'' with $T = n + m$. We can write

$$\mathbf{z}_0 = \mathbf{X}\boldsymbol{\omega} + \mathbf{z} \tag{13.3.15}$$

where $\mathbf{X}$ is a $T \times m$ matrix with columns that are ''pulse'' unit vectors to indicate the m missing values, specifically, $\mathbf{X} = [\mathbf{0}, \mathbf{I}_m]'$ under the rearrangement of the data. Thus, (13.3.15) can be interpreted as a model that allows for AOs, with parameters $\boldsymbol{\omega}$, at all time points where a missing value occurs. Note that $\mathbf{z}_n = \mathbf{H}'\mathbf{z} \equiv \mathbf{H}'\mathbf{z}_0$ where $\mathbf{H}' = [\mathbf{I}_n, \mathbf{0}]$ is the $n \times T$ matrix whose rows are pulse unit vectors to indicate the n observed values.

From one perspective, (13.3.15) can be viewed as a ''regression model'' for the extended data vector $\mathbf{z}_0$ with $\boldsymbol{\omega}$ treated as unknown parameters and ARMA noise process $\{z_t\}$. (Note in fact that $\boldsymbol{\omega} = -\mathbf{z}_m$ by actual definition.) Let $\sigma_a^2\mathbf{V}_* = \text{cov}[\mathbf{z}]$ denote the $T \times T$ covariance matrix of the complete series of values. Then, the form of the restricted likelihood function for the extended data vector $\mathbf{z}_0$ under this regression model is given as in (9.5.11) of Section 9.5.2,

$$L_*\left(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma_a^2; \mathbf{z}_0\right) \propto \left(\sigma_a^2\right)^{-n/2} |\mathbf{V}_*|^{-1/2} |\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X}|^{-1/2} \times \exp\left[-\frac{1}{2\sigma_a^2}(\mathbf{z}_0 - \mathbf{X}\hat{\boldsymbol{\omega}})'\mathbf{V}_*^{-1}(\mathbf{z}_0 - \mathbf{X}\hat{\boldsymbol{\omega}})\right] \tag{13.3.16}$$

where $\hat{\boldsymbol{\omega}} = (\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{z}_0$. Recall from discussion in Section 9.5.2, however, that (13.3.16) has an equivalent representation as the density of the ''error contrast vector'' $\mathbf{H}'\mathbf{z}_0$, since $\mathbf{H}'$ is a full rank $(T - m) \times T$ matrix such that $\mathbf{H}'\mathbf{X} = \mathbf{0}$. Then noting that $\mathbf{H}'\mathbf{z}_0 = \mathbf{z}_n$, the observed data vector, expression (13.3.16) also represents the density of $\mathbf{z}_n$ and hence represents the exact likelihood based on the *observed* data vector $\mathbf{z}_n$, essentially by definition. However, we would now like to directly verify the equivalence between (13.3.16) and the exact likelihood (density) function of the observed data vector $\mathbf{z}_n$.

For this, we express the covariance matrix of $\mathbf{z} = (\mathbf{z}_n', \mathbf{z}_m')'$ in partitioned form as

$$\text{cov}[\mathbf{z}] = \sigma_a^2 \mathbf{V}_* = \sigma_a^2 \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix}$$

where $\sigma_a^2 \mathbf{V}_{11} = \text{cov}[\mathbf{z}_n]$ in particular. We let $\mathbf{V}^{ij}$, $i, j = 1, 2$, denote the block matrices of $\mathbf{V}_*^{-1}$ corresponding to the above partitioning of $\mathbf{V}_*$. Then using basic results for partitioned matrices (e.g., Rao (1965), or Appendix A7.1.1), we can readily derive that the restricted likelihood expression in (13.3.16) is the same as the likelihood (density) for the observed data vector $\mathbf{z}_n$. That is, from results on partitioned matrices, we first have that

$$\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X} \equiv \mathbf{V}^{22} = \left(\mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\right)^{-1} \tag{13.3.17}$$

and $|\mathbf{V}_*| = |\mathbf{V}_{11}||\mathbf{V}_{22} - \mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}|$. Hence, the determinant factor in (13.3.16) is $|\mathbf{V}_*|^{-1/2}|\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X}|^{-1/2} = |\mathbf{V}_{11}|^{-1/2}$. Also, the quadratic form in (13.3.16) is expressible as

$$\begin{aligned} &\mathbf{z}_0'[\mathbf{V}_*^{-1} - \mathbf{V}_*^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_*^{-1}]\mathbf{z}_0 \\ &\quad = \mathbf{z}_n'[\mathbf{V}^{11} - \mathbf{V}^{12}(\mathbf{V}^{22})^{-1}\mathbf{V}^{21}]\mathbf{z}_n = \mathbf{z}_n'\mathbf{V}_{11}^{-1}\mathbf{z}_n \end{aligned}$$

again using a basic result on the inverse of a partitioned matrix. Therefore, expression (13.3.16) is equal to

$$p(\mathbf{z}_n) \propto \left(\sigma_a^2\right)^{-n/2} |\mathbf{V}_{11}|^{-1/2} \exp\left[-\frac{1}{2\sigma_a^2}\mathbf{z}_n'\mathbf{V}_{11}^{-1}\mathbf{z}_n\right] \tag{13.3.18}$$

which, since $\mathbf{z}_n$ is distributed as normal $N(\mathbf{0}, \sigma_a^2\mathbf{V}_{11})$, is the likelihood based on the observed data vector $\mathbf{z}_n$.

This equivalence establishes a device for obtaining ML estimates in ARMA models with missing values by using an REML estimation routine for the extended data vector $\mathbf{z}_0$ by setting up a regression component $\mathbf{X}\boldsymbol{\omega}$ that includes an indicator variable (AO term) for each missing observation. Estimation of the ''extended data'' regression model (13.3.15) with ARMA errors by the method of REML then results in ML estimates of the ARMA model parameters based on the observed data $\mathbf{z}_n$. Finally, we note that the GLS estimate of $\boldsymbol{\omega}$ in model (13.3.15) is

$$\begin{aligned} \hat{\boldsymbol{\omega}} &= (\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{z}_0 \\ &= (\mathbf{V}^{22})^{-1}\mathbf{V}^{21}\mathbf{z}_n \equiv -\mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{z}_n = -E[\mathbf{z}_m|\mathbf{z}_n] \end{aligned} \tag{13.3.19}$$

so the estimates of the missing values $\mathbf{z}_m$ are obtained as $\hat{\mathbf{z}}_m = -\hat{\boldsymbol{\omega}}$ immediately as a by-product of the fitting of the model (13.3.15), with estimation error covariance matrix $\text{cov}[\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}] \equiv \text{cov}[\mathbf{z}_m - \hat{z}_m] = \sigma_a^2(\mathbf{X}'\mathbf{V}_*^{-1}\mathbf{X})^{-1}$ directly available as well. In addition, for a complete data vector situation, if there were additive outliers specified at the given times corresponding to $\mathbf{z}_m$, then model (13.3.15) could be used to estimate ''smoothed values'' of the observations at all times where an AO is proposed to occur, $\hat{\mathbf{z}}_m = -\hat{\boldsymbol{\omega}}$ as

given in (13.3.19), and magnitudes of the outliers can be estimated by the differences between the observed values and the interpolated values, $\mathbf{z}_m - \hat{\mathbf{z}}_m$.

EXERCISES

13.1. In an analysis (Box and Tiao, 1975) of monthly data Y_t on smog-producing oxidant, allowance was made for two possible ''interventions'' I_1 and I_2 as follows:

I_1: In early 1960, diversion of traffic from the opening of the Golden State Freeway and the coming into effect of a law reducing reactive hydrocarbons in gasoline sold locally.

I_2: In 1966, the coming into effect of a law requiring all new cars to have modified engine design. In the case of this intervention, allowance was made for the well-known fact that the smog phenomenon is different in summer and winter months.

In a pilot analysis of the data, the following intervention model was used:

$$Y_t = \omega_1 \xi_{1t} + \frac{\omega_2}{1 - B^{12}} \xi_{2t} + \frac{\omega_3}{1 - B^{12}} \xi_{3t} + \frac{(1 - \theta B)(1 - \Theta B^{12})}{1 - B^{12}} a_t$$

where

$$\xi_{1t} = \begin{cases} 0 & t < \text{Jan. } 1960 \\ 1 & t \geq \text{Jan. } 1960 \end{cases} \qquad \xi_{2t} = \begin{cases} 0 & t < \text{Jan. } 1966 \\ 1 & t \geq \text{Jan. } 1966 \end{cases} \quad \text{(summer months)}$$

$$\xi_{3t} = \begin{cases} 0 & t < \text{Jan. } 1966 \\ 1 & t \geq \text{Jan. } 1966 \end{cases} \quad \text{(winter months)}$$

(a) Show that the model allows for the following:

(1) A possible step change in January 1960 of size ω_1, possibly produced by I_1.

(2) A ''staircase function'' of annual step size ω_2 to allow for possible summer effect of cumulative influx of cars with new engine design.

(3) A ''staircase function'' of annual step size ω_3 to allow for possible winter effect of cumulative influx of cars with new engine design.

(b) Describe what steps you would take to check the representational adequacy of the model.

(c) Assuming you were satisfied with the checking after (b), what conclusions would you draw from the following results? (Estimates are shown with their standard errors below in parentheses.)

$$\hat{\omega}_1 = -1.09 \quad \hat{\omega}_2 = -0.25 \quad \hat{\omega}_3 = -0.07 \quad \hat{\theta} = -0.24 \quad \hat{\Theta} = 0.55$$
$$(\pm 0.13) \qquad (\pm 0.07) \qquad (\pm 0.06) \qquad (\pm 0.03) \qquad (\pm 0.04)$$

(d) The data for this analysis are listed as Series R in the Collection of Time Series in Part Five. Use these data to perform your own intervention analysis.

13.2. A general transfer function model of the form

$$Y_t = \sum_{j=1}^{k} \delta_j^{-1}(B)\omega_j(B)\xi_{jt} + \phi^{-1}(B)\theta(B)a_t \equiv \mathcal{Y}_t + N_t$$

can include input variables ξ_j, which are themselves time series, and other inputs ξ_i, which are indicator variables. The latter can estimate (and eliminate) the effects of interventions of the kind described in Exercise 13.1 and, in particular, are often useful in the analysis of sales data.

Let $\xi_t^{(T)}$ be an indicator variable that takes the form of a unit pulse at time T, that is

$$\xi_t^{(T)} = \begin{cases} 0 & t \neq T \\ 1 & t = T \end{cases}$$

For illustration, consider the models

$$(1)\ \mathcal{Y}_t = \frac{\omega_1 B}{1 - \delta B}\xi_t^{(T)} \qquad (\text{with } \omega_1 = 1.0, \delta = 0.5)$$

$$(2)\ \mathcal{Y}_t = \left(\frac{\omega_1 B}{1 - \delta B} + \frac{\omega_2 B}{1 - B}\right)\xi_t^{(T)} \qquad (\text{with } \omega_1 = 1.0, \delta = 0.5, \omega_2 = 0.3)$$

$$(3)\ \mathcal{Y}_t = \left(\omega_0 + \frac{\omega_1 B}{1 - \delta B} + \frac{\omega_2 B}{1 - B}\right)\xi_t^{(T)} \qquad (\text{with } \omega_0 = 1.5, \omega_1 = -1.0, \delta = 0.5, \omega_2 = -0.5)$$

Compute recursively the response $\mathcal{Y}_t$ for each of these models at times $t = T, T + 1, T + 2, \ldots$ and comment on their possible usefulness in the estimation and/or elimination of effects due to such phenomena as advertising campaigns, promotions, and price changes.

13.3. Figure 13.2 shows the residuals before and after an outlier adjustment for the temperature data in Series C. Construct a similar graph for the viscosity data in Series D.

13.4. A time series defined as $z_t = 1000 \log_{10}(H_t)$, where H_t is the price of hogs recorded annually by the U.S. Census of Agriculture over the period 1867–1948, was considered in Exercise 6.6.

(a) Estimate the parameters of the model identified for this series. Perform diagnostic check to determine the adequacy of the fitted model.

(b) Are additive or innovational outliers present in this series?

(c) If outliers are found, perform the appropriate adjustments to the basic ARIMA model and evaluate the results.

13.5. Daily air quality measurements in New York, May–September 1973, are available in the data file ''airquality'' in the R datasets package. The file provides data on four air quality variables, including the solar radiation measured from 8 a.m. to 12 noon at Central Park. The solar radiation series has a few missing values.

(a) Assuming that an AR(1) is appropriate for the series, derive an expression for the conditional expectation of the missing values, given the available data.

(b) Repeat the derivation in part (a) assuming that an AR(2) model is appropriate for the series.

(c) How would you evaluate the AR assumptions and proceed to develop a suitable model for this series?

14

MULTIVARIATE TIME SERIES ANALYSIS

Multivariate time series analysis involves the use of stochastic models to describe and analyze the relationships among *several* time series. While the focus in most of the earlier chapters has been on univariate methods, we will now assume that k time series, denoted as $z_{1t}, z_{2t}, \ldots, z_{kt}$, are to be analyzed, and we let $\boldsymbol{Z}_t = (z_{1t}, \ldots, z_{kt})'$ denote the time series vector at time t, for $t = 0, \pm 1, \ldots$. Such multivariate processes are of interest in a variety of fields such as economics, business, the social sciences, earth sciences (e.g., meteorology and geophysics), environmental sciences, and engineering. For example, in an engineering setting, one may be interested in the study of the simultaneous behavior over time of current and voltage, or of pressure, temperature, and volume. In economics, we may be interested in the variations of interest rates, money supply, unemployment, and so on, while sales volume, prices, and advertising expenditures for a particular commodity may be of interest in a business context. Multiple time series of this type may be contemporaneously related, some series may lead other series, or there may exist feedback relationships between the series.

In the study of multivariate processes, a framework is needed for describing not only the properties of the individual series but also the possible cross relationships among the series. Two key purposes for analyzing and modeling the series jointly are:

1. To understand the dynamic relationships over time among the series.
2. To improve accuracy of forecasts for individual series by utilizing the additional information available from the related series in the forecasts for each series.

With these objectives in mind, we begin this chapter by introducing some basic concepts and tools that are needed for modeling multivariate time series. We then describe the vector autoregressive, or VAR, models that are widely used in applied work. The properties of

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

these models are examined and methods for model identification, parameter estimation, and model checking are described. This is followed by a discussion of vector moving average and mixed vector autoregressive–moving average models, along with associated modeling tools. A brief discussion of nonstationary unit-root models and cointegration among vector time series is also included. We find that most of the basic concepts and results from univariate time series analysis extend to the multivariate case. However, new problems and challenges arise in the modeling of multivariate time series due to the greater complexity of models and parametrizations in the vector case. Methods designed to overcome such challenges are discussed. For a more detailed coverage of various aspects of multivariate time series analysis, see for example, Reinsel (1997), Lütkepohl (2006), and Tsay (2014).

14.1 STATIONARY MULTIVARIATE TIME SERIES

Let $\boldsymbol{Z}_t = (z_{1t}, \ldots, z_{kt})', t = 0, \pm 1, \pm 2, \ldots$, denote a k-dimensional time series vector of random variables of interest. The choice of the univariate component time series z_{it} that are included in $\boldsymbol{Z}_t$ will depend on the subject matter area and an understanding of the system under study, but it is implicit that the component series will be interrelated both contemporaneously and across time lags. The representation and modeling of these dynamic interrelationships is of main interest in multivariate time series analysis. Similar to the univariate case, an important concept in the model representation and analysis, which enables useful modeling results to be obtained from a finite sample realization of the series, is that of stationarity.

The vector process $\{\boldsymbol{Z}_t\}$ is (strictly) *stationary* if the probability distributions of the random vectors $(\boldsymbol{Z}_{t_1}, \boldsymbol{Z}_{t_2}, \ldots, \boldsymbol{Z}_{t_m})$ and $(\boldsymbol{Z}_{t_1+l}, \boldsymbol{Z}_{t_2+l}, \ldots, \boldsymbol{Z}_{t_m+l})$ are the same for arbitrary times $t_1, t_2, \ldots, t_m$, all m, and all lags or leads $l = 0, \pm 1, \pm 2, \ldots$. Thus, the probability distribution of observations from a stationary vector process is invariant with respect to shifts in time. Hence, assuming finite first and second moments exist, for a stationary process we must have $E[\boldsymbol{Z}_t] = \boldsymbol{\mu}$, constant for all t, where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)'$ is the mean vector of the process. Also, the vectors $\boldsymbol{Z}_t$ must have a constant covariance matrix for all t, which we denote by $\boldsymbol{\Sigma}_z \equiv \boldsymbol{\Gamma}(0) = E[(\boldsymbol{Z}_t - \boldsymbol{\mu})(\boldsymbol{Z}_t - \boldsymbol{\mu})']$. A less stringent definition of second-order, or covariance stationarity will be provided below.

14.1.1 Cross-Covariance and Cross-Correlation Matrices

For a stationary process $\{\boldsymbol{Z}_t\}$ the covariance between z_{it} and $z_{j,t+l}$ must depend only on the lag l, not on time t, for $i, j = 1, \ldots, k, l = 0, \pm 1, \pm 2, \ldots$. Hence, similar to definitions used in Section 12.1.1, we define the cross-covariance between the series z_{it} and z_{jt} at lag l as

$$\gamma_{ij}(l) = \text{cov}[z_{it}, z_{j,t+l}] = E[(z_{it} - \mu_i)(z_{j,t+l} - \mu_j)]$$

and denote the $k \times k$ matrix of *cross-covariances* at lag l as

$$\boldsymbol{\Gamma}(l) = E[(\boldsymbol{Z}_t - \boldsymbol{\mu})(\boldsymbol{Z}_{t+l} - \boldsymbol{\mu})'] = \begin{bmatrix} \gamma_{11}(l) & \gamma_{12}(l) & \cdots & \gamma_{1k}(l) \\ \gamma_{21}(l) & \gamma_{22}(l) & \cdots & \gamma_{2k}(l) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{k1}(l) & \gamma_{k2}(l) & \cdots & \gamma_{kk}(l) \end{bmatrix} \tag{14.1.1}$$

for $l = 0, \pm 1, \pm 2, \ldots$. The corresponding *cross-correlations* at lag l are

$$\rho_{ij}(l) = \text{corr}[z_{it}, z_{j,t+l}] = \frac{\gamma_{ij}(l)}{\{\gamma_{ii}(0)\gamma_{jj}(0)\}^{1/2}}$$

with $\gamma_{ii}(0) = \text{var}[z_{it}]$. Thus, for $i = j$, $\rho_{ii}(l) = \rho_{ii}(-l)$ denotes the autocorrelation function of the ith series z_{it}, and for $i \neq j$, $\rho_{ij}(l) = \rho_{ji}(-l)$ denotes the cross-correlation function between the series z_{it} and z_{jt}. The $k \times k$ *cross-correlation* matrix $\boldsymbol{\rho}(l)$ at lag l, with (i, j)th element equal to $\rho_{ij}(l)$, is given by

$$\boldsymbol{\rho}(l) = \mathbf{V}^{-1/2}\boldsymbol{\Gamma}(l)\mathbf{V}^{-1/2} = \{\rho_{ij}(l)\} \tag{14.1.2}$$

for $l = 0, \pm 1, \pm 2, \ldots$, where $\mathbf{V}^{-1/2} = \text{diag}\{\gamma_{11}(0)^{-1/2}, \ldots, \gamma_{kk}(0)^{-1/2}\}$. Note that $\boldsymbol{\Gamma}(l)' = \boldsymbol{\Gamma}(-l)$ and $\boldsymbol{\rho}(l)' = \boldsymbol{\rho}(-l)$, since $\gamma_{ij}(l) = \gamma_{ji}(-l)$. In addition, the cross-covariance matrices $\boldsymbol{\Gamma}(l)$ and cross-correlation matrices $\boldsymbol{\rho}(l)$ are nonnegative definite, since

$$\text{var}\left[\sum_{i=1}^{n} \boldsymbol{b}_i' \boldsymbol{Z}_{t-i}\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{b}_i' \boldsymbol{\Gamma}(i-j)\boldsymbol{b}_j \geq 0$$

for all positive integers n and all k-dimensional constant vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$.

14.1.2 Covariance Stationarity

The definition of stationarity given above is usually referred to as strict or strong stationarity. In general, a process $\{\boldsymbol{Z}_t\}$ that possesses finite first and second moments and that satisfies the conditions that $E[\boldsymbol{Z}_t] = \boldsymbol{\mu}$ does not depend on t and $E[(\boldsymbol{Z}_t - \boldsymbol{\mu})(\boldsymbol{Z}_{t+l} - \boldsymbol{\mu})']$ depends only on l is referred to as *weak, second-order*, or *covariance stationary*. In this chapter, the term stationary will generally be used in this latter sense of weak stationarity. For a stationary vector process, the cross-covariance and cross-correlation matrices provide useful summary information on the dynamic interrelations among the components of the process. However, because of the higher dimensionality $k > 1$ of the vector process, the cross-correlation matrices generally have more complicated structures and can be much more difficult to interpret than the autocorrelation functions in the univariate case. In Sections 4.2-4.4, we will examine the covariance properties implied by vector autoregressive, moving average, and mixed autoregressive-moving average models.

14.1.3 Vector White Noise Process

The simplest example of a stationary vector process is the *vector white noise process*, which plays a fundamental role as a building block for general vector processes. The vector white noise process is defined as a sequence of random vectors $\ldots, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_t, \ldots$ with $\boldsymbol{a}_t = (a_{1t}, \ldots, a_{kt})'$, such that $E[\boldsymbol{a}_t] = \mathbf{0}$, $E[\boldsymbol{a}_t \boldsymbol{a}_t'] = \boldsymbol{\Sigma}$, and $E[\boldsymbol{a}_t \boldsymbol{a}_{t+l}'] = \mathbf{0}$, for $l \neq 0$. Hence, its covariance matrices $\boldsymbol{\Gamma}(l)$ are given by

$$\boldsymbol{\Gamma}(l) = E[\boldsymbol{a}_t \boldsymbol{a}_{t+l}'] = \begin{cases} \boldsymbol{\Sigma} & \text{for} \quad l = 0 \\ \mathbf{0} & \text{for} \quad l \neq 0 \end{cases} \tag{14.1.3}$$

The $k \times k$ covariance matrix $\boldsymbol{\Sigma}$ is assumed to be positive definite, since the dimension k of the process could be reduced otherwise. Sometimes, additional properties will be assumed for the $\boldsymbol{a}_t$, such as normality or mutual independence over different time periods.

14.1.4 Moving Average Representation of a Stationary Vector Process

A multivariate generalization of Wold's theorem states that if $\{\boldsymbol{Z}_t\}$ is a purely nondeterministic (i.e., $\boldsymbol{Z}_t$ does not contain a purely deterministic component process whose future values can be perfectly predicted from the past values) stationary process with mean vector $\boldsymbol{\mu}$, then $\boldsymbol{Z}_t$ can be represented as an *infinite vector moving average (MA) process*,

$$\boldsymbol{Z}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t-j} = \boldsymbol{\mu} + \boldsymbol{\Psi}(B)\boldsymbol{a}_t \qquad \boldsymbol{\Psi}_0 = \mathbf{I} \tag{14.1.4}$$

where $\boldsymbol{\Psi}(B) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j B^j$ is a $k \times k$ matrix in the backshift operator B such that $B^j \boldsymbol{a}_t = \boldsymbol{a}_{t-j}$ and the $k \times k$ coefficient matrices $\boldsymbol{\Psi}_j$ satisfy the condition $\sum_{j=0}^{\infty} \|\boldsymbol{\Psi}_j\|^2 < \infty$, where $\|\boldsymbol{\Psi}_j\|$ denotes the norm of $\boldsymbol{\Psi}_j$. The $\boldsymbol{a}_t$ form a vector white noise process with mean $\mathbf{0}$ and covariances given by (14.1.3). The covariance matrix of $\boldsymbol{Z}_t$ is then given by

$$\text{Cov}(\boldsymbol{Z}_t) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\Sigma} \boldsymbol{\Psi}_j'$$

The Wold representation in (14.1.4) is obtained by defining $\boldsymbol{a}_t$ as the error $\boldsymbol{a}_t = \boldsymbol{Z}_t - \hat{\boldsymbol{Z}}_{t-1}(1)$ of the best (i.e., minimum mean square error) one-step-ahead linear predictor $\hat{\boldsymbol{Z}}_{t-1}(1)$ of $\boldsymbol{Z}_t$ based on the infinite past $\boldsymbol{Z}_{t-1}, \boldsymbol{Z}_{t-2}, \ldots$. Thus, the $\boldsymbol{a}_t$ are mutually uncorrelated by construction since $\boldsymbol{a}_t$ is uncorrelated with $\boldsymbol{Z}_{t-j}$ for all $j \geq 1$ and, hence, is uncorrelated with $\boldsymbol{a}_{t-j}$ for all $j \geq 1$, and the $\boldsymbol{a}_t$ have a constant covariance matrix by stationarity of the process $\{\boldsymbol{Z}_t\}$. The best one-step-ahead linear predictor can be expressed as

$$\hat{\boldsymbol{Z}}_{t-1}(1) = \boldsymbol{\mu} + \sum_{j=1}^{\infty} \boldsymbol{\Psi}_j \{\boldsymbol{Z}_{t-j} - \hat{\boldsymbol{Z}}_{t-j-1}(1)\} = \boldsymbol{\mu} + \sum_{j=1}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t-j}$$

Consequently, the coefficient matrices $\boldsymbol{\Psi}_j$ in (14.1.4) have the interpretation of the linear regression matrices of $\boldsymbol{Z}_t$ on the $\boldsymbol{a}_{t-j}$ in that $\boldsymbol{\Psi}_j = \text{cov}[\boldsymbol{Z}_t, \boldsymbol{a}_{t-j}]\boldsymbol{\Sigma}^{-1}$.

In what follows, we will assume that $\boldsymbol{\Psi}(B)$ can be represented (at least approximately, in practice) as the product $\boldsymbol{\Phi}^{-1}(B)\boldsymbol{\Theta}(B)$, where $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$ are finite autoregressive and moving average matrix polynomials of orders p and q, respectively. This leads to a class of linear models for vector time series $\boldsymbol{Z}_t$ defined by a relation of the form $\boldsymbol{\Phi}(B)(\boldsymbol{Z}_t - \mu) = \boldsymbol{\Theta}(B)\mathbf{a}_t$, or

$$(\boldsymbol{Z}_t - \boldsymbol{\mu}) - \sum_{j=1}^{p} \boldsymbol{\Phi}_j (\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) = \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j} \tag{14.1.5}$$

A process $\{\boldsymbol{Z}_t\}$ is referred to as a *vector autoregressive--moving average*, or VARMA(p, q), process if it satisfies the relations (14.1.5) for a given white noise sequence $\{\boldsymbol{a}_t\}$.

We begin the discussion of this class of vector models by examining the special case when q is zero so that the process follows a pure vector autoregressive model of order p.

The discussion will focus on time-domain methods for analyzing vector time series and spectral methods will not be used. However, a brief summary of the spectral characteristics of stationary vector processes is provided in Appendix A14.1.

14.2 VECTOR AUTOREGRESSIVE MODELS

Among multivariate time series models, vector autoregressive models are the most widely used in practice. A major reason for this is their similarity to ordinary regression models and the relative ease of fitting these models to actual time series. For example, the parameters can be estimated using least-squares methods that yield closed-form expressions for the estimates. Other methods from multivariate regression analysis can be used at other steps of the analysis. Vector autoregressive models are widely used in econometrics, for example, to describe the dynamic behavior of economic and financial time series and to produce forecasts. This section examines the properties of vector autoregressive models and describes methods for order specification, parameter estimation, and model checking that can be used to develop these models in practice.

14.2.1 VAR(p) Model

A *vector autoregressive* model of order p, or VAR(p) model, is defined as

$$\boldsymbol{\Phi}(B)(\boldsymbol{Z}_t - \boldsymbol{\mu}) = \boldsymbol{a}_t$$

where $\boldsymbol{\Phi}(B) = \mathbf{I} - \boldsymbol{\Phi}_1 B - \boldsymbol{\Phi}_2 B^2 - \cdots - \boldsymbol{\Phi}_p B^p$, $\boldsymbol{\Phi}_i$ is a $k \times k$ parameter matrix, and $\boldsymbol{a}_t$ is a white noise sequence with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The model can equivalently be written as

$$(\boldsymbol{Z}_t - \boldsymbol{\mu}) = \sum_{j=1}^{p} \boldsymbol{\Phi}_j(\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) + \boldsymbol{a}_t \tag{14.2.1}$$

The behavior of the process is determined by the roots of the determinantal equation $\det\{\boldsymbol{\Phi}(B)\} = 0$. In particular, the process is stationary if all the roots of this equation are greater than one in absolute value; that is, lie outside the unit circle (e.g., Reinsel,1997, Chapter 2). When this condition is met, $\{\boldsymbol{Z}_t\}$ has the infinite moving average representation

$$\boldsymbol{Z}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t-j} \tag{14.2.2}$$

or $\boldsymbol{Z}_t = \boldsymbol{\mu} + \boldsymbol{\Psi}(B)\boldsymbol{a}_t$, where $\boldsymbol{\Psi}(B) = \boldsymbol{\Phi}^{-1}(B)$ and the coefficient matrices $\boldsymbol{\Psi}_j$ satisfy the condition $\sum_{j=0}^{\infty} \|\boldsymbol{\Psi}_j\| < \infty$. Then, since $\boldsymbol{\Phi}(B)\boldsymbol{\Psi}(B) = \mathbf{I}$, the coefficient matrices can be calculated recursively from

$$\boldsymbol{\Psi}_j = \boldsymbol{\Phi}_1 \boldsymbol{\Psi}_{j-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{\Psi}_{j-p} \tag{14.2.3}$$

with $\boldsymbol{\Psi}_0 = \mathbf{I}$ and $\boldsymbol{\Psi}_j = \mathbf{0}$, for $j < 0$.

The moving average representation (14.2.2) is useful for examining the covariance properties of the process and it has a number of other applications. As in the univariate case, it is useful for studying forecast errors when the VAR(p) model is used for forecasting.

It is also used in *impulse response analysis* to determine how current or future values of the series are impacted by past changes or ''shocks'' to the system. The coefficient matrix $\boldsymbol{\Psi}_j$ shows the expected impact of a past shock $\boldsymbol{a}_{t-j}$ on the current value $\boldsymbol{Z}_t$. The response of a specific variable to a shock in another variable is often of interest in applied work. However, since the components of $\boldsymbol{a}_{t-j}$ are typically correlated, the individual elements of the $\boldsymbol{\Psi}_j$ can be difficult to interpret. To aid the interpretation, the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{a}_t$ can be diagonalized using a Cholesky decomposition $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}'$, where $\boldsymbol{L}$ is a lower triangular matrix with positive diagonal elements. Then, letting $\boldsymbol{b}_t = \boldsymbol{L}^{-1}\boldsymbol{a}_t$, we have $\mathrm{Cov}(\boldsymbol{b}_t) = \mathbf{I}_k$, and the model can be rewritten as

$$\boldsymbol{Z}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j^* \boldsymbol{b}_{t-j}$$

where $\boldsymbol{\Psi}_0^* = \boldsymbol{L}$ and $\boldsymbol{\Psi}_j^* = \boldsymbol{\Psi}_j \boldsymbol{L}$ for $j > 0$. The matrices $\boldsymbol{\Psi}_j^*$ are called the *impulse response weights* with respect to the orthogonal innovations $\boldsymbol{b}_t$. Since $\boldsymbol{L}$ is a lower triangular matrix, the ordering of the variables will, however, matter in this case. For further discussion and for applications of impulse response analysis, see Lütkepohl (2006, Chapter 2) and Tsay (2014, Chapter 2).

Reduced and Structural Forms. It is sometimes useful to express the VAR(p) process in (14.2.1) in the following slightly different form. Since the matrix $\boldsymbol{\Sigma} = E[\boldsymbol{a}_t \boldsymbol{a}_t']$ is assumed to be positive definite, there exists a lower triangular matrix $\boldsymbol{\Phi}_0^{\#}$ with ones on the diagonal such that $\boldsymbol{\Phi}_0^{\#}\boldsymbol{\Sigma}\boldsymbol{\Phi}_0^{\#\prime} = \boldsymbol{\Sigma}^{\#}$ is a diagonal matrix with positive diagonal elements. Hence, by premultiplying (14.2.1) by $\boldsymbol{\Phi}_0^{\#}$, we obtain the following representation:

$$\boldsymbol{\Phi}_0^{\#}(\boldsymbol{Z}_t - \boldsymbol{\mu}) = \sum_{j=1}^{p} \boldsymbol{\Phi}_j^{\#}(\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) + \boldsymbol{b}_t \tag{14.2.4}$$

where $\boldsymbol{\Phi}_j^{\#} = \boldsymbol{\Phi}_0^{\#}\boldsymbol{\Phi}_j$ and $\boldsymbol{b}_t = \boldsymbol{\Phi}_0^{\#}\boldsymbol{a}_t$ with $\mathrm{Cov}[\boldsymbol{b}_t] = \boldsymbol{\Sigma}^{\#}$. This model displays the concurrent dependence among the components of $\boldsymbol{Z}_t$ through the lower triangular matrix $\boldsymbol{\Phi}_0^{\#}$ and is sometimes referred to as the *structural* form of the VAR(p) model. The model (14.2.1) that includes the concurrent relationships in the covariance matrix $\boldsymbol{\Sigma}$ of the errors and does not show them explicitly is referred to as the standard or *reduced form* of the VAR(p) model. Note that a diagonalizing transformation of this type was already used in the impulse response analysis described above, where the innovations $\boldsymbol{b}_t$'s were further normalized to have unit variance.

14.2.2 Moment Equations and Yule–Walker Estimates

For the VAR(p) model, the covariance matrices $\boldsymbol{\Gamma}(l) = \mathrm{Cov}(\boldsymbol{Z}_t, \boldsymbol{Z}_{t+l}) = \mathrm{Cov}(\boldsymbol{Z}_{t-l}, \boldsymbol{Z}_t) = E[(\boldsymbol{Z}_{t-l} - \boldsymbol{\mu})(\boldsymbol{Z}_t - \boldsymbol{\mu})']$ satisfy the matrix equations

$$\boldsymbol{\Gamma}(l) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(l-j)\boldsymbol{\Phi}_j' \tag{14.2.5}$$

for $l = 1, 2, \ldots$, with $\boldsymbol{\Gamma}(0) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(-j)\boldsymbol{\Phi}_j' + \boldsymbol{\Sigma}$. This result is readily derived using (14.2.1), noting that $E[(\boldsymbol{Z}_{t-l} - \boldsymbol{\mu})\boldsymbol{a}_{t-j}'] = \mathbf{0}$, for $j < l$. The matrix equations (14.2.5) are commonly referred to as the multivariate *Yule–Walker equations* for the VAR(p) model. For $l = 0, \ldots, p$, these equations can be used to solve for the $\boldsymbol{\Gamma}(l)$ simultaneously in terms of the AR parameter matrices $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Sigma}$.

Conversely, the AR coefficient matrices $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p$ and $\boldsymbol{\Sigma}$ can also be determined from the $\boldsymbol{\Gamma}$'s by first solving the Yule–Walker equations, for $l = 1, \ldots, p$, to obtain the parameters $\boldsymbol{\Phi}_j$. These equations can be written in matrix form as $\boldsymbol{\Gamma}_p \boldsymbol{\Phi}_{(p)} = \boldsymbol{\Gamma}_{(p)}$, with solution $\boldsymbol{\Phi}_{(p)} = \boldsymbol{\Gamma}_p^{-1} \boldsymbol{\Gamma}_{(p)}$, where

$$\boldsymbol{\Phi}_{(p)} = [\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p]' \qquad \boldsymbol{\Gamma}_{(p)} = [\boldsymbol{\Gamma}(1)', \ldots, \boldsymbol{\Gamma}(p)']'$$

and $\boldsymbol{\Gamma}_p$ is a $kp \times kp$ matrix with (i, j)th block of elements equal to $\boldsymbol{\Gamma}(i - j)$. Once the $\boldsymbol{\Phi}_j$ are determined from this, $\boldsymbol{\Sigma}$ can be obtained as

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(0) - \sum_{j=1}^{p} \boldsymbol{\Gamma}(-j)\boldsymbol{\Phi}_j' \equiv \boldsymbol{\Gamma}(0) - \boldsymbol{\Gamma}_{(p)}' \boldsymbol{\Phi}_{(p)} = \boldsymbol{\Gamma}(0) - \boldsymbol{\Phi}_{(p)}' \boldsymbol{\Gamma}_p \boldsymbol{\Phi}_{(p)}$$

In practical applications, these results can be used to derive *Yule--Walker estimates* of the parameters in the VAR(p) model by replacing the variance and covariance matrices by their estimates.

14.2.3 Special Case: VAR(1) Model

To examine the properties of VAR models in more detail, we will consider the VAR(1) model,

$$\boldsymbol{Z}_t = \boldsymbol{\Phi}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t$$

where the mean vector $\boldsymbol{\mu}$ is assumed to be zero for convenience. For $k = 2$, we have the bivariate VAR(1) process

$$\boldsymbol{Z}_t = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \boldsymbol{Z}_{t-1} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}$$

or equivalently

$$\begin{aligned} z_{1t} &= \phi_{11} z_{1,t-1} + \phi_{12} z_{2,t-1} + a_{1t} \\ z_{2t} &= \phi_{21} z_{1,t-1} + \phi_{22} z_{2,t-1} + a_{2t} \end{aligned}$$

where ϕ_{11} and ϕ_{22} reflect the dependence of each component on its own past. The parameter ϕ_{12} shows the dependence of z_{1t} on $z_{2,t-1}$ in the presence of $z_{1,t-1}$, while ϕ_{21} shows the dependence of z_{2t} on $z_{1,t-1}$ in the presence of $z_{2,t-1}$. Thus, if $\phi_{12} \neq 0$ and $\phi_{21} \neq 0$, then there is a feedback relationship between the two components. On the other hand, if the off-diagonal elements of the parameter matrix Φ are zero, that is, $\phi_{12} = \phi_{21} = 0$, then z_{1t} and z_{2t} are not dynamically correlated. However, they are still contemporaneously correlated unless $\boldsymbol{\Sigma}$ is a diagonal matrix.

Relationship to Transfer Function Model. If $\phi_{12} = 0$, but $\phi_{21} \neq 0$, then z_{1t} does not depend on past values of z_{2t} but z_{2t} depends on past values of z_{1t}. A *transfer function relationship* then exists with z_{1t} acting as an input variable and z_{2t} as an output variable. However, unless z_{1t} is uncorrelated with a_{2t}, the resulting model is not in the standard transfer function form discussed in Chapter 12. To obtain the standard transfer function model, we let $a_{1t} = b_{1t}$ and $a_{2t} = \beta a_{1t} + b_{2t}$, where β is the regression coefficient of a_{2t} on a_{1t}. Under normality, the error term b_{2t} is then independent of a_{1t} and hence of b_{1t}. The unidirectional transfer function model is obtained by rewriting the equations for z_{1t} and z_{2t} above in terms of the orthogonal innovations b_{1t} and b_{2t}. This yields

$$(1 - \phi_{22}B)z_{2t} = \{\beta + (\phi_{21} - \beta\phi_{11})B\}z_{1,t-1} + b_{2t}$$

where the input variable z_{1t} does not depend on the noise term b_{2t}.

Hence, the bivariate transfer function model emerges as a special case of the bivariate AR model, in which a unidirectional relationship exists between the variables. In general, for a VAR(1) model in higher dimensions, $k > 2$, if the k series can be arranged so that the matrix $\mathbf{\Phi}$ is lower triangular, then the VAR(1) model can also be expressed in the form of unidirectional transfer function equations.

Stationarity Conditions for VAR(1) Model. The VAR(1) process is stationary if the roots of $\det\{\mathbf{I} - \mathbf{\Phi}B\} = 0$ exceed one in absolute value. Since $\det\{\mathbf{I} - \mathbf{\Phi}B\} = 0$ if and only if $\det\{\lambda\mathbf{I} - \mathbf{\Phi}\} = 0$ with $\lambda = 1/B$, it follows that the stationarity condition for the AR(1) model is equivalent to requiring that the eigenvalues of $\mathbf{\Phi}$ be less than one in absolute value. When this condition is met, the process has the convergent infinite MA representation (14.2.2) with MA coefficient matrices $\mathbf{\Psi}_j = \mathbf{\Phi}^j$, since from (14.2.3) the $\mathbf{\Psi}_j$ now satisfy

$$\mathbf{\Psi}_j = \mathbf{\Phi}\mathbf{\Psi}_{j-1} \equiv \mathbf{\Phi}^j\mathbf{\Psi}_0$$

To look at the stationarity for a k-dimensional VAR(1) model further, we note that for arbitrary $n > 0$, by $t + n$ successive substitutions in the right-hand side of $\boldsymbol{Z}_t = \mathbf{\Phi}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t$ we obtain

$$\boldsymbol{Z}_t = \sum_{j=0}^{t+n} \mathbf{\Phi}^j \boldsymbol{a}_{t-j} + \mathbf{\Phi}^{t+n+1}\boldsymbol{Z}_{-n-1}$$

Hence, provided that all eigenvalues of $\mathbf{\Phi}$ are less than one in absolute value, as $n \to \infty$ this will converge to the infinite MA representation $\boldsymbol{Z}_t = \sum_{j=0}^{\infty} \mathbf{\Phi}^j \boldsymbol{a}_{t-j}$, with $\sum_{j=0}^{\infty} \|\mathbf{\Phi}^j\| < \infty$, which is stationary. For example, suppose that $\mathbf{\Phi}$ has k distinct eigenvalues $\lambda_1, \ldots, \lambda_k$, so there is a $k \times k$ nonsingular matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{\Phi}\mathbf{P} = \mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_k)$. Then $\mathbf{\Phi} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$ and $\mathbf{\Phi}^j = \mathbf{P}\mathbf{\Lambda}^j\mathbf{P}^{-1}$, where $\mathbf{\Lambda}^j = \text{diag}(\lambda_1^j, \ldots, \lambda_k^j)$, so when all $|\lambda_i| < 1$, $\sum_{j=0}^{\infty} \|\mathbf{\Phi}^j\| < \infty$ since then $\sum_{j=0}^{\infty} \|\mathbf{\Lambda}^j\| < \infty$.

Moment Equations. For the VAR(1) model, the matrix Yule–Walker equations (14.2.5) simplify to

$$\mathbf{\Gamma}(l) = \mathbf{\Gamma}(l-1)\mathbf{\Phi}' \qquad \text{for } l \geq 1$$

so $\boldsymbol{\Gamma}(1) = \boldsymbol{\Gamma}(0)\boldsymbol{\Phi}'$, in particular, with

$$\boldsymbol{\Gamma}(0) = \boldsymbol{\Gamma}(-1)\boldsymbol{\Phi}' + \Sigma = \boldsymbol{\Phi}\boldsymbol{\Gamma}(0)\boldsymbol{\Phi}' + \boldsymbol{\Sigma}$$

Hence, $\boldsymbol{\Phi}'$ can be determined from $\boldsymbol{\Gamma}(0)$ and $\boldsymbol{\Gamma}(1)$ as $\boldsymbol{\Phi}' = \boldsymbol{\Gamma}(0)^{-1}\boldsymbol{\Gamma}(1)$ and also $\boldsymbol{\Gamma}(l) = \boldsymbol{\Gamma}(0)\boldsymbol{\Phi}'^{l}$. This last relation illustrates that the behavior of all correlations in $\boldsymbol{\rho}(l)$, obtained using (14.1.2), will be controlled by the behavior of the $\lambda_i^l, i = 1, \ldots, k$, where $\lambda_1, \ldots, \lambda_k$ are the eigenvalues of $\boldsymbol{\Phi}$, and shows that even the simple VAR(1) model is capable of fairly general correlation structures (e.g., mixtures of exponential decaying and damping sinusoidal terms) for dimensions $k > 1$. (For more details, see Reinsel, 1997, Section 2.2.3).

14.2.4 Numerical Example

Consider the bivariate ($k = 2$) AR(1) model $(\mathbf{I} - \boldsymbol{\Phi}B)\boldsymbol{Z}_t = \boldsymbol{a}_t$ with

$$\boldsymbol{\Phi} = \begin{bmatrix} 0.8 & 0.7 \\ -0.4 & 0.6 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$$

The roots of $\det\{\lambda\mathbf{I} - \boldsymbol{\Phi}\} = \lambda^2 - 1.4\lambda + 0.76 = 0$ are $\lambda = 0.7 \pm 0.5196i$, with absolute value equal to $(0.76)^{1/2}$; hence, the AR(1) model is stationary. Since the roots are complex, the correlations of this AR(1) process will exhibit damped sinusoidal behavior. The covariance matrix $\boldsymbol{\Gamma}(0)$ is determined by solving the linear equations $\boldsymbol{\Gamma}(0) - \boldsymbol{\Phi}\boldsymbol{\Gamma}(0)\boldsymbol{\Phi}' = \boldsymbol{\Sigma}$. Together with $\boldsymbol{\Gamma}(l) = \boldsymbol{\Gamma}(l-1)\boldsymbol{\Phi}'$, these lead to the covariance matrices

$$\boldsymbol{\Gamma}(0) = \begin{bmatrix} 18.536 & -1.500 \\ -1.500 & 8.884 \end{bmatrix} \qquad \boldsymbol{\Gamma}(1) = \begin{bmatrix} 13.779 & -8.315 \\ 5.019 & 5.931 \end{bmatrix}$$

$$\boldsymbol{\Gamma}(2) = \begin{bmatrix} 5.203 & -10.500 \\ 8.166 & 1.551 \end{bmatrix} \qquad \boldsymbol{\Gamma}(3) = \begin{bmatrix} -3.188 & -8.381 \\ 7.619 & -2.336 \end{bmatrix}$$

$$\boldsymbol{\Gamma}(4) = \begin{bmatrix} -8.417 & -3.754 \\ 4.460 & -4.449 \end{bmatrix} \qquad \boldsymbol{\Gamma}(5) = \begin{bmatrix} -9.361 & 1.115 \\ 0.453 & -4.453 \end{bmatrix}$$

The corresponding correlation matrices are obtained from $\boldsymbol{\rho}(l) = \mathbf{V}^{-1/2}\boldsymbol{\Gamma}(l)\mathbf{V}^{-1/2}$, where $\mathbf{V}^{-1/2} = \text{diag}(18.536^{-1/2}, 8.884^{-1/2})$. The autocorrelations and cross-correlations of this process are displayed up to 18 lags in Figure 14.1. We note that the correlation patterns are rather involved and correlations do not die out very quickly. The coefficients $\boldsymbol{\Psi}_j = \Phi^j$, $j \geq 1$, in the infinite MA representation for this AR(1) process are

$$\boldsymbol{\Psi}_1 = \begin{bmatrix} 0.80 & 0.70 \\ -0.40 & 0.60 \end{bmatrix} \qquad \boldsymbol{\Psi}_2 = \begin{bmatrix} 0.36 & 0.98 \\ -0.56 & 0.08 \end{bmatrix} \qquad \boldsymbol{\Psi}_3 = \begin{bmatrix} -0.10 & 0.84 \\ -0.48 & -0.34 \end{bmatrix}$$

$$\boldsymbol{\Psi}_4 = \begin{bmatrix} -0.42 & 0.43 \\ -0.25 & -0.54 \end{bmatrix} \qquad \boldsymbol{\Psi}_5 = \begin{bmatrix} -0.51 & -0.03 \\ 0.02 & -0.50 \end{bmatrix} \qquad \boldsymbol{\Psi}_6 = \begin{bmatrix} -0.39 & -0.38 \\ 0.22 & -0.28 \end{bmatrix}$$

So the elements of the $\boldsymbol{\Psi}_j$ matrices are also persistent and exhibit damped sinusoidal behavior similar to that of the correlations.

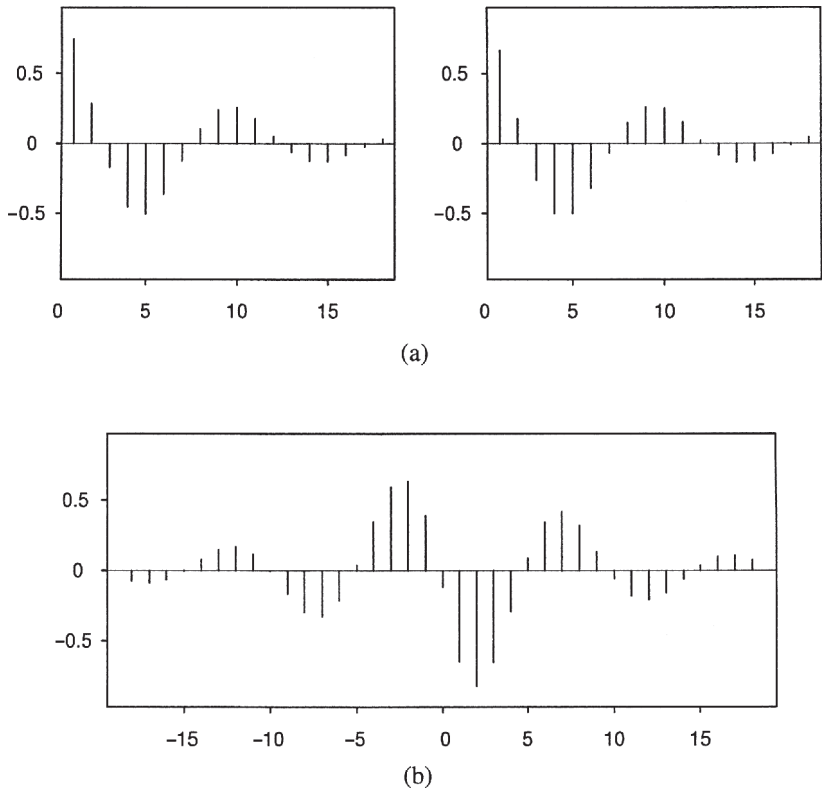


FIGURE 14.1 Theoretical autocorrelations and cross-correlations, $\rho_{ij}(l)$, for the bivariate VAR(1) process example: (a) autocorrelations $\rho_{11}(l)$ and $\rho_{22}(l)$ and (b) cross-correlations $\rho_{12}(l)$.

Finally, since $\det\{\lambda\mathbf{I} - \mathbf{\Phi}\} = \lambda^2 - 1.4\lambda + 0.76 = 0$, it follows from Reinsel (1997, Section 2.2.4) that each individual series z_{it} has a univariate ARMA(2, 1) model representation as $(1 - 1.4B + 0.76B^2)z_{it} = (1 - \eta_i B)\varepsilon_{it}$, $\sigma^2_{\varepsilon_i} = \text{var}[\varepsilon_{it}]$, where η_i and $\sigma^2_{\varepsilon_i}$ are readily determined. For a k-dimensional VAR(p) model, it can be shown that each individual component z_{it} follows a univariate ARMA of maximum order $(kp, (k-1)p)$. The order can be much less if the AR and MA polynomials have common factors (e.g., Wei, 2006, Chapter 16).

Computations in R. The covariance matrices $\mathbf{\Gamma}(l)$ and the $\mathbf{\Psi}$ matrices shown above can be reproduced using the MTS package in R as follows:

```
> library(MTS)
> phi1=matrix(c(0.8,-0.4,0.7,0.6),2,2)
> sig=matrix(c(4,1,1,2),2,2)
> eigen(phi1)
> m1=VARMAcov(Phi=phi1, Sigma=sig, lag=5)
> names(m1)
  [1] "autocov" "ccm"
> autocov=t(m1$autocov)
```

```
> m2=PSIwgt(Phi=phi1)
> names(m2)
  [1] "psi.weight" "irf"
> m2$psi.weight
```

The command VARMAcov() computes the covariance and cross-correlation matrices up to 12 lags by default. These matrices need to be transposed using the command t() since MTS defines the lag l covariance matrix $\boldsymbol{\Gamma}(l)$ as $E[(\boldsymbol{Z}_t - \boldsymbol{\mu})(\boldsymbol{Z}_{t-l} - \boldsymbol{\mu})']$, whereas the definition $E[(\boldsymbol{Z}_{t-l} - \boldsymbol{\mu})(\boldsymbol{Z}_t - \boldsymbol{\mu})']$ is used in this chapter. Transposing the matrices makes the results from R consistent with our definition. The command eigen(phi1) included in the code gives the eigenvalues of the matrix $\boldsymbol{\Phi}$.

14.2.5 Initial Model Building and Least-Squares Estimation for VAR Models

Given an observed vector time series $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N$ of length N from a multivariate process, the development of an appropriate VAR model for the series can be performed iteratively using a three-stage procedure of model specification, parameter estimation, and diagnostic checking. In the VAR case, the model specification involves choosing a suitable value for the order p. Some useful tools at this stage include the sample covariance and correlation matrices described below and the sample partial autoregression matrices discussed, for example, by Tiao and Box (1981). The latter quantities are analogous to the partial autocorrelations used in the univariate case and are estimated as the last autoregressive matrix, $\boldsymbol{\Phi}_m$, in a VAR(m) model with $m = 1, 2, \ldots$. The estimates $\hat{\boldsymbol{\Phi}}_m$ can be derived from the Yule–Walker equations or by least-squares estimation of the parameter matrices. Statistical tests are used to determine the significance of the estimates for each value of m. The partial autoregression matrices are zero for all lags greater than p and are thus particularly useful for identifying the autoregressive model order. Additional methods for model selection include the use of information criteria such as AIC, BIC, and HC, as well as methods based on canonical correlation analysis described later in this chapter.

Sample Covariance and Correlation Matrices. Given an observed time series, the sample covariance matrix of the $\boldsymbol{Z}_t$ at lag l is defined as

$$\hat{\boldsymbol{\Gamma}}(l) = \mathbf{C}(l) = \frac{1}{N}\sum_{t=1}^{N-l}(\boldsymbol{Z}_t - \overline{\boldsymbol{Z}})(\boldsymbol{Z}_{t+l} - \overline{\boldsymbol{Z}})' \quad l = 0, 1, 2, \ldots \tag{14.2.6}$$

where $\overline{\boldsymbol{Z}} = (\overline{z}_1, \ldots, \overline{z}_k)' = N^{-1}\sum_{t=1}^{N}\boldsymbol{Z}_t$ is the sample mean vector, which is a natural estimator of the process mean vector $\boldsymbol{\mu} = E[\boldsymbol{Z}_t]$ in the stationary case. In particular, $\hat{\boldsymbol{\Gamma}}(0) = \mathbf{C}(0) = N^{-1}\sum_{t=1}^{N}(\boldsymbol{Z}_t - \overline{\boldsymbol{Z}})(\boldsymbol{Z}_t - \overline{\boldsymbol{Z}})'$ is the sample covariance matrix of the $\boldsymbol{Z}_t$. The (i, j)th element of $\hat{\boldsymbol{\Gamma}}(l)$ is given by

$$\hat{\gamma}_{ij}(l) = c_{ij}(l) = \frac{1}{N}\sum_{t=1}^{N-l}(z_{it} - \overline{z}_i)(z_{j,t+l} - \overline{z}_j)$$

The *sample cross-correlations* are defined as

$$\hat{\rho}_{ij}(l) = r_{ij}(l) = \frac{c_{ij}(l)}{\{c_{ii}(0)c_{jj}(0)\}^{1/2}} \quad i, j = 1, \ldots, k$$

For a stationary series, the $\hat{\rho}_{ij}(l)$ are sample estimates of the theoretical $\rho_{ij}(l)$. The asymptotic sampling properties of sample correlations $\hat{\rho}_{ij}(l)$ were discussed earlier in Section 12.1.3. The expressions for the asymptotic variances and covariances of the estimates are complicated but simplify in certain cases. For example, in the special case where $\boldsymbol{Z}_t$ is a white noise process, the results give $\text{var}[\hat{\rho}_{ij}(l)] \simeq 1/(N-l)$.

The sample cross-correlation matrices are important tools for the initial specification of a model for the series $\boldsymbol{Z}_t$. They are particularly useful in the model specification for a low-order pure vector moving average model, which has the property that $\rho_{ij}(l) = 0$ for all $l > q$, as discussed in Section 14.3 below. However, similar to the univariate case, a slowly decaying pattern in the estimated autocorrelation and cross-correlation matrices would indicate that autoregressive terms are needed.

Estimation of the Partial Autoregression Matrices. Consider the vector autoregressive model of order m, $\boldsymbol{Z}_t = \boldsymbol{\delta} + \sum_{j=1}^{m} \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j} + \boldsymbol{a}_t$, where $\boldsymbol{\delta} = (\mathbf{1} - \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_m)\boldsymbol{\mu}$ accounts for the non-zero mean vector. Estimates of the partial autoregressive matrices can be obtained from the Yule–Walker equations in (14.2.5) as

$$\hat{\boldsymbol{\Phi}}_{(m)} = [\hat{\boldsymbol{\Phi}}_{1m}, \ldots, \hat{\boldsymbol{\Phi}}_{mm}]' = \hat{\boldsymbol{\Gamma}}_m^{-1} \hat{\boldsymbol{\Gamma}}_{(m)}$$

The estimate of the error covariance matrix estimate is $\hat{\boldsymbol{\Sigma}}_m = \hat{\boldsymbol{\Gamma}}(0) - \sum_{j=1}^{m} \hat{\boldsymbol{\Gamma}}(-j)\hat{\boldsymbol{\Phi}}_{jm}'$. The estimation is performed for $m = 1, 2, \ldots$, yielding a sequence of estimates $\hat{\boldsymbol{\Phi}}_{mm}$ of the last parameter matrix in the VAR(m) model. These matrices are referred to as partial autoregression matrices by Tiao and Box (1981).

An asymptotically equivalent procedure is to estimate the partial autoregression matrices using multivariate linear least-squares (LS) estimation described, for example, by Johnson and Wichern (2007). Using this approach, the components of $\boldsymbol{Z}_t$ are regressed on the lagged vector values $\boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_{t-m}$, by first writing the VAR(m) model in regression form as

$$\boldsymbol{Z}_t = \boldsymbol{\delta} + \sum_{j=1}^{m} \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j} + \boldsymbol{a}_t = \boldsymbol{\delta} + \boldsymbol{\Phi}_{(m)}' \boldsymbol{X}_t + \boldsymbol{a}_t \tag{14.2.7}$$

with $\boldsymbol{X}_t = (\boldsymbol{Z}_{t-1}', \ldots, \boldsymbol{Z}_{t-m}')'$. The LS estimates for the AR parameters are then given by

$$\hat{\boldsymbol{\Phi}}_{(m)} = [\hat{\boldsymbol{\Phi}}_{1m}, \ldots, \hat{\boldsymbol{\Phi}}_{mm}]' = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\boldsymbol{Z}} \tag{14.2.8}$$

where the matrices $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{X}}$, respectively, have typical rows $(\boldsymbol{Z}_t - \overline{\boldsymbol{Z}}_{(0)})'$ and

$$[(\boldsymbol{Z}_{t-1} - \overline{\boldsymbol{Z}}_{(1)})', \ldots, (\boldsymbol{Z}_{t-m} - \overline{\boldsymbol{Z}}_{(m)})'] \quad t = m+1, \ldots, N$$

with $\overline{\boldsymbol{Z}}_{(i)} = n^{-1} \sum_{t=m+1}^{N} \boldsymbol{Z}_{t-i}$ and $n = N - m$. The estimate of the error covariance matrix $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}}_m = [n - (km+1)]^{-1} \mathbf{S}_m \tag{14.2.9}$$

where

$$\mathbf{S}_m = \sum_{t=m+1}^{N} \hat{\boldsymbol{a}}_t \hat{\boldsymbol{a}}_t'$$

is the residual sum-of-squares matrix and

$$\hat{\boldsymbol{a}}_t = (\boldsymbol{Z}_t - \overline{\boldsymbol{Z}}_{(0)}) - \sum_{j=1}^{m} \hat{\boldsymbol{\Phi}}_j (\boldsymbol{Z}_{t-j} - \overline{\boldsymbol{Z}}_{(j)})$$

are the residual vectors. These LS estimators $\hat{\boldsymbol{\Phi}}_j$ are also the conditional maximum likelihood (ML) estimators under the normality assumption. Asymptotic distribution theory for the LS estimators in the stationary VAR model was provided by Hannan (1970, Chapter 6). Under a stationary VAR(m) model, the distribution of vec[$\hat{\boldsymbol{\Phi}}_{(m)}$] is approximately multivariate normal with mean vector vec[$\boldsymbol{\Phi}_{(m)}$] and covariance matrix estimated by $\hat{\boldsymbol{\Sigma}}_m \otimes (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$, where $\otimes$ denotes the Kronecker product of $\hat{\boldsymbol{\Sigma}}_m$ and $(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$.

Sequential Likelihood Ratio Tests. The estimation of the partial autoregression matrices is supplemented by likelihood ratio tests that are applied sequentially to help determine the model order p. (e.g., see Tiao and Box (1981) and Reinsel (1997, Chapter 4)). Thus, after fitting a VAR(m) model, we test the null hypothesis H_0: $\boldsymbol{\Phi}_{mm} = \mathbf{0}$ against the alternative $\boldsymbol{\Phi}_{mm} \neq \mathbf{0}$, using the likelihood ratio (LR) statistic

$$M_m = -\left(n - mk - \frac{1}{2}\right) \ln\left[\frac{|\mathbf{S}_m|}{|\mathbf{S}_{m-1}|}\right] \tag{14.2.10}$$

where $\mathbf{S}_m$ is the residual sum-of-squares matrix defined above, and $n = N - m - 1$ is the effective number of observations assuming that the model includes a constant term. For large n, when H_0: $\boldsymbol{\Phi}_{mm} = \mathbf{0}$ is true, the statistic M_m has an approximate χ^2 distribution with k^2 degrees of freedom, and we reject H_0 for large values of M_m. The LR test statistic in (14.2.10) is asymptotically equivalent to a Wald statistic formed in terms of the LS estimator $\hat{\boldsymbol{\Phi}}_{mm}$ of $\boldsymbol{\Phi}_{mm}$.

This procedure is a natural extension of the use of the sample PACF $\hat{\phi}_{mm}$ for identification of the order of an AR model in the univariate case as described in Section 6.2. However, unlike the univariate case, the partial autoregression matrices are not partial autocorrelation matrices (or correlations of any kind) in the vector case. Similar tests based on the sample partial autocorrelation matrices, whose elements are proper correlation coefficients, are described by Reinsel (1997, Chapter 4) and Wei (2006, Chapter 16).

Use of Information Criteria. Model selection criteria such as AIC, BIC, and HQ can also be employed for model specification. Here, AIC represents Akaike's information criterion (Akaike, 1974a), BIC is the Bayesian information criterion due to Schwarz (1978), and HQ is the model selection criterion proposed by Hannan and Quinn (1979); see also Quinn (1980). These criteria are likelihood based and include under normality the determinant of the innovations covariance matrix that reflects the goodness of fit of the model. A second term is a function of the number of fitted parameters and penalizes models that are unnecessarily complex. For the VAR model, we have

$$\begin{aligned}
\text{AIC}_m &= \ln\{|\tilde{\boldsymbol{\Sigma}}_m|\} + 2mk^2/N \\
\text{BIC}_m &= \ln\{|\tilde{\boldsymbol{\Sigma}}_m|\} + mk^2 \ln(N)/N \\
\text{HQ}_m &= \ln\{|\tilde{\boldsymbol{\Sigma}}_m|\} + 2mk^2 \ln(\ln(N))/N
\end{aligned}$$

where N is the sample size, m is the VAR order, and $\tilde{\boldsymbol{\Sigma}}_m$ is the corresponding ML residual covariance matrix estimate of $\boldsymbol{\Sigma}$. It can be seen that BIC imposes a greater ''penalty factor'' for the number of estimated parameters than does AIC, while HQ is intermediate between AIC and BIC. Other similar measures include the final prediction error (FPE) criterion suggested by Akaike (1971). These criteria can be used to compare models fitted using maximum likelihood and the model that gives the lowest value for a given criterion would be selected. For a discussion of the properties and performance of different model selection criteria, see, for example, Quinn (1980) and Lütkepohl (2006).

14.2.6 Parameter Estimation and Model Checking

Parameter Estimation. With the order of the VAR model specified, the model parameters can be estimated using the least-squares procedure described above. For a stationary process, the Yule–Walker estimates are asymptotically equivalent to the least-squares estimates. However, when the process is nonstationary or near nonstationary, it is known that the least-squares estimator still performs consistently, whereas the Yule–Walker estimator may have a considerable bias. Hence, the least-squares method is generally to be preferred (e.g., Reinsel, 1997, Section 4.4). Under the normality assumption, the least-squares estimates are equivalent to conditional maximum likelihood estimates. Exact maximum likelihood estimates can be derived using the unconditional likelihood function described for VARMA models in Section 14.4.5. However, use of the conditional likelihood function simplifies the calculations and is often adequate for VAR models in practice.

Model Checking. Model diagnostics of the estimated VAR model are primarily based on examination of the residual vectors $\hat{\boldsymbol{a}}_t$ from the estimated model and their sample covariance matrices. The residuals $\hat{\boldsymbol{a}}_t$ are calculated from (14.2.1) with the parameters replaced by their estimates $\hat{\boldsymbol{\Phi}}_j$. Useful diagnostic checks include plots of the residuals against time and/or against other variables, and detailed examination of the cross-correlation matrices of the residuals. Approximate two-standard-error limits can be imposed to assess the statistical significance of the residual correlations.

In addition, overall portmanteau or ''goodness-of-fit'' tests based on the residual covariance matrices at several lags can be employed for model checking; see, for example, Hosking (1980), Li and McLeod (1981), Poskitt and Tremayne (1982), and Ali (1989). Specifically, using s lags, an overall goodness-of-fit test statistic, analogous to that proposed by Ljung and Box (1978) for the univariate case, is given by

$$Q_s = N^2 \sum_{l=1}^{s} (N-l)^{-1} \mathrm{tr}[\hat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{a}}}(l) \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{a}}}(l)' \hat{\boldsymbol{\Sigma}}^{-1}] \tag{14.2.11}$$

where

$$\hat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{a}}}(l) = N^{-1} \sum_{t=1}^{N-l} \hat{\boldsymbol{a}}_t \hat{\boldsymbol{a}}'_{t+l} \quad l = 0, 1, \ldots, s$$

with $\hat{\boldsymbol{\Gamma}}_{\hat{\boldsymbol{a}}}(0) \approx \hat{\boldsymbol{\Sigma}}$. Under the null hypothesis of model adequacy, the test statistic Q_s is approximately distributed as chi-squared with $k^2(s-p)$ degrees of freedom. The fitted model is rejected as inadequate for large values of Q_s. Mahdi and McLeod (2012) extended the portmanteau test of Peña and Rodríguez (2002, 2006) described in Chapter 8 to the

multivariate case and proposed a test based on the determinant of the autocorrelation matrix of the multivariate residuals. Alternative tests such as score or Lagrange multiplier (LM) tests have also been proposed in the literature. For a discussion of the LM tests and their relationship to portmanteau tests, see, for example, Reinsel (1997) and Lütkepohl (2006).

14.2.7 An Empirical Example

To illustrate the model building procedure for a vector process outlined above, we consider the bivariate time series of U.S. fixed investment and change in business inventories. These data are quarterly, seasonally adjusted, and are given in Lütkepohl (2006). The fixed investment data for the time period 1947 to 1971 are shown in Figure 14.2, and the changes in business inventories series for the same period are shown in Figure 14.3(b). Since the investment series is clearly nonstationary, the first differences of this series, which are displayed in Figure 14.3(a), are considered as series z_{1t} together with the change in business inventories as series z_{2t}, resulting in $N = 99$ quarterly observations.

Sample cross-correlation matrices of the series $\boldsymbol{Z}_t = (z_{1t}, z_{2t})'$ for lags 1 through 12 are shown in Table 14.1, and these sample autocorrelations and cross-correlations $\hat{\rho}_{ij}(l)$ are also displayed up to 18 lags in Figure 14.4. Included in Figure 14.4 are the rough guidelines of the two-standard-error limits $\pm 2/\sqrt{N} \simeq \pm 0.2$, which are appropriate for the $\hat{\rho}_{ij}(l)$ from a vector white noise process as noted in Section 14.2.5. These sample correlations show exponentially decaying and damped sinusoidal behavior as a function of lag l, indicative of autoregressive dependence structure in the series.

To select a suitable model, we apply the sequential likelihood ratio test and the three information criteria discussed above to the data. The calculations are performed using the MTS package in R and the results are summarized in Table 14.2. We note that the three criteria AIC_m, BIC_m, and HQ_m all attain a minimum at $m = 2$. The likelihood ratio statistic M_m also supports the value $m = 2$, although a slight discrepancy occurs at $m = 4$. These results therefore indicate that, among pure autoregressive models, a second-order VAR(2) model may be the most appropriate for these data.

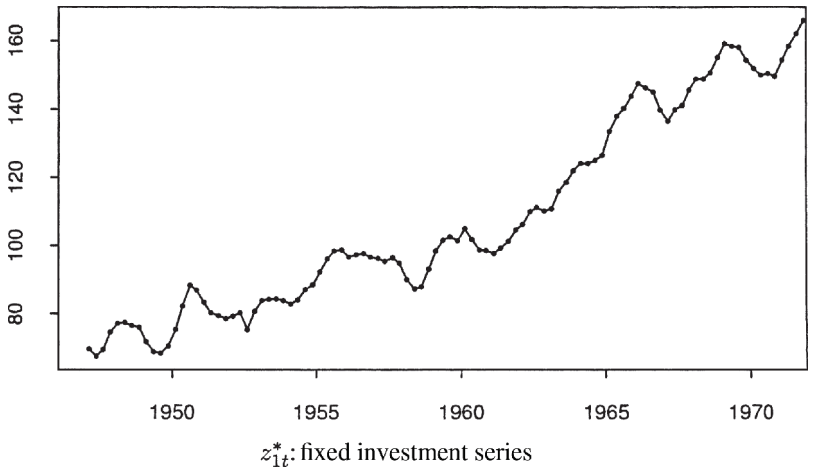


FIGURE 14.2 Quarterly (seasonally adjusted) U.S. fixed investment data for 1947 through 1971.

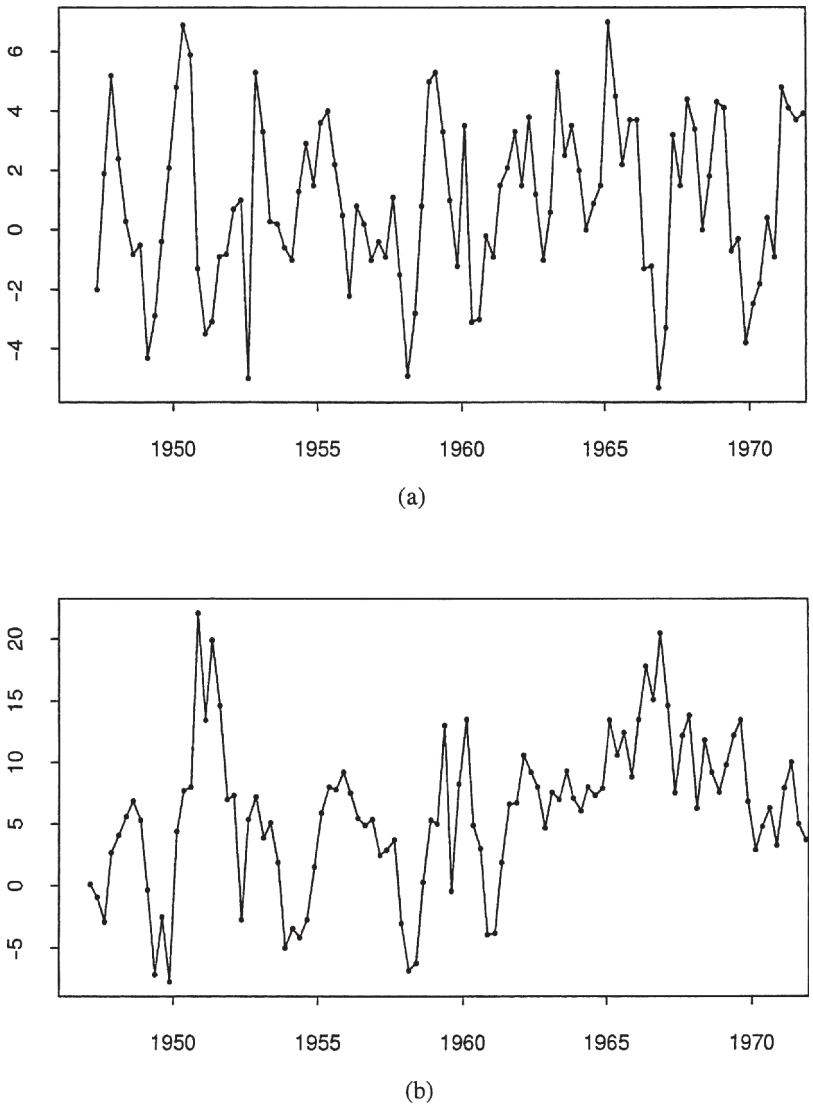


FIGURE 14.3 Quarterly (seasonally adjusted) first differences of U.S. fixed investment data and changes in business inventories data (in billions) for the period 1947 through 1971: (a) z_{1t}: first differences of investment series, $z_{1t} = z_{1t}^* - z_{1,t-1}^*$; and (b) z_{2t}: changes in business inventories series.

TABLE 14.1 Sample Correlation Matrices $\hat{\rho}(l)$ for the Bivariate Quarterly Series of First Differences of U.S. Fixed Investment and U.S. Changes in Business Inventories

l	1		2		3		4		5		6	
$\hat{\rho}(l)$	0.47	0.27	0.10	0.35	−0.12	0.29	−0.31	0.27	−0.30	0.19	−0.21	0.04
	−0.06	0.68	−0.33	0.50	−0.29	0.32	−0.21	0.07	−0.10	0.07	0.10	0.01
l	7		8		9		10		11		12	
$\hat{\rho}(l)$	−0.14	−0.04	−0.09	−0.11	0.13	−0.03	0.19	0.07	0.13	0.12	0.02	0.20
	0.15	0.04	0.20	0.05	0.12	0.05	0.05	0.11	0.01	0.09	−0.04	0.05

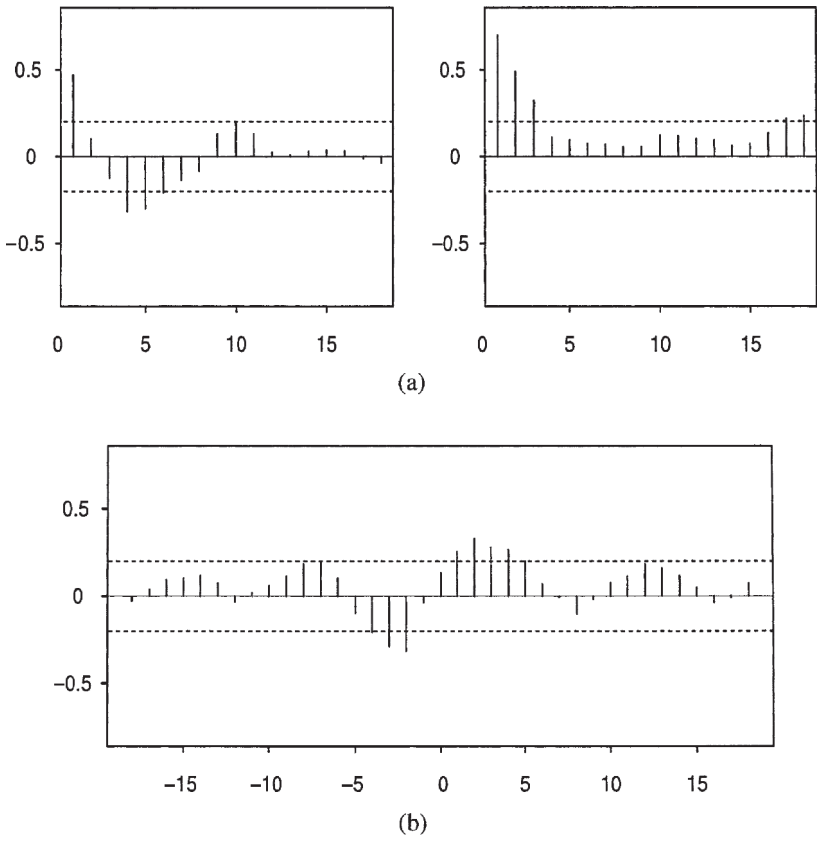


FIGURE 14.4 Sample auto- and cross-correlations $\hat{\rho}_{ij}(l)$ for the bivariate series of first differences of U.S. fixed investment and U.S. changes in business inventories: (a) sample autocorrelations $\hat{\rho}_{11}(l)$ and $\hat{\rho}_{22}(l)$ and (b) sample cross-correlations $\hat{\rho}_{12}(l)$.

TABLE 14.2 Order Selection Statistics for the U.S. Business Investment and Inventories Data

m (VAR Order)	AIC_m	BIC_m	HQ_m	M_m	p-Value
0	5.539	5.539	5.539	0.000	0.000
1	4.723	4.828	4.766	73.997	0.000
2	4.597	4.807	4.682	16.652	0.002
3	4.659	4.974	4.786	1.483	0.830
4	4.614	5.033	4.784	9.628	0.047
5	4.624	5.148	4.836	5.283	0.260
6	4.703	5.332	4.958	0.113	0.999
7	4.759	5.493	5.056	1.785	0.775
8	4.785	5.623	5.124	3.755	0.440

The LS estimates from the AR(2) model (with estimated standard errors in parentheses), as well as the ML estimate of $\boldsymbol{\Sigma}$, are given as

$$\hat{\boldsymbol{\Phi}}_1 = \begin{bmatrix} 0.504 & 0.108 \\ (0.096) & (0.056) \\ 0.345 & 0.531 \\ (0.177) & (0.103) \end{bmatrix} \quad \hat{\boldsymbol{\Phi}}_2 = \begin{bmatrix} -0.146 & -0.205 \\ (0.099) & (0.054) \\ 0.256 & 0.139 \\ (0.181) & (0.099) \end{bmatrix}$$

$$\tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} 5.0270 & 1.6958 \\ 1.6958 & 16.9444 \end{bmatrix}$$

with $|\tilde{\boldsymbol{\Sigma}}| = 82.3032$. The estimates of the two constant terms are 1.217 and 1.527, with respective standard errors of 0.354 and 0.650. In the matrix $\hat{\boldsymbol{\Phi}}_2$, the coefficient estimate in the (1, 2) position is statistically significant, while the rest are insignificant and might perhaps be omitted.

We now examine the residuals $\hat{\boldsymbol{a}}_t$ from the fitted VAR(2) model. The residual autocorrelations and cross-correlations are displayed in Figure 14.5. The approximate two-standard-error limits are also included in the graphs. The individual elements of the residual correlation matrices are generally quite small for all lags through $l = 12$, with $|\hat{\rho}_{\hat{a},ij}(l)| \ll 2/\sqrt{N} = 0.2$ in nearly all cases. One notable feature of these residual correlations, however, is the (marginally) significant correlation of $\hat{\rho}_{\hat{a},22}(4) = -0.20$ at lag 4 for the second residual series $\hat{a}_{2t}$ (see lower right panel of Figure 14.5). This feature, which also appears visible from the p-values of the portmanteau test shown in Figure 14.6, may be a consequence of the seasonal adjustment procedure, related to a weak seasonal structure that may still exist in the quarterly (''seasonally adjusted'') series $\boldsymbol{Z}_t$. To accommodate this feature, we could consider a modification to the VAR(2) model by inclusion of an MA coefficient matrix $\boldsymbol{\Theta}_4$ at the quarterly seasonal lag of 4 in the model. Although this could lead to a small improvement, we do not pursue this modification here.

As a benchmark for comparison against the bivariate AR(2) model fitted above, comparable *univariate* models for z_{1t} and z_{2t} that were found to be adequate, estimated by the conditional ML method, were obtained as

$$(1 - 1.275B + 0.545B^2)z_{1t} = 0.251 + (1 - 0.769B)\varepsilon_{1t}$$

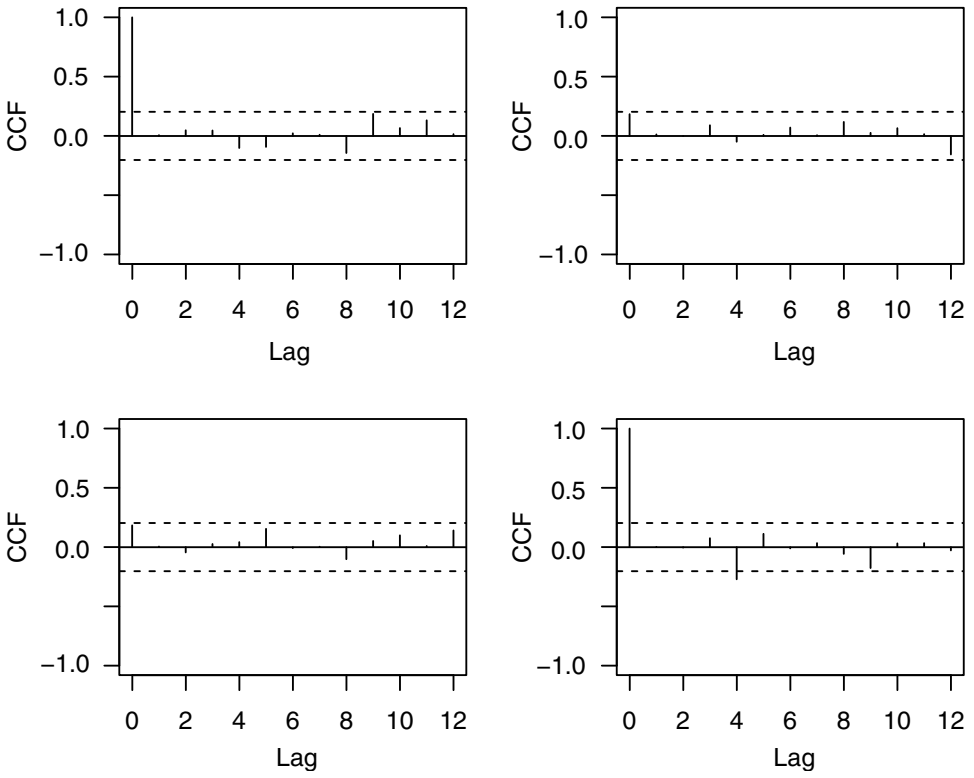


FIGURE 14.5 Cross-correlation matrices for the residuals from the VAR(2) model fitted to the U.S. business investment and inventories data.

with $\hat{\sigma}^2_{\varepsilon_1} = 5.44$, and $(1 - 0.690B)z_{2t} = 1.808 + \varepsilon_{2t}$, with $\hat{\sigma}^2_{\varepsilon_2} = 19.06$. Note that the residual variances are slightly larger in this case. The fitted bivariate models imply that the changes in business inventories series z_{2t} have a modest but significant influence on the (first differences of) investments z_{1t}, but there appears to be less influence in the feedback from investments to the changes in inventories series. In addition, there is only a small degree of contemporaneous correlation suggested, since the correlation between the residual series $\hat{a}_{1t}$ and $\hat{a}_{2t}$ in the bivariate models estimated from $\tilde{\boldsymbol{\Sigma}}$ equals 0.184.

Remark. The bivariate analysis described above was performed using the multivariate time series package MTS in R. Letting zz denote the data after differencing the investments series, the relevant commands are

```
> ccm(zz)            % Cross-correlation analysis
> m1=VARorder(zz)    % Order selection
> m2=VAR(zz,2)       % Estimation of VAR(2) model
> MTSdiag(m2)        % Model checking
> ccm(m2$residuals) % Residual cross-correlation analysis
```

For more detailed discussion and for demonstrations of the analysis capabilities of the MTS package in R, see Tsay (2014). Multivariate time series tools are also available in other

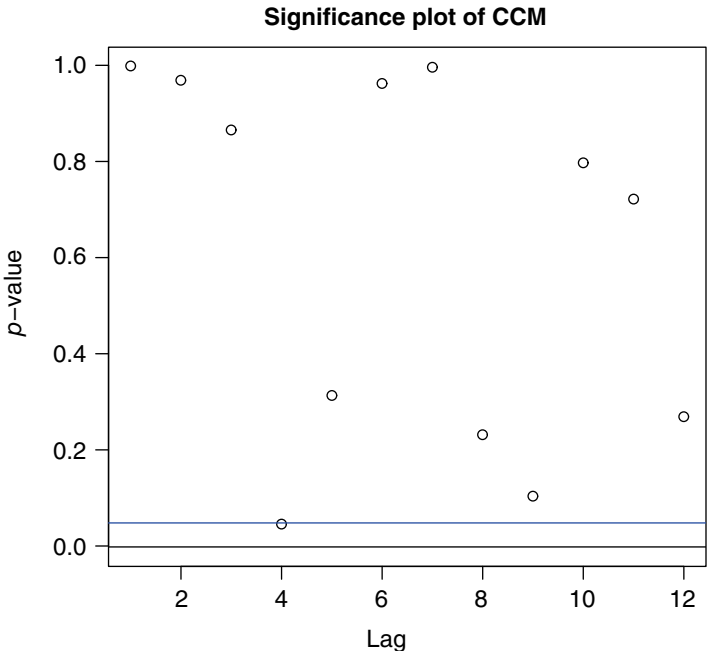


FIGURE 14.6 Plot of p-values of the multivariate portmanteau statistic applied to the residuals from the fitted VAR(2) model.

packages such as the SCA package released by Scientific Computing Associates Corp., and the S-Plus software package available from TIBCO Software, Inc.

14.3 VECTOR MOVING AVERAGE MODELS

The vector autoregressive models described above provide an adequate representation to many applied time series and are widely used in practice. However, pure autoregressive models have a disadvantage in that the model order needed to obtain a satisfactory representation can in some cases be rather high. Analogous to the univariate case, a more parsimonious representation can sometimes be achieved by adding moving average terms to the model. This would result in the vector ARMA (or VARMA) model form mentioned briefly in Section 14.1.4. Aggregation of vector series across time or in space also creates a need for VARMA models as noted e.g. by Lütkepohl and Poskitt (1996). In addition, trend or seasonal adjustments may change the dependence structure and make a pure VAR model inadequate (e.g., Maravall, 1993). Prior to discussing the VARMA model in more detail, we will briefly examine the special case when no autoregressive terms are present and the series follows a pure moving average model.

14.3.1 Vector MA(q) Model

A vector moving average model of order q, or VMA(q) model, is defined as

$$\boldsymbol{Z}_t = \boldsymbol{\mu} + \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j} \tag{14.3.1}$$

or equivalently, $\boldsymbol{Z}_t = \boldsymbol{\mu} + \boldsymbol{\Theta}(B)\boldsymbol{a}_t$, where $\boldsymbol{\mu}$ is the mean of the process, $\boldsymbol{\Theta}(B) = \mathbf{I} - \boldsymbol{\Theta}_1 B - \cdots - \boldsymbol{\Theta}_q B^q$ is a matrix polynomial of order q, and the $\boldsymbol{\Theta}_i$ are $k \times k$ matrices with $\boldsymbol{\Theta}_q \neq 0$.

Invertibility. A vector MA(q) process is said to be *invertible* if it can be represented in the form

$$(\boldsymbol{Z}_t - \boldsymbol{\mu}) - \sum_{j=1}^{\infty} \boldsymbol{\Pi}_j (\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) = \boldsymbol{a}_t \tag{14.3.2}$$

or equivalently as $\boldsymbol{\Pi}(B)(\boldsymbol{Z}_t - \boldsymbol{\mu}) = \boldsymbol{a}_t$ where $\boldsymbol{\Pi}(B) = \mathbf{I} - \sum_{j=1}^{\infty} \boldsymbol{\Pi}_j B^j$, with $\sum_{j=1}^{\infty} \|\boldsymbol{\Pi}_j\| < \infty$. The process is invertible if all the roots of $\det\{\boldsymbol{\Theta}(B)\} = 0$ are greater than one in absolute value. The process then has the infinite VAR representation given by (14.3.2) with $\boldsymbol{\Pi}(B) = \boldsymbol{\Theta}^{-1}(B)$ so that $\boldsymbol{\Theta}(B)\boldsymbol{\Pi}(B) = \mathbf{I}$. As in the univariate case, this form is particularly useful for determining how forecasts of future observations depend on current and past values of the k series.

Moment Equations. For the VMA(q) model, the covariance matrices $\boldsymbol{\Gamma}(l)$ are given by

$$\boldsymbol{\Gamma}(l) = \sum_{h=0}^{q-l} \boldsymbol{\Theta}_h \boldsymbol{\Sigma} \boldsymbol{\Theta}'_{h+l} \tag{14.3.3}$$

for $l = 0, 1, \ldots, q$, with $\boldsymbol{\Theta}_0 = -\mathbf{I}$, and $\boldsymbol{\Gamma}(l) = \mathbf{0}$, for $l > q$. The result is readily verified since the $\{\boldsymbol{a}_t\}$ form a white noise sequence and $\mathrm{Cov}[\boldsymbol{\Theta}\boldsymbol{a}_t] = \boldsymbol{\Theta}\boldsymbol{\Sigma}\boldsymbol{\Theta}'$.

14.3.2 Special Case: Vector MA(1) Model

To examine the properties further, we consider the VMA(1) model, $\boldsymbol{Z}_t - \boldsymbol{\mu} = \boldsymbol{a}_t - \boldsymbol{\Theta}\boldsymbol{a}_{t-1}$. From the same reasoning as given concerning the stationarity condition for the VAR(1) process, the invertibility condition for the VMA(1) model is equivalent to all eigenvalues of $\boldsymbol{\Theta}$ being less than one in absolute value. Then we have the convergent infinite VAR representation (14.3.2) with infinite VAR coefficient matrices $\boldsymbol{\Pi}_j = -\boldsymbol{\Theta}^j$, $j \geq 1$. This follows since $\boldsymbol{\Theta}(B)\boldsymbol{\Pi}(B) = \mathbf{I}$ now simplifies to $\boldsymbol{\Pi}_j = \boldsymbol{\Theta}\boldsymbol{\Pi}_{j-l} \equiv \boldsymbol{\Theta}^j \boldsymbol{\Pi}_0$ with $\boldsymbol{\Pi}_0 = -\mathbf{I}$. Also, from (14.3.3) the covariance matrices of the VMA(1) process simplify to

$$\boldsymbol{\Gamma}(0) = \boldsymbol{\Sigma} + \boldsymbol{\Theta}\boldsymbol{\Sigma}\boldsymbol{\Theta}', \quad \boldsymbol{\Gamma}(1) = -\boldsymbol{\Sigma}\boldsymbol{\Theta}' = \boldsymbol{\Gamma}(-1)'$$

and $\boldsymbol{\Gamma}(l) = \mathbf{0}$ for $|l| > 1$. Thus, as in the univariate MA(1) case, all covariances are zero for lags greater than one.

14.3.3 Numerical Example

Consider the bivariate ($k = 2$) VMA(1) model $\boldsymbol{Z}_t = (\mathbf{I} - \boldsymbol{\Theta}B)\boldsymbol{a}_t$ with

$$\boldsymbol{\Theta} = \begin{bmatrix} 0.8 & 0.7 \\ -0.4 & 0.6 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$$

Similar to results for the VAR(1) example, the roots of $\det\{\lambda\mathbf{I} - \boldsymbol{\Theta}\} = \lambda^2 - 1.4\lambda + 0.76 = 0$ are $\lambda = 0.7 \pm 0.5196i$, with absolute value equal to $(0.76)^{1/2}$; hence, the VMA(1) model is invertible. The coefficient matrices $\boldsymbol{\Pi}_j = -\boldsymbol{\Theta}^j$ in the infinite VAR form are of the same

magnitudes as the $\boldsymbol{\Psi}_j$ coefficient matrices in the previous AR(1) example. The covariance matrices of the MA(1) at lags 0 and 1 are

$$\boldsymbol{\Gamma}(0) = \boldsymbol{\Sigma} + \boldsymbol{\Theta}\boldsymbol{\Sigma}\boldsymbol{\Theta}' = \begin{bmatrix} 8.66 & 0.76 \\ 0.76 & 2.88 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Gamma}(1) = -\boldsymbol{\Sigma}\boldsymbol{\Theta}' = \begin{bmatrix} -3.9 & 1.0 \\ -2.2 & -0.8 \end{bmatrix}$$

with corresponding correlation matrices

$$\boldsymbol{\rho}(0) = \mathbf{V}^{-1/2}\boldsymbol{\Gamma}(0)\mathbf{V}^{-1/2} = \begin{bmatrix} 1.000 & 0.152 \\ 0.152 & 1.000 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\rho}(1) = \begin{bmatrix} -0.450 & 0.200 \\ -0.441 & -0.278 \end{bmatrix}$$

The above calculations are conveniently performed in R as follows:

```
> library(MTS)
> theta1=matrix(c(0.8,-0.4,0.7,0.6),2,2)
> sig=matrix(c(4,1,1,2),2,2)
> eigen(theta1)
> PIwgt(Theta=theta1)
> m1=VARMAcov(Theta=theta1, Sigma=sig, lag=1)
> names(m1)
 [1] "autocov" "ccm"
> autocov=t(m1$autocov)
> autocorr=t(m1$ccm)
```

For the bivariate MA(1) model, it follows from the autocovariance structure that each series has a univariate MA(1) model representation as $z_{it} = (1 - \eta_i B)\varepsilon_{it}$, $\sigma^2_{\varepsilon_i} = \text{var}(\varepsilon_{it})$. From Appendix A4.3, the parameter values η_i and $\sigma^2_{\varepsilon_i}$ of the component series can be determined directly by solving the relations $\rho_{ii}(1) = -\eta_i/(1 + \eta_i^2)$, $\gamma_{ii}(0) = \sigma^2_{\varepsilon_i}(1 + \eta_i^2)$, $i = 1, 2$, which lead to the values $\eta_1 = 0.628$, $\sigma^2_{\varepsilon 1} = 6.211$, and $\eta_2 = 0.303$, $\sigma^2_{\varepsilon 2} = 2.637$, respectively.

14.3.4 Model Building for Vector MA Models

The model building tools discussed for VAR models in Section 14.2 extend in a straightforward way to moving average models. As noted, the estimated cross-covariance and cross-correlation matrices are particularly useful for specifying the model order q since from (14.3.3) the corresponding theoretical quantities are zero for lags greater than q. The partial autoregression matrices, on the other hand, would show a decaying pattern for a moving average process. The parameter estimates can be obtained using the least-squares method that is equivalent to conditional likelihood method under the normality assumption. However, analogous to the univariate case, the unknown presample values can have a larger impact on the parameter estimates for VMA models. In particular, if the true parameter values are close to the boundary of the invertibility region, the conditional likelihood approach can result in biased estimates, especially for relatively short series. Because of this, the use of the *unconditional likelihood function* is typically recommended for models with moving average terms. We return to the parameter estimation in Section 14.4.5 where the exact likelihood function is discussed for the general VARMA case.

14.4 VECTOR AUTOREGRESSIVE–MOVING AVERAGE MODELS

We now assume that the matrix $\boldsymbol{\Psi}(B)$ can be represented as the product $\boldsymbol{\Psi}(B) = \boldsymbol{\Phi}^{-1}(B)\boldsymbol{\Theta}(B)$, where $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$ are the autoregressive and moving average matrix polynomials defined above. This leads to the vector model

$$(\boldsymbol{Z}_t - \boldsymbol{\mu}) - \sum_{j=1}^{p} \boldsymbol{\Phi}_j(\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) = \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j} \tag{14.4.1}$$

where $\boldsymbol{a}_t$ again is a vector white noise process with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = E[\boldsymbol{a}_t \boldsymbol{a}_t']$. The resulting process $\{\boldsymbol{Z}_t\}$ is referred to as a vector *autoregressive–moving average*, or VARMA(p, q), process regardless of whether $\{\boldsymbol{Z}_t\}$ is stationary or not.

As for the VAR(p) model, the VARMA(p, q) process can be expressed in *structural* form by premultiplying both sides of (14.4.1) by a lower triangular matrix $\boldsymbol{\Phi}_0^{\#}$ with ones on the diagonal such that $\boldsymbol{\Phi}_0^{\#}\boldsymbol{\Sigma}\boldsymbol{\Phi}_0^{\#\prime} = \boldsymbol{\Sigma}^{\#}$ is a diagonal matrix with positive diagonal elements. This gives the following representation:

$$\boldsymbol{\Phi}_0^{\#}(\boldsymbol{Z}_t - \boldsymbol{\mu}) - \sum_{j=1}^{p} \boldsymbol{\Phi}_j^{\#}(\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) = \boldsymbol{b}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j^{\#} \boldsymbol{b}_{t-j} \tag{14.4.2}$$

where $\boldsymbol{\Phi}_j^{\#} = \boldsymbol{\Phi}_0^{\#}\boldsymbol{\Phi}_j$, $\boldsymbol{\Theta}_j^{\#} = \boldsymbol{\Phi}_0^{\#}\boldsymbol{\Theta}_j\boldsymbol{\Phi}_0^{\#-1}$, and $\boldsymbol{b}_t = \boldsymbol{\Phi}_0^{\#}\boldsymbol{a}_t$. This model displays the concurrent dependence among the components of $\boldsymbol{Z}_t$ through the lower triangular matrix $\boldsymbol{\Phi}_0^{\#}$, with diagonal elements for $\boldsymbol{\Sigma}^{\#}$, whereas the standard or *reduced form* (14.4.1) places the concurrent relationships in the covariance matrix $\boldsymbol{\Sigma}$ of the errors. More generally, premultiplication of (14.4.1) by an arbitrary nonsingular matrix $\boldsymbol{\Phi}_0^{\#}$ yields a form similar to (14.4.2) that is useful in some cases. For example, representation of a VARMA model in this general form, but with a special structure imposed on the parameter matrices, will sometimes be more useful for model specification than the standard form (14.4.1). This is discussed further in Section 14.7.

14.4.1 Stationarity and Invertibility Conditions

The stationarity conditions for a VARMA(p, q) process are the same as for the VAR(p) process discussed in Section 14.2. Hence it can be shown that the process is stationary and has an infinite moving average representation $\boldsymbol{Z}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t-j}$ if all the roots of $\det\{\boldsymbol{\Phi}(B)\} = 0$ are greater than one in absolute value. The coefficient matrices $\boldsymbol{\Psi}_j$ are determined from the relation $\boldsymbol{\Phi}(B)\boldsymbol{\Psi}(B) = \boldsymbol{\Theta}(B)$, and satisfy the recursion

$$\boldsymbol{\Psi}_j = \boldsymbol{\Phi}_1\boldsymbol{\Psi}_{j-1} + \boldsymbol{\Phi}_2\boldsymbol{\Psi}_{j-2} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{\Psi}_{j-p} - \boldsymbol{\Theta}_j \qquad j = 1, 2, \ldots \tag{14.4.3}$$

where $\boldsymbol{\Psi}_0 = \mathbf{I}, \boldsymbol{\Psi}_j = \mathbf{0}$ for $j < 0$, and $\boldsymbol{\Theta}_j = \mathbf{0}$ for $j > q$.

Conversely, the VARMA(p, q) process is invertible with an infinite AR representation similar to (14.3.2) if all the roots of $\det\{\boldsymbol{\Theta}(B)\} = 0$ are greater than one in absolute value. The coefficient weights $\boldsymbol{\Pi}_j$ in the infinite AR representation are given by the relation

$\boldsymbol{\Theta}(B)\boldsymbol{\Pi}(B) = \boldsymbol{\Phi}(B)$, and satisfy the recursion

$$\boldsymbol{\Pi}_j = \boldsymbol{\Theta}_1\boldsymbol{\Pi}_{j-1} + \boldsymbol{\Theta}_2\boldsymbol{\Pi}_{j-2} + \cdots + \boldsymbol{\Theta}_q\boldsymbol{\Pi}_{j-q} + \boldsymbol{\Phi}_j \qquad j = 1, 2, \ldots \tag{14.4.4}$$

where $\boldsymbol{\Pi}_0 = -\mathbf{I}, \boldsymbol{\Pi}_j = \mathbf{0}$ for $j < 0$, and $\boldsymbol{\Phi}_j = \mathbf{0}$ for $j > p$.

In addition, using the moving average representation, the covariance matrices for $\boldsymbol{Z}_t$ can be written as $\boldsymbol{\Gamma}(l) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j\boldsymbol{\Sigma}\boldsymbol{\Psi}'_{j+l}, l \geq 0$. From this it follows that the covariance matrix-generating function is given by $\mathbf{G}(z) = \sum_{l=-\infty}^{\infty} \boldsymbol{\Gamma}(l)z^l = \boldsymbol{\Psi}(z^{-1})\boldsymbol{\Sigma}\boldsymbol{\Psi}(z)'$; hence, the spectral density matrix of the VARMA(p, q) process is given as in (A14.1.7) with $\boldsymbol{\Psi}(z) = \boldsymbol{\Phi}^{-1}(z)\boldsymbol{\Theta}(z)$.

14.4.2 Covariance Matrix Properties of VARMA Models

For the general stationary VARMA(p, q) process $\{\boldsymbol{Z}_t\}$, it follows from the infinite MA representation $\boldsymbol{Z}_t = \boldsymbol{\mu} + \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j\boldsymbol{a}_{t-j}$ that

$$E[\boldsymbol{Z}_{t-l}\boldsymbol{a}'_{t-j}] = \begin{cases} \mathbf{0} & \text{for } j < l \\ \boldsymbol{\Psi}_{j-l}\boldsymbol{\Sigma} & \text{for } j \geq l \end{cases}$$

Therefore, it is easy to determine from (14.4.1) that the covariance matrices $\boldsymbol{\Gamma}(l) = E[(\boldsymbol{Z}_{t-l} - \boldsymbol{\mu})(\boldsymbol{Z}_t - \boldsymbol{\mu})']$ of $\{\boldsymbol{Z}_t\}$ satisfy the relations

$$\boldsymbol{\Gamma}(l) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(l-j)\boldsymbol{\Phi}'_j - \sum_{j=1}^{q} \boldsymbol{\Psi}_{j-l}\boldsymbol{\Sigma}\boldsymbol{\Theta}'_j \qquad l = 0, 1, \ldots, q \tag{14.4.5}$$

and $\boldsymbol{\Gamma}(l) = \sum_{j=1}^{p} \boldsymbol{\Gamma}(l-j)\boldsymbol{\Phi}'_j$ for $l > q$, with the convention that $\boldsymbol{\Theta}_0 = -\mathbf{I}$. Thus, the $\boldsymbol{\Gamma}(l)$ can be evaluated in terms of the AR and MA parameter matrices $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Theta}_j$, and $\boldsymbol{\Sigma}$, using these recursions.

14.4.3 Nonuniqueness and Parameter Identifiability for VARMA Models

Although the VARMA(p, q) model appears to be a straightforward extension of the univariate ARMA(p, q) model, a number of issues are associated with this extension. For example, since each AR or MA term contributes $k \times k$ parameters, the total number of parameters in the model increases rapidly as the order increases. The overflow of parameters, whose estimates can be highly correlated, makes the interpretation of the modeled results very difficult. An additional problem that arises in the VARMA case relates to the nonuniqueness of the parameters and the lack of an identifiable model representation. This issue does not arise for the pure VAR(p) model or the pure VMA(q) model discussed earlier in this chapter. But in the vector case it is possible to have two ARMA representations, $\boldsymbol{\Phi}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}(B)\boldsymbol{a}_t$ and $\boldsymbol{\Phi}_*(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}_*(B)\boldsymbol{a}_t$ with different parameters, that give rise to the same coefficients $\boldsymbol{\Psi}_j$ in the infinite MA representation, such that

$$\boldsymbol{\Psi}(B) = \boldsymbol{\Phi}^{-1}(B)\boldsymbol{\Theta}(B) = \boldsymbol{\Phi}_*^{-1}(B)\boldsymbol{\Theta}_*(B)$$

Thus, the two models also give rise to the same covariance matrix structure $\{\boldsymbol{\Gamma}(l)\}$ and hence the same process.

Two VARMA models with this property are said to be *observationally equivalent*, or the models are said to be *exchangeable*. As a basic example, the bivariate VARMA(1, 1)

model $(\mathbf{I} - \boldsymbol{\Phi}_* B)\boldsymbol{Z}_t = (\mathbf{I} - \boldsymbol{\Theta}_* B)\boldsymbol{a}_t$ with parameters

$$\boldsymbol{\Phi}_* = \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} \qquad \boldsymbol{\Theta}_* = \begin{bmatrix} 0 & \beta \\ 0 & 0 \end{bmatrix}$$

is observationally equivalent to both a VAR(1) model $(\mathbf{I} - \boldsymbol{\Phi}B)\boldsymbol{Z}_t = \boldsymbol{a}_t$ and a VMA(1) model $\boldsymbol{Z}_t = (\mathbf{I} - \boldsymbol{\Theta}B)\boldsymbol{a}_t$, with

$$\boldsymbol{\Phi} \equiv -\boldsymbol{\Theta} = \begin{bmatrix} 0 & (\alpha - \beta) \\ 0 & 0 \end{bmatrix}$$

since, for example, $(\mathbf{I} - \boldsymbol{\Phi}_* B)^{-1}(\mathbf{I} - \boldsymbol{\Theta}_* B) = (\mathbf{I} + \boldsymbol{\Phi}_* B)(\mathbf{I} - \boldsymbol{\Phi}_* B) = (\mathbf{I} - \boldsymbol{\Theta}\boldsymbol{B})$. Hence, the parameters $\boldsymbol{\Phi}_*$ and $\boldsymbol{\Theta}_*$ in the ARMA(1, 1) model representation are not identifiable, since the properties of the process depend only on the value of $\alpha - \beta$.

In general, observationally equivalent ARMA(p, q) representations can exist because matrix AR and MA operators could be related by a common left matrix factor $\mathbf{U}(B)$ as

$$\boldsymbol{\Phi}_*(B) = \mathbf{U}(B)\boldsymbol{\Phi}(B) \quad \text{and} \quad \boldsymbol{\Theta}_*(B) = \mathbf{U}(B)\boldsymbol{\Theta}(B)$$

but such that the orders of $\boldsymbol{\Phi}_*(B)$ and $\boldsymbol{\Theta}_*(B)$ are not increased over those of $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$. This common left factor $\mathbf{U}(B)$ would cancel when $\boldsymbol{\Phi}_*^{-1}(B)\boldsymbol{\Theta}_*(B)$ is formed, resulting in the same parameter matrices in $\boldsymbol{\Psi}(\boldsymbol{B})$. A particular ARMA model specification and its parameters are said to be *identifiable* if the $\boldsymbol{\Phi}_j$ and the $\boldsymbol{\Theta}_j$ are uniquely determined by the set of impulse response matrices $\boldsymbol{\Psi}_j$ in the infinite MA representation, or equivalently by the set of covariance matrices $\boldsymbol{\Gamma}(l)$ in the stationary case.

For the mixed VARMA(p, q) model, certain conditions are needed on the matrix operators $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$ to ensure uniqueness of the parameters in the ARMA representation. In addition to the stationarity and invertibility conditions, the following two conditions are sufficient for identifiability:

1. The matrices $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$ have no common left factors other than unimodular ones. That is, if $\boldsymbol{\Phi}(B) = \mathbf{U}(B)\boldsymbol{\Phi}_1(B)$ and $\boldsymbol{\Theta}(B) = \mathbf{U}(B)\boldsymbol{\Theta}_1(B)$, then the common factor $\mathbf{U}(B)$ must be unimodular, that is, $\det\{\mathbf{U}(B)\}$ is a nonzero constant. When this property holds, $\boldsymbol{\Phi}(B)$ and $\boldsymbol{\Theta}(B)$ are called left-coprime.
2. With q as small as possible and p as small as possible for that q, the joint matrix $[\boldsymbol{\Phi}_p, \boldsymbol{\Theta}_q]$ must be of rank k, the dimension of $\boldsymbol{Z}_t$.

Notice that through the relation $\mathbf{U}(B)^{-1} = [1/\det\{\mathbf{U}(B)\}]\text{adj}\{\mathbf{U}(B)\}$, the operator $\mathbf{U}(B)$ is a unimodular matrix if and only if $\mathbf{U}(B)^{-1}$ is a matrix polynomial of finite order. The operator $\mathbf{U}(B) = \mathbf{I} - \boldsymbol{\Phi}_* B$ in the simple ARMA(1, 1) example above is an illustration of a unimodular matrix. For further discussion of the identifiability conditions for the VARMA(p, q) model, see, for example, Hannan and Deistler (1988, Chapter 2) and Reinsel (1997, Chapter 2).

14.4.4 Model Specification for VARMA Processes

The model specification tools discussed for VAR(p) models in Section 14.2 extend in principle to the VARMA case. This includes the examination of the cross-correlation and partial autoregression matrices as discussed by Tiao and Box (1981). Additional tools

include the information criteria for model specification examined earlier, and the use of extended cross-correlation matrices for VARMA models discussed by Tiao and Tsay (1983). However, because of the identifiability issue and the overflow of parameters in the vector case, additional model specification tools focusing on the parameter structure of the VARMA representation are now needed.

Kronecker Indices. Beyond the specification of overall orders p and q, the structure of the VARMA(p, q) model can be characterized by a set of Kronecker indices $K_1, \ldots, K_k$ and the McMillan degree $M = \sum_{i=1}^{k} K_i$ of the process. The Kronecker indices, also known as structural indices, represent the maximal row degrees of the individual equations of the VARMA model. The use of these indices leads to the specification of a VARMA process of order $p = q = \max\{K_i\}$ with certain simplifying structure in the parameter matrices $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Theta}_j$. A Kronecker index equal to K_i, in particular, implies that a VARMA representation can be constructed for the process such that the ith rows of the matrices $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Theta}_j$ are zero for $j > K_i$ and with zero constraints imposed on certain other elements of $\boldsymbol{\Phi}_j$. The resulting model is referred to as the *echelon canonical form* of the VARMA model. The set of Kronecker indices is unique for a given VARMA process and the identifiability issue discussed above is thus avoided. The echelon form structure and identifiability conditions in terms of the echelon form have been examined extensively by Hannan and Deistler (1988) and others.

The Kronecker indices can be estimated using canonical correlation analysis methods introduced by Akaike (1976) and further elaborated upon by Cooper and Wood (1982) and Tsay (1989a). These methods, which are extensions of the canonical correlation analysis procedures discussed for the univariate case in Section 6.2.4, are employed to determine the nonzero canonical correlations between the past and present values of the process, $\{\boldsymbol{Z}_{t-j}, j \geq 0\}$, and the future values $\{\boldsymbol{Z}_{t+j}, j > 0\}$. In this way, the Kronecker indices K_i can be deduced, which then provide the overall model order as well as the maximum order of the AR and MA polynomials for each individual component. Further details of this approach will be given in Section 14.7. More extensive accounts of the Kronecker index approach to model specification have been provided by Solo (1986), Reinsel (1997), Lütkepohl (2006), and Tsay (1989b, 1991, 2014), among others.

Scalar Component Models. Tiao and Tsay (1989) proposed an alternative way to identify the order structure of the VARMA model based on the concept of *scalar component models* (SCMs). This approach examines linear combinations of the observed series with the goal of arriving at a parsimonious model representation that overcomes the identification issue and that may reveal meaningful structures in the data. Using this approach, k independent linear combinations $y_{it} = \boldsymbol{v}_i' \boldsymbol{Z}_t$ of orders (p_i, q_i), $i = 1, \ldots, k$, are sought such that the orders $p_i + q_i$ are as small as possible. Given a k-dimensional VARMA(p, q) process, a nonzero linear combination $y_t = \boldsymbol{v}' \boldsymbol{Z}_t$ follows SCM(p_1, q_1) if

$$y_t - \sum_{j=1}^{p_1} \boldsymbol{v}' \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j} = \boldsymbol{v}' \boldsymbol{a}_t - \sum_{j=1}^{q_1} \boldsymbol{v}' \boldsymbol{\Theta}_j a_{t-j}$$

where $0 \leq p_1 \leq p$, $0 \leq q_1 \leq q$, and $u_t = y_t - \sum_{j=1}^{p_1} \boldsymbol{v}' \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j}$ is uncorrelated with $\boldsymbol{a}_{t-j}$ for $j > q_1$. Notice that the scalar component y_t depends only on lags 1 to p_1 of all variables

$\boldsymbol{Z}_t$, and lags 1 to q_1 of all the innovations $\boldsymbol{a}_t$. Starting from SCM(0, 0), the SCM method uses a sequence of canonical correlation tests to discover k such linear combinations.

Once such a set has been found, the specification of the ARMA structure for $\boldsymbol{Z}_t$ can be determined through the relations

$$\boldsymbol{T}\boldsymbol{Z}_t - \sum_{j=1}^{p} \boldsymbol{G}_j \boldsymbol{Z}_{t-j} = \boldsymbol{T}\boldsymbol{a}_t + \sum_{j=1}^{q} \boldsymbol{H}_j \boldsymbol{a}_{t-j} \tag{14.4.6}$$

where $\boldsymbol{T} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k]'$ is a $k \times k$ nonsingular matrix, $\boldsymbol{G}_j = \boldsymbol{T}\boldsymbol{\Phi}_j, j = 1, \ldots, p$, $\boldsymbol{H}_j = \boldsymbol{T}\boldsymbol{\Theta}_j, j = 1, \ldots, q$, $p = \max\{p_i\}$ and $q = \max\{q_i\}$. Moreover, the ith row of $\boldsymbol{G}_j$ is specified to be zero for $j > p_i$ and the ith row of $\boldsymbol{H}_j$ is zero for $j > q_i$. Premultiplication of (14.4.6) by $\boldsymbol{T}^{-1}$ thus leads to a VARMA(p, q) model for $\boldsymbol{Z}_t$ in standard form but such that the coefficient matrices $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Theta}_j$ have a reduced-rank structure. On the other hand, inserting the factor $\boldsymbol{T}^{-1}\boldsymbol{T}$ in front of the $\boldsymbol{Z}_{t-j}$ and $\boldsymbol{a}_{t-j}$ in (14.4.6) yields a VARMA(p, q) representation for the *transformed process* $\boldsymbol{Y}_t = \boldsymbol{T}\boldsymbol{Z}_t$ as

$$\boldsymbol{Y}_t - \sum_{j=1}^{p} \boldsymbol{\Phi}_j^* Y_{t-j} = \boldsymbol{e}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j^* e_{t-j}$$

where $\boldsymbol{\Phi}_j^* = \boldsymbol{G}_j\, \boldsymbol{T}^{-1} = \boldsymbol{T}\boldsymbol{\Phi}_j\, \boldsymbol{T}^{-1}, \boldsymbol{\Theta}_j^* = \boldsymbol{H}_j\, \boldsymbol{T}^{-1} = \boldsymbol{T}\boldsymbol{\Theta}_j\, \boldsymbol{T}^{-1}$, and $\boldsymbol{e}_t = \boldsymbol{T}\boldsymbol{a}_t$. This VARMA representation for the transformed process is parsimonious in the sense that the ith row of $\boldsymbol{\Phi}_j^*$ is zero for $j > p_i$ and the ith row of $\boldsymbol{\Theta}_j^*$ is zero for $j > q_i$. In addition, some elements of the ith row of $\boldsymbol{\Theta}_j^*$, for $i = 1, \ldots, q_i$, are specified to be zero to remove possible redundancy of the parameters in the AR and MA matrices. The method used to identify and eliminate redundant parameters is referred to as *the rule of elimination*.

The approach of Tiao and Tsay (1989) thus identifies the scalar component processes $\boldsymbol{Y}_t = \boldsymbol{T}\boldsymbol{Z}_t$ and their associated orders (p_j, q_j) through canonical correlation methods, and then estimates a VARMA process for the transformed variables $\boldsymbol{Y}_t$ with zero constraints imposed on some of the parameters. By comparison, the Kronecker index approach estimates Kronecker indices that lead to the echelon model form for the original series $\boldsymbol{Z}_t$ directly. Also, the scalar component allows the orders of the AR and MA polynomials to differ while the orders are the same for the Kronecker index approach. The scalar component approach may in this regard be viewed as a refinement over the Kronecker index approach.

More detailed comparisons of the Kronecker index and the SCM model specification methods are provided by Reinsel (1997) and Tsay (1989b, 1991, 2014). A comparison of the forecasting performance of models specified by the two approaches was reported by Athanasopoulos et al. (2012), who found the results for SCM more favorable. Software modeling tools are available for both methods in the MTS package in R; for details and demonstrations, see Tsay (2014).

Order Determination Using Linear Least Squares. Before we proceed to discuss parameter estimation in the next section, we will mention another method that has been considered for VARMA model specification. This is a multivariate extension of the two-stage linear least-squares regression approach presented for the univariate case by Hannan and Rissanen (1982) and briefly discussed in Section 6.2.4. At the first stage of this procedure, the VARMA model is approximated by a high-order pure VAR model and the least squares method is used to obtain an estimate $\hat{\boldsymbol{a}}_t$ of the white noise error process $\boldsymbol{a}_t$. In the second

stage, one regresses $\boldsymbol{Z}_t$ on the lagged $\boldsymbol{Z}_{t-j}$ and lagged $\hat{\boldsymbol{a}}_{t-j}$ for various combinations of p and q. A model selection criterion such as BIC is then employed to help select appropriate orders for the VARMA model. Use of this procedure may lead to one or two models that seem highly promising, which are later estimated by more efficient procedures such as the maximum likelihood method. Similar linear estimation methods have been proposed by Hannan and Kavalieris (1984), Poskitt (1992), and Lütkepohl and Poskitt (1996), among others, for determining the Kronecker index structure of the VARMA model.

14.4.5 Estimation and Model Checking for VARMA Models

Once a well-defined VARMA model has been specified, the estimation of the parameters is typically performed using maximum likelihood methods assuming normality. In the past, conditional likelihood approaches were often employed for computational convenience. In the VARMA(p, q) model, this corresponds to treating the unknown presample values of $\boldsymbol{Z}_t$ and $\boldsymbol{a}_t$ as fixed constants with the $\boldsymbol{a}_t, t = 0, \ldots, 1-q$, typically set equal to zero. However, for many mixed models with an MA operator $\boldsymbol{\Theta}(B)$ having roots near the unit circle, the conditional likelihood approach has been shown to produce estimates with poorer finite sample properties than the unconditional, or exact, ML estimates.

Various approaches to the construction of the exact Gaussian likelihood function have been considered in the literature. Earlier classical approaches to evaluate the exact likelihood were presented by Hillmer and Tiao (1979) and Nicholls and Hall (1979). Given N observations $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$, the exact likelihood of a stationary VARMA(p, q) model $\boldsymbol{\Phi}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}(B)\boldsymbol{a}_t$ has the form

$$L = |\boldsymbol{\Sigma}|^{-N/2}|\boldsymbol{\Omega}|^{-1/2}|\mathbf{D}|^{-1/2} \exp\left\{ -\left(\frac{1}{2}\right)\left[\sum_{t=1}^{N} \hat{\boldsymbol{a}}_t'\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{a}}_t + \hat{\mathbf{a}}_*'\boldsymbol{\Omega}^{-1}\hat{\mathbf{a}}_*\right]\right\} \tag{14.4.7}$$

where $\mathbf{a}_* = (\boldsymbol{Z}_{1-p}', \ldots, \boldsymbol{Z}_0', \boldsymbol{a}_{1-q}', \ldots, \boldsymbol{a}_0')'$ denotes the vector of presample values, $\hat{\mathbf{a}}_* = E[\mathbf{a}_* | \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N]$ represents the conditional expectation of $\mathbf{a}_*$ given the data, $\boldsymbol{\Omega} = \text{cov}[\mathbf{a}_*]$ denotes the covariance matrix of $\mathbf{a}_*$, and $\mathbf{D}^{-1} = \text{cov}[\mathbf{a}_* - \hat{\mathbf{a}}_*]$. The $\hat{\mathbf{a}}_t$ satisfy the recursion

$$\hat{\boldsymbol{a}}_t = \boldsymbol{Z}_t - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j} + \sum_{j=1}^{q} \boldsymbol{\Theta}_j \hat{\boldsymbol{a}}_{t-j} \quad t = 1, \ldots, N \tag{14.4.8}$$

where the presample values are the estimated values $\hat{\boldsymbol{Z}}_t, t = 1-p, \ldots, 0$, and $\hat{\boldsymbol{a}}_t, t = 1-q, \ldots, 0$. Details of the calculations are given in the papers referenced above. Explicit expressions for the quantities $\boldsymbol{\Omega}$, $\boldsymbol{D}$, and $\hat{\mathbf{a}}_*$ are also provided by Reinsel (1997, Section 5.3.1).

Other approaches to likelihood evaluation emphasize the innovations form of the exact likelihood and the use of the state-space model representation of the VARMA model and the associated Kalman filtering methods; see, for example, Ansley and Kohn (1983), Solo (1984a), and Shea (1987). The innovations form of the exact likelihood is

$$L = \left(\prod_{t=1}^{N} |\boldsymbol{\Sigma}_{t|t-1}|^{-1/2}\right) \exp\left\{ -\left(\frac{1}{2}\right) \sum_{t=1}^{N} \boldsymbol{a}_{t|t-1}'\boldsymbol{\Sigma}_{t|t-1}^{-1}\boldsymbol{a}_{t|t-1}\right\} \tag{14.4.9}$$

where $\boldsymbol{a}_{t|t-1} = \boldsymbol{Z}_t - \hat{\boldsymbol{Z}}_{t|t-1}$ is the one-step prediction error, or innovation,

$$\hat{\boldsymbol{Z}}_{t|t-1} = E[\boldsymbol{Z}_t | \boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_1]$$

denotes the linear predictor of $\boldsymbol{Z}_t$ based on $\boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_1$, and $\boldsymbol{\Sigma}_{t|t-1} = \text{cov}[\boldsymbol{a}_{t|t-1}]$ is the one-step prediction error covariance matrix. The $\boldsymbol{a}_{t|t-1}$ and $\boldsymbol{\Sigma}_{t|t-1}$, for $t = 1, \ldots, N$, can be computed recursively using the innovations algorithm described by Brockwell and Davis (1991) and Reinsel (1997). Equivalently, the quantities $\boldsymbol{a}_{t|t-1} = \boldsymbol{Z}_t - \hat{\boldsymbol{Z}}_{t|t-1}$ and $\boldsymbol{\Sigma}_{t|t-1}$ are also obtained naturally as outputs from the Kalman filtering algorithm applied to the state-space representation of the VARMA model, which is discussed in more detail in Section 14.6. Asymptotic theory of the resulting maximum likelihood estimators for VARMA models has been studied by Dunsmuir and Hannan (1976), Deistler et al. (1978), and Hannan and Deistler (1988).

Diagnostic Checking. The checking of the fitted model can be performed using the tools described for VAR models in Section 14.2.6. These include plots of the residuals against time and/or against other variables and detailed examination of the autocorrelation and cross-correlation functions of the residuals. These tools can provide valuable information about possible lack of fit and suggest directions for model improvement. Useful supplementary tools include the portmanteau test and similar statistical tests. These tools also extend to fitted models with constraints imposed on the parameter coefficient matrices (i.e., structured parameterizations), such as echelon canonical form and reduced-rank models discussed in more detail in Section 14.7. For example, the statistic Q_s will then have $k^2 s - b$ degrees of freedom in its limiting chi-squared distribution, where b denotes the number of unconstrained parameters involved in the estimation of the ARMA model coefficients $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Theta}_j$.

14.4.6 Relation of VARMA Models to Transfer Function and ARMAX Models

The relationship between a bivariate VAR(1) model and a transfer function model was mentioned in Section 14.2.1. We will now briefly examine the relationship between subcomponents in a more general VARMA(p, q) process. We begin by partitioning the k-dimensional vector process $\boldsymbol{Z}_t$ into two groups of subcomponents of dimensions k_1 and k_2, respectively, as $\boldsymbol{Z}_t = (\boldsymbol{Z}'_{1t}, \boldsymbol{Z}'_{2t})'$. The innovations vector $\boldsymbol{a}_t$ and the AR and MA matrix polynomials are partitioned accordingly as $\boldsymbol{a}_t = (\boldsymbol{a}'_{1t}, \boldsymbol{a}'_{2t})'$ and

$$\boldsymbol{\Phi}(B) = \begin{bmatrix} \boldsymbol{\Phi}_{11}(B) & \boldsymbol{\Phi}_{12}(B) \\ \boldsymbol{\Phi}_{21}(B) & \boldsymbol{\Phi}_{22}(B) \end{bmatrix} \qquad \boldsymbol{\Theta}(B) = \begin{bmatrix} \boldsymbol{\Theta}_{11}(B) & \boldsymbol{\Theta}_{12}(B) \\ \boldsymbol{\Theta}_{21}(B) & \boldsymbol{\Theta}_{22}(B) \end{bmatrix}$$

Suppose now that $\boldsymbol{\Phi}_{12}(B)$ and $\boldsymbol{\Theta}_{12}(B)$ are both identically zero, and for convenience also assume that $\boldsymbol{\Theta}_{21}(B) = 0$. The equations for the VARMA model can then be expressed in two distinct groups as

$$\boldsymbol{\Phi}_{11}(B)\boldsymbol{Z}_{1t} = \boldsymbol{\Theta}_{11}(B)\boldsymbol{a}_{1t} \tag{14.4.10a}$$

and

$$\boldsymbol{\Phi}_{22}(B)\boldsymbol{Z}_{2t} = -\boldsymbol{\Phi}_{21}(B)\boldsymbol{Z}_{1t} + \boldsymbol{\Theta}_{22}(B)\boldsymbol{a}_{2t} \tag{14.4.10b}$$

We see from these expressions that future values of the process $\boldsymbol{Z}_{1t}$ are only influenced by its own past and not by the past of $\boldsymbol{Z}_{2t}$, whereas future values of $\boldsymbol{Z}_{2t}$ are influenced by the past of both $\boldsymbol{Z}_{1t}$ and $\boldsymbol{Z}_{2t}$. Notice that even if $\boldsymbol{\Theta}_{21}(B) \neq 0$, this conclusion still holds since the additional term in (14.4.10b) would then be $\boldsymbol{\Theta}_{21}(B)\boldsymbol{a}_{1t} = \boldsymbol{\Theta}_{21}(B)\boldsymbol{\Theta}_{11}^{-1}(B)\boldsymbol{\Phi}_{11}(B)\boldsymbol{Z}_{1t}$.

In the terminology of causality from econometrics, under (14.4.10a) and (14.4.10b), the variables $\boldsymbol{Z}_{1t}$ are said to cause $\boldsymbol{Z}_{2t}$, but $\boldsymbol{Z}_{2t}$ do not cause $\boldsymbol{Z}_{1t}$. The variables $\boldsymbol{Z}_{1t}$ are referred to as *exogenous variables*, and (14.4.10b) is often referred to as an ARMAX model or ARMAX system for the output variables $\boldsymbol{Z}_{2t}$ with $\boldsymbol{Z}_{1t}$ serving as input variables. The **X** in ARMAX stands for exogenous. The model (14.4.10b) can be rewritten as

$$\boldsymbol{Z}_{2t} = \boldsymbol{\Psi}_*(B)\boldsymbol{Z}_{1t} + \boldsymbol{\Psi}_{22}(B)\mathbf{a}_{2t}$$

where

$$\boldsymbol{\Psi}_*(B) = -\boldsymbol{\Phi}_{22}^{-1}(B)\boldsymbol{\Phi}_{21}(B) \quad \text{and} \quad \boldsymbol{\Psi}_{22}(B) = \boldsymbol{\Phi}_{22}^{-1}(B)\boldsymbol{\Theta}_{22}(B)$$

This equation provides a representation for the output process $\boldsymbol{Z}_{2t}$ as a causal linear filter of the input process $\boldsymbol{Z}_{1t}$ with added unobservable noise, that is,

$$\boldsymbol{Z}_{2t} = \boldsymbol{\Psi}_*(B)\boldsymbol{Z}_{1t} + \boldsymbol{N}_t \tag{14.4.11}$$

where the noise process $\boldsymbol{N}_t$ follows a VARMA model $\boldsymbol{\Phi}_{22}(B)\mathbf{N}_t = \boldsymbol{\Theta}_{22}(B)\mathbf{a}_{2t}$. Since the ARMAX model can be viewed as a special case of the VARMA model, the methods for model building are quite similar to those used for the VARMA model. These include the use of model selection criteria and least-squares estimation methods for model specification and examination of the residuals from the fitted model for model checking. For further discussion, see, for example, Hannan and Deistler (1988, Chapter 4) and Reinsel (1997, Chapter 8).

In the special case of bivariate time series, $\boldsymbol{Z}_{1t} \equiv z_{1t}$ and $\boldsymbol{Z}_{2t} \equiv z_{2t}$ are each univariate time series. Then we see from the above that when $\Phi_{12}(B) = 0$ and $\Theta_{12}(B) = 0$, the model reduces to the structure of the ''unidirectional'' instantaneous transfer function model with z_{1t} as the ''input'' process and z_{2t} as the output, assuming independence between z_{2t} and the noise term of z_{1t}. More generally, assuming independence between $\boldsymbol{Z}_{1t}$ and $\boldsymbol{N}_t$ above, (14.4.11) can be viewed as a multivariate generalization of the univariate (single-equation) transfer function model discussed in Chapters 11 and 12.

14.5 FORECASTING FOR VECTOR AUTOREGRESSIVE–MOVING AVERAGE PROCESSES

14.5.1 Calculation of Forecasts from ARMA Difference Equation

For forecasting in the VARMA(p, q) model

$$\boldsymbol{Z}_t = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j} + \boldsymbol{\delta} + \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j} \tag{14.5.1}$$

where $\boldsymbol{\delta} = (\mathbf{I} - \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_p)\boldsymbol{\mu}$ for stationary processes, we assume that the white noise series $\boldsymbol{a}_t$ are mutually independent random vectors. From general principles of prediction, the predictor of a future value $\boldsymbol{Z}_{t+l}, l = 1, 2, \ldots$, based on observations available at time

$t, \{\boldsymbol{Z}_s, s \le t\}$, that yields the minimum mean squared error (MSE) matrix is given by $\hat{\boldsymbol{Z}}_t(l) = E[\boldsymbol{Z}_{t+l} | \boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots]$. So from a computational view, forecasts are determined by applying conditional expectations to both sides of the VARMA(p, q) relation

$$\boldsymbol{\Phi}(B)\boldsymbol{Z}_{t+l} = \boldsymbol{\delta} + \boldsymbol{\Theta}(B)\boldsymbol{a}_{t+l}$$

using the result that $E[\boldsymbol{a}_{t+h} | \boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots] = \boldsymbol{0}, h > 0$, since $\boldsymbol{a}_{t+h}$ is independent of present and past values of the series. Thus, forecasts $\hat{\boldsymbol{Z}}_t(l)$ can be computed recursively from the VARMA model difference equation as

$$\hat{\boldsymbol{Z}}_t(l) = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \hat{\boldsymbol{Z}}_t(l-j) + \boldsymbol{\delta} - \sum_{j=l}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t+l-j} \quad l = 1, 2, \ldots, q \tag{14.5.2}$$

with $\hat{\boldsymbol{Z}}_t(l) = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \hat{\boldsymbol{Z}}_t(l-j) + \boldsymbol{\delta}$, for $l > q$, where $\hat{\boldsymbol{Z}}_t(l-j) = \boldsymbol{Z}_{t+l-j}$ for $l \le j$. Note that for pure VAR models with $q = 0$

$$\hat{\boldsymbol{Z}}_t(l) = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \hat{\boldsymbol{Z}}_t(l-j) + \boldsymbol{\delta}, \quad \text{for all } l = 1, 2, \ldots$$

So the p initial forecast values are completely determined by the last p observations $\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_{t-p+1}$; hence, for AR models all forecasts depend only on these last p observations in the series.

For models that involve an MA term, in practice it is necessary to generate the white noise sequence $\boldsymbol{a}_t$ recursively from the past data $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_t$, as

$$\boldsymbol{a}_s = \boldsymbol{Z}_s - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{Z}_{s-j} - \boldsymbol{\delta} + \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{s-j} \quad s = 1, 2, \ldots, t$$

using appropriate starting values for $\boldsymbol{a}_0, \ldots, \boldsymbol{a}_{1-q}$ and $\boldsymbol{Z}_0, \ldots, \boldsymbol{Z}_{1-p}$. One way to estimate the starting values is to use the *backcasting* technique described earlier for evaluation of the exact likelihood function for ARMA models. This method yields $\hat{\boldsymbol{a}}_{1-j} = E[\boldsymbol{a}_{1-j} | \boldsymbol{Z}_t, \ldots, \boldsymbol{Z}_1]$, $j = 1, \ldots, q$, and $\hat{\boldsymbol{Z}}_{1-j} = E[\boldsymbol{Z}_{1-j} | \boldsymbol{Z}_t, \ldots, \boldsymbol{Z}_1], j = 1, \ldots, p$. The resulting forecasts $\hat{\boldsymbol{Z}}_t(l)$ are then equal to

$$\hat{\boldsymbol{Z}}_t(l) \equiv E[\boldsymbol{Z}_{t+l} | \boldsymbol{Z}_t, \ldots, \boldsymbol{Z}_1]$$

These are optimal forecasts based on the finite past history $\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_1$, although the analysis of forecast properties given below assumes that the forecasts are based on the infinite past history $\boldsymbol{Z}_s$, all $s \le t$. However, these two forecasts will be nearly identical for any moderate or large value of t, the number of past values available for forecasting. Alternative methods to obtain the ''exact'' finite sample forecasts, as well as the exact covariance matrices of the forecast errors, based on the finite sample data $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t$, in a convenient computational manner are through an innovations approach or through the closely related state-space model and Kalman filter approach that will be discussed briefly in Section 14.6.

14.5.2 Forecasts from Infinite VMA Form and Properties of Forecast Errors

To establish the theoretical MSE properties of the forecast errors, we use the ''infinite'' moving average representation $\boldsymbol{Z}_t = \boldsymbol{\Psi}(B)\boldsymbol{a}_t$ of the VARMA(p, q) model, where $\boldsymbol{\Psi}(B) = \boldsymbol{\Phi}^{-1}(B)\boldsymbol{\Theta}(B) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j B^j$. A future value $\boldsymbol{Z}_{t+l}$, relative to the forecast origin t, can then be expressed as

$$\boldsymbol{Z}_{t+l} = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t+l-j} = \boldsymbol{a}_{t+l} + \boldsymbol{\Psi}_1 \boldsymbol{a}_{t+l-1} + \cdots + \boldsymbol{\Psi}_{l-1} \boldsymbol{a}_{t+1} + \boldsymbol{\Psi}_l \boldsymbol{a}_t + \cdots$$

Thus, since $E[\boldsymbol{a}_{t+h} | \boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots] = 0$, $h > 0$, the minimum MSE matrix predictor of $\boldsymbol{Z}_{t+l}$ based on $\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots$ can be represented as

$$\hat{\boldsymbol{Z}}_t(l) = E[\boldsymbol{Z}_{t+l} | \boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots] = \sum_{j=l}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t+l-j} \tag{14.5.3}$$

The l-step-ahead forecast error is $\boldsymbol{e}_t(l) = \boldsymbol{Z}_{t+l} - \hat{\boldsymbol{Z}}_t(l) = \sum_{j=0}^{l-1} \boldsymbol{\Psi}_j \boldsymbol{a}_{t+l-j}$ has zero mean and covariance matrix:

$$\boldsymbol{\Sigma}(l) = \text{cov}[\boldsymbol{e}_t(l)] = E[\boldsymbol{e}_t(l)\boldsymbol{e}_t(l)'] = \sum_{j=0}^{l-1} \boldsymbol{\Psi}_j \boldsymbol{\Sigma} \boldsymbol{\Psi}_j' \quad \boldsymbol{\Psi}_0 = \mathbf{I} \tag{14.5.4}$$

In particular, for one step ahead, $\boldsymbol{e}_t(1) = \boldsymbol{Z}_{t+1} - \hat{\boldsymbol{Z}}_t(1) = \boldsymbol{a}_{t+1}$ with error covariance matrix $\boldsymbol{\Sigma}$, so that the white noise series $\boldsymbol{a}_t$ can be interpreted as a sequence of one-step-ahead forecast errors for the process.

It follows from the infinite MA representation of the forecasts given by (14.5.3) that we obtain the multivariate version of the updating formula (5.2.5) as

$$\hat{\boldsymbol{Z}}_{t+1}(l) = E[\boldsymbol{Z}_{t+l+1} | \boldsymbol{Z}_{t+1}, \boldsymbol{Z}_t, \ldots] = \sum_{j=l}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t+l+1-j} = \hat{\boldsymbol{Z}}_t(l+1) + \boldsymbol{\Psi}_l \boldsymbol{a}_{t+1} \tag{14.5.5}$$

where $\boldsymbol{a}_{t+1} = \boldsymbol{Z}_{t+1} - \hat{\boldsymbol{Z}}_t(1)$ is the one-step-ahead forecast error. This provides a simple relationship to indicate how the forecast $\hat{\boldsymbol{Z}}_t(l)$ with forecast origin t is adjusted or updated to incorporate the information available from a new observation $\boldsymbol{Z}_{t+1}$ at time $t + 1$.

For the case of unit-root nonstationary processes to be discussed in Section 14.8, similar forecasting topics as presented above can also be developed and results such as (14.5.2) and (14.5.4) continue to apply.

14.6 STATE-SPACE FORM OF THE VARMA MODEL

The state-space model was introduced for univariate ARMA models in Section 5.5. Similar to the univariate case, the VARMA model can be represented in the equivalent state-space form, which is of interest for purposes of prediction as well as for model specification and maximum likelihood estimation of parameters. The state-space model consists of a transition or state equation

$$\boldsymbol{Y}_t = \boldsymbol{\Phi} \boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t$$

and an observation equation

$$\boldsymbol{Z}_t = \mathbf{H}\boldsymbol{Y}_t + \boldsymbol{N}_t$$

where $\boldsymbol{Y}_t$ is an $r \times 1$ (unobservable) time series vector called the state vector, and $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{N}_t$ are independent white noise processes. In this representation, the state vector $\boldsymbol{Y}_t$ conceptually contains all information from the past of the process $\boldsymbol{Z}_t$, which is relevant for the future of the process, and, hence, the dynamics of the system can be represented in the simple first-order or Markovian transition equation for the state vector. The above state-space model is said to be stable if all the eigenvalues of the matrix $\boldsymbol{\Phi}$ are less than one in absolute value, and conversely, it can be shown that any stationary process $\boldsymbol{Z}_t$ that has a stable state-space representation of the above form can also be represented in the form of a stationary VARMA(p, q) model; see, for example, Akaike (1974b). Hence, it follows that any process $\boldsymbol{Z}_t$ that satisfies a stable state-space representation can be expressed in the causal convergent infinite moving average form $\boldsymbol{Z}_t = \boldsymbol{\Psi}(B)\boldsymbol{a}_t$. The stability condition for the matrix $\boldsymbol{\Phi}$ in the state-space model is equivalent to the stability condition for the matrix coefficients $\boldsymbol{\Psi}_j$ of the linear filter $\boldsymbol{\Psi}(B)$ (see Appendix A14.1.2), since it ensures that $\sum_{j=0}^{\infty} ||\boldsymbol{\Psi}_j|| < \infty$ in the representation $\boldsymbol{Z}_t = \boldsymbol{\Psi}(B)\boldsymbol{a}_t$.

For the VARMA(p, q) model (14.5.1) (with $\boldsymbol{\delta} = \mathbf{0}$), define the predictors $\hat{\boldsymbol{Z}}_t(j) = E[\boldsymbol{Z}_{t+j}|\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots]$ as in Section 14.5.1 for $j = 0, 1, \ldots, r-1$, with $r = \max(p, q+1)$, and $\hat{\boldsymbol{Z}}_t(0) = \boldsymbol{Z}_t$. From the updating equations (14.5.5), we have $\hat{\boldsymbol{Z}}_t(j-1) = \hat{\boldsymbol{Z}}_{t-1}(j) + \boldsymbol{\Psi}_{j-1}\boldsymbol{a}_t$, $j = 1, 2, \ldots, r-1$. Also, for $j = r > q$ we find using (14.5.2) that

$$\hat{\boldsymbol{Z}}_t(j-1) = \hat{\boldsymbol{Z}}_{t-1}(j) + \boldsymbol{\Psi}_{j-1}\boldsymbol{a}_t = \sum_{i=1}^{p} \boldsymbol{\Phi}_i \hat{\boldsymbol{Z}}_{t-1}(j-i) + \boldsymbol{\Psi}_{j-1}\boldsymbol{a}_t$$

Let us define the ''state'' vector at time t, with r vector components, as $\boldsymbol{Y}_t = [\hat{\boldsymbol{Z}}_t(0)', \hat{\boldsymbol{Z}}_t(1)', \ldots, \hat{\boldsymbol{Z}}_t(r-1)']'$. Then, from the relations above, the state vector $\boldsymbol{Y}_t$ satisfies the state-space (transition) equations

$$\boldsymbol{Y}_t = \begin{bmatrix} \mathbf{0} & \mathbf{I} & \mathbf{0} \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdot \cdots & \mathbf{I} \\ \boldsymbol{\Phi}_r & \boldsymbol{\Phi}_{r-1} & \cdot \cdots & \boldsymbol{\Phi}_1 \end{bmatrix} \boldsymbol{Y}_{t-1} + \begin{bmatrix} \mathbf{I} \\ \boldsymbol{\Psi}_1 \\ \vdots \\ \boldsymbol{\Psi}_{r-2} \\ \boldsymbol{\Psi}_{r-1} \end{bmatrix} \boldsymbol{a}_t \tag{14.6.1}$$

where $\boldsymbol{\Phi}_i = \mathbf{0}$ if $i > p$. Thus, we have

$$\boldsymbol{Y}_t = \boldsymbol{\Phi}\boldsymbol{Y}_{t-1} + \boldsymbol{\Psi}\boldsymbol{a}_t \tag{14.6.2}$$

together with the observation equation

$$\boldsymbol{Z}_t^* = \boldsymbol{Z}_t + \boldsymbol{N}_t = [\mathbf{I}, \mathbf{0}, \ldots, \mathbf{0}]\boldsymbol{Y}_t + \boldsymbol{N}_t = \mathbf{H}\boldsymbol{Y}_t + \boldsymbol{N}_t \tag{14.6.3}$$

where the vector noise $\boldsymbol{N}_t$ would be present only if the process $\boldsymbol{Z}_t$ is observed subject to additional white noise error; otherwise, we simply have $\boldsymbol{Z}_t = \boldsymbol{Z}_t^* = \mathbf{H}\boldsymbol{Y}_t$. For convenience, we assume in the remainder of this section that the additional white noise is not present.

The state or transition equation (14.6.2) and the observation equation (14.6.3) constitute a state-space representation of the VARMA model. There are many other constructions

of the state vector $\boldsymbol{Y}_t$ that will give rise to state-space equations of the general form (14.6.2) and (14.6.3); that is, the state-space form of the VARMA model is not unique. Specifically, if we transform the state vector $\boldsymbol{Y}_t$ into $\overline{\boldsymbol{Y}}_t = \mathbf{P}\boldsymbol{Y}_t$, where $\mathbf{P}$ is an arbitrary nonsingular matrix, then models (14.6.2) and (14.6.3) can be written in a similar form in terms of $\overline{\boldsymbol{Y}}_t$ with $\bar{\boldsymbol{\Phi}} = \mathbf{P}\boldsymbol{\Phi}\mathbf{P}^{-1}, \overline{\mathbf{H}} = \mathbf{H}\mathbf{P}^{-1}$, and $\overline{\boldsymbol{\Psi}} = \mathbf{P}\boldsymbol{\Psi}$. The particular form given above has the state vector $\boldsymbol{Y}_t$, which can be viewed as generating the space of predictions of all future values of the process $\boldsymbol{Z}_t$, since $\hat{\boldsymbol{Z}}_t(l) = \sum_{i=1}^{r} \boldsymbol{\Phi}_i \hat{\boldsymbol{Z}}_t(l-i)$ for $l > r-1$.

In the state-space model, the unobservable state vector $\boldsymbol{Y}_t$ constitutes a summary of the state of the dynamic system through time t, and the state equation (14.6.2) describes the evolution of the dynamic system in time. The minimal dimension of the state vector $\boldsymbol{Y}_t$ in a state-space representation needs to be sufficiently large so that the dynamics of the system can be represented by the simple Markovian first-order structure. State-space representations for the VARMA model can exist with a state vector of minimal dimension smaller than the dimension in (14.6.1). This minimal dimension is the dimension of the set of basis predictors that generate the linear space of predictors of all future values; it is of smaller dimension than in (14.6.1) whenever the state vector $\boldsymbol{Y}_t$ can be represented linearly in terms of a smaller number of basis elements. Specifically, suppose that $\boldsymbol{Y}_t$ in (14.6.1) can be expressed as $\boldsymbol{Y}_t = \mathbf{A}\boldsymbol{Y}_t^*$, where $\boldsymbol{Y}_t^*$ is an $M \times 1$ vector whose elements form a subset of the elements of $\boldsymbol{Y}_t$, with $M < rk$ being the smallest possible such dimension. Then $\mathbf{A}$ is a $rk \times M$ matrix of full rank M, with $\boldsymbol{Y}_t^* = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\boldsymbol{Y}_t$, and we assume the first $k \times M$ block row of $\mathbf{A}$ is $[\mathbf{I}, \mathbf{0}, \ldots, \mathbf{0}]$. Thus, multiplying (14.6.2) on the left by $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$, we obtain the equivalent representation of minimal dimension M given by $\boldsymbol{Y}_t^* = \boldsymbol{\Phi}^*\boldsymbol{Y}_{t-1}^* + \boldsymbol{\Psi}^*\boldsymbol{a}_t$, where $\boldsymbol{\Phi}^* = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\boldsymbol{\Phi}\mathbf{A}$ and $\boldsymbol{\Psi}^* = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\boldsymbol{\Psi}$, with $\boldsymbol{Z}_t = \mathbf{H}\mathbf{A}\boldsymbol{Y}_t^* \equiv \mathbf{H}\boldsymbol{Y}_t^*$. This minimal dimension M is in fact the McMillan degree of the process $\{\boldsymbol{Z}_t\}$ as described in Section 14.7.1 below.

One important use of the state-space form of the VARMA model is that it enables exact finite sample forecasts of the process $\{\boldsymbol{Z}_t\}$ to be obtained through Kalman filtering and the associated prediction algorithm. This provides a convenient computational procedure to obtain the minimum MSE matrix estimate of the state vector $\boldsymbol{Y}_{t+l}$ based on observations $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t$ as $\hat{\boldsymbol{Y}}_{t+l|t} = E[\boldsymbol{Y}_{t+l} | \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t]$, with

$$\mathbf{P}_{t+l|t} = E[(\boldsymbol{Y}_{t+l} - \hat{\boldsymbol{Y}}_{t+l|t})(\boldsymbol{Y}_{t+l} - \hat{\boldsymbol{Y}}_{t+l|t})']$$

equal to the error covariance matrix. The recursions for the Kalman filter procedure have been presented as equations (5.5.6) to (5.5.9) in Section 5.5.2. It follows that optimal forecasts $\hat{\boldsymbol{Z}}_{t+l|t} = E[\boldsymbol{Z}_{t+l} | \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t]$ of future observations $\boldsymbol{Z}_{t+l}$ are then available as $\hat{\boldsymbol{Z}}_{t+l|t} = \mathbf{H}\hat{\boldsymbol{Y}}_{t+l|t}$, since $\boldsymbol{Z}_{t+l} = \mathbf{H}\boldsymbol{Y}_{t+l}$, with forecast error covariance matrix

$$\boldsymbol{\Sigma}_{t+l|t} = E[(\boldsymbol{Z}_{t+l} - \hat{\boldsymbol{Z}}_{t+l|t})(\boldsymbol{Z}_{t+l} - \hat{\boldsymbol{Z}}_{t+l|t})'] = \mathbf{H}\mathbf{P}_{t+l|t}\mathbf{H}'$$

The ''steady-state'' values of the Kalman filtering lead l forecast error covariance matrices, obtained as t increases, equal the expressions in (14.5.4) of Section 14.5.2, $\boldsymbol{\Sigma}(l) = \sum_{j=0}^{l-1} \boldsymbol{\Psi}_j \boldsymbol{\Sigma} \boldsymbol{\Psi}_j'$. That is, $\boldsymbol{\Sigma}_{t+l|t}$ approaches $\boldsymbol{\Sigma}(l)$ as $t \to \infty$.

Thus, the Kalman filtering procedure provides a convenient method to obtain exact finite sample forecasts for future values in the VARMA process, based on observations $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_t$, subject to specification of appropriate initial conditions to use in (5.5.6) to (5.5.9). In particular, for the VARMA process represented in state-space form, the

exact finite-sample one-step-ahead forecasts $\hat{\boldsymbol{Z}}_{t|t-1} = \mathbf{H}\hat{\boldsymbol{Y}}_{t|t-1}$, and their error covariance matrices $\boldsymbol{\Sigma}_{t|t-1} = \mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}'$, can be obtained conveniently through the Kalman filtering equations. This can be particularly useful for evaluation of the exact Gaussian likelihood function, based on N vector observations $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ from the VARMA process, as mentioned earlier in Section 14.4.5.

14.7 FURTHER DISCUSSION OF VARMA MODEL SPECIFICATION

In this section, we return to the issue of model specification for a vector ARMA process. As noted in Section 14.4.4, extending the ARMA model to the vector case involves some difficulties that are not present in the univariate case. One problem in the vector case is the overflow of parameters, whose estimates can be highly correlated. A second issue is that of identifiability, which refers to the fact that two different sets of parameters can give rise to the same probability structure and hence the same process. This causes problems at the parameter estimation stage, in particular, since the likelihood function will not have a uniquely defined maximum in this case. Two methods designed to overcome these issues are the Kronecker index approach that originates in the engineering literature and the SCM method developed by Tiao and Tsay (1989). Both methods make use of canonical correlation analysis methods to arrive at a parsimonious and well-defined VARMA model.

In this section, we will discuss the VARMA model specification in more detail focusing on the Kronecker index approach to model specification. We first discuss the estimation of the Kronecker indices and the McMillan degree of a vector process. We then describe the specification of the echelon canonical form of the VARMA model through the Kronecker indices. A brief discussion of the use of partial canonical correlation analysis to identify models with reduced rank structure is also included.

14.7.1 Kronecker Structure for VARMA Models

The VARMA(p, q) model (14.4.1) can always be expressed in the equivalent form

$$\boldsymbol{\Phi}_0^{\#}(\boldsymbol{Z}_t - \boldsymbol{\mu}) - \sum_{j=1}^{p} \boldsymbol{\Phi}_j^{\#}(\boldsymbol{Z}_{t-j} - \boldsymbol{\mu}) = \boldsymbol{\Theta}_0^{\#}\boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j^{\#}\boldsymbol{a}_{t-j} \tag{14.7.1}$$

where $\boldsymbol{\Phi}_0^{\#}$ is an arbitrary nonsingular matrix, $\boldsymbol{\Phi}_j^{\#} = \boldsymbol{\Phi}_0^{\#}\boldsymbol{\Phi}_j$, $\boldsymbol{\Theta}_0^{\#} = \boldsymbol{\Phi}_0^{\#}$, and $\boldsymbol{\Theta}_j^{\#} = \boldsymbol{\Phi}_0^{\#}\boldsymbol{\Theta}_j$. For purposes of parsimony, we are interested in model forms that lead to the simplest structure in some sense, such as in terms of the number of unknown parameters in the matrices $\boldsymbol{\Phi}_0^{\#}, \boldsymbol{\Phi}_1^{\#}, \ldots, \boldsymbol{\Phi}_p^{\#}, \boldsymbol{\Theta}_1^{\#}, \ldots, \boldsymbol{\Theta}_q^{\#}$. For unique identifiability of the parameters, it is necessary to normalize the form of $\boldsymbol{\Phi}_0^{\#}$ at least to be lower triangular with ones on the diagonal.

As discussed in detail by Hannan and Deistler (1988, Chapter 2), a representation of a VARMA model in a certain special form of (14.7.1) can sometimes be more useful for model specification than the standard or reduced VARMA form (14.4.1), and this form of (14.7.1) is referred to as the *echelon canonical form* of the VARMA model. To specify the echelon canonical form, k *Kronecker indices* or structural indices, $K_1, \ldots, K_k$, must be determined beyond the overall orders p and q. The echelon (canonical) form is such that $[\boldsymbol{\Phi}^{\#}(B), \boldsymbol{\Theta}^{\#}(B)]$ has the smallest possible row degrees, and K_i denotes the degree of the ith row of $[\boldsymbol{\Phi}^{\#}(B), \boldsymbol{\Theta}^{\#}(B)]$, that is, the maximum of the degrees of the polynomials

in the ith row of $[\boldsymbol{\Phi}^{\#}(B),\ \boldsymbol{\Theta}^{\#}(B)]$, for $i = 1, \ldots, k$, and with $p = q = \max\{K_1, \ldots, K_k\}$. The specification of these Kronecker indices or ''row orders'' $\{K_i\}$, which are unique for any given equivalence class of ARMA models, that is, models with the same infinite MA operator $\boldsymbol{\Psi}(B)$, then determines a unique echelon canonical form of the VARMA model (14.7.1) in which the unknown parameters are uniquely identifiable.

Kronecker Indices and McMillan Degree of VARMA Process. For any stationary vector process $\{\boldsymbol{Z}_t\}$ with covariance matrices $\boldsymbol{\Gamma}(l) = \text{cov}[\boldsymbol{Z}_t, \boldsymbol{Z}_{t+l}]$, we define the infinite-dimensional (block) Hankel matrix of the covariances as

$$\mathbf{H} = \begin{bmatrix} \boldsymbol{\Gamma}(1)' & \boldsymbol{\Gamma}(2)' & \boldsymbol{\Gamma}(3)' & \cdots \\ \boldsymbol{\Gamma}(2)' & \boldsymbol{\Gamma}(3)' & \boldsymbol{\Gamma}(4)' & \cdots \\ \boldsymbol{\Gamma}(3)' & \boldsymbol{\Gamma}(4)' & \boldsymbol{\Gamma}(5)' & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \tag{14.7.2}$$

Then, in particular, the *McMillan degree* M of the process is defined as the rank of the Hankel matrix $\mathbf{H}$. The process $\{\boldsymbol{Z}_t\}$ follows a finite-order VARMA model if and only if the rank of $\mathbf{H}$ is finite. For a stationary VARMA(p, q) process, the moment relations (14.4.5) yield that

$$\boldsymbol{\Gamma}(l)' - \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{\Gamma}(l-j)' = \mathbf{0} \qquad \text{for } l > q \tag{14.7.3}$$

It can be seen directly from this that the rank of $\mathbf{H}$, the McMillan degree M, will then satisfy $M \leq ks$, where $s = \max\{p, q\}$, since all the $k \times k$ block rows of $\mathbf{H}$ beyond the sth block row will be linearly dependent on the preceding block rows. But the McMillan degree M of a VARMA(p, q) could be considerably smaller than ks due to rank deficiencies in the AR and MA coefficient matrices.

The McMillan degree M has the interpretation as the number of linearly independent linear combinations of the present and past vectors $\boldsymbol{Z}_t, \boldsymbol{Z}_{t-1}, \ldots$ that are needed for optimal prediction of all future vectors within the ARMA structure. Note that

$$\mathbf{H} = \text{cov}[\mathbf{F}_{t+1}, \mathbf{P}_t] = \text{cov}[\mathbf{F}_{t+1|t}, \mathbf{P}_t] \tag{14.7.4}$$

is the covariance between the collection of all present and past vectors, $\mathbf{P}_t = (\boldsymbol{Z}_t', \boldsymbol{Z}_{t-1}', \ldots)'$, and the collection of all future vectors $\mathbf{F}_{t+1} = (\boldsymbol{Z}_{t+1}', \boldsymbol{Z}_{t+2}', \ldots)'$ or the collection of predicted values of all future vectors, $\mathbf{F}_{t+1|t} = E[\mathbf{F}_{t+1}|\mathbf{P}_t]$. Hence, if the rank of $\mathbf{H}$ is equal to M, then the (linear) predictor space formed from the collection $\mathbf{F}_{t+1|t}$ of predicted values $\hat{\boldsymbol{Z}}_t(l) = E[\boldsymbol{Z}_{t+l}|\mathbf{P}_t]$, $l > 0$, of all future vectors is of finite dimension M. Sometimes (e.g., Hannan and Deistler, 1988, Chapter 2) the Hankel matrix $\mathbf{H}$ is defined in terms of the coefficients $\boldsymbol{\Psi}_j$ in the infinite MA form $\boldsymbol{Z}_t - \boldsymbol{\mu} = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \mathbf{a}_{t-j}$ of the ARMA process, instead of the covariance matrices $\boldsymbol{\Gamma}(j)'$, but all main conclusions hold in either case.

In addition, the ith *Kronecker index* K_i, $i = 1, \ldots, k$, of the process $\{\boldsymbol{Z}_t\}$ is the smallest value such that the $(kK_i + i)$th row of $\mathbf{H}$, that is, the ith row in the $(K_i + 1)$th block of rows of $\mathbf{H}$, is linearly dependent on the previous rows of $\mathbf{H}$. This also implies, through the structure of the Hankel matrix $\mathbf{H}$, that all rows $kl + i$, for every $l \geq K_i$, will also be linearly dependent on the rows preceding the $(kK_i + i)$th row. The set of Kronecker indices $\{K_1, \ldots, K_k\}$ is unique for any given VARMA process; hence, it is not dependent

on any one particular form of the observationally equivalent ARMA model representations of the process. As indicated in Section 14.6, the VARMA model can be represented in its equivalent minimal dimension state-space form, with minimal dimension, the McMillan degree

$$M = \sum_{i=1}^{k} K_i = K_1 + K_2 + \cdots + K_k$$

being the number of linearly independent predictors required to generate the linear prediction space $\{\hat{\boldsymbol{Z}}_t(l),\ l \geq 1\}$ of all future vectors $\{\boldsymbol{Z}_{t+l}, l \geq 1\}$. This minimal dimension state-space representation is one way to reveal the special structure of the VARMA parameters associated with the Kronecker indices. Canonical correlation analysis methods between past and future vectors of a VARMA process $\{\boldsymbol{Z}_t\}$ are useful as a means to determine the Kronecker indices of the process. We will now indicate, in particular, the direct connections that the Kronecker indices have with the second moment equations as in (14.4.5) and (14.7.3), since these equations exhibit the row dependencies among the covariance matrices $\boldsymbol{\Gamma}(j)'$. Hence, knowledge of these Kronecker indices can be used to deduce special structure among the AR and MA parameter matrices and lead to specification of the special (echelon) form of the VARMA model.

Echelon Canonical Form Implied by Kronecker Indices. Specifically, if VARMA models similar to the form in (14.7.1) are considered, with $\boldsymbol{\Phi}_0^{\#} = \boldsymbol{\Theta}_0^{\#}$ lower triangular (and having ones on the diagonal), then equations similar to (14.4.5) for the cross-covariance matrices $\boldsymbol{\Gamma}(l)$ of the process are obtained as

$$\boldsymbol{\Phi}_0^{\#}\boldsymbol{\Gamma}(l)' - \sum_{j=1}^{p} \boldsymbol{\Phi}_j^{\#}\boldsymbol{\Gamma}(l-j)' = -\sum_{j=l}^{q} \boldsymbol{\Theta}_j^{\#}\boldsymbol{\Sigma}\boldsymbol{\Psi}_{j-l}' \tag{14.7.5}$$

Thus, if $\boldsymbol{\phi}_j(i)'$ denotes the ith row of $\boldsymbol{\Phi}_j^{\#}$, then the ith Kronecker index equal to K_i implies the linear dependence in the rows of the Hankel matrix $\mathbf{H}$ of the form

$$\boldsymbol{\phi}_0(i)'\boldsymbol{\Gamma}(l)' - \sum_{j=1}^{K_i} \boldsymbol{\phi}_j(i)'\boldsymbol{\Gamma}(l-j)' = \mathbf{0}' \qquad \text{for all } l \geq K_i + 1 \tag{14.7.6}$$

that is, $\boldsymbol{b}_i'\mathbf{H} = \mathbf{0}'$ with $\boldsymbol{b}_i' = (-\boldsymbol{\phi}_{K_i}(i)', \ldots, -\boldsymbol{\phi}_1(i)',\ \boldsymbol{\phi}_0(i)', \mathbf{0}', \ldots)$. Note that by definition of the ith Kronecker index K_i, the row vector $\boldsymbol{\phi}_0(i)'$ in (14.7.6) can be taken to have a one in the ith position and zeros for positions greater than the ith. Therefore, a Kronecker index equal to K_i implies, in particular, that an ARMA model representation of the form (14.7.1) can be constructed for the process such that the ith rows of the matrices $\boldsymbol{\Phi}_j^{\#}$ and $\boldsymbol{\Theta}_j^{\#}$ will be zero for $j > K_i$.

In addition to these implications from (14.7.6), additional zero constraints on certain elements in the ith rows of the matrices $\boldsymbol{\Phi}_j^{\#}$ for $j \leq K_i$ can be specified. Specifically, the lth element of the ith row $\boldsymbol{\phi}_j(i)'$ can be specified to be zero whenever $j + K_l \leq K_i$ because for $K_l \leq K_i$ the rows $k(K_l + j) + l,\ j = 0, \ldots, (K_i - K_l)$, of the Hankel matrix $\mathbf{H}$ are all

linearly dependent on the previous rows of $\mathbf{H}$. Hence, the (i, l)th element of the AR operator

$$\mathbf{\Phi}^{\#}(B) = \mathbf{\Phi}_0^{\#} - \sum_{j=1}^{p} \mathbf{\Phi}_j^{\#} B^j$$

in model (14.7.1) can be specified to have nonzero coefficients only for the lags $j = K_i - K_{il} + 1, \ldots, K_i$, with zero coefficients specified for any lower lags of j (when $i \neq l$), where we define

$$K_{il} = \begin{cases} \min(K_i + 1, K_l) & \text{for } i > l \\ \min(K_i, K_l) & \text{for } i \leq l \end{cases} \tag{14.7.7}$$

(so that whenever $K_l \leq K_i$ we have $K_{il} = K_l$). Thus, the corresponding number of unknown AR parameters in the (i, l)th element of $\mathbf{\Phi}^{\#}(B)$ is equal to K_{il}. Hence, the AR operator $\mathbf{\Phi}^{\#}(B)$ in model (14.7.1) can be specified such that the total number of unknown parameters of $\mathbf{\Phi}^{\#}(B)$ is equal to $\sum_{i=1}^{k} \sum_{l=1}^{k} K_{il} = M + \sum \sum_{i \neq l}^{k} K_{il}$, while the number of unknown parameters in the MA operator $\mathbf{\Theta}^{\#}(B)$, excluding those parameters in $\mathbf{\Theta}_0^{\#} = \mathbf{\Phi}_0^{\#}$, is equal to $\sum_{i=1}^{k} kK_i = kM$.

In summary, for a stationary linear process $\{\boldsymbol{Z}_t\}$ with Kronecker indices $K_1, \ldots, K_k$, a VARMA representation as in (14.7.1) with $p = q = \{\max K_i\}$ can be specified to describe the process, with the matrices $\mathbf{\Phi}_j^{\#}$ and $\mathbf{\Theta}_j^{\#}$ possessing the structure that their ith rows are zero for $j > K_i$ and the additional zero constraints structure noted above. Moreover, for a stationary vector process with given covariance matrix structure $\mathbf{\Gamma}(l)$, or equivalently with given infinite MA coefficients $\mathbf{\Psi}_j$, Hannan and Deistler (1988, Theorem 2.5.1) have shown that this model provides a unique VARMA representation, with AR and MA operators $\mathbf{\Phi}^{\#}(B)$ and $\mathbf{\Theta}^{\#}(B)$ being left-coprime, and where all unknown parameters are identified. This (canonical) ARMA representation is referred to as a (reversed) *echelon ARMA form*. In particular, the VAR coefficient matrices $\mathbf{\Phi}_j^{\#}$ in the echelon canonical representation (14.7.1) are uniquely determined from the $\mathbf{\Gamma}(l)$ by the requirement that their ith rows $\phi_j(i)'$, $j = 0, \ldots, K_i, i = 1, \ldots, k$, satisfy the conditions (14.7.6).

Examples. For simple illustrative examples, consider a bivariate ($k = 2$) process $\{\boldsymbol{Z}_t\}$. When this process has Kronecker indices $K_1 = K_2 = 1$, then a general VARMA(1, 1) representation $\boldsymbol{Z}_t - \mathbf{\Phi}_1 \boldsymbol{Z}_{t-1} = \boldsymbol{a}_t - \mathbf{\Theta}_1 \boldsymbol{a}_{t-1}$ is implied. However, notice that a pure VAR(1) process with full-rank VAR matrix $\mathbf{\Phi}_1$ and a pure VMA(1) process with full-rank VMA matrix $\mathbf{\Theta}_1$ would both also possess Kronecker indices equal to $K_1 = K_2 = 1$. This simple example thus illustrates that specification of the Kronecker indices alone does not necessarily lead to the specification of a VARMA representation where all the simplifying structure in the parameters is directly revealed. For a second case, suppose the bivariate process has Kronecker indices $K_1 = 1$ and $K_2 = 0$. Then, the implied structure for the process is VARMA(1, 1) as in (14.7.1), with (note, in particular, that $K_{12} = 0$ in (14.7.7))

$$\mathbf{\Phi}_0^{\#} = \begin{bmatrix} 1 & 0 \\ X & 1 \end{bmatrix} \quad \mathbf{\Phi}_1^{\#} = \begin{bmatrix} X & 0 \\ 0 & 0 \end{bmatrix} \quad \mathbf{\Theta}_1^{\#} = \begin{bmatrix} X & X \\ 0 & 0 \end{bmatrix}$$

where the X's denote unknown parameters that need estimation and 0's indicate values that are known to be specified as zero. On multiplication of the VARMA(1, 1) relation $\mathbf{\Phi}_0^{\#} \boldsymbol{Z}_t - \mathbf{\Phi}_1^{\#} \boldsymbol{Z}_{t-1} = \mathbf{\Theta}_0^{\#} \boldsymbol{a}_t - \mathbf{\Theta}_1^{\#} \boldsymbol{a}_{t-1}$ on the left by $\mathbf{\Phi}_0^{\#-1}$, we obtain a VARMA(1, 1) representation

$\boldsymbol{Z}_t - \boldsymbol{\Phi}_1 \boldsymbol{Z}_{t-1} = \boldsymbol{a}_t - \boldsymbol{\Theta}_1 \boldsymbol{a}_{t-1}$ in the standard VARMA form (14.4.1), but with a reduced-rank structure for the coefficient matrices such that rank $[\boldsymbol{\Phi}_1, \boldsymbol{\Theta}_1] = 1$. For a third situation, suppose the bivariate process has Kronecker indices $K_1 = 2$ and $K_2 = 1$. Then, the echelon form structure for the process is VARMA(2, 2) as in (14.7.1), with (note $K_{12} = 1$ in this case)

$$\boldsymbol{\Phi}_0^{\#} = \begin{bmatrix} 1 & 0 \\ X & 1 \end{bmatrix} \quad \boldsymbol{\Phi}_1^{\#} = \begin{bmatrix} X & 0 \\ X & X \end{bmatrix} \quad \boldsymbol{\Phi}_2^{\#} = \begin{bmatrix} X & X \\ 0 & 0 \end{bmatrix} \quad \boldsymbol{\Theta}_1^{\#} = \begin{bmatrix} X & X \\ X & X \end{bmatrix} \quad \boldsymbol{\Theta}_2^{\#} = \begin{bmatrix} X & X \\ 0 & 0 \end{bmatrix}$$

Again, on multiplication of the echelon form VARMA(2, 2) relation on the left by $\boldsymbol{\Phi}_0^{\#-1}$, we obtain a VARMA(2, 2) representation in standard form, but with reduced-rank structure for the coefficient matrices such that rank $[\boldsymbol{\Phi}_2, \boldsymbol{\Theta}_2] = 1$.

Software Implementation. In practical applications, the Kronecker index approach to model specification can be implemented using the commands Kronid, Kronfit, and refKronfit available in the MTS package of R. The specification of the Kronecker indices is performed using the command Kronid and is based on canonical correlation analysis. With the Kronecker indices specified, the VARMA parameters are estimated using the command Kronfit. Parameters with nonsignificant estimates can be removed using the command refKronfit. For further discussion and for demonstrations of the individual commands, see Tsay (2014).

14.7.2 An Empirical Example

To illustrate model specification approach described above, we return to the bivariate time series of U.S. fixed investment and change in business inventories analyzed earlier in this chapter. A bivariate VAR(2) model was fitted to the series in Section 14.2.7. As an alternative, we now consider the possibility of a mixed VARMA model for these data through determination of the echelon canonical ARMA model for the two series. The Kronecker indices $\{K_i\}$ for the process are determined using the canonical correlation method suggested by Akaike (1976) and Cooper and Wood (1982); see also Tsay (2014, Section 4.4). For the vector of present and past values, we use a maximum of three time-lagged vector variables and set $\mathbf{P}_t = (\boldsymbol{Z}_t', \boldsymbol{Z}_{t-1}', \boldsymbol{Z}_{t-2}')'$. Then, for various vectors $\mathbf{F}_{t+1}^*$ of future variables, the squared sample canonical correlations between $\mathbf{F}_{t+1}^*$ and $\mathbf{P}_t$ are determined as the eigenvalues of the matrix similar to the matrix in (6.2.6) of Section 6.2.4. The canonical correlation analysis calculations are performed sequentially by adding variables to $\mathbf{F}_{t+1}^*$ one at a time, starting with $\mathbf{F}_{t+1}^* = (z_{1,t+1})$, until $k = 2$ near zero sample canonical correlations between $\mathbf{P}_t$ and $\mathbf{F}_{t+1}^*$ are determined. At each step, a likelihood ratio test is used to determine the significance of the smallest squared canonical correlation.

The calculations can be performed using the MTS package in R. If zz denotes the two time series, the command for determining the Kronecker indices is Kronfit(zz, plag=3), where plag represents the number of elements in $\mathbf{P}_t$. The resulting squared sample canonical correlations between $\mathbf{P}_t$ and various future vectors $\mathbf{F}_{t+1}^*$ are presented in Table 14.3. From these results, we note that the first occurrence of a small squared sample canonical correlation value (0.044), indicative of a zero canonical correlation between the future and the present and past, is obtained when $\mathbf{F}_{t+1}^* = (z_{1,t+1}, z_{2,t+1}, z_{1,t+2})'$. This indicates that the Kronecker index $K_1 = 1$, since it implies that a linear combination involving $z_{1,t+2}$ in terms of the remaining variables in $\mathbf{F}_{t+1}^*$, that is, of the form $z_{1,t+2} - \boldsymbol{\phi}_1(1)' \boldsymbol{Z}_{t+1}$, is uncorrelated

TABLE 14.3 Specification of Kronecker Indices for First Differences of U.S. Fixed Investment Data and Changes in Business Inventories Data

Future Vector $\mathbf{F}^*_{t+1}$	Smallest Squared Canonical Correlation	LR Test	Degrees of Freedom	p-Value	Kronecker Index
$z_{1,t+1}$	0.371	44.02	6	0.000	
$z_{1,t+1}, z_{2,t+1}$	0.369	43.50	5	0.000	
$z_{1,t+1}, z_{2,t+1}, z_{1,t+2}$	0.044	4.13	4	0.389	$K_1 = 1$
$z_{1,t+1}, z_{2,t+1}, z_{2,t+2}$	0.069	6.20	4	0.185	$K_2 = 1$

with the present and past vector $\mathbf{P}_t$. An additional small squared canonical correlation value of 0.069 occurs when $\mathbf{F}^*_{t+1} = (z_{1,t+1}, z_{2,t+1}, z_{2,t+2})'$, and this implies that we may have $K_2 = 1$. Hence, this leads to specification of a VARMA(1, 1) model in the echelon form of equation (14.7.1) with Kronecker indices $K_1 = K_2 = 1$. This echelon model form is, in fact, the same as the standard VARMA(1, 1) model in (14.4.1); that is, $K_1 = K_2 = 1$ implies that we have $\mathbf{\Phi}^{\#}_0 = \mathbf{\Theta}^{\#}_0 = \mathbf{I}$ in (14.7.1).

The canonical correlation analysis suggests that a VARMA(1, 1) model might be essentially equivalent to the VAR(2) model in terms of fit, and that these two models are likely superior to other models considered. The parameters of the VARMA(1, 1) model were estimated using the Kronfit routine available in the MTS package of R, and the results are given as

$$\hat{\mathbf{\Phi}}_1 = \begin{bmatrix} 0.440 & -0.200 \\ (0.176) & (0.063) \\ 0.637 & 0.775 \\ (0.210) & (0.076) \end{bmatrix} \quad \hat{\mathbf{\Theta}}_1 = \begin{bmatrix} -0.030 & -0.309 \\ (0.209) & (0.081) \\ 0.313 & 0.227 \\ (0.284) & (0.129) \end{bmatrix}$$

$$\tilde{\mathbf{\Sigma}} = \begin{bmatrix} 5.0239 & 1.6697 \\ 1.6697 & 16.8671 \end{bmatrix}$$

with $|\tilde{\mathbf{\Sigma}}| = 81.9498$, and AIC = 4.608. Again, the coefficient estimate in the (1, 1) position of the matrix $\hat{\mathbf{\Theta}}_1$, as well as estimates in the second row of $\hat{\mathbf{\Theta}}_1$, is not significant and might be omitted from the model.

It is clear from these estimation results, particularly from the estimates $\tilde{\mathbf{\Sigma}}$ and associated summary measures, that the VARMA(1, 1) model provides a nearly equivalent fit to the VAR(2) model. For instance, we consider the coefficient matrices $\mathbf{\Psi}_j$ in the infinite VMA representation for $\boldsymbol{Z}_t$ implied by the VAR(2) and VARMA(1, 1) models. For the VAR(2) model, the $\mathbf{\Psi}_j$ are determined from $\mathbf{\Psi}_1 = \mathbf{\Phi}_1$:

$$\mathbf{\Psi}_j = \mathbf{\Phi}_1\mathbf{\Psi}_{j-1} + \mathbf{\Phi}_2\mathbf{\Psi}_{j-2} \quad \text{for} \quad j > 1\,(\mathbf{\Psi}_0 = \mathbf{I})$$

hence, the $\mathbf{\Psi}_j$ are given as

$$\mathbf{\Psi}_1 = \begin{bmatrix} 0.50 & 0.11 \\ 0.34 & 0.53 \end{bmatrix} \quad \mathbf{\Psi}_2 = \begin{bmatrix} 0.15 & -0.09 \\ 0.61 & 0.46 \end{bmatrix} \quad \mathbf{\Psi}_3 = \begin{bmatrix} -0.00 & -0.12 \\ 0.55 & 0.31 \end{bmatrix}$$

$$\mathbf{\Psi}_4 = \begin{bmatrix} -0.09 & -0.11 \\ 0.41 & 0.16 \end{bmatrix} \quad \mathbf{\Psi}_5 = \begin{bmatrix} -0.11 & -0.08 \\ 0.26 & 0.06 \end{bmatrix} \quad \mathbf{\Psi}_6 = \begin{bmatrix} -0.10 & -0.05 \\ 0.14 & -0.00 \end{bmatrix}$$

and so on, while those for the VARMA(1, 1) model are determined from $\boldsymbol{\Psi}_1 = \boldsymbol{\Phi}_1 - \boldsymbol{\Theta}_1, \boldsymbol{\Psi}_j = \boldsymbol{\Phi}_1 \boldsymbol{\Psi}_{j-1}, j > 1$, and so are given as

$$\boldsymbol{\Psi}_1 = \begin{bmatrix} 0.47 & 0.11 \\ 0.32 & 0.55 \end{bmatrix} \quad \boldsymbol{\Psi}_2 = \begin{bmatrix} 0.14 & -0.06 \\ 0.55 & 0.49 \end{bmatrix} \quad \boldsymbol{\Psi}_3 = \begin{bmatrix} -0.05 & -0.13 \\ 0.52 & 0.34 \end{bmatrix}$$
$$\boldsymbol{\Psi}_4 = \begin{bmatrix} -0.12 & -0.12 \\ 0.37 & 0.19 \end{bmatrix} \quad \boldsymbol{\Psi}_5 = \begin{bmatrix} -0.13 & -0.09 \\ 0.21 & 0.07 \end{bmatrix} \quad \boldsymbol{\Psi}_6 = \begin{bmatrix} -0.10 & -0.05 \\ 0.08 & -0.01 \end{bmatrix}$$

Thus, we see that the $\boldsymbol{\Psi}_j$ coefficient matrices are very similar for both models, implying, in particular, that forecasts $\hat{\boldsymbol{Z}}_t(l)$ and the covariance matrices $\boldsymbol{\Sigma}(l) + \sum_{j=0}^{l-1} \boldsymbol{\Psi}_j \boldsymbol{\Sigma} \boldsymbol{\Psi}_j'$ of the l-step-ahead forecast errors $\boldsymbol{e}_t(l) = \boldsymbol{Z}_{t+l} - \hat{\boldsymbol{Z}}_t(l)$ obtained from the two models, VAR(2) and VARMA(1, 1), are nearly identical.

14.7.3 Partial Canonical Correlation Analysis for Reduced-Rank Structure

Another approach to allow for simplifying structure in the parameterization of the VAR and VARMA models is to incorporate certain *reduced-rank* structure in the coefficient matrices. For the VAR(p) model (14.2.1), Ahn and Reinsel (1988) proposed a particular nested reduced-rank model structure, such that

$$\text{rank}(\boldsymbol{\Phi}_j) = r_j \geq \text{rank}\,(\boldsymbol{\Phi}_{j+1}) = r_{j+1} \quad j = 1, 2, \ldots, p-1$$

and it is also specified that range($\boldsymbol{\Phi}_j$) $\supset$ range($\boldsymbol{\Phi}_{j+1}$). Then the $\boldsymbol{\Phi}_j$ can be represented in reduced-rank factorization form as $\boldsymbol{\Phi}_j = \mathbf{A}_j \mathbf{B}_j$, where $\mathbf{A}_j$ and $\mathbf{B}_j$ are full-rank matrices of dimensions $k \times r_j$ and $r_j \times k$, respectively, with range($\mathbf{A}_j$) $\supset$ range($\mathbf{A}_{j+1}$). One fundamental consequence for this model is that there then exists a full-rank $(k - r_j) \times k$ matrix $\mathbf{F}_j'$, such that $\mathbf{F}_j' \boldsymbol{\Phi}_j = \mathbf{0}$ and hence $\mathbf{F}_j' \boldsymbol{\Phi}_i = \mathbf{0}$ for all $i \geq j$ because of the nested structure. Therefore, the vector

$$\mathbf{F}_j' \left(\boldsymbol{Z}_t - \sum_{i=1}^{j-1} \boldsymbol{\Phi}_i \boldsymbol{Z}_{t-i} \right) = \mathbf{F}_j' \left(\boldsymbol{Z}_t - \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{Z}_{t-i} \right) \equiv \mathbf{F}_j' \boldsymbol{\delta} + \mathbf{F}_j' \boldsymbol{a}_t$$

is uncorrelated with the past values $\boldsymbol{Z}_{j-1,t-1} = (\boldsymbol{Z}_{t-1}', \ldots, \boldsymbol{Z}_{t-j}')'$ and consists of $k - r_j$ linear combinations of $\boldsymbol{Z}_{j-1,t} = (\boldsymbol{Z}_t', \ldots, \boldsymbol{Z}_{t-j+1}')'$. Thus, it follows that $k - r_j$ *zero partial canonical correlations* will occur between $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_{t-j}$, given $\boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_{t-j+1}$. Hence, performing a (partial) canonical correlation analysis for the various values of $j = 1, 2, \ldots$ can identify the simplifying nested reduced-rank structure, as well as the overall order p, of the VAR model.

The sample statistic that can be used to (tentatively) specify the ranks is

$$C(j, r) = -(N - j - jk - 1) \sum_{t=r+1}^{k} \ln[1 - \hat{\rho}_i^2(j)] \tag{14.7.8}$$

for $r = k - 1, k - 2, \ldots, 0$, where $1 \geq \hat{\rho}_1(j) \geq \cdots \geq \hat{\rho}_k(j) > 0$ are the *sample partial canonical correlations* between $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_{t-j}$, given $\boldsymbol{Z}_{t-1}, \ldots, \boldsymbol{Z}_{t-j+1}$. (Calculation of sample canonical correlations was discussed previously in Section 6.2.4.) Under the null hypothesis that rank($\boldsymbol{\Phi}_j$) $\leq r$ within the nested reduced-rank model framework, the statistic $C(j, r)$ is

asymptotically distributed as chi-squared with $(k-r)^2$ degrees of freedom. Hence, if the value of the test statistic is not ''significantly'' large, we would not reject the null hypothesis and might conclude that $\boldsymbol{\Phi}_j$ has reduced rank equal to the smallest value r_j for which the test does not reject the null hypothesis. Note, in particular, that when $r = 0$ the statistic in (14.7.8) is (essentially) the same as the LR test statistic given in (14.2.10) for testing $H_0\colon \boldsymbol{\Phi}_j = \mathbf{0}$ in an VAR(j) model, since it can be verified that $\ln[|\mathbf{S}_j|/|\mathbf{S}_{j-1}|] = \sum_{i=1}^{k} \ln[1-\hat{\rho}_i^2(j)]$.

Once the ranks in the nested reduced-rank VAR model have been specified, the parameters in the restricted model can be estimated by maximum likelihood methods. Some normalization conditions on the $\mathbf{A}_j$ and $\mathbf{B}_j$ in $\boldsymbol{\Phi}_j = \mathbf{A}_j\mathbf{B}_j$ are required to ensure a unique set of parameters. Assuming the components of $\boldsymbol{Z}_t$ are arranged suitably, this parameterization can be obtained as $\boldsymbol{\Phi}_j = \mathbf{A}_1\mathbf{D}_j\mathbf{B}_j$, where $\mathbf{A}_1$ is $k \times r_1$ lower triangular with ones on the main diagonal and may have certain other elements ''normalized'' to fixed values of zero, $\mathbf{B}_j$ contains unrestricted parameters, and $\mathbf{D}_j = [\mathbf{I}_{r_j}, \mathbf{0}]'$ is $r_1 \times r_j$. Asymptotic distribution theory for the ML estimators of parameters of this model extends from theory for the LS estimators in a stationary VAR(p) model in a fairly direct manner.

The structure of the reduced-rank VAR model relates directly to the concepts of Kronecker indices, McMillan degree, and echelon canonical form of VARMA models discussed earlier. In particular, it can be easily verified that the McMillan degree of a nested reduced-rank AR process is equal to $M = \sum_{j=1}^{p} r_j$, the sum of the ranks of the AR coefficient matrices $\boldsymbol{\Phi}_j$. In addition, from the nested reduced-rank structure it follows that the model can also be represented as

$$\boldsymbol{\Phi}_0^{\#}\boldsymbol{Z}_t - \sum_{j=1}^{p} \boldsymbol{\Phi}_j^{\#}\boldsymbol{Z}_{t-j} = \boldsymbol{\delta}^{\#} + \boldsymbol{\Phi}_0^{\#}\boldsymbol{a}_t$$

with $\boldsymbol{\Phi}_0^{\#} = \mathbf{A}^{-1}$, where $\mathbf{A}$ is the $k \times k$ matrix formed by augmenting the $k \times r_1$ matrix $\mathbf{A}_1$ with the last $k - r_1$ columns of the $k \times k$ identity matrix, and

$$\boldsymbol{\Phi}_j^{\#} = \mathbf{A}^{-1}\boldsymbol{\Phi}_j = \mathbf{A}^{-1}\mathbf{A}_1\mathbf{D}_j\mathbf{B}_j \equiv [\mathbf{B}_j', \mathbf{0}']'$$

having its last $k - r_j$ rows equal to zero. This relation can be viewed as an echelon canonical form representation, as in (14.7.1), for the nested reduced-rank vector VAR(p) model. Also, as noted by Reinsel (1997, p. 66), the notion of a nested reduced-rank model and its relationship to the echelon form representation can be directly extended to the VARMA model leading to the specification of a reduced-rank VARMA model for the vector process.

14.8 NONSTATIONARITY AND COINTEGRATION

14.8.1 Vector ARIMA Models

Time series encountered in practice will frequently exhibit nonstationary behavior. To generalize stationary VARMA models to nonstationary processes, we can consider a general form of the VARMA model, $\boldsymbol{\Phi}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}(B)\mathbf{a}_t$, where some of the roots of $\det\{\boldsymbol{\Phi}(B)\} = 0$ are allowed to have absolute value equal to one. More specifically, because of the prominent role of the differencing operator $(1 - B)$ in univariate models, for nonseasonal time series we might only allow some roots to equal one (unit roots) while the remaining roots are all

greater than one in absolute value. A particular restrictive class of models of this type for nonstationary series are of the form

$$\boldsymbol{\Phi}_1(B)\boldsymbol{D}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}(B)\mathbf{a}_t \tag{14.8.1}$$

where $\mathbf{D}(B) = \text{diag}[(1-B)^{d_1}, \ldots, (1-B)^{d_k}]$ is a diagonal matrix, $d_1, \ldots, d_k$ are nonnegative integers, and $\det\{\boldsymbol{\Phi}_1(B)\} = 0$ has all roots greater than one in absolute value. Thus, this model, which is referred to as a vector ARIMA model, simply states that after each series z_{it} is individually differenced an appropriate number (d_i) of times to reduce it to a stationary series, the resulting vector series $\mathbf{W}_t = \mathbf{D}(B)\boldsymbol{Z}_t$ is a stationary VARMA(p, q) process. For vector time series, however, simultaneous differencing of all component series can lead to unnecessary complications in modeling and estimation as a result of ''overdifferencing,'' including noninvertible model representations, so differencing needs to be examined with particular care in the vector case.

14.8.2 Cointegration in Nonstationary Vector Processes

The nonstationary unit-root aspects of a vector process $\boldsymbol{Z}_t$ become more complicated in the multivariate case compared with the univariate case, due in part to the possibility of *cointegration* among the component series z_{it} of a nonstationary vector process $\boldsymbol{Z}_t$. For instance, the possibility exists for each component series z_{it} to be nonstationary with its first difference $(1-B)z_{it}$ stationary (in which case z_{it} is said to be integrated of order one), but such that certain linear combinations $y_{it} = \boldsymbol{b}_i'\boldsymbol{Z}_t$ of $\boldsymbol{Z}_t$ will be stationary. That this possibility exists was demonstrated by Box and Tiao (1977) in their analysis of a five-dimensional dataset from Quenouille (1957). A process $\boldsymbol{Z}_t$ that displays this behavior is said to be *cointegrated* with cointegrating vectors $\boldsymbol{b}_i$ (e.g., Engle and Granger, 1987). An interpretation of cointegrated vector processes $\boldsymbol{Z}_t$ is that the individual components z_{it} share some common nonstationary components or ''common trends''; hence, they tend to have certain similar movements in their longer term behavior. These common trend components will be eliminated upon taking suitable linear combinations of the components of the process $\boldsymbol{Z}_t$. A related interpretation is that the component series z_{it}, although they may exhibit nonstationary behavior, satisfy a long-run equilibrium relation $\boldsymbol{b}_i'\boldsymbol{Z}_t \simeq 0$ such that the process $y_{it} = \boldsymbol{b}'_i\boldsymbol{Z}_t$, which represents the deviation from the equilibrium, exhibits stable behavior and so forms a stationary process. Properties of nonstationary cointegrated systems have been investigated by Engle and Granger (1987) and Johansen (1988), among others.

An Error Correction Form. A specific nonstationary VARMA model structure for which cointegration occurs is the model $\boldsymbol{\Phi}(B)\boldsymbol{Z}_t = \boldsymbol{\Theta}(B)\boldsymbol{a}_t$, where $\det\{\boldsymbol{\Phi}(B)\} = 0$ has $d < k$ roots equal to one and all other roots are greater than one in absolute value, and also the matrix

$$\boldsymbol{\Phi}(1) = \mathbf{I} - \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_p$$

has rank $r = k - d$. Because the process has unit roots fewer than the number of components, this type of process is called *partially nonstationary* by Ahn and Reinsel (1990). For such a process, it can be established that r linearly independent vectors $\boldsymbol{b}_i$ exist such that $\boldsymbol{b}_i'\boldsymbol{Z}_t$ is stationary, and $\boldsymbol{Z}_t$ is said to have cointegrating rank r. A useful approach to the investigation of this model is to express it in its equivalent *error correction* (EC) form

given by

$$\boldsymbol{W}_t = \mathbf{C}\boldsymbol{Z}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_j^* \boldsymbol{W}_{t-j} + \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j} \tag{14.8.2}$$

where $\boldsymbol{W}_t = (1 - B)\boldsymbol{Z}_t, \boldsymbol{\Phi}_j^* = -\sum_{i=j+1}^{p} \boldsymbol{\Phi}_i$, and

$$\mathbf{C} = -\boldsymbol{\Phi}(1) = -\left(\boldsymbol{I} - \sum_{j=1}^{p} \boldsymbol{\Phi}_j\right) \tag{14.8.3}$$

For instance, by subtracting $\boldsymbol{Z}_{t-1}$ from both sides of the VAR(1) model $\boldsymbol{Z}_t = \boldsymbol{\Phi}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t$, we see that the model can be expressed as $(\boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}) = -(\mathbf{I} - \boldsymbol{\Phi})\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t \equiv \mathbf{C}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t$, with $\mathbf{C} = -(\mathbf{I} - \boldsymbol{\Phi})$. The VAR(2) model can be expressed as

$$\begin{aligned}(\boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}) &= -(\mathbf{I} - \boldsymbol{\Phi}_1 - \boldsymbol{\Phi}_2)\boldsymbol{Z}_{t-1} - \boldsymbol{\Phi}_2(\boldsymbol{Z}_{t-1} - \boldsymbol{Z}_{t-2}) + \boldsymbol{a}_t \\ &\equiv \mathbf{C}\boldsymbol{Z}_{t-1} + \boldsymbol{\Phi}_1^*(\boldsymbol{Z}_{t-1} - \boldsymbol{Z}_{t-2}) + \boldsymbol{a}_t\end{aligned}$$

with $\mathbf{C} = -(\mathbf{I} - \boldsymbol{\Phi}_1 - \boldsymbol{\Phi}_2)$ and $\boldsymbol{\Phi}_1^* = -\boldsymbol{\Phi}_2$, and similarly for higher order VAR models.

We note that the error correction form (14.8.2) has an invertible moving average operator but introduces $\mathbf{C}\boldsymbol{Z}_{t-1}$ on the right-hand side of the model. Since the moving average operator remains unchanged, problems associated with noninvertibility are now avoided. The term $\mathbf{C}\boldsymbol{Z}_{t-1}$ is referred to as the error correction term and the rank $r = k - d$ of the coefficient matrix $\mathbf{C}$ represents the number of cointegrating vectors in the system.

To derive an alternative form, we note that the reduced-rank matrix $\mathbf{C}$ can be written as $\mathbf{C} = \mathbf{AB}$, where $\mathbf{A}$ and $\mathbf{B}$ are full-rank matrices of dimensions $k \times r$ and $r \times k$, respectively. We can also determine a full-rank $k \times (k - r)$ matrix $\mathbf{Q}_1$ such that $\mathbf{Q}_1'\mathbf{A} = \mathbf{0}$, hence also $\mathbf{Q}_1'\mathbf{C} = \mathbf{0}$. Hence, it can be established that the r linear combinations $\boldsymbol{Y}_{2t} = \mathbf{B}\boldsymbol{Z}_t$ are stationary, the r rows of $\mathbf{B}$ are linearly independent cointegrating vectors, whereas the $d = k - r$ components $\boldsymbol{Y}_{1t} = \mathbf{Q}'_1\boldsymbol{Z}_t$ are ''purely'' nonstationary and are often referred to as the ''common trends'' among the components of the nonstationary process $\boldsymbol{Z}_t$. Therefore, the error correction form (14.8.2) can also be expressed as

$$\begin{aligned}\boldsymbol{W}_t &= \mathbf{AB}\mathbf{Z}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_j^* \boldsymbol{W}_{t-j} + \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j} \\ &\equiv \mathbf{A}\boldsymbol{Y}_{2,t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_j^* \boldsymbol{W}_{t-j} + \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j}\end{aligned} \tag{14.8.4}$$

Issues of estimation of cointegrated VAR models and testing for the rank r of cointegration will be discussed briefly in Section 14.8.3.

Illustration: Nonstationary VAR(1) Model. To illustrate some of the preceding points, consider the VAR(1) process $\boldsymbol{Z}_t = \boldsymbol{\Phi}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t$ with d eigenvalues of $\boldsymbol{\Phi}$ equal to one and the remaining $r = k - d$ eigenvalues less than one in absolute value, and suppose the d unit eigenvalues have d linearly independent eigenvectors. Then there is a $k \times k$ nonsingular

matrix $\mathbf{P}$ such that $\mathbf{P}^{-1}\mathbf{\Phi}\mathbf{P} = \mathbf{\Lambda}$ with $\mathbf{\Lambda} = \text{diag}(\mathbf{I}_d, \mathbf{\Lambda}_2)$, where $\mathbf{\Lambda}_2 = \text{diag}(\lambda_{d+1}, \ldots, \lambda_k)$ is an $r \times r$ diagonal matrix with $|\lambda_i| < 1$. Letting $\mathbf{P} = [\mathbf{P}_1, \mathbf{P}_2]$ and $\mathbf{Q} = \mathbf{P}^{-1} = [\mathbf{Q}_1, \mathbf{Q}_2]'$, where $\mathbf{P}_1$ and $\mathbf{Q}_1$ are $k \times d$ matrices, define $\boldsymbol{Y}_t = \mathbf{Q}\boldsymbol{Z}_t = (\boldsymbol{Y}'_{1t}, \boldsymbol{Y}'_{2t})'$, that is, $\boldsymbol{Y}_{1t} = \mathbf{Q}'_1\boldsymbol{Z}_t$ and $\boldsymbol{Y}_{2t} = \mathbf{Q}'_2\boldsymbol{Z}_t$, and similarly $\boldsymbol{\varepsilon}_t = \mathbf{Q}\boldsymbol{a}_t = (\boldsymbol{\varepsilon}'_{1t}, \boldsymbol{\varepsilon}'_{2t})'$. Then we have

$$\mathbf{Q}\boldsymbol{Z}_t = \mathbf{Q}\mathbf{\Phi}\mathbf{P}\mathbf{Q}\boldsymbol{Z}_{t-1} + \mathbf{Q}\boldsymbol{a}_t$$

or $\boldsymbol{Y}_t = \mathbf{\Lambda}\boldsymbol{Y}_{t-1} + \boldsymbol{\varepsilon}_t$. Therefore, the model in terms of $\boldsymbol{Y}_t$ reduces to

$$(1 - B)\boldsymbol{Y}_{1t} = \boldsymbol{\varepsilon}_{1t} \quad \text{and} \quad (\mathbf{I} - \mathbf{\Lambda}_2 B)\boldsymbol{Y}_{2t} = \boldsymbol{\varepsilon}_{2t}$$

so $\{\boldsymbol{Y}_{1t}\}$ is a d-dimensional purely nonstationary series, whereas $\{\boldsymbol{Y}_{2t}\}$ is an r-dimensional stationary series. Thus, $\{\boldsymbol{Z}_t\}$ is nonstationary but has r linearly independent linear combinations $\boldsymbol{Y}_{2t} = \mathbf{Q}'_2\boldsymbol{Z}_t$, which are stationary, so $\boldsymbol{Z}_t$ is cointegrated with cointegrating rank r and linearly independent cointegrating vectors, which are the rows of $\mathbf{Q}'_2$. Conversely, since $\boldsymbol{Z}_t = \mathbf{P}\boldsymbol{Y}_t = \mathbf{P}_1\boldsymbol{Y}_{1t} + \mathbf{P}_2\boldsymbol{Y}_{2t}$, the components of the vector $\boldsymbol{Z}_t$ are linear combinations of a nonstationary vector (random walk) component $\boldsymbol{Y}_{1t}$ and a stationary VAR(1) component $\boldsymbol{Y}_{2t}$. Also notice that the error correction form of this VAR(1) model as in (14.8.2) is

$$\boldsymbol{W}_t = \boldsymbol{Z}_t - \boldsymbol{Z}_{t-1} = \mathbf{C}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t$$

where

$$\mathbf{C} = -(\mathbf{I} - \mathbf{\Phi}) = -\mathbf{P}(\mathbf{I} - \mathbf{\Lambda})\mathbf{Q} = -\mathbf{P}_2(\mathbf{I}_r - \mathbf{\Lambda}_2)\mathbf{Q}'_2$$

which is clearly of reduced rank r.

14.8.3 Estimation and Inferences for Cointegrated VAR Models

As noted above, when the vector series $\boldsymbol{Z}_t$ is unit-root nonstationary but with cointegration features, it is not appropriate to difference all component series and model the resulting series $\mathbf{W}_t = (1 - B)\boldsymbol{Z}_t$ by a VAR or VARMA model. Instead, we may prefer to incorporate the unit-root and cointegration features into the analysis using the model (14.8.2). This can provide a better understanding on the nature of the nonstationarity and improve the forecasting performance of the model. This section examines the estimation and statistical inference for this model focusing on the special case of an error correction VAR(p) model

$$\boldsymbol{W}_t = \mathbf{C}\boldsymbol{Z}_{t-1} + \sum_{j=1}^{p-1} \mathbf{\Phi}^*_j \boldsymbol{W}_{t-j} + \boldsymbol{a}_t \tag{14.8.5}$$

where $\boldsymbol{W}_t = \boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}$ with $\text{rank}(\mathbf{C}) = r < k$. Note that a *special case* of (14.8.5), at one extreme, occurs with $r = 0$ (i.e., $d = k$ unit roots and $\mathbf{C} = \mathbf{0}$) and leads to a usual VAR model of order $p - 1$ for the series of first differences $\boldsymbol{W}_t$.

The least-squares and Gaussian maximum likelihood estimation of cointegrated VAR models and likelihood ratio testing for the rank of cointegration, generally utilizing the error correction form (14.8.5) of the model, have been examined by several authors including Johansen (1988, 1991), Johansen and Juselius (1990), Ahn and Reinsel (1990), and Reinsel and Ahn (1992). Estimation of the cointegrated model, which imposes the restriction on the number d of unit roots in $\mathbf{\Phi}(B)$ (or the number $r = k - d$ of cointegrating relations), is equivalent to *reduced-rank* estimation, which imposes the restriction on the rank r of

the coefficient matrix $\mathbf{C}$, which can be written as $\mathbf{C} = \mathbf{AB}$ as noted in Section 14.8.2. So techniques from reduced-rank estimation of multivariate regression models can be utilized.

When there are no additional constraints on the coefficient matrices $\mathbf{\Phi}_j^*$ in (14.8.5), that is, the only parameter constraints involved in the model are $\text{rank}(\mathbf{C}) \leq r$, it follows from the original work by Anderson (1951) on reduced-rank regression that the Gaussian (ML) reduced-rank estimation can be obtained explicitly through the partial canonical correlation analysis between $\boldsymbol{W}_t$ and $\boldsymbol{Z}_{t-1}$, given $\boldsymbol{W}_{t-1}, \ldots, \boldsymbol{W}_{t-p+1}$. When there are additional constraints, however, iterative numerical techniques are needed for the Gaussian ML estimation. Specifically, in the partial canonical correlation analysis approach, let $\tilde{\boldsymbol{W}}_t$ and $\tilde{\boldsymbol{Z}}_{t-1}$ denote the residual vectors from least-squares regressions of $\boldsymbol{W}_t$ and $\boldsymbol{Z}_{t-1}$, respectively, on the lagged values $\boldsymbol{W}_{t-1}, \ldots, \boldsymbol{W}_{t-p+1}$, and let

$$\mathbf{S}_{\tilde{w}\tilde{w}} = \sum_{t=1}^{N} \tilde{\boldsymbol{W}}_t \tilde{\boldsymbol{W}}_t' \quad \mathbf{S}_{\tilde{w}\tilde{z}} = \sum_{t=1}^{N} \tilde{\boldsymbol{W}}_t \tilde{\boldsymbol{Z}}_{t-1}' \quad \mathbf{S}_{\tilde{z}\tilde{z}} = \sum_{t=1}^{N} \tilde{\boldsymbol{Z}}_{t-1} \tilde{\boldsymbol{Z}}_{t-1}'$$

Then the *Gaussian reduced-rank estimator* of $\mathbf{C}$ in model (14.8.5) can be expressed explicitly as

$$\tilde{\mathbf{C}} = \hat{\mathbf{\Sigma}}\hat{\mathbf{V}}\hat{\mathbf{V}}'\hat{\mathbf{C}} \tag{14.8.6}$$

where $\hat{\mathbf{C}} = \mathbf{S}_{\tilde{w}\tilde{z}}\mathbf{S}_{\tilde{z}\tilde{z}}^{-1}$ is the full-rank LS estimator of $\mathbf{C}$, $\hat{\mathbf{\Sigma}}$ is the corresponding residual covariance matrix estimate of $\mathbf{\Sigma} = \text{cov}[\mathbf{a}_t]$ from the full-rank LS estimation of (14.8.5), and $\hat{\mathbf{V}} = [\hat{\boldsymbol{V}}_1, \ldots, \hat{\boldsymbol{V}}_r]$ are the vectors corresponding to the r largest partial canonical correlations $\hat{\rho}_i(p), i = 1, \ldots, r$. The vectors $\hat{\boldsymbol{V}}_i$ are normalized so that $\hat{\mathbf{V}}'\hat{\mathbf{\Sigma}}\hat{\mathbf{V}} = \mathbf{I}_r$. Note that the form of the estimator (14.8.6) provides the reduced-rank factorization as $\tilde{\mathbf{C}} = (\hat{\mathbf{\Sigma}}\hat{\mathbf{V}})(\hat{\mathbf{V}}'\hat{\mathbf{C}}) \equiv \hat{\mathbf{A}}\hat{\mathbf{B}}$, with $\hat{\mathbf{A}} = \hat{\mathbf{\Sigma}}\hat{\mathbf{V}}$ satisfying the normalization $\hat{\mathbf{A}}'\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{A}} = \mathbf{I}_r$.

The asymptotic distribution theory of the LS and reduced-rank estimators, $\hat{\mathbf{C}}$ and $\tilde{\mathbf{C}}$, and of LR test statistics for rank has been established and the limiting distributions represented as functionals of vector Brownian motion processes, extending the ''nonstandard'' unit-root asymptotic distribution theory for univariate AR models as outlined in Section 10.1. In particular, we discuss the LR statistic for the test of the hypothesis H_0: $\text{rank}(\mathbf{C}) \leq r$ for model (14.8.5). The LR test statistic is given by

$$-N \ln(U) = -N \ln\left(\frac{|\mathbf{S}|}{|\mathbf{S}_0|}\right)$$

where $\mathbf{S}$ denotes the residual sum-of-squares matrix in the full-rank LS estimation (such that $\hat{\mathbf{\Sigma}} = N^{-1}\mathbf{S}$), while $\mathbf{S}_0$ is the residual sum-of-squares matrix obtained under the reduced-rank restriction that $\text{rank}(\mathbf{C}) = r$. Again from the work of Anderson (1951), it is established that

$$\mathbf{S}_0 = \mathbf{S} + (\hat{\mathbf{C}} - \tilde{\mathbf{C}})\mathbf{S}_{\tilde{z}\tilde{z}}(\hat{\mathbf{C}} - \tilde{\mathbf{C}})'$$

and it follows that

$$|\mathbf{S}_0| = |\mathbf{S}| \prod_{i=r+1}^{k} [1 - \hat{\rho}_i^2(p)]^{-1}$$

where the $\hat{\rho}_i(p)$ are the $d = k - r$ *smallest* sample partial canonical correlations between $\boldsymbol{W}_t$ and $\boldsymbol{Z}_{t-1}$, given $\boldsymbol{W}_{t-1} \ldots, \boldsymbol{W}_{t-p+1}$. Therefore, the LR statistic can be expressed equivalently as

$$-N \ln(U) = -N \sum_{i=r+1}^{k} \ln[1 - \hat{\rho}_i^2(p)] \tag{14.8.7}$$

The limiting distribution for the LR statistic has been derived, based on limiting distribution properties for the LS and reduced-rank estimators $\hat{\mathbf{C}}$ and $\tilde{\mathbf{C}}$, and its limiting distribution is represented by

$$-N \ln(U) \xrightarrow{\mathcal{D}} \operatorname{tr}\left\{ \left[\int_0^1 \mathbf{B}_d(u) d\mathbf{B}_d(u)' \right]' \left[\int_0^1 \mathbf{B}_d(u)\mathbf{B}_d(u)' du \right]^{-1} \times \left[\int_0^1 \mathbf{B}_d(u) d\mathbf{B}_d(u)' \right] \right\} \tag{14.8.8}$$

where $\mathbf{B}_d(u)$ is a d-dimensional standard Brownian motion process, with $d = k - r$. The limiting distribution of the LR statistic under H_0 depends only on d and not on any nuisance parameters or the order p of the VAR model. Note that in the special case of testing for (at least) one unit root, $d = 1$, the limiting distribution in (14.8.8) reduces to

$$-N \ln(U) \xrightarrow{\mathcal{D}} \frac{\left[\int_0^1 B_1(u) d B_1(u) \right]^2}{\int_0^1 B_1(u)^2 du}$$

which is the asymptotic distribution for the (univariate) unit root statistic $\hat{\tau}^2$ in the univariate AR(1) model as discussed in Section 10.1.1.

Critical values of the limiting distribution in (14.8.8) have been obtained by simulation by Johansen (1988) and Reinsel and Ahn (1992) and can be used in the test of H_0. Similar to other LR testing procedures in multivariate linear models, it is suggested that the LR statistic in (14.8.7) be modified to $-(N - kp) \sum_{i=r+1}^{k} \ln[1 - \hat{\rho}_i^2(p)]$ for practical use in finite samples, as this may provide a test statistic whose finite sample distribution is closer to the limiting distribution in (14.8.8) than the ''unmodified'' LR test statistic. The ML estimation and LR testing procedures and asymptotic theory are also extended to the more practical case where a constant term $\boldsymbol{\delta}$ is included in the estimation of the VAR(p) model in error correction form, $\boldsymbol{W}_t = \mathbf{C}\boldsymbol{Z}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_j^* \boldsymbol{W}_{t-j} + \boldsymbol{\delta} + \mathbf{a}_t$. A recommended procedure to be used in specification of the rank r or $\mathbf{C}$ in model (14.8.5) is thus based on performing LR tests of H_0: rank$(\mathbf{C}) \le r$ for a sequence of values of $r = k - 1, k - 2, \ldots, 1, 0$, and an appropriate value of r can be chosen as the smallest value for which H_0 is not rejected. For further discussion of the model building process and for software demonstrations using the R package, the readers are referred to Tsay (2014).

APPENDIX A14.1 SPECTRAL CHARACTERISTICS AND LINEAR FILTERING RELATIONS FOR STATIONARY MULTIVARIATE PROCESSES

A14.1.1 Spectral Characteristics for Stationary Multivariate Processes

The covariance-generating matrix function (provided $\sum_{l=-\infty}^{\infty} |\gamma_{i,j}(l)| < \infty, i, j = 1, \ldots, k$) is defined as $\mathbf{G}(z) = \sum_{l=-\infty}^{\infty} \mathbf{\Gamma}(l) z^l$, and the *spectral density* matrix of the stationary process $\{\boldsymbol{Z}_t\}$ as a function of frequency f is defined as

$$\mathbf{P}(f) = 2\mathbf{G}(e^{-i2\pi f}) = 2 \sum_{l=-\infty}^{\infty} \mathbf{\Gamma}(l) e^{-i2\pi f l} \qquad 0 \le f \le \frac{1}{2} \tag{A14.1.1}$$

The (h, j)th element of $\mathbf{P}(f)$, denoted as $p_{hj}(f)$, is

$$p_{hj}(f) = 2 \sum_{l=-\infty}^{\infty} \gamma_{hj}(l) e^{-i2\pi f l}$$

For $h = j, p_{jj}(f)$ is the (auto)spectral density function of the series z_{jt}, while for $h \neq j$, $p_{hj}(f)$ is the cross-spectral density function of z_{ht} and z_{jt}. Notice that $p_{jj}(f)$ is real valued and nonnegative, but since $\gamma_{hj}(l) \neq \gamma_{hj}(-l)$ for $h \neq j$, the cross-spectral density function $p_{hj}(f)$ is in general complex valued, with $p_{hj}(f)$ being equal to $p_{jh}(-f)$, the complex conjugate of $p_{jh}(f)$. Therefore, the spectral density matrix $\mathbf{P}(f)$ is Hermitian, that is, $\mathbf{P}(f) = \mathbf{P}(-f)'$. Moreover, $\mathbf{P}(f)$ is a nonnegative-definite matrix in the sense that $\boldsymbol{b}'\mathbf{P}(f)\boldsymbol{b} \ge 0$ for any k-dimensional (real-valued) vector $\boldsymbol{b}$, since $\boldsymbol{b}'\mathbf{P}(f)\boldsymbol{b}$ is the spectral density function of the linear combination $\boldsymbol{b}'\boldsymbol{Z}_t$ and hence must be nonnegative. Note also that

$$\mathbf{\Gamma}(l) = \frac{1}{2} \int_{-1/2}^{1/2} e^{i2\pi f l} \mathbf{P}(f)\, df \quad l = 0, \pm 1, \pm 2, \ldots \tag{A14.1.2}$$

that is, $\gamma_{hj}(l) = \frac{1}{2} \int_{-1/2}^{1/2} e^{i2\pi f l} p_{hj}(f)\, df$.

The real part of $p_{hj}(f)$, denoted as $c_{hj}(f) = \mathrm{Re}\{p_{hj}(f)\}$, is called the *co-spectrum*, and the negative of the imaginary part, denoted as $q_{hj}(f) = -\mathrm{Im}\{p_{hj}(f)\}$, is called the *quadrature spectrum*. We can also express $p_{hj}(f)$ in polar form as $p_{hj}(f) = \alpha_{hj}(f) e^{i\phi_{hj}(f)}$, where

$$\alpha_{hj}(f) = |p_{hj}(f)| = \{c_{hj}^2(f) + q_{hj}^2(f)\}^{1/2}$$

and $\phi_{hj}(f) = \tan^{-1}\{-q_{hj}(f)/c_{hj}(f)\}$. The function $\alpha_{hj}(f)$ is called the *cross-amplitude spectrum* and $\phi_{hj}(f)$ is the *phase spectrum*.

Similar to the univariate case, the spectral density matrix $\mathbf{P}(f)$ represents the covariance matrix of the random vector of components at frequency f in the theoretical spectral representations of the components z_{jt} of the vector process $\{\boldsymbol{Z}_t\}$ corresponding to the finite-sample Fourier representations of the time series z_{jt} as in (2.2.1). The (*squared*) coherency spectrum of a pair of series z_{ht} and z_{jt} is defined as

$$k_{hj}^2(f) = \frac{|p_{hj}(f)|^2}{\{p_{hh}(f) p_{jj}(f)\}}$$

The coherency $k_{hj}(f)$ at frequency f can thus be interpreted as the correlation coefficient between the random components at frequency f in the theoretical spectral representations of z_{ht} and z_{jt}. Hence, $k_{hj}(f)$ as a function of f measures the extent to which the two processes z_{ht} and z_{jt} are linearly related in terms of the degree of linear association of their random components at different frequencies f. When spectral relations that involve more than two time series are considered, the related concepts of partial coherency and multiple coherency are also of interest. Detailed accounts of the spectral theory and analysis of multivariate time series may be found in the books by Jenkins and Watts (1968), Hannan (1970), Priestley (1981), and Bloomfield (2000).

A14.1.2 Linear Filtering Relations for Stationary Multivariate Processes

The representation of dynamic linear relationships through the formulation of linear filters is fundamental to the study of stationary multivariate time series. An important example is the moving average representation of the k-dimensional process $\boldsymbol{Z}_t$ in (14.1.4). More generally, a multivariate *linear* (*time-invariant*) *filter* relating an r-dimensional input series $\boldsymbol{X}_t$ to a k-dimensional output series $\boldsymbol{Z}_t$ is given by the form

$$\boldsymbol{Z}_t = \sum_{j=-\infty}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{X}_{t-j} \tag{A14.1.3}$$

where the $\boldsymbol{\Psi}_j$ are $k \times r$ matrices. The filter is physically realizable or *causal* when the $\boldsymbol{\Psi}_j = 0$ for $j < 0$, so that $\boldsymbol{Z}_t = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j X_{t-j}$ is expressible in terms of only present and past values of the input process $\{\boldsymbol{X}_t\}$. The filter is said to be *stable* if $\sum_{j=-\infty}^{\infty} \|\boldsymbol{\Psi}_j\| < \infty$, where $\|\mathbf{A}\|$ denotes a norm for the matrix $\mathbf{A}$ such as $\|\mathbf{A}\|^2 = \text{tr}\{\mathbf{A}'\mathbf{A}\}$. When the filter is stable and the input series $\boldsymbol{X}_t$ is stationary with cross-covariance matrices $\boldsymbol{\Gamma}_{xx}(l)$, the output $\boldsymbol{Z}_t = \sum_{j=-\infty}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{X}_{t-j}$ is a stationary process. The cross-covariance matrices of the stationary process $\{\boldsymbol{Z}_t\}$ are then given by

$$\boldsymbol{\Gamma}_{zz}(l) = \text{cov}[\boldsymbol{Z}_t, \boldsymbol{Z}_{t+l}] = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \boldsymbol{\Psi}_i \boldsymbol{\Gamma}_{xx}(l + i - j) \boldsymbol{\Psi}_j' \tag{A14.1.4}$$

It also follows, from (A14.1.1), that the spectral density matrix of the output $\boldsymbol{Z}_t$ has the representation

$$\mathbf{P}_{zz}(f) = \boldsymbol{\Psi}(e^{i2\pi f}) \mathbf{P}_{xx}(f) \boldsymbol{\Psi}(e^{-i2\pi f})' \tag{A14.1.5}$$

where $\mathbf{P}_{xx}(f)$ is the spectral density matrix of $\boldsymbol{X}_t$, and $\boldsymbol{\Psi}(z) = \sum_{j=-\infty}^{\infty} \boldsymbol{\Psi}_j z^j$ is the *transfer function* (matrix) of the linear filter. In addition, the cross-covariance matrices between $\boldsymbol{Z}_t$ and $\boldsymbol{X}_t$ are

$$\boldsymbol{\Gamma}_{zx}(l) = \text{cov}[\boldsymbol{Z}_t, \ \boldsymbol{X}_{t+l}] = \sum_{j=-\infty}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\Gamma}_{xx}(l + j)$$

and the cross-spectral density matrix between $\boldsymbol{Z}_t$ and $\boldsymbol{X}_t$ is

$$\mathbf{P}_{zx}(f) = 2 \sum_{l=-\infty}^{\infty} \boldsymbol{\Gamma}_{zx}(l) e^{-i2\pi f l} = \boldsymbol{\Psi}(e^{i2\pi f}) \mathbf{P}_{xx}(f)$$

so the transfer function $\boldsymbol{\Psi}(z)$ satisfies the relation $\boldsymbol{\Psi}(e^{i2\pi f}) = \mathbf{P}_{zx}(f)\mathbf{P}_{xx}(f)^{-1}$. In practice, when a causal linear filter is used to represent the relation between an observable input process $\boldsymbol{X}_t$ and an output process $\boldsymbol{Z}_t$ in a dynamic system, there will be added unobserved noise $\boldsymbol{N}_t$ in the system and a dynamic model of the form $\boldsymbol{Z}_t = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{X}_{t-j} + \boldsymbol{N}_t$ will be useful.

For a special example of the above linear filtering results, consider the basic stationary vector *white noise process* $\{\boldsymbol{a}_t\}$ defined in Section 14.1.3, with the properties that $E[\boldsymbol{a}_t] = 0$, $E[\boldsymbol{a}_t \boldsymbol{a}_t'] = \boldsymbol{\Sigma}$, and $E[\boldsymbol{a}_t \boldsymbol{a}_{t+l}'] = 0$ for $l \neq 0$. Hence, $\boldsymbol{a}_t$ has spectral density matrix $\mathbf{P}_{aa}(f) = 2\boldsymbol{\Sigma}$. Then the process $\boldsymbol{Z}_t = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{a}_{t-j}$, with $\sum_{j=0}^{\infty} \|\boldsymbol{\Psi}_j\| < \infty$, is stationary and has cross-covariance matrices

$$\boldsymbol{\Gamma}_{zz}(l) = \sum_{j=0}^{\infty} \boldsymbol{\Psi}_j \boldsymbol{\Sigma} \boldsymbol{\Psi}_{j+l}' \tag{A14.1.6}$$

and spectral density matrix

$$\mathbf{P}_{zz}(f) = 2\boldsymbol{\Psi}(e^{i2\pi f})\boldsymbol{\Sigma}\boldsymbol{\Psi}(e^{-i2\pi f})' \tag{A14.1.7}$$

and the cross-covariance matrices between $\{\boldsymbol{Z}_t\}$ and $\{\boldsymbol{a}_t\}$ are $\boldsymbol{\Gamma}_{za}(l) = \boldsymbol{\Psi}_{-l}\boldsymbol{\Sigma}$ for $l \leq 0$ and zero for $l > 0$.

In addition, for the stationary VARMA(p, q) process with infinite MA representation (14.1.4), the covariance matrix-generating function is given by $\mathbf{G}(z) = \sum_{l=-\infty}^{\infty} \boldsymbol{\Gamma}(l) z^l = \boldsymbol{\Psi}(z^{-1})\boldsymbol{\Sigma}\boldsymbol{\Psi}(z)'$; hence, the spectral density matrix of the VARMA(p, q) process is given as in (A14.1.7) with $\boldsymbol{\Psi}(z) = \boldsymbol{\Phi}^{-1}(z)\boldsymbol{\Theta}(z)$.

EXERCISES

14.1. Consider the bivariate VMA(1) process $\boldsymbol{Z}_t = (\mathbf{I} - \boldsymbol{\Theta}B)\boldsymbol{a}_t$, with

$$\boldsymbol{\Theta} = \begin{bmatrix} 0.4 & 0.3 \\ -0.5 & 0.8 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$$

(a) Find the lag 0 and lag 1 autocorrelations and cross-correlations of $\boldsymbol{Z}_t$; that is, find the matrices $\boldsymbol{\Gamma}(0), \boldsymbol{\Gamma}(1)$, and $\boldsymbol{\rho}(0), \boldsymbol{\rho}(1)$.

(b) Find the individual univariate MA(1) models for z_{1t} and z_{2t}; that is, in the models $z_{it} = (1 - \eta_i B)\varepsilon_{it}, i = 1, 2$, find the values of the parameters η_i and $\sigma_{\varepsilon_i}^2 = \text{var}[\varepsilon_{it}]$, from $\boldsymbol{\Gamma}(0)$ and $\boldsymbol{\Gamma}(1)$.

(c) Show that the bivariate VMA(1) model above is invertible, state the matrix difference equation satisfied by the matrix weights $\boldsymbol{\Pi}_j$ in the infinite AR form of the VMA(1) model, and explicitly evaluate the $\boldsymbol{\Pi}_j$ for $j = 1, 2, 3, 4$.

(d) It follows from Section 14.5.2 that the diagonal elements of $\boldsymbol{\Sigma}$ represent the one-step-ahead forecast error variances for the two series when each series is forecast from the past history of both series, that is, when each series is forecast based on the bivariate model. Compare these one-step forecast error variances

in the bivariate model with the one-step forecast error variances $\sigma_{\varepsilon_i}^2$ based on the individual univariate models in (b).

14.2. For the stationary multivariate VAR(1) model $(\mathbf{I} - \mathbf{\Phi}B)\boldsymbol{Z}_t = \boldsymbol{a}_t$, it is known that $\mathbf{\Gamma}(0) - \mathbf{\Phi}\mathbf{\Gamma}(0)\mathbf{\Phi}' = \mathbf{\Sigma}$. Hence, if the model parameters $\mathbf{\Phi}$ and $\mathbf{\Sigma}$ are given, this matrix equation may be solved to determine $\mathbf{\Gamma}(0)$. In the bivariate case, this leads to three linear equations in the unknowns $\gamma_{11}(0), \gamma_{12}(0)$, and $\gamma_{22}(0)$. If these equations are expressed in matrix form as $\boldsymbol{A}[\gamma_{11}(0), \gamma_{12}(0), \gamma_{22}(0)]' = \boldsymbol{b}$, give explicitly the expressions for $\boldsymbol{A}$ and $\boldsymbol{b}$. Consider the specific case

$$\mathbf{\Phi} = \begin{bmatrix} 0.2 & 0.3 \\ -0.6 & 1.1 \end{bmatrix} \qquad \mathbf{\Sigma} = \begin{bmatrix} 4 & 1 \\ 1 & 1 \end{bmatrix}$$

(a) Show that

$$\mathbf{\Gamma}(0) = \begin{bmatrix} 5.667 & 4.000 \\ 4.000 & 10.667 \end{bmatrix}$$

Also, determine the stationarity of the VAR(1) model above, state the difference equation satisfied by the $\mathbf{\Gamma}(j), j \geq 1$, and find the values of $\mathbf{\Gamma}(1), \mathbf{\Gamma}(2)$, and $\mathbf{\Gamma}(3)$. In addition, compute the cross-correlation matrices $\boldsymbol{\rho}(0), \boldsymbol{\rho}(1)$, $\boldsymbol{\rho}(2)$, and $\boldsymbol{\rho}(3)$.

(b) Find the matrix coefficients $\mathbf{\Psi_1}, \mathbf{\Psi_2}$, and $\mathbf{\Psi_3}$ in the infinite MA representation for $\boldsymbol{Z}_t$, and hence, compute the covariance matrix of the bivariate lead l forecast errors from the bivariate model using the formula $\mathbf{\Sigma}(l) = \sum_{j=0}^{l-1} \mathbf{\Psi}_j \mathbf{\Sigma} \mathbf{\Psi}_j'$, for $l = 1, 2, 3$.

(c) For a bivariate VAR(1) model, indicate what simplifications occur in the model when $\mathbf{\Phi}$ is lower triangular (i.e., $\phi_{12} = 0$). In particular, show in this case that the bivariate system can be expressed equivalently in the form of a ''unidirectional'' transfer function model, as in Chapter 12, with z_{1t} as input series and z_{2t} as output. In addition, indicate the specific nature of the univariate ARMA model for the series z_{2t} implied by this situation.

(d) For a bivariate VAR(1) model, show that the case $\det(\mathbf{\Phi}) = 0$ implies that there exists a linear combination of z_{1t} and $z_{2t}, Y_{1t} = c_{11}z_{1t} + c_{12}z_{2t}$, which is a white noise series, and a second linear combination $Y_{2t} = c_{21}z_{1t} + c_{22}z_{2t}$, which is again a univariate AR(1) process.

Hint: If $\det(\mathbf{\Phi}) = 0$, then $\mathbf{\Phi}$ has rank at most one and can be written as

$$\mathbf{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \alpha\phi_{11} & \alpha\phi_{12} \end{bmatrix} = \begin{bmatrix} 1 \\ \alpha \end{bmatrix} \begin{bmatrix} \phi_{11} & \phi_{12} \end{bmatrix}$$

14.3. Consider the VAR(p) model

$$\boldsymbol{Z}_t = \mathbf{\Phi}_1 \boldsymbol{Z}_{t-1} + \mathbf{\Phi}_2 \boldsymbol{Z}_{t-2} + \cdots + \mathbf{\Phi}_p \boldsymbol{Z}_{t-p} + \boldsymbol{a}_t$$

Verify that the model can be expressed as a VAR(1) model in terms of the kp-dimensional vector $\boldsymbol{Y}_t = (\boldsymbol{Z}_t', \boldsymbol{Z}_{t-1}', \ldots, \boldsymbol{Z}_{t-p+1}')'$, $\boldsymbol{Y}_t = \boldsymbol{\Phi}\boldsymbol{Y}_{t-1} + \boldsymbol{e}_t$, using the $kp \times kp$ *companion matrix* $\boldsymbol{\Phi}$ for the VAR(p) operator $\boldsymbol{\Phi}(B) = \boldsymbol{I} - \boldsymbol{\Phi}_1 B - \cdots - \boldsymbol{\Phi}_p B^p$,

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 & \cdots & \cdots & \boldsymbol{\Phi}_p \\ \mathbf{I} & 0 & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \cdot \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix}$$

In addition, show that $\det\{\mathbf{I} - \boldsymbol{\Phi}B\} = \det\{\mathbf{I} - \boldsymbol{\Phi}_1 B - \cdots - \boldsymbol{\Phi}_p B^p\}$, and hence the stationarity condition for the VAR(p) process is equivalent to the condition that all eigenvalues of the companion matrix $\boldsymbol{\Phi}$ be less than one in absolute value. (*Hint*: To evaluate $\det\{\mathbf{I} - \boldsymbol{\Phi}B\}$, multiply the ith column of $\mathbf{I} - \boldsymbol{\Phi}B$ by B and add to the $(i-1)$st column, successively, for $i = p, p-1, \ldots, 2$.)

14.4. For a bivariate VAR(2) model $\boldsymbol{Z}_t = \boldsymbol{\Phi}_1 \boldsymbol{Z}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{Z}_{t-2} + \boldsymbol{a}_t$, with

$$\boldsymbol{\Phi}_1 = \begin{bmatrix} 1.5 & -0.6 \\ 0.3 & 0.2 \end{bmatrix} \qquad \boldsymbol{\Phi}_2 = \begin{bmatrix} -0.5 & 0.3 \\ 0.7 & -0.2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$$

(a) Verify that this model is stationary based on the nature of the roots of $\det\{\mathbf{I} - \boldsymbol{\Phi}_1 B - \boldsymbol{\Phi}_2 B^2\} = 0$. (Note that you may want to make use of the result of Exercise 14.3 for computational convenience.)

(b) Calculate forecasts $\hat{\boldsymbol{Z}}_n(l)$ for $l = 1, \ldots, 5$ steps ahead, given that $\boldsymbol{Z}_n = (1.2, 0.6)'$ and $\boldsymbol{Z}_{n-1} = (0.5, 0.9)'$.

(c) Find the coefficient matrices $\boldsymbol{\Psi}_j$, $j = 1, \ldots, 4$, in the infinite MA representation of the process, and find the forecast error covariance matrices $\boldsymbol{\Sigma}(l)$ for $l = 1, \ldots, 5$.

14.5. Consider the simple transfer function model

$$(1 - B)z_{1t} = \varepsilon_{1t} - \theta\varepsilon_{1,t-1} \qquad z_{2t} = \omega z_{1t} + \varepsilon_{2t}$$

where ε_{1t} and ε_{2t} are independent white noise processes.

(a) Determine the univariate ARIMA model for z_{2t}, and note that z_{2t} is nonstationary.

(b) Express the bivariate model for $\boldsymbol{Z}_t = (z_{1t}, z_{2t})'$ in the general form of a ''generalized'' ARMA(1, 1) model, $(\mathbf{I} - \boldsymbol{\Phi}_1 B)\boldsymbol{Z}_t = (\mathbf{I} - \boldsymbol{\Theta}_1 B)\boldsymbol{a}_t$, and determine that one of the eigenvalues of $\boldsymbol{\Phi}_1$ is equal to one.

(c) Determine the bivariate model for the first differences $(1 - B)\boldsymbol{Z}_t$, and show that it has the form of a bivariate IMA(1, 1) model, $(1 - B)\boldsymbol{Z}_t = (\mathbf{I} - \boldsymbol{\Theta}^* B)\boldsymbol{a}_t$,

where the MA operator $(\mathbf{I} - \boldsymbol{\Theta}^* B)$ is not invertible. Hence, this model represents an ''overdifferencing'' of the bivariate series $\boldsymbol{Z}_t$.

14.6. Suppose $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$, with $N = 60$, is a sample from a bivariate VAR(1) process, with sample covariance matrices obtained as

$$\hat{\boldsymbol{\Gamma}}(0) = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 2.0 \end{bmatrix} \qquad \hat{\boldsymbol{\Gamma}}(1) = \begin{bmatrix} 0.6 & 0.4 \\ 0.7 & 1.2 \end{bmatrix} \qquad \hat{\boldsymbol{\Gamma}}(2) = \begin{bmatrix} 0.30 & 0.10 \\ 0.42 & 0.64 \end{bmatrix}$$

(a) Obtain the corresponding estimated correlation matrices $\hat{\boldsymbol{\rho}}(0), \hat{\boldsymbol{\rho}}(1)$, and $\hat{\boldsymbol{\rho}}(2)$.

(b) Find the sample Yule–Walker estimates for $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ in the VAR(1) model, and find an estimate for the approximate covariance matrix of the estimator $\hat{\boldsymbol{\Phi}}$, that is, for the covariance matrix of $\text{vec}[\hat{\boldsymbol{\Phi}}']$.

(c) Based on the results in (b), test whether the matrix $\boldsymbol{\Phi}$ has a lower triangular structure; that is, test whether $\phi_{12} = 0$.

14.7. Suppose that a three-dimensional VARMA process $\boldsymbol{Z}_t$ has Kronecker indices $K_1 = 3, K_2 = 1$, and $K_3 = 2$.

(a) Write the form of the coefficient matrices $\boldsymbol{\Phi}_j^{\#}$ and $\boldsymbol{\Theta}_j^{\#}$ in the echelon canonical ARMA model structure of equation (14.7.1) for this process.

(b) For this process $\{\boldsymbol{Z}_t\}$, describe the nature of the zero canonical correlations that occur in the canonical correlation analysis of the past vector $\boldsymbol{P}_t = (\boldsymbol{Z}_t', \boldsymbol{Z}_{t-1}', \ldots)'$ and various future vectors $\boldsymbol{F}_{t+1}^*$.

(c) Write the form of the minimal dimension echelon state-space model corresponding to the echelon canonical ARMA model for this process.

14.8. Verify that any VAR(p) model $\boldsymbol{Z}_t = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{Z}_{t-j} + \boldsymbol{a}_t$ can be expressed equivalently in the error correction form of equation (14.8.2) as $\boldsymbol{W}_t = \boldsymbol{C}\boldsymbol{Z}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_j^* \boldsymbol{W}_{t-j} + a_t$, where $\boldsymbol{W}_t = \boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}, \boldsymbol{\Phi}_j^* = -\sum_{i=j+1}^{p} \boldsymbol{\Phi}_i$, and $\boldsymbol{C} = -\boldsymbol{\Phi}(1) = -\left(\boldsymbol{I} - \sum_{j=1}^{p} \boldsymbol{\Phi}_j\right)$.

14.9. Express the model for the nonstationary bivariate process $\boldsymbol{Z}_t$ given in Exercise 14.5 in an error correction form, similar to equation (14.8.2), as $\boldsymbol{W}_t = \mathbf{C}\boldsymbol{Z}_{t-1} + \boldsymbol{a}_t - \boldsymbol{\Theta}a_{t-1}$, where $\boldsymbol{W}_t = (1 - B)\boldsymbol{Z}_t$. Determine the structure (and the ranks) of the matrices $\boldsymbol{C}$ and $\boldsymbol{\Theta}$ explicitly.

14.10. Consider analysis of the logarithms of monthly flour price indices from three U.S. cities. The raw (unlogged) data were given and analyzed by Tiao and Tsay (1989). To first investigate a VAR model for these data, with possible reduced-rank structure, the results of the partial canonical correlation analysis of Section 14.7.3, in terms of the (squared) partial canonical correlations $\hat{\rho}_i^2(j)$ between $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_{t-j}$ for lags $j = 1, \ldots, 6$, and the associated test statistic values computed using (14.7.8) are displayed in the following table:

j	Squared Correlations			$C(j, r)$ $r = 2$	$r = 1$	$r = 0$
1	0.747,	0.814,	0.938	129.93	288.89	551.67
2	0.003,	0.081,	0.274	0.29	7.97	36.91
3	0.001,	0.007,	0.035	0.07	0.69	3.76
4	0.000,	0.015,	0.047	0.03	1.29	5.31
5	0.017,	0.036,	0.073	1.36	4.24	10.22
6	0.000,	0.020,	0.077	0.00	1.51	7.49

In addition, values of $|\tilde{\mathbf{\Sigma}}_j|$ and of the AIC_j and HQ_j model selection criteria for the full-rank VAR(j) models are given as follows:

j (AR order)	1	2	3	4	5	6
$\lvert\tilde{\mathbf{\Sigma}}_j\rvert(\times 10^{-10})$	1.66213	1.12396	1.10523	1.06784	0.88963	0.81310
AIC_j	−22.336	−22.542	−22.369	−22.210	−22.195	−22.084
HQ_j	−22.240	−22.350	−22.079	−21.822	−21.707	−21.494

Interpret the preliminary model specification information above, and specify the structure (order and possible reduced ranks) of a VAR model that may seem appropriate for these data based on these results.

PART FOUR

DESIGN OF DISCRETE CONTROL SCHEMES

In earlier chapters we studied the modeling of discrete univariate time series and dynamic systems involving two or more time series. We saw how once adequate models have been developed, they can be used to generate forecasts of future observations, to characterize the transfer function of a dynamic system, and to represent the interrelationships among several time series of a multivariate dynamic system. Examples involving real-world applications have been used for illustration. However, the models and the methodology are of much wider importance than even these applications indicate. The ideas we have outlined are of importance in the analysis of a wide class of stochastic–dynamic systems occurring, for example, in economics, engineering, commerce, hydrology, meteorology, and in organizational studies.

It is obviously impossible to illustrate every application. Rather, it is hoped that the theory and examples of this book will help the reader to adapt the general methodology to their own particular problems. In doing this, the dynamic and stochastic models we have discussed will often act as *building blocks* that can be linked together to represent the particular system under study. The techniques of identification, estimation, and diagnostic checking, similar to those we have illustrated, should be useful to establish the model. Finally, recursive calculations and the ideas considered under the general heading of forecasting will have wider application in evaluating the adequacy and the usefulness of a model for a specific purpose once the model has been fitted.

We shall conclude this book by illustrating these possibilities in one further application—the design of feedback and feedforward control schemes. In working through Chapter 15, it is the task of bringing together the previously discussed ideas in a fresh application, quite as much as the detailed results, that we hope will be of value.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

15

ASPECTS OF PROCESS CONTROL

The term *process control* is used in different ways. Shewhart charts and other quality control charts are frequently employed in industries concerned with the manufacture of discrete ''parts'' in what is called *statistical process control* (SPC). By contrast, various forms of feedback and feedforward adjustment are used, particularly in the process and chemical industries, in what we call *engineering process control* (EPC). Because the adjustments made by engineering process control are usually computed and applied automatically, this type of control is sometimes called *automatic process control* (APC). However, the manner in which adjustments are applied is a matter of convenience, so we will not use that terminology here. The object of this chapter is to draw on the earlier discussions in this book to provide insight into the statistical aspects of these control methods and to appreciate better their relationships and objectives.

We first discuss *process monitoring* using, for example, Shewhart control charts and contrast this with techniques for *process adjustment*. In particular, a common adjustment problem is to maintain an output variable close to a target value in a dynamic system subject to disturbances by manipulation of an input variable, to obtain feedback control. *Feedback control* schemes use only the observed deviation of the output from target as a basis for adjustment of the input variable. We consider this problem first in a purely intuitive way and then relate this to some of the previously discussed stochastic and transfer function models to yield feedback control schemes producing minimum mean square error (MMSE) at the output. This leads to a discussion of discrete schemes, which are analogs of the proportional–integral (PI) schemes of engineering control, and we show how simple charts may be devised for *manually adjusting* processes with PI control.

It turns out that minimum mean square error control often requires excessively large adjustments of the input variable. ''Optimal'' constrained schemes are, therefore, introduced that require much smaller adjustments at the expense of only minor increases in

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

the output mean square error. These constrained schemes are generally not PI schemes, but in certain important cases it turns out that appropriately chosen PI schemes can often closely approximate their behavior. Particularly, in industries concerned with the manufacture of parts, there may be a fixed cost associated with adjusting the process and, in some cases, a monitoring cost associated with obtaining an observation. We therefore also discuss bounded adjustment schemes for feedback control that minimize overall cost in these circumstances.

In some instances, one or more *sources* of disturbance may be measured, and these measurements may be used to compensate potential deviations in the output. This type of adjustment action is called *feedforward control*, as compared to feedback control where only the observed deviation of output from target is used as a basis for adjustment. In certain instances it may also be desirable to use a combination of these two modes of control, and this is referred to as *feedforward–feedback control*. We therefore also present feedforward and feedforward–feedback types of control schemes for a general dynamic system that yield minimum mean square error at the output. Finally, we consider a general procedure for monitoring control schemes for possible changes in parameter values using Cuscore charts. More general discussion is given in the appendices and references.

15.1 PROCESS MONITORING AND PROCESS ADJUSTMENT

Process control is no less than an attempt to cancel out the effect of a fundamental physical law—the second law of thermodynamics, which implies that if left to itself, the entropy or disorganization of any system can never decrease and will usually increase. SPC and EPC are two complementary approaches to combat this law. SPC attempts to *remove* disturbances using *process monitoring*, while EPC attempts to *compensate* them using *process adjustment* (see also Box and Kramer, 1992).

15.1.1 Process Monitoring

The SPC strategy for stabilization of a process is to standardize procedures and raw materials and to use hypothesis-generating devices (such as graphs, check sheets, Pareto charts, cause–effect diagrams, etc.) to track down and eliminate causes of trouble (see, for example, Ishikawa, 1976). Since searching for assignable causes is tedious and expensive, it usually makes sense to wait until ''statistically significant'' deviations from the stable model occur before instituting this search. This is achieved by the use of process *monitoring charts* such as Shewhart charts, Cusum charts, and Roberts' EWMA charts. The philosophy is ''don't fix it when it ain't broke''—don't needlessly tamper with the process (see, for example, Deming, 1986).

Figure 15.1 shows an example of process monitoring using a Shewhart control chart. Condoms were tested by taking a sample of 50 items every 2 hours from routine production, inflating them to a very high fixed pressure, and noting the proportion that burst. Figure 15.1 shows data taken during the startup of a machine making these articles. Studies from similar machines had shown that a high-quality product was produced if the proportion failing this very severe test was $p = 0.20$.

The *reference* distribution indicated by the bars on the right of Figure 15.1(a) characterizes desired process behavior. It is a binomial distribution showing the probabilities of getting various proportions failing in random samples of $n = 50$ when p stays constant at a

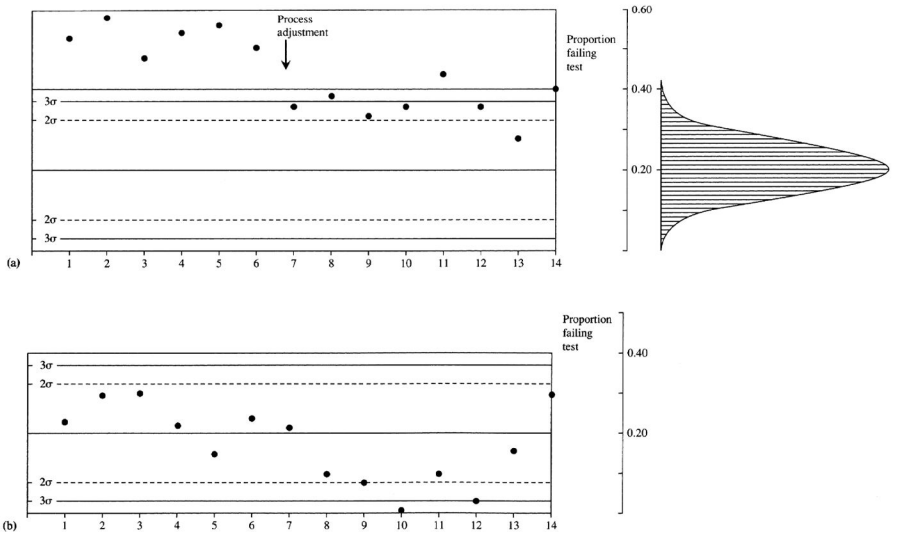


FIGURE 15.1 Shewhart charts for the proportion of condoms failing an inflation test. (a) Data taken before the process was brought to a state of control. (b) Data taken at a later stage of development.

value of 0.20. If the data behaved like a random sequence from this reference distribution, we should say the process appeared to be in a *state of control* and no action would be called for. By contrast, if the data did not have this appearance, showing outliers or suspicious patterns, we might have reason to suppose that something else was going on. In practice, the whole reference distribution would not usually be shown. Instead, upper and lower control limits and warning lines would be set. When, as in this case, a normal curve (shown as a continuous line) provides a close approximation to the reference distribution, these are usually set at $\pm 2\sigma$ and $\pm 3\sigma$ with $\sigma = \sqrt{p(1-p)/n}$, the standard deviation of the sample proportion from a binomial distribution. In this example, with $p = 0.20$ and $n = 50$, this gives $\sigma = 0.057$. Figure 15.1(a) shows that during the startup phase, the process was badly out of control, with the proportion of items failing the test initially as high as 50%. A process adjustment made after 12 hours of operation brought the proportion of defectives down to around 40%, but further changes were needed to get the process to a state of control at the desired level of $p = 0.20$. By a series of management actions, this was eventually achieved and Figure 15.1(b) shows the operation of the process at a later stage of development. Although for the most part the system now appears to be in the desired state of control, notice that the 10th point on the chart fell *below the lower* $\pm 3\sigma$ line. Subsequent investigation showed that the testing procedure was responsible for this aberrant point. A fault in the air line had developed and the condoms tested at about this time were inadvertently submitted to a much reduced air pressure, resulting in a falsely low value of the proportion defective. Corrective action was taken and the system was modified so that the testing machine would not function unless the air pressure was at the correct setting, ensuring that this particular fault could not occur again.

Monitoring procedures of this kind are obviously of great value. Following Shewhart (1931) and Deming (1986), we refer to the natural variation in the process when in state of control (binomial variation for a sample of $n = 50$ with $p = 0.20$ in this case) as due to *common causes*. The common cause system can only be changed by management action that alters the system. Thus, a new type of testing machine might be introduced for which the acceptable proportion of defects should be 10%. Common cause variation would then be binomial about the value $p = 0.10$.

The fault in the air line that was discovered by using the chart is called a *special*[1] *cause*. By suitable ''detective'' work, it is often possible for the plant operators to track down and eliminate special causes. The objectives of process monitoring are thus (1) to establish and continually confirm that the desired common cause system remains in operation and (2) to look for deviations unlikely to be due to chance that can lead to the tracking and elimination of assignable causes of trouble.

15.1.2 Process Adjustment

Although we must always make a dedicated endeavor to remove causes of variation such as unsatisfactory testing methods, differences in raw materials, differences in operators, and so on, some processes cannot be fully brought to a satisfactory state of stability in this way. Despite our best efforts, there remains a tendency for the process to wander off

[1] Also called an ''assignable'' cause. However, we are sometimes faced with a system that is demonstrably not in a state of control and yet no causative reason can be found. So we will stay with Deming in his less optimistic word ''special.''

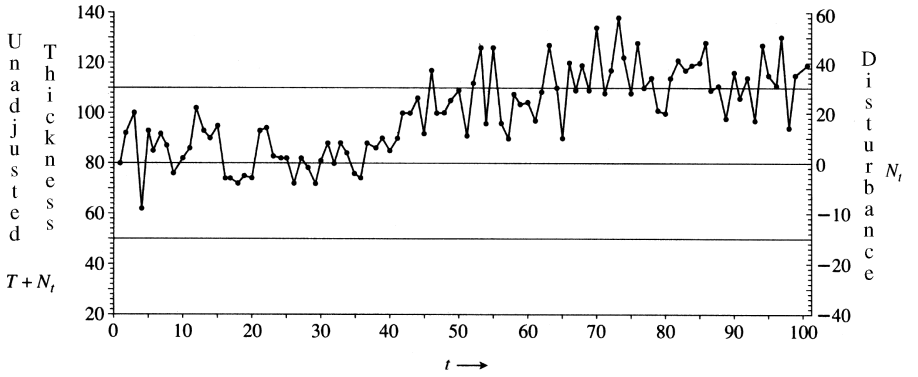


FIGURE 15.2 One hundred successive values of the thickness of a metallic film when no adjustment was applied.

target. This may be due to known but uncontrollable phenomena such as variations in ambient temperature, humidity, and feedstock quality, or due to causes currently unknown. In such circumstances, some system of process *adjustment* or *regulation* may be necessary in which manipulation of some additional variable is used to *compensate* for deviations in the quality characteristic.

To fix ideas, we first introduce a simple feedback adjustment scheme relying on a purely empirical argument and leave theoretical justification until later. Consider the measurements shown in Figure 15.2 of the thickness of a very thin metallic film taken at equally spaced units of time. The quality characteristic was badly out of control, but standard procedures failed to stabilize it (Box, 1991a). Suppose that the *disturbance* N_t is defined as the deviation of this quality characteristic from its target value T when *no adjustment is made*; that is, N_t is the underlying noise process. Suppose also that there is a manipulable variable—deposition rate X—which can be used conveniently to adjust the thickness, and that a unit change in X will produce g units of change in thickness and will take full effect in one time interval. If at time t, X was set equal to X_t, then at time $t + 1$ the deviation from target, $\varepsilon_{t+1} = Y_{t+1} - T$, after adjustment would be

$$\varepsilon_{t+1} = gX_t + N_{t+1} \tag{15.1.1}$$

Now suppose that *at time t* you can, in some way or the other, compute an estimate (forecast) $\hat{N}_t(1)$ of N_{t+1} and that this forecast has an error $e_t(1)$, so that

$$N_{t+1} = \hat{N}_t(1) + e_t(1) \tag{15.1.2}$$

Then using (15.1.1) and (15.1.2),

$$\varepsilon_{t+1} = gX_t + \hat{N}_t(1) + e_t(1) \tag{15.1.3}$$

If, in particular, X can be adjusted so that at time t,

$$X_t = -\frac{1}{g}\hat{N}_t(1) \tag{15.1.4}$$

then for the adjusted process

$$\varepsilon_{t+1} = e_t(1) \tag{15.1.5}$$

Thus, the deviation from target ε_{t+1} for the *adjusted* process would now be the error $e_t(1)$ in *forecasting* N_{t+1}, instead of the deviation N_{t+1} measured when the process is not adjusted.

If we used measurements of one or more of the known disturbing *input* factors (e.g., ambient temperature) to calculate the estimate $\hat{N}_t(1)$ of N_{t+1}, we would have an example of feedforward control. If the estimate $\hat{N}_t(1)$ of N_{t+1} directly or indirectly used only present and past values of the *output* disturbance $N_t, N_{t-1}, N_{t-2}, \ldots$, equation (15.1.4) would define a system of *feedback* control. A system of mixed *feedback–feedforward* control would employ both kinds of data. For simplicity, we will focus on the feedback case in the next three sections, and consider feedforward and mixed control in Section 15.5.

15.2 PROCESS ADJUSTMENT USING FEEDBACK CONTROL

Empirical Introduction. It might often be reasonable to use for the estimate $\hat{N}_t(\mathrm{l})$ in (15.1.4) some kind of weighted average of past values $N_t, N_{t-1}, N_{t-2}, \ldots$. In particular, an *exponentially* weighted moving average (EWMA) has intuitive appeal since recently occurring data are given most weight. Suppose, then, that $\hat{N}_t(\mathrm{l})$ is an EWMA,

$$\hat{N}_t(\mathrm{l}) = \lambda(N_t + \theta N_{t-1} + \theta^2 N_{t-2} + \cdots) \qquad 0 \le \theta \le 1 \tag{15.2.1}$$

where θ is the smoothing constant and $\lambda = 1 - \theta$,

We first consider the situation where, as has usually been the case in the process industries, adjustments are continually made as each observation comes to hand. Then using equation (15.1.4), the *adjustment* (*change* in deposition rate) made at time t would be given by

$$X_t - X_{t-l} = -\frac{1}{g}[\hat{N}_t(\mathrm{l}) - \hat{N}_{t-1}(1)] \tag{15.2.2}$$

Now with $e_{t-1}(1) = N_t - \hat{N}_{t-1}(1)$ the forecast error, the updating formula for an EWMA forecast can be written as

$$\hat{N}_t(1) - \hat{N}_{t-1}(1) = \lambda e_{t-1}(1) \tag{15.2.3}$$

Therefore, for any feedback scheme in which the compensatory variable X was set so as to cancel out an EWMA of the noise $\{N_t\}$, the required adjustment should be such that

$$X_t - X_{t-1} = -\frac{\lambda}{g} e_{t-1}(1) = -\frac{\lambda}{g}\varepsilon_t \tag{15.2.4}$$

For the metal deposition process, $g = 1.2$, $\lambda = 0.2$, and $T = 80$, so that the adjustment equation is

$$X_t - X_{t-1} = -\frac{1}{6}\varepsilon_t \tag{15.2.5}$$

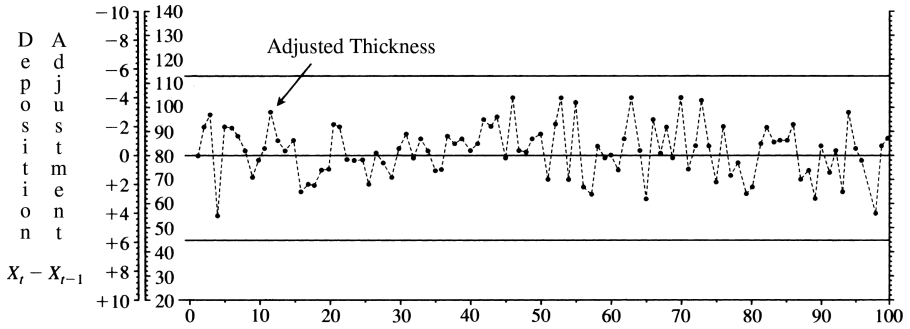


FIGURE 15.3 Manual adjustment chart for thickness that allows the operator to read off the appropriate change in deposition rate.

15.2.1 Feedback Adjustment Chart

This kind of adjustment is very easily applied, as is shown in Figure 15.3. This shows a manual feedback adjustment chart (Box and Jenkins, 1976) for the metallic thickness example given previously. To use it, the operator records the latest value of thickness and reads off on the adjustment scale the appropriate amount by which he or she should now increase or decrease the deposition rate. For example, the first recorded thickness of 80 is on target, so no action is called for. The second value of 92 is 12 units above the target, so $\varepsilon_2 = 12$, corresponding on the left-hand scale to a deposition rate adjustment of $X_2 - X_1 = -2$. Thus, the operator should now reduce the deposition rate by 2 units from its previous level.

Notice that the successive recorded thickness values shown on this chart are the readings that would actually occur *after adjustment*; the underlying disturbance is, of course, not seen on this chart. In this example, over the recorded period of observation, the chart produces a more than fivefold reduction in mean square error; the standard deviation of the adjusted thickness being now only about $\sigma_\varepsilon = 11$. Notice the following:

1. The chart is no more difficult to use than a Shewhart chart.
2. While the ''intuitive'' adjustment would be $-(1/g)\varepsilon_t = -(5/6)\varepsilon_t$ (corresponding to what Deming called ''tinkering''), the adjustment given by equation (15.2.4) is $-(\lambda/g)\varepsilon_t = -(1/6)\varepsilon_t$. Thus, it uses a discounted or ''damped'' estimate $\lambda\varepsilon_t$ of the deviation from target to determine the appropriate adjustment, where the discount factor λ is $1 - \theta$, with θ being the smoothing constant of the EWMA estimate of the noise.
3. By summing equation (15.2.4), we see that the *total* adjustment at time t is

$$X_t = k_0 + k_I \sum_{i=1}^{t} \varepsilon_i \tag{15.2.6}$$

with $k_0 = X_0$ and $k_I = -\lambda/g$. This adjustment procedure thus depends on the *cumulative* sum of the adjustment errors ε_i and the constant k_I determines how much the ''intuitive'' adjustment is discounted.

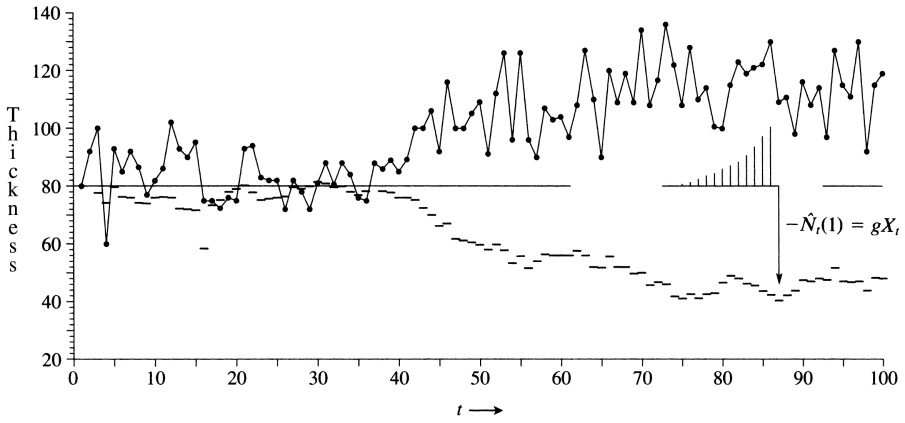


FIGURE 15.4 Dashes indicate the total adjustment $\hat{N}_t(1) = -gX_t$ achieved by the manual adjustment chart of Figure 15.3.

4. It follows from the previous argument that the adjustment is also equivalent to estimating at each time t the next value of the total unadjusted disturbance N_{t+1} by an *exponentially weighted average* of its past values and using this estimate to make an appropriate adjustment. This is illustrated for the metallic thickness example in Figure 15.4. Notice that in this preliminary discussion we have not explicitly assumed any particular time series model or claimed any particular optimal properties for the procedure. That the procedure can be discussed in such terms accounts, to some extent, for its remarkable robustness, which we discuss later.

In summary, then:

1. By *process monitoring* we mean the use of, for example, Shewhart charts and/or Cusum or Cuscore charts, as discussed by Box and Ramírez (1992). These are devices for continually checking a model that represents the desired ideal stable state of the system: for example, normal, independent, identically distributed (iid) variation about a fixed target T. The use of such charts can lead to the elimination of special causes pointed to by discrepant behavior. The judgment that behavior is sufficiently discrepant to merit attention is decided by a process analogous to *hypothesis testing*. Its properties are described in terms of probabilities (e.g., the probability of a point falling outside the 3σ limits of a Shewhart chart).
2. By *process adjustment* we mean the use of feedback and feedforward control or some combination of these to maintain the process as close as possible to some desired target value. Process adjustment employs a system of statistical *estimation* (forecasting) rather than of hypothesis testing, and its properties are described, for example, by output *mean square error*. Process monitoring and process adjustment are complementary rather than competitive corresponding to the complementary roles of hypothesis testing and estimation (see, for example, Box, 1980). We discuss this point more fully later in the chapter.

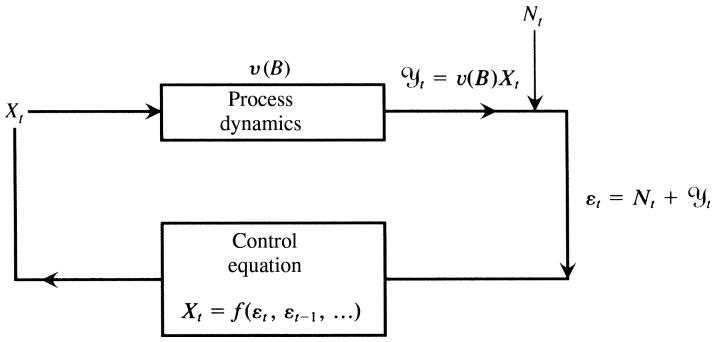


FIGURE 15.5 Feedback control loop.

15.2.2 Modeling the Feedback Loop

A somewhat more general system of feedback control is shown in Figure 15.5. The process is affected by a disturbance that in the absence of compensatory action would cause the output quality characteristic to deviate from target by an amount N_t. Thus, $\{N_t\}$ is a time series exemplifying what would happen at the output if no control were applied. In fact, a compensating variable X_t (deposition rate in our example) can be manipulated to cancel out this disturbance as far as possible. Changes in X will pass through the process and be acted on by its dynamics to produce at time t an amount or compensation $\mathcal{Y}_t$ at the output (again measured as a deviation from target). To the extent that this compensation $\mathcal{Y}_t$ fails to cancel out the disturbance N_t, there will be an error, or deviation from target $\varepsilon_t = Y_t - T$, equal to $\varepsilon_t = N_t + \mathcal{Y}_t$. The controller is some means (automatic or manual) that brings into effect the control equation $X_t = f(\varepsilon_t, \varepsilon_{t-1}, \ldots)$, which adjusts the output depending on present and past errors.

A device that has been used in the process industries for many years is the *three-term controller*. Controllers of this kind are usually operated automatically and employ continuous rather than discrete measurement and adjustment. If ε_t is the error at the output at time t, control action could, in particular, be made proportional to ε itself, to its integral with respect to time, or to its derivative with respect to time. A three-term controller uses a linear combination of these modes of control action, so that if X_t indicates the level of the manipulated variable at time t, the control equation is of the form

$$X_t = k_0 + k_D \frac{d\varepsilon_t}{dt} + k_P \varepsilon_t + k_I \int \varepsilon_t \, dt \tag{15.2.7}$$

where k_D, k_P, and k_I are constants.

Frequently, only one or two of these three modes of action are used. In particular, if only k_I is nonzero ($k_D = 0, k_P = 0$), we have *integral* control. If only k_I and k_P are nonzero ($k_D = 0$), we have *proportional--integral* (PI) control.

Notice that in the example we have just discussed, where the result of any adjustment fully takes effect at the output in one time interval, the dynamics of the process are represented by $\mathcal{Y}_t = gX_{t-1} = gBX_t$. The control equation $X_t = k_0 + k_I \sum_{i=1}^{t} \varepsilon_i$ in (15.2.6) is then the discrete analog of the control engineer's *integral* control.

In general, the discrete analog of (15.2.7) is

$$X_t = k_0 + k_D \nabla \varepsilon_t + k_P \varepsilon_t + k_I \sum_{i=1}^{t} \varepsilon_i$$

or in terms of the adjustment to be made,

$$\begin{aligned} x_t = X_t - X_{t-1} &= k_D \nabla^2 \varepsilon_t + k_P \nabla \varepsilon_t + k_I \varepsilon_t \\ &= c_1 \varepsilon_t + c_2 \varepsilon_{t-1} + c_3 \varepsilon_{t-2} \end{aligned}$$

where c_1, c_2, and c_3 are suitable constants. Not unexpectedly, control equations of this type are of considerable practical value.

15.2.3 Simple Models for Disturbances and Dynamics

So far we introduced a simple system of feedback control on purely empirical grounds. The *efficiency* of any such system will depend on the nature of the disturbance and the dynamics of the process. From a theoretical point of view, we can consider very general models for noise and dynamics and then proceed to find the control equation that ''optimizes'' the system in accordance with some criterion. However, the practical effectiveness of such models is usually determined by whether they, and the ''optimization'' criterion, make broad *scientific* sense and by their robustness to likely deviations from the ideal. We have already kept this in mind when discussing control procedures from a purely commonsense point of view and we will continue to do so when choosing models for the disturbance and for process dynamics.

Characterizing Appropriate Disturbance Models with a Variogram. A tool that helps to characterize process disturbances is the standardized variogram, which measures the variance of the difference between observations m steps apart compared to the variance of the difference of observations one step apart:

$$G_m = \frac{\text{var}[N_{t+m} - N_t]}{\text{var}[N_{t+1} - N_t]} \equiv \frac{V_m}{V_1} \tag{15.2.8}$$

For a stationary process, G_m is a simple function of the autocorrelation function. In fact, then, $G_m = (1 - \rho_m)/(1 - \rho_1)$. However, the variogram can be used to characterize nonstationary as well as stationary behavior. Figure 15.6 shows realizations of 100 observations initially on target generated by (a) a white noise process, (b) a first-order autoregressive process, and (c)–(f) IMA(0, 1, 1) processes with $\lambda = 0.1, 0.2, 0.3, 0.4$, respectively. The corresponding theoretical standardized variograms for these time series models are also shown.

In some imaginary world we might, once and for all, set the controls of a machine and give a set of instructions to an ever-alert and never-forgetting operator, and this would yield a perfectly stable process from that point on. In such a case the disturbance might be represented by a ''white noise'' series, and its corresponding standardized variogram G_m would be independent of m and equal to 1. But, in reality, left to themselves, machines involved in production are slowly losing adjustment and wearing out, and left to themselves, people tend, gradually, to forget instructions and miscommunicate. Thus, for an *uncontrolled* disturbance, some kind of monotonically increasing variogram would be expected. We cannot obtain such a variogram from a linear *stationary model*, for although

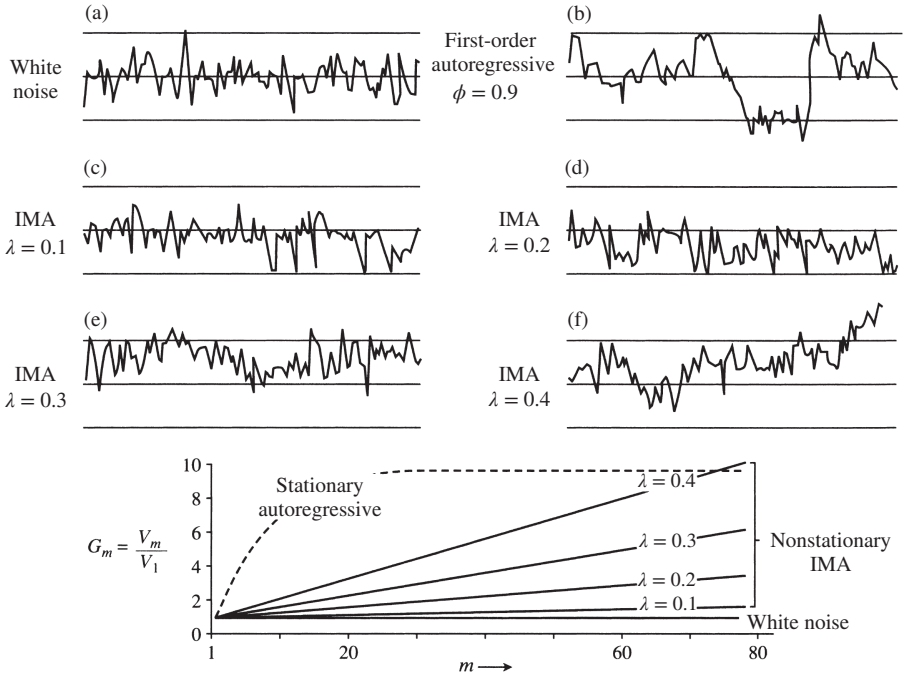


FIGURE 15.6 Realization of white noise, autoregressive, and IMA(0, 1, 1) time series with theoretical variograms.

G_m can initially increase with m, it will always approach an asymptote for such a process. That this can happen quite quickly, even when successive observations are highly positively correlated, is illustrated by the variogram shown in the figure for the first-order stationary autoregressive time series model $N_t = 0.9N_{t-1} + a_t$. In this example, even though successive deviations N_t from the target value have autocorrelation 0.9, G_m is already within 5% of its asymptotic value after only 20 lags. This implies that, for example, when generated by such a model, observations 100 steps apart differ little more than those 20 steps apart.

A model that can approximate the behavior of an uncontrolled system that *continuously* increases its entropy may be arrived at by thinking of the disturbance as containing two parts, a transitory part b_t and a nontransitory part z_t:

$$N_t = b_t + z_t \tag{15.2.9}$$

The transitory part b_t is associated only with the tth observation and is supposed independent of observations taken at every other time. Typical sources contributing to b_t are measurement and sampling errors. We represent this transitory part by random drawings from a distribution having mean zero and variance σ_b^2, that is, $\{b_t\}$ is a white noise process.

Sticky Innovation Model. The evolving nontransitory part z_t represents innovations that enter the system from time to time and *get stuck* there. These ''sticky'' innovations can arise from a multitude of causes, such as wear, corrosion, and human miscommunication. Thus, a car tire hits a sharp stone and *from that point onward* the tread is slightly damaged; a tiny crater caused by corrosion appears on the surface of a driving shaft and *remains*

there; certain details in the standard procedure for taking blood pressure in a hospital are forgotten and from that point on *permanently omitted or changed*. It is these nontransitory or sticky innovations that constitute the unwanted ''signal'' we wish to cancel out. Every system is subject to such influences. They continuously drive the increase in entropy if nothing is done to combat them. Such a sticky innovation model was suggested by Barnard (1959) and has a variogram that increases linearly with m. A special case of this model, which may also be used to approximate it, is the IMA(0, 1, 1) model:

$$N_t - N_{t-1} = a_t - \theta a_{t-1} \tag{15.2.10}$$

Recall, also, from Appendix A4.3 that if the nontransitory process Z_t is an IMA(0, 1, 1) process, then the disturbance process $N_t = b_t + z_t$ in (15.2.9) with the added white noise b_t, will again follow an IMA(0, 1, 1) model. Since for the IMA model (15.2.10) the EWMA of equation (15.2.1) with smoothing parameter θ provides a minimum mean square error (MMSE) forecast with forecast error $e_{t-1}(1) = a_t$, the corresponding discrete ''integral'' controller of (15.2.6) with $k_I = -\lambda/g$ produces MMSE control with $\varepsilon_t = a_t$. As we discuss later more formally, this is then a special case of the general MMSE linear feedback control scheme.

Dynamics. In discussion of the integral control scheme of equation (15.2.6), we assumed that any change made at the input of the system would have its full effect at the output in one time interval. The assumed dynamic equation for the response $\mathcal{Y}_t$ was, therefore,

$$\mathcal{Y}_t = gBX_{t+} \tag{15.2.11}$$

where we now denote the fixed level of the ''pulsed'' input X in the time interval from t until $t + 1$ by X_{t+}. A somewhat more general assumption is that the system can be described by the first-order difference equation

$$(1 + \xi\nabla)\mathcal{Y}_t = gBX_{t+} \tag{15.2.12}$$

(see, for example, (11.3.6)) or, equivalently,

$$(1 - \delta B)\mathcal{Y}_t = (1 - \delta)gBX_{t+} \qquad -1 < \delta < 1 \tag{15.2.13}$$

where $\xi = \delta/(1 - \delta)$ or, equivalently, $\delta = \xi/(1 + \xi)$. In that case at time $t + 1$ [cf. (15.1.1)], the deviation from target after adjustment is

$$\varepsilon_{t+1} = \mathcal{Y}_{t+1} + N_{t+1}$$

so that

$$\varepsilon_{t+1} = \frac{(1 - \delta)g}{1 - \delta B}X_{t+} + \hat{N}_t(1) + e_t(1)$$

where $\hat{N}_t(1)$ is some forecast of N_{t+1} made at time t with forecast error $e_t(1)$. Then, if we use the adjustment equation

$$X_{t+} - X_{t-1+} = x_t = -\frac{1 - \delta B}{(1 - \delta)g}[\hat{N}_t(1) - \hat{N}_{t-1}(1)]$$

the deviation ε_{t+1} from the target is equal to the forecast error $e_t(1)$. Thus, again we substitute the error in *forecasting* N_{t+1} for the deviation N_{t+1} itself. In particular, if $\hat{N}_t(1)$

is an EWMA forecast with smoothing parameter θ and if $\lambda = 1 - \theta$, then using (15.2.3)

$$x_t = (1 - B)X_{t+} = -\frac{\lambda(1 - \delta B)}{g(1 - \delta)}\varepsilon_t = -\frac{\lambda(1 - \delta) + \lambda\delta\nabla}{g(1 - \delta)}\varepsilon_t \tag{15.2.14}$$

Finally, if N_t can be represented by an IMA(0, 1, 1) process with parameter θ, then $\varepsilon_t = a_t$, and this adjustment will yield MMSE control. After summing (15.2.14), we obtain

$$X_t = k_0 + k_P\varepsilon_t + k_I \sum_{i=1}^{t} \varepsilon_i \tag{15.2.15}$$

in which

$$k_P = -\frac{\lambda}{g}\xi \quad \text{and} \quad k_I = -\frac{\lambda}{g}$$

The control equation (15.2.15) yields the discrete analog of continuous PI control mentioned earlier and will hereafter be referred to as (discrete) PI control.

Notice that despite their interesting ramifications, the adjustment equations corresponding to discrete integral control and PI control are extremely simple and intuitive. For discrete integral control

$$x_t = c_1\varepsilon_t \qquad (\text{with } c_1 = k_I)$$

and for PI control

$$x_t = c_1\varepsilon_t + c_2\varepsilon_{t-1} \quad (\text{with } c_1 = k_I + k_P \text{ and } c_2 = -k_P)$$

They, thus, make the adjustment x_t depend linearly on the last error and the last two errors, respectively.

15.2.4 General Minimum Mean Square Error Feedback Control Schemes

Arguing as earlier, it is not difficult to derive theoretical minimum mean square error feedback control schemes for the more general stochastic and linear dynamic models discussed in Chapters 4 and 11. Suppose the response to the series of adjustments in the manipulable input variable X_t is represented by the dynamic transfer function relation (11.2.3), written as

$$\mathcal{Y}_t = L_1^{-1}(B)L_2(B)B^{f+1}X_{t+}$$

where $L_1(B)$ and $L_2(B)$ are polynomials in B. This relation allows for f periods of pure dead time in the response. In addition, assume the noise or process disturbances $\{N_t\}$ may be represented by the linear stochastic ARIMA process defined by

$$N_t = \varphi^{-1}(B)\theta(B)a_t = \left(1 + \sum_{i=1}^{\infty} \psi_i B^i\right) a_t$$

where a_t is a white noise process. Then the error at the output, $\varepsilon_{t+f+1} = Y_{t+f+1} - T$, at time $t + f + 1$ can be written

$$\varepsilon_{t+f+1} = \mathcal{Y}_{t+f+1} + N_{t+f+1} = L_1^{-1}(B)L_2(B)X_{t+} + N_{t+f+1}$$

Clearly, the effect of the disturbance at time $t + f + 1$ would be canceled if it were possible to set $X_{t+} = -L_1(B)L_2^{-1}(B)N_{t+t+1}$. Since $f + 1$ is positive, this is not possible, but intuitively we can obtain minimum mean square error control by replacing N_{t+f+1} by its optimal forecast $\hat{N}_t(f + 1)$ at origin t. Now we can write $N_{t+f+1} = \hat{N}_t(f + 1) + e_t(f + 1)$, where $\hat{N}_t(f + 1)$ is the *forecast* at time t of N_{t+f+1} and $e_t(f + 1)$ is the error of the forecast for $f + 1$ steps ahead. The noise N_{t+f+1} is not known at time t, but its minimum mean square error forecast $\hat{N}_t(f + 1)$ can be deduced from the error sequence $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2} \ldots$, which is observed. Thus, it follows that the control equation $X_{t+} = -L_1(B)L_2^{-1}(B)\hat{N}_t(f + 1)$ will produce at time $t + f + 1$ a level at the output that will cancel out the forecast of the noise $f + 1$ periods ahead, and the error at the output will then be $\varepsilon_{t+f+1} = e_t(f + 1)$, the error of the forecast. To express the control equation in terms of the error sequence ε_t's, we can write

$$\varepsilon_t = e_{t-f-1}(f + 1) = a_t + \psi_1 a_{t-1} + \cdots + \psi_f a_{t-f} = L_4(B)a_t$$

and

$$\hat{N}_t(f + 1) = \psi_{f+1}a_t + \psi_{f+2}a_{t-1} + \cdots = L_3(B)a_t$$

where the operators $L_3(B)$ and $L_4(B)$ are determined from knowledge of the model $N_t = \varphi^{-1}(B)\theta(B)a = \psi(B)a_t$ for the noise process. Hence, we have

$$\hat{N}_t(f + 1) = L_3(B)L_4^{-1}(B)\varepsilon_t$$

Therefore, the MMSE feedback control equation is then

$$X_{t+} = -\frac{L_1(B)L_3(B)}{L_2(B)L_4(B)}\varepsilon_t \tag{15.2.16}$$

Alternatively, as is usually convenient, we can define the control action in terms of the *adjustment* $x_t = X_{t+} - X_{t-1+}$ to be made at time t as

$$x_t = -\frac{L_1(B)L_3(B)(1 - B)}{L_2(B)L_4(B)}\varepsilon_t$$

Example: Model with Dead Time. In particular, one more general dynamic model used above allows for ''dead time''—that is, pure delay in response to adjustment. To illustrate the application of equation (15.2.16), consider a first-order system affected by between f and $f + 1$ unit intervals of pure delay so that

$$(1 - \delta B)\mathcal{Y}_t = g(1 - \delta)[(1 - \upsilon) + \upsilon B]B^f X_{t-1} \tag{15.2.17}$$

Combining this with the IMA(0, 1, 1) disturbance model of equation (15.2.10), we can use the general derivation above to obtain the MMSE control scheme. In terms of the general model, we have $L_2(B)/L_1(B) = g(1 - \delta)(1 - \upsilon\nabla)/(1 - \delta B)$, and the IMA noise model yields $\hat{N}(f + 1) - \hat{N}_{t-1}(f + 1) = \lambda a_t$, so that $L_3(B) = \lambda/(1 - B)$, and also

$$e_{t-f-1}(f + 1) = [1 + \lambda(B + B^2 + \cdots + B^f)]a_t \equiv L_4(B)a_t$$

Hence, for the adjustment x_t, we have the relation

$$L_2(B)L_4(B)x_t = -L_1(B)L_3(B)(1 - B)\varepsilon_t$$

and we obtain the MMSE control equation as

$$(1 - \nu\nabla)[1 + \lambda(B + B^2 + \cdots + B^f)]x_t = -\frac{\lambda}{g(1-\delta)}(1 - \delta B)\varepsilon_t$$

Thus, this optimal control scheme is not PI but is of the form

$$x_t = c_1 x_{t-1} + c_3 x_{t-2} + \cdots + c_f x_{t-f-1} + c(\varepsilon_t - \delta\varepsilon_{t-1}) \tag{15.2.18}$$

where $c = -\lambda/[g(1-\delta)] = k_I + k_P$.

An interesting example by Fearn and Maris (1991) describes an MMSE scheme of this kind applied to the control of gluten addition to bread-making flour in a flour mill where the object was to maintain the protein content of the flour as close as possible to the target value. A careful process study showed that to an adequate approximation for this process $\delta = 0$, $\nu = 0$, $f = 1$, and $\lambda = 0.25$ ($\theta = 0.75$). The adjustment equation was thus

$$x_t = -0.25x_{t-1} - \frac{0.25}{g}\varepsilon_t \tag{15.2.19}$$

The scheme was tested extensively and the authors remarked that it worked well over a wide range of manufacturing conditions and was robust to moderate changes in the parameters.

The flour milling example does not yield a PI scheme. Notice, however, that the adjustment equation can be written $x_t = -(1 + \lambda B)^{-1}(\lambda/g)\varepsilon_t = -(1 - \lambda B + \lambda^2 B^2 - \cdots)(\lambda/g)\varepsilon_t$. For the rather small value $\lambda = 0.25$, if we truncate the expansion after the first-order term, we obtain the PI scheme $x_t = c_1\varepsilon_t + c_2\varepsilon_{t-1}$ with $c_1 = -\lambda/g$ and $c_2 = \lambda^2/g$. In practice, the behavior of this PI scheme will be almost identical to that of (15.2.19). More generally, we will find that PI schemes have an importance in addition to that conferred on them by their producing MMSE schemes for certain simple models. We therefore next consider how PI schemes can be put in effect using simple *feedback control charts*.

15.2.5 Manual Adjustment for Discrete Proportional–Integral Schemes

The equation for the adjustment $x_t = X_t - X_{t-1}$ for the discrete PI scheme (15.2.15) may also be written

$$x_t = -G(1 + P\nabla)\varepsilon_t \tag{15.2.20}$$

where

$$-G = k_I \quad \text{and} \quad P = \frac{k_P}{k_I} \tag{15.2.21}$$

or equivalently, $k_I = -G$ and $k_P = -PG$, and P is zero for pure integral control. In the special case where the stochastic and dynamic models are defined by (15.2.10) and (15.2.12), respectively, the PI control equation (15.2.15) yields MMSE when $G = \lambda/g$ and $P = \xi$.

Equation (15.2.20) shows how we can make a manual adjustment chart to put PI control into effect. We have already illustrated the use of such a chart for the metallic thickness example in Figure 15.3. For further illustration, we adapt an example discussed by Box et al. (1978). In a dyeing process, the quality characteristic of interest was the color index. Deviations ε_t from the desired target value of $T = 9$ were compensated by changing the dye addition rate X. For this example, the disturbance in the color index was approximated

by an IMA(0, 1, 1) model with $\lambda = 0.3$, and a change of 1 unit in the dye addition rate X eventually produced a change of 0.06 unit in the color index so that $g = 0.06$.

Suppose at first that ξ were zero so that the dynamic model was simply $\mathcal{Y}_t = gBX_{t+}$, implying that a change in the input X_t was fully effective at the output in one time interval. Then,

$$-G = k_I = -\frac{\lambda}{g} = -\frac{0.30}{0.06} = -5 \quad \text{and} \quad k_P = 0 \tag{15.2.22}$$

The MMSE integral feedback equation would be

$$X_t = k_0 - G\sum_{i=1}^{t} \varepsilon_t = k_0 - 5\sum_{i=1}^{t} \varepsilon_i \tag{15.2.23}$$

and at time t the corresponding *adjustment* would be

$$x_t = -G\varepsilon_t = -5\varepsilon_t \tag{15.2.24}$$

Appropriate action is read off the manual adjustment chart in Figure 15.7 with scales such that one unit deviation in the color index corresponds to $-G = -5$ units of adjustment of the dye addition rate. Action is taken after each observation by recording the value of the color index (indicated by a filled dot) and reading off on the left-hand scale the required adjustment to the dye addition rate. Thus, in the diagram at time 1:30 p.m., the color index was 9.14 calling for a reduction of -0.7 in the dye addition rate.

Now consider the case where, due perhaps to incomplete mixing of the dye, the process was subject to inertia, which was approximated by a first-order dynamic system as in (15.2.13) with $\delta = 0.2$ and consequently $\xi = \delta/(1 - \delta) = 0.25$. Thus, as before, $G = 0.3/0.06 = 5$ and now $P = \xi = 0.25$. Thus, the appropriate MMSE control equation (15.2.15) would call for proportional–integral action such that

$$X_t = k_0 - 1.25\varepsilon_t - 5\sum_{i=1}^{t} \varepsilon_i \tag{15.2.25}$$

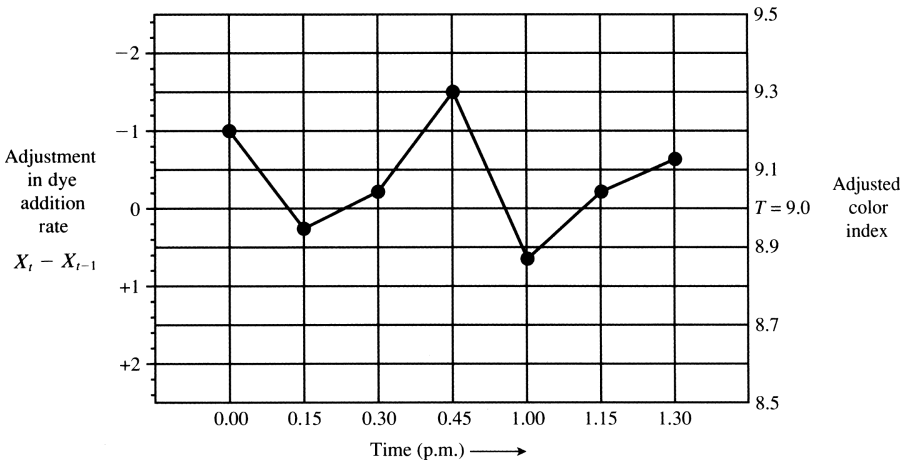


FIGURE 15.7 Manual adjustment chart for discrete integral control.

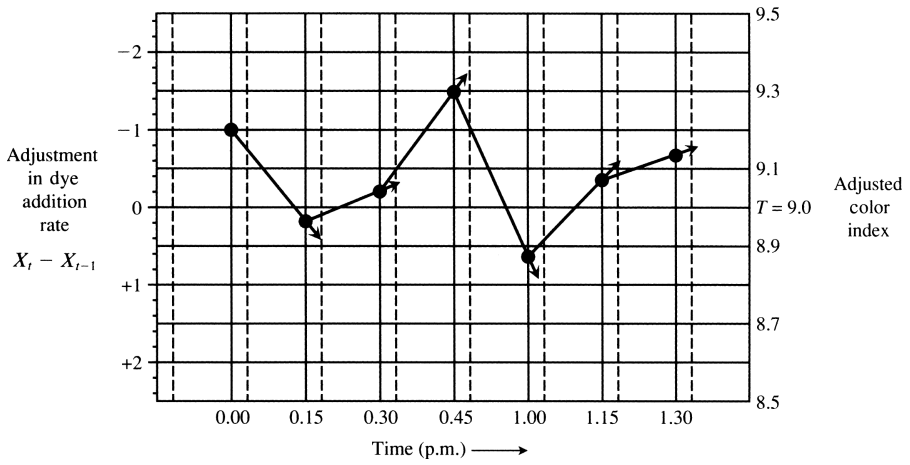


FIGURE 15.8 Manual adjustment chart putting into effect discrete integral plus proportional control.

The corresponding adjustment equation is

$$x_t = -5(1 + 0.25\nabla)\varepsilon_t \tag{15.2.26}$$

To put this into effect manually, the chart in Figure 15.8 may be employed with the vertical dashed lines placed at a fraction $P = k_P/k_I = 0.25$ within each sampling interval. At each step the operator extrapolates the line through the last two points to the next dashed line and reads off the appropriate adjustment. Thus, in this figure, the last two readings, at 1:15 and 1:30 p.m., were 9.06 and 9.14. The projected value of 9.16 requires reduction of the dye addition rate by −0.8 unit. No exactness is required. A line extrapolated by eye is good enough. As we later explore other uses of PI charts, we will sometimes use schemes in which P is negative. This calls for *interpolation* between the last two points rather than extrapolation.

Rounded Adjustment. The feedback schemes as so far discussed require that we take *some* action at every opportunity—in this example, every 15 minutes. In practice, usually little is lost if the ''rounded'' adjustment chart indicated in Figure 15.9 is used. Such a chart is easily constructed from the original chart by dividing the action scale into bands. The adjustment made when an observation falls within the band is that appropriate to the middle point of the band on an ordinary chart. Figure 15.9 shows a rounded chart in which possible action is limited to −2-, −1-, 0-, 1-, or 2-unit catalyst formulation changes. The increase in mean square error (usually small), which results from using the rounded scheme, is often outweighed by the convenience of working with a small number of standard adjustments. A convenient width for the rounded bands is about one standard deviation σ_ε or a little less. Justification for the use of such charts was provided by Box and Jenkins (1976, Section 13.1), where consideration is given to the effects of errors in the adjustment x_t. Note that the use of all these manual adjustment charts requires no calculation—they are simple and entirely graphical.

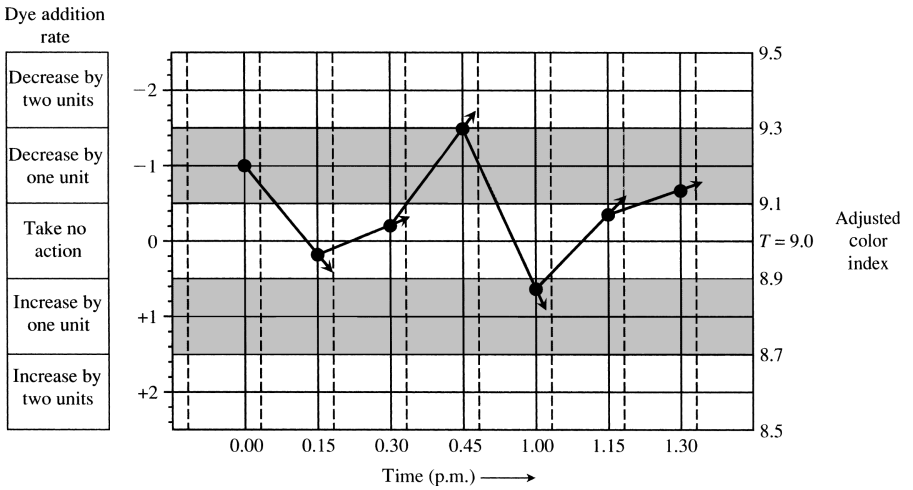


FIGURE 15.9 Rounded adjustment chart for proportional–integral control.

15.2.6 Complementary Roles of Monitoring and Adjustment

It is sometimes complained that feedback control can conceal the nature of a compensated disturbance that otherwise might be eliminated. However, when combined with appropriate monitoring, this need not happen. Adjustment schemes and monitoring schemes are complementary and should be used in consort. Figure 15.10 illustrates the point. This shows the behavior of a simulated feedback scheme in which the disturbance is an IMA(0, 1, 1) process with $\lambda = 0.2$ and the process dynamics are represented by a first-order system (15.2.13) with $\delta = 0.5$ and $g = 1.0$. The calculations were made assuming that the system is controlled by the PI controller,

$$-X_t = \text{constant} + 0.20\varepsilon_t + 0.20 \sum_{i=1}^{t} \varepsilon_i \tag{15.2.27}$$

which, for these stated parameter values, produces MMSE. Although this is not usually done, the control action X_t in Figure 15.10(b), as well as the deviation from target $\{\varepsilon_t\}$ in Figure 15.10(d), can be charted (or better still, displayed on the screen of a process computer). Assuming the dynamics known, the exact compensation $\mathcal{Y}_t$ shown in Figure 15.10(c) can also be computed and hence the original disturbance N_t of Figure 15.10(a) can be reconstructed.

Examination of these monitoring displays motivates a generalized concept of common and special causes. The disturbance and the dynamic system together define the *common cause* system, which is taken account of in the design of the controller. But management action could change the system and hence the appropriate form of control. For example, suppose it was discovered that in the operation of the system, the pattern of the feedback control action X_t shown in Figure 15.10(b) mirrored that of a particular impurity in the feedstock. If this correlation checked out as a causative relation, management might decide to change the control system either by removing the impurity from the feedstock before it reached the process, or if that were impossible or too expensive, by measuring it and compensating for it by appropriate feedforward control.

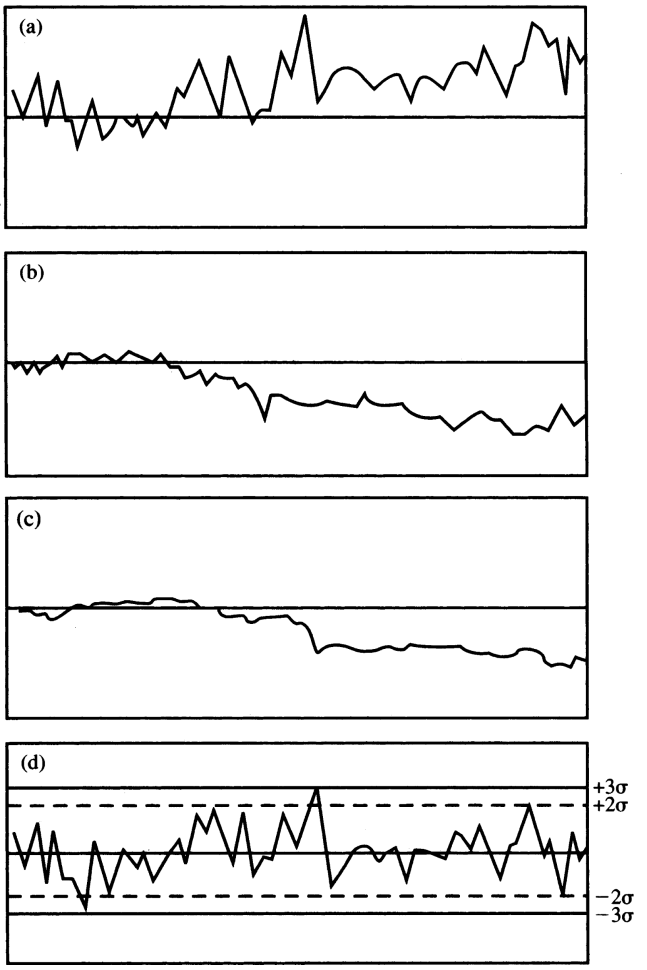


FIGURE 15.10 (a) Disturbance N_t, (b) feedback control action X_t, (c) compensation of the disturbance $\mathcal{Y}_t$, and (d) resulting deviation ε_t from the target value.

In addition, a *special cause* producing a temporary deviation from the underlying system model, induced perhaps by misoperation of the controller or a mistake by the operator, can be evidenced in the residual sequence $\{\varepsilon_t\}$ leading to remedial action. To illustrate this, we have added a deviation of size $3\sigma_a$ to the 30th value of the disturbance N_t in Figure 15.10(a). After the disturbance has been subjected to feedback control, this outlier is clearly visible in the record of the deviations ε_t from target plotted as a Shewhart chart in Figure 15.10(d). The control limits can be calculated directly from the models used to design the controller or from the record of the ε_t's during stable operation. Also, as noted later in Section 15.6, more specific checks may be applied to detect possible changes in the system parameters.

Assuming the models correct, in this particular example the residual ε_t's will be a white noise sequence. For control schemes that are not MMSE or that allow for dead time, however, the sequence $\{\varepsilon_t\}$ will, in general, be autocorrelated. One way to allow for this is to filter $\{\varepsilon_t\}$ suitably to produce a sequence that, given the assumed model, will be white noise. Appropriate checks may then be applied to that series.

15.3 EXCESSIVE ADJUSTMENT SOMETIMES REQUIRED BY MMSE CONTROL

One rationalization for the use of integral control and proportional–integral control is that for perhaps the simplest models for disturbance [equation (15.2.10)] and dynamics [equations (15.2.12) and (15.2.13)], which approximate reality, these forms of feedback adjustment can produce minimum mean square error.[2] Unfortunately, MMSE control sometimes requires unacceptably large manipulations of the compensating variable X_t. For illustration, consider again the situation where to an adequate approximation the disturbance model is the IMA(0, 1, 1) model of equation (15.2.10) with parameter θ and the dynamic model is the first-order difference equation (15.2.13) with parameters δ and g. Then, the MMSE feedback control adjustment scheme can be written (see (15.2.14)) as

$$x_t = -\frac{\lambda}{g}\frac{1-\delta B}{1-\delta}\varepsilon_t = -\frac{\lambda}{g(1-\delta)}(\varepsilon_t - \delta\varepsilon_{t-1}) \tag{15.3.1}$$

where $\lambda = 1 - \theta$ and $\varepsilon_t = a_t$. If δ is negligibly small, MMSE control will be obtained with $x_t = -(\lambda/g)\varepsilon_t$ and let us then write

$$\sigma_x^2 = \text{var}[x_t] = \frac{\lambda^2}{g^2}\sigma_a^2 = k \tag{15.3.2}$$

But then, when δ is *not* negligible,

$$\sigma_x^2 = k\left[\frac{1+\delta^2}{(1-\delta)^2}\right]$$

Thus, if δ were near its upper limit of unity, σ_x^2 could become very large. For example, with $\delta = 0.9$ (so that only 1/10 of the eventual change produced by a step input is experienced in the first interval), $\sigma_x^2 = 181k$. In fact, as δ approaches unity, the MMSE control action in equation (15.3.1) takes on more and more of an ''alternating'' character,[3] the adjustment made at time t reversing a substantial portion of the adjustment made at time $t - 1$. The reason for such alternating and variable adjustment can also be understood from the consideration that with $\delta = 0.9$, the constant $P = \xi = 9$ of the manual adjustment chart for MMSE control would call for *extrapolation* of the line joining ε_{t-1} and ε_t by *nine sampling intervals!* In practice, constrained schemes can be used that at the expense of rather small increases in MSE at the output require much less compensatory manipulation.

[2]This theoretical formulation, which results in a discrete PI controller yielding MMSE, is, however, not unique. For example, a PI controller giving MMSE can be obtained from the models $\mathcal{Y}_t = gBX_t$ and $N_t = (1 - \theta_1 B - \theta_2 B^2)a_t$, as well as the dynamics model (15.2.13) with IMA(0, 1, 1) noise model (15.2.10).

[3]A value of $\delta = 0.9$ corresponds to a time constant for the system of over nine sampling intervals. The occurrence of such a value would immediately raise the question as to whether the sampling interval being taken was too short; whether in fact the inertia of the process was so large that little would be lost by less frequent surveillance. Now (see Appendix A15.2) the question of the choice of sampling interval must depend on the nature of the noise that infects the system. Because the properties of the noise usually reflect system inertia as well, in many cases it would be concluded that the sampling interval should be increased.

15.3.1 Constrained Control

When the adjustments x_t form a stationary time series, such constrained control schemes can be obtained by finding an unconstrained minimum of the expression

$$\sigma_\varepsilon^2 + \alpha\sigma_x^2 \tag{15.3.3}$$

where α can be regarded as an undetermined multiplier that allocates the relative *quadratic costs* of variations of ε_t and x_t. Such a scheme will be called a constrained MMSE scheme or CMMSE scheme. In particular, we have seen that for an IMA(0, 1, 1) disturbance and first-order dynamics, the *unconstrained* MMSE scheme calls for an adjustment of

$$x_t = -\frac{\lambda}{g}(1 + \xi\nabla)\varepsilon_t = -\frac{\lambda(1 - \delta B)}{g(1 - \delta)}\varepsilon_t \tag{15.3.4}$$

It is shown in Appendix A15.1 (see equation (A15.1.27)) that the corresponding CMMSE is of the form

$$x_t = [k_1 + (1 - \lambda)k_0]x_{t-1} - (1 - \lambda)k_1 x_{t-2} - \frac{\lambda(1 - k_0)(1 - \delta B)}{g(1 - \delta)}\varepsilon_t \tag{15.3.5}$$

where k_0 and k_1 are fairly complicated functions of the parameters g, λ, δ, and α. A table for applying such control is also given in Appendix A15.1.

For illustration suppose that $\lambda = 0.6$, $\delta = 0.5$, and $g = 1$; then the optimal *unconstrained* MMSE scheme is

$$x_t = -1.2(1 - 0.5B)\varepsilon_t \tag{15.3.6}$$

with

$$\sigma_x^2 = (0.6)^2\left[\frac{1 + (0.5)^2}{(1 - 0.5)^2}\right]\sigma_a^2 = 1.80\sigma_a^2$$

from (15.3.2)–(15.3.2a), and $\sigma_\varepsilon^2 = \sigma_a^2$. Suppose that this amount of variation in the adjustment x_t produced difficulties in process operation and it was desired to reduce it so that σ_x^2 was about $0.50\sigma_a^2$. Use of Table A15.2 shows that this can be achieved with the scheme

$$x_t = 0.32x_{t-1} - 0.06x_{t-2} - (0.57 \times 1.2)(1 - 0.5B)\varepsilon_t \tag{15.3.7}$$

which reduces σ_x^2 to $0.47\sigma_a^2$ with $\sigma_\varepsilon^2 = 1.07\sigma_a^2$. Thus, an almost fourfold reduction in σ_x^2 is produced for an increase of only 7% in the output variance. Such optimal constrained schemes are extremely attractive since they often produce a very large reduction in σ_x^2 for only a small increase in σ_ε^2. See, for example, Whittle (1963), Tunnicliffe Wilson (1970a, 1970b), MacGregor (1972), Box and Jenkins (1976), Harris et al. (1982), Aström and Wittenmark (1984), Rivera et al. (1986), and Bergh and MacGregor (1987). Unfortunately, such schemes can become complicated.

In practice, however, exact ''optimality'' is to some extent an illusion because assumptions are never true. It turns out that a form of constrained control, which is almost as good as CMMSE control, can often be obtained using an *appropriately tuned* PI controller. Such a controller has the advantage that it is simple and, in particular, is easily adapted to manual control. The following example shows how suitably tuned PI controllers can do almost as

TABLE 15.1 Illustrative Results Comparing Different Control Schemes for Models (15.2.13) and (15.2.10), with g= 0.4, $\delta = 0.5$, $\lambda = 0.4$, and $\sigma_a^2 = 1$

		σ_ε^2	σ_x^2
(a) MMSE control	$-x_t = (1 + \nabla)\varepsilon_t$	1	5
(b) Optimal constrained control	$-x_t = -0.82x_{t-1} - 0.21x_{t-2} - 0.39\varepsilon_t + 0.19\varepsilon_{t-1}$	1.20	0.25
(c) Optimal constrained PI control	$-x_t = 0.52(1 - 0.25\nabla)\varepsilon_t$	1.20	0.25

well as optimal constrained schemes in producing great reductions in the variance σ_x^2 of the adjustment for only modest increases in the output variance σ_ε^2.

As an illustration, consider once again the situation where the process disturbance is represented by an IMA(0, 1, 1) process of (15.2.10) and the process dynamics by the first-order system (15.2.13), that is,

$$(1 - \delta B)\mathcal{Y}_t = (1 - \delta)gBX_{t+}$$

and suppose that $\lambda = 0.4$, $\sigma_a^2 = 1$, $g = 0.4$, and $\delta = 0.5$, so that $\xi = \delta/(1 - \delta) = 1$. Then minimum mean square error control is achieved by the PI scheme (a) shown in Table 15.1, yielding an output variance σ_ε^2 of 1.00 with $\sigma_x^2 = 5$. Using the optimal constrained control equation (b) in Table 15.1, it is possible to achieve a 20-fold reduction in σ_x^2 (to 0.25) at the expense of a 20% increase in σ_ε^2 to 1.20. But almost nothing is lost by, instead, using the much simpler optimal constrained PI controller (c) in Table 15.1 for which, to two-decimal accuracy, the same result is obtained. Notice that if we use a manual adjustment chart for the MMSE PI scheme (a), it would be necessary to extrapolate one whole time period ahead from the current time t. However, for the constrained PI control (c), we must *interpolate* a quarter of a period back from the current time t. This accounts for the much greater stability of the latter scheme. A fuller discussion of this topic can be found in Box and Luceño (1993).

15.4 MINIMUM COST CONTROL WITH FIXED COSTS OF ADJUSTMENT AND MONITORING

From the point of view of cost, we can summarize the discussion so far as follows. If we assume that the *only* control cost we need to consider is that of being off target and that this cost is proportional to the square of the deviation from target, unconstrained minimum mean square error control implies minimization of the total cost of the scheme. Suppose, however, that there is an additional quadratic loss associated with the size of the adjustment x_t, and that α is some measure of the *relative* cost of being off target and of making adjustments. Then, $\sigma_\varepsilon^2 + \alpha\sigma_x^2$ can be a measure of the overall cost of the scheme, and minimization of this quantity can produce a control scheme yielding minimum cost, and, as we have seen, suitably chosen PI schemes can often do almost as well. In either case, in practice, it is rarely easy to gauge α, in terms of relative costs. Instead, choice of a suitable scheme can be made by empirical judgment of what constitutes a satisfactory reduction of σ_x^2 in exchange for an acceptable increase in σ_ε^2. The same kinds of considerations apply to systems for which there are fixed adjustment and monitoring costs.

15.4.1 Bounded Adjustment Scheme for Fixed Adjustment Cost

Especially in the ''parts'' industries, situations occur where an adjustment often has immediate effect but entails a *fixed* cost incurred, for example, by stopping a machine or changing a tool.

Bounded Adjustment Charts. It was shown by Box and Jenkins (1963) that in the latter case, on the assumption of a quadratic off-target loss and an IMA disturbance, the minimum cost feedback control is *not* achieved by repeated adjustment after each observation. Instead, it requires that an adjustment be made only when an exponentially weighted average $\hat{\varepsilon}_t(1)$ of the deviations from target falls outside some fixed limits, $\pm L$, say. We call this *bounded adjustment*. The adjustment that should then be made is the one that will produce a change $-\hat{\varepsilon}_t(1)$ at the output. Such an adjustment can be put into effect manually using a ''bounded adjustment chart'' such as that discussed below, or automatically.

A bounded adjustment chart such as that shown in Figure 15.11 is superficially similar to that proposed for process monitoring by Roberts (1959). However, its purpose and design are different. The purpose is to decide when, and by how much, to *adjust* the process. The boundary lines are designed to minimize the overall cost, taking into account both the cost of making adjustments and the cost of being off target. Their purpose is *not* to discover statistically significant deviations from target. As the cost of adjustment approaches zero, the lines come closer together, converging on the target value when the cost of adjustment is zero and so yielding the ''repeated adjustment'' MMSE scheme.

Figure 15.11 shows an example of such a chart for the metallic thickness control problem that would be appropriate if there had been a fixed cost for changing the deposition rate X. As before, $\lambda = 0.2$, $g = 1.2$, and $\sigma_a = 11$. At time t, an open circle represents the deviation from target ε_t obtained after periodically changing the deposition rate X_t as required by the chart. A filled circle represents an appropriate exponentially weighted moving average

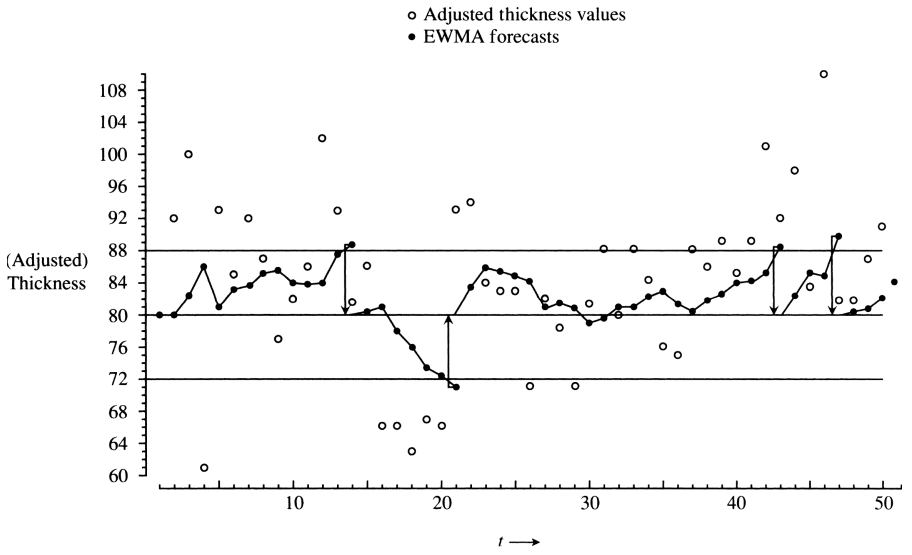


FIGURE 15.11 Bounded adjustment chart: the open circles are the thickness deviations ε_t (after adjustment), the filled circles are their EWMA forecasts $\hat{\varepsilon}_{t-1}(1)$ of these deviations.

forecast. This is conveniently updated using the formula

$$\hat{\varepsilon}_t(1) = \lambda\varepsilon_t + \theta\hat{\varepsilon}_{t-1}(1)$$

The particular chart shown has boundary lines at 80 ± 8, that is, at $T \pm 0.720\sigma_a$. We discuss the rationale for this choice below. To understand how the chart operates, suppose initially that the deposition rate is some value X_0. This will remain unchanged until time $t = 13$, when the forecasted value 88.7 (i.e., $\hat{\varepsilon}_t(1) = 8.7$) falls outside the upper limit and the chart signals that a change is needed in the deposition rate that will reduce the thickness by -8.7. An adjustment of

$$X_{13} - X_0 = -8.7/1.2$$

is now made in the deposition rate. Notice that such an adjustment does not upset the calculation of the next EWMA. For example, the forecasted thickness at time $t = 14$ is

$$(0.2 \times 81.3) + (0.8 \times 80.0) = 80.3$$

where 80 is the appropriate previous forecasted value *after the adjustment has been made to bring the process on target*.

15.4.2 Indirect Approach for Obtaining a Bounded Adjustment Scheme

Tables for calculating the positions of the appropriate limit lines for minimum cost schemes in terms of the *cost of being off target* and the *cost of adjustment* were provided by Box and Jenkins (1963), Box et al. (1974), and Box and Kramer (1992). However, as we said earlier, these costs are not always easy to assess, and it seems more practical to use these results to provide an envelope of minimum cost schemes and then to choose among them empirically by considering the increased standard deviation at the output obtained in exchange for a longer interval between making adjustments. This approach was illustrated by Box (1991b). Table 15.2 shows theoretical average adjustment intervals (AAIs) and percent increase in standard deviation (ISD) of the adjusted process for various values of λ and L/σ_a, where limit lines of the bounded adjustment scheme are at $T \pm L$.

For illustration, consider again the thickness adjustment example. Entering Table 15.2 with $\lambda = 0.2$ shows how much inflation in the error standard deviation would occur for a bounded scheme for various choice of L/σ_a. Thus, if L/σ_a were set equal to 0.5, a 2.6% increase in the standard deviation would occur, but on the average, adjustments would be needed only every 10 intervals. If L/σ_a were set equal to 1.0, a 9% increase in standard deviation would result, but the AAI would be 32. The scheme depicted in Figure 15.11 is a compromise in which L/σ_a was set equal to 0.72, which rough interpolation shows would give a 5% increase in the standard deviation with an AAI of about 20. To achieve this, L was set equal to $8 \approx 0.72 \times 11$. A Monte Carlo study using the 100 observations of metallic thickness graphed in Figure 15.2 shows an actual inflation of the standard deviation of 8.5% for this example with an AAI of 14. In view of the rather limited sample size, the agreement must be considered quite good.

Interpolation Chart. Any degree of technological sophistication can be used in applying these ideas: anything from transducers taking actions calculated by computers to operators taking actions based on a simple interpolation chart such as that shown in Figure 15.12, which used a pushpin and a piece of thread to indicate the appropriate *manual* adjustment.

TABLE 15.2 Average Adjustment Interval (AAI) and Percent Increase in Standard Deviation of Output (ISD) for Various Choice of L/σ_a Where the Limit Lines Are at $T \pm L$

λ	L/σ_a	AAI	Percent Increase in Standard Deviation ISD
0.1	0.5	32	2.4
	1.0	112	9
	1.5	243	18
	2.0	423	30
0.2	0.5	10	2.6
	1.0	32	9
	1.5	66	20
	2.0	112	32
0.3	0.5	5	2.6
	1.0	16	10
	1.5	32	20
	2.0	52	33
0.4	0.5	4	2.6
	1.0	10	10
	1.5	19	21
	2.0	32	34
0.5	0.5	3	2.5
	1.0	7	10
	1.5	13	21
	2.0	21	35

Source: Box (1991b).

In the situation depicted, a previous forecast made at time $t - 1$ was 86 and the observation, which has just been made at time t, is 66. Just before the current time t, therefore, the location of the pushpin on the current forecast scale would be at 86 with the thread hanging down from the pin. As soon as the actual value 66 became available, the thread would be pulled tightly to join the point 66 on the right-hand scale. The updated forecast of 82 would then be read off on the intermediate scale. This value lies within the boundaries, so that pushpin would be moved down to this new current forecast value with the thread hanging loose again until the next observation became available to produce a new updated forecast. As soon as an updated forecast fell outside either boundary, the appropriate adjustment in deposition rate to cancel out the forecasted deviation would be made, and the pushpin would then be *placed on the target value* ready for the next interpolation.

15.4.3 Inclusion of the Cost of Monitoring

It was shown by Box and Kramer (1992) how these results could be extended to the case where the cost of monitoring the process had also to be taken into account. They considered the possibility of further reducing cost by less frequent monitoring at an interval m instead of at a unit interval. They provided charts for obtaining minimum cost schemes given that in addition to σ_a and λ (estimated from plant data), three cost constants were known:

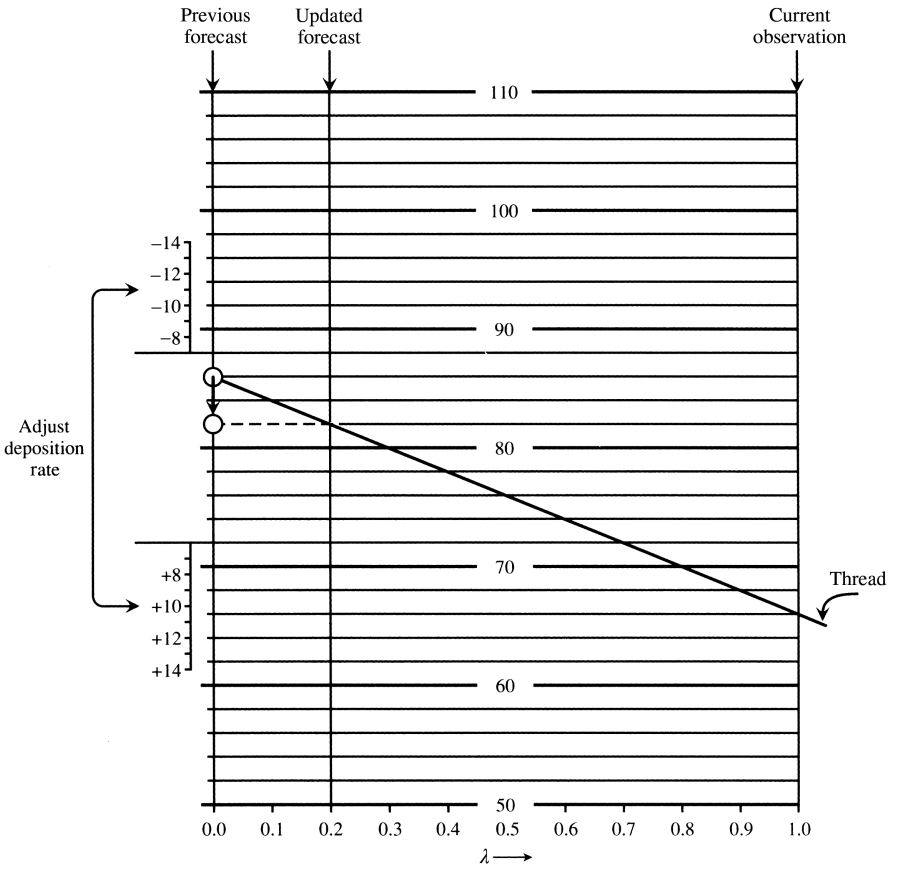


FIGURE 15.12 Interpolation chart to update the forecasted value of thickness and to indicate when and by how much the deposition rate should be adjusted.

(1) the (assumed quadratic) cost of being off target, (2) the fixed cost of making a change, and (3) the fixed monitoring cost of taking an observation. Given this information, the corresponding values of L/σ_a and of m yielding minimum cost could be read off their charts.

Again, these three individual costs may not be easy to determine, and Box and Luceño (1993) used their results to allow the choice of scheme to be based on empirical judgment. The charts shown in Figure 15.13 give the values of the AAI and the percent ISD with respect to σ_a corresponding to value of the nonstationarity measure $\lambda = 0.1(0.1)0.6, 0.8$, and 1.0, the standardized action limit $L/\sigma_a = 0.0(0.25)\ 2.5$, and the monitoring interval $m = 1, 2, 3, \ldots$. The charts cover small to moderate increases in the output standard deviation such as might be needed in practice. Thus, the larger values of m appear only with smaller values of λ.

For example, we saw earlier that by using a bounded adjustment chart with $L/\sigma_a = 0.72$ instead of a continuous scheme, the average adjustment interval could be increased to about 20 at the cost of an increase of 5% in the standard deviation. This is confirmed by the chart of Figure 15.13 for $\lambda = 0.2$, which also shows, for example, that if we monitor the process

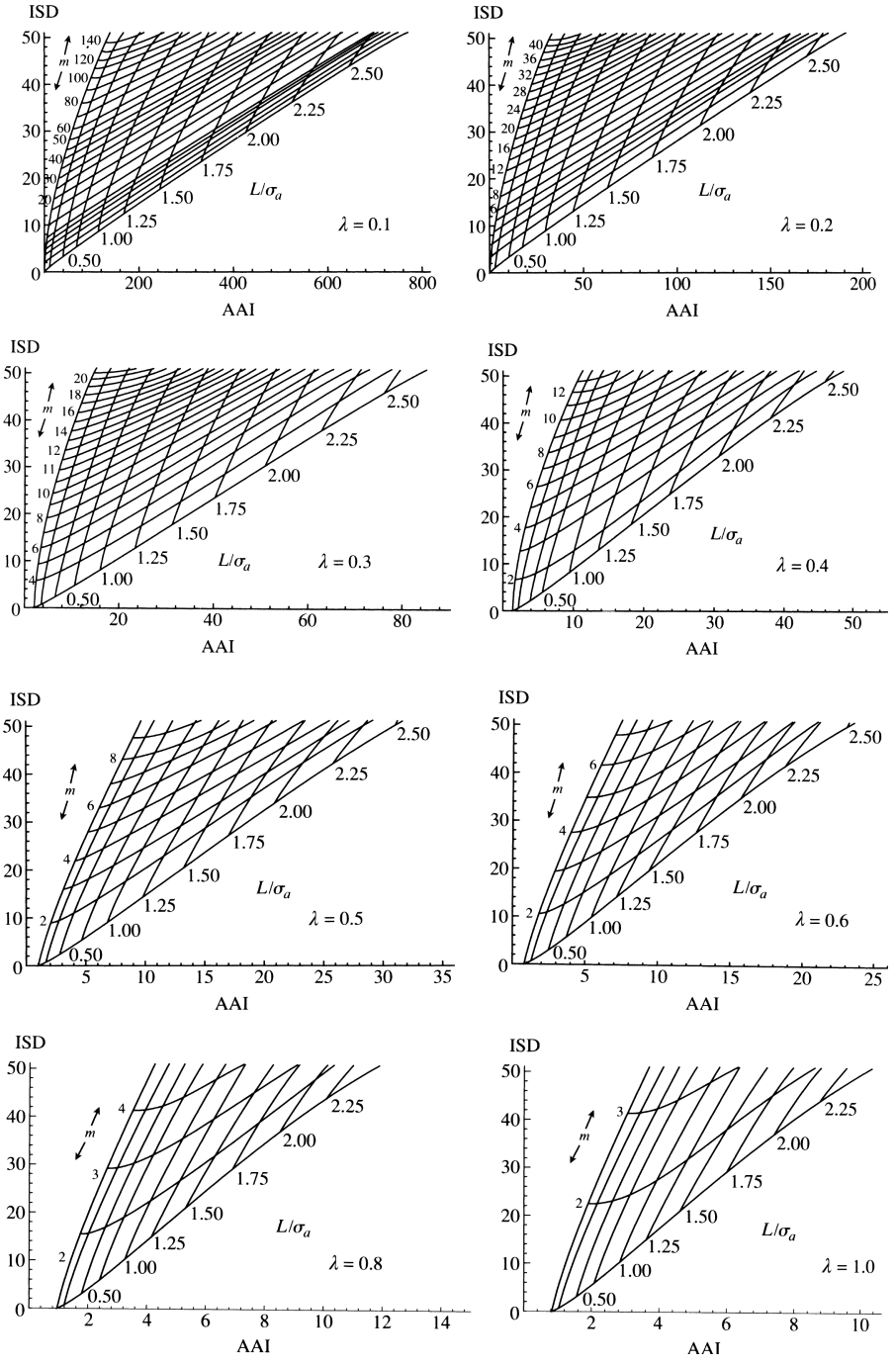


FIGURE 15.13 Charts for $\lambda = 0.1$, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, and 1.0 showing AAIs and ISDs obtained from various choices of L/σ_a and m.

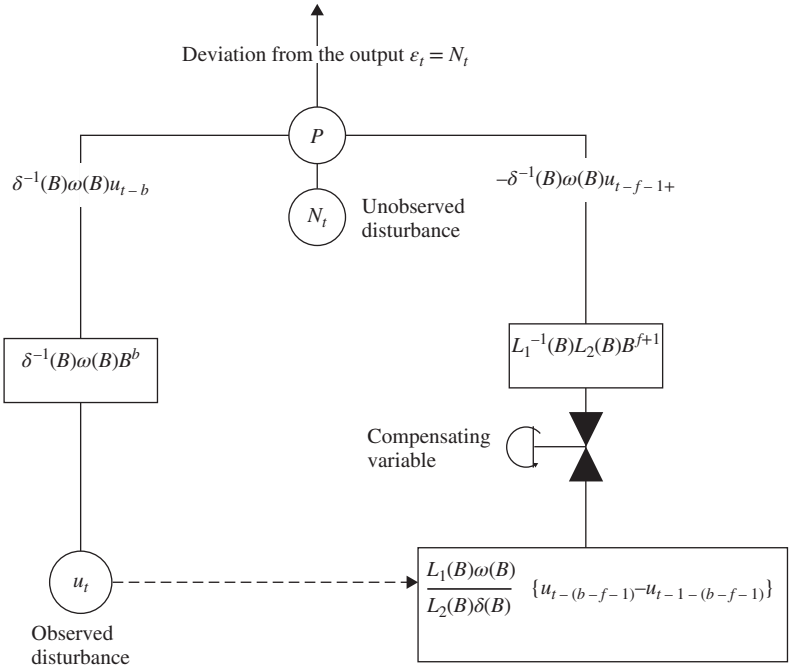


FIGURE 15.14 System at time t subject to an observed input disturbance u_t and unobserved disturbance N_t, with potential compensating variable X_t.

half as frequently ($m = 2$) and we again set $L/\sigma = 0.72$, we could obtain about the same average adjustment interval (20) but with an 8% increase in the standard deviation.

15.5 FEEDFORWARD CONTROL

We now consider the design of discrete *feedforward* control schemes that give minimum mean square error at the output. A situation arising in the manufacture of a polymer is illustrated in Figure 15.14. The viscosity Y_t of the product is known to vary in part due to fluctuations in the feed concentration u_t, which can be observed but not changed. The steam pressure X_t is a control variable that is measured, can be manipulated, and is potentially available to alter the viscosity by any desired amount and hence compensate potential deviations from target. The total effect in the output viscosity of all *other* sources of disturbance at time t is denoted by N_t.

15.5.1 Feedforward Control to Minimize Mean Square Error at the Output

We can suppose that Y_t, u_t, X_t, N_t are deviations from reference values, which are such that if the conditions $u = 0, X = 0, N = 0$ were continuously maintained, then the process would remain in an equilibrium state such that the output was exactly on the target value $Y = 0$.

The transfer function model, which connects the observed but uncontrollable input disturbance u_t (feed concentration) and the output Y_t (viscosity), is assumed to be

$$\mathcal{Y}_{1t} = \delta^{-1}(B)\omega(B)B^b u_t$$

Now, changes will be made in X at times $t, t-1, t-2, \ldots$ immediately after the observations $u_t, u_{t-1}, u_{t-2}, \ldots$ are taken. Hence, we obtain a ''pulsed'' input, and we denote the level of X in the interval t to $t+1$ by X_{t+}. For this pulsed input, it is assumed that the transfer function model, which connects the compensating variable X_t (steam pressure) and the output Y_t (viscosity), has the effect

$$\mathcal{Y}_{2t} = L_1^{-1}(B)L_2(B)B^{f+1}X_{t+}$$

where $L_1(B)$ and $L_2(B)$ are polynomials in B. Then, if no control is exerted (the potential compensating variable X_t is held fixed at $X_t = 0$), the total error or deviation from target value $T = 0, \varepsilon_t = Y_t - T$, in the output viscosity will be

$$\varepsilon_t = \delta^{-1}(B)\omega(B)u_{t-b} + N_t$$

Clearly, it ought to be possible to compensate the effect of the measured parts of the overall disturbance by manipulating X_t. Now at time t, and at the point P in Figure 15.14,

1. The total effect of the input disturbance (u) is

$$\delta^{-1}(B)\omega(B)u_{t-b}$$

2. The total effect of the compensation (X) is

$$L_1^{-1}(B)L_2(B)X_{t-f-1+}$$

and we assume that the effects of the input influences u and X on the output Y are additive. Then, the effect of the observed input disturbance u will be canceled if we set

$$L^{-1}(B)L_2(B)X_{t-f-1+} = -\delta^{-1}(B)\omega(B)u_{t-b}$$

Thus, the control action at time t should be such that

$$L_1^{-1}(B)L_2(B)X_{t+} = -\delta^{-1}(B)\omega(B)u_{t-(b-f-1)} \tag{15.5.1}$$

Case 1: $b \geq f + 1$. Now at time t, the values $u_{t+1}, u_{t+2} \ldots$ are unknown. The control action (15.5.1) is directly realizable, therefore, only if $(b - f - 1) \geq 0$, in which case the desired control action at time t is to set the manipulated variable X to the level

$$X_{t+} = -\frac{L_1(B)\omega(B)}{L_2(B)\delta(B)}u_{t-(b-f-1)} \tag{15.5.2}$$

Alternatively, it is often more convenient to define the control action in terms of the *change* $x_t = X_{t+} - X_{t-1+}$, which is to be made in the level of X immediately after the observation u_t has come to hand. This is

$$x_t = -\frac{L_1(B)\omega(B)}{L_2(B)\delta(B)}(u_{t-(b-f-1)} - u_{t-1-(b-f-1)}) \tag{15.5.3}$$

The situation is illustrated in Figure 15.14. The effect at P from the control action is $-\delta^{-1}(B)\omega(B)u_{t-b}$, and this exactly cancels the effect at P of the input disturbance. The component of the deviation from target due to u_t is (theoretically at least) exactly eliminated at the observation times, and only the component N_t due to the unobserved disturbance remains.

Case 2:* $(b - f - 1)$ *Negative. It can happen that $f + 1 > b$. This means that an observed input disturbance reaches the output before it is possible for compensating action to become effective. In this case the action in (15.5.2) is not realizable because at time t, when the action is to be taken, the relevant value $u_{t+(f+1-b)}$ of the input disturbance is not yet available. One would usually avoid this situation if one could (if some quicker acting compensating variable could be used instead of X), but sometimes such an alternative is not available.

Now with $u'_t = \delta^{-1}(B)\omega(B)u_t$ represented by the linear model (see, for example, Box et al. (1974))

$$u'_t = \left(1 + \sum_{i=1}^{\infty} \psi'_i B^i\right)\alpha_t$$

where α_t is a white noise process with mean zero and variance σ_α^2, then

$$u'_{t+f+1-b} = \hat{u}'_t(f + 1 - b) + e'_t(f + 1 - b)$$

In this expression

$$e'_t(f + 1 - b) = \alpha_{t+f+1-b} + \psi'_1\alpha_{t+f-b} + \cdots + \psi'_{f-b}\alpha_{t+1}$$

is the forecast error. Then, we can write the right-hand side of (15.5.2) in the form

$$-L_1(B)L_2^{-1}(B)\hat{u}'_t(f + 1 - b) - L_1(B)L_2^{-1}(B)e'_t(f + 1 - b)$$

Now, $e'_f(f + 1 - b)$ is a function of the uncorrelated random variates $\alpha_{t+h}(h \geq 1)$, which have not yet occurred at time t and which are uncorrelated with any variable known at time t (and the α_{t+h} are therefore not forecastable). It follows that the optimal (minimum mean square error) action is achieved by setting

$$X_{t+} = -\frac{L_1(B)}{L_2(B)}\hat{u}'_t(f + 1 - b) \tag{15.5.4}$$

that is, by making the *change* in the compensating variable at time t equal to

$$x_t = -\frac{L_1(B)}{L_2(B)}\{\hat{u}'_t(f + 1 - b) - \hat{u}'_{t-1}(f + 1 - b)\} \tag{15.5.5}$$

This results in an additional component in the deviation ε_t from the target, which now becomes

$$\varepsilon_t = N_t + e'_{t-f-1}(f + 1 - b)$$

If the model for the input disturbance is $\varphi_u(B)u_t = \theta_u(B)\alpha_t$, then the model for $u'_t = \delta^{-1}(B)\omega(B)u_t$ can be written

$$\varphi'_u(B)u'_t = \theta'_u(B)\alpha_t$$

with

$$\varphi'_u(B) = \varphi_u(B)\delta(B) \text{ and } \theta'_u(B) = \theta_u(B)\omega(B)$$

The needed forecasts $\hat{u}'_t(f + 1 - b)$, obtained as in Chapter 5, can then be written conveniently in terms of previous u's and α's obtainable from the u series itself.

15.5.2 An Example: Control of the Specific Gravity of an Intermediate Product

In the manufacture of an intermediate product, used for the production of a synthetic resin, the specific gravity Y_t of the product had to be maintained as close as possible to the value 1.260. This was actually achieved by a mixed scheme of feedforward and feedback control. We consider the complete scheme later and discuss here only the feedforward part. The process has rather slow dynamics, and also the disturbance is known to change slowly, so that observations and adjustments are made at 2-hour intervals. The uncontrolled input disturbance that is fed forward is the feed concentration u_t, which is measured as deviations from an origin of 30 g/L. The relation between specific gravity and feed concentration over the range of normal operation has the effect

$$\mathcal{Y}_{1t} = 0.0016u_t$$

where the effect $\mathcal{Y}_{1t}$ is measured from the target value 1.260.

This relation contains ''no dynamics'' because the feed concentration can only be measured at the inlet to the reactor, so that in our general notation $\delta(B) = 1, \omega(B) = 0.0016, b = 0$. Control is achieved by varying pressure, which is referred to a convenient origin of 25 psi. The transfer function model relating specific gravity and pressure X_t was estimated as having the effect

$$(1 - 0.7B)\mathcal{Y}_{2t} = 0.0024X_{t-1+}$$

so that $L_1(B) = (1 - 0.7B), L_2(B) = 0.0024, f = 0$. So far as could be ascertained, the effects of pressure and feed concentration were approximately additive in the region of normal operation. Therefore, the control equation (15.5.4) is used, since $b - f - 1$ is negative, and yields

$$X_{t+} = -\frac{(1 - 0.7B)0.0016}{0.0024}\hat{u}_t(1) \tag{15.5.6}$$

for, in this particular example, $u'_t = 0.0016u_t$ and hence $\hat{u}'_t(1) = 0.0016\hat{u}_t(1)$. Study of the feed concentration showed that it could be represented by the linear stochastic model of order (0, 1, 1),

$$\nabla u_t = (1 - \theta_u B)\alpha_t$$

with $\theta_u = 0.5$. For such a process,

$$\hat{u}_t(1) = (1 - \theta_u)u_t + \theta_u\hat{u}_{t-1}(1)$$

that is, $(1 - \theta_u B)\hat{u}_1(1) = (1 - \theta_u)u_t$ or

$$\hat{u}_t(1) = \frac{1 - \theta_u}{1 - \theta_u B}u_t$$

TABLE 15.3 Calculation of Adjustments for Feedforward Control Scheme (15.5.7)

	Concentration			Pressure	
t	$u_t + 30$	u_t	X_{t+}	$X_{t+} + 25$	x_t
0	31.6	1.6	−0.63	24.4	
1	31.1	1.1	−0.31	24.7	0.3
2	34.4	4.4	−1.36	23.6	−1.1
3	32.0	2.0	−0.32	24.7	1.1
4	28.2	−1.8	0.90	25.9	1.2

Thus, the control equation (15.5.6) can be written finally as

$$X_{t+} = -\frac{(1 - 0.7B)0.0016(0.5)}{0.0024(1 - 0.5B)}u_t$$

or

$$X_{t+} = 0.5X_{t-1+} - 0.333(u_t - 0.7u_{t-1}) \tag{15.5.7}$$

Table 15.3 shows the calculation of the first few of a series of settings of the pressure required to compensate the variations in feed concentration, given the starting conditions for time $t = 0$ of $u_0 = 1.6, X_{0+} = -0.63$. Once the calculation has been started off, it is sometimes more convenient to work directly with the changes x_t to be made at time t using

$$x_t = 0.5x_{t-1} - 0.333(\nabla u_t - 0.7\nabla u_{t-1}) \tag{15.5.8}$$

Figure 15.15a shows a section of the feed concentration. Figure 15.15b shows the output after applying feedforward control. Figure 15.15c shows the specific gravity if no control had been applied. These values Y_t are, of course, not directly available but may be obtained in general from the values Y_t', which actually occurred using

$$Y_t = Y_t' + \hat{u}'_{t-f-1}(f + 1 - b)$$

For this example then

$$Y_t = Y_t' + \frac{0.0008}{1 - 0.5B}u_{t-1}$$

that is,

$$Y_t = 0.5Y_{t-1} + Y_t' - 0.5Y'_{t-1} + 0.0008u_{t-1}$$

As a result of feedforward control, the root mean square error deviation of the output from the target value over the sample record shown is 0.003. Over the same period, the root mean square error of the uncorrected series would have been 0.008. The improvement is marked and extremely worthwhile. However, it appears that other unidentified sources of disturbance exist in the process, as evidenced by the drift away from target. This kind of tendency is frequently met in pure feedforward control schemes, but may be compensated by the addition of feedback control, as discussed in Section 15.2. We will briefly indicate the details of the combined scheme later in Section 15.5.4.

Control action is effected in whatever manner is most suited to the situation. If changes are made infrequently, and if the control equation is fairly simple as in the above example,

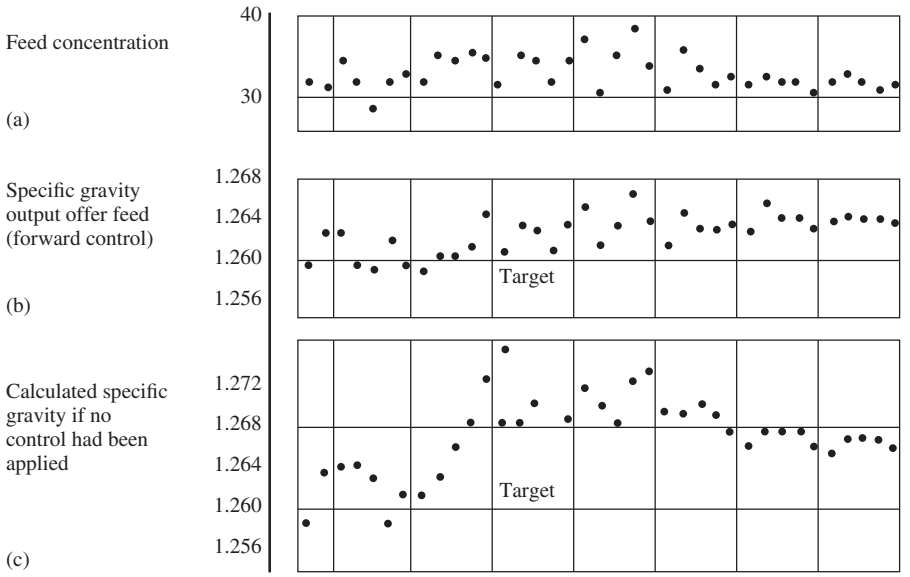


FIGURE 15.15 (a) Feed concentration, (b) Specific gravity after feedforward control, (c) Specific gravity if no control had been applied.

the theory we have outlined may be used to obtain optimal control *manually*. It is then convenient to use some form of control chart or nomogram that can be easily understood by the process operator, similar to charts illustrated in Section 15.2 regarding feedback control.

15.5.3 Feedforward Control with Multiple Inputs

No difficulty arises in principle when the effects of several additive input disturbances $u_1, u_2, \ldots, u_m$ are to be compensated by changes in X using feedforward control. Suppose the combined effect at the output of all the input disturbances is given by

$$\mathcal{Y}_t = \sum_{j=1}^{m} \delta_j^{-1}(B)\omega_j(B)B^{b_j}u_{j,t} = \sum_{j=1}^{m} B^{b_j}u'_{j,t}$$

where $u'_{j,t} = \delta_j^{-1}(B)\omega_j(B)u_{j,t}$, and, as before, the transfer function model for the compensating variable contributes the effect

$$\mathcal{Y}_{2t} = L_1^{-1}(B)L_2(B)B^{f+1}X_{t+}$$

Then, proceeding precisely as before, the required control action is to change X at time t by an amount

$$x_t = -L_1(B)L_2^{-1}(B)\sum_{j=1}^{m}[u'_{j,t+f+1-b_j} - u'_{j,t+f-b_j}] \tag{15.5.9}$$

where

$$[u'_{j,t+f+1-b_j} - u'_{j,t+f-b_j}]$$
$$= \begin{cases} u'_{j,t+f+1-b_j} - u'_{j,t+f-b_j} & f+1-b_j \le 0 \\ \hat{u}'_{j,t}(f+1-b_j) - \hat{u}'_{j,t-1}(f+1-b_j) & f+1-b_j > 0 \end{cases} \tag{15.5.10}$$

If, as before, N_t is an unmeasurable disturbance, then the error or deviation from target at the output from this control action in the compensating variable X_t+ will be

$$\varepsilon_t = N_t + \sum_{j=1}^{m} e'_{j,t-f-1}(f+1-b_j) \tag{15.5.11}$$

where $e'_{j,t-f-1}(f+1-b_j) = 0$ if $f+1-b_j \le 0$, and is the forecast error corresponding to the jth input variable $u_{j,t}$ if $f+1-b_j > 0$.

On the one hand, feedforward control allows us to take prompt action to cancel the effect of input disturbance variables, and if $f+1-b_j \le 0$, to anticipate completely such disturbances, at least in theory. On the other hand, to use this type of control we must be able to measure the disturbing variables and possess complete knowledge—or at least a good estimate—of the relationship between each input disturbance variable and the output. In practice, we could never measure *all* of the disturbances that affected the system. The remaining disturbances, which we have denoted by N_t and which are not affected by feedforward control, could of course increase the variance at the output or cause the process to wander off target, as in fact occurred in the example discussed in Section 15.5.2. Clearly, we can prevent this from happening by using the deviations ε_t themselves to indicate an appropriate adjustment, that is, by using feedback control as discussed in earlier sections of this chapter. In fact, a combined feedforward–feedback control scheme can be used, which provides for the elimination of identifiable input disturbances by feedforward control and for the reduction of the remaining disturbance by feedback control.

15.5.4 Feedforward–Feedback Control

A combined feedforward–feedback control scheme provides for the elimination of identifiable input disturbances by feedforward control and for the reduction of the remaining disturbance by feedback control. We briefly discuss a combined feedforward–feedback scheme in which m identifiable input disturbances $u_1, u_2, \ldots, u_m$ are fed forward. The combined effects on the output of all the input disturbances and of the compensating input variable X_t are assumed to be additive of the same form as given previously in Section 15.5.3. It is assumed also that N'_t is a further unidentified disturbance and that the *augmented noise* N_t is made up of N'_t plus that part of the feedforward disturbance that cannot be predicted at time t. Thus, using (15.5.11),

$$N_t = N'_t + \sum_{j=1}^{m} e'_{j,t-f-1}(f+1-b_j)$$

where $e'_{j,t-f-1}(f+1-b_j) = 0$ if $f+1-b_j \le 0$, and includes any further contributions from errors in forecasting the identifiable inputs. It is assumed that N_t can be represented by a linear stochastic process so that, in the notation of Section 15.2.4, it follows that the

relationship between the forecasts of this noise process and the forecast errors may be written as

$$\frac{L_3(B)(1-B)}{L_4(B)}\varepsilon_t = \hat{N}_t(f+1) - \hat{N}_{t-1}(f+1)$$

where $\varepsilon_t = e_{t-f-1}(f+1) = N_t - \hat{N}_{t-f-1}(f+1)$.

Arguing as in (15.2.16) and (15.5.9), the optimal control action for the compensating input variable X_t to minimize the mean square error at the output is

$$x_t = -\frac{L_1(B)}{L_2(B)}\left\{\sum_{i=1}^{m}[u'_{j,t+f+1-b_j} - u'_{j,t+f-b_j}] + \frac{L_3(B)(1-B)}{L_4(B)}\varepsilon_t\right\} \tag{15.5.12}$$

where the $[u'_{j,t+f+1-b_j} - u'_{j,t+f-b_j}]$ are as given in equation (15.5.10). The first term in the control equation (15.5.12) is the same as in (15.5.9) and compensates for changes in the feedforward input variables. The second term in (15.5.12) corresponds exactly to (15.2.16) and compensates for that part N'_t of the augmented noise, which can be predicted at time t.

An Example of Feedforward–Feedback Control. We illustrate by discussing further the example used in Section 15.5.2, where it was desired to control specific gravity as close as possible to a target value 1.260. Study of the deviations from target occurring *after feedforward control* showed that they could be represented by the IMA(0, 1, 1) process

$$\nabla N_t = (1 - 0.5B)a_t$$

where a_t is a white noise process. Thus,

$$\frac{L_3(B)(1-B)}{L_4(B)}a_t = \hat{N}_t(1) - \hat{N}_{t-1}(1) = 0.5a_t$$

and $\varepsilon_t = e_{t-1}(1) = a_t$. As in Section 15.5.2, the remaining parameters are

$$\delta^{-1}(B)\omega(B) = 0.0016 \quad b = 0$$
$$L_2^{-1}(B)L_1(B) = \frac{1-0.7B}{0.0024} \quad f = 0$$

and

$$\hat{u}_t(1) - \hat{u}_{t-1}(1) = \frac{0.5}{1-0.5B}(u_t - u_{t-1})$$

Using (15.5.12), the minimum mean square error adjustment incorporating feedforward and feedback control is

$$x_t = -\frac{1-0.7B}{0.0024}\left[\frac{(0.0016)(0.5)}{1-0.5B}(u_t - u_{t-1}) + 0.5\varepsilon_t\right] \tag{15.5.13}$$

that is,

$$x_t = 0.5x_{t-1} - 0.333(1-0.7B)(u_t - u_{t-1}) - 208(1-0.7B)(1-0.5B)\varepsilon_t$$

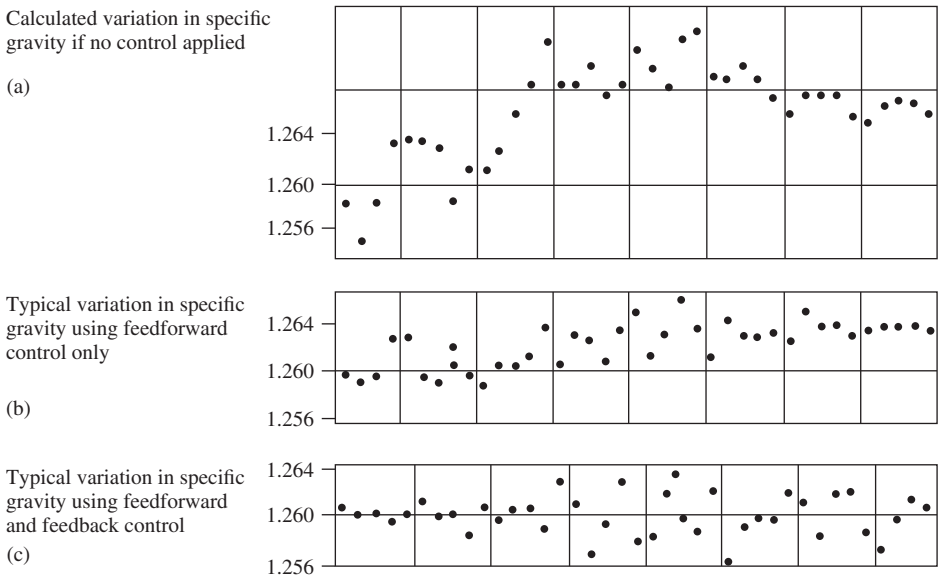


FIGURE 15.16 Typical variation in specific gravity with (a) no control, (b) feedforward control only, and (c) feedforward with feedback control.

or

$$x_t = 0.5x_{t-1} - 0.333u_t + 0.566u_{t-1} - 0.233u_{t-2} - 208\epsilon_t + 250\epsilon_{t-1} - 73\epsilon_{t-2} \quad (15.5.14)$$

Figure 15.16 shows the section of record previously given in Figure 15.15, when only feedforward control was employed, and the corresponding calculated variation that would have occurred if no control had been applied. This is now compared with a record from a scheme using both feedforward and feedback control. The introduction of feedback control resulted in a further substantial reduction in mean square error and corrected the tendency to drift from the target, which was experienced with the feedforward scheme.

Note that with a feedback scheme, the correction employs a forecast having lead time $f + 1$, whereas with a feedforward scheme the forecast has lead time $f + 1 - b$ and no forecasting is involved if $f + 1 - b$ is zero or negative. Thus, feedforward control gains in the immediacy of possible adjustment whenever b is greater than zero. The example we have quoted is an exception in that $b = 0$, and consequently no advantage of immediacy is gained, in this case, by feedforward control. It might be true in this case that equally good control could have been obtained by a feedback scheme alone. In practice, possibly because of error transmission problems, the mixed scheme did rather better than the pure feedback system.

15.5.5 Advantages and Disadvantages of Feedforward and Feedback Control

With feedback control, it is the total disturbance, as evidenced by the error at the output, that actuates compensation. Therefore, it is not necessary to be able to identify and measure the sources of disturbance. All that is needed is that we *characterize* the disturbance N_t at the output by an appropriate stochastic model (and as we have seen in earlier sections,

an IMA(0, 1, 1) model would often provide adequate approximation to the noise model). Because we are not relying on ''dead reckoning,'' unexpected disturbances and moderate errors in identifying and estimating the system's characteristics will normally result only in greater variation about the target value and not (as may occur with feedforward control) in a consistent drift away from the target value. On the other hand, especially if the delay $f + 1$ is large, the errors about the target (since they are then the errors of a remote forecast) may be large, although they have zero mean. Clearly, if identifiable sources of input disturbance can be partially or wholly eliminated by feedforward control, then this should be done. Then, only the unidentifiable error has to be dealt with by feedback control.

In summary, although we can design a feedback control scheme that is optimal, in the sense that it is the best possible feedback scheme, it will not usually be as good as a combined feedforward–feedback scheme in which sources of error that can be are eliminated before the feedback loop.

15.5.6 Remarks on Fitting Transfer Function–Noise Models Using Operating Data

It is desirable that the parameters of a control system be estimated from data collected under as nearly as possible the conditions that will apply when the control scheme is in actual operation. The calculated control action, using estimates so obtained, properly takes account of noise in the system, which will be characterized as if it entered at the point provided for in the model. This being so, it is desirable to proceed iteratively in the development of a control scheme. Using technical knowledge of the process, together with whatever can be learned from past operating data, preliminary transfer function and noise models are postulated and used to design a pilot control scheme. The operation of this pilot scheme can then be used to supply further data, which may be analyzed to give improved estimates of the transfer function and noise models, and then used to plan an improved control scheme.

For example, consider a feedforward–feedback scheme with a single feedforward input, as in Section 15.5.1, and the case with $b - f - 1$ nonnegative. Then for any inputs u_t and X_{t+}, the output deviation from target is given by

$$\varepsilon_t = \delta^{-1}(B)\omega(B)u_{t-b} + L_1^{-1}(B)L_2(B)X_{t-f-1+} + N_t \tag{15.5.15}$$

and it is assumed that the noise N_t may be described by an ARIMA(p, d, q) model. It is supposed that time series data are available for ε_t, u_t, and X_{t+} during a sufficiently long period of actual plant operation. Often, although not necessarily, this would be a period during which some preliminary pilot control scheme was being operated. Then for specified orders of transfer function operators and noise model, the methods of Sections 12.3 and 12.4 may be used directly to construct the sums of squares and likelihood function and to obtain estimates of the model parameters in the standard way through nonlinear estimation using numerical iterative calculation.

Consider now a pure feedback system that may be represented in the transfer function–noise model form

$$\varepsilon_t = v(B)X_{t+} + N_t \tag{15.5.16}$$

$$X_{t+} = c(B)\varepsilon_t\{+d_t\} \tag{15.5.17}$$

with

$$v(B) = L_1^{-1}(B)L_2(B)B^{f+1}$$

where $c(B)$ is the known operator of the controller, not necessarily optimal, and d_t is either an additional unintended error or an added ''dither'' signal that has been deliberately introduced. The curly brackets in (15.5.17) emphasize that the added term may or may not be present. In either case, estimation of the unknown transfer function and noise model parameters can be performed, as described in Chapter 12.

However, difficulties in estimation of the model under feedback conditions can arise when the added term d_t is not present. To better understand the nature of issues involved in fitting of the model, we can substitute (15.5.17) in (15.5.16) to obtain

$$[1 - v(B)c(B)]\varepsilon_t = \psi(B)a_t\{+v(B)d_t\} \tag{15.5.18}$$

First consider the case where d_t is *zero*. Because, from (15.5.17), X_{t+} is then a deterministic function of the ε_t's, the model (which appears in (15.5.16) to be of the transfer function form) is seen in (15.5.18) to be equivalent to an ARIMA model whose coefficients are functions of the known parameters of $c(B)$ and of the unknown dynamic and stochastic noise parameters of the model. It is then apparent that, with d_t absent, estimation difficulties can arise, as all dynamic and stochastic noise model forms $v_0(B)$ and $\psi_0(B)$, which are such that

$$\psi_0^{-1}(B)[1 - v_0(B)c(B)] = \psi^{-1}(B)[1 - v(B)c(B)] \tag{15.5.19}$$

will fit equally well in theory. In particular, it can be shown (Box and MacGregor, 1976) that as the pilot feedback controller used during the generation of the data approaches near optimality, near singularities occur in the sum-of-squares surface used for estimation of model parameters. The individual parameters may then be estimated only very imprecisely or will be nonestimable in the limit. In these circumstances, however, accurate estimates of those functions of the parameters that are the constants of the feedback control equation may be obtainable. Thus, while data collected under feedback conditions may be inadequate for estimating the *individual* dynamic and stochastic noise parameters of the system, it may nevertheless be used for updating the estimates of the constants of a control equation whose mathematical form is assumed known.

The situation can be much improved by the deliberate introduction during data generation of a random signal d_t as in (15.5.17). To achieve this, the action $c(B)\varepsilon_t$ is first computed according to the control equation and then d_t is added on. The added signal can, for example, be a random normal variate or a random binary variable and should have mean zero and variance small enough so as not to unduly upset the process. We see from (15.5.18) that with d_t present, the estimation procedure based on fitting model (15.5.16) now involves a genuine transfer function model form in which ε_t depends on the random input d_t as well as on the random shocks a_t. Thus, with d_t present, the fitting procedure tacitly employs not only information arising from the autocorrelations of the ε_t's but also additional information associated with the cross-correlations of the ε_t's and the d_t's.

In many examples, data from a pilot scheme are used to re-estimate parameters with the model form *already identified* from open-loop (no feedback control loop) data and from previous knowledge of the system. Considerable caution and care is needed in using closed-loop data in the model identification/specification process itself. In the first place, if d_t is absent, it is apparent from (15.5.16) that cross-correlation of the ''output'' ε_t and

the ''input'' X_{t+} with or without prewhitening will tell us (what we already know) about $c(B)$ and not, as might appear if (15.5.16) were treated as defining an open-loop system, about $v(B)$. Furthermore, since the autocorrelations of the ε_t will be the same for all model forms satisfying (15.5.19), unique identification is not possible if nothing is known about the form of either $\psi(B)$ or $v(B)$. On the other hand, if either $\psi(B)$ or $v(B)$ is known, the autocorrelation function can be used for the identification of the other. With d_t present, the form of (15.5.18) is that of a genuine transfer function–noise model considered in Chapter 12 and corresponding methods may be used for identification.

15.6 MONITORING VALUES OF PARAMETERS OF FORECASTING AND FEEDBACK ADJUSTMENT SCHEMES

Earlier we mentioned the complementary roles of process adjustment and process monitoring. This symbiosis is further illustrated if we again consider the need to monitor the adjustment scheme itself. It has often been proposed that the series of residual deviations from the target from such schemes (and similarly the errors from forecasting schemes) should be studied and that a Shewhart chart or more generally a cumulative sum or other monitoring chart should be run on the residual errors to warn of changes. The cumulative sum is, of course, appropriate to look for small changes in mean level, but often other kinds of discrepancies may be feared. A general theory of sequential directional monitoring based on a cumulative Fisher score statistic (Cuscore) was proposed by Box and Ramírez (1992) (see also Bagshaw and Johnson, 1977).

Suppose that a model can be written in the form of deviations e_t that depend on an unknown parameter θ as

$$e_t = e_t(\theta) \tag{15.6.1}$$

and *that if the correct value of the parameter* $\theta = \theta_0$ is employed in the model, $\{e_t\} = \{a_t\}$ is a sequence of Normal iid random variables. Then, the cumulative score statistic appropriate to detect a departure from the value θ_0 may be written

$$Q_t = \sum_{i=1}^{t} e_i r_i \tag{15.6.2}$$

where $r_t = -(de_t/d\theta)|_{\theta=\theta_0}$ may be called the detector signal.

For example, suppose that we wished to detect a shift in a mean from a value θ_0 for the simple model $y_t = \theta + e_t$. We can write

$$e_t = e_t(\theta) = y_t - \theta \quad a_t = y_t - \theta_0 \tag{15.6.3}$$

Then, in this example, the detector signal is $r_t = 1$ and $Q_t = \sum_{i=1}^{t} e_i$, the well-known *cumulative sum* statistic.

In general, for some value of θ close to θ_0, since e_t may be approximated by $e_t = a_t - (\theta - \theta_0)r_t$, the cumulative product in (15.6.2) will contain a part

$$-(\theta - \theta_0) \sum_{i=1}^{t} r_i^2 \tag{15.6.4}$$

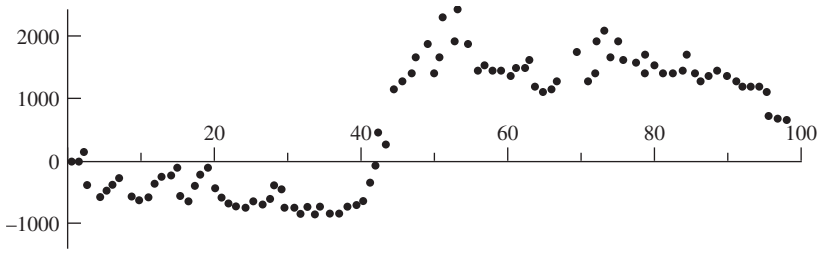


FIGURE 15.17 Cuscore monitoring for detecting a change in the parameter θ used in conjunction with the adjustment chart of Figure 15.3.

which systematically increases in magnitude with sample size t when θ differs from θ_0. For illustration, consider the possibility that in the feedback control scheme for metallic thickness of Section 15.2.1, the value of λ (estimated as 0.2) may have changed during the period $t = 1$ to $t = 100$. For this example,

$$e_t = e_t(\theta) = \frac{1 - B}{1 - \theta B} N_t \tag{15.6.5}$$

Thus,

$$r_t = -\frac{1 - B}{(1 - \theta B)^2} N_{t-1} = -\frac{e_{t-1}}{1 - \theta B} = -\frac{\hat{e}_{t-1}(1)}{\lambda} \tag{15.6.6}$$

where $\hat{e}_{t-1}(1) = \lambda(1 - \theta B)^{-1} e_{t-1}$ is an EWMA of past e_t's. The cumulative score (Cuscore) statistic for detecting this departure is, therefore,

$$Q_t = -\frac{1}{\lambda} \sum_{i=1}^{t} e_i \hat{e}_{i-1}(1) \tag{15.6.7}$$

where the detector signal $\hat{e}_{t-1}(1)$ is, in this case, the EWMA of past values of the *residuals*. These residuals are the deviations from the target plotted on the feedback adjustment chart of Figure 15.3. The criterion agrees with the commonsense idea that if the model is true, then $e_t = a_t$ and e_t is not predictable from previous values. The Cuscore chart shown in Figure 15.17 suggests that a change in parameter may have occurred at about $t = 40$. However, we see from the original data of Figure 15.2 that this is very close to the point at which the level of the original series appears to have changed, and further data and analysis would be needed to confirm this finding.

The important point is that this example shows the *partnership* of two types of control (adjustment and monitoring) and the corresponding two types of statistical inference (estimation and criticism). A further development is to feed back the filtered Cuscore statistic to ''self-tune'' the control equation, but we do not pursue this further here.

APPENDIX A15.1 FEEDBACK CONTROL SCHEMES WHERE THE ADJUSTMENT VARIANCE IS RESTRICTED

Consider now the feedback control situation where the models for the noise and system dynamics are again given by (15.2.10) and (15.2.13), so that $\varepsilon_t = \mathcal{Y}_t + N_t$ with

$$(1 - B)N_t = (1 - \theta B)a_t \text{ and } (1 - \delta B)\mathcal{Y}_t = (1 - \delta)g X_{t-1+}$$

but some restriction of the input variance var$[x_t]$ is necessary, where $x_t = (1 - B)X_t$. The unrestricted optimal scheme has the property that the errors in the output $\varepsilon_1, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$ are the uncorrelated random variables $a_t, a_{t-1}, a_{t-2}, \ldots$ and the variance of the output σ_ε^2 has the minimum possible value σ_a^2. With the restricted schemes, the variance σ_ε^2 will necessarily be greater than σ_a^2, and the errors $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$ at the output will be correlated.

We will pose our problem as follows: Given that σ_t^2 be allowed to increase to some value $\sigma_\varepsilon^2 = (1 + c)\sigma_a^2$, where c is a positive constant, we want to find the control scheme that produces the minimum value for $\sigma_x^2 = \text{var}[x_t]$. Equivalently, the problem is to find an (unconstrained) minimum of the expression $\sigma_\varepsilon^2 + \alpha\sigma_x^2$, where α is some specified multiplier that allocates the relative costs of variations in ε_t and x_t.

A15.1.1 Derivation of Optimal Adjustment

Let the optimal adjustment, *expressed in terms of the* a_t's, be

$$x_t = -\frac{1}{g}L(B)a_t \tag{A15.1.1}$$

where

$$L(B) = l_0 + l_1 B + l_2 B^2 + \cdots$$

Then, we see that the error ε_t at the output is given by

$$\begin{aligned}\varepsilon_t &= \frac{(1-\delta)g}{1-\delta B}X_{t-1+} + N_t \\ &= -\frac{1-\delta}{1-\delta B}(1-B)^{-1}L(B)a_{t-1} + (1-B)^{-1}(1-\theta B)a_t \\ &= a_t + \left[\lambda - \frac{L(B)(1-\delta)}{1-\delta B}\right]Sa_{t-1}\end{aligned} \tag{A15.1.2}$$

where $S = (1 - B)^{-1}$. The coefficient of a_t in this expression is unity, so we can write

$$\varepsilon_t = [1 + B\mu(B)]a_t \tag{A15.1.3}$$

where

$$\mu(B) = \mu_1 + \mu_2 B + \mu_3 B^2 + \cdots$$

Furthermore, in practice, control would need to be exerted in terms of the observed output errors ε_t rather than in terms of the a_t's, so that the control equation actually used would be of the form

$$x_t = -\frac{1}{g}\frac{L(B)}{1 + B\mu(B)}\varepsilon_t \tag{A15.1.4}$$

Equating (A15.1.2) and (A15.1.3), we obtain

$$(1 - \delta)L(B) = [\lambda - (1 - B)\mu(B)](1 - \delta B) \tag{A15.1.5}$$

Since δ, g, and σ_a^2 are constants, we can proceed conveniently by finding an unrestricted minimum of

$$C = \frac{(1-\delta)^2 g^2 V[x_t] + \upsilon V[\varepsilon_t]}{\sigma_a^2} \tag{A15.1.6}$$

where, for example,

$$V[x_t] = \text{var}[x_t]$$

and $\upsilon = (1-\delta)^2 g^2/\alpha$. Now, from (A15.1.3), $V[\varepsilon_t]/\sigma_a^2 = 1 + \sum_{j=1}^{\infty} \mu_j^2$, while from (A15.1.1), $(1-\delta)g x_t = -(1-\delta)L(B)a_t = -\tau(B)a_t$, so that

$$\frac{(1-\delta)^2 g^2 V[x_t]}{\sigma_a^2} = \sum_{j=0}^{\infty} \tau_j^2$$

where

$$\tau(B) = \sum_{j=0}^{\infty} \tau_j B^j = (1-\delta)L(B) = [\lambda - (1-B)\mu(B)](1-\delta B)$$

from (A15.1.5). The coefficients $\{\tau_i\}$ are thus seen to be functionally related to the μ_i by the difference equation

$$\mu_i - (1+\delta)\mu_{i-1} + \delta\mu_{i-2} = -\tau_{i-1} \qquad \text{for } i > 2 \tag{A15.1.7}$$

with $\tau_0 = -(\mu_1 - \lambda)$, $\tau_1 = -[\mu_2 - (1+\delta)\mu_1 + \lambda\delta]$. Hence, we require an unrestricted minimum, with respect to the μ_i, of the expression

$$C = \sum_{j=0}^{\infty} \tau_j^2 + \upsilon\left(1 + \sum_{j=1}^{\infty} \mu_j^2\right) \tag{A15.1.8}$$

This can be obtained by differentiating C with respect to each μ_i $(i = 1, 2, \ldots)$, equating these derivatives to zero and solving the resulting equations. Now, a given μ_i only influences the values τ_{i+1}, τ_i, and τ_{i-1} through (A15.1.7), and we see that

$$\frac{\partial \tau_j}{\partial \mu_i} = \begin{cases} -1 & j = i-1 \\ 1+\delta & j = i \\ -\delta & j = i+1 \\ 0 & \text{otherwise} \end{cases} \tag{A15.1.9}$$

Therefore, from (A15.1.8) and (A15.1.9), we obtain

$$\begin{aligned} \frac{\partial}{\partial \mu_i} C &= 2\left(\tau_{i+1}\frac{\partial \tau_{i+1}}{\partial \mu_i} + \tau_i \frac{\partial \tau_i}{\partial \mu_i} + \tau_{i-1}\frac{\partial \tau_{i-1}}{\partial \mu_i} + \upsilon\mu_i\right) \\ &= 2[-\delta\tau_{i+1} + (1+\delta)\tau_i - \tau_{i-1} + \upsilon\mu_i] \qquad \text{for } i = 1, 2, \ldots \end{aligned} \tag{A15.1.10}$$

Then, after substituting the expressions for the τ_j in terms of the μ_i from equation (A15.1.7) in (A15.1.10) and setting each of these equal to zero, we obtain the following equations:

$$(i = 1): \quad -\lambda(1 + \delta + \delta^2) + 2(1 + \delta + \delta^2)\mu_1 - (1 + \delta)^2\mu_2 + \delta\mu_3 + \nu\mu_1 = 0 \tag{A15.1.11}$$

$$(i = 2): \quad \lambda\delta - (1 + \delta)^2\mu_1 + 2(1 + \delta + \delta^2)\mu_2 - (1 + \delta)^2\mu_3 + \delta\mu_4 + \nu\mu_2 = 0 \tag{A15.1.12}$$

$$(i > 2): \quad [\delta B^2 - (1 + \delta)^2 B + 2(1 + \delta + \delta^2) - (1 + \delta)^2 F + \delta F^2 + \nu]\mu_i = 0 \tag{A15.1.13}$$

A15.1.2 Case Where δ Is Negligible

Consider first the simpler case where δ is negligibly small and can be set equal to zero. Then the equations above can be written as

$$(i = 1): \quad -(\lambda - \mu_1) + (\mu_1 - \mu_2) + \nu\mu_1 = 0 \tag{A15.1.14}$$

$$(i > 1): \quad [B - (2 + \nu) + F]\mu_i = 0 \tag{A15.1.15}$$

These difference equations have a solution of the form

$$\mu_i = A_1\kappa_1^i + A_2\kappa_2^i$$

where κ_1 and κ_2 are the roots of the characteristic equation

$$B^2 - (2 + \upsilon)B + 1 = 0 \tag{A15.1.16}$$

that is, of

$$B + B^{-1} = 2 + \nu$$

Evidently, if κ is a root, so is κ^{-1}. Thus, the solution is of the form $\mu_i = A_1\kappa^i + A_2\kappa^{-i}$. Now if κ has modulus less than or equal to 1, κ^{-1} has modulus greater than or equal to 1, and since $\varepsilon_t = [1 + B\mu(B)]a_t$ must have finite variance, A_2 must be zero with $|\kappa| < 1$. By substituting the solution $\mu_i = A_1\kappa^i$ in (A15.1.14), we find that $A_1 = \lambda$.

Finally, then, $\mu_i = \lambda\kappa^i$, and since μ_i and λ must be real, so must the root κ. Hence,

$$\mu(B) = \frac{\lambda\kappa}{1 - \kappa B} \qquad 0 < \kappa < 1 \tag{A15.1.17}$$

$$1 + B\mu(B) = 1 + \frac{\lambda\kappa B}{1 - \kappa B} = \frac{1 - \theta\kappa B}{1 - \kappa B} \tag{A15.1.18}$$

where $\theta = 1 - \lambda$. Thus,

$$\varepsilon_t = \frac{1 - \theta\kappa B}{1 - \kappa B}a_t$$

so that

$$\frac{V[\varepsilon_t]}{\sigma_a^2} = 1 + \frac{\lambda^2\kappa^2}{1 - \kappa^2} \tag{A15.1.19}$$

Also, using (A15.1.5) with $\delta = 0$,

$$L(B) = \lambda - \frac{(1-B)\lambda\kappa}{1-\kappa B} = \frac{\lambda(1-\kappa)}{1-\kappa B} \tag{A15.1.20}$$

Thus,

$$x_t = -\frac{\lambda}{g}\frac{1-\kappa}{1-\kappa B}a_t$$

and

$$\frac{V[x_t]}{\sigma_a^2} = \frac{\lambda^2}{g^2}\frac{(1-\kappa)^2}{1-\kappa^2} = \frac{\lambda^2}{g^2}\frac{1-\kappa}{1+\kappa} \tag{A15.1.21}$$

Using (A15.1.4) with (A15.1.18) and (A15.1.20), we now find that the optimal control action, in terms of the observed output error ε_t, is

$$x_t = -\frac{1}{g}\frac{\lambda(1-\kappa)}{1-\theta\kappa B}\varepsilon_t$$

that is,

$$x_t = (1-\lambda)\kappa x_{t-1} - \frac{1}{g}\lambda(1-\kappa)\varepsilon_t \tag{A15.1.22}$$

Note that the constrained control equation differs from the unconstrained one in two respects:

1. A new factor $(1-\lambda)\kappa x_{t-1}$ is introduced, thus making present action depend partly on previous action.
2. The constant determining the amount of integral control is reduced by a factor $1-\kappa$.

We have supposed that the output variance is allowed to increase to some value $\sigma_a^2(1+c)$. It follows from (A15.1.19) that

$$c = \frac{\lambda^2\kappa^2}{1-\kappa^2}$$

that is,

$$\kappa = \sqrt{\frac{c}{\lambda^2 + c}}$$

where the positive square root is to be taken. It is convenient to write $Q = c/\lambda^2$. Then, $Q = \kappa^2/(1-\kappa^2)$ and $\kappa^2 = Q/(1+Q)$ and the output variance becomes $\sigma_a^2(1+\lambda^2 Q)$.

In summary, suppose that we are prepared to tolerate an increase in variance in the output to some value $\sigma_a^2(1+\lambda^2 Q)$; then

1. We compute $\kappa = \sqrt{Q/(1+Q)}$.
2. Optimal control will be achieved by taking action given by (A15.1.22).

Table A15.1 Values of Parameters for a Simple Constrained Control Scheme

$c/\lambda^2 = Q$	κ	W	$c/\lambda^2 = Q$	κ	W
0.10	0.302	53.7	0.60	0.612	24.0
0.20	0.408	42.0	0.70	0.641	21.9
0.30	0.480	35.1	0.80	0.667	20.0
0.40	0.535	30.3	0.90	0.688	18.5
0.50	0.577	26.8	1.00	0.707	17.2

3. The variance of the input will be reduced to

$$V[x_t] = \frac{\lambda^2}{g^2}\frac{1-\kappa}{1+\kappa}\sigma_a^2$$

that is, it will reduce to a value that is $W\%$ of that for the unconstrained scheme, where

$$W = 100\left(\frac{1-\kappa}{1+\kappa}\right)$$

Table A15.1 shows κ and W for values of Q between 0.1 and 1.0. For illustration, suppose that $\lambda = 0.4$. Then the optimal unconstrained scheme will employ the control action

$$x_t = -\frac{0.4}{g}\varepsilon_t$$

with $\varepsilon_t = a_t$. The variance of x_t would be $V[x_t] = (\sigma_a^2/g^2)0.16$. Suppose that it was desired to reduce this by a factor of 4, to the value $(\sigma_a^2/g^2)0.04$. Thus, we require W to be 25%. Table A15.1 shows that a reduction of the input variance to 24% of its unconstrained value is possible with $Q = 0.60$ and $\kappa = 0.612$. If we use this scheme, the output variance will be

$$\sigma_\varepsilon^2 = \sigma_a^2(1 + 0.16 \times 0.60) = 1.10\sigma_a^2$$

Thus, by the use of the control action

$$x_t = 0.37x_{t-1} - \frac{1}{g}0.16\varepsilon_t$$

instead of $x_t = -(0.4/g)\varepsilon_t$, the variance of the input is reduced to about 1/4 of its previous value, while the variance of the output is increased by only 10%.

Case Where δ Is Not Negligible. Consider now the more general situation where δ is not negligible and the system dynamics must be taken account of. The difference equation (A15.1.13) is of the form

$$(\alpha B^{-2} + \beta B^{-1} + \gamma + \beta B + \alpha B^2)\mu_i = 0$$

and if κ is a root of the characteristic equation, so is κ^{-1}. Suppose that the roots are $\kappa_1, \kappa_2, \kappa_1^{-1}, \kappa_2^{-1}$ and that κ_1 and κ_2 are a pair of roots with modulus < 1. Then, in the

solution

$$\mu_i = A_1\kappa_1^i + A_2\kappa_2^i + A_3\kappa_1^{-i} + A_4\kappa_2^{-i}$$

A_3 and A_4 must be zero, because ε_t is required to have a finite variance.

Hence, the solution is of the form

$$\mu_i = A_1\kappa_1^i + A_2\kappa_2^i \qquad |\kappa_1| < 1 \quad |\kappa_2| < 1$$

The A's satisfying the initial conditions, defined by (A15.1.11) and (A15.1.12), are obtained by substitution to give

$$A_1 = \frac{\lambda\kappa_1(1-\kappa_2)}{\kappa_1 - \kappa_2} \qquad A_2 = -\frac{\lambda\kappa_2(1-\kappa_1)}{\kappa_1 - \kappa_2}$$

If we write $k_0 = \kappa_1 + \kappa_2 - \kappa_1\kappa_2$, $k_1 = \kappa_1\kappa_2$, then

$$\mu(B) = \lambda\left[\frac{k_0 - k_1 B}{1 - (k_0 + k_1)B + k_1 B^2}\right] \tag{A15.1.23}$$

and

$$1 + B\mu(B) = \frac{1 - k_1 B - (1-\lambda)(k_0 B - k_1 B^2)}{1 - (k_0 + k_1)B + k_1 B^2} \tag{A15.1.24}$$

Now substituting (A15.1.23) in (A15.1.5),

$$L(B) = \frac{\lambda(1-\delta B)(1-k_0)}{(1-\delta)[1 - (k_0 + k_1)B + k_1 B^2]} \tag{A15.1.25}$$

and

$$\frac{L(B)}{1 + B\mu(B)} = \frac{\lambda(1-\delta B)(1-k_0)}{(1-\delta)[1 - k_1 B - (1-\lambda)(k_0 B - k_1 B^2)]}$$

Therefore, using (A15.1.4), we find that the optimal control action in terms of the error ε_t is

$$x_t = -\frac{\lambda}{g}\frac{(1-\delta B)(1-k_0)}{(1-\delta)[1 - k_1 B - (1-\lambda)(k_0 B - k_t B^2)]}\varepsilon_t \tag{A15.1.26}$$

or

$$x_t = [k_1 + (1-\lambda)k_0]x_{t-1} - (1-\lambda)k_1 x_{t-2} - \frac{\lambda(1-k_0)(1-\delta B)}{g(1-\delta)}\varepsilon_t \tag{A15.1.27}$$

Thus, the modified control scheme makes x_t depend on both x_{t-1} and x_{t-2} (only on x_{t-1} if $\lambda = 1$) and reduces the standard integral and proportional action by a factor $1 - k_0$.

Variances of Output and Input. The actual variances for the output and input are readily found since

$$\varepsilon_t = a_t + \lambda\left[\frac{k_0 - k_1 B}{1 - (k_0 + k_1)B + k_1 B^2}\right] a_{t-1}$$

The second term on the right defines a mixed autoregressive–moving average process of order (2, 0, 1), the variance for which is readily obtained to give

$$\frac{V[\varepsilon_t]}{\sigma_a^2} = 1 + \lambda^2 \left\{ \frac{(k_0 + k_1)^2(1 - k_1) - 2k_1(k_0 - k_1^2)}{(1 - k_1)(1 + k_1)^2 - (k_0 + k_1)^2]} \right\} = 1 + \lambda^2 Q \qquad \text{(A15.1.28)}$$

Also,

$$\frac{V[x_t]}{\sigma_a^2} = \frac{\lambda^2}{g^2(1 - \delta)^2} \frac{(1 - k_0)[(1 + \delta^2)(1 + k_1) - 2\delta(k_0 + k_1)]}{(1 + k_0 + 2k_1)(1 - k_1)} \qquad \text{(A15.1.29)}$$

Computation of k_0 and k_1. Returning to the difference equations (A15.1.13), the characteristic equation may be written

$$B^4 - MB^3 - NB^2 - MB + 1 = 0$$

where $M = (1 + \delta)^2/\delta$ and $N = [(1 + \delta^2) + (1 + \delta^2) + \upsilon]/\delta$. It may also be written in the form

$$(B^2 - TB + P)(B^2 - P^{-1}TB + P^{-1}) = 0$$

where

$$T = \kappa_1 + \kappa_2 \quad \text{and} \quad P = \kappa_1\kappa_2$$

Equating coefficients of B gives

$$T + P^{-1}T = M$$

that is, $T = PM/(1 + P)$, and

$$P + P^{-1} + P^{-1}T^2 = N$$

Thus, $P + P^{-1} + PM^2/(1 + P)^2 = N$, that is,

$$(P + 2 + P^{-1})(P + P^{-1}) + M^2 = N(P + 2 + P^{-1})$$

or

$$(P + P^{-1})^2 + (2 - N)(P + P^{-1}) + M^2 - 2N = 0$$

For suitable vales of υ, this quadratic equation will have two real roots:

$$u_1 = \kappa_1\kappa_2 + \kappa_1^{-1}\kappa_2^{-1} \quad u_2 = \kappa_1\kappa_2^{-1} + \kappa_1^{-1}\kappa_2$$

the root u_1 being the larger. The required quantity P is now the smaller root of the quadratic equation

$$P^2 - u_1P + 1 = 0$$

and T is given by

$$T = [P(u_2 + 2)]^{1/2}$$

Table A15.2 Table to Facilitate the Calculation of Optimal Constrained Control Schemes

		100Q				
δ		20	40	60	80	100
0.9	100 W	21.7	11.3	6.7	4.5	3.1
	k_0	0.44	0.585	0.68	0.74	0.78
	k_1	0.18	0.27	0.34	0.39	0.44
0.8	100 W	22.0	11.7	7.2	4.8	3.4
	k_0	0.44	0.585	0.68	0.74	0.78
	k_1	0.18	0.27	0.33	0.38	0.43
0.7	100 W	22.7	12.4	8.0	5.6	4.1
	k_0	0.44	0.585	0.68	0.74	0.78
	k_1	0.17	0.25	0.32	0.36	0.40
0.6	100 W	24.1	13.6	9.0	6.6	5.0
	k_0	0.44	0.58	0.67	0.73	0.78
	k_1	0.16	0.24	0.29	0.33	0.365
0.5	100 W	26.5	15.5	10.5	7.9	6.2
	k_0	0.43	0.58	0.67	0.72	0.77
	k_1	0.15	0.21	0.26	0.29	0.32
0.4	100 W	28.5	17.7	12.7	9.8	7.9
	k_0	0.43	0.57	0.66	0.72	0.76
	k_1	0.13	0.18	0.22	0.245	0.265
0.3	100 W	31.5	20.5	15.2	12.0	9.9
	k_0	0.43	0.57	0.65	0.71	0.75
	k_1	0.105	0.145	0.17	0.19	0.20
0.2	100 W	34.8	23.6	18.0	14.5	12.2
	k_0	0.42	0.56	0.64	0.69	0.73
	k_1	0.07	0.10	0.12	0.13	0.14
0.1	100 W	38.2	26.7	21.0	17.3	14.6
	k_0	0.42	0.55	0.63	0.68	0.72
	k_1	0.04	0.05	0.06	0.065	0.07

Table of Optimal Values for Constrained Schemes

Construction of the Table. Table A15.2 is provided to facilitate the selection of an optimal control scheme. The tabled values were obtained as follows for each chosen value of the parameter δ in the transfer function model:

1. Compute $M = (1+\delta)^2/\delta$ and $N = ((1+\delta)^2 + (1+\delta^2) + v)/\delta$ for a series of values of v chosen to provide a suitable range for Q.
2. Compute $u_1 = 1/2(N-2) + \left[((N-2)/2)^2 + 2N - M^2\right]^{1/2}$ and $u_2 = 1/2(N-2) - \left[((N-2)/2)^2 + 2N - M^2\right]^{1/2}$
3. Compute $k_1 = P = 1/2u_1 - \left[\left(1/2u_1\right)^2 - 1\right]^{1/2}$ and $k_0 = T - P = [k_1(u_2+2)]^{1/2} - k_1$.
4. Compute $Q = \dfrac{(k_0+k_1)^2(1-k_1) - 2k_1(k_0-k_1^2)}{(1-k_1)[(1+k_1)^2 - (k_1+k_1)^2]}$.

5. Compute $W = \dfrac{(1-k_0)[(1+\delta^2)(1+k_1)-2\delta(k_0+k_1)]}{(1+k_0+2k_1)(1-k_1)(1+\delta^2)}$.

6. Interpolate among the W, k_0, k_1 values at convenient values of Q.

Use of the Table. Table A15.2 may be used as follows. The value of δ is entered in the vertical margin. Using the fact that $V[\varepsilon_t] = (1+\lambda^2 Q)\sigma_a^2$, the percentage increase in output variance is $100Q\lambda^2$. A suitable value of Q is entered in the horizontal margin. The entries in the table are then (1) $100W$, the percentage reduction in the variance of x_t, (2) k_0, and (3) k_1.

For illustration, suppose that $\lambda = 0.6, \delta = 0.5$, and $g = 1$. The optimal unconstrained control equation is then

$$x_t = -1.2(1-0.5B)\varepsilon_t = -1.2(1-0.5B)a_t$$

and $\text{var}[x_t] = 1.80\sigma_a^2$. Suppose that this amount of variation in the input variable produces difficulties in process operation and it is desired to reduce $\text{var}[x_t]$ to about $0.50\sigma_a^2$, that is, to about 28% of the value for the unconstrained scheme. Inspection of Table A15.2 in the row labeled $\delta = 0.5$ shows that a reduction to 26.5% can be achieved by using a control scheme with constants $k_0 = 0.43, k_1 = 0.15$, that is, by employing the control equation (A15.1.27) to give

$$x_t = 0.32x_{t-1} - 0.06x_{t-2} - (0.57 \times 1.2)(1-0.5B)\varepsilon_t$$

This solution corresponds to a value $Q = 0.20$. Therefore, the variance at the output will be increased by a factor of

$$1 + \lambda^2 Q = 1 + 0.6^2(0.2) = 1.072$$

that is, by about 7%.

APPENDIX A15.2 CHOICE OF THE SAMPLING INTERVAL

In comparison to continuous systems, discrete systems of control, such as those discussed here, can be very efficient provided that the sampling interval is suitably chosen. Roughly speaking, we want the interval to be such that not too much change can occur during the sampling interval. Usually, the behavior of the disturbance that has to pass through all or part of the system reflects the inertia or dynamic properties of the system, so that the sampling interval will often be chosen tacitly or explicitly to be proportional to the time constant or constants of the system. In chemical processes involving reaction and mixing of liquids, rather infrequent sampling, say at hourly intervals and possibly with operator surveillance and manual adjustment, will be sufficient. By contrast, where reactions between gases are involved, a suitable sampling interval may be measured in seconds and automatic monitoring and adjustment may be essential.

In some cases, experimentation may be needed to arrive at a satisfactory sampling interval, and in others rather simple calculations will show how the choice of sampling interval will affect the degree of control that is possible.

A15.2.1 Illustration of the Effect of Reducing Sampling Frequency

To illustrate the kind of calculation that is helpful, suppose again that we have a simple system in which, using a particular sampling interval, the noise is represented by a (0, 1, 1) process $\nabla N_t = (1 - \theta B)a_t$ and the transfer function model by the first-order system $(1 - \delta B)\mathcal{Y}_t = g(1 - \delta)X_{t-1}$. In this case, if we employ the MMSE adjustment

$$x_t = -\frac{1-\theta}{g(1-\delta)}(1 - \delta B)\varepsilon_t \tag{A15.2.1}$$

then the deviation from target is $\varepsilon_t = a_t$ and has variance $\sigma_a^2 = \sigma_1^2$, say.

In practice, the question has often arisen: How much worse off would we be if we took samples less frequently? To answer this question, we consider the effect of sampling the stochastic process involved.

A15.2.2 Sampling an IMA(0, 1, 1) Process

Suppose that with observations being made at some ''unit'' interval, we have a noise model

$$\nabla N_t = (1 - \theta_1 B)a_t$$

with $\text{var}[a_t] = \sigma_a^2 = \sigma_1^2$, where the subscript 1 is used in this context to denote the choice of sampling interval. Then, for the differences ∇N_t, the autocovariances γ_k are given by

$$\begin{aligned} \gamma_0 &= (1 + \theta_1^2)\sigma_1^2 \\ \gamma_1 &= -\theta_1\sigma_1^2 \\ \gamma_j &= 0 \qquad j \geq 2 \end{aligned} \tag{A15.2.2}$$

Writing $\zeta = (\gamma_0 + 2\gamma_1)/\gamma_1$, we obtain

$$\zeta = -\frac{(1-\theta_1)^2}{\theta_1}$$

so that, given γ_0 and γ_1, the parameter $\lambda = 1 - \theta_1$ of the IMA process may be obtained by solving the quadratic equation

$$(1 - \theta_1)^2 - \zeta(1 - \theta_1) + \zeta = 0$$

selecting that root for which $-1 < \theta_1 < 1$. Also,

$$\sigma_1^2 = -\frac{\gamma_1}{\theta_1} \tag{A15.2.3}$$

Suppose now that the process N_t is observed at intervals of h units (where h is a positive integer) and the resulting process is denoted by M_t. Then,

$$\begin{aligned} \nabla M_t = N_t - N_{t-h} &= (a_t + a_{t-1} + \cdots + a_{t-h+1}) \\ &\quad - \theta_1(a_{t-1} + a_{t-2} + \cdots + a_{t-h}) \\ \nabla M_{t-h} = N_{t-h} - N_{t-2h} &= (a_{t-h} + a_{t-h-1} + \cdots + a_{t-2h+1}) \\ &\quad - \theta_1(a_{t-h-1} + \cdots + a_{t-2h}) \end{aligned}$$

and so on. Then, for the differences ∇M_t, the autocovariances $\gamma_k(h)$ are

$$\begin{aligned}\gamma_0(h) &= [(1+\theta_1^2) + (h-1)(1-\theta_1)^2]\sigma_1^2 \\ \gamma_1(h) &= -\theta_1\sigma_1^2 \\ \gamma_j(h) &= 0 \qquad j \geq 2\end{aligned} \tag{A15.2.4}$$

It follows that the process M_t is also an IMA process of order (0, 1, 1),

$$\nabla M_t = (1-\theta_h B)e_t$$

where e_t is a white noise process with variance σ_h^2. Now

$$\frac{\gamma_0(h)+2\gamma_1(h)}{\gamma_1(h)} = -\frac{h(1-\theta_1)^2}{\theta_1}$$

so that

$$\frac{h(1-\theta_1)^2}{\theta_1} = \frac{(1-\theta_h)^2}{\theta_h} \tag{A15.2.5}$$

Also, since $\gamma_1(h) = -\theta_h\sigma_h^2 = -\theta_1\sigma_1^2$, it follows that

$$\frac{\sigma_h^2}{\sigma_1^2} = \frac{\theta_1}{\theta_h} \tag{A15.2.6}$$

Therefore, we have shown that the sampling of an IMA process of order (0, 1, 1) at interval h produces another IMA process of order (0, 1, 1). From (A15.2.5), we can obtain the value of the parameter θ_h for the sampled process, and from (A15.2.6) we can obtain the variance $\sigma_h^2 = \text{var}[e_t]$ of the corresponding white noise generating process in terms of the parameters θ_1 and $\sigma_t^2 = \text{var}[a_t]$ of the original process.

In Figure A15.1, θ_n is plotted against log h, a scale of h being appended. The graph enables one to find the effect of increasing the sampling interval of a (0, 1, 1) process by any given multiple. For illustration, suppose that we have a process for which $\theta_1 = 0.5$ and $\sigma_1^2 = 1$. Let us use the graph to find the values of the corresponding parameters $\theta_2, \theta_4, \sigma_2^2, \sigma_4^2$ when the sampling interval is (a) doubled and (b) quadrupled. Marking on the edge of a piece of paper the points $h = 1, h = 2, h = 4$ from the scale of the graph, we set the paper

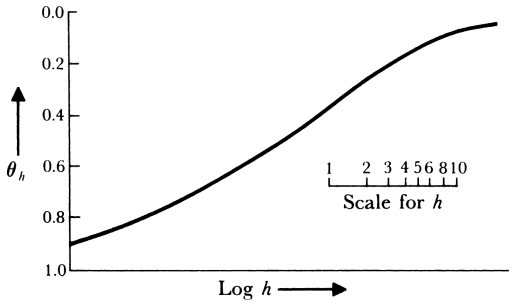


FIGURE A15.1 Sampling of IMA(0, 1, 1) process: parameter θ_h plotted against log h.

horizontally so that $h = 1$ corresponds to the point on the curve for which $\theta_1 = 0.5$. We then read off the ordinates for θ_2 and θ_4 corresponding to $h = 2$ and $h = 4$. We find that

$$\theta_1 = 0.5 \qquad \theta_2 = 0.38 \qquad \theta_4 = 0.27$$

Using (A15.2.6), the variances are in inverse proportion to the values of θ, so that

$$\sigma_1^2 = 1.00 \qquad \sigma_2^2 = 1.32 \qquad \sigma_4^2 = 2.17$$

Suppose now that for the original scheme with unit interval, the dynamic constant was δ_1 (again we will use the subscript to denote the sampling interval). Then, since in real time the same fixed time constant $T = -h/\ln(\delta)$ applies to all the schemes, we have

$$\delta_2 = \delta_1^2 \qquad \delta_4 = \delta_1^4$$

The scheme giving minimum mean square error for a *particular* sampling interval h would be

$$x_t(h) = -\frac{1-\theta_h}{g(1-\delta_1^h)}(1-\delta_1^h B)\varepsilon_t(h)$$

or

$$x_t(h) = -\frac{1-\theta_h}{g}\left(1+\frac{\delta_1^h}{1-\delta_1^h}\nabla\right)\varepsilon_t(h) \tag{A15.2.7}$$

Suppose, for example, with $\theta_1 = 0.5$ as above, $\delta_1 = 0.8$, so that $\delta_2 = 0.64$, $\delta_4 = 0.41$. Then the optimal schemes would be

$$h = 1: \quad x_t(1) = -\frac{0.5}{g}(1+4\nabla)\varepsilon_t(1) \qquad \sigma_\varepsilon^2 = 1.00 \qquad g^2\sigma_x^2 = 10.25$$

$$h = 2: \quad x_t(2) = -\frac{0.62}{g}(1+1.78\nabla)\varepsilon_t(2) \qquad \sigma_\varepsilon^2 = 1.32 \qquad g^2\sigma_x^2 = 5.50$$

$$h = 4: \quad x_t(4) = -\frac{0.73}{g}(1+0.69\nabla)\varepsilon_t(4) \qquad \sigma_\varepsilon^2 = 2.17 \qquad g^2\sigma_x^2 = 3.84$$

In accordance with expectation, as the sampling interval is increased and the dynamics of the system have relatively less importance, the amount of ''integral'' control is increased and the ratio of proportional to integral control is markedly reduced. We noted earlier that an excessively large adjustment variance σ_x^2 would usually be a disadvantage. The values of $g^2\sigma_x^2$ are indicated to show how the schemes differ in this respect. The smaller value for σ_x^2 would not of itself, of course, justify the choice $h = 4$. Using an optima1 constrained scheme, as is described in Appendix A15.1, with $h = 1$, a very large reduction in σ_x^2 would be produced with only a small increase in the output variance. For example, entering Table A15.2 with $\delta = 0.8, 100Q = 20$, we find that for a 5% increase of output variance to the value $(1+\lambda^2 Q)\sigma_1^2 = 1.05\sigma_1^2$, the input variance for the scheme with $h = 1$ could be reduced to 22% of its unconstrained value, so that $g^2\sigma_x^2 = 10.25 \times 0.22 = 2.26$.

Using (A15.1.27), we obtain for the constrained scheme with $h = 1$,

$$x_t = 0.40x_{t-1} - 0.09x_{t-2} - 0.56\left[\frac{0.5}{g}(1 + 4\nabla)\right]\varepsilon_t(1)$$
$$\sigma_\varepsilon^2 = 1.05 \qquad g^2\sigma_x^2 = 2.26$$

In practice, various alternative schemes could be set out with their accompanying characteristics and an economic choice made to suit the particular problem. In general, the increase in output variance that comes with the larger interval would have to be balanced off against the economic advantage, if any, of less frequent surveillance.

EXERCISES

15.1. In a chemical process, 30 successive values of viscosity N_t that occurred during a period when the control variable (gas rate) X_t *was held fixed* at its standard reference origin were recorded as follows:

Time	Viscosities									
1–10	92	92	96	96	96	98	98	100	100	94
11–20	98	88	88	88	96	96	92	92	90	90
21–30	90	94	90	90	94	94	96	96	96	96

Reconstruct and plot the error sequence (deviations from target) ε_t and adjustments x_t, which would have occurred if the optimal feedback control scheme

$$x_t = -10\varepsilon_t + 5\varepsilon_{t-1} \tag{1}$$

had been applied during this period. It is given that the dynamic model is

$$y_t = 0.5y_{t-1} + 0.10x_{t-1} \tag{2}$$

and that the error signal may be obtained from

$$\varepsilon_t = \varepsilon_{t-1} + \nabla N_t + y_t \tag{3}$$

Your calculation sequence should proceed in the order (2), (3), and (1) and initially you should assume that $\varepsilon_1 = 0, y_1 = 0, x_1 = 0$. Can you devise a more direct way to compute ε_t from N_t?

15.2. Given the following combinations of disturbance and transfer function models:

$$\begin{aligned}
&(1) && \nabla N_t = (1 - 0.7B)a_t \\
& && (1 - 0.4B)\mathcal{Y}_t = 5.0X_{t-1+} \\
&(2) && \nabla N_t = (1 - 0.5B)a_t \\
& && (1 - 1.2B + 0.4B^2)\mathcal{Y}_t = (20 - 8.5)X_{t-1+} \\
&(3) && \nabla^2 N_t = (1 - 0.9B + 0.5B^2)a_t \\
& && (1 - 0.7B)\mathcal{Y}_t = 3.0X_{t-1+} \\
&(4) && \nabla N_t = (1 - 0.7B)a_t \\
& && (1 - 0.4B)\mathcal{Y}_t = 5.0X_{t-2+}
\end{aligned}$$

(a) Design the minimum mean square error feedback control schemes associated with each combination of disturbance and transfer function model.

(b) For case (4), derive an expression for the error ε_t and for its variance in terms of σ_a^2.

(c) For case (4), design a nomogram suitable for carrying out the control action manually by a process operator.

15.3. In a treatment plant for industrial waste, the strength u_t of the influent is measured every 30 minutes and can be represented by the model $\nabla u_t = (1 - 0.5B)\alpha_t$. In the absence of control, the strength of the effluent Y_t is related to that of the influent u_t by an effect $\mathcal{Y}_{1t}$ that can be represented as

$$\mathcal{Y}_{1t} = \frac{0.3B}{1 - 0.2B}\tilde{u}_t$$

An increase in strength in the waste may be compensated by an increase in the flow X_t of a chemical to the plant, whose effect on Y_t is represented by the effect

$$\mathcal{Y}_{2t} = \frac{21.6B^2}{1 - 0.7B}\tilde{X}_t$$

Show that minimum mean square error feedforward control is obtained with the control equation

$$\tilde{X}_t = -\frac{0.3}{21.6}\left[\frac{(0.7 - 0.2B)(1 - 0.7B)}{(1 - 0.2B)(1 - 0.5B)}\right]\tilde{u}_t$$

that is, $\tilde{X}_t = 0.7\tilde{X}_{t-1} - 0.1\tilde{X}_{t-2} - 0.0139(0.7\tilde{u}_t - 0.69\tilde{u}_{t-1} + 0.14\tilde{u}_{t-2})$.

15.4. A pilot feedback control scheme, based on the following disturbance and transfer function models:

$$\begin{aligned}
\nabla N_t &= a_t \\
(1 - \delta B)\mathcal{Y}_t &= \omega_0 X_{t-1+} - \omega_1 X_{t-2+}
\end{aligned}$$

was operated, leading to a series of adjustments x_t and errors ε_t. It was believed that the noise model was reasonably accurate, but that the parameters of the transfer function model were of questionable accuracy.

(a) Given the first 10 values of the x_t, ε_t series shown below:

t	x_t	ε_t	t	x_t	ε_t
1	25	−7	6	−30	1
2	42	−7	7	−25	3
3	3	−6	8	−25	4
4	20	−7	9	20	0
5	5	−4	10	40	−3

set out the calculation of the residuals a_t $(t = 2, 3, \ldots, 10)$ for $\delta = 0.5$, $\omega_0 = 0.3, \omega_1 = 0.2$, and for arbitrary starting values y_1^0 and x_0^0.

(b) Calculate the values $y_1, \hat{x}_0$ of y_1^0 and x_0^0 that minimize the sum of squares $\sum_{t=2}^{10}(a_t|\delta = 0.5, \omega_0 = 0.3, \omega_1 = 0.2, y_1^0, x_0^0)^2$ and the value of this minimum sum of squares.

15.5. Consider (Box and MacGregor, 1976) a system for which the process transfer function is gB and the noise model is $(1 - B)N_t = (1 - \theta B)a_t$ so that the error ε_t at the output satisfies

$$(1 - B)\varepsilon_t = g(1 - B)X_{t-1+} + (1 - \theta B)a_t$$

Suppose that the system is controlled by a known discrete ''integral'' controller

$$(1 - B)X_{t+} = -c\varepsilon_t$$

(a) Show that the errors ε_t at the output will follow the ARMA(1, 1) process

$$(1 - \phi B)\varepsilon_t = (1 - \theta B)a_t \qquad \phi = 1 - gc$$

and hence that the problem of estimating g and θ using data from a pilot control scheme is equivalent to that of estimating the parameters in this ARMA(1, 1) model.

(b) Show also that the optimal control scheme is such that $c = c_0 = (1 - \theta)/g$ and hence that if the pilot scheme used in collecting the data happens to be optimal already, then $1 - \theta$ and g cannot be separately estimated.

PART FIVE

CHARTS AND TABLES

This part of the book is a collection of auxiliary material useful in the analysis of time series. This includes tables and charts for obtaining preliminary estimates of the parameters in autoregressive–moving-average models, together with the usual tail area tables of the normal, χ^2, and t distributions. This is followed by a listing of the time series analyzed in the book, as well as some additional time series that are discussed in the exercises located at the end of the individual chapters.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

COLLECTION OF TABLES AND CHARTS

TABLE A Table relating ρ_1 to θ for a first-order moving average process

CHART B Chart relating ρ_1 and ρ_2 to ϕ_1 and ϕ_2 for a second-order autoregressive process

CHART C Chart relating ρ_1 and ρ_2 to θ_1 and θ_2 for a second-order moving average process

CHART D Chart relating ρ_1 and ρ_2 to ϕ and θ for a mixed first-order autoregressive–moving average process

TABLE E Tail areas and ordinates of unit normal distribution

TABLE F Tail areas of the chi-square distribution

TABLE G Tail areas of the *t* distribution

Charts B, C, and D are adapted and reproduced from Stralkowski (1968) with permission of the author. Tables E, F, and G are condensed and adapted from *Biometrika Tables for Statisticians,* Volume I, with permission from the trustees of Biometrika.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

TABLE A Table Relating ρ_1 to θ for a First-Order Moving Average Process

θ	ρ_1	θ	ρ_1
0.00	0.000	0.00	0.000
0.05	−0.050	−0.05	0.050
0.10	−0.099	−0.10	0.099
0.15	−0.147	−0.15	0.147
0.20	−0.192	−0.20	0.192
0.25	−0.235	−0.25	0.235
0.30	−0.275	−0.30	0.275
0.35	−0.315	−0.35	0.315
0.40	−0.349	−0.40	0.349
0.45	−0.374	−0.45	0.374
0.50	−0.400	−0.50	0.400
0.55	−0.422	−0.55	0.422
0.60	−0.441	−0.60	0.441
0.65	−0.457	−0.65	0.457
0.70	−0.468	−0.70	0.468
0.75	−0.480	−0.75	0.480
0.80	−0.488	−0.80	0.488
0.85	−0.493	−0.85	0.493
0.90	−0.497	−0.90	0.497
0.95	−0.499	−0.95	0.499
1.00	−0.500	−1.00	0.500

Table A may be used to obtain first estimates of the parameters in the $(0, d, 1)$ model $w_t = (1 - \theta B)a_t$, where $w_t = \nabla^d z_t$, by substituting $r_1(w)$ for ρ_1.

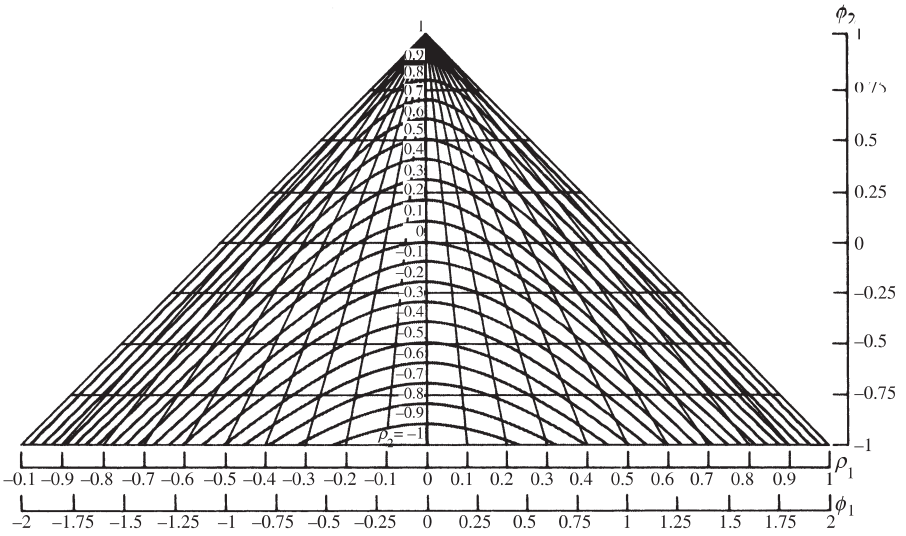


CHART B Chart relating ρ_1 and ρ_1 to ϕ_1 and ϕ_2 for a second-order autoregressive process.

The chart may be used to obtain estimates of the parameters in the $(2, d, 0)$ model $(1 - \phi_1 B - \phi_2 B^2)w_t = a_t$, where $w_t = \nabla^d z_t$, by substituting $r_1(w)$ and $r_2(w)$ for ρ_1 and ρ_2.

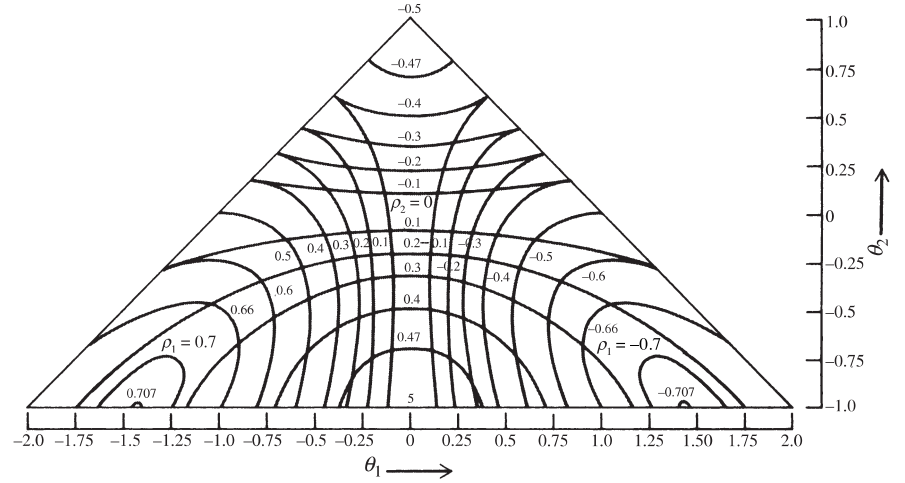


CHART C Chart relating ρ_1 and ρ_1 to θ_1 and θ_2 for a second-order autoregressive process.

The chart may be used to obtain estimates of the parameters in the $(0, d, 2)$ model $w_t = (1 - \theta_1 B - \theta_2 B^2)a_t$, where $w_t = \nabla^d z_t$, by substituting $r_1(w)$ and $r_2(w)$ for ρ_1 and ρ_2.

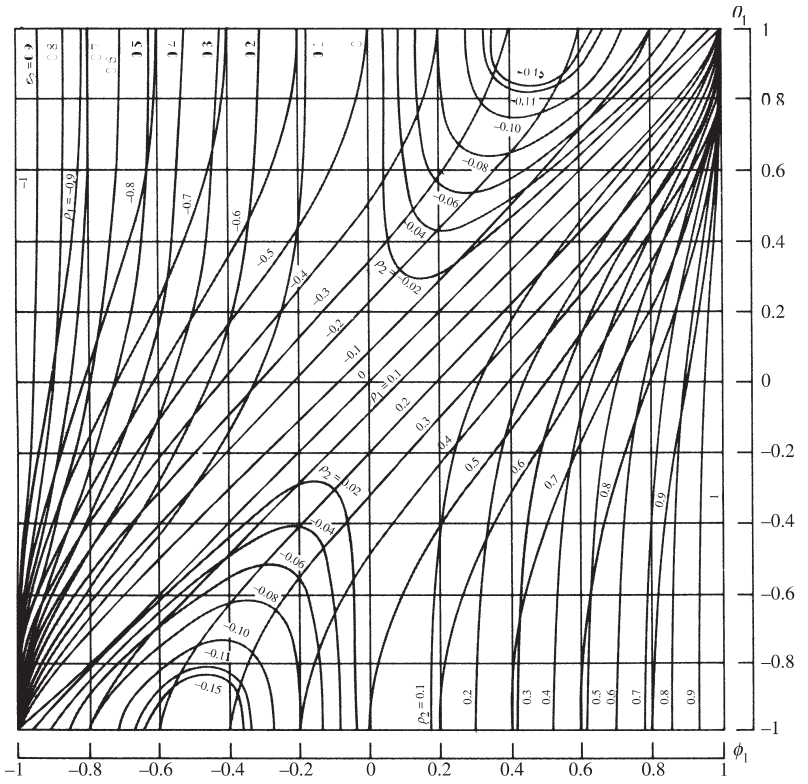


CHART D Chart relating ρ_1 and ρ_2 to ϕ and θ for a mixed first-order autoregressive–moving average process.

The chart may be used to obtain estimates of the parameters in the $(1, d, 1)$ model $(1 - \phi B)w_t = (1 - \theta B)a_t$, where $w_t = \nabla^d z_t$, by substituting $r_1(w)$ and $r_2(w)$ for ρ_1 and ρ_2.

TABLE E Tail Areas and Ordinates of Unit Normal Distribution[a]

u_ε	ε	$p(u_\varepsilon)$	u_ε	ε	$p(u_\varepsilon)$
0.0	0.500	0.3989	1.6	0.055	0.1109
0.1	0.460	0.3969	1.7	0.045	0.0940
0.2	0.421	0.3910	1.8	0.036	0.0790
0.3	0.382	0.3814	1.9	0.029	0.0656
0.4	0.345	0.3683	2.0	0.023	0.0540
0.5	0.309	0.3521	2.1	0.018	0.0440
0.6	0.274	0.3322	2.2	0.014	0.0355
0.7	0.242	0.3123	2.3	0 011	0.0283
0.8	0.212	0.2897	2.4	0.008	0.0224
0.9	0.184	0.2661	2.5	0.006	0.0175
1.0	0.159	0.2420	2.6	0.005	0.0136
1.1	0.136	0.2179	2.7	0.003	0.0104
1.2	0.115	0.1942	2.8	0.003	0.0079
1.3	0.097	0.1714	2.9	0.002	0.0059
1.4	0.081	0.1497	3.0	0.001	0.0044
1.5	0.067	0.1295			

[a] Shown are the values of the unit normal deviate u_ε such that $\Pr\{u > u_\varepsilon\} = \varepsilon$; also shown are the ordinates $p(u = u_\varepsilon)$.

TABLE F Tail Areas of the Chi-Square Distribution[a]

	ε														
m	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005	0.001	m
1	—	—	—	—	0.016	0.102	0.455	1.32	2.71	3.84	5.02	6.63	7.88	10.8	1
2	0.010	0.020	0.051	0.103	0.211	0.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6	13.8	2
3	0.072	0.115	0.216	0.352	0.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8	16.3	3
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13 3	14.9	18.5	4
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7	20.5	5
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5	22.5	6
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3	24.3	7
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0	26.1	8
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6	27.9	9
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2	29.6	10
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8	31.3	11
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3	32.9	12
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8	34.5	13
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3	36.1	14
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8	37.7	15
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3	39.3	16
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7	40.8	17
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2	42.3	18
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6	43.8	19
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0	45.3	20
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4	46.8	21
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8	48.3	22
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2	49.7	23
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6	51.2	24
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9	52.6	25
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3	54.1	26
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6	55.5	27
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0	56.9	28
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3	58.3	29
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7	59.7	30

[a]Shown are the values of $x_\varepsilon^2(m)$ such that $\Pr\{x^2(m) > x_\varepsilon^2(m)\} = \varepsilon$, where m is the number of degrees of freedom.

TABLE G Tail Areas of the *t* Distribution[a]

	ε					
nu	0.25	0.10	0.05	0.025	0.01	0.005
1	1.00	3.08	6.31	12.71	31.82	63.66
2	0.82	1.89	2.92	4.30	6.96	9.92
3	0.76	1.64	2.35	3.18	4.54	5.84
4	0.74	1.53	2.13	2.78	3.75	4.60
5	0.73	1.48	2.02	2.57	3.36	4.03
6	0.72	1.44	1.94	2.45	3.14	3.71
7	0.71	1.42	1.90	2.36	3.00	3.50
8	0.71	1.40	1.86	2.31	2.90	3.36
9	0.70	1.38	1.83	2.26	2.82	3.25
10	0.70	1.37	1.81	2.23	2.76	3.17
11	0.70	1.36	1.80	2.20	2.72	3.11
12	0.70	1.36	1.78	2.18	2.68	3.06
13	0.69	1.35	1.77	2.16	2.65	3.01
14	0.69	1.34	1.76	2.14	2.62	2.98
15	0.69	1.34	1.75	2.13	2.60	2.95
16	0.69	1.34	1.75	2.12	2.58	2.92
17	0.69	1.33	1.74	2.11	2.57	2.90
18	0.69	1.33	1.73	2.10	2.55	2.88
19	0.69	1.33	1.73	2.09	2.54	2.86
20	0.69	1.33	1.72	2.09	2.53	2.84
30	0.68	1.31	1.70	2.04	2.46	2.75
40	0.68	1.30	1.68	2.02	2.42	2.70
60	0.68	1.30	1.67	2.00	2.39	2.66
120	0.68	1.29	1.66	1.98	2.36	2.62
∞	0.67	1.28	1.64	1.96	2.33	2.58

[a] Shown are the values of $t_\varepsilon(\upsilon)$ such that $\Pr\{t(\upsilon) > t_\varepsilon(\upsilon)\} = \varepsilon$, where υ is the number of degrees of freedom.

COLLECTION OF TIME SERIES USED FOR EXAMPLES IN THE TEXT AND IN EXERCISES

SERIES A Chemical process concentration readings: every 2 hours

SERIES B IBM common stock closing prices: daily, May 17, 1961–November 2, 1962

SERIES B′ IBM common stock closing prices: daily, June 29, 1959–June 30, 1960

SERIES C Chemical process temperature readings: every minute

SERIES D Chemical process viscosity readings: every hour

SERIES E Wölfer sunspot numbers: yearly

SERIES F Yields from a batch chemical process: consecutive

SERIES G International airline passengers: monthly totals (thousands of passengers) January 1949–December 1960

SERIES J Gas furnace data

SERIES K Simulated dynamic data with two inputs

SERIES L Pilot scheme data

SERIES M Sales data with leading indicator

SERIES N Mink fur sales of the Hudson's Bay Company: annual for 1850–1911

SERIES P Unemployment and GDP data in UK: quarterly for 1955–1969

SERIES Q Logged and coded U.S. hog price data: annual for 1867–1948

SERIES R Monthly averages of hourly readings of ozone in downtown Los Angeles

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

SERIES A Chemical Process Concentration Readings: Every 2 Hours[a]

1	17.0	41	17.6	81	16.8	121	16.9	161	17.1
2	16.6	42	17.5	82	16.7	122	17.1	162	17.1
3	16.3	43	16.5	83	16.4	123	16.8	163	17.1
4	16.1	44	17.8	84	16.5	124	17.0	164	17.4
5	17.1	45	17.3	85	16.4	125	17.2	165	17.2
6	16.9	46	17.3	86	16.6	126	17.3	166	16.9
7	16.8	47	17.1	87	16.5	127	17.2	167	16.9
8	17.4	48	17.4	88	16.7	128	17.3	168	17.0
9	17.1	49	16.9	89	16.4	129	17.2	169	16.7
10	17.0	50	17.3	90	16.4	130	17.2	170	16.9
11	16.7	51	17.6	91	16.2	131	17.5	171	17.3
12	17.4	52	16.9	92	16.4	132	16.9	172	17.8
13	17.2	53	16.7	93	16.3	133	16.9	173	17.8
14	17.4	54	16.8	94	16.4	134	16.9	174	17.6
15	17.4	55	16.8	95	17.0	135	17.0	175	17.5
16	17.0	56	17.2	96	16.9	136	16.5	176	17.0
17	17.3	57	16.8	97	17.1	137	16.7	177	16.9
18	17.2	58	17.6	98	17.1	138	16.8	178	17.1
19	17.4	59	17.2	99	16.7	139	16.7	179	17.2
20	16.8	60	16.6	100	16.9	140	16.7	180	17.4
21	17.1	61	17.1	101	16.5	141	16.6	181	17.5
22	17.4	62	16.9	102	17.2	142	16.5	182	17.9
23	17.4	63	16.6	103	16.4	143	17.0	183	17.0
24	17.5	64	18.0	104	17.0	144	16.7	184	17.0
25	17.4	65	17.2	105	17.0	145	16.7	185	17.0
26	17.6	66	17.3	106	16.7	146	16.9	186	17.2
27	17.4	67	17.0	107	16.2	147	17.4	187	17.3
28	17.3	68	16.9	108	16.6	148	17.1	188	17.4
29	17.0	69	17.3	109	16.9	149	17.0	189	17.4
30	17.8	70	16.8	110	16.5	150	16.8	190	17.0
31	17.5	71	17.3	111	16.6	151	17.2	191	18.0
32	18.1	72	17.4	112	16.6	152	17.2	192	18.2
33	17.5	73	17.7	113	17.0	153	17.4	193	17.6
34	17.4	74	16.8	114	17.1	154	17.2	194	17.8
35	17.4	75	16.9	115	17.1	155	16.9	195	17.7
36	17.1	76	17.0	116	16.7	156	16.8	196	17.2
37	17.6	77	16.9	117	16.8	157	17.0	197	17.4
38	17.7	78	17.0	118	16.3	158	17.4		
39	17.4	79	16.6	119	16.6	159	17.2		
40	17.8	80	16.7	120	16.8	160	17.2		

[a] 197 observations.

SERIES B IBM Common Stock Closing Prices: Daily, May 17, 1961–November 2, 1962[a]

460	471	527	580	551	523	333	394	330
457	467	540	579	551	516	330	393	340
452	473	542	584	552	511	336	409	339
459	481	538	581	553	518	328	411	331
462	488	541	581	557	517	316	409	345
459	490	541	577	557	520	320	408	352
463	489	547	577	548	519	332	393	346
479	489	553	578	547	519	320	391	352
493	485	559	580	545	519	333	388	357
490	491	557	586	545	518	344	396	
492	492	557	583	539	513	339	387	
498	494	560	581	539	499	350	383	
499	499	571	576	535	485	351	388	
497	498	571	571	537	454	350	382	
496	500	569	575	535	462	345	384	
490	497	575	575	536	473	350	382	
489	494	580	573	537	482	359	383	
478	495	584	577	543	486	375	383	
487	500	585	582	548	475	379	388	
491	504	590	584	546	459	376	395	
487	513	599	579	547	451	382	392	
482	511	603	572	548	453	370	386	
479	514	599	577	549	446	365	383	
478	510	596	571	553	455	367	377	
479	509	585	560	553	452	372	364	
477	515	587	549	552	457	373	369	
479	519	585	556	551	449	363	355	
475	523	581	557	550	450	371	350	
479	519	583	563	553	435	369	353	
476	523	592	564	554	415	376	340	
476	531	592	567	551	398	387	350	
478	547	596	561	551	399	387	349	
479	551	596	559	545	361	376	358	
477	547	595	553	547	383	385	360	
476	541	598	553	547	393	385	360	
475	545	598	553	537	385	380	366	
475	549	595	547	539	360	373	359	
473	545	595	550	538	364	382	356	
474	549	592	544	533	365	377	355	
474	547	588	541	525	370	376	367	
474	543	582	532	513	374	379	357	
465	540	576	525	510	359	386	361	
466	539	578	542	521	335	387	355	
467	532	589	555	521	323	386	348	
471	517	585	558	521	306	389	343	

[a] 369 observations (read down).

SERIES B′ IBM Common Stock Closing Prices: Daily, June 29, 1959–June 30, 1960[a]

445	425	406	441	415	461
448	421	407	437	420	463
450	414	410	427	420	463
447	410	408	423	424	461
451	411	408	424	426	465
453	406	409	428	423	473
454	406	410	428	423	473
454	413	409	431	425	475
459	411	405	425	431	499
440	410	406	423	436	485
446	405	405	420	436	491
443	409	407	426	440	496
443	410	409	418	436	504
440	405	407	416	443	504
439	401	409	419	445	509
435	401	425	418	439	511
435	401	425	416	443	524
436	414	428	419	445	525
435	419	436	425	450	541
435	425	442	421	461	531
435	423	442	422	471	529
433	411	433	422	467	530
429	414	435	417	462	531
428	420	433	420	456	527
425	412	435	417	464	525
427	415	429	418	463	519
425	412	439	419	465	514
422	412	437	419	464	509
409	411	439	417	456	505
407	412	438	419	460	513
423	409	435	422	458	525
422	407	433	423	453	519
417	408	437	422	453	519
421	415	437	421	449	522
424	413	444	421	447	522
414	413	441	419	453	
419	410	440	418	450	
429	405	441	421	459	
426	410	439	420	457	
425	412	439	413	453	
424	413	438	413	455	
425	411	437	408	453	
425	411	441	409	450	
424	409	442	415	456	

[a] 255 observations (read down).

SERIES C Chemical Process Temperature Readings: Every Minute[a]

26.6	19.6	24.4	21.1	24.4
27.0	19.6	24.4	20.9	24.2
27.1	19.6	24.4	20.8	24.2
27.1	19.6	24.4	20.8	24.1
27.1	19.6	24.5	20.8	24.1
27.1	19.7	24.5	20.8	24.0
26.9	19.9	24.4	20.9	24.0
26.8	20.0	24.3	20.8	24.0
26.7	20.1	24.2	20.8	23.9
26.4	20.2	24.2	20.7	23.8
26.0	20.3	24.0	20.7	23.8
25.8	20.6	23.9	20.8	23.7
25.6	21.6	23.7	20.9	23.7
25.2	21.9	23.6	21.2	23.6
25.0	21.7	23.5	21.4	23.7
24.6	21.3	23.5	21.7	23.6
24.2	21.2	23.5	21.8	23.6
24.0	21.4	23.5	21.9	23.6
23.7	21.7	23.5	22.2	23.5
23.4	22.2	23.7	22.5	23.5
23.1	23.0	23.8	22.8	23.4
22.9	23.8	23.8	23.1	23.3
22.8	24.6	23.9	23.4	23.3
22.7	25.1	23.9	23.8	23.3
22.6	25.6	23.8	24.1	23.4
22.4	25.8	23.7	24.6	23.4
22.2	26.1	23.6	24.9	23.3
22.0	26.3	23.4	24.9	23.2
21.8	26.3	23.2	25.1	23.3
21.4	26.2	23.0	25.0	23.3
20.9	26.0	22.8	25.0	23.2
20.3	25.8	22.6	25.0	23.1
19.7	25.6	22.4	25.0	22.9
19.4	25.4	22.0	24.9	22.8
19.3	25.2	21.6	24.8	22.6
19.2	24.9	21.3	24.7	22.4
19.1	24.7	21.2	24.6	22.2
19.0	24.5	21.2	24.5	21.8
18.9	24.4	21.1	24.5	21.3
18.9	24.4	21.0	24.5	20.8
19.2	24.4	20.9	24.5	20.2
19.3	24.4	21.0	24.5	19.7
19.3	24.4	21.0	24.5	19.3
19.4	24.3	21.1	24.5	19.1
19.5	24.4	21.2	24.4	19.0
				18.8

[a]226 observations (read down).

SERIES D Chemical Process Viscosity Readings: Every Hour[a]

8.0	8.8	9.3	9.1	9.0	10.0	9.6
8.0	8.6	9.9	9.5	9.0	9.8	8.6
7.4	8.6	9.7	9.4	9.4	9.8	8.0
8.0	8.4	9.1	9.5	9.0	9.7	8.0
8.0	8.3	9.3	9.6	9.0	9.6	8.0
8.0	8.4	9.5	10.2	9.4	9.4	8.0
8.0	8.3	9.4	9.8	9.4	9.2	8.4
8.8	8.3	9.0	9.6	9.6	9.0	8.8
8.4	8.1	9.0	9.6	9.4	9.4	8.4
8.4	8.2	8.8	9.4	9.6	9.6	8.4
8.0	8.3	9.0	9.4	9.6	9.6	9.0
8.2	8.5	8.8	9.4	9.6	9.6	9.0
8.2	8.1	8.6	9.4	10.0	9.6	9.4
8.2	8.1	8.6	9.6	10.0	9.6	10.0
8.4	7.9	8.0	9.6	9.6	9.6	10.0
8.4	8.3	8.0	9.4	9.2	9.0	10.0
8.4	8.1	8.0	9.4	9.2	9.4	10.2
8.6	8.1	8.0	9.0	9.2	9.4	10.0
8.8	8.1	8.6	9.4	9.0	9.4	10.0
8.6	8.4	8.0	9.4	9.0	9.6	9.6
8.6	8.7	8.0	9.6	9.6	9.4	9.0
8.6	9.0	8.0	9.4	9.8	9.6	9.0
8.6	9.3	7.6	9.2	10.2	9.6	8.6
8.6	9.3	8.6	8.8	10.0	9.8	9.0
8.8	9.5	9.6	8.8	10.0	9.8	9.6
8.9	9.3	9.6	9.2	10.0	9.8	9.6
9.1	9.5	10.0	9.2	9.4	9.6	9.0
9.5	9.5	9.4	9.6	9.2	9.2	9.0
8.5	9.5	9.3	9.6	9.6	9.6	8.9
8.4	9.5	9.2	9.8	9.7	9.2	8.8
8.3	9.5	9.5	9.8	9.7	9.2	8.7
8.2	9.5	9.5	10.0	9.8	9.6	8.6
8.1	9.9	9.5	10.0	9.8	9.6	8.3
8.3	9.5	9.9	9.4	9.8	9.6	7.9
8.4	9.7	9.9	9.8	10.0	9.6	8.5
8.7	9.1	9.5	8.8	10.0	9.6	8.7
8.8	9.1	9.3	8.8	8.6	9.6	8.9
8.8	8.9	9.5	8.8	9.0	10.0	9.1
9.2	9.3	9.5	8.8	9.4	10.0	9.1
9.6	9.1	9.1	9.6	9.4	10.4	9.1
9.0	9.1	9.3	9.6	9.4	10.4	
8.8	9.3	9.5	9.6	9.4	9.8	
8.6	9.5	9.3	9.2	9.4	9.0	
8.6	9.3	9.1	9.2	9.6	9.6	
8.8	9.3	9.3	9.0	10.0	9.8	

[a] 310 observations (read down).

SERIES E Wölfer Sunspot Numbers: Yearly[a]

1770	101	1795	21	1820	16	1845	40
1771	82	1796	16	1821	7	1846	62
1772	66	1797	6	1822	4	1847	98
1773	35	1798	4	1823	2	1848	124
1774	31	1799	7	1824	8	1849	96
1775	7	1800	14	1825	17	1850	66
1776	20	1801	34	1826	36	1851	64
1777	92	1802	45	1827	50	1852	54
1778	154	1803	43	1828	62	1853	39
1779	125	1804	48	1829	67	1854	21
1780	85	1805	42	1830	71	1855	7
1781	68	1806	28	1831	48	1856	4
1782	38	1807	10	1832	28	1857	23
1783	23	1808	8	1833	8	1858	55
1784	10	1809	2	1834	13	1859	94
1785	24	1810	0	1835	57	1860	96
1786	83	1811	1	1836	122	1861	77
1787	132	1812	5	1837	138	1862	59
1788	131	1813	12	1838	103	1863	44
1789	118	1814	14	1839	86	1864	47
1790	90	1815	35	1840	63	1865	30
1791	67	1816	46	1841	37	1866	16
1792	60	1817	41	1842	24	1867	7
1793	47	1818	30	1843	11	1868	37
1794	41	1819	24	1844	15	1869	74

[a] 100 observations.

SERIES F Yields from a Batch Chemical Process: Consecutive[a]

47	44	50	62	68
64	80	71	44	38
23	55	56	64	50
71	37	74	43	60
38	74	50	52	39
64	51	58	38	59
55	57	45	59	40
41	50	54	55	57
59	60	36	41	54
48	45	54	53	23
71	57	48	49	
35	50	55	34	
57	45	45	35	
40	25	57	54	
58	59	50	45	

[a] 70 Observations (read down).

SERIES G International Airline Passengers: Monthly Totals (Thousands of Passengers) January 1949–December 1960[a]

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	264	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	284	277	317	313	318	374	413	405	355	306	271	306
1957	315	301	356	348	355	422	465	467	404	347	305	336
1958	340	318	362	348	363	435	491	505	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	405
1960	417	391	419	461	472	535	622	606	508	461	390	432

[a] 144 observations.

SERIES J Series J Gas Furnace Data[a]

t	X_t	Y_t	t	X_t	Y_t	t	X_t	Y_t
1	−0.109	53.8	51	1.608	46.9	101	−0.288	51.0
2	0.000	53.6	52	1.905	47.8	102	−0.153	51.8
3	0.178	53.5	53	2.023	48.2	103	−0.109	52.4
4	0.339	53.5	54	1.815	48.3	104	−0.187	53.0
5	0.373	53.4	55	0.535	47.9	105	−0.255	53.4
6	0.441	53.1	56	0.122	47.2	106	−0.229	53.6
7	0.461	52.7	57	0.009	47.2	107	−0.007	53.7
8	0.348	52.4	58	0.164	48.1	108	0.254	53.8
9	0.127	52.2	59	0.671	49.4	109	0.330	53.8
10	−0.180	52.0	60	1.019	50.6	110	0.102	53.8
11	−0.588	52.0	61	1.146	51.5	111	−0.423	53.3
12	−1.055	52.4	62	1.155	51.6	112	−1.139	53.0
13	−1.421	53.0	63	1.112	51.2	113	−2.275	52.9
14	−1.520	54.0	64	1.121	50.5	114	−2.594	53.4
15	−1.302	54.9	65	1.223	50.1	115	−2.716	54.6
16	−0.814	56.0	66	1.257	49.8	116	−2.510	56.4
17	−0.475	56.8	67	1.157	49.6	117	−1.790	58.0
18	−0.193	56.8	68	0.913	49.4	118	−1.346	59.4
19	0.088	56.4	69	0.620	49.3	119	−1.081	60.2
20	0.435	55.7	70	0.255	49.2	120	−0.910	60.0
21	0.771	55.0	71	−0.280	49.3	121	−0.876	59.4
22	0.866	54.3	72	−1.080	49.7	122	−0.885	58.4
23	0.875	53.2	73	−1.551	50.3	123	−0.800	57.6
24	0.891	52.3	74	−1.799	51.3	124	−0.544	56.9
25	0.987	51.6	75	−1.825	52.8	125	−0.416	56.4
26	1.263	51.2	76	−1.456	54.4	126	−0.271	56.0
27	1.775	50.8	77	−0.944	56.0	127	0.000	55.7
28	1.976	50.5	78	−0.570	56.9	128	0.403	55.3
29	1.934	50.0	79	−0.431	57.5	129	0.841	55.0

SERIES J (*continued*)

t	X_t	Y_t	t	X_t	Y_t	t	X_t	Y_t
30	1.866	49.2	80	−0.577	57.3	130	1.285	54.4
31	1.832	48.4	81	−0.960	56.6	131	1.607	53.7
32	1.767	47.9	82	−1.616	56.0	132	1.746	52.8
33	1.608	47.6	83	−1.875	55.4	133	1.683	51.6
34	1.265	47.5	84	−1.891	55.4	134	1.485	50.6
35	0.790	47.5	85	−1.746	56.4	135	0.993	49.4
36	0.360	47.6	86	−1.474	57.2	136	0.648	48.8
37	0.115	48.1	87	−1.201	58.0	137	0.577	48.5
38	0.088	49.0	88	−0.927	58.4	138	0.577	48.7
39	0.331	50.0	89	−0.524	58.4	139	0.632	49.2
40	0.645	51.1	90	0.040	58.1	140	0.747	49.8
41	0.960	51.8	91	0.788	57.7	141	0.900	50.4
42	1.409	51.9	92	0.943	57.0	142	0.993	50.7
43	2.670	51.7	93	0.930	56.0	143	0.968	50.9
44	2.834	51.2	94	1.006	54.7	144	0.790	50.7
45	2.812	50.0	95	1.137	53.2	145	0.399	50.5
46	2.483	48.3	96	1.198	52.1	146	−0.161	50.4
47	1.929	47.0	97	1.054	51.6	147	−0.553	50.2
48	1.485	45.8	98	0.595	51.0	148	−0.603	50.4
49	1.214	45.6	99	−0.080	50.5	149	−0.424	51.2
50	1.239	46.0	100	−0.314	50.4	150	−0.194	52.3
151	−0.049	53.2	201	−2.473	55.6	251	0.185	56.3
152	0.060	53.9	202	−2.330	58.0	252	0.662	56.4
153	0.161	54.1	203	−2.053	59.5	253	0.709	56.4
154	0.301	54.0	204	−1.739	60.0	254	0.605	56.0
155	0.517	53.6	205	−1.261	60.4	255	0.501	55.2
156	0.566	53.2	206	−0.569	60.5	256	0.603	54.0
157	0.560	53.0	207	−0.137	60.2	257	0.943	53.0
158	0.573	52.8	208	−0.024	59.7	258	1.223	52.0
159	0.592	52.3	209	−0.050	59.0	259	1.249	51.6
160	0.671	51.9	210	−0.135	57.6	260	0.824	51.6
161	0.933	51.6	211	−0.276	56.4	261	0.102	51.1
162	1.337	51.6	212	−0.534	55.2	262	0.025	50.4
163	1.460	51.4	213	−0.871	54.5	263	0.382	50.0
164	1.353	51.2	214	−1.243	54.1	264	0.922	50.0
165	0.772	50.7	215	−1.439	54.1	265	1.032	52.0
166	0.218	50.0	216	−1.422	54.4	266	0.866	54.0
167	−0.237	49.4	217	−1.175	55.5	267	0.527	55.1
168	−0.714	49.3	218	−0.813	56.2	268	0.093	54.5
169	−1.099	49.7	219	−0.634	57.0	269	−0.458	52.8
170	−1.269	50.6	220	−0.582	57.3	270	−0.748	51.4
171	−1.175	51.8	221	−0.625	57.4	271	−0.947	50.8
172	−0.676	53.0	222	−0.713	57.0	272	−1.029	51.2
173	0.033	54.0	223	−0.848	56.4	273	−0.928	52.0
174	0.556	55.3	224	−1.039	55.9	274	−0.645	52.8
175	0.643	55.9	225	−1.346	55.5	275	−0.424	53.8
176	0.484	55.9	226	−1.628	55.3	276	−0.276	54.5
177	0.109	54.6	227	−1.619	55.2	277	−0.158	54.9
178	−0.310	53.5	228	−1.149	55.4	278	−0.033	54.9

SERIES J (*continued*)

t	X_t	Y_t	t	X_t	Y_t	t	X_t	Y_t
179	−0.697	52.4	229	−0.488	56.0	279	0.102	54.8
180	−1.047	52.1	230	−0.160	56.5	280	0.251	54.4
181	−1.218	52.3	231	−0.007	57.1	281	0.280	53.7
182	−1.183	53.0	232	−0.092	57.3	282	0.000	53.3
183	−0.873	53.8	233	−0.620	56.8	283	−0.493	52.8
184	−0.336	54.6	234	−1.086	55.6	284	−0.759	52.6
185	0.063	55.4	235	−1.525	55.0	285	−0.824	52.6
186	0.084	55.9	236	−1.858	54.1	286	−0.740	53.0
187	0.000	55.9	237	−2.029	54.3	287	−0.528	54.3
188	0.001	55.2	238	−2.024	55.3	288	−0.204	56.0
189	0.209	54.4	239	−1.961	56.4	289	0.034	57.0
190	0.556	53.7	240	−1.952	57.2	290	0.204	58.0
191	0.782	53.6	241	−1.794	57.8	291	0.253	58.6
192	0.858	53.6	242	−1.302	58.3	292	0.195	58.5
193	0.918	53.2	243	−1.030	58.6	293	0.131	58.3
194	0.862	52.5	244	−0.918	58.8	294	0.017	57.8
195	0.416	52.0	245	−0.798	58.8	295	−0.182	57.3
196	−0.336	51.4	246	−0.867	58.6	296	−0.262	57.0
197	−0.959	51.0	247	−1.047	58.0			
198	−1.813	50.9	248	−1.123	57.4			
199	−2.378	52.4	249	−0.876	57.0			
200	−2.499	53.5	250	−0.395	56.4			

[a] Sampling interval 9 seconds; observations for 296 pairs of data points. X, 0.60 – 0.04 (input gas rate in cubic feet per minute); Y, $\%CO_2$ in outlet gas.

SERIES K Simulated Dynamic Data with Two Inputs[a]

t	X_{1t}	X_{2t}	Y_t	t	X_{1t}	X_{2t}	Y_t
−2	0	0	58.3				
−1			61.8				
0			64.2	30			65.8
1			62.1	31			67.4
2	−1	1	55.1	32	−1	−1	64.7
3			50.6	33			65.7
4			47.8	34			67.5
5			49.7	35			58.2
6			51.6	36			57.0
7	1	−1	58.5	37	−1	1	54.7
8			61.5	38			54.9
9			63.3	39			48.4
10			65.9	40			49.7
11			70.9	41			53.1
12	−1	−1	65.8	42	1	−1	50.2
13			57.6	43			51.7
14			56.1	44			57.4
15			58.2	45			62.6
16			61.7	46			65.8
17	1	1	59.2	47	−1	−1	61.5
18			57.9	48			61.5
19			61.3	49			56.8
20			60.8	50			62.3
21			63.6	51			57.7
22	1	−1	69.5	52	−1	1	54.0
23			69.3	53			45.2
24			70.5	54			51.9
25			68.0	55			45.6
26			68.1	56			46.2
27	1	1	65.0	57	1	1	50.2
28			71.9	58			54.6
29			64.8	59			55.6
				60	0	0	60.4
				61			59.4

[a] 64 observations.

SERIES L Pilot Scheme Data[a]

t	x_t	ε_t	t	x_t	ε_t	t	x_t	ε_t
1	30	−4	53	−60	6	105	55	−4
2	0	−2	54	50	−2	106	0	2
3	−10	0	55	−10	0	107	−90	8
4	0	0	56	40	−4	108	40	0
5	−40	4	57	40	−6	109	0	0
6	0	2	58	−30	0	110	80	−8
7	−10	2	59	20	−2	111	−20	−2
8	10	0	60	−30	2	112	−10	0
9	20	−2	61	10	0	113	−70	6
10	50	−6	62	−20	2	114	−30	6
11	−10	−2	63	30	−2	115	−10	4
12	−55	4	64	−50	4	116	30	−1
13	0	2	65	10	−2	117	−5	0
14	10	0	66	10	−2	118	−60	6
15	0	−2	67	10	−2	119	70	−4
16	10	−2	68	−30	0	120	40	−6
17	−70	6	69	0	0	121	10	−4
18	30	0	70	−10	2	122	20	−4
19	−20	2	71	−10	3	123	10	−3
20	10	0	72	15	0	124	0	−2
21	0	0	73	20	−2	125	−70	6
22	0	0	74	−50	4	126	50	−2
23	20	−2	75	20	0	127	30	−4
24	30	−4	76	0	0	128	0	−2
25	0	−2	77	0	0	129	−10	0
26	−10	0	78	0	0	130	0	0
27	−20	2	79	0	0	131	−40	4
28	−30	4	80	−40	4	132	0	2
29	0	2	81	−100	12	133	−10	2
30	10	0	82	0	8	134	10	0
31	20	−2	83	0	−12	135	0	0
32	−10	0	84	50	−15	136	80	−8
33	0	0	85	85	−15	137	−80	4
34	20	−2	86	5	−12	138	20	4
35	10	−2	87	40	−14	139	20	0
36	−10	0	88	10	−8	140	−10	2
37	0	0	89	−60	2	141	10	0
38	0	0	90	−50	6	142	0	0
39	0	0	91	−50	8	143	−20	2
40	0	0	92	40	0	144	20	−1
41	0	0	93	0	0	145	55	−6
42	0	0	94	0	0	146	0	−3
43	20	−2	95	−20	2	147	25	−4
44	−50	4	96	−30	4	148	20	−4
45	20	0	97	−60	8	149	−60	4
46	0	0	98	−20	6	150	−40	6
47	0	0	99	−30	6	151	10	4
48	40	−4	100	30	0	152	20	0

SERIES L (*continued*)

t	x_t	ε_t	t	x_t	ε_t	t	x_t	ε_t
49	0	−2	101	−40	4	153	60	−6
50	50	−6	102	80	−6	154	−50	2
51	−40	0	103	−40	0	155	−10	2
52	−50	3	104	−20	2	156	−30	4
157	20	0	209	−40	4	261	−25	4
158	0	0	210	40	−2	262	35	−2
159	20	−2	211	−90	8	263	70	8
160	10	−2	212	40	0	264	−10	−5
161	10	−2	213	0	0	265	100	−20
162	10	−22	214	0	0	266	−20	−8
163	50	−6	215	0	0	267	−40	0
164	−30	0	216	20	−2	268	−20	2
165	−30	6	217	90	−10	269	10	0
166	90	12	218	30	−8	270	0	0
167	60	0	219	20	−6	271	0	0
168	−40	4	220	30	−6	272	−20	2
169	20	0	221	30	−6	273	−50	6
170	0	0	222	30	−6	274	50	−2
171	20	−2	223	30	−6	275	30	−4
172	10	−2	224	−90	6	276	60	−8
173	−30	2	225	10	2	277	−40	0
174	−30	4	226	10	2	278	−20	2
175	0	2	227	−30	4	279	−10	2
176	50	−4	228	−20	4	280	10	0
177	−60	4	229	40	−2	281	−110	13
178	20	0	230	10	−2	282	15	4
179	0	0	231	10	−2	283	30	−2
180	40	−8	232	10	−2	284	0	−1
181	80	−12	233	−100	12	285	25	−3
182	20	−8	234	10	6	286	−5	−1
183	−100	6	235	45	−2	287	−15	1
184	−30	6	236	30	−4	288	45	−4
185	30	0	237	30	−5	289	40	−6
186	−20	2	238	−15	−1	290	−50	2
187	−30	4	239	−5	0	291	−10	2
188	20	0	240	10	−1	292	−50	6
189	60	−6	241	−85	8	293	20	1
190	−10	−2	242	0	4	294	5	0
191	30	−4	243	0	0	295	−40	4
192	−40	2	244	60	−4	296	0	6
193	30	−2	245	40	−6	297	−60	8
194	−20	1	246	−30	0	298	40	0
195	5	0	247	−40	4	299	−20	2
196	−20	2	248	−40	6	300	130	−12
197	−30	4	249	50	−2	301	−20	−4
198	20	0	250	10	−2	302	0	−2
199	10	−1	251	30	−4	303	30	−4

SERIES L (*continued*)

t	x_t	ε_t	t	x_t	ε_t	t	x_t	ε_t
200	−15	1	252	−40	2	304	−20	0
201	−75	8	253	10	0	305	60	6
202	−40	8	254	−40	4	306	10	−4
203	−40	6	255	40	−2	307	−10	1
204	90	−6	256	−30	2	308	−25	2
205	90	−12	257	−50	6	309	0	1
206	80	−14	258	0	3	310	15	−1
207	−45	−2	259	−45	6	311	−5	0
208	−10	0	260	−20	5	312	0	0

[a] 312 observations.

SERIES M Sales Data with Leading Indicator[a]

t	Leading Indicator X_t	Sales Y_t	t	Leading Indicator X_t	Sales Y_t	t	Leading Indicator X_t	Sales Y_t
1	10.01	200.1	51	10.77	220.0	101	12.90	249.4
2	10.07	199.5	52	10.88	218.7	102	13.12	249.0
3	10.32	199.4	53	10.49	217.0	103	12.47	249.9
4	9.75	198.9	54	10.50	215.9	104	12.47	250.5
5	10.33	199.0	55	11.00	215.8	105	12.94	251.5
6	10.13	200.2	56	10.98	214.1	106	13.10	249.0
7	10.36	198.6	57	10.61	212.3	107	12.91	247.6
8	10.32	200.0	58	10.48	213.9	108	13.39	248.8
9	10.13	200.3	59	10.53	214.6	109	13.13	250.4
10	10.16	201.2	60	11.07	213.6	110	13.34	250.7
11	10.58	201.6	61	10.61	212.1	111	13.34	253.0
12	10.62	2013	62	10.86	211.4	112	13.14	253.7
13	10.86	201.5	63	10.34	213.1	113	13.49	255.0
14	11.20	203.5	64	10.78	212.9	114	13.87	256.2
15	10.74	204.9	65	10.80	213.3	115	13.39	256.0
16	10.56	207.1	66	10.33	211.5	116	13.59	257.4
17	10.48	210.5	67	10.44	212.3	117	13.27	260.4
18	10.77	210.5	68	10.50	213.0	118	13.70	260.0
19	11.33	209.8	69	10.75	211.0	119	13.20	261.3
20	10.96	208.8	70	10.40	210.7	120	13.32	260.4
21	11.16	209.5	71	10.40	210.1	121	13.15	261.6
22	11.70	213.2	72	10.34	211.4	122	13.30	260.8
23	11.39	213.7	73	10.55	210.0	123	12.94	259.8
24	11.42	215.1	74	10.46	209.7	124	13.29	259.0
25	11.94	218.7	75	10.82	208.8	125	13.26	258.9
26	11.24	219.8	76	10.91	208.8	126	13.08	257.4
27	11.59	220.5	77	10.87	208.8	127	13.24	257.7
28	10.96	223.8	78	10.67	210.6	128	13.31	257.9
29	11.40	222.8	79	11.11	211.9	129	13.52	257.4
30	11.02	223.8	80	10.88	212.8	130	13.02	257.3
31	11.01	221.7	81	11.28	212.5	131	13.25	257.6
32	11.23	222.3	82	11.27	214.8	132	13.12	258.9
33	11.33	220.8	83	11.44	215.3	133	13.26	257.8
34	10.83	219.4	84	11.52	217.5	134	13.11	257.7
35	10.84	220.1	85	12.10	218.8	135	13.30	257.2
36	11.14	220.6	86	11.83	220.7	136	13.06	257.5
37	10.38	218.9	87	12.62	222.2	137	13.32	256.8
38	10.90	217.8	88	12.41	226.7	138	13.10	257.5
39	11.05	217.7	89	12.43	228.4	139	13.27	257.0
40	11.11	215.0	90	12.73	233.2	140	13.64	257.6
41	11.01	215.3	91	13.01	235.7	141	13.58	257.3
42	11.22	215.9	92	12.74	237.1	142	13.87	257.5
43	11.21	216.7	93	12.73	240.6	143	13.53	259.6
44	11 91	216.7	94	12.76	243.8	144	13.41	261.1
45	11.69	217.7	95	12.92	245.3	145	13.25	262.9
46	10.93	218.7	96	12.64	246.0	146	13.50	263.3
47	10.99	222.9	97	12.79	246.3	147	13.58	262.8
48	11.01	224.9	98	13.05	247.7	148	13.51	261.8
49	10.84	222.2	99	12.69	247.6	149	13.77	262.2
50	10.76	220.7	100	13.01	247.8	150	13.40	262.7

[a] 150 observations.

SERIES N Mink Fur Sales of the Hudson's Bay Company: Annual for 1850–1911[a]

1850	29,619	1866	51,404	1882	45,600	1897	76,365
1851	21,151	1867	58,451	1883	47,508	1898	70,407
1852	24,859	1868	73,575	1884	52,290	1899	41,839
1853	25,152	1869	74,343	1885	110,824	1900	45,978
1854	42,375	1870	27,708	1886	76,503	1901	47,813
1855	50,839	1871	31,985	1887	64,303	1902	57,620
1856	61,581	1872	39,266	1888	83,023	1903	66,549
1857	61,951	1873	44,740	1889	40,748	1904	54,673
1858	76,231	1874	60,429	1890	35,596	1905	55,996
1859	63,264	1875	72,273	1891	29,479	1906	60,053
1860	44,730	1876	79,214	1892	42,264	1907	39,169
1861	31,094	1877	79,060	1893	58,171	1908	21,534
1862	49,452	1878	84,244	1894	50,815	1909	17,857
1863	43,961	1879	62,590	1895	51,285	1910	21,788
1864	61,727	1880	35,072	1896	70,229	1911	33,008
1865	60,334	1881	36,160				

[a]62 observations.

SERIES P Unemployment and GDP Data in UK: Quarterly for 1955–1969[a]

		UN	GDP			UN	GDP			UN	GDP
1955	1	225	81.37	1960	1	363	92.30	1965	1	306	108.07
	2	208	82.60		2	342	92.13		2	304	107.64
	3	201	82.30		3	325	93.17		3	321	108.87
	4	199	83.00		4	312	93.50		4	305	109.75
1956	1	207	82.87	1961	1	291	94.77	1966	1	279	110.20
	2	215	83.60		2	293	95.37		2	282	110.20
	3	240	83.33		3	304	95.03		3	318	110.90
	4	245	83.53		4	330	95.23		4	414	110.40
1957	1	295	84.27	1962	1	357	95.07	1967	1	463	111.00
	2	293	85.50		2	401	96.40		2	506	112.10
	3	279	84.33		3	447	96.97		3	538	112.50
	4	287	84.30		4	483	96.50		4	536	113.00
1958	1	331	85.07	1963	1	535	96.16	1968	1	544	114.30
	2	396	83.60		2	520	99.79		2	541	115.10
	3	432	84.37		3	489	101.14		3	547	116.40
	4	462	84.50		4	456	102.95		4	532	117.80
1959	1	454	85.20	1964	1	386	103.96	1969	1	532	116.80
	2	446	87.07		2	368	105.28		2	519	117.80
	3	426	88.40		3	358	105.81		3	547	119.00
	4	402	90.03		4	330	107.14		4	544	119.60

Source: Bray (1971).

[a]60 pairs of data; data are seasonally adjusted; unemployment (UN) in thousands; gross domestic product (GDP) is composite estimate (1963 = 100).

SERIES Q Logged and Coded U.S. Hog Price Data: Annual for 1867–1948[a]

1867	597	1888	709	1909	810	1929	1112
1868	509	1889	763	1910	957	1930	1129
1869	663	1890	681	1911	970	1931	1055
1870	751	1891	627	1912	903	1932	787
1871	739	1892	667	1913	995	1933	624
1872	598	1893	804	1914	1022	1934	612
1873	556	1894	782	1915	998	1935	800
1874	594	1895	707	1916	928	1936	1104
1875	667	1896	653	1917	1073	1937	1075
1876	776	1897	639	1918	1294	1938	1052
1877	754	1898	672	1919	1346	1939	1048
1878	689	1899	669	1920	1301	1940	891
1879	498	1900	729	1921	1134	1941	921
1880	643	1901	784	1922	1024	1942	1193
1881	681	1902	842	1923	1090	1943	1352
1882	778	1903	886	1924	1013	1944	1243
1883	829	1904	784	1925	1119	1945	1314
1884	751	1905	770	1926	1195	1946	1380
1885	704	1906	783	1927	1235	1947	1556
1886	633	1907	877	1928	1120	1948	1632
1887	663	1908	777				

Source: Quenouille (1957).

[a]82 observations; values are $1000 \log_{10}(H_t)$, where H_t is the price, in dollars, per head on January 1 of the year.

SERIES R Monthly Averages of Hourly Readings of Ozone in Downtown Los Angeles[a]

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1955	2.63	1.94	3.38	4.92	6.29	5.58	5.50	4.71	6.04	7.13	7.79	3.83
1956	3.83	4.25	5.29	3.75	4.67	5.42	6.04	5.71	8.13	4.88	5.42	5.50
1957	3.00	3.42	4.50	4.25	4.00	5.33	5.79	6.58	7.29	5.04	5.04	4.48
1958	3.33	2.88	2.50	3.83	4.17	4.42	4.25	4.08	4.88	4.54	4.25	4.21
1959	2.75	2.42	4.50	5.21	4.00	7.54	7.38	5.96	5.08	5.46	4.79	2.67
1960	1.71	1.92	3.38	3.98	4.63	4.88	5.17	4.83	5.29	3.71	2.46	2.17
1961	2.15	2.44	2.54	3.25	2.81	4.21	4.13	4.17	3.75	3.83	2.42	2.17
1962	2.33	2.00	2.13	4.46	3.17	3.25	4.08	5.42	4.50	4.88	2.83	2.75
1963	1.63	3.04	2.58	2.92	3.29	3.71	4.88	4.63	4.83	3.42	2.38	2.33
1964	1.50	2.25	2.63	2.96	3.46	4.33	5.42	4.79	4.38	4.54	2.04	1.33
1965	2.04	2.81	2.67	4.08	3.90	3.96	4.50	5.58	4.52	5.88	3.67	1.79
1966	1.71	1.92	3.58	4.40	3.79	5.52	5.50	5.00	5.48	4.81	2.42	1.46
1967	1.71	2.46	2.42	1.79	3.63	3.54	4.88	4.96	3.63	5.46	3.08	1.75
1968	2.13	2.58	2.75	3.15	3.46	3.33	4.67	4.13	4.73	3.42	3.08	1.79
1969	1.96	1.63	2.75	3.06	4.31	3.31	3.71	5.25	3.67	3.10	2.25	2.29
1970	1.25	2.25	2.67	3.23	3.58	3.04	3.75	4.54	4.46	2.83	1.63	1.17
1971	1.79	1.92	2.25	2.96	2.38	3.38	3.38	3.21	2.58	2.42	1.58	1.21
1972	1.42	1.96	3.04	2.92	3.58	3.33	4.04	3.92	3.08	2.00	1.58	1.21

[a]216 observations; values are in pphm.

REFERENCES

Abraham, B. (1981). Missing observations in time series, *Commun. Stat.*, **A10**, 1643–1653.

Abraham, B. and Box, G. E. P. (1978). Deterministic and forecast-adaptive time-dependent models, *Appl. Stat.*, **27**, 120–130.

Adler, J. (2010). *R in a Nutshell*, O'Reilly Media, Sebastopol, CA.

Ahn, S. K. (1993). Some tests for unit roots in autoregressive-integrated-moving average models with deterministic trends, *Biometrika*, **80**, 855–868.

Ahn, S. K. and Reinsel, G. C. (1988). Nested reduced-rank autoregressive models for multiple time series, *J. Am. Stat. Assoc.*, **83**, 849–856.

Ahn, S. K. and Reinsel, G. C. (1990). Estimation for partially nonstationary multivariate autoregressive models, *J. Am. Stat. Assoc.*, **85**, 813–823.

Akaike, H. (1971). Autoregressive model fitting for control, *Ann. Inst. Stat. Math.*, **23**, 163–180.

Akaike, H. (1974a). A new look at the statistical model identification, *IEEE Trans. Autom. Control*, **AC-19**, 716–723.

Akaike, H. (1974b). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes, *Ann. Inst. Stat. Math.*, **26**, 363–387.

Akaike, H. (1976). Canonical correlation analysis of time series and the use of an information criterion, in *Systems Identification: Advances and Case Studies* (eds. R. K. Mehra and D. G. Lainiotis), Academic Press, New York, pp. 27–96.

Ali, M. M. (1989). Tests for autocorrelation and randomness in multiple time series, *J. Am. Stat. Assoc.*, **84**, 533–540.

Andersen, T. G., Davis, R. A., Kreiss, J.-P., and Mikosch, T. (2009). *Handbook of Financial Time Series*, Springer, Berlin.

Anderson, B. D. O. and Moore, J. B. (1979). *Optimal Filtering*, Prentice Hall, Englewood Cliffs, NJ.

Anderson, R. L. (1942). Distribution of the serial correlation coefficient, *Ann. Math. Stat.*, **13**, 1–13.

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions, *Ann. Math. Stat.*, **22**, 327–351.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series*, Wiley, New York.

Anscombe, F. J. (1961). Examination of residuals, *Proceedings of the 4th Berkeley Symposium*, Vol. 1, pp. 1–36.

Anscombe, F. J. and Tukey, J. W. (1963). The examination and analysis of residuals, *Technometrics*, **5**, 141–160.

Ansley, C. F. (1979). An algorithm for the exact likelihood of a mixed autoregressive moving average process, *Biometrika*, **66**, 59–65.

Ansley, C. F. and Kohn, R. (1982). A geometrical derivation of the fixed interval smoothing algorithm, *Biometrika*, **69**, 486–487.

Ansley, C. F. and Kohn, R. (1983). Exact likelihood of vector autoregressive-moving average process with missing or aggregated data, *Biometrika*, **70**, 275–278.

Ansley, C. F. and Kohn, R. (1985). Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions, *Ann. Stat.*, **13**, 1286–1316.

Ansley, C. F. and Newbold, P. (1980). Finite sample properties of estimators for autoregressive moving average models, *J. Econom.*, **13**, 159–183.

Aström, K. J. and Bohlin, T. (1966). Numerical identification of linear dynamic systems from normal operating records, in *Theory of Self-Adaptive Control Systems* (ed. P. H. Hammond), Plenum Press, New York, pp. 96–111.

Aström, K. J. and Wittenmark, B. (1984). *Computer Controlled Systems*, Prentice Hall, Englewood Cliffs, NJ.

Athanasopoulos, G., Poskitt, D. S., and Vahid, F. (2012). Two canonical VARMA forms: Scalar component models vis-à-vis the Echelon form, *Econometric Rev.*,**31**, 60–83.

Bachelier, L. (1900). Théorie de la spéculation, *Ann. Sci. École Norm. Sup., Paris, Ser. 3*, **17**, 21–86.

Bagshaw, M. and Johnson, R. A. (1977). Sequential procedures for detecting parameter changes in a time-series model, *J. Am. Stat. Assoc.*, **72**, 593–597.

Baillie, R. T. (1979). The asymptotic mean squared error of multistep prediction from the regression model with autoregressive errors, *J. Am. Stat. Assoc.*, **74**, 175–184.

Baillie, R. T. and Bollerslev, T. (1992). Prediction in dynamic models with time-dependent conditional variances, *J. Econom.*, **52**, 91–113.

Barnard, G. A. (1949). Statistical inference, *J. R. Stat. Soc.*, **B11**, 115–139.

Barnard, G. A. (1959). Control charts and stochastic processes, *J. R. Stat. Soc.*, B21, 239–257.

Barnard, G. A. (1963). The logic of least squares, *J. R. Stat. Soc.*, **B25**, 124–127.

Barnard, G. A., Jenkins, G. M., and Winsten, C. B. (1962). Likelihood inference and time series, *J. R. Stat. Soc.*, **A125**, 321–352.

Bartlett, M. S. (1946). On the theoretical specification and sampling properties of autocorrelated time-series, *J. R. Stat. Soc.*, **B8**, 27–41.

Bartlett, M. S. (1955). *Stochastic Processes*, Cambridge University Press, Cambridge.

Basu, S. and Reinsel, G. C. (1996). Relationship between missing data likelihoods and complete data restricted data likelihoods for regression time series models: an application to total ozone data, *Appl. Stat.*, **45**, 63–72.

Bell, W. R. (1984). Signal extraction for nonstationary time series, *Ann. Stat.*, **12**, 646–664; correction, **19**, 2280, 1991.

Bell, W. R. (1987). A note on overdifferencing and the equivalence of seasonal time series models with monthly means and models with $(0, 1, 1)_{12}$ seasonal parts when $\Theta = 1$, *J. Bus. Econ. Stat.*, **5**, 383–387.

Bell, W. R., Chu, Y.-J., and Tiao, G. C. (2012). Comparing mean squared errors of X-12-ARIMA and canonical ARIMA model-based seasonal adjustments, in *Economic Time Series: Modeling and Seasonality* (eds. W. R. Bell, S. H. Holan, and T. S. McElroy), Chapman & Hall, Boca Raton, FL, pp. 161–184.

Bell, W. R. and Hillmer, S. C. (1983). Modeling time series with calendar variation, *J. Am. Stat. Assoc.*, **78**, 526–534.

Bell, W. R. and Hillmer, S. C. (1987). Initializing the Kalman filter in the nonstationary case, *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, pp. 693–697.

Bell, W. R. and Sotiris, E. (2010). Seasonal adjustment to facilitate forecasting: empirical results, *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, Alexandria, VA.

Bera, J. and Higgins, M. (1993). ARCH models: properties, estimation, and testing, *J. Econ. Surv.*, **7**, 305–362.

Beran, J. (1994). *Statistics for Long Memory Processes*, Chapman & Hall, New York.

Beran, J. (1995). Maximum likelihood estimation of the differencing parameter for invertible short and long memory autoregressive integrated moving average models, *J. R. Stat. Soc.*, **B57**, 659–672.

Bergh, L. G. and MacGregor, J. F. (1987). Constrained minimum variance controllers: internal model structure and robustness properties, *Ind. Eng. Chem. Res.*, **26**, 1558–1564.

Berndt, E. K., Hall, B. H., Hall, R. E., and Hausman, J. A. (1974). Estimation inference in nonlinear structural models, *Ann. Econ. Soc. Meas.*, **4**, 653–665.

Bhargava, A. (1986). On the theory of testing for unit roots in observed time series, *Rev. Econ. Stud.*, **53**, 369–384.

Bhattacharyya, M. N. and Layton, A. P. (1979). Effectiveness of seat belt legislation on the Queensland road toll—an Australian case study in intervention analysis, *J. Am. Stat. Assoc.*, **74**, 596–603.

Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd ed., Wiley, New York.

Birnbaum, A. (1962). On the foundations of statistical inference, *J. Am. Stat. Assoc.*, **57**, 269–306.

Bloomfield, P. (2000). *Fourier Analysis of Time Series: An Introduction*, 2nd ed., Wiley, Hoboken, NJ.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *J. Econom.*, **31**, 307–327.

Bollerslev, T. (1987). A conditional heteroskedastic time series model for speculative prices and rates of return, *Rev. Econ. Stat.*, **69**, 542–547.

Bollerslev, T. (1988). On the correlation structure in the generalized autoregressive conditional heteroskedastic process, *J. Time Ser. Anal.*, **9**, 121–31.

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992). ARCH modelling in finance: a review of the theory and empirical evidence. *J. Econom.*, **52**, 5–59.

Bollerslev, T., Engle, R. F., and Nelson, D. B. (1994). ARCH model, in *Handbook of Econometrics IV* (eds. R. F. Engle and D. C. McFadden), Elsevier Science, Amsterdam, pp. 2959–3038.

Bollerslev, T. and Wooldridge, J. M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances, *Econom. Rev.*, **11**, 143–172.

Box, G. E. P. (1966). Use and abuse of regression, *Technometrics*, **8**, 625–629.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness, *J. R. Stat. Soc.*, **A143**, 383–430.

Box, G. E. P. (1991a). Feedback control by manual adjustment, *Qual. Eng.*, **4**(1), 143–151.

Box, G. E. P. (1991b). Bounded adjustment charts, *Qual. Eng.*, **4**(a), 331–338.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *J. R. Stat. Soc.*, **B26**, 211–243.

Box, G. E. P. and Hunter, W. G. (1965). The experimental study of physical mechanisms, *Technometrics*, **7**, 23–42.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*, Wiley, New York.

Box, G. E. P. and Jenkins, G. M. (1962). Some statistical aspects of adaptive optimization and control, *J. R. Stat. Soc.*, **B24**, 297–331.

Box, G. E. P. and Jenkins, G. M. (1963). Further contributions to adaptive quality control: simultaneous estimation of dynamics: non-zero costs, *Bulletin of the International Statistical Institute, 34th Session*, Ottawa, Canada, pp. 943–974.

Box, G. E. P. and Jenkins, G. M. (1965). Mathematical models for adaptive control and optimization, *AIChE J. Chem. E Symp. Ser.*, **4**, 61.

Box, G. E. P. and Jenkins, G. M. (1968a). Discrete models for feedback and feedforward control, in *The Future of Statistics* (ed. D. G. Watts), Academic Press, New York, pp. 201–240.

Box, G. E. P. and Jenkins, G. M. (1968b). Some recent advances in forecasting and control, I, *Appl. Stat.*, **17**, 91–109.

Box, G. E. P. and Jenkins, G. M. (1969). Discrete models for forecasting and control, in *Encyclopaedia of Linguistics, Information and Control* (eds. A. R. Meaham and R. A. Hudson), Pergamon Press, Elmsford, NY, p. 162.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*, revised ed., Holden-Day, San Francisco, CA.

Box, G. E. P., Jenkins, G. M., and Bacon, D. W. (1967a). Models for forecasting seasonal and nonseasonal time series, in *Spectral Analysis of Time Series* (ed. B. Harris), Wiley, New York, pp. 271–311.

Box, G. E. P., Jenkins, G. M., and MacGregor, J. F. (1974). Some recent advances in forecasting and control, Part II, *Appl. Stat.*, **23**, 158–179.

Box, G. E. P., Jenkins, G. M., and Wichern, D. W. (1967b). Least squares analysis with a dynamic model, Technical Report 105, Department of Statistics, University of Wisconsin–Madison.

Box, G. E. P. and Kramer, T. (1992). Statistical process monitoring and feedback adjustments—a discussion, *Technometrics*, **34**, 251–285.

Box, G. E. P. and Luceño, A. (1993). Charts for optimal feedback control with recursive sampling and adjustment, Report 89, Center for Quality and Productivity Improvement, University of Wisconsin–Madison.

Box, G. E. P. and MacGregor, J. F. (1976). Parameter estimation with closed-loop operating data, *Technometrics*, **18**, 371–380.

Box, G. E. P. and Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models, *J. Am. Stat. Assoc.*, **65**, 1509–1526.

Box, G. E. P. and Ramírez, J. (1992). Cumulative score charts, *Qual. Reliab. Eng.*, **8**, 17–27.

Box, G. E. P. and Tiao, G. C. (1965). Multiparameter problems from a Bayesian point of view, *Ann. Math. Stat.*, **36**, 1468–1482.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference*, Addison-Wesley, Reading, MA.

Box, G. E. P. and Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems, *J. Am. Stat. Assoc.*, **70**, 70–79.

Box, G. E. P. and Tiao, G. C. (1976). Comparison of forecast and actuality, *Appl. Stat.*, **25**, 195–200.

Box, G. E. P. and Tiao, G. C. (1977). A canonical analysis of multiple time series, *Biometrika*, **64**, 355–365.

Bray, J. (1971). Dynamic equations for economic forecasting with the G.D.P.–unemployment relation and the growth of G.D.P. in the United Kingdom as an example, *J. R. Stat. Soc.*, **A134**, 167–209.

Briggs, P. A. N., Hammond, P. H., Hughes, M. T. G., and Plumb, G. O. (1965). Correlation analysis of process dynamics using pseudo-random binary test perturbations, *Proceedings of the Institution of Mechanical Engineers, Advances in Automatic Control*, Paper 7, Nottingham, UK, April.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, 2nd ed., Springer, New York.

Brown, R. G. (1962). *Smoothing, Forecasting and Prediction of Discrete Time Series*, Prentice Hall, Englewood Cliffs, NJ.

Brown, R. G. and Meyer, R. F. (1961). The fundamental theorem of exponential smoothing, *Oper. Res.*, **9**, 673–685.

Brubacher, S. R. and Tunnicliffe Wilson, G. (1976). Interpolating time series with applications to the estimation of holiday effects on electricity demand, *Appl. Stat.*, **25**, 107–116.

Bruce, A. G. and Martin, R. D. (1989). Leave-*k*-out diagnostics for time series (with discussion), *J. R. Stat. Soc.*, **B51**, 363–424.

Campbell, S. D. and Diebold, F. X. (2005). Weather forecasting for weather derivatives, *J. Am. Stat. Assoc.*, **100**, 6–16.

Chan, N. H. and Wei, C. Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes, *Ann. Stat.*, **16**, 367–401.

Chang, I., Tiao, G. C., and Chen, C. (1988). Estimation of time series parameters in the presence of outliers, *Technometrics*, **30**, 193–204.

Chatfield, C. (1979). Inverse autocorrelations, *J. R. Stat. Soc.*, **A142**, 363–377.

Cheang, W. K. and Reinsel, G. C. (2000). Bias reduction of autoregressive estimates in time series regression model through restricted maximum likelihood, *J. Am. Stat. Assoc.*, **95**, 1173–1184.

Cheang, W. K. and Reinsel, G. C. (2003). Finite sample properties of ML and REML estimators in time series regression models with long memory noise, *J. Stat. Comput. Simul.*, **73**, 233–259.

Chen, C. and Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series, *J. Am. Stat. Assoc.*, **88**, 284–297.

Chen, C. and Tiao, G. C. (1990). Random level shift time series models, ARIMA approximation, and level shift detection, *J. Bus. Econ. Stat.*, **8**, 170–186.

Chen, R. and Tsay, R. S. (1993). Nonlinear additive ARX models, *J. Am. Stat. Assoc.*, **88**, 955–967.

Chu, Y-J., Tiao, G. C., and Bell, W. R. (2012). A mean squared error criterion for comparing X-12-ARIMA and model-based seasonal adjustment filters, *Taiwan Econ. Forecast Policy*, **43**, 1–32.

Cleveland, W. S. (1972). The inverse autocorrelations of a time series and their applications, *Technometrics*, **14**, 277–293.

Cleveland, W. S. and Tiao, G. C. (1976). Decomposition of seasonal time series: a model for the Census X-11 Program, *J. Am. Stat. Assoc.*, **71**, 581–587.

Cooper, D. M. and Thompson, R. (1977). A note on the estimation of the parameters of the autoregressive moving average process, *Biometrika*, **64**, 625–628.

Cooper, D. M. and Wood, E. F. (1982). Identifying multivariate time series models, *J. Time Ser. Anal.*, **3**, 153–164.

Coutie, G. A. (1964) *Short Term Forecasting*, ICI Monograph 2, Oliver & Boyd, Edinburgh.

Crawley, M. J. (2007). *The R Book*, Wiley, Hoboken, NJ.

Cryer, J. D. and Chan, K.-S. (2010). *Time Series Analysis with Applications in R*, 2nd ed., Springer, New York.

Damsleth, E. (1980). Interpolating missing values in a time series, *Scand. J. Stat.*, **7**, 33–39.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments, *Technometrics*, **1**, 311–341.

Davies, N., Triggs, C. M., and Newbold, P. (1977). Significance levels of the Box–Pierce portmanteau statistic in finite samples, *Biometrika*, **64**, 517–522.

Deistler, M., Dunsmuir, W., and Hannan, E. J. (1978). Vector linear time series models: corrections and extensions, *Adv. Appl. Probab.*, **10**, 360–372.

Deming, W. E. (1986). *Out of the Crisis*, Center for Advanced Engineering Study, MIT, Cambridge, MA.

Dent, W. and Min, A. S. (1978). A Monte Carlo study of autoregressive integrated moving average processes, *J. Econom.*, **7**, 23–55.

Dickey, D. A., Bell, W. R., and Miller, R. B. (1986). Unit roots in time series models: tests and implications, *Am. Stat.*, **40**, 12–26.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimates for autoregressive time series with a unit root, *J. Am. Stat. Assoc.*, **74**, 427–431.

Dickey, D. A. and Fuller, W. A. (1981). Likelihood ratio tests for autoregressive time series with a unit root, *Econometrica*, **49**, 1057–1072.

Dickey, D. A. and Pantula, S. G. (1987). Determining the order of differencing in autoregressive processes, *J. Bus. Econ. Stat.*, **5**, 455–461.

Diebold, F. X. and Nerlove, M. (1989). The dynamics of exchange rate volatility: a multivariate latent factor ARCH model, *J. Appl. Econom.*, **4**, 1–28.

Doob, J. L. (1953). *Stochastic Processes*, Wiley, New York.

Dudding, B. P. and Jennet, W. J. (1942). Quality control charts, British Standard 600R.

Dunsmuir, W. and Hannan, E. J. (1976). Vector linear time series models, *Adv. Appl. Probab.*, **8**, 339–364.

Durbin, J. (1960). The fitting of time-series models, *Rev. Int. Stat. Inst.*, **28**, 233–244.

Durbin, J. (1970). Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables, *Econometrica*, **38**, 410–421.

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods*, 2nd ed., Oxford University Press.

Elliott, G., Rothenberg, T. J., and Stock, J. H. (1996). Efficient tests for an autoregressive unit root, *Econometrica*, **64**, 813–836.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica*, **50**, 987–1008.

Engle, R. F. (1983). Estimates of the variance of U.S. inflation based on the ARCH model, *J. Money Credit Banking*, **15**, 286–301.

Engle, R. F. and Bollerslev, T. (1986). Modeling the persistence of conditional variances, *Econom. Rev.*, **5**, 1–50.

Engle, R. F. and Granger, C. W. J. (1987). Co-integration and error correction: representation, estimation, and testing, *Econometrica*, **55**, 251–276.

Engle, R. F., Lilien, D. M., and Robins, R. P. (1987). Estimating time-varying risk premia in the term structure: the ARCH-M model, *Econometrica*, **55**, 391–407.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series*, Springer, New York.

Fearn, T. and Maris, P. I. (1991). An application of Box–Jenkins methodology to the control of gluten addition in a flour mill, *Appl. Stat.*, **40**, 477–484.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*, Oliver & Boyd, Edinburgh.

Fisher, T. J. and Gallagher, C. M. (2012). New weighted portmanteau statistics for time series goodness of fit testing, *J. Am. Stat. Assoc.* **107**, 777–787.

Fox, A. J. (1972). Outliers in time series, *J. R. Stat. Soc.*, **B34**, 350–363.

Francq, C. and Zakoïan, J.-M. (2009). A tour in the asymptotic theory of GARCH estimation, in *Handbook of Financial Time Series* (eds. T. G. Andersen, R. A. Davis, J.-P. Kreiss, and T. Mikosch), Springer, New York, pp. 85–111.

Francq, C. and Zakoïan, J.-M. (2010). *GARCH Models*, Wiley, New York.

Franses, P. H. and van Dijk, D. (2000). *Non-Linear Time Series Models in Empirical Finance*, Cambridge University Press, Cambridge.

Fuller, W. A. (1996). *Introduction to Statistical Time Series*, 2nd ed., Wiley, New York.

Gao, J. (2007). *Nonlinear Time Series: Semiparametric and Nonparametric Methods*, Chapman & Hall/CRC.

Gardner, G., Harvey, A. C., and Phillips, G. D. A. (1980). Algorithm AS 154. An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering, *Appl. Stat.*, **29**, 311–322.

Gersch, W. and Kitagawa, G. (1983). The prediction of time series with trends and seasonalities, *J. Bus. Econ. Stat.*, **1**, 253–264.

Geweke, J. and Porter-Hudak, S. (1983). The estimation and application of long memory time series models, *J. Time Ser. Anal.*, **4**, 221–238.

Ghysels, E. and Osborn, D. R. (2001). *The Econometric Analysis of Seasonal Time Series*, Cambridge University Press, Cambridge.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of nominal excess return on stocks, *J. Finance*, **48**, 1779–1801.

Godfrey, L. G. (1979). Testing the adequacy of a time series model, *Biometrika*, **66**, 67–72.

González-Rivera, G. (1998). Smooth transition GARCH models, *Stud. Nonlinear Dyn. Econom.*, **3**, 161–178.

Granger, C. W. J. and Anderson, A. P. (1978). *An Introduction to Bilinear Time Series Models*, Vandenhoek & Ruprecht, Gottingen.

Granger, C. W. J. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing, *J. Time Ser. Anal.*, **1**, 15–29.

Gray, H. L., Kelley, G. D., and McIntire, D. D. (1978). A new approach to ARMA modelling, *Commun. Stat.*, **B7**, 1–77.

Grenander, U. and Rosenblatt, M. (1957). *Statistical Analysis of Stationary Time Series*, Wiley, New York.

Hagerud, G. E. (1997). A new non-linear GARCH model, EFI, Stockholm School of Economics.

Haggan, V. and Ozaki, T. (1981). Modeling nonlinear vibrations using an amplitude-dependent autoregressive time series model, *Biometrika*, **68**, 189–196.

Haldrup, N., Kruse, R., Terävirta, T., and Varneskov, R. T. (2013). Unit roots, non-linearities and structural breaks, in *Handbook of Research Methods and Applications in Empirical Macroeconomics* (eds. N. Hashimzade and M. A. Thornton), Edward Elgar Publishing, Ltd., UK

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*, Academic Press, New York.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton University Press, Princeton, NJ.

Hannan, E. J. (1960). *Time Series Analysis*, Methuen, London.

Hannan, E. J. (1970). *Multiple Time Series*, Wiley, New York.

Hannan, E. J. and Deistler, M. (1988). *The Statistical Theory of Linear Systems*, Wiley, New York.

Hannan, E. J., Dunsmuir, W. M., and Deistler, M. (1979). Estimation of vector ARMAX models, *J. Multivariate Anal.*, **10**, 275–295.

Hannan, E. J. and Kavalieris, L. (1984). Multivariate linear time series models, *Adv. Appl. Probab.*, **16**, 492–561.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression, *J. R. Stat. Soc.*, **B41**, 190–195.

Hannan, E. J. and Rissanen, J. (1982). Recursive estimation of mixed autoregressive moving average order, *Biometrika*, **69**, 81–94; correction, **70**, 303, 1983.

Harris, T. J., MacGregor, J. F., and Wright, J. D. (1982). An overview of discrete stochastic controllers: generalized PID algorithms with dead-time compensation, *Can. J. Chem. Eng.*, **60**, 425–432.

Harrison, P. J. (1965). Short-term sales forecasting, *Appl. Stat.*, **14**, 102–139.

Harvey, A. C. (1981). Finite sample prediction and overdifferencing, *J. Time Ser. Anal.*, **2**, 221–232.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

Harvey, A. C. and Phillips, G. D. A. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances, *Biometrika*, **66**, 49–58.

Harvey, A. C. and Pierse, R. G. (1984). Estimating missing observations in economic time series, *J. Am. Stat. Assoc.*, **79**, 125–131.

Harvey, A. C., Ruiz, E., and Shephard, N. (1994). Multivariate stochastic variance models, *Rev. Econ. Stud.*, **61**, 247–264.

Harvey, A. C. and Todd, P. H. J. (1983). Forecasting economic time series with structural and Box–Jenkins models: a case study, *J. Bus. Econ. Stat.*, **1**, 299–307.

Harvey, D. I., Leybourne, S. J., and Taylor, A. M. R. (2009). Unit root testing in practice: dealing with uncertainty over trend and initial conditions (with commentaries and rejoiner), *Econom. Theory*, **25**, 587–667.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts, *Biometrika*, **61**, 383–385.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems, *J. Am. Stat. Assoc.*, **72**, 320–340.

Haugh, L. D. and Box, G. E. P. (1977). Identification of dynamic regression (distributed lag) models connecting two time series, *J. Am. Stat. Assoc.*, **72**, 121–130.

Hauser, M. A. (1999). Maximum likelihood estimators for ARMA and ARFIMA models: a Monte Carlo study, *J. Stat. Plan. Infer.*, **80**, 229–255.

He, C., Silvennoinen, A., and Teräsvirta, T. (2008). Parameterizing unconditional skewness in non-linear time series with conditional heteroscedasticity, *J. Financ. Econ.*, **6**, 208–230.

He, C. and Teräsvirta, T. (1999). Fourth moment structure of the GARCH(p, q) process, *Econom. Theory*, **15**, 824–846.

Hillmer, S. C. and Tiao, G. C. (1979). Likelihood function of stationary multiple autoregressive moving average models, *J. Am. Stat. Assoc.*, **74**, 652–660.

Hillmer, S. C. and Tiao, G. C. (1982). An ARIMA-model-based approach to seasonal adjustment, *J. Am. Stat. Assoc.*, **77**, 63–70.

Hinich, M. J. (1982). Testing for Gaussianity and linearity of a stationary time series, *J. Time Ser. Anal.*, **3**, 169–176.

Holt, C. C. (1957). Forecasting trends and seasonals by exponentially weighted moving averages, O.N.R. Memorandum 52, Carnegie Institute of Technology, Pittsburgh, PA.

Holt, C. C., Modigliani, F., Muth, J. F., and Simon, H. A. (1963). *Planning Production, Inventories and Work Force*, Prentice Hall, Englewood Cliffs, NJ.

Hosking, J. R. M. (1980). The multivariate portmanteau statistic, *J. Am. Stat. Assoc.*, **75**, 602–608.

Hosking, J. R. M. (1981). Fractional differencing, *Biometrika*, **68**, 165–176.

Hosking, J. R. M. (1996). Asymptotic distributions of the sample mean, autocovariances, and auto-correlations of long memory time series, *J. Econom.*, **73**, 261–284.

Hougen, J. O. (1964). *Experience and Experiments with Process Dynamics*, Chemical Engineering Progress Monograph Series, Vol. 60, No. 4, American Institute Chemical Engineers.

Hurst, H. (1951). Long term storage capacity of reservoirs, *Trans. Am. Soc. Civil Eng.*, **116**, 778–808.

Hutchinson, A. W. and Shelton, R. J. (1967). Measurement of dynamic characteristics of full-scale plant using random perturbing signals: an application to a refinery distillation column, *Trans. Inst. Chem. Eng.*, **45**, 334–342.

Ishikawa, K. (1976). *Guide to Quality Control*, Asian Productivity Organization, Tokyo.

Jacquier, E., Polson, N. G., and Rossi, P. (1994). Bayesian analysis of stochastic volatility models (with discussion), *J. Bus. Econ. Stat.*, **12**, 371–417.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed., Clarendon Press, Oxford.

Jenkins, G. M. (1956). Tests of hypotheses in the linear autoregressive model, I, *Biometrika*, **41**, 405–419, 1954; II, *Biometrika*, **43**, 186–199.

Jenkins, G. M. (1964). Contribution to the discussion of the paper ''Relationships between Bayesian and confidence limits for predictors,'' by A. R. Thatcher, *J. R. Stat. Soc.*, **B26**, 176–210.

Jenkins, G. M. (1975). The interaction between the muskrat and mink cycles in North Canada, *Proceedings of the 8th International Biometric Conference*, Editura Aca- demiei Republicii Socialiste Romania, Bucharest, pp. 55–71.

Jenkins, G. M. (1979). *Practical Experiences with Modelling and Forecasting Time Series*, Gwilym Jenkins & Partners Ltd., Jersey, Channel Islands.

Jenkins, G. M. and Watts, D. G. (1968). *Spectral Analysis and Its Applications*, Holden-Day, San Francisco, CA.

Johansen, S. (1988). Statistical analysis of cointegration vectors, *J. Econ. Dyn. Control*, **12**, 231–254.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models, *Econometrica*, **59**, 1551–1580.

Johansen, S. and Juselius, K. (1990). Maximum likelihood estimation and inference on cointegration: with applications to the demand for money, *Oxf. Bull. Econ. Stat.*, **52**, 169–210.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, 6th ed., Pearson, Prentice Hall, Upper Saddle River, NJ.

Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations, *Technometrics*, **22**, 389–395.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems, *J. Basic Eng.*, **82**, 35–45.

Kalman, R. E. and Bucy, R.S. (1961). New results in linear filtering and prediction theory, *J. Basic Eng.*, **83**, 95–108.

Keenan, D. M. (1985). A Tukey non-additivity-type test for time series nonlinearity, *Biometrika*, **72**, 39–44.

Kendall, M. G. (1945). On the analysis of oscillatory time-series, *J. R. Stat. Soc.*, **A108**, 93–129.

Kitagawa, G. and Gersch, W. (1984). A smoothness priors-state space modeling of time series with trend and seasonality, *J. Am. Stat. Assoc.*, **79**, 378–389.

Kohn, R. and Ansley, C. F. (1986). Estimation, prediction, and interpolation for ARIMA models with missing data, *J. Am. Stat. Assoc.*, **81**, 751–761.

Kolmogoroff, A. (1939). Sur l'interpolation et l'extrapolation des suites stationnaires, *C. R. Acad. Sci. Paris*, **208**, 2043–2045.

Kolmogoroff, A. (1941a). Stationary sequences in Hilbert space, *Bull. Math. Univ. Moscow*, **2**(6), 1–40.

Kolmogoroff, A. (1941b). Interpolation und Extrapolation von stationären zufälligen folgen, *Bull. Acad. Sci.* (*Nauk*) *USSR, Ser. Math.*, **5**, 3–14.

Koopmans, L. H. (1974). *The Spectral Analysis of Time Series*, Academic Press, New York.

Kotnour, K. D., Box, G. E. P., and Altpeter, R. J. (1966). A discrete predictor-controller applied to sinusoidal perturbation adaptive optimization, *Inst. Soc. Am. Trans.*, **5**, 255–262.

Lanne, M. and Saikkonen, P. (2005). Nonlinear GARCH models for highly persistent volatility, *Econom. J.*, **8**, 251–276.

Le, N. D., Martin, R. D., and Raftery, A. E. (1996). Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models, *J. Am. Stat. Assoc.*, **91**, 1504–1515.

Ledolter, J. (1981). Recursive estimation and adaptive forecasting in ARIMA models with time varying coefficients, in *Applied Time Series II* (ed. D. F. Findley), Academic Press, New York, pp. 449–472.

Levinson, N. (1947). The Wiener (root mean square) error criterion in filter design and prediction, *Math. Phys.*, **25**, 262–278.

Lewis, P. A. W. and Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS), *J. Am. Stat. Assoc.*, **86**, 864–877.

Leybourne, S. J. and McCabe, B. P. M. (1994). A consistent test for a unit root, *J. Bus. Econ. Stat.*, **12**, 157–166.

Li, W. K. (2004). *Diagnostic Checks for Time Series*, Chapman & Hall/CRC.

Li, W. K., Ling, S., and McAleer, M. (2003). Recent theoretical results for time series models with GARCH errors, *J. Econ. Surv.*, **16**, 245–269.

Li, W. K. and Mak, T. K. (1994). On the squared residual autocorrelations in non-linear time series models with conditional heteroscedasticity, *J. Time Ser. Anal.*, **15**, 627–636.

Li, W. K. and McLeod, A. I. (1981). Distribution of the residual autocorrelations in multivariate ARMA time series models, *J. R. Stat. Soc.*, **B43**, 231–239.

Li, W. K. and McLeod, A. I. (1986). Fractional time series modelling, *Biometrika*, **73**, 217–221.

Ling, S. and McAleer, M. (2002). Stationarity and the existence of moments of a family of GARCH models, *J. Econom.*, **106**, 109–117.

Ling, S. and Tong, H. (2011). Score based goodness-of-fit tests for time series, *Stat. Sin.*, **21**, 1807–1829.

Liu, J. and Brockwell, P. J. (1988). On the general bilinear time series model, *J. Appl. Probab.*, **25**, 553–564.

Liu, L.-M. and Hanssens, D. M. (1982). Identification of multiple-input transfer function models, *Commun. Stat.*, **A11**, 297–314.

Ljung, G. M. (1986). Diagnostic testing of univariate time series models, *Biometrika*, **73**, 725–730.

Ljung, G. M. (1993). On outlier detection in time series, *J. R. Stat. Soc.*, **B55**, 559–567.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models, *Biometrika*, **65**, 297–303.

Ljung, G. M. and Box, G. E. P. (1979). The likelihood function of stationary autoregressive-moving average models, *Biometrika*, **66**, 265–270.

Loève, M. (1977). *Probability Theory I*, Springer, New York.

Lundbergh, S. and Teräsvirta, T. (2002). Evaluating GARCH models, *J. Econom.*, **110**, 417–435.

Lütkepohl, H. (2006). *New Introduction to Multiple Time Series Analysis*, Springer, Berlin.

Lütkepohl, H. and Poskitt, D. S. (1996). Specification of echelon-form VARMA models, *J. Bus. Econ. Stat.*, **14**, 69–79.

Luukkonen, R., Saikkonen, P., and Teräsvirta, T. (1988a). Testing linearity against smooth transition autoregressive models, *Biometrika*, **75**, 491–499.

Luukkonen, R., Saikkonen, P., and Teräsvirta, T. (1988b). Testing linearity in univariate time series, *Scand. J. Stat.*, **15**, 161–175.

MacGregor, J. F. (1972). Topics in the control of linear processes with stochastic disturbances, Ph.D. thesis, University of Wisconsin–Madison.

Mahdi, E. and McLeod, I. A. (2012). Improved multivariate portmanteau test, *J. Time Ser. Anal.*, **33**, 211–222.

Mann, H. B. and Wald, A. (1943). On the statistical treatment of linear stochastic difference equations, *Econometrica*, **11**, 173–220.

Maravall, A. (1993). Stochastic linear trends: models and estimators. *J. Econom.*, **56**, 5–37.

Martin, R. D. and Yohai, V. J. (1986). Influence functionals for time series (with discussion), *Ann. Stat.*, **14**, 781–855.

McLeod, A. I. and Hipel, K. W. (1978). Preservation of the rescaled adjusted range. I. A reassessment of the Hurst phenomenon, *Water Resour. Res.*, **14**, 491–508.

McLeod, A. I. and Li, W. K. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations, *J. Time Ser. Anal.*, **4**, 269–273.

Melino, A. and Turnbull, S. M. (1990). Pricing foreign currency options with stochastic volatility, *J. Econom.*, **45**, 239–265.

Milhøj, A. (1985). The moment structure of ARCH processes, *Scand. J. Stat.*, **12**, 281–292.

Mills, T. C. and Markellos, R. N. (2008). *The Econometric Modelling of Financial Time Series*, 3rd ed., Cambridge University Press, Cambridge.

Montgomery, D. C. and Weatherby, G. (1980). Modeling and forecasting time series using transfer function and intervention methods, *AIIE Trans.*, 289–307.

Monti, A. C. (1994). A proposal for residual autocorrelation test in linear models, *Biometrika*, **81**, 776–780.

Moran, P. A. P. (1954). Some experiments on the prediction of sunspot numbers, *J. R. Stat. Soc.*, **B16**, 112–117.

Muth, J. F. (1960). Optimal properties of exponentially weighted forecasts, *J. Am. Stat. Assoc.*, **55**, 299–306.

Nelson, D. B. (1990). Stationarity and persistence in the GARCH(1, 1) model, *Econom. Theory*, **6**, 318–334.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: a new approach, *Econometrica*, **59**, 347–370.

Nelson, D. B. and Cao, C. Q. (1992). Inequality constraints in the univariate GARCH model, *J. Bus. Econ. Stat.*, **10**, 229–235.

Newbold, P. (1973). Bayesian estimation of Box–Jenkins transfer function-noise models, *J. R. Stat. Soc.*, **B35**, 323–336.

Newbold, P. (1974). The exact likelihood function for a mixed autoregressive-moving average process, *Biometrika*, **61**, 423–426.

Newbold, P. (1980). The equivalence of two tests of time series model adequacy, *Biometrika*, **67**, 463–465.

Ng, S. and Perron, P. (2001). Lag length selection and the construction of unit root tests with good size and power, *Econometrica*, **69**, 1519–1554.

Nicholls, D. F. and Hall, A. D. (1979). The exact likelihood function of multivariate autoregressive moving average models, *Biometrika*, **66**, 259–264.

Nicholls, D. F. and Quinn, B. G. (1982). *Random Coefficient Autoregressive Models: An Introduction*, *Lecture Notes in Statistics*, Vol. 11, Springer, New York.

Osborn, D. R. (1982). On the criteria functions used for the estimation of moving average processes, *J. Am. Stat. Assoc.*, **77**, 388–392.

Page, E. S. (1957). On problems in which a change in a parameter occurs at an unknown point, *Biometrika*, **44**, 248–252.

Page, E. S. (1961). Cumulative sum charts, *Technometrics*, **3**, 1–9.

Palm, F. C., Smeekes, S., and Urbain, J. P. (2008). Bootstrap unit-root tests: comparison and extensions, *J. Time Ser. Anal.*, **29**, 371–401.

Palma, W. (2007). *Long-Memory Time Series: Theory and Methods*, Wiley, New York.

Pankratz, A. (1991). *Forecasting with Dynamic Regression Models*, Wiley, New York.

Pantula, S. G., Gonzalez-Farias, G., and Fuller, W. A. (1994). A comparison of unit root test criteria, *J. Bus. Econ. Stat.*, **12**, 449–459.

Peña, D. and Rodríguez, J. (2002). A powerful portmanteau test of lack of fit for time series, *J. Am. Stat. Assoc.*, **97**, 601–610.

Peña, D. and Rodríguez, J. (2006). The log of the determinant of the autocorrelation matrix for testing of lack of fit in time series, *J. Stat. Plan. Infer.*, **136**, 2706–2718.

Perron, P. and Qu, Z. (2007). A simple modification to improve the finite sample properties of Ng and Perron's unit root tests, *Econ. Lett.*, **94**, 12–19.

Petruccelli, J. and Davies, N. (1986). A portmanteau test for self-exciting threshold autoregressive-type nonlinearity in time series, *Biometrika*, **73**, 687–694.

Phillips, P. C. B. (1987). Time series regression with a unit root, *Econometrica*, **55**, 277–301.

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression, *Biometrika*, **75**, 335–346.

Phillips, P. C. B. and Xiao, Z. (1998). A primer on unit root testing, *J. Econ. Surv.*, **12**, 423–470.

Pierce, D. A. (1972a). Least squares estimation in dynamic-disturbance time series models, *Biometrika*, **59**, 73–78.

Pierce, D. A. (1972b). Residual correlations and diagnostic checking in dynamic-disturbance time series models, *J. Am. Stat. Assoc.*, **67**, 636–640.

Poskitt, D. S. (1992). Identification of echelon canonical forms for vector linear processes using least squares, *Ann. Stat.*, **20**, 195–215.

Poskitt, D. S. and Tremayne, A. R. (1980). Testing the specification of a fitted autoregressive-moving average model, *Biometrika*, **67**, 359–363.

Poskitt, D. S. and Tremayne, A. R. (1981). An approach to testing linear time series models, *Ann. Stat.*, **9**, 974–986.

Poskitt, D. S. and Tremayne, A. R. (1982). Diagnostic tests for multiple time series models, *Ann. Stat.*, **10**, 114–120.

Priestley, M. B. (1980). State-dependent models: a general approach to non-linear time series analysis, *J. Time Ser. Anal.*, **1**, 47–71.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*, Academic Press, New York.

Priestley, M. B. (1988). *Non-Linear and Non-Stationary Time Series Analysis*, Academic Press, London.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series, *J. R. Stat. Soc.*, **B11**, 68–84.

Quenouille, M. H. (1952). *Associated Measurements*, Butterworth, London.

Quenouille, M. H. (1957). *Analysis of Multiple Time Series*, Hafner, New York.

Quinn, B. G. (1980). Order determination for a multivariate autoregression, *J. R. Stat. Soc.*, **B42**, 182–185.

Ramsey, J. B. (1969). Tests for specification errors in classical least-squares regression analysis, *J. R. Stat. Soc.*, **B31**, 350–371.

Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*, Wiley, New York.

Rao, J. N. K. and Tintner, G. (1963). On the variate difference method, *Aust. J. Stat.*, **5**, 106–116.

Reinsel, G. C. (1979). Maximum likelihood estimation of stochastic linear difference equations with autoregressive moving average errors, *Econometrica*, **47**, 129–151.

Reinsel, G. C. (1997). *Elements of Multivariate Time Series Analysis*, 2nd ed., Springer, New York.

Reinsel, G. C. and Ahn, V. (1992). Vector autoregressive models with unit roots and reduced rank structure: estimation, likelihood ratio test, and forecasting, *J. Time Ser. Anal.*, **13**, 353–375.

Reinsel, G. C. and Tiao, G. C. (1987). Impact of chlorofluoromethanes on stratospheric ozone: a statistical analysis of ozone data for trends, *J. Am. Stat. Assoc.*, **82**, 20–30.

Reinsel, G. C. and Wincek, M. A. (1987). Asymptotic distribution of parameter estimators for nonconsecutively observed time series, *Biometrika*, **74**, 115–124.

Rivera, D. E., Morari, M., and Skogestad, S. (1986). Internal model control. 4. PID controller design, *Ind. Eng. Chem. Process Des. Dev.*, **25**, 252–265.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages, *Technometrics*, **1**, 239–250.

Robinson, E. A. (1967). *Multichannel Time Series Analysis*, Holden-Day, San Francisco, CA.

Robinson, P. M. (2003). *Time Series with Long Memory*, Oxford University Press, Oxford.

Said, S. E. and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order, *Biometrika*, **71**, 599–607.

Said, S. E. and Dickey, D. A. (1985). Hypothesis testing in ARIMA($p, 1, q$) models, *J. Am. Stat. Assoc.*, **80**, 369–374.

Saikkonen, P. and Luukkonen, R. (1993). Testing for a moving average unit root in autoregressive integrated moving average models, *J. Am. Stat. Assoc.*, **88**, 596–601.

Savage, L. J. (1962). *The Foundations of Statistical Inference*, Methuen, London.

Schmidt, P. and Phillips, P. C. B. (1992). LM tests for a unit root in the presence of deterministic trends, *Oxf. Bull. Econ. Stat.*, **54**, 257–287.

Schuster, A. (1898). On the investigation of hidden periodicities, *Terr. Magn. Atmos. Electr.*, **3**, 13–41.

Schuster, A. (1906). On the periodicities of sunspots, *Philos. Trans. R. Soc.*, **A206**, 69–100.

Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Stat.*, **6**, 461–464.

Shea, B. L. (1987). Estimation of multivariate time series, *J. Time Ser. Anal.*, **8**, 95–109.

Shephard, N. G. (2005). *Stochastic Volatility: Selected Readings*, Oxford University Press, Oxford.

Shewhart, W. A. (1931). *The Economic Control of the Quality of Manufactured Product*, Macmillan, New York.

Shin, D. W. and Fuller, W. A. (1998). Unit root tests based on unconditional maximum likelihood estimation for the autoregressive moving average, *J. Time Ser. Anal.*, **19**, 591–599.

Shumway, R. and Stoffer, D. (2011). *Time Series Analysis and Its Applications*, 3rd ed., Springer, New York.

Silvey, S. D. (1959). The Lagrangian multiplier test, *Ann. Math. Stat.*, **30**, 389–407.

Slutsky, E. (1937). The summation of random causes as the source of cyclic processes (Russian), *Probl. Econ. Conditions*, **3**, 1, 1927; English translation, *Econometrica*, **5**, 105–146.

Solo, V. (1984a). The exact likelihood for a multivariate ARMA model, *J. Multivariate Anal.*, **15**, 164–173.

Solo, V. (1984b). The order of differencing in ARIMA models, *J. Am. Stat. Assoc.*, **79**, 916–921.

Solo, V. (1986). Topics in advanced time series analysis, in *Lectures in Probability and Statistics* (eds. G. del Pino and R. Rebolledo), Springer, New York, pp. 165–328.

Sowell, F. (1992). Maximum likelihood estimation of stationary univariate fractionally integrated time series models, *J. Econom.*, **53**, 165–188.

Stralkowski, C. M. (1968). Lower order autoregressive-moving average stochastic models and their use for the characterization of abrasive cutting tools, Ph.D. thesis, University of Wisconsin–Madison.

Straumann, D. and Mikosch, T. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: a stochastic recurrence equations approach, *Ann. Stat.*, **34**, 2449–2495.

Subba Rao, T. (1981). On the theory of bilinear models, *J. R. Stat. Soc.*, **B43**, 244–255.

Subba Rao, T. and Gabr, M. M. (1980). A test for nonlinearity of stationary time series, *J. Time Ser. Anal.*, **1**, 145–158.

Subba Rao, T. and Gabr, M. M. (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models*, Springer, Berlin.

Tam, W. K. and Reinsel, G. C. (1997). Tests for seasonal moving average unit root in ARIMA models, *J. Am. Stat. Assoc.*, **92**, 725–738.

Tam, W. K. and Reinsel, G. C. (1998). Seasonal moving-average unit root tests in the presence of a linear trend, *J. Time Ser. Anal.*, **19**, 609–625.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models, *J. Am. Stat. Assoc.*, **89**, 208–218.

Teräsvirta, T. (2009). An introduction to univariate GARCH models, in *Handbook of Financial Time Series* (eds. T. G. Andersen, R. A. Davis, J.-P Kreiss, J.-P., and T. Mikosch), Springer, New York, pp. 85–111.

Teräsvirta, T., Tjøstheim, D, and Granger, W. J. (2010). *Modelling Nonlinear Economic Time Series*, Oxford University Press, Oxford, UK.

Thompson, H. E. and Tiao, G. C. (1971). Analysis of telephone data: a case study of forecasting seasonal time series, *Bell J. Econ. Manage. Sci.*, **2**, 515–541.

Tiao, G. C. and Box, G. E. P. (1981). Modeling multiple time series with applications, *J. Am. Stat. Assoc.*, **76**, 802–816.

Tiao, G. C., Box, G. E. P., and Hamming, W. J. (1975). Analysis of Los Angeles photochemical smog data: a statistical overview, *J. Air Pollut. Control Assoc.*, **25**, 260–268.

Tiao, G. C. and Tsay, R. S. (1983). Multiple time series modeling and the extended sample cross-correlations, *J. Bus. Econ. Stat.*, **1**, 43–56.

Tiao, G. C. and Tsay, R. S. (1989). Model specification in multivariate time series (with discussion), *J. R. Stat. Soc.*, **B51**, 157–213.

Tintner, G. (1940). *The Variate Difference Method*, Principia Press, Bloomington, IN.

Tjøstheim, D. (1994). Non-linear time series: a selective review, *Scand. J. Stat.*, **21**, 97–130.

Tong, H. (1978). On a threshold model, in *Pattern Recognition and Signal Processing* (ed. C. H. Chen), Sijthoff & Noordhoff, Amsterdam, pp. 101–141.

Tong, H. (1983). *Threshold Models in Non-Linear Time Series Analysis*, Springer, New York.

Tong, H. (1990). *Non-Linear Time Series: A Dynamical System Approach*, Oxford University Press, Oxford, UK.

Tong, H. (2007). Birth of the threshold autoregressive model, *Stat. Sin.*, **17**, 8–14.

Tong, H. and Lim, K. S. (1980). Threshold autoregression, limit cycles, and cyclical data (with discussion), *J. R. Stat. Soc.*, **B42**, 245–292.

Tsay, R. S. (1986a). Nonlinearity tests for time series, *Biometrika*, **73**, 461–466.

Tsay, R. S. (1986b). Time series model specification in the presence of outliers, *J. Am. Stat. Assoc.*, **81**, 132–141.

Tsay, R. S. (1987). Conditional heteroskedastic time series models, *J. Am. Stat. Assoc.*, **82**, 590–604.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series, *J. Forecasting*, **7**, 1–20.

Tsay, R. S. (1989a). Identifying multivariate time series models, *J. Time Ser. Anal.*, **10**, 357–372.

Tsay, R. S. (1989b). Parsimonious parametrization of vector autoregressive moving average models, *J. Bus. Econ. Stat.*, **7**, 327–341.

Tsay, R. S. (1991). Two canonical forms for vector ARMA processes, *Stat. Sin.*, **1**, 247–269.

Tsay, R. S. (2010). *Analysis of Financial Time Series*, 3rd ed., Wiley, Hoboken, NJ.

Tsay, R. S. (2014). *Multivariate Time Series Analysis*, Wiley, Hoboken, NJ.

Tsay, R. S. and Tiao, G. C. (1984). Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models, *J. Am. Stat. Assoc.*, **79**, 84–96.

Tsay, R. S. and Tiao, G. C. (1985). Use of canonical analysis in time series model identification, *Biometrika*, **72**, 299–315.

Tuan, P.-D. (1985). Bilinear Markovian representation and bilinear models, *Stoch. Process. Appl.*, **20**, 295–306.

Tuan, P.-D. (1986). The mixing property of bilinear and generalized random coefficient autoregressive models, *Stoch. Process. Appl.*, **21**, 291–300.

Tukey, J. W. (1961). Discussion, emphasizing the connection between analysis of variance and spectrum analysis, *Technometrics*, **3**, 191–219.

Tunnicliffe Wilson, G. (1970a). Optimal control: a general method of obtaining the feedback scheme which minimizes the output variance, subject to a constraint on the variability of the control variable, Technical Report 20, Department of Systems Engineering, University of Lancaster, Lancaster, UK.

Tunnicliffe Wilson, G. (1970b). Modelling Linear Systems for Multivariate Control, Ph.D. thesis, University of Lancaster, UK.

Tunnicliffe Wilson, G. (1989). On the use of marginal likelihood in time series model estimation, *J. R. Stat. Soc.*, **B51**, 15–27.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*, 4th ed., Springer, New York.

Walker, A. M. (1964). Asymptotic properties of least-squares estimates of parameters of the spectrum of a stationary non-deterministic time-series, *J. Aust. Math. Soc.*, **4**, 363–384.

Walker, G. (1931). On periodicity in series of related terms, *Proc. R. Soc.*, **A131**, 518–532.

Wei, W. W. S. (2006). *Time Series Analysis, Univariate and Multivariate Methods*, 2nd ed., Pearson, Addison-Wesley.

Weiss, A. A. (1984). ARMA models with ARCH errors, *J. Time Ser. Anal.*, **5**, 129–143.

Weiss, A. A. (1986). Asymptotic theory for ARCH models: estimation and testing. *Econom. Theory*, **2**, 107–131.

Whittle, P. (1953). Estimation and information in stationary time series, *Ark. Math.*, **2**, 423–434.

Whittle, P. (1963). *Prediction and Regulation by Linear Least-Squares Methods*, English Universities Press, London.

Wichern, D. W. (1973). The behaviour of the sample autocorrelation function for an integrated moving average process, *Biometrika*, **60**, 235–239.

Wiener, N. (1949). *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, Wiley, New York.

Wincek, M. A. and Reinsel, G. C. (1986). An exact maximum likelihood estimation procedure for regression-ARMA time series models with possibly nonconsecutive data, *J. R. Stat. Soc.*, **B48**, 303–313.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages, *Manage. Sci.*, **6**, 324–342.

Wold, H. O. (1938). *A Study in the Analysis of Stationary Time Series*, Almqvist & Wiksell, Uppsala, Sweden; 2nd ed., 1954.

Wong, H. and Ling, S. (2005). Mixed portmanteau tests for time series, *J. Time Ser. Anal.*, **26**, 569–579.

Woodward, W. A. and Gray, H. L. (1981). On the relationship between the S array and the Box–Jenkins method of ARMA model identification, *J. Am. Stat. Assoc.*, **76**, 579–587.

Xekalaki, E. and Degiannakis, S. (2010). *ARCH Models for Financial Applications*, Wiley, New York.

Yaglom, A. M. (1955). The correlation theory of processes whose *n*th difference constitute a stationary process, *Mat. Sb.*, **37**(79), 141.

Yamamoto, T. (1976). Asymptotic mean square prediction error for an autoregressive model with estimated coefficients, *Appl. Stat.*, **25**, 123–127.

Yap, S. F. and Reinsel, G. C. (1995). Results on estimation and testing for a unit root in the nonstationary autoregressive moving-average model, *J. Time Ser. Anal.*, **16**, 339–353.

Young, A. J. (1955). *An Introduction to Process Control Systems Design*, Longmans Green, New York.

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers, *Philos. Trans. R. Soc.*, **A226**, 267–298.

Zadeh, L. A. and Ragazzini, J. R. (1950). An extension of Wiener's theory of prediction, *J. Appl. Phys.*, **21**, 645.

Zakoïan, J. M. (1994). Threshold heteroscedastic models, *J. Econ. Dyn. Control*, **18**, 931–955.

INDEX

Time Series Analysis: Forecasting and Control, Fifth Edition. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

The *Wiley Series in Probability and Statistics* is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

† ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data, *Second Edition*
AGRESTI · An Introduction to Categorical Data Analysis, *Second Edition*
AGRESTI · Categorical Data Analysis, *Third Edition*
AGRESTI · *Foundations of Linear and Generalized Linear Models*
ALSTON, MENGERSEN and PETTITT (editors) · Case Studies in Bayesian Statistical Modelling and Analysis
ALTMAN, GILL, and McDONALD · Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA · Exploration and Analysis of DNA Microarray and Protein Array Data
AMARATUNGA, CABRERA, and SHKEDY · Exploration and Analysis of DNA Microarray and Other High-Dimensional Data, *Second Edition*
ANDĚL · Mathematics of Chance
ANDERSON · An Introduction to Multivariate Statistical Analysis, *Third Edition*
* ANDERSON · The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ANDERSON and LOYNES · The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) · Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA · Records
* ARTHANARI and DODGE · Mathematical Programming in Statistics
AUGUSTIN, COOLEN, DE COOMAN and TROFFAES (editors) · Introduction to Imprecise Probabilities
* BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BAJORSKI · Statistics for Imaging, Optics, and Photonics
BALAKRISHNAN and KOUTRAS · Runs and Scans with Applications
BALAKRISHNAN and NG · Precedence-Type Tests and Applications
BARNETT · Comparative Statistical Inference, *Third Edition*
BARNETT · Environmental Statistics
BARNETT and LEWIS · Outliers in Statistical Data, *Third Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley Interscience Paperback Series.

BARTHOLOMEW, KNOTT, and MOUSTAKI · Latent Variable Models and Factor Analysis: A Unified Approach, *Third Edition*

BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference, *Second Edition*

BASILEVSKY · Statistical Factor Analysis and Related Methods: Theory and Applications

BATES and WATTS · Nonlinear Regression Analysis and Its Applications

BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons

BEH and LOMBARDO · Correspondence Analysis: Theory, Practice and New Strategies

BEIRLANT, GOEGEBEUR, SEGERS, TEUGELS, and DE WAAL · Statistics of Extremes: Theory and Applications

BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression

† BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity

BENDAT and PIERSOL · Random Data: Analysis and Measurement Procedures, *Fourth Edition*

BERNARDO and SMITH · Bayesian Theory

BHAT and MILLER · Elements of Applied Stochastic Processes, *Third Edition*

BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications

BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN · Measurement Errors in Surveys

BILLINGSLEY · Convergence of Probability Measures, *Second Edition*

BILLINGSLEY · Probability and Measure, *Anniversary Edition*

BIRKES and DODGE · Alternative Methods of Regression

BISGAARD and KULAHCI · Time Series Analysis and Forecasting by Example

BISWAS, DATTA, FINE, and SEGAL · Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics

BLISCHKE and MURTHY (editors) · Case Studies in Reliability and Maintenance

BLISCHKE and MURTHY · Reliability: Modeling, Prediction, and Optimization

BLOOMFIELD · Fourier Analysis of Time Series: An Introduction, *Second Edition*

BOLLEN · Structural Equations with Latent Variables

BOLLEN and CURRAN · Latent Curve Models: A Structural Equation Perspective

BONNINI, CORAIN, MAROZZI and SALMASO · Nonparametric Hypothesis Testing: Rank and Permutation Methods with Applications in R

BOROVKOV · Ergodicity and Stability of Stochastic Processes

BOSQ and BLANKE · Inference and Prediction in Large Dimensions

BOULEAU · Numerical Methods for Stochastic Processes

* BOX and TIAO · Bayesian Inference in Statistical Analysis

BOX · Improving Almost Anything, *Revised Edition*

* BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement

BOX and DRAPER · Response Surfaces, Mixtures, and Ridge Analyses, *Second Edition*

BOX, HUNTER, and HUNTER · Statistics for Experimenters: Design, Innovation, and Discovery, *Second Editon*

BOX, JENKINS, REINSEL, and LJUNG · Time Series Analysis: Forecasting and Control, *Fifth Edition*

BOX, LUCEÑO, and PANIAGUA-QUIÑONES · Statistical Control by Monitoring and Adjustment, *Second Edition*

* BROWN and HOLLANDER · Statistics: A Biomedical Introduction

CAIROLI and DALANG · Sequential Stochastic Optimization

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley Interscience Paperback Series.

CASTILLO, HADI, BALAKRISHNAN, and SARABIA · Extreme Value and Related Models with Applications in Engineering and Science
CHAN · Time Series: Applications to Finance with R and S-Plus®, *Second Edition*
CHARALAMBIDES · Combinatorial Methods in Discrete Distributions
CHATTERJEE and HADI · Regression Analysis by Example, *Fourth Edition*
CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
CHEN · The Fitness of Information: Quantitative Assessments of Critical Evidence
CHERNICK · Bootstrap Methods: A Guide for Practitioners and Researchers, *Second Edition*
CHERNICK and FRIIS · Introductory Biostatistics for the Health Sciences
CHILÈS and DELFINER · Geostatistics: Modeling Spatial Uncertainty, *Second Edition*
CHIU, STOYAN, KENDALL and MECKE · Stochastic Geometry and Its Applications, *Third Edition*
CHOW and LIU · Design and Analysis of Clinical Trials: Concepts and Methodologies, *Third Edition*
CLARKE · Linear Models: The Theory and Application of Analysis of Variance
CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
* COCHRAN and COX · Experimental Designs, *Second Edition*
COLLINS and LANZA · Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences
CONGDON · Applied Bayesian Modelling, *Second Edition*
CONGDON · Bayesian Models for Categorical Data
CONGDON · Bayesian Statistical Modelling, *Second Edition*
CONOVER · Practical Nonparametric Statistics, *Third Edition*
COOK · Regression Graphics
COOK and WEISBERG · An Introduction to Regression Graphics
COOK and WEISBERG · Applied Regression Including Computing and Graphics
CORNELL · A Primer on Experiments with Mixtures
CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*
COX · A Handbook of Introductory Statistical Methods
CRESSIE · Statistics for Spatial Data, *Revised Edition*
CRESSIE and WIKLE · Statistics for Spatio-Temporal Data
CSÖRGŐ and HORVÁTH · Limit Theorems in Change Point Analysis
DAGPUNAR · Simulation and Monte Carlo: With Applications in Finance and MCMC
DANIEL · Applications of Statistics to Industrial Experimentation
DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*
* DANIEL · Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
DASU and JOHNSON · Exploratory Data Mining and Data Cleaning
DAVID and NAGARAJA · Order Statistics, *Third Edition*
DAVINO, FURNO and VISTOCCO · Quantile Regression: Theory and Applications
* DEGROOT, FIENBERG, and KADANE · Statistics and the Law
DEL CASTILLO · Statistical Process Adjustment for Quality Control
DEMARIS · Regression with Social Data: Modeling Continuous and Limited Response Variables
DEMIDENKO · Mixed Models: Theory and Applications with R, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley Interscience Paperback Series.

DENISON, HOLMES, MALLICK, and SMITH · Bayesian Methods for Nonlinear Classification and Regression
DETTE and STUDDEN · The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
DEY and MUKERJEE · Fractional Factorial Plans
DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications
* DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*
* DOOB · Stochastic Processes
DOWDY, WEARDEN, and CHILKO · Statistics for Research, *Third Edition*
DRAPER and SMITH · Applied Regression Analysis, *Third Edition*
DRYDEN and MARDIA · Statistical Shape Analysis
DUDEWICZ and MISHRA · Modern Mathematical Statistics
DUNN and CLARK · Basic Statistics: A Primer for the Biomedical Sciences, *Fourth Edition*
DUPUIS and ELLIS · A Weak Convergence Approach to the Theory of Large Deviations
EDLER and KITSOS · Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment
* ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis
ENDERS · Applied Econometric Time Series, *Third Edition*
† ETHIER and KURTZ · Markov Processes: Characterization and Convergence
EVANS, HASTINGS, and PEACOCK · Statistical Distributions, *Third Edition*
EVERITT, LANDAU, LEESE, and STAHL · Cluster Analysis, *Fifth Edition*
FEDERER and KING · Variations on Split Plot and Split Block Experiment Designs
FELLER · An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition,* Revised; Volume II, *Second Edition*
FITZMAURICE, LAIRD, and WARE · Applied Longitudinal Analysis, *Second Edition*
* FLEISS · The Design and Analysis of Clinical Experiments
FLEISS · Statistical Methods for Rates and Proportions, *Third Edition*
† FLEMING and HARRINGTON · Counting Processes and Survival Analysis
FUJIKOSHI, ULYANOV, and SHIMIZU · Multivariate Statistics: High-Dimensional and Large-Sample Approximations
FULLER · Introduction to Statistical Time Series, *Second Edition*
† FULLER · Measurement Error Models
GALLANT · Nonlinear Statistical Models
GEISSER · Modes of Parametric Statistical Inference
GELMAN and MENG · Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives
GEWEKE · Contemporary Bayesian Econometrics and Statistics
GHOSH, MUKHOPADHYAY, and SEN · Sequential Estimation
GIESBRECHT and GUMPERTZ · Planning, Construction, and Statistical Analysis of Comparative Experiments
GIFI · Nonlinear Multivariate Analysis
GIVENS and HOETING · Computational Statistics
GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems
GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*
GOLDSTEIN · Multilevel Statistical Models, *Fourth Edition*
GOLDSTEIN and LEWIS · Assessment: Problems, Development, and Statistical Issues
GOLDSTEIN and WOOFF · Bayes Linear Statistics

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley Interscience Paperback Series.

GRAHAM · Markov Chains: Analytic and Monte Carlo Computations
GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing
GROSS, SHORTLE, THOMPSON, and HARRIS · Fundamentals of Queueing Theory, *Fourth Edition*
GROSS, SHORTLE, THOMPSON, and HARRIS · Solutions Manual to Accompany Fundamentals of Queueing Theory, *Fourth Edition*
* HAHN and SHAPIRO · Statistical Models in Engineering
HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners
HALD · A History of Probability and Statistics and their Applications Before 1750
† HAMPEL · Robust Statistics: The Approach Based on Influence Functions
HARTUNG, KNAPP, and SINHA · Statistical Meta-Analysis with Applications
HEIBERGER · Computation for the Analysis of Designed Experiments
HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
HEDEKER and GIBBONS · Longitudinal Data Analysis
HELLER · MACSYMA for Statisticians
HERITIER, CANTONI, COPT, and VICTORIA-FESER · Robust Methods in Biostatistics
HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design, *Second Edition*
HINKELMANN and KEMPTHORNE · Design and Analysis of Experiments, Volume 2: Advanced Experimental Design
HINKELMANN (editor) · Design and Analysis of Experiments, Volume 3: Special Designs and Applications
HOAGLIN, MOSTELLER, and TUKEY · Fundamentals of Exploratory Analysis of Variance
* HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes
* HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis
HOCHBERG and TAMHANE · Multiple Comparison Procedures
HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Third Edition*
HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
HOGG and KLUGMAN · Loss Distributions
HOLLANDER, WOLFE, and CHICKEN · Nonparametric Statistical Methods, *Third Edition*
HOSMER and LEMESHOW · Applied Logistic Regression, *Second Edition*
HOSMER, LEMESHOW, and MAY · Applied Survival Analysis: Regression Modeling of Time-to-Event Data, *Second Edition*
HUBER · Data Analysis: What Can Be Learned From the Past 50 Years
HUBER · Robust Statistics
† HUBER and RONCHETTI · Robust Statistics, *Second Edition*
HUBERTY · Applied Discriminant Analysis, *Second Edition*
HUBERTY and OLEJNIK · Applied MANOVA and Discriminant Analysis, *Second Edition*
HUITEMA · The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, *Second Edition*
HUNT and KENNEDY · Financial Derivatives in Theory and Practice, *Revised Edition*
HURD and MIAMEE · Periodically Correlated Random Sequences: Spectral Theory and Practice
HUSKOVA, BERAN, and DUPAC · Collected Works of Jaroslav Hajek— with Commentary
HUZURBAZAR · Flowgraph Models for Multistate Time-to-Event Data

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley Interscience Paperback Series.

JACKMAN · Bayesian Analysis for the Social Sciences
† JACKSON · A User's Guide to Principle Components
JOHN · Statistical Methods in Engineering and Quality Assurance
JOHNSON · Multivariate Statistical Simulation
JOHNSON and BALAKRISHNAN · Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
JOHNSON, KEMP, and KOTZ · Univariate Discrete Distributions, *Third Edition*
JOHNSON and KOTZ (editors) · Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 1, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Continuous Univariate Distributions, Volume 2, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN · Discrete Multivariate Distributions
JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE · The Theory and Practice of Econometrics, *Second Edition*
JUREK and MASON · Operator-Limit Distributions in Probability Theory
KADANE · Bayesian Methods and Ethics in a Clinical Trial Design
KADANE AND SCHUM · A Probabilistic Analysis of the Sacco and Vanzetti Evidence
KALBFLEISCH and PRENTICE · The Statistical Analysis of Failure Time Data, *Second Edition*
KARIYA and KURATA · Generalized Least Squares
KASS and VOS · Geometrical Foundations of Asymptotic Inference
† KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis
KEDEM and FOKIANOS · Regression Models for Time Series Analysis
KENDALL, BARDEN, CARNE, and LE · Shape and Shape Theory
KHURI · Advanced Calculus with Applications in Statistics, *Second Edition*
KHURI, MATHEW, and SINHA · Statistical Tests for Mixed Linear Models
* KISH · Statistical Design for Research
KLEIBER and KOTZ · Statistical Size Distributions in Economics and Actuarial Sciences
KLEMELÄ · Smoothing of Multivariate Data: Density Estimation and Visualization
KLUGMAN, PANJER, and WILLMOT · Loss Models: From Data to Decisions, *Third Edition*
KLUGMAN, PANJER, and WILLMOT · Loss Models: Further Topics
KLUGMAN, PANJER, and WILLMOT · Solutions Manual to Accompany Loss Models: From Data to Decisions, *Third Edition*
KOSKI and NOBLE · Bayesian Networks: An Introduction
KOTZ, BALAKRISHNAN, and JOHNSON · Continuous Multivariate Distributions, Volume 1, *Second Edition*
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
KOTZ and JOHNSON (editors) · Encyclopedia of Statistical Sciences: Supplement Volume
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 1
KOTZ, READ, and BANKS (editors) · Encyclopedia of Statistical Sciences: Update Volume 2
KOWALSKI and TU · Modern Applied U-Statistics
KRISHNAMOORTHY and MATHEW · Statistical Tolerance Regions: Theory, Applications, and Computation

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley Interscience Paperback Series.

KROESE, TAIMRE, and BOTEV · Handbook of Monte Carlo Methods
KROONENBERG · Applied Multiway Data Analysis
KULINSKAYA, MORGENTHALER, and STAUDTE · Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence
KULKARNI and HARMAN · An Elementary Introduction to Statistical Learning Theory
KUROWICKA and COOKE · Uncertainty Analysis with High Dimensional Dependence Modelling
KVAM and VIDAKOVIC · Nonparametric Statistics with Applications to Science and Engineering
LACHIN · Biostatistical Methods: The Assessment of Relative Risks, *Second Edition*
LAD · Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*
LAWLESS · Statistical Models and Methods for Lifetime Data, *Second Edition*
LAWSON · Statistical Methods in Spatial Epidemiology, *Second Edition*
LE · Applied Categorical Data Analysis, *Second Edition*
LE · Applied Survival Analysis
LEE · Structural Equation Modeling: A Bayesian Approach
LEE and WANG · Statistical Methods for Survival Data Analysis, *Fourth Edition*
LEPAGE and BILLARD · Exploring the Limits of Bootstrap
LESSLER and KALSBEEK · Nonsampling Errors in Surveys
LEYLAND and GOLDSTEIN (editors) · Multilevel Modelling of Health Statistics
LIAO · Statistical Group Comparison
LIN · Introductory Stochastic Analysis for Finance and Insurance
LINDLEY · Understanding Uncertainty, *Revised Edition*
LITTLE and RUBIN · Statistical Analysis with Missing Data, *Second Edition*
LLOYD · The Statistical Analysis of Categorical Data
LOWEN and TEICH · Fractal-Based Point Processes
MAGNUS and NEUDECKER · Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*
MALLER and ZHOU · Survival Analysis with Long Term Survivors
MARCHETTE · Random Graphs for Statistical Pattern Recognition
MARDIA and JUPP · Directional Statistics
MARKOVICH · Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice
MARONNA, MARTIN and YOHAI · Robust Statistics: Theory and Methods
MASON, GUNST, and HESS · Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*
McCULLOCH, SEARLE, and NEUHAUS · Generalized, Linear, and Mixed Models, *Second Edition*
McFADDEN · Management of Data in Clinical Trials, *Second Edition*
* McLACHLAN · Discriminant Analysis and Statistical Pattern Recognition
McLACHLAN, DO, and AMBROISE · Analyzing Microarray Gene Expression Data
McLACHLAN and KRISHNAN · The EM Algorithm and Extensions, *Second Edition*
McLACHLAN and PEEL · Finite Mixture Models
McNEIL · Epidemiological Research Methods
MEEKER and ESCOBAR · Statistical Methods for Reliability Data
MEERSCHAERT and SCHEFFLER · Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
MENGERSEN, ROBERT, and TITTERINGTON · Mixtures: Estimation and Applications

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley Interscience Paperback Series.

MICKEY, DUNN, and CLARK · Applied Statistics: Analysis of Variance and Regression, *Third Edition*

* MILLER · Survival Analysis, *Second Edition*

MONTGOMERY, JENNINGS, and KULAHCI · Introduction to Time Series Analysis and Forecasting, *Second Edition*

MONTGOMERY, PECK, and VINING · Introduction to Linear Regression Analysis, *Fifth Edition*

MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness

MUIRHEAD · Aspects of Multivariate Statistical Theory

MULLER and STOYAN · Comparison Methods for Stochastic Models and Risks

MURTHY, XIE, and JIANG · Weibull Models

MYERS, MONTGOMERY, and ANDERSON-COOK · Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Third Edition*

MYERS, MONTGOMERY, VINING, and ROBINSON · Generalized Linear Models. With Applications in Engineering and the Sciences, *Second Edition*

NATVIG · Multistate Systems Reliability Theory With Applications

† NELSON · Accelerated Testing, Statistical Models, Test Plans, and Data Analyses

† NELSON · Applied Life Data Analysis

NEWMAN · Biostatistical Methods in Epidemiology

NG, TAIN, and TANG · Dirichlet Theory: Theory, Methods and Applications

OKABE, BOOTS, SUGIHARA, and CHIU · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*

OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis

PALTA · Quantitative Methods in Population Health: Extensions of Ordinary Regressions

PANJER · Operational Risk: Modeling and Analytics

PANKRATZ · Forecasting with Dynamic Regression Models

PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases

PARDOUX · Markov Processes and Applications: Algorithms, Networks, Genome and Finance

PARMIGIANI and INOUE · Decision Theory: Principles and Approaches

* PARZEN · Modern Probability Theory and Its Applications

PEÑA, TIAO, and TSAY · A Course in Time Series Analysis

PESARIN and SALMASO · Permutation Tests for Complex Data: Applications and Software

PIANTADOSI · Clinical Trials: A Methodologic Perspective, *Second Edition*

POURAHMADI · Foundations of Time Series Analysis and Prediction Theory

POURAHMADI · High-Dimensional Covariance Estimation

POWELL · Approximate Dynamic Programming: Solving the Curses of Dimensionality, *Second Edition*

POWELL and RYZHOV · Optimal Learning

PRESS · Subjective and Objective Bayesian Statistics, *Second Edition*

PRESS and TANUR · The Subjectivity of Scientists and the Bayesian Approach

PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics

† PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming

QIU · Image Processing and Jump Regression Analysis

* RAO · Linear Statistical Inference and Its Applications, *Second Edition*

RAO · Statistical Inference for Fractional Diffusion Processes

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley Interscience Paperback Series.

RAUSAND and HØYLAND · System Reliability Theory: Models, Statistical Methods, and Applications, *Second Edition*
RAYNER, THAS, and BEST · Smooth Tests of Goodnes of Fit: Using R, *Second Edition*
RENCHER and SCHAALJE · Linear Models in Statistics, *Second Edition*
RENCHER and CHRISTENSEN · Methods of Multivariate Analysis, *Third Edition*
RENCHER · Multivariate Statistical Inference with Applications
RIGDON and BASU · Statistical Methods for the Reliability of Repairable Systems
* RIPLEY · Spatial Statistics
* RIPLEY · Stochastic Simulation
ROHATGI and SALEH · An Introduction to Probability and Statistics, *Third Edition*
ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS · Stochastic Processes for Insurance and Finance
ROSENBERGER and LACHIN · Randomization in Clinical Trials: Theory and Practice
ROSSI, ALLENBY, and McCULLOCH · Bayesian Statistics and Marketing
† ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
ROYSTON and SAUERBREI · Multivariate Model Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modeling Continuous Variables
* RUBIN · Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN and KROESE · Simulation and the Monte Carlo Method, *Second Edition*
RUBINSTEIN and MELAMED · Modern Simulation and Modeling
RUBINSTEIN, RIDDER, and VAISMAN · Fast Sequential Monte Carlo Methods for Counting and Optimization
RYAN · Modern Engineering Statistics
RYAN · Modern Experimental Design
RYAN · Modern Regression Methods, *Second Edition*
RYAN · Sample Size Determination and Power
RYAN · Statistical Methods for Quality Improvement, *Third Edition*
SALEH · Theory of Preliminary Test and Stein-Type Estimation with Applications
SALTELLI, CHAN, and SCOTT (editors) · Sensitivity Analysis
SCHERER · Batch Effects and Noise in Microarray Experiments: Sources and Solutions
* SCHEFFE · The Analysis of Variance
SCHIMEK · Smoothing and Regression: Approaches, Computation, and Application
SCHOTT · Matrix Analysis for Statistics, *Second Edition*
SCHOUTENS · Levy Processes in Finance: Pricing Financial Derivatives
SCOTT · Multivariate Density Estimation
SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
* SEARLE · Linear Models
† SEARLE · Linear Models for Unbalanced Data
† SEARLE · Matrix Algebra Useful for Statistics
† SEARLE, CASELLA, and McCULLOCH · Variance Components
SEARLE and WILLETT · Matrix Algebra for Applied Economics
SEBER · A Matrix Handbook For Statisticians
† SEBER · Multivariate Observations
SEBER and LEE · Linear Regression Analysis, *Second Edition*
† SEBER and WILD · Nonlinear Regression
SENNOTT · Stochastic Dynamic Programming and the Control of Queueing

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley Interscience Paperback Series.

Systems
* SERFLING · Approximation Theorems of Mathematical Statistics
SHAFER and VOVK · Probability and Finance: It's Only a Game!
SHERMAN · Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties
SILVAPULLE and SEN · Constrained Statistical Inference: Inequality, Order, and Shape Restrictions
SINGPURWALLA · Reliability and Risk: A Bayesian Perspective
SMALL and McLEISH · Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA · Methods of Multivariate Statistics
STAPLETON · Linear Statistical Models, *Second Edition*
STAPLETON · Models for Probability and Statistical Inference: Theory and Applications
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYAN · Counterexamples in Probability, *Second Edition*
STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
STREET and BURGESS · The Construction of Optimal Stated Choice Experiments: Theory and Methods
STYAN · The Collected Papers of T. W. Anderson: 1943–1985
SUTTON, ABRAMS, JONES, SHELDON, and SONG · Methods for Meta-Analysis in Medical Research
TAKEZAWA · Introduction to Nonparametric Regression
TAMHANE · Statistical Analysis of Designed Experiments: Theory and Applications
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON · Empirical Model Building: Data, Models, and Reality, *Second Edition*
THOMPSON · Sampling, *Third Edition*
THOMPSON · Simulation: A Modeler's Approach
THOMPSON and SEBER · Adaptive Sampling
THOMPSON, WILLIAMS, and FINDLAY · Models for Investors in Real World Markets
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing
and Dynamic Graphics
TROFFAES and DE COOMAN · Lower Previsions
TSAY · Analysis of Financial Time Series, *Third Edition*
TSAY · An Introduction to Analysis of Financial Data with R
TSAY · Multivariate Time Series Analysis: With R and Financial Applications
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
† VAN BELLE · Statistical Rules of Thumb, *Second Edition*
VAN BELLE, FISHER, HEAGERTY, and LUMLEY · Biostatistics: A Methodology for the Health Sciences, *Second Edition*
VESTRUP · The Theory of Measures and Integration
VIDAKOVIC · Statistical Modeling by Wavelets
VIERTL · Statistical Methods for Fuzzy Data
VINOD and REAGLE · Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments
WALLER and GOTWAY · Applied Spatial Statistics for Public Health Data
WEISBERG · Applied Linear Regression, *Fourth Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley Interscience Paperback Series.

WEISBERG · Bias and Causation: Models and Judgment for Valid Comparisons
WELSH · Aspects of Statistical Inference
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment
* WHITTAKER · Graphical Models in Applied Multivariate Statistics
WINKER · Optimization Heuristics in Economics: Applications of Threshold Accepting
WOODWORTH · Biostatistics: A Bayesian Introduction
WOOLSON and CLARKE · Statistical Methods for the Analysis of Biomedical Data, *Second Edition*
WU and HAMADA · Experiments: Planning, Analysis, and Parameter Design Optimization, *Second Edition*
WU and ZHANG · Nonparametric Regression Methods for Longitudinal Data Analysis
YAKIR · Extremes in Random Fields
YIN · Clinical Trial Design: Bayesian and Frequentist Adaptive Methods
YOUNG, VALERO-MORA, and FRIENDLY · Visual Statistics: Seeing Data with Dynamic Interactive Graphics
ZACKS · Examples and Problems in Mathematical Statistics
ZACKS · Stage-Wise Adaptive Designs
* ZELLNER · An Introduction to Bayesian Inference in Econometrics
ZELTERMAN · Discrete Distributions Applications in the Health Sciences
ZHOU, OBUCHOWSKI, and McCLISH · Statistical Methods in Diagnostic Medicine, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley Interscience Paperback Series.

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.