Springer Texts in Statistics

Advisors:
George Casella Stephen Fienberg Ingram Olkin

Springer
New York
Berlin
Heidelberg
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Peter J. Brockwell Richard A. Davis

Introduction to Time Series and Forecasting

Second Edition

With 126 Illustrations

Includes CD-ROM

Springer

Peter J. Brockwell
Department of Statistics
Colorado State University
Fort Collins, CO 80523
USA
pjbrock@stat.colostate.edu

Richard A. Davis
Department of Statistics
Colorado State University
Fort Collins, CO 80523
USA
rdavis@stat.colostate.edu

Library of Congress Cataloging-in-Publication Data
Brockwell, Peter J.
Introduction to time series and forecasting / Peter J. Brockwell and Richard A. Davis.—2nd ed.
p. cm. — (Springer texts in statistics)
Includes bibliographical references and index.
ISBN 0-387-95351-5 (alk. paper)
1. Time-series analysis. I. Davis, Richard A. II. Title. III. Series.
QA280.B757 2002
519.5′5—dc21 2001049262

Printed on acid-free paper.

Production managed by MaryAnn Brickner; manufacturing supervised by Joe Quatela.
Typeset by The Bartlett Press, Inc., Marietta, GA.
Printed and bound by R.R. Donnelley and Sons, Harrisonburg, VA.
Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-95351-5 SPIN 10850334

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

To Pam and Patti

Preface

This book is aimed at the reader who wishes to gain a working knowledge of time series and forecasting methods as applied in economics, engineering and the natural and social sciences. Unlike our earlier book, *Time Series*: *Theory and Methods*, referred to in the text as TSTM, this one requires only a knowledge of basic calculus, matrix algebra and elementary statistics at the level (for example) of Mendenhall, Wackerly and Scheaffer (1990). It is intended for upper-level undergraduate students and beginning graduate students.

The emphasis is on methods and the analysis of data sets. The student version of the time series package ITSM2000, enabling the reader to reproduce most of the calculations in the text (and to analyze further data sets of the reader's own choosing), is included on the CD-ROM which accompanies the book. The data sets used in the book are also included. The package requires an IBM-compatible PC operating under Windows 95, NT version 4.0, or a later version of either of these operating systems. The program ITSM can be run directly from the CD-ROM or installed on a hard disk as described at the beginning of Appendix D, where a detailed introduction to the package is provided.

Very little prior familiarity with computing is required in order to use the computer package. Detailed instructions for its use are found in the on-line help files which are accessed, when the program ITSM is running, by selecting the menu option `Help>Contents` and selecting the topic of interest. Under the heading `Data` you will find information concerning the data sets stored on the CD-ROM. The book can also be used in conjunction with other computer packages for handling time series. Chapter 14 of the book by Venables and Ripley (1994) describes how to perform many of the calculations using S-plus.

There are numerous problems at the end of each chapter, many of which involve use of the programs to study the data sets provided.

To make the underlying theory accessible to a wider audience, we have stated some of the key mathematical results without proof, but have attempted to ensure that the logical structure of the development is otherwise complete. (References to proofs are provided for the interested reader.)

Since the upgrade to ITSM2000 occurred after the first edition of this book appeared, we have taken the opportunity, in this edition, to coordinate the text with the new software, to make a number of corrections pointed out by readers of the first edition and to expand on several of the topics treated only briefly in the first edition.

Appendix D, the software tutorial, has been rewritten in order to be compatible with the new version of the software.

Some of the other extensive changes occur in (i) Section 6.6, which highlights the role of the innovations algorithm in generalized least squares and maximum likelihood estimation of regression models with time series errors, (ii) Section 6.4, where the treatment of forecast functions for ARIMA processes has been expanded and (iii) Section 10.3, which now includes GARCH modeling and simulation, topics of considerable importance in the analysis of financial time series. The new material has been incorporated into the accompanying software, to which we have also added the option `Autofit`. This streamlines the modeling of time series data by fitting maximum likelihood ARMA(p, q) models for a specified range of (p, q) values and automatically selecting the model with smallest AICC value.

There is sufficient material here for a full-year introduction to univariate and multivariate time series and forecasting. Chapters 1 through 6 have been used for several years in introductory one-semester courses in univariate time series at Colorado State University and Royal Melbourne Institute of Technology. The chapter on spectral analysis can be excluded without loss of continuity by readers who are so inclined.

We are greatly indebted to the readers of the first edition and especially to Matthew Calder, coauthor of the new computer package, and Anthony Brockwell for their many valuable comments and suggestions. We also wish to thank Colorado State University, the National Science Foundation, Springer-Verlag and our families for their continuing support during the preparation of this second edition.

Fort Collins, Colorado
August 2001

Peter J. Brockwell
Richard A. Davis

Contents

Preface **vii**

1. Introduction **1**

1.1. Examples of Time Series 1
1.2. Objectives of Time Series Analysis 6
1.3. Some Simple Time Series Models 7
1.3.1. Some Zero-Mean Models 8
1.3.2. Models with Trend and Seasonality 9
1.3.3. A General Approach to Time Series Modeling 14
1.4. Stationary Models and the Autocorrelation Function 15
1.4.1. The Sample Autocorrelation Function 18
1.4.2. A Model for the Lake Huron Data 21
1.5. Estimation and Elimination of Trend and Seasonal Components 23
1.5.1. Estimation and Elimination of Trend in the Absence of Seasonality 24
1.5.2. Estimation and Elimination of Both Trend and Seasonality 31
1.6. Testing the Estimated Noise Sequence 35
Problems 40

2. Stationary Processes **45**

2.1. Basic Properties 45
2.2. Linear Processes 51
2.3. Introduction to ARMA Processes 55
2.4. Properties of the Sample Mean and Autocorrelation Function 57
2.4.1. Estimation of μ 58
2.4.2. Estimation of $\gamma(\cdot)$ and $\rho(\cdot)$ 59
2.5. Forecasting Stationary Time Series 63
2.5.1. The Durbin–Levinson Algorithm 69
2.5.2. The Innovations Algorithm 71
2.5.3. Prediction of a Stationary Process in Terms of Infinitely Many Past Values 75

2.6. The Wold Decomposition 77
Problems 78

3. ARMA Models 83

3.1. ARMA(p, q) Processes 83
3.2. The ACF and PACF of an ARMA(p, q) Process 88
3.2.1. Calculation of the ACVF 88
3.2.2. The Autocorrelation Function 94
3.2.3. The Partial Autocorrelation Function 94
3.2.4. Examples 96
3.3. Forecasting ARMA Processes 100
Problems 108

4. Spectral Analysis 111

4.1. Spectral Densities 112
4.2. The Periodogram 121
4.3. Time-Invariant Linear Filters 127
4.4. The Spectral Density of an ARMA Process 132
Problems 134

5. Modeling and Forecasting with ARMA Processes 137

5.1. Preliminary Estimation 138
5.1.1. Yule–Walker Estimation 139
5.1.2. Burg's Algorithm 147
5.1.3. The Innovations Algorithm 150
5.1.4. The Hannan–Rissanen Algorithm 156
5.2. Maximum Likelihood Estimation 158
5.3. Diagnostic Checking 164
5.3.1. The Graph of $\{\hat{R}_t, t = 1, \ldots, n\}$ 165
5.3.2. The Sample ACF of the Residuals 166
5.3.3. Tests for Randomness of the Residuals 166
5.4. Forecasting 167
5.5. Order Selection 169
5.5.1. The FPE Criterion 170
5.5.2. The AICC Criterion 171
Problems 174

6. Nonstationary and Seasonal Time Series Models 179

6.1. ARIMA Models for Nonstationary Time Series 180
6.2. Identification Techniques 187

6.3. Unit Roots in Time Series Models 193
6.3.1. Unit Roots in Autoregressions 194
6.3.2. Unit Roots in Moving Averages 196
6.4. Forecasting ARIMA Models 198
6.4.1. The Forecast Function 200
6.5. Seasonal ARIMA Models 203
6.5.1. Forecasting SARIMA Processes 208
6.6. Regression with ARMA Errors 210
6.6.1. OLS and GLS Estimation 210
6.6.2. ML Estimation 213
Problems 219

7. Multivariate Time Series 223

7.1. Examples 224
7.2. Second-Order Properties of Multivariate Time Series 229
7.3. Estimation of the Mean and Covariance Function 234
7.3.1. Estimation of $\boldsymbol{\mu}$ 234
7.3.2. Estimation of $\Gamma(h)$ 235
7.3.3. Testing for Independence of Two Stationary Time Series 237
7.3.4. Bartlett's Formula 238
7.4. Multivariate ARMA Processes 241
7.4.1. The Covariance Matrix Function of a Causal ARMA Process 244
7.5. Best Linear Predictors of Second-Order Random Vectors 244
7.6. Modeling and Forecasting with Multivariate AR Processes 246
7.6.1. Estimation for Autoregressive Processes Using Whittle's Algorithm 247
7.6.2. Forecasting Multivariate Autoregressive Processes 250
7.7. Cointegration 254
Problems 256

8. State-Space Models 259

8.1. State-Space Representations 260
8.2. The Basic Structural Model 263
8.3. State-Space Representation of ARIMA Models 267
8.4. The Kalman Recursions 271
8.5. Estimation For State-Space Models 277
8.6. State-Space Models with Missing Observations 283
8.7. The EM Algorithm 289
8.8. Generalized State-Space Models 292
8.8.1. Parameter-Driven Models 292

8.8.2. Observation-Driven Models 299
Problems 311

9. Forecasting Techniques 317

9.1. The ARAR Algorithm 318
9.1.1. Memory Shortening 318
9.1.2. Fitting a Subset Autoregression 319
9.1.3. Forecasting 320
9.1.4. Application of the ARAR Algorithm 321
9.2. The Holt–Winters Algorithm 322
9.2.1. The Algorithm 322
9.2.2. Holt–Winters and ARIMA Forecasting 324
9.3. The Holt–Winters Seasonal Algorithm 326
9.3.1. The Algorithm 326
9.3.2. Holt–Winters Seasonal and ARIMA Forecasting 328
9.4. Choosing a Forecasting Algorithm 328
Problems 330

10. Further Topics 331

10.1. Transfer Function Models 331
10.1.1. Prediction Based on a Transfer Function Model 337
10.2. Intervention Analysis 340
10.3. Nonlinear Models 343
10.3.1. Deviations from Linearity 344
10.3.2. Chaotic Deterministic Sequences 345
10.3.3. Distinguishing Between White Noise and iid Sequences 347
10.3.4. Three Useful Classes of Nonlinear Models 348
10.3.5. Modeling Volatility 349
10.4. Continuous-Time Models 357
10.5. Long-Memory Models 361
Problems 365

A. Random Variables and Probability Distributions 369

A.1. Distribution Functions and Expectation 369
A.2. Random Vectors 374
A.3. The Multivariate Normal Distribution 377
Problems 381

B. Statistical Complements 383

B.1. Least Squares Estimation 383
B.1.1. The Gauss-Markov Theorem 385
B.1.2. Generalized Least Squares 386
B.2. Maximum Likelihood Estimation 386
B.2.1. Properties of Maximum Likelihood Estimators 387
B.3. Confidence Intervals 388
B.3.1. Large-Sample Confidence Regions 388
B.4. Hypothesis Testing 389
B.4.1. Error Probabilities 390
B.4.2. Large-Sample Tests Based on Confidence Regions 390

C. Mean Square Convergence 393

C.1. The Cauchy Criterion 393

D. An ITSM Tutorial 395

D.1. Getting Started 396
D.1.1. Running ITSM 396
D.2. Preparing Your Data for Modeling 396
D.2.1. Entering Data 397
D.2.2. Information 397
D.2.3. Filing Data 397
D.2.4. Plotting Data 398
D.2.5. Transforming Data 398
D.3. Finding a Model for Your Data 403
D.3.1. Autofit 403
D.3.2. The Sample ACF and PACF 403
D.3.3. Entering a Model 404
D.3.4. Preliminary Estimation 406
D.3.5. The AICC Statistic 408
D.3.6. Changing Your Model 408
D.3.7. Maximum Likelihood Estimation 409
D.3.8. Optimization Results 410
D.4. Testing Your Model 411
D.4.1. Plotting the Residuals 412
D.4.2. ACF/PACF of the Residuals 412
D.4.3. Testing for Randomness of the Residuals 414
D.5. Prediction 415
D.5.1. Forecast Criteria 415
D.5.2. Forecast Results 415

D.6. Model Properties 416
D.6.1. ARMA Models 417
D.6.2. Model ACF, PACF 418
D.6.3. Model Representations 419
D.6.4. Generating Realizations of a Random Series 420
D.6.5. Spectral Properties 421
D.7. Multivariate Time Series 421

References 423

Index 429

1 Introduction

1.1 Examples of Time Series
1.2 Objectives of Time Series Analysis
1.3 Some Simple Time Series Models
1.4 Stationary Models and the Autocorrelation Function
1.5 Estimation and Elimination of Trend and Seasonal Components
1.6 Testing the Estimated Noise Sequence

In this chapter we introduce some basic ideas of time series analysis and stochastic processes. Of particular importance are the concepts of stationarity and the autocovariance and sample autocovariance functions. Some standard techniques are described for the estimation and removal of trend and seasonality (of known period) from an observed time series. These are illustrated with reference to the data sets in Section 1.1. The calculations in all the examples can be carried out using the time series package ITSM, the student version of which is supplied on the enclosed CD. The data sets are contained in files with names ending in .TSM. For example, the Australian red wine sales are filed as WINE.TSM. Most of the topics covered in this chapter will be developed more fully in later sections of the book. The reader who is not already familiar with random variables and random vectors should first read Appendix A, where a concise account of the required background is given.

1.1 Examples of Time Series

A **time series** is a set of observations x_t, each one being recorded at a specific time t. A *discrete-time time series* (the type to which this book is primarily devoted) is one in which the set T_0 of times at which observations are made is a discrete set, as is the

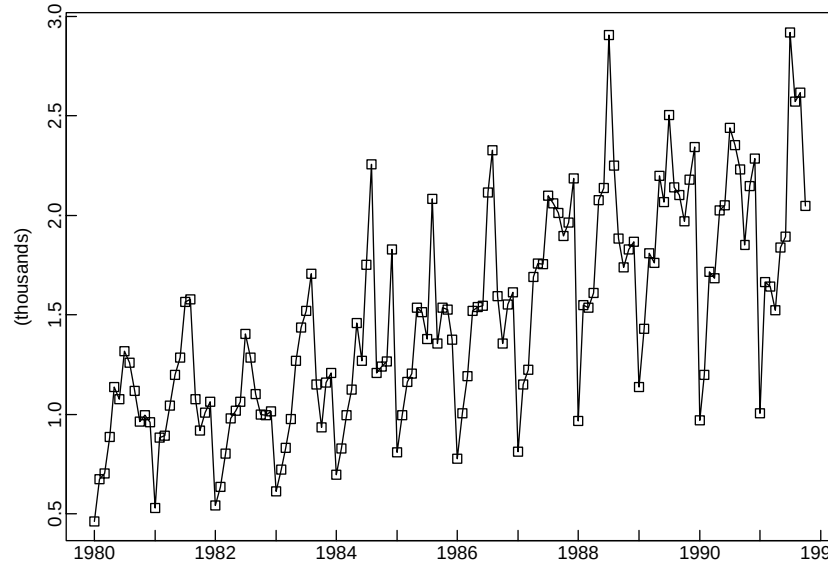

Figure 1-1
The Australian red wine sales, Jan. '80 – Oct. '91.

case, for example, when observations are made at fixed time intervals. *Continuous-time time series* are obtained when observations are recorded continuously over some time interval, e.g., when $T_0 = [0, 1]$.

Example 1.1.1 Australian red wine sales; WINE.TSM

Figure 1.1 shows the monthly sales (in kiloliters) of red wine by Australian winemakers from January 1980 through October 1991. In this case the set T_0 consists of the 142 times {(Jan. 1980), (Feb. 1980), . . . ,(Oct. 1991)}. Given a set of n observations made at uniformly spaced time intervals, it is often convenient to rescale the time axis in such a way that T_0 becomes the set of integers $\{1, 2, \ldots, n\}$. In the present example this amounts to measuring time in months with (Jan. 1980) as month 1. Then T_0 is the set $\{1, 2, \ldots, 142\}$. It appears from the graph that the sales have an upward trend and a seasonal pattern with a peak in July and a trough in January. To plot the data using ITSM, run the program by double-clicking on the ITSM icon and then select the option `File>Project>Open>Univariate`, click OK, and select the file WINE.TSM. The graph of the data will then appear on your screen. □

Example 1.1.2 All-star baseball games, 1933–1995

Figure 1.2 shows the results of the all-star games by plotting x_t, where

$$x_t = \begin{cases} 1 & \text{if the National League won in year } t, \\ -1 & \text{if the American League won in year } t. \end{cases}$$

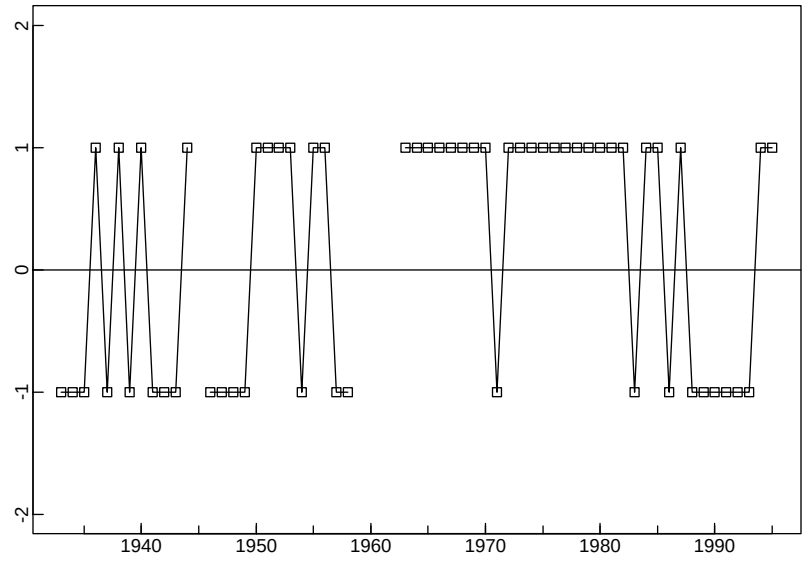

Figure 1-2
Results of the all-star baseball games, 1933–1995.

This is a series with only two possible values, ± 1. It also has some missing values, since no game was played in 1945, and two games were scheduled for each of the years 1959–1962. □

Example 1.1.3 Accidental deaths, U.S.A., 1973–1978; DEATHS.TSM

Like the red wine sales, the monthly accidental death figures show a strong seasonal pattern, with the maximum for each year occurring in July and the minimum for each year occurring in February. The presence of a trend in Figure 1.3 is much less apparent than in the wine sales. In Section 1.5 we shall consider the problem of representing the data as the sum of a trend, a seasonal component, and a residual term. □

Example 1.1.4 A signal detection problem; SIGNAL.TSM

Figure 1.4 shows simulated values of the series

$$X_t = \cos\left(\frac{t}{10}\right) + N_t, \quad t = 1, 2, \ldots, 200,$$

where $\{N_t\}$ is a sequence of independent normal random variables, with mean 0 and variance 0.25. Such a series is often referred to as *signal plus noise*, the signal being the smooth function, $S_t = \cos(\frac{t}{10})$ in this case. Given only the data X_t, how can we determine the unknown signal component? There are many approaches to this general problem under varying assumptions about the signal and the noise. One simple approach is to *smooth* the data by expressing X_t as a sum of sine waves of various frequencies (see Section 4.2) and eliminating the high-frequency components. If we do this to the values of $\{X_t\}$ shown in Figure 1.4 and retain only the lowest

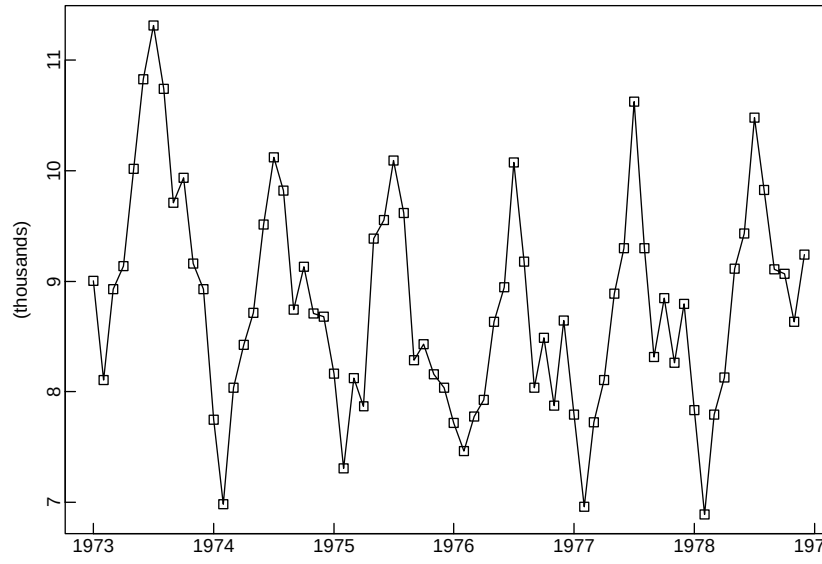

Figure 1-3
The monthly accidental deaths data, 1973–1978.

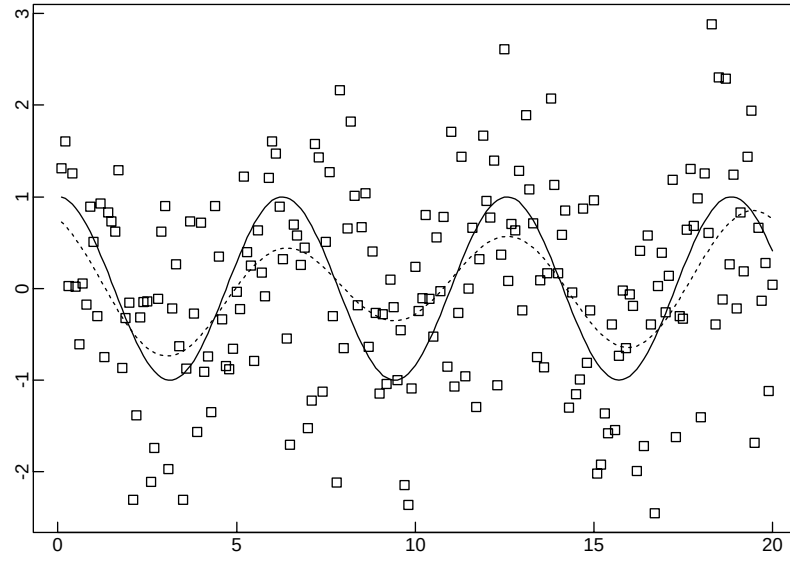

Figure 1-4
The series $\{X_t\}$ of Example 1.1.4.

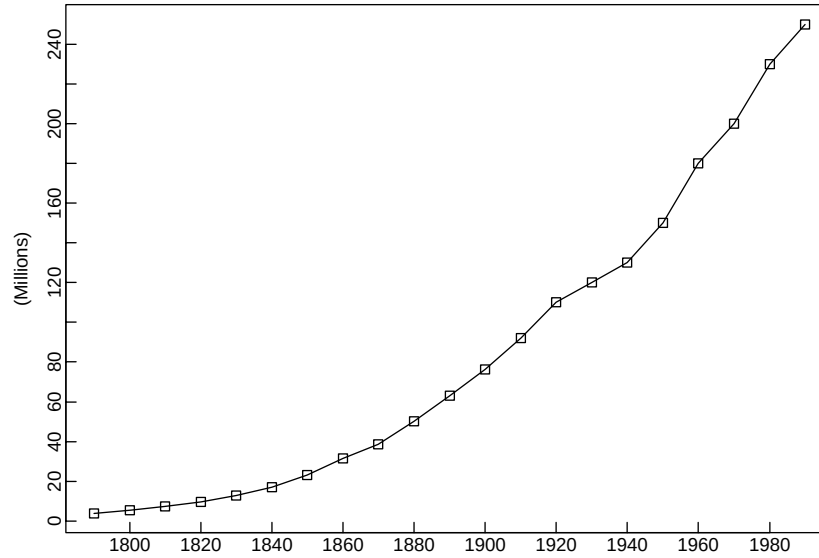

Figure 1-5
Population of the U.S.A. at ten-year intervals, 1790–1990.

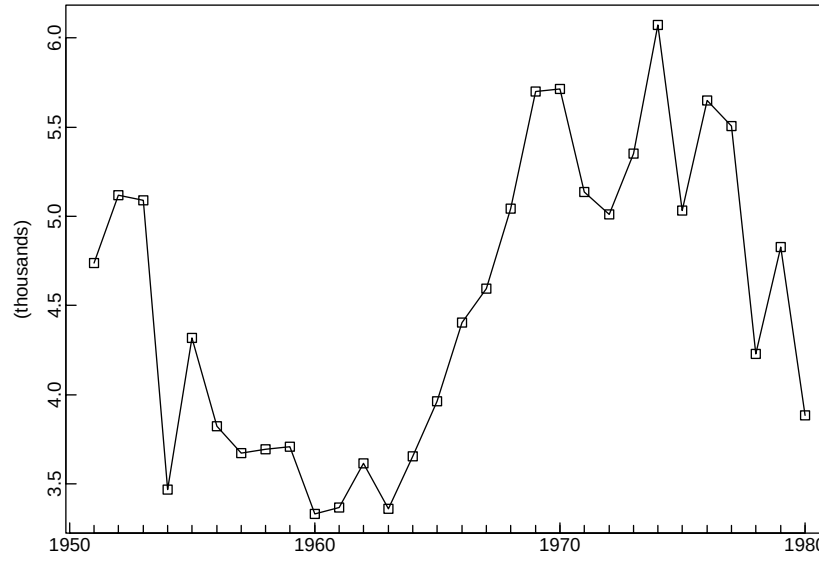

Figure 1-6
Strikes in the U.S.A., 1951–1980.

3.5% of the frequency components, we obtain the estimate of the signal also shown in Figure 1.4. The waveform of the signal is quite close to that of the true signal in this case, although its amplitude is somewhat smaller. □

Example 1.1.5 Population of the U.S.A., 1790–1990; USPOP.TSM

The population of the U.S.A., measured at ten-year intervals, is shown in Figure 1.5. The graph suggests the possibility of fitting a quadratic or exponential trend to the data. We shall explore this further in Section 1.3. □

Example 1.1.6 Number of strikes per year in the U.S.A., 1951–1980; STRIKES.TSM

The annual numbers of strikes in the U.S.A. for the years 1951–1980 are shown in Figure 1.6. They appear to fluctuate erratically about a slowly changing level. □

1.2 Objectives of Time Series Analysis

The examples considered in Section 1.1 are an extremely small sample from the multitude of time series encountered in the fields of engineering, science, sociology, and economics. Our purpose in this book is to study techniques for drawing inferences from such series. Before we can do this, however, it is necessary to set up a hypothetical probability model to represent the data. After an appropriate family of models has been chosen, it is then possible to estimate parameters, check for goodness of fit to the data, and possibly to use the fitted model to enhance our understanding of the mechanism generating the series. Once a satisfactory model has been developed, it may be used in a variety of ways depending on the particular field of application.

The model may be used simply to provide a compact description of the data. We may, for example, be able to represent the accidental deaths data of Example 1.1.3 as the sum of a specified trend, and seasonal and random terms. For the interpretation of economic statistics such as unemployment figures, it is important to recognize the presence of seasonal components and to remove them so as not to confuse them with long-term trends. This process is known as **seasonal adjustment**. Other applications of time series models include separation (or filtering) of noise from signals as in Example 1.1.4, prediction of future values of a series such as the red wine sales in Example 1.1.1 or the population data in Example 1.1.5, testing hypotheses such as global warming using recorded temperature data, predicting one series from observations of another, e.g., predicting future sales using advertising expenditure data, and controlling future values of a series by adjusting parameters. Time series models are also useful in simulation studies. For example, the performance of a reservoir depends heavily on the random daily inputs of water to the system. If these are modeled as a time series, then we can use the fitted model to simulate a large number of independent sequences of daily inputs. Knowing the size and mode of operation

of the reservoir, we can determine the fraction of the simulated input sequences that cause the reservoir to run out of water in a given time period. This fraction will then be an estimate of the probability of emptiness of the reservoir at some time in the given period.

1.3 Some Simple Time Series Models

An important part of the analysis of a time series is the selection of a suitable probability model (or class of models) for the data. To allow for the possibly unpredictable nature of future observations it is natural to suppose that each observation x_t is a realized value of a certain random variable X_t.

Definition 1.3.1

A **time series model** for the observed data $\{x_t\}$ is a specification of the joint distributions (or possibly only the means and covariances) of a sequence of random variables $\{X_t\}$ of which $\{x_t\}$ is postulated to be a realization.

Remark. We shall frequently use the term *time series* to mean both the data and the process of which it is a realization. □

A complete probabilistic time series model for the sequence of random variables $\{X_1, X_2, \ldots\}$ would specify all of the **joint distributions** of the random vectors $(X_1, \ldots, X_n)'$, $n = 1, 2, \ldots$, or equivalently all of the probabilities

$$P[X_1 \leq x_1, \ldots, X_n \leq x_n], \quad -\infty < x_1, \ldots, x_n < \infty, \quad n = 1, 2, \ldots.$$

Such a specification is rarely used in time series analysis (unless the data are generated by some well-understood simple mechanism), since in general it will contain far too many parameters to be estimated from the available data. Instead we specify only the **first-** and **second-order moments** of the joint distributions, i.e., the expected values EX_t and the expected products $E(X_{t+h}X_t)$, $t = 1, 2, \ldots$, $h = 0, 1, 2, \ldots$, focusing on properties of the sequence $\{X_t\}$ that depend only on these. Such properties of $\{X_t\}$ are referred to as **second-order properties**. In the particular case where all the joint distributions are multivariate normal, the second-order properties of $\{X_t\}$ completely determine the joint distributions and hence give a complete probabilistic characterization of the sequence. In general we shall lose a certain amount of information by looking at time series "through second-order spectacles"; however, as we shall see in Chapter 2, the theory of minimum mean squared error linear prediction depends only on the second-order properties, thus providing further justification for the use of the second-order characterization of time series models.

Figure 1.7 shows one of many possible realizations of $\{S_t, t = 1, \ldots, 200\}$, where $\{S_t\}$ is a sequence of random variables specified in Example 1.3.3 below. In most practical problems involving time series we see only *one* realization. For example,

there is only one available realization of Fort Collins's annual rainfall for the years 1900–1996, but we imagine it to be one of the many sequences that *might* have occurred. In the following examples we introduce some simple time series models. One of our goals will be to expand this repertoire so as to have at our disposal a broad range of models with which to try to match the observed behavior of given data sets.

1.3.1 Some Zero-Mean Models

Example 1.3.1 iid noise

Perhaps the simplest model for a time series is one in which there is no trend or seasonal component and in which the observations are simply independent and identically distributed (iid) random variables with zero mean. We refer to such a sequence of random variables $X_1, X_2, \ldots$ as iid noise. By definition we can write, for any positive integer n and real numbers $x_1, \ldots, x_n$,

$$P[X_1 \le x_1, \ldots, X_n \le x_n] = P[X_1 \le x_1] \cdots P[X_n \le x_n] = F(x_1) \cdots F(x_n),$$

where $F(\cdot)$ is the cumulative distribution function (see Section A.1) of each of the identically distributed random variables $X_1, X_2, \ldots$. In this model there is no dependence between observations. In particular, for all $h \ge 1$ and all $x, x_1, \ldots, x_n$,

$$P[X_{n+h} \le x | X_1 = x_1, \ldots, X_n = x_n] = P[X_{n+h} \le x],$$

showing that knowledge of $X_1, \ldots, X_n$ is of no value for predicting the behavior of X_{n+h}. Given the values of $X_1, \ldots, X_n$, the function f that minimizes the mean squared error $E\big[(X_{n+h} - f(X_1, \ldots, X_n))^2\big]$ is in fact identically zero (see Problem 1.2). Although this means that iid noise is a rather uninteresting process for forecasters, it plays an important role as a building block for more complicated time series models. □

Example 1.3.2 A binary process

As an example of iid noise, consider the sequence of iid random variables $\{X_t, t = 1, 2, \ldots, \}$ with

$$P[X_t = 1] = p, \quad P[X_t = -1] = 1 - p,$$

where $p = \frac{1}{2}$. The time series obtained by tossing a penny repeatedly and scoring $+1$ for each head and -1 for each tail is usually modeled as a realization of this process. A priori we might well consider the same process as a model for the all-star baseball games in Example 1.1.2. However, even a cursory inspection of the results from 1963–1982, which show the National League winning 19 of 20 games, casts serious doubt on the hypothesis $P[X_t = 1] = \frac{1}{2}$. □

Example 1.3.3 Random walk

The random walk $\{S_t, t = 0, 1, 2, \ldots\}$ (starting at zero) is obtained by cumulatively summing (or "integrating") iid random variables. Thus a random walk with zero mean is obtained by defining $S_0 = 0$ and

$$S_t = X_1 + X_2 + \cdots + X_t, \quad \text{for } t = 1, 2, \ldots,$$

where $\{X_t\}$ is iid noise. If $\{X_t\}$ is the binary process of Example 1.3.2, then $\{S_t, t = 0, 1, 2, \ldots, \}$ is called a **simple symmetric random walk**. This walk can be viewed as the location of a pedestrian who starts at position zero at time zero and at each integer time tosses a fair coin, stepping one unit to the right each time a head appears and one unit to the left for each tail. A realization of length 200 of a simple symmetric random walk is shown in Figure 1.7. Notice that the outcomes of the coin tosses can be recovered from $\{S_t, t = 0, 1, \ldots\}$ by differencing. Thus the result of the tth toss can be found from $S_t - S_{t-1} = X_t$. □

1.3.2 Models with Trend and Seasonality

In several of the time series examples of Section 1.1 there is a clear trend in the data. An increasing trend is apparent in both the Australian red wine sales (Figure 1.1) and the population of the U.S.A. (Figure 1.5). In both cases a zero-mean model for the data is clearly inappropriate. The graph of the population data, which contains no apparent periodic component, suggests trying a model of the form

$$X_t = m_t + Y_t,$$

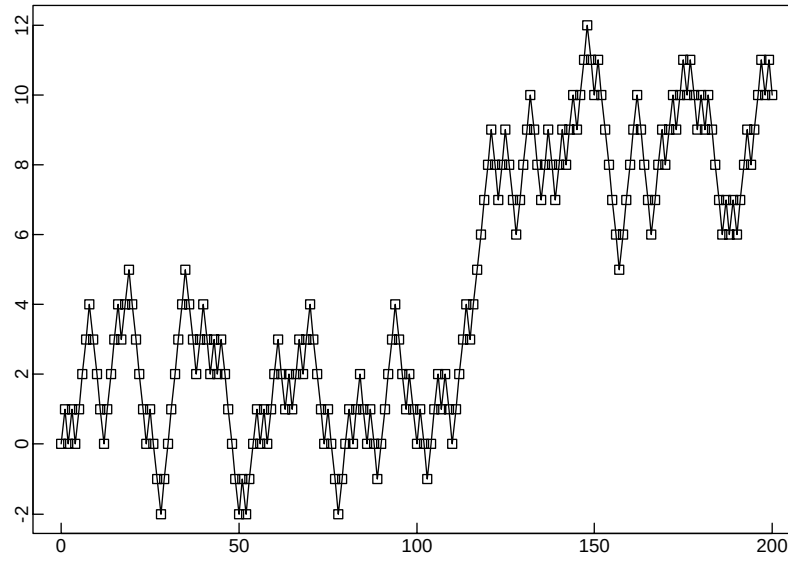

Figure 1-7
One realization of a simple random walk $\{S_t, t = 0, 1, 2, \ldots, 200\}$

where m_t is a slowly changing function known as the **trend component** and Y_t has zero mean. A useful technique for estimating m_t is the method of least squares (some other methods are considered in Section 1.5).

In the least squares procedure we attempt to fit a parametric family of functions, e.g.,

$$m_t = a_0 + a_1 t + a_2 t^2, \tag{1.3.1}$$

to the data $\{x_1, \ldots, x_n\}$ by choosing the parameters, in this illustration a_0, a_1, and a_2, to minimize $\sum_{t=1}^{n}(x_t - m_t)^2$. This method of curve fitting is called **least squares regression** and can be carried out using the program ITSM and selecting the `Regression` option.

Example 1.3.4 Population of the U.S.A., 1790–1990

To fit a function of the form (1.3.1) to the population data shown in Figure 1.5 we relabel the time axis so that $t = 1$ corresponds to 1790 and $t = 21$ corresponds to 1990. Run ITSM, select `File>Project>Open>Univariate`, and open the file US-POP.TSM. Then select `Regression>Specify`, choose `Polynomial Regression` with order equal to 2, and click OK. Then select `Regression>Estimation>Least Squares`, and you will obtain the following estimated parameter values in the model (1.3.1):

$$\hat{a}_0 = 6.9579 \times 10^6,$$

$$\hat{a}_1 = -2.1599 \times 10^6,$$

and

$$\hat{a}_2 = 6.5063 \times 10^5.$$

A graph of the fitted function is shown with the original data in Figure 1.8. The estimated values of the noise process Y_t, $1 \le t \le 21$, are the residuals obtained by subtraction of $\hat{m}_t = \hat{a}_0 + \hat{a}_1 t + \hat{a}_2 t^2$ from x_t.

The estimated trend component $\hat{m}_t$ furnishes us with a natural predictor of future values of X_t. For example, if we estimate the noise Y_{22} by its mean value, i.e., zero, then (1.3.1) gives the estimated U.S. population for the year 2000 as

$$\hat{m}_{22} = 6.9579 \times 10^6 - 2.1599 \times 10^6 \times 22 + 6.5063 \times 10^5 \times 22^2 = 274.35 \times 10^6.$$

However, if the residuals $\{Y_t\}$ are highly correlated, we may be able to use their values to give a better estimate of Y_{22} and hence of the population X_{22} in the year 2000. □

Example 1.3.5 Level of Lake Huron 1875–1972; LAKE.DAT

A graph of the level in feet of Lake Huron (reduced by 570) in the years 1875–1972 is displayed in Figure 1.9. Since the lake level appears to decline at a roughly linear rate, ITSM was used to fit a model of the form

$$X_t = a_0 + a_1 t + Y_t, \quad t = 1, \ldots, 98 \tag{1.3.2}$$

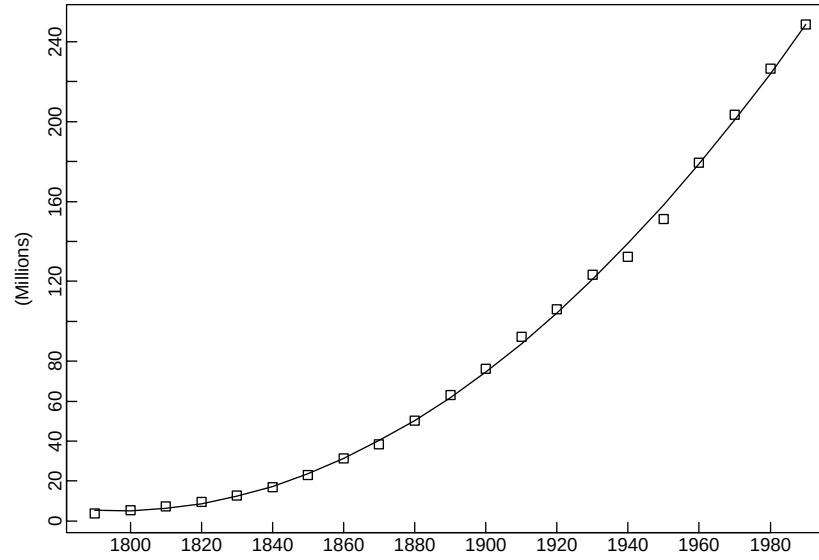

Figure 1-8
Population of the U.S.A. showing the quadratic trend fitted by least squares.

(with the time axis relabeled as in Example 1.3.4). The least squares estimates of the parameter values are

$$\hat{a}_0 = 10.202 \quad \text{and} \quad \hat{a}_1 = -.0242.$$

(The resulting least squares line, $\hat{a}_0 + \hat{a}_1 t$, is also displayed in Figure 1.9.) The estimates of the noise, Y_t, in the model (1.3.2) are the residuals obtained by subtracting the least squares line from x_t and are plotted in Figure 1.10. There are two interesting

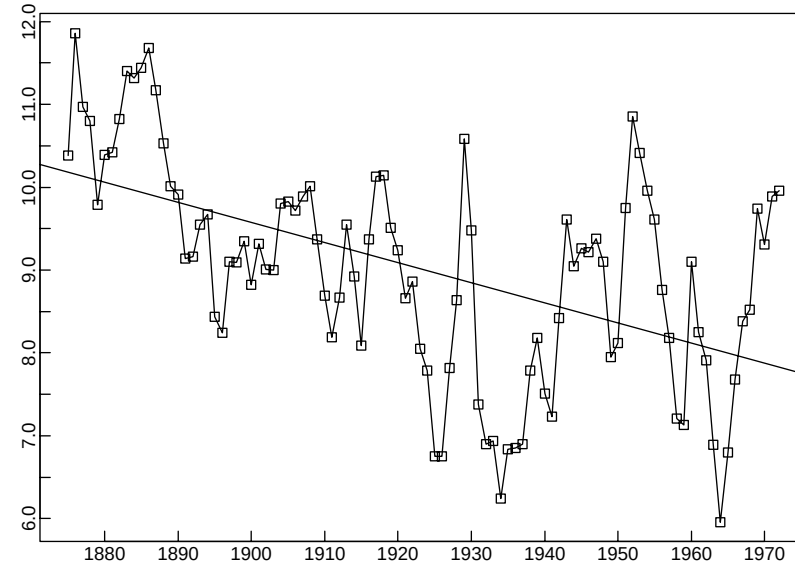

Figure 1-9
Level of Lake Huron 1875–1972 showing the line fitted by least squares.

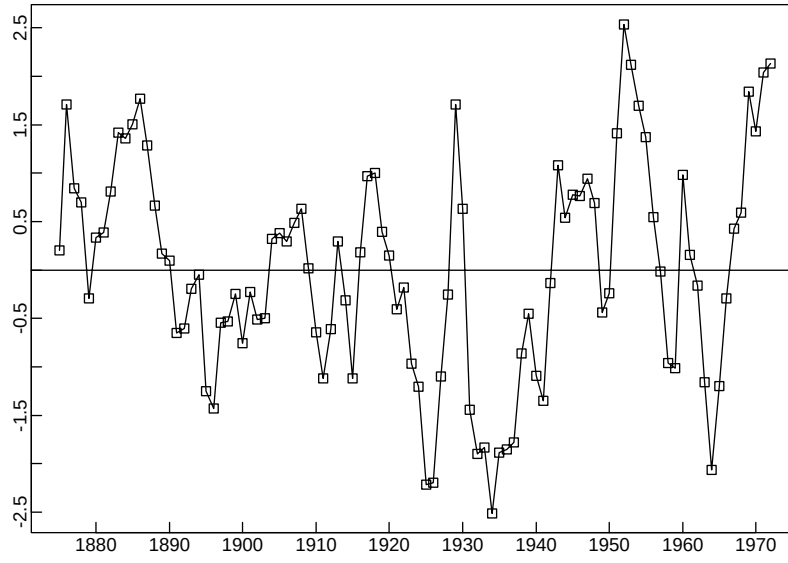

Figure 1-10
Residuals from fitting a line to the Lake Huron data in Figure 1.9.

features of the graph of the residuals. The first is the absence of any discernible trend. The second is the smoothness of the graph. (In particular, there are long stretches of residuals that have the same sign. This would be very unlikely to occur if the residuals were observations of iid noise with zero mean.) Smoothness of the graph of a time series is generally indicative of the existence of some form of dependence among the observations.

Such dependence can be used to advantage in forecasting future values of the series. If we were to assume the validity of the fitted model with iid residuals $\{Y_t\}$, then the minimum mean squared error predictor of the next residual (Y_{99}) would be zero (by Problem 1.2). However, Figure 1.10 strongly suggests that Y_{99} will be positive.

How then do we quantify dependence, and how do we construct models for forecasting that incorporate dependence of a particular type? To deal with these questions, Section 1.4 introduces the autocorrelation function as a measure of dependence, and stationary processes as a family of useful models exhibiting a wide variety of dependence structures. □

Harmonic Regression

Many time series are influenced by seasonally varying factors such as the weather, the effect of which can be modeled by a periodic component with fixed known period. For example, the accidental deaths series (Figure 1.3) shows a repeating annual pattern with peaks in July and troughs in February, strongly suggesting a seasonal factor with period 12. In order to represent such a seasonal effect, allowing for noise but assuming no trend, we can use the simple model,

$$X_t = s_t + Y_t,$$

where s_t is a periodic function of t with period d ($s_{t-d} = s_t$). A convenient choice for s_t is a sum of harmonics (or sine waves) given by

$$s_t = a_0 + \sum_{j=1}^{k} (a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t)), \tag{1.3.3}$$

where $a_0, a_1, \ldots, a_k$ and $b_1, \ldots, b_k$ are unknown parameters and $\lambda_1, \ldots, \lambda_k$ are fixed frequencies, each being some integer multiple of $2\pi/d$. To carry out harmonic regression using ITSM, select `Regression>Specify` and check `Include intercept term` and `Harmonic Regression`. Then specify the number of harmonics (k in (1.3.3)) and enter k integer-valued Fourier indices $f_1, \ldots, f_k$. For a sine wave with period d, set $f_1 = n/d$, where n is the number of observations in the time series. (If n/d is not an integer, you will need to delete a few observations from the beginning of the series to make it so.) The other $k-1$ Fourier indices should be positive integer multiples of the first, corresponding to harmonics of the fundamental sine wave with period d. Thus to fit a single sine wave with period 365 to 365 daily observations we would choose $k = 1$ and $f_1 = 1$. To fit a linear combination of sine waves with periods $365/j$, $j = 1, \ldots, 4$, we would choose $k = 4$ and $f_j = j$, $j = 1, \ldots, 4$. Once k and $f_1, \ldots, f_k$ have been specified, click OK and then select `Regression>Estimation >Least Squares` to obtain the required regression coefficients. To see how well the fitted function matches the data, select `Regression>Show fit`.

Example 1.3.6 Accidental deaths

To fit a sum of two harmonics with periods twelve months and six months to the monthly accidental deaths data $x_1, \ldots, x_n$ with $n = 72$, we choose $k = 2$, $f_1 =$

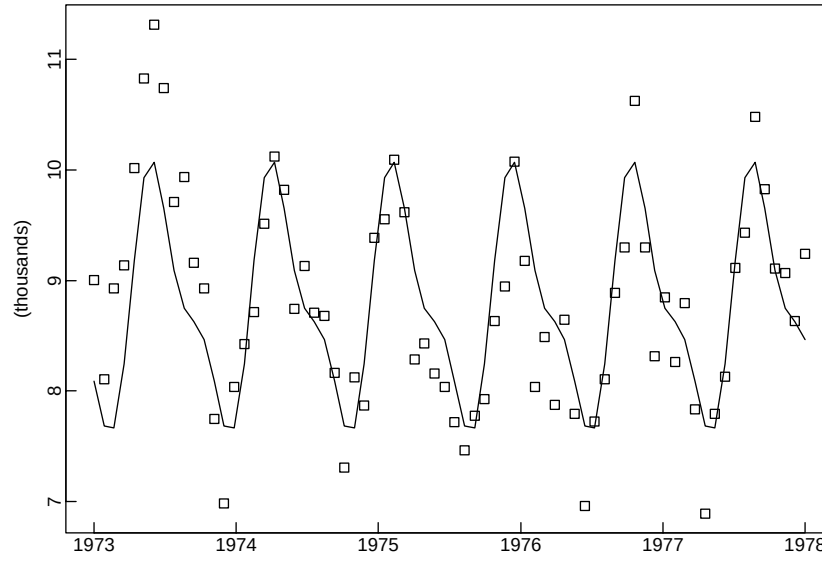

Figure 1-11
The estimated harmonic component of the accidental deaths data from ITSM.

$n/12 = 6$, and $f_2 = n/6 = 12$. Using ITSM as described above, we obtain the fitted function shown in Figure 1.11. As can be seen from the figure, the periodic character of the series is captured reasonably well by this fitted function. In practice, it is worth experimenting with several different combinations of harmonics in order to find a satisfactory estimate of the seasonal component. The program ITSM also allows fitting a linear combination of harmonics and polynomial trend by checking both `Harmonic Regression` and `Polynomial Regression` in the `Regression>Specification` dialog box. Other methods for dealing with seasonal variation in the presence of trend are described in Section 1.5. □

1.3.3 A General Approach to Time Series Modeling

The examples of the previous section illustrate a general approach to time series analysis that will form the basis for much of what is done in this book. Before introducing the ideas of dependence and stationarity, we outline this approach to provide the reader with an overview of the way in which the various ideas of this chapter fit together.

- Plot the series and examine the main features of the graph, checking in particular whether there is
  (a) a trend,
  (b) a seasonal component,
  (c) any apparent sharp changes in behavior,
  (d) any outlying observations.
- Remove the trend and seasonal components to get *stationary* residuals (as defined in Section 1.4). To achieve this goal it may sometimes be necessary to apply a preliminary transformation to the data. For example, if the magnitude of the fluctuations appears to grow roughly linearly with the level of the series, then the transformed series $\{\ln X_1, \ldots, \ln X_n\}$ will have fluctuations of more constant magnitude. See, for example, Figures 1.1 and 1.17. (If some of the data are negative, add a positive constant to each of the data values to ensure that all values are positive before taking logarithms.) There are several ways in which trend and seasonality can be removed (see Section 1.5), some involving estimating the components and subtracting them from the data, and others depending on *differencing* the data, i.e., replacing the original series $\{X_t\}$ by $\{Y_t := X_t - X_{t-d}\}$ for some positive integer d. Whichever method is used, the aim is to produce a stationary series, whose values we shall refer to as residuals.
- Choose a model to fit the residuals, making use of various sample statistics including the sample autocorrelation function to be defined in Section 1.4.
- Forecasting will be achieved by forecasting the residuals and then inverting the transformations described above to arrive at forecasts of the original series $\{X_t\}$.

- An extremely useful alternative approach touched on only briefly in this book is to express the series in terms of its Fourier components, which are sinusoidal waves of different frequencies (cf. Example 1.1.4). This approach is especially important in engineering applications such as signal processing and structural design. It is important, for example, to ensure that the resonant frequency of a structure does not coincide with a frequency at which the loading forces on the structure have a particularly large component.

1.4 Stationary Models and the Autocorrelation Function

Loosely speaking, a time series $\{X_t, t = 0, \pm 1, \ldots\}$ is said to be stationary if it has statistical properties similar to those of the "time-shifted" series $\{X_{t+h}, t = 0, \pm 1, \ldots\}$, for each integer h. Restricting attention to those properties that depend only on the first- and second-order moments of $\{X_t\}$, we can make this idea precise with the following definitions.

Definition 1.4.1

Let $\{X_t\}$ be a time series with $E(X_t^2) < \infty$. The **mean function** of $\{X_t\}$ is

$$\mu_X(t) = E(X_t).$$

The **covariance function** of $\{X_t\}$ is

$$\gamma_X(r, s) = \text{Cov}(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))]$$

for all integers r and s.

Definition 1.4.2

$\{X_t\}$ is **(weakly) stationary** if

(i) $\mu_X(t)$ is independent of t,

and

(ii) $\gamma_X(t + h, t)$ is independent of t for each h.

Remark 1. Strict stationarity of a time series $\{X_t, t = 0, \pm 1, \ldots\}$ is defined by the condition that $(X_1, \ldots, X_n)$ and $(X_{1+h}, \ldots, X_{n+h})$ have the same joint distributions for all integers h and $n > 0$. It is easy to check that if $\{X_t\}$ is strictly stationary and $EX_t^2 < \infty$ for all t, then $\{X_t\}$ is also weakly stationary (Problem 1.3). Whenever we use the term *stationary* we shall mean weakly stationary as in Definition 1.4.2, unless we specifically indicate otherwise. □

Remark 2. In view of condition (ii), whenever we use the term covariance function with reference to a *stationary* time series $\{X_t\}$ we shall mean the function γ_X of *one*

variable, defined by

$$\gamma_X(h) := \gamma_X(h, 0) = \gamma_X(t + h, t).$$

The function $\gamma_X(\cdot)$ will be referred to as the autocovariance function and $\gamma_X(h)$ as its value at *lag* h. □

Definition 1.4.3

Let $\{X_t\}$ be a stationary time series. The **autocovariance function** (ACVF) of $\{X_t\}$ at lag h is

$$\gamma_X(h) = \text{Cov}(X_{t+h}, X_t).$$

The **autocorrelation function** (ACF) of $\{X_t\}$ at lag h is

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = \text{Cor}(X_{t+h}, X_t).$$

In the following examples we shall frequently use the easily verified **linearity property of covariances**, that if $\text{E}X^2 < \infty$, $\text{E}Y^2 < \infty$, $\text{E}Z^2 < \infty$ and a, b, and c are any real constants, then

$$\text{Cov}(aX + bY + c, Z) = a\,\text{Cov}(X, Z) + b\,\text{Cov}(Y, Z).$$

Example 1.4.1 iid noise

If $\{X_t\}$ is iid noise and $E(X_t^2) = \sigma^2 < \infty$, then the first requirement of Definition 1.4.2 is obviously satisfied, since $E(X_t) = 0$ for all t. By the assumed independence,

$$\gamma_X(t + h, t) = \begin{cases} \sigma^2, & \text{if } h = 0, \\ 0, & \text{if } h \neq 0, \end{cases}$$

which does not depend on t. Hence iid noise with finite second moment is stationary. We shall use the notation

$$\{X_t\} \sim \text{IID}\left(0, \sigma^2\right)$$

to indicate that the random variables X_t are independent and identically distributed random variables, each with mean 0 and variance σ^2. □

Example 1.4.2 White noise

If $\{X_t\}$ is a sequence of uncorrelated random variables, each with zero mean and variance σ^2, then clearly $\{X_t\}$ is stationary with the same covariance function as the iid noise in Example 1.4.1. Such a sequence is referred to as **white noise** (with mean 0 and variance σ^2). This is indicated by the notation

$$\{X_t\} \sim \text{WN}\left(0, \sigma^2\right).$$

Clearly, every IID$(0, \sigma^2)$ sequence is WN$(0, \sigma^2)$ but not conversely (see Problem 1.8 and the ARCH(1) process of Section 10.3). □

Example 1.4.3 The random walk

If $\{S_t\}$ is the random walk defined in Example 1.3.3 with $\{X_t\}$ as in Example 1.4.1, then $ES_t = 0$, $E(S_t^2) = t\sigma^2 < \infty$ for all t, and, for $h \geq 0$,

$$\begin{aligned}\gamma_S(t+h, t) &= \text{Cov}(S_{t+h}, S_t)\\ &= \text{Cov}(S_t + X_{t+1} + \cdots + X_{t+h}, S_t)\\ &= \text{Cov}(S_t, S_t)\\ &= t\sigma^2.\end{aligned}$$

Since $\gamma_S(t+h, t)$ depends on t, the series $\{S_t\}$ is *not* stationary. □

Example 1.4.4 First-order moving average or MA(1) process

Consider the series defined by the equation

$$X_t = Z_t + \theta Z_{t-1}, \quad t = 0, \pm 1, \ldots, \tag{1.4.1}$$

where $\{Z_t\} \sim \text{WN}\left(0, \sigma^2\right)$ and θ is a real-valued constant. From (1.4.1) we see that $EX_t = 0$, $EX_t^2 = \sigma^2(1+\theta^2) < \infty$, and

$$\gamma_X(t+h, t) = \begin{cases} \sigma^2\left(1+\theta^2\right), & \text{if } h = 0,\\ \sigma^2\theta, & \text{if } h = \pm 1,\\ 0, & \text{if } |h| > 1.\end{cases}$$

Thus the requirements of Definition 1.4.2 are satisfied, and $\{X_t\}$ is stationary. The autocorrelation function of $\{X_t\}$ is

$$\rho_X(h) = \begin{cases} 1, & \text{if } h = 0,\\ \theta/\left(1+\theta^2\right), & \text{if } h = \pm 1,\\ 0, & \text{if } |h| > 1.\end{cases}$$

□

Example 1.4.5 First-order autoregression or AR(1) process

Let us *assume* now that $\{X_t\}$ is a stationary series satisfying the equations

$$X_t = \phi X_{t-1} + Z_t, \quad t = 0, \pm 1, \ldots, \tag{1.4.2}$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, $|\phi| < 1$, and Z_t is uncorrelated with X_s for each $s < t$. (We shall show in Section 2.2 that there is in fact exactly one such solution of (1.4.2).) By taking expectations on each side of (1.4.2) and using the fact that $EZ_t = 0$, we see

at once that

$$EX_t = 0.$$

To find the autocorrelation function of $\{X_t\}$ we multiply each side of (1.4.2) by X_{t-h} ($h > 0$) and then take expectations to get

$$\begin{aligned}\gamma_X(h) &= \text{Cov}(X_t, X_{t-h}) \\ &= \text{Cov}(\phi X_{t-1}, X_{t-h}) + \text{Cov}(Z_t, X_{t-h}) \\ &= \phi\gamma_X(h-1) + 0 = \cdots = \phi^h\gamma_x(0).\end{aligned}$$

Observing that $\gamma(h) = \gamma(-h)$ and using Definition 1.4.3, we find that

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = \phi^{|h|}, \quad h = 0, \pm 1, \ldots.$$

It follows from the linearity of the covariance function in each of its arguments and the fact that Z_t is uncorrelated with X_{t-1} that

$$\gamma_X(0) = \text{Cov}(X_t, X_t) = \text{Cov}(\phi X_{t-1} + Z_t, \phi X_{t-1} + Z_t) = \phi^2\gamma_X(0) + \sigma^2$$

and hence that $\gamma_X(0) = \sigma^2/\left(1-\phi^2\right)$. □

1.4.1 The Sample Autocorrelation Function

Although we have just seen how to compute the autocorrelation function for a few simple time series models, in practical problems we do not start with a model, but with *observed data* $\{x_1, x_2, \ldots, x_n\}$. To assess the degree of dependence in the data and to select a model for the data that reflects this, one of the important tools we use is the **sample autocorrelation function** (sample ACF) of the data. If we believe that the data are realized values of a stationary time series $\{X_t\}$, then the sample ACF will provide us with an estimate of the ACF of $\{X_t\}$. This estimate may suggest which of the many possible stationary time series models is a suitable candidate for representing the dependence in the data. For example, a sample ACF that is close to zero for all nonzero lags suggests that an appropriate model for the data might be iid noise. The following definitions are natural *sample* analogues of those for the autocovariance and autocorrelation functions given earlier for stationary time series *models*.

Definition 1.4.4

Let $x_1, \ldots, x_n$ be observations of a time series. The **sample mean** of $x_1, \ldots, x_n$ is

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t.$$

The **sample autocovariance function** is

$$\hat{\gamma}(h) := n^{-1}\sum_{t=1}^{n-|h|}(x_{t+|h|} - \bar{x})(x_t - \bar{x}), \quad -n < h < n.$$

The **sample autocorrelation function** is

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad -n < h < n.$$

Remark 3. For $h \geq 0$, $\hat{\gamma}(h)$ is approximately equal to the sample covariance of the $n - h$ pairs of observations $(x_1, x_{1+h}), (x_2, x_{2+h}), \ldots, (x_{n-h}, x_n)$. The difference arises from use of the divisor n instead of $n - h$ and the subtraction of the *overall* mean, $\bar{x}$, from each factor of the summands. Use of the divisor n ensures that the sample covariance matrix $\hat{\Gamma}_n := [\hat{\gamma}(i-j)]_{i,j=1}^{n}$ is nonnegative definite (see Section 2.4.2). □

Remark 4. Like the sample covariance matrix defined in Remark 3, the sample correlation matrix $\hat{R}_n := [\hat{\rho}(i-j)]_{i,j=1}^{n}$ is nonnegative definite. Each of its diagonal elements is equal to 1, since $\hat{\rho}(0) = 1$. □

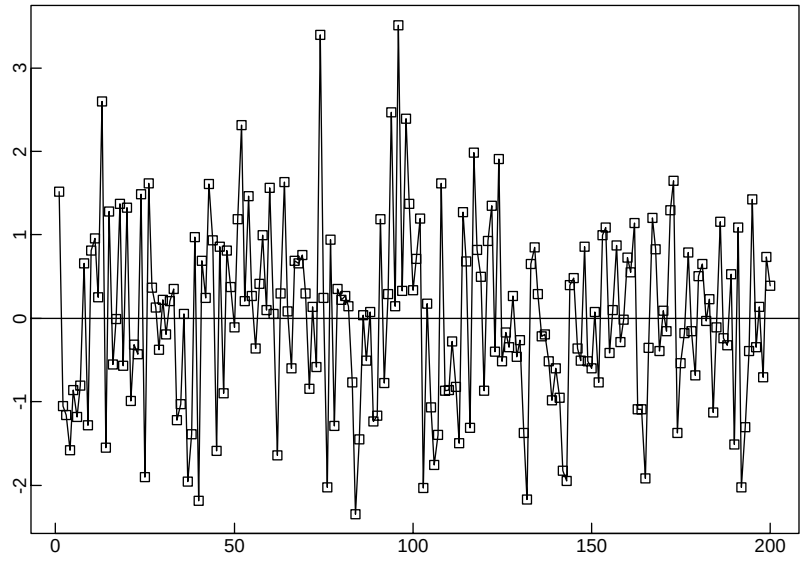

Figure 1-12
200 simulated values of iid N(0,1) noise.

Example 1.4.6 Figure 1.12 shows 200 simulated values of normally distributed iid $(0, 1)$, denoted by IID N$(0, 1)$, noise. Figure 1.13 shows the corresponding sample autocorrelation function at lags $0, 1, \ldots, 40$. Since $\rho(h) = 0$ for $h > 0$, one would also expect the corresponding sample autocorrelations to be near 0. It can be shown, in fact, that for iid noise with finite variance, the sample autocorrelations $\hat{\rho}(h), h > 0$, are approximately IID N$(0, 1/n)$ for n large (see TSTM p. 222). Hence, approximately 95% of the sample autocorrelations should fall between the bounds $\pm 1.96/\sqrt{n}$ (since 1.96 is the .975 quantile of the standard normal distribution). Therefore, in Figure 1.13 we would expect roughly $40(.05) = 2$ values to fall outside the bounds. To simulate 200 values of IID N$(0, 1)$ noise using ITSM, select `File>Project>New>Univariate` then `Model>Simulate`. In the resulting dialog box, enter 200 for the required `Number of Observations`. (The remaining entries in the dialog box can be left as they are, since the model assumed by ITSM, until you enter another, is IID N$(0, 1)$ noise. If you wish to reproduce exactly the same sequence at a later date, record the `Random Number Seed` for later use. By specifying different values for the random number seed you can generate independent realizations of your time series.) Click on `OK` and you will see the graph of your simulated series. To see its sample autocorrelation function together with the autocorrelation function of the model that generated it, click on the third yellow button at the top of the screen and you will see the two graphs superimposed (with the latter in red.) The horizontal lines on the graph are the bounds $\pm 1.96/\sqrt{n}$. □

Remark 5. The sample autocovariance and autocorrelation functions can be computed for *any* data set $\{x_1, \ldots, x_n\}$ and are not restricted to observations from a

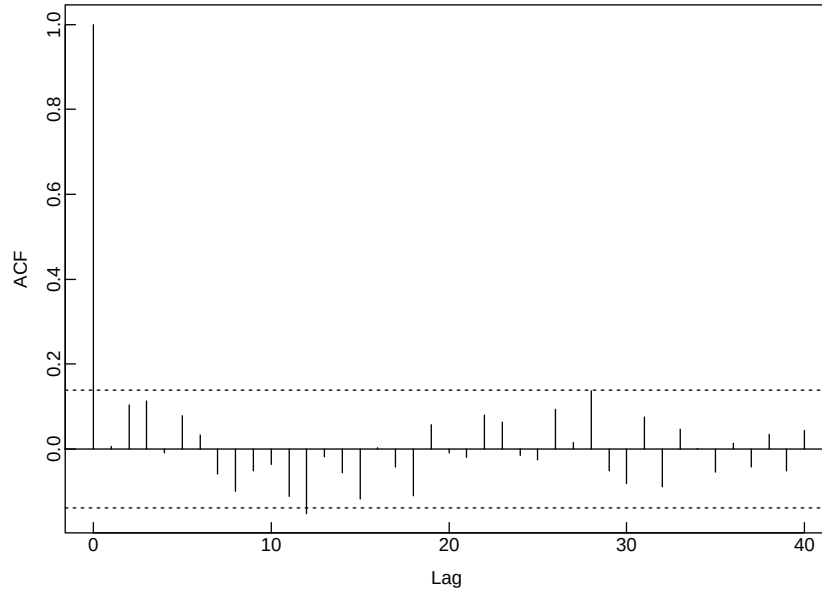

Figure 1-13
The sample autocorrelation function for the data of Figure 1.12 showing the bounds $\pm 1.96/\sqrt{n}$.

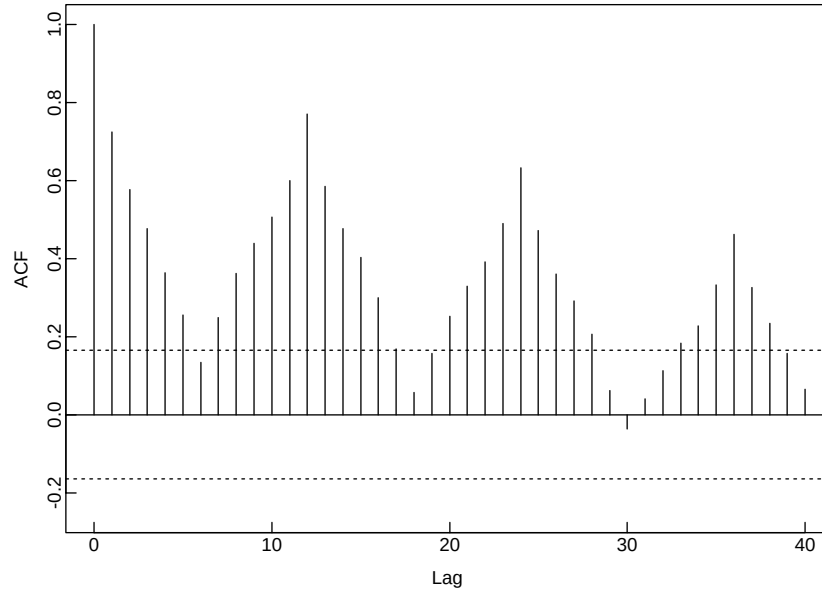

Figure 1-14
The sample autocorrelation function for the Australian red wine sales showing the bounds $\pm 1.96/\sqrt{n}$.

stationary time series. For data containing a trend, $|\hat{\rho}(h)|$ will exhibit slow decay as h increases, and for data with a substantial deterministic periodic component, $|\hat{\rho}(h)|$ will exhibit similar behavior with the same periodicity. (See the sample ACF of the Australian red wine sales in Figure 1.14 and Problem 1.9.) Thus $\hat{\rho}(\cdot)$ can be useful as an indicator of nonstationarity (see also Section 6.1). □

1.4.2 A Model for the Lake Huron Data

As noted earlier, an iid noise model for the residuals $\{y_1, \ldots, y_{98}\}$ obtained by fitting a straight line to the Lake Huron data in Example 1.3.5 appears to be inappropriate. This conclusion is confirmed by the sample ACF of the residuals (Figure 1.15), which has three of the first forty values well outside the bounds $\pm 1.96/\sqrt{98}$.

The roughly geometric decay of the first few sample autocorrelations (with $\hat{\rho}(h+1)/\hat{\rho}(h) \approx 0.7$) suggests that an AR(1) series (with $\phi \approx 0.7$) might provide a reasonable model for these residuals. (The form of the ACF for an AR(1) process was computed in Example 1.4.5.)

To explore the appropriateness of such a model, consider the points (y_1, y_2), $(y_2, y_3), \ldots, (y_{97}, y_{98})$ plotted in Figure 1.16. The graph does indeed suggest a linear relationship between y_t and y_{t-1}. Using simple least squares estimation to fit a straight line of the form $y_t = a y_{t-1}$, we obtain the model

$$Y_t = .791 Y_{t-1} + Z_t, \tag{1.4.3}$$

where $\{Z_t\}$ is iid noise with variance $\sum_{t=2}^{98}(y_t - .791 y_{t-1})^2/97 = .5024$. The sample ACF of the estimated noise sequence $z_t = y_t - .791 y_{t-1}$, $t = 2, \ldots, 98$, is slightly

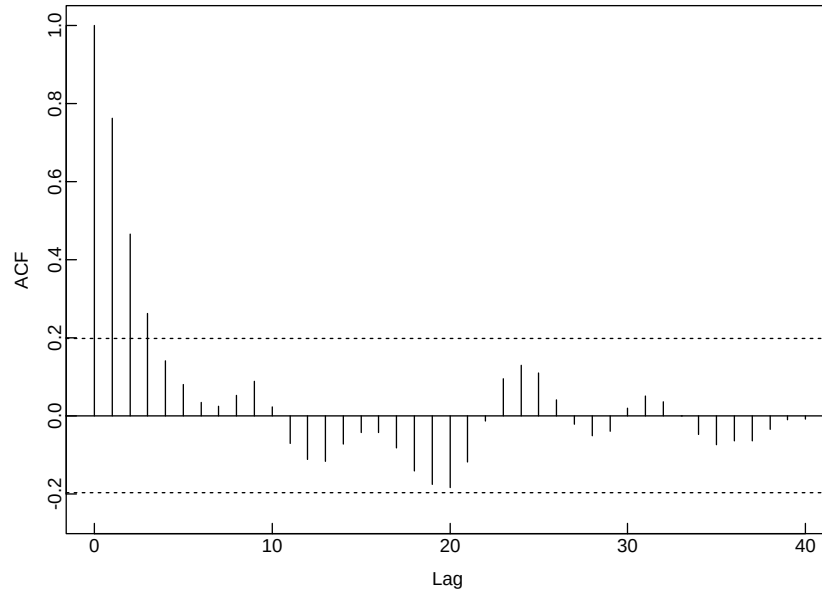

Figure 1-15
The sample autocorrelation function for the Lake Huron residuals of Figure 1.10 showing the bounds $\pm 1.96/\sqrt{n}$.

outside the bounds $\pm 1.96/\sqrt{97}$ at lag 1 ($\hat{\rho}(1) = .216$), but it is inside the bounds for all other lags up to 40. This check that the estimated noise sequence is consistent with the iid assumption of (1.4.3) reinforces our belief in the fitted model. More *goodness of fit* tests for iid noise sequences are described in Section 1.6. The estimated noise sequence $\{z_t\}$ in this example passes them all, providing further support for the model (1.4.3).

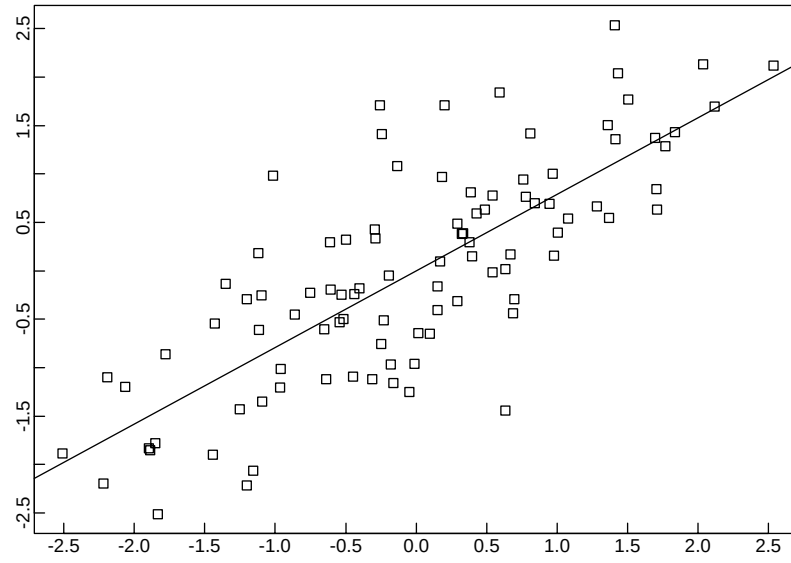

Figure 1-16
Scatter plot of (y_{t-1}, y_t), $t = 2, \ldots, 98$, for the data in Figure 1.10 showing the least squares regression line $y = .791x$.

A better fit to the residuals in equation (1.3.2) is provided by the second-order autoregression

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + Z_t, \tag{1.4.4}$$

where $\{Z_t\}$ is iid noise with variance σ^2. This is analogous to a linear model in which Y_t is regressed on the previous *two* values Y_{t-1} and Y_{t-2} of the time series. The least squares estimates of the parameters ϕ_1 and ϕ_2, found by minimizing $\sum_{t=3}^{98}(y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2})^2$, are $\hat{\phi}_1 = 1.002$ and $\hat{\phi}_2 = -.2834$. The estimate of σ^2 is $\hat{\sigma}^2 = \sum_{t=3}^{98}(y_t - \hat{\phi}_1 y_{t-1} - \hat{\phi}_2 y_{t-2})^2/96 = .4460$, which is approximately 11% smaller than the estimate of the noise variance for the AR(1) model (1.4.3). The improved fit is indicated by the sample ACF of the estimated residuals, $y_t - \hat{\phi}_1 y_{t-1} - \hat{\phi}_2 y_{t-2}$, which falls well within the bounds $\pm 1.96/\sqrt{96}$ for *all* lags up to 40.

1.5 Estimation and Elimination of Trend and Seasonal Components

The first step in the analysis of any time series is to plot the data. If there are any apparent discontinuities in the series, such as a sudden change of level, it may be advisable to analyze the series by first breaking it into homogeneous segments. If there are outlying observations, they should be studied carefully to check whether there is any justification for discarding them (as for example if an observation has been incorrectly recorded). Inspection of a graph may also suggest the possibility of representing the data as a realization of the process (the **classical decomposition** model)

$$X_t = m_t + s_t + Y_t, \tag{1.5.1}$$

where m_t is a slowly changing function known as a **trend component**, s_t is a function with known period d referred to as a **seasonal component**, and Y_t is a **random noise component** that is stationary in the sense of Definition 1.4.2. If the seasonal and noise fluctuations appear to increase with the level of the process, then a preliminary transformation of the data is often used to make the transformed data more compatible with the model (1.5.1). Compare, for example, the red wine sales in Figure 1.1 with the transformed data, Figure 1.17, obtained by applying a logarithmic transformation. The transformed data do not exhibit the increasing fluctuation with increasing level that was apparent in the original data. This suggests that the model (1.5.1) is more appropriate for the transformed than for the original series. In this section we shall assume that the model (1.5.1) is appropriate (possibly after a preliminary transformation of the data) and examine some techniques for estimating the components m_t, s_t, and Y_t in the model.

Our aim is to estimate and extract the deterministic components m_t and s_t in the hope that the residual or noise component Y_t will turn out to be a stationary time series. We can then use the theory of such processes to find a satisfactory probabilistic

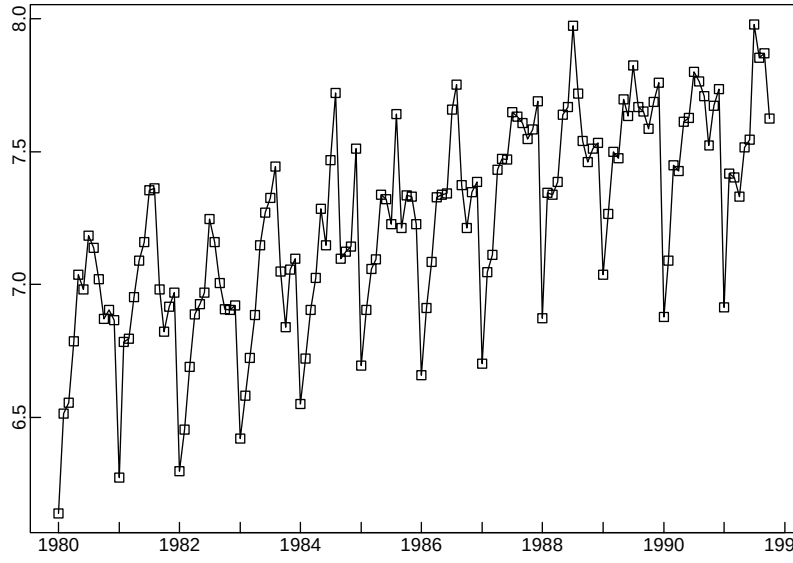

Figure 1-17
The natural logarithms of the red wine data.

model for the process Y_t, to analyze its properties, and to use it in conjunction with m_t and s_t for purposes of prediction and simulation of $\{X_t\}$.

Another approach, developed extensively by Box and Jenkins (1976), is to apply differencing operators repeatedly to the series $\{X_t\}$ until the differenced observations resemble a realization of some stationary time series $\{W_t\}$. We can then use the theory of stationary processes for the modeling, analysis, and prediction of $\{W_t\}$ and hence of the original process. The various stages of this procedure will be discussed in detail in Chapters 5 and 6.

The two approaches to trend and seasonality removal, (1) by estimation of m_t and s_t in (1.5.1) and (2) by differencing the series $\{X_t\}$, will now be illustrated with reference to the data introduced in Section 1.1.

1.5.1 Estimation and Elimination of Trend in the Absence of Seasonality

In the absence of a seasonal component the model (1.5.1) becomes the following.

Nonseasonal Model with Trend:

$$X_t = m_t + Y_t, \quad t = 1, \ldots, n, \tag{1.5.2}$$

where $EY_t = 0$.

(If $EY_t \neq 0$, then we can replace m_t and Y_t in (1.5.2) with $m_t + EY_t$ and $Y_t - EY_t$, respectively.)

Method 1: Trend Estimation

Moving average and spectral smoothing are essentially nonparametric methods for trend (or signal) estimation and not for model building. Special smoothing filters can also be designed to remove periodic components as described under Method S1 below. The choice of smoothing filter requires a certain amount of subjective judgment, and it is recommended that a variety of filters be tried in order to get a good idea of the underlying trend. Exponential smoothing, since it is based on a moving average of *past* values only, is often used for forecasting, the smoothed value at the present time being used as the forecast of the next value.

To construct a *model* for the data (with no seasonality) there are two general approaches, both available in ITSM. One is to fit a polynomial trend (by least squares) as described in Method 1(d) below, then to subtract the fitted trend from the data and to find an appropriate stationary time series model for the residuals. The other is to eliminate the trend by differencing as described in Method 2 and then to find an appropriate stationary model for the differenced series. The latter method has the advantage that it usually requires the estimation of fewer parameters and does not rest on the assumption of a trend that remains fixed throughout the observation period. The study of the residuals (or of the differenced series) is taken up in Section 1.6.

(a) *Smoothing with a finite moving average filter.* Let q be a nonnegative integer and consider the two-sided moving average

$$W_t = (2q+1)^{-1} \sum_{j=-q}^{q} X_{t-j} \tag{1.5.3}$$

of the process $\{X_t\}$ defined by (1.5.2). Then for $q+1 \le t \le n-q$,

$$W_t = (2q+1)^{-1} \sum_{j=-q}^{q} m_{t-j} + (2q+1)^{-1} \sum_{j=-q}^{q} Y_{t-j} \approx m_t, \tag{1.5.4}$$

assuming that m_t is approximately linear over the interval $[t-q, t+q]$ and that the average of the error terms over this interval is close to zero (see Problem 1.11).

The moving average thus provides us with the estimates

$$\hat{m}_t = (2q+1)^{-1} \sum_{j=-q}^{q} X_{t-j}, \quad q+1 \le t \le n-q. \tag{1.5.5}$$

Since X_t is not observed for $t \le 0$ or $t > n$, we cannot use (1.5.5) for $t \le q$ or $t > n-q$. The program ITSM deals with this problem by defining $X_t := X_1$ for $t < 1$ and $X_t := X_n$ for $t > n$.

Example 1.5.1 The result of applying the moving-average filter (1.5.5) with $q = 2$ to the strike data of Figure 1.6 is shown in Figure 1.18. The estimated noise terms $\hat{Y}_t = X_t - \hat{m}_t$ are shown in Figure 1.19. As expected, they show no apparent trend. To apply this filter using ITSM, open the project STRIKES.TSM, select `Smooth>Moving Average`, specify

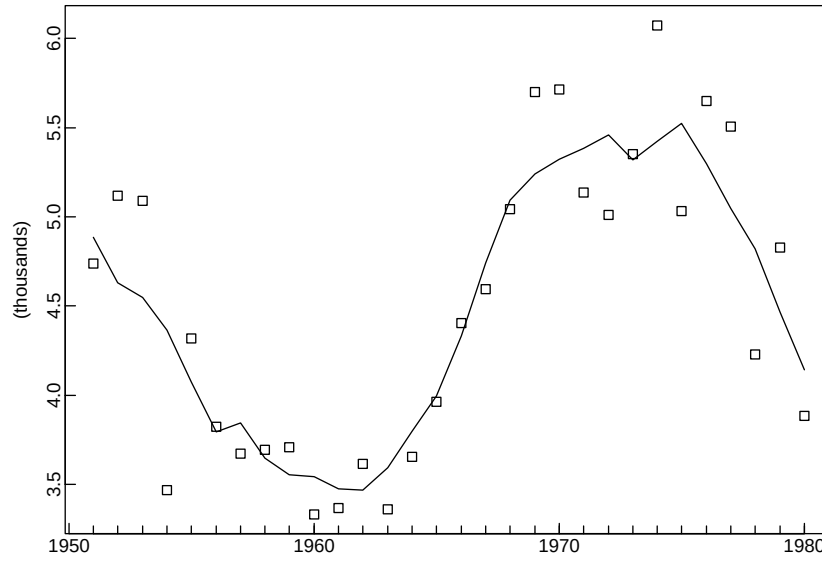

Figure 1-18
Simple 5-term moving average $\hat{m}_t$ of the strike data from Figure 1.6.

2 for the filter order, and enter the weights 1,1,1 for Theta(0), Theta(1), and Theta(2) (these are automatically normalized so that the sum of the weights is one). Then click OK. □

It is useful to think of $\{\hat{m}_t\}$ in (1.5.5) as a process obtained from $\{X_t\}$ by application of a linear operator or linear filter $\hat{m}_t = \sum_{j=-\infty}^{\infty} a_j X_{t-j}$ with weights $a_j = (2q + 1)^{-1}$, $-q \le j \le q$. This particular filter is a **low-pass** filter in the sense that it takes the

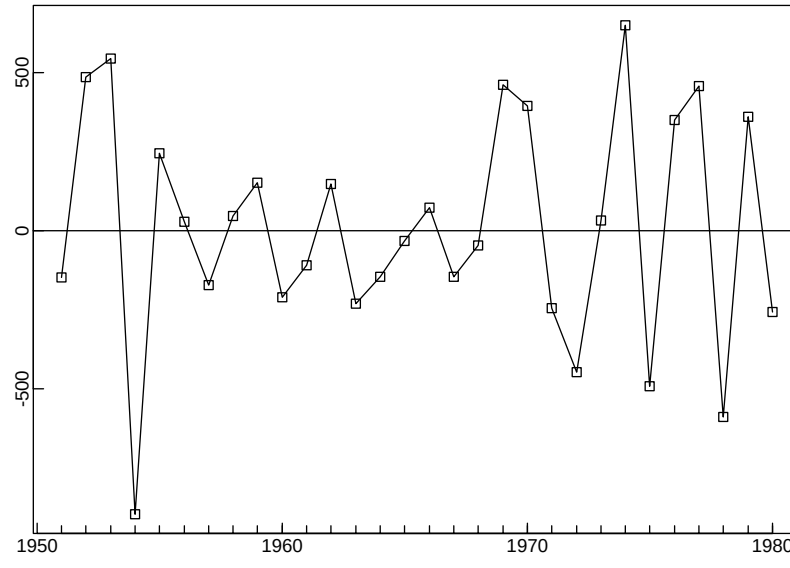

Figure 1-19
Residuals $\hat{Y}_t = X_t - \hat{m}_t$ after subtracting the 5-term moving average from the strike data

Figure 1-20
Smoothing with a low-pass linear filter.

data $\{X_t\}$ and removes from it the rapidly fluctuating (or high frequency) component $\{\hat{Y}_t\}$ to leave the slowly varying estimated trend term $\{\hat{m}_t\}$ (see Figure 1.20).

The particular filter (1.5.5) is only one of many that could be used for smoothing. For large q, provided that $(2q+1)^{-1}\sum_{j=-q}^{q} Y_{t-j} \approx 0$, it not only will attenuate noise but at the same time will allow linear trend functions $m_t = c_0 + c_1 t$ to pass without distortion (see Problem 1.11). However, we must beware of choosing q to be too large, since if m_t is not linear, the filtered process, although smooth, will not be a good estimate of m_t. By clever choice of the weights $\{a_j\}$ it is possible (see Problems 1.12–1.14 and Section 4.3) to design a filter that will not only be effective in attenuating noise in the data, but that will also allow a larger class of trend functions (for example all polynomials of degree less than or equal to 3) to pass through without distortion. The Spencer 15-point moving average is a filter that passes polynomials of degree 3 without distortion. Its weights are

$$a_j = 0, \quad |j| > 7,$$

with

$$a_j = a_{-j}, \quad |j| \leq 7,$$

and

$$[a_0, a_1, \ldots, a_7] = \frac{1}{320}[74, 67, 46, 21, 3, -5, -6, -3]. \tag{1.5.6}$$

Applied to the process (1.5.2) with $m_t = c_0 + c_1 t + c_2 t^2 + c_3 t^3$, it gives

$$\sum_{j=-7}^{7} a_j X_{t-j} = \sum_{j=-7}^{7} a_j m_{t-j} + \sum_{j=-7}^{7} a_j Y_{t-j} \approx \sum_{j=-7}^{7} a_j m_{t-j} = m_t,$$

where the last step depends on the assumed form of m_t (Problem 1.12). Further details regarding this and other smoothing filters can be found in Kendall and Stuart (1976), Chapter 46.

(b) *Exponential smoothing.* For any fixed $\alpha \in [0, 1]$, the one-sided moving averages $\hat{m}_t$, $t = 1, \ldots, n$, defined by the recursions

$$\hat{m}_t = \alpha X_t + (1-\alpha)\hat{m}_{t-1}, \quad t = 2, \ldots, n, \tag{1.5.7}$$

and

$$\hat{m}_1 = X_1 \tag{1.5.8}$$

can be computed using ITSM by selecting `Smooth>Exponential` and specifying the value of α. Application of (1.5.7) and (1.5.8) is often referred to as exponential smoothing, since the recursions imply that for $t \geq 2$, $\hat{m}_t = \sum_{j=0}^{t-2} \alpha(1-\alpha)^j X_{t-j} + (1-\alpha)^{t-1} X_1$, a weighted moving average of $X_t, X_{t-1}, \ldots$, with weights decreasing exponentially (except for the last one).

(c) *Smoothing by elimination of high-frequency components.* The option `Smooth>FFT` in the program ITSM allows us to smooth an arbitrary series by elimination of the high-frequency components of its Fourier series expansion (see Section 4.2). This option was used in Example 1.1.4, where we chose to retain the fraction $f = .035$ of the frequency components of the series in order to estimate the underlying signal. (The choice $f = 1$ would have left the series unchanged.)

Example 1.5.2 In Figures 1.21 and 1.22 we show the results of smoothing the strike data by exponential smoothing with parameter $\alpha = 0.4$ (see (1.5.7)) and by high-frequency elimination with $f = 0.4$, i.e., by eliminating a fraction 0.6 of the Fourier components at the top of the frequency range. These should be compared with the simple 5-term moving average smoothing shown in Figure 1.18. Experimentation with different smoothing parameters can easily be carried out using the program ITSM. The exponentially smoothed value of the last observation is frequently used to forecast the next data value. The program automatically selects an optimal value of α for this purpose if α is specified as -1 in the exponential smoothing dialog box. □

(d) *Polynomial fitting.* In Section 1.3.2 we showed how a trend of the form $m_t = a_0 + a_1 t + a_2 t^2$ can be fitted to the data $\{x_1, \ldots, x_n\}$ by choosing the parameters

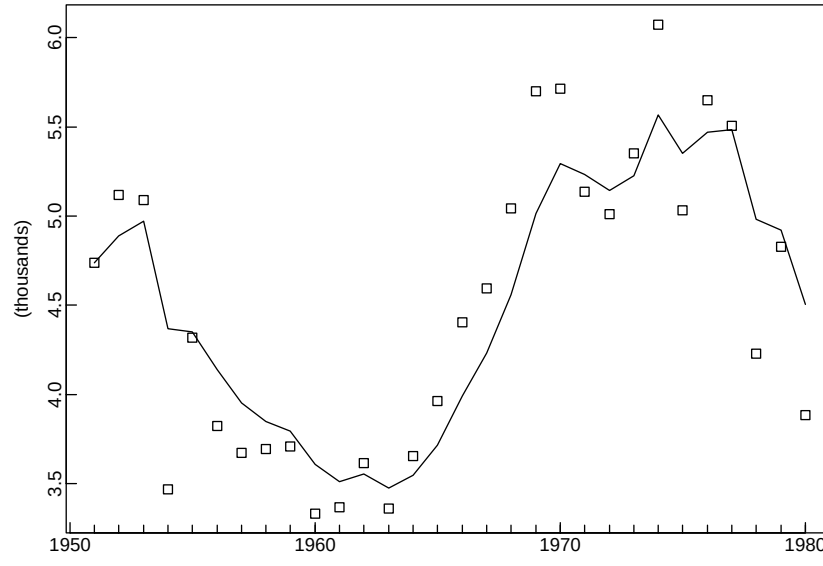

Figure 1-21
Exponentially smoothed strike data with $\alpha = 0.4$.

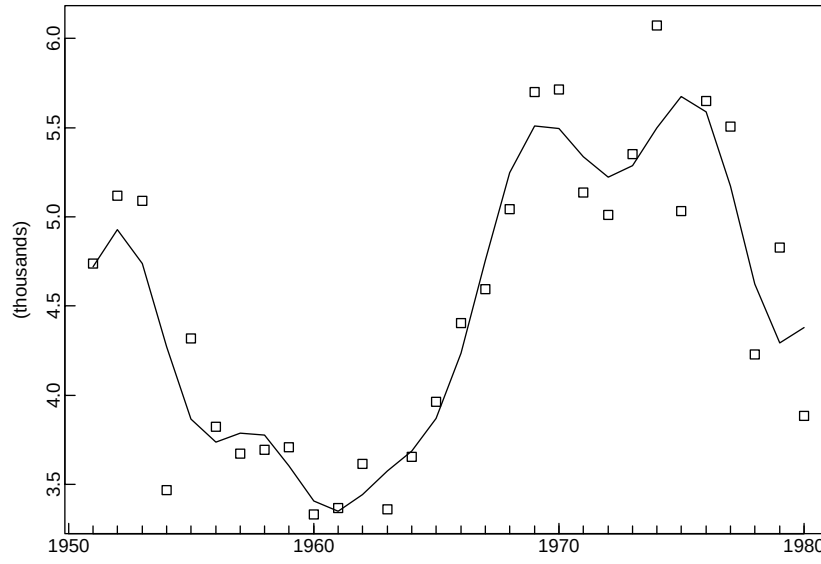

Figure 1-22
Strike data smoothed by elimination of high frequencies with $f = 0.4$.

a_0, a_1, and a_2 to minimize the sum of squares, $\sum_{t=1}^{n}(x_t - m_t)^2$ (see Example 1.3.4). The method of least squares estimation can also be used to estimate higher-order polynomial trends in the same way. The `Regression` option of ITSM allows least squares fitting of polynomial trends of order up to 10 (together with up to four harmonic terms; see Example 1.3.6). It also allows generalized least squares estimation (see Section 6.6), in which correlation between the residuals is taken into account.

Method 2: Trend Elimination by Differencing

Instead of attempting to remove the noise by smoothing as in Method 1, we now attempt to eliminate the trend term by differencing. We define the lag-1 difference operator ∇ by

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t, \tag{1.5.9}$$

where B is the backward shift operator,

$$BX_t = X_{t-1}. \tag{1.5.10}$$

Powers of the operators B and ∇ are defined in the obvious way, i.e., $B^j(X_t) = X_{t-j}$ and $\nabla^j(X_t) = \nabla(\nabla^{j-1}(X_t))$, $j \geq 1$, with $\nabla^0(X_t) = X_t$. Polynomials in B and ∇ are manipulated in precisely the same way as polynomial functions of real variables. For example,

$$\begin{aligned}\nabla^2 X_t &= \nabla(\nabla(X_t)) = (1 - B)(1 - B)X_t = (1 - 2B + B^2)X_t \\ &= X_t - 2X_{t-1} + X_{t-2}.\end{aligned}$$

If the operator ∇ is applied to a linear trend function $m_t = c_0 + c_1 t$, then we obtain the constant function $\nabla m_t = m_t - m_{t-1} = c_0 + c_1 t - (c_0 + c_1(t-1)) = c_1$. In the same way any polynomial trend of degree k can be reduced to a constant by application of the operator ∇^k (Problem 1.10). For example, if $X_t = m_t + Y_t$, where $m_t = \sum_{j=0}^{k} c_j t^j$ and Y_t is stationary with mean zero, application of ∇^k gives

$$\nabla^k X_t = k! c_k + \nabla^k Y_t,$$

a stationary process with mean $k!c_k$. These considerations suggest the possibility, given any sequence $\{x_t\}$ of data, of applying the operator ∇ repeatedly until we find a sequence $\{\nabla^k x_t\}$ that can plausibly be modeled as a realization of a stationary process. It is often found in practice that the order k of differencing required is quite small, frequently one or two. (This relies on the fact that many functions can be well approximated, on an interval of finite length, by a polynomial of reasonably low degree.)

Example 1.5.3 Applying the operator ∇ to the population values $\{x_t, t = 1, \ldots, 20\}$ of Figure 1.5, we find that two differencing operations are sufficient to produce a series with no apparent trend. (To carry out the differencing using ITSM, select `Transform>Difference`, enter the value 1 for the differencing lag, and click `OK`.) This replaces the original series $\{x_t\}$ by the once-differenced series $\{x_t - x_{t-1}\}$. Repetition of these steps gives the twice-differenced series $\nabla^2 x_t = x_t - 2x_{t-1} + x_{t-2}$, plotted in Figure 1.23. Notice that the magnitude of the fluctuations in $\nabla^2 x_t$ increases with the value of x_t. This effect can be suppressed by first taking natural logarithms, $y_t = \ln x_t$, and then applying the operator ∇^2 to the series $\{y_t\}$. (See also Figures 1.1 and 1.17.) □

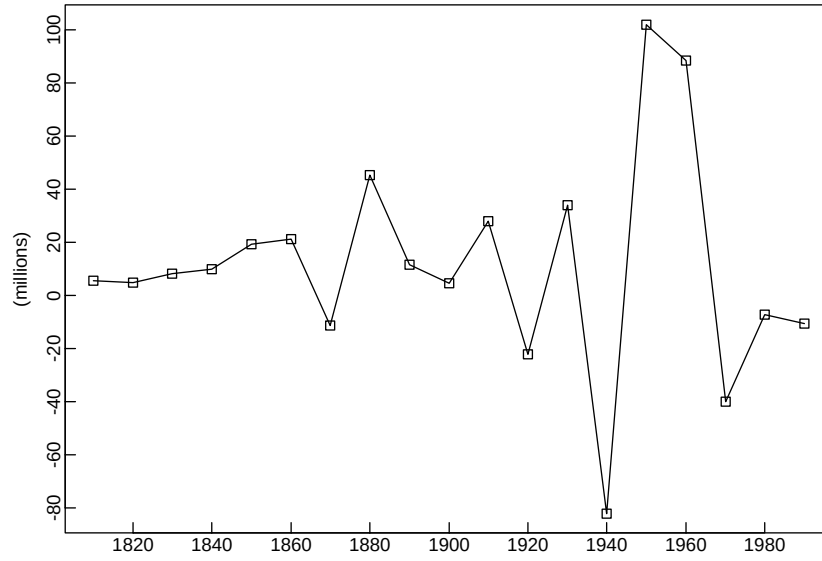

Figure 1-23
The twice-differenced series derived from the population data of Figure 1.5.

1.5.2 Estimation and Elimination of Both Trend and Seasonality

The methods described for the estimation and elimination of trend can be adapted in a natural way to eliminate both trend and seasonality in the general model, specified as follows.

Classical Decomposition Model

$$X_t = m_t + s_t + Y_t, \quad t = 1, \ldots, n, \tag{1.5.11}$$

where $EY_t = 0$, $s_{t+d} = s_t$, and $\sum_{j=1}^{d} s_j = 0$.

We shall illustrate these methods with reference to the accidental deaths data of Example 1.1.3, for which the period d of the seasonal component is clearly 12.

Method S1: Estimation of Trend and Seasonal Components

The method we are about to describe is used in the `Transform>Classical` option of ITSM.

Suppose we have observations $\{x_1, \ldots, x_n\}$. The trend is first estimated by applying a moving average filter specially chosen to eliminate the seasonal component and to dampen the noise. If the period d is even, say $d = 2q$, then we use

$$\hat{m}_t = (0.5x_{t-q} + x_{t-q+1} + \cdots + x_{t+q-1} + 0.5x_{t+q})/d, \quad q < t \le n - q. \tag{1.5.12}$$

If the period is odd, say $d = 2q + 1$, then we use the simple moving average (1.5.5).

The second step is to estimate the seasonal component. For each $k = 1, \ldots, d$, we compute the average w_k of the deviations $\{(x_{k+jd} - \hat{m}_{k+jd}), q < k + jd \le n - q\}$. Since these average deviations do not necessarily sum to zero, we estimate the seasonal component s_k as

$$\hat{s}_k = w_k - d^{-1} \sum_{i=1}^{d} w_i, \quad k = 1, \ldots, d, \tag{1.5.13}$$

and $\hat{s}_k = \hat{s}_{k-d}$, $k > d$.

The *deseasonalized* data is then defined to be the original series with the estimated seasonal component removed, i.e.,

$$d_t = x_t - \hat{s}_t, \quad t = 1, \ldots, n. \tag{1.5.14}$$

Finally, we reestimate the trend from the deseasonalized data $\{d_t\}$ using one of the methods already described. The program ITSM allows you to fit a least squares polynomial trend $\hat{m}$ to the deseasonalized series. In terms of this reestimated trend and the estimated seasonal component, the estimated noise series is then given by

$$\hat{Y}_t = x_t - \hat{m}_t - \hat{s}_t, \quad t = 1, \ldots, n.$$

The reestimation of the trend is done in order to have a parametric form for the trend that can be extrapolated for the purposes of prediction and simulation.

Example 1.5.4 Figure 1.24 shows the deseasonalized accidental deaths data obtained from ITSM by reading in the series DEATHS.TSM, selecting `Transform>Classical`, checking *only* the box marked `Seasonal Fit`, entering 12 for the period, and clicking `OK`. The estimated seasonal component $\hat{s}_t$, shown in Figure 1.25, is obtained by selecting `Transform>Show Classical Fit`. (Except for having a mean of zero, this estimate is very similar to the harmonic regression function with frequencies $2\pi/12$ and $2\pi/6$ displayed in Figure 1.11.) The graph of the deseasonalized data suggests the presence of an additional quadratic trend function. In order to fit such a trend to the deseasonalized data, select `Transform>Undo Classical` to retrieve the original data and then select `Transform>Classical` and check the boxes marked `Seasonal Fit` and `Polynomial Trend`, entering 12 for the period and selecting `Quadratic` for the trend. Then click OK and you will obtain the trend function

$$\hat{m}_t = 9952 - 71.82t + 0.8260t^2, \quad 1 \le t \le 72.$$

At this point the data stored in ITSM consists of the estimated noise

$$\hat{Y}_t = x_t - \hat{m}_t - \hat{s}_t, \quad t = 1, \ldots, 72,$$

obtained by subtracting the estimated seasonal and trend components from the original data. □

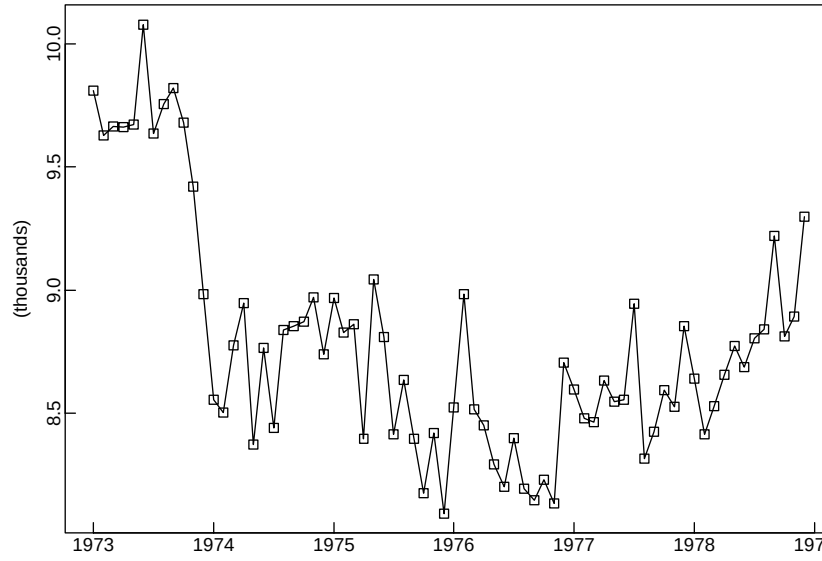

Figure 1-24
The deseasonalized accidental deaths data from ITSM.

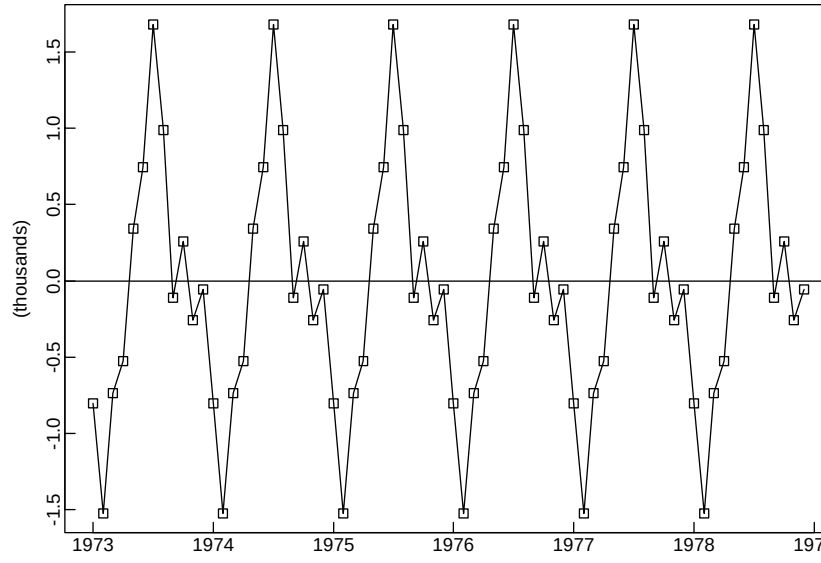

Figure 1-25
The estimated seasonal component of the accidental deaths data from ITSM.

Method S2: Elimination of Trend and Seasonal Components by Differencing
The technique of differencing that we applied earlier to nonseasonal data can be adapted to deal with seasonality of period d by introducing the lag-d differencing operator ∇_d defined by

$$\nabla_d X_t = X_t - X_{t-d} = (1 - B^d)X_t. \tag{1.5.15}$$

(This operator should not be confused with the operator $\nabla^d = (1 - B)^d$ defined earlier.)

Applying the operator ∇_d to the model

$$X_t = m_t + s_t + Y_t,$$

where $\{s_t\}$ has period d, we obtain

$$\nabla_d X_t = m_t - m_{t-d} + Y_t - Y_{t-d},$$

which gives a decomposition of the difference $\nabla_d X_t$ into a trend component $(m_t - m_{t-d})$ and a noise term $(Y_t - Y_{t-d})$. The trend, $m_t - m_{t-d}$, can then be eliminated using the methods already described, in particular by applying a power of the operator ∇.

Example 1.5.5 Figure 1.26 shows the result of applying the operator ∇_{12} to the accidental deaths data. The graph is obtained from ITSM by opening DEATHS.TSM, selecting `Transform>Difference`, entering lag 12, and clicking `OK`. The seasonal component evident in Figure 1.3 is absent from the graph of $\nabla_{12}x_t$, $13 \le t \le 72$. However, there still appears to be a nondecreasing trend. If we now apply the operator ∇ to $\{\nabla_{12}x_t\}$ by again selecting `Transform>Difference`, this time with lag one, we obtain the graph

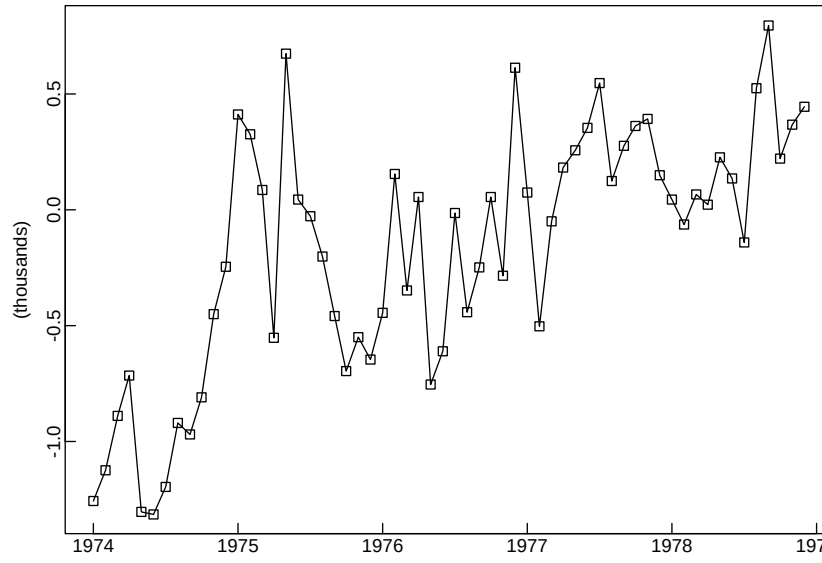

Figure 1-26
The differenced series $\{\nabla_{12}x_t, t = 13, \ldots, 72\}$ derived from the monthly accidental deaths $\{x_t, t = 1, \ldots, 72\}$.

of $\nabla\nabla_{12}x_t$, $14 \le t \le 72$, shown in Figure 1.27, which has no apparent trend or seasonal component. In Chapter 5 we shall show that this doubly differenced series can in fact be well represented by a stationary time series model. □

In this section we have discussed a variety of methods for estimating and/or removing trend and seasonality. The particular method chosen for any given data set will depend on a number of factors including whether or not estimates of the

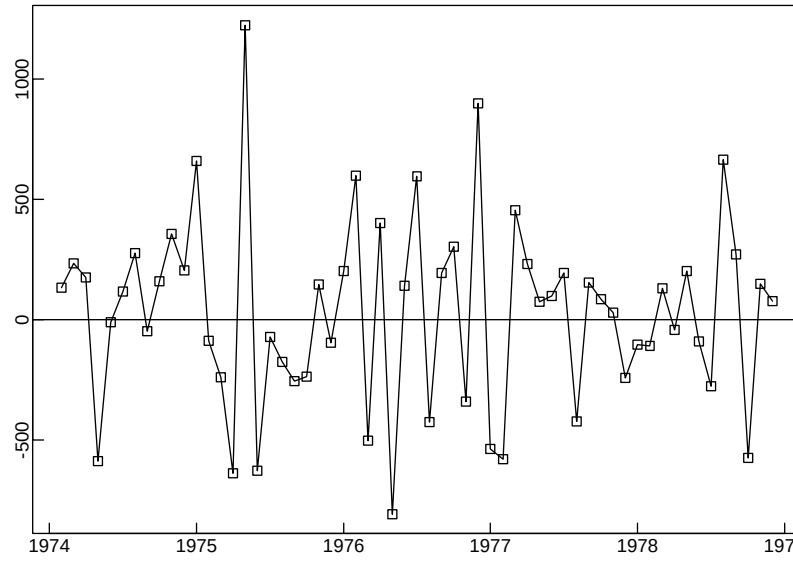

Figure 1-27
The differenced series $\{\nabla\nabla_{12}x_t, t = 14, \ldots, 72\}$ derived from the monthly accidental deaths $\{x_t, t = 1, \ldots, 72\}$.

components of the series are required and whether or not it appears that the data contain a seasonal component that does not vary with time. The program ITSM allows two options under the `Transform` menu:

1. "classical decomposition," in which trend and/or seasonal components are estimated and subtracted from the data to generate a noise sequence, and
2. "differencing," in which trend and/or seasonal components are removed from the data by repeated differencing at one or more lags in order to generate a noise sequence.

A third option is to use the `Regression` menu, possibly after applying a Box–Cox transformation. Using this option we can (see Example 1.3.6)

3. fit a sum of harmonics and a polynomial trend to generate a noise sequence that consists of the residuals from the regression.

In the next section we shall examine some techniques for deciding whether or not the noise sequence so generated differs significantly from iid noise. If the noise sequence *does* have sample autocorrelations significantly different from zero, then we can take advantage of this serial dependence to forecast future noise values in terms of past values by modeling the noise as a stationary time series.

1.6 Testing the Estimated Noise Sequence

The objective of the data transformations described in Section 1.5 is to produce a series with no apparent deviations from stationarity, and in particular with no apparent trend or seasonality. Assuming that this has been done, the next step is to model the estimated noise sequence (i.e., the **residuals** obtained either by differencing the data or by estimating and subtracting the trend and seasonal components). If there is no dependence among between these residuals, then we can regard them as observations of independent random variables, and there is no further modeling to be done except to estimate their mean and variance. However, if there is significant dependence among the residuals, then we need to look for a more complex stationary time series model for the noise that accounts for the dependence. This will be to our advantage, since dependence means in particular that past observations of the noise sequence can assist in predicting future values.

In this section we examine some simple tests for checking the hypothesis that the residuals from Section 1.5 are observed values of independent and identically distributed random variables. If they are, then our work is done. If not, then we must use the theory of stationary processes to be developed in later chapters to find a more appropriate model.

(a) *The sample autocorrelation function.* For large n, the sample autocorrelations of an iid sequence $Y_1, \ldots, Y_n$ with finite variance are approximately iid with distribution $\mathrm{N}(0, 1/n)$ (see TSTM p. 222). Hence, if $y_1, \ldots, y_n$ is a realization of such an iid sequence, about 95% of the sample autocorrelations should fall between the bounds $\pm 1.96/\sqrt{n}$. If we compute the sample autocorrelations up to lag 40 and find that more than two or three values fall outside the bounds, or that one value falls far outside the bounds, we therefore reject the iid hypothesis. The bounds $\pm 1.96/\sqrt{n}$ are automatically plotted when the sample autocorrelation function is computed by the program ITSM.

(b) *The portmanteau test.* Instead of checking to see whether each sample autocorrelation $\hat{\rho}(j)$ falls inside the bounds defined in (a) above, it is also possible to consider the single statistic

$$Q = n \sum_{j=1}^{h} \hat{\rho}^2(j).$$

If $Y_1, \ldots, Y_n$ is a finite-variance iid sequence, then by the same result used in (a), Q is approximately distributed as the sum of squares of the independent $\mathrm{N}(0, 1)$ random variables, $\sqrt{n}\hat{\rho}(j)$, $j = 1, \ldots, h$, i.e., as chi-squared with h degrees of freedom. A large value of Q suggests that the sample autocorrelations of the data are too large for the data to be a sample from an iid sequence. We therefore reject the iid hypothesis at level α if $Q > \chi^2_{1-\alpha}(h)$, where $\chi^2_{1-\alpha}(h)$ is the $1-\alpha$ quantile of the chi-squared distribution with h degrees of freedom. The program ITSM conducts a refinement of this test, formulated by Ljung and Box (1978), in which Q is replaced by

$$Q_{\mathrm{LB}} = n(n+2) \sum_{j=1}^{h} \hat{\rho}^2(j)/(n-j),$$

whose distribution is better approximated by the chi-squared distribution with h degrees of freedom.

Another portmanteau test, formulated by McLeod and Li (1983), can be used as a further test for the iid hypothesis, since if the data are iid, then the squared data are also iid. It is based on the same statistic used for the Ljung–Box test, except that the sample autocorrelations of the data are replaced by the sample autocorrelations of the *squared* data, $\hat{\rho}_{WW}(h)$, giving

$$Q_{\mathrm{ML}} = n(n+2) \sum_{k=1}^{h} \hat{\rho}^2_{WW}(k)/(n-k).$$

The hypothesis of iid data is then rejected at level α if the observed value of Q_{ML} is larger than the $1-\alpha$ quantile of the $\chi^2(h)$ distribution.

(c) *The turning point test.* If $y_1, \ldots, y_n$ is a sequence of observations, we say that there is a turning point at time i, $1 < i < n$, if $y_{i-1} < y_i$ and $y_i > y_{i+1}$ or if $y_{i-1} > y_i$ and $y_i < y_{i+1}$. If T is the number of turning points of an iid sequence of

length n, then, since the probability of a turning point at time i is $\frac{2}{3}$, the expected value of T is

$$\mu_T = E(T) = 2(n-2)/3.$$

It can also be shown for an iid sequence that the variance of T is

$$\sigma_T^2 = \mathrm{Var}(T) = (16n - 29)/90.$$

A large value of $T - \mu_T$ indicates that the series is fluctuating more rapidly than expected for an iid sequence. On the other hand, a value of $T - \mu_T$ much smaller than zero indicates a positive correlation between neighboring observations. For an iid sequence with n large, it can be shown that

$$T \text{ is approximately } \mathrm{N}\big(\mu_T, \sigma_T^2\big).$$

This means we can carry out a test of the iid hypothesis, rejecting it at level α if $|T - \mu_T|/\sigma_T > \Phi_{1-\alpha/2}$, where $\Phi_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. (A commonly used value of α is .05, for which the corresponding value of $\Phi_{1-\alpha/2}$ is 1.96.)

(d) *The difference-sign test.* For this test we count the number S of values of i such that $y_i > y_{i-1}$, $i = 2, \ldots, n$, or equivalently the number of times the differenced series $y_i - y_{i-1}$ is positive. For an iid sequence it is clear that

$$\mu_S = ES = \frac{1}{2}(n-1).$$

It can also be shown, under the same assumption, that

$$\sigma_S^2 = \mathrm{Var}(S) = (n+1)/12,$$

and that for large n,

$$S \text{ is approximately } \mathrm{N}\big(\mu_S, \sigma_S^2\big).$$

A large positive (or negative) value of $S - \mu_S$ indicates the presence of an increasing (or decreasing) trend in the data. We therefore reject the assumption of no trend in the data if $|S - \mu_S|/\sigma_S > \Phi_{1-\alpha/2}$.

The difference-sign test must be used with caution. A set of observations exhibiting a strong cyclic component will pass the difference-sign test for randomness, since roughly half of the observations will be points of increase.

(e) *The rank test.* The rank test is particularly useful for detecting a linear trend in the data. Define P to be the number of pairs (i, j) such that $y_j > y_i$ and $j > i$, $i = 1, \ldots, n-1$. There is a total of $\binom{n}{2} = \frac{1}{2}n(n-1)$ pairs (i, j) such that $j > i$. For an iid sequence $\{Y_1, \ldots, Y_n\}$, each event $\{Y_j > Y_i\}$ has probability $\frac{1}{2}$, and the mean of P is therefore

$$\mu_P = \frac{1}{4}n(n-1).$$

It can also be shown for an iid sequence that the variance of P is

$$\sigma_P^2 = n(n-1)(2n+5)/72$$

and that for large n,

$$P \text{ is approximately } \mathrm{N}\left(\mu_P, \sigma_P^2\right)$$

(see Kendall and Stuart, 1976). A large positive (negative) value of $P - \mu_P$ indicates the presence of an increasing (decreasing) trend in the data. The assumption that $\{y_j\}$ is a sample from an iid sequence is therefore rejected at level $\alpha = 0.05$ if $|P - \mu_P|/\sigma_P > \Phi_{1-\alpha/2} = 1.96$.

(f) *Fitting an autoregressive model.* A further test that can be carried out using the program ITSM is to fit an autoregressive model to the data using the Yule–Walker algorithm (discussed in Section 5.1.1) and choosing the order which minimizes the AICC statistic (see Section 5.5). A selected order equal to zero suggests that the data is white noise.

(g) *Checking for normality.* If the noise process is Gaussian, i.e., if all of its joint distributions are normal, then stronger conclusions can be drawn when a model is fitted to the data. The following test enables us to check whether it is reasonable to assume that observations from an iid sequence are also Gaussian.

Let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(n)}$ be the order statistics of a random sample $Y_1, \ldots, Y_n$ from the distribution $\mathrm{N}(\mu, \sigma^2)$. If $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ are the order statistics from a $\mathrm{N}(0, 1)$ sample of size n, then

$$EY_{(j)} = \mu + \sigma m_j,$$

where $m_j = EX_{(j)}$, $j = 1, \ldots, n$. The graph of the points $\left(m_1, Y_{(1)}\right), \ldots, \left(m_n, Y_{(n)}\right)$ is called a Gaussian **qq plot**) and can be displayed in ITSM by clicking on the yellow button labeled QQ. If the normal assumption is correct, the Gaussian qq plot should be approximately linear. Consequently, the squared correlation of the points $(m_i, Y_{(i)})$, $i = 1, \ldots, n$, should be near 1. The assumption of normality is therefore rejected if the squared correlation R^2 is sufficiently small. If we approximate m_i by $\Phi^{-1}((i-.5)/n)$ (see Mage, 1982 for some alternative approximations), then R^2 reduces to

$$R^2 = \frac{\left(\sum_{i=1}^n (Y_{(i)} - \overline{Y})\Phi^{-1}\left(\frac{i-.5}{n}\right)\right)^2}{\sum_{i=1}^n (Y_{(i)} - \overline{Y})^2 \sum_{i=1}^n \left(\Phi^{-1}\left(\frac{i-.5}{n}\right)\right)^2},$$

where $\overline{Y} = n^{-1}(Y_1 + \cdots + Y_n)$. Percentage points for the distribution of R^2, assuming normality of the sample values, are given by Shapiro and Francia (1972) for sample sizes $n < 100$. For $n = 200$, $P(R^2 < .987) = .05$ and $P(R^2 < .989) = .10$. For larger values of n the Jarque-Bera test for normality can be used (see Section 5.3.3).

Example 1.6.1 If we did not know in advance how the signal plus noise data of Example 1.1.4 were generated, we might suspect that they came from an iid sequence. We can check this hypothesis with the aid of the tests (a)–(f) introduced above.

(a) The sample autocorrelation function (Figure 1.28) is obtained from ITSM by opening the project SIGNAL.TSM and clicking on the second yellow button at the top of the ITSM window. Observing that 25% of the autocorrelations are outside the bounds $\pm 1.96/\sqrt{200}$, we reject the hypothesis that the series is iid.

The remaining tests (b), (c), (d), (e), and (f) are performed by choosing the option `Statistics>Residual Analysis>Tests of Randomness`. (Since no model has been fitted to the data, the residuals are the same as the data themselves.)

(b) The sample value of the Ljung–Box statistic Q_{LB} with $h = 20$ is 51.84. Since the corresponding p-value (displayed by ITSM) is .00012 $<$ .05, we reject the iid hypothesis at level .05. The p-value for the McLeod–Li statistic Q_{ML} is 0.717. The McLeod–Li statistic does therefore not provide sufficient evidence to reject the iid hypothesis at level .05.

(c) The sample value of the turning-point statistic T is 138, and the asymptotic distribution under the iid hypothesis (with sample size $n = 200$) is N(132, 35.3). Thus $|T - \mu_T|/\sigma_T = 1.01$, corresponding to a computed p-value of .312. On the basis of the value of T there is therefore not sufficient evidence to reject the iid hypothesis at level .05.

(d) The sample value of the difference-sign statistic S is 101, and the asymptotic distribution under the iid hypothesis (with sample size $n = 200$) is N(99.5, 16.7). Thus $|S-\mu_S|/\sigma_S = 0.38$, corresponding to a computed p-value of 0.714. On the basis of the value of S there is therefore not sufficient evidence to reject the iid hypothesis at level .05.

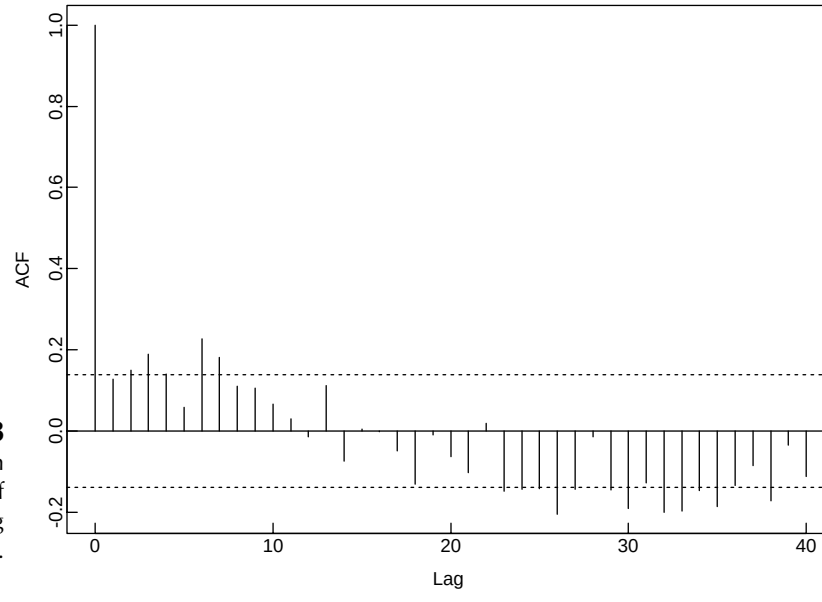

Figure 1-28
The sample autocorrelation function for the data of Example 1.1.4 showing the bounds $\pm 1.96/\sqrt{n}$.

(e) The sample value of the rank statistic P is 10310, and the asymptotic distribution under the iid hypothesis (with $n = 200$) is $\mathrm{N}(9950, 2.239 \times 10^5)$. Thus $|P - \mu_P|/\sigma_P = 0.76$, corresponding to a computed p-value of 0.447. On the basis of the value of P there is therefore not sufficient evidence to reject the iid hypothesis at level .05.

(f) The minimum-AICC Yule–Walker autoregressive model for the data is of order seven, supporting the evidence provided by the sample ACF and Ljung–Box tests against the iid hypothesis.

Thus, although not all of the tests detect significant deviation from iid behavior, the sample autocorrelation, the Ljung–Box statistic, and the fitted autoregression provide strong evidence against it, causing us to reject it (correctly) in this example. □

The general strategy in applying the tests described in this section is to check them all and to proceed with caution if any of them suggests a serious deviation from the iid hypothesis. (Remember that as you increase the number of tests, the probability that *at least one* rejects the null hypothesis when it is true increases. You should therefore not necessarily reject the null hypothesis on the basis of one test result only.)

Problems

1.1. Let X and Y be two random variables with $E(Y) = \mu$ and $EY^2 < \infty$.

a. Show that the constant c that minimizes $E(Y - c)^2$ is $c = \mu$.

b. Deduce that the random variable $f(X)$ that minimizes $E\big[(Y - f(X))^2 | X\big]$ is

$$f(X) = E[Y|X].$$

c. Deduce that the random variable $f(X)$ that minimizes $E(Y - f(X))^2$ is also

$$f(X) = E[Y|X].$$

1.2. (Generalization of Problem 1.1.) Suppose that $X_1, X_2, \ldots$ is a sequence of random variables with $E(X_t^2) < \infty$ and $E(X_t) = \mu$.

a. Show that the random variable $f(X_1, \ldots, X_n)$ that minimizes $E\big[(X_{n+1} - f(X_1, \ldots, X_n))^2 | X_1, \ldots, X_n\big]$ is

$$f(X_1, \ldots, X_n) = E[X_{n+1}|X_1, \ldots, X_n].$$

b. Deduce that the random variable $f(X_1, \ldots, X_n)$ that minimizes $E\big[(X_{n+1} - f(X_1, \ldots, X_n))^2\big]$ is also

$$f(X_1, \ldots, X_n) = E[X_{n+1}|X_1, \ldots, X_n].$$

c. If $X_1, X_2, \ldots$ is iid with $E(X_i^2) < \infty$ and $EX_i = \mu$, where μ is known, what is the minimum mean squared error predictor of X_{n+1} in terms of $X_1, \ldots, X_n$?

d. Under the conditions of part (c) show that the best linear unbiased estimator of μ in terms of $X_1, \ldots, X_n$ is $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. ($\hat{\mu}$ said to be an unbiased estimator of μ if $\mathrm{E}\hat{\mu} = \mu$ for all μ.)

e. Under the conditions of part (c) show that $\bar{X}$ is the best linear predictor of X_{n+1} that is unbiased for μ.

f. If $X_1, X_2, \ldots$ is iid with $E(X_i^2) < \infty$ and $EX_i = \mu$, and if $S_0 = 0$, $S_n = X_1 + \cdots + X_n$, $n = 1, 2, \ldots$, what is the minimum mean squared error predictor of S_{n+1} in terms of $S_1, \ldots, S_n$?

1.3. Show that a strictly stationary process with $E(X_i^2) < \infty$ is weakly stationary.

1.4. Let $\{Z_t\}$ be a sequence of independent normal random variables, each with mean 0 and variance σ^2, and let a, b, and c be constants. Which, if any, of the following processes are stationary? For each stationary process specify the mean and autocovariance function.

a. $X_t = a + bZ_t + cZ_{t-2}$

b. $X_t = Z_1 \cos(ct) + Z_2 \sin(ct)$

c. $X_t = Z_t \cos(ct) + Z_{t-1} \sin(ct)$

d. $X_t = a + bZ_0$

e. $X_t = Z_0 \cos(ct)$

f. $X_t = Z_t Z_{t-1}$

1.5. Let $\{X_t\}$ be the moving-average process of order 2 given by

$$X_t = Z_t + \theta Z_{t-2},$$

where $\{Z_t\}$ is WN(0, 1).

a. Find the autocovariance and autocorrelation functions for this process when $\theta = .8$.

b. Compute the variance of the sample mean $(X_1 + X_2 + X_3 + X_4)/4$ when $\theta = .8$.

c. Repeat (b) when $\theta = -.8$ and compare your answer with the result obtained in (b).

1.6. Let $\{X_t\}$ be the AR(1) process defined in Example 1.4.5.

a. Compute the variance of the sample mean $(X_1 + X_2 + X_3 + X_4)/4$ when $\phi = .9$ and $\sigma^2 = 1$.

b. Repeat (a) when $\phi = -.9$ and compare your answer with the result obtained in (a).

1.7. If $\{X_t\}$ and $\{Y_t\}$ are uncorrelated stationary sequences, i.e., if X_r and Y_s are uncorrelated for every r and s, show that $\{X_t + Y_t\}$ is stationary with autocovariance function equal to the sum of the autocovariance functions of $\{X_t\}$ and $\{Y_t\}$.

1.8. Let $\{Z_t\}$ be IID N(0, 1) noise and define

$$X_t = \begin{cases} Z_t, & \text{if } t \text{ is even,} \\ (Z_{t-1}^2 - 1)/\sqrt{2}, & \text{if } t \text{ is odd.} \end{cases}$$

a. Show that $\{X_t\}$ is WN(0, 1) but not iid(0, 1) noise.

b. Find $E(X_{n+1}|X_1, \ldots, X_n)$ for n odd and n even and compare the results.

1.9. Let $\{x_1, \ldots, x_n\}$ be observed values of a time series at times $1, \ldots, n$, and let $\hat{\rho}(h)$ be the sample ACF at lag h as in Definition 1.4.4.

a. If $x_t = a + bt$, where a and b are constants and $b \neq 0$, show that for each fixed $h \geq 1$,

$$\hat{\rho}(h) \to 1 \text{ as } n \to \infty.$$

b. If $x_t = c\cos(\omega t)$, where c and ω are constants ($c \neq 0$ and $\omega \in (-\pi, \pi]$), show that for each fixed h,

$$\hat{\rho}(h) \to \cos(\omega h) \text{ as } n \to \infty.$$

1.10. If $m_t = \sum_{k=0}^{p} c_k t^k$, $t = 0, \pm 1, \ldots$, show that ∇m_t is a polynomial of degree $p - 1$ in t and hence that $\nabla^{p+1} m_t = 0$.

1.11. Consider the simple moving-average filter with weights $a_j = (2q+1)^{-1}$, $-q \leq j \leq q$.

a. If $m_t = c_0 + c_1 t$, show that $\sum_{j=-q}^{q} a_j m_{t-j} = m_t$.

b. If Z_t, $t = 0, \pm 1, \pm 2, \ldots$, are independent random variables with mean 0 and variance σ^2, show that the moving average $A_t = \sum_{j=-q}^{q} a_j Z_{t-j}$ is "small" for large q in the sense that $EA_t = 0$ and $\text{Var}(A_t) = \sigma^2/(2q+1)$.

1.12. a. Show that a linear filter $\{a_j\}$ passes an arbitrary polynomial of degree k without distortion, i.e., that

$$m_t = \sum_j a_j m_{t-j}$$

for all kth-degree polynomials $m_t = c_0 + c_1 t + \cdots + c_k t^k$, if and only if

$$\begin{cases} \sum_j a_j = 1 & \text{and} \\ \sum_j j^r a_j = 0, & \text{for } r = 1, \ldots, k. \end{cases}$$

b. Deduce that the Spencer 15-point moving-average filter $\{a_j\}$ defined by (1.5.6) passes arbitrary third-degree polynomial trends without distortion.

1.13. Find a filter of the form $1 + \alpha B + \beta B^2 + \gamma B^3$ (i.e., find α, β, and γ) that passes linear trends without distortion and that eliminates arbitrary seasonal components of period 2.

1.14. Show that the filter with coefficients $[a_{-2}, a_{-1}, a_0, a_1, a_2] = \frac{1}{9}[-1, 4, 3, 4, -1]$ passes third-degree polynomials and eliminates seasonal components with period 3.

1.15. Let $\{Y_t\}$ be a stationary process with mean zero and let a and b be constants.

a. If $X_t = a + bt + s_t + Y_t$, where s_t is a seasonal component with period 12, show that $\nabla\nabla_{12}X_t = (1 - B)(1 - B^{12})X_t$ is stationary and express its autocovariance function in terms of that of $\{Y_t\}$.

b. If $X_t = (a + bt)s_t + Y_t$, where s_t is a seasonal component with period 12, show that $\nabla_{12}^2 X_t = (1 - B^{12})^2 X_t$ is stationary and express its autocovariance function in terms of that of $\{Y_t\}$.

1.16. (Using ITSM to smooth the strikes data.) Double-click on the ITSM icon, select `File>Project>Open>Univariate`, click OK, and open the file STRIKES.TSM. The graph of the data will then appear on your screen. To smooth the data select `Smooth>Moving Ave`, `Smooth>Exponential`, or `Smooth>FFT`. Try using each of these to reproduce the results shown in Figures 1.18, 1.21, and 1.22.

1.17. (Using ITSM to plot the deaths data.) In ITSM select `File>Project>Open>Univariate`, click OK, and open the project DEATHS.TSM. The graph of the data will then appear on your screen. To see a histogram of the data, click on the sixth yellow button at the top of the ITSM window. To see the sample autocorrelation function, click on the second yellow button. The presence of a strong seasonal component with period 12 is evident in the graph of the data and in the sample autocorrelation function.

1.18. (Using ITSM to analyze the deaths data.) Open the file DEATHS.TSM, select `Transform>Classical`, check the box marked `Seasonal Fit`, and enter 12 for the period. Make sure that the box labeled `Polynomial Fit` is not checked, and click, `OK`. You will then see the graph (Figure 1.24) of the deseasonalized data. This graph suggests the presence of an additional quadratic trend function. To fit such a trend to the deseasonalized data, select `Transform>Undo Classical` to retrieve the original data. Then select `Transform>Classical` and check the boxes marked `Seasonal Fit` and `Polynomial Trend`, entering 12 for the period and `Quadratic` for the trend. Click `OK` and you will obtain the trend function

$$\hat{m}_t = 9952 - 71.82t + 0.8260t^2, \quad 1 \le t \le 72.$$

At this point the data stored in ITSM consists of the estimated noise

$$\hat{Y}_t = x_t - \hat{m}_t - \hat{s}_t, \quad t = 1, \ldots, 72,$$

obtained by subtracting the estimated seasonal and trend components from the original data. The sample autocorrelation function can be plotted by clicking on the second yellow button at the top of the ITSM window. Further tests for dependence can be carried out by selecting the options `Statistics>Residual Analysis>Tests of Randomness`. It is clear from these that there is substantial dependence in the series $\{Y_t\}$.

To forecast the data without allowing for this dependence, select the option `Forecasting>ARMA`. Specify 24 for the number of values to be forecast, and the program will compute forecasts based on the assumption that the estimated seasonal and trend components are true values and that $\{Y_t\}$ is a white noise sequence with zero mean. (This is the default model assumed by ITSM until a more complicated stationary model is estimated or specified.) The original data are plotted with the forecasts appended.

Later we shall see how to improve on these forecasts by taking into account the dependence in the series $\{Y_t\}$.

1.19. Use a text editor, e.g., WORDPAD or NOTEPAD, to construct and save a text file named TEST.TSM, which consists of a single column of 30 numbers, $\{x_1, \ldots, x_{30}\}$, defined by

$$x_1, \ldots, x_{10} : 486, 474, 434, 441, 435, 401, 414, 414, 386, 405;$$

$$x_{11}, \ldots, x_{20} : 411, 389, 414, 426, 410, 441, 459, 449, 486, 510;$$

$$x_{21}, \ldots, x_{30} : 506, 549, 579, 581, 630, 666, 674, 729, 771, 785.$$

This series is in fact the sum of a quadratic trend and a period-three seasonal component. Use the program ITSM to apply the filter in Problem 1.14 to this time series and discuss the results.

(Once the data have been typed, they can be imported directly into ITSM by coping and pasting to the clipboard, and then in ITSM selecting `File>Project>New>Univariate`, clicking on OK and selecting `File>Import Clipboard`.)

2 Stationary Processes

2.1 Basic Properties
2.2 Linear Processes
2.3 Introduction to ARMA Processes
2.4 Properties of the Sample Mean and Autocorrelation Function
2.5 Forecasting Stationary Time Series
2.6 The Wold Decomposition

A key role in time series analysis is played by processes whose properties, or some of them, do not vary with time. If we wish to make predictions, then clearly we must assume that *something* does not vary with time. In extrapolating deterministic functions it is common practice to assume that either the function itself or one of its derivatives is constant. The assumption of a constant first derivative leads to linear extrapolation as a means of prediction. In time series analysis our goal is to predict a series that typically is not deterministic but contains a random component. If this random component is stationary, in the sense of Definition 1.4.2, then we can develop powerful techniques to forecast its future values. These techniques will be developed and discussed in this and subsequent chapters.

2.1 Basic Properties

In Section 1.4 we introduced the concept of stationarity and defined the autocovariance function (ACVF) of a stationary time series $\{X_t\}$ as

$$\gamma(h) = \text{Cov}(X_{t+h}, X_t), \quad h = 0, \pm 1, \pm 2, \ldots.$$

The autocorrelation function (ACF) of $\{X_t\}$ was defined similarly as the function $\rho(\cdot)$ whose value at lag h is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}.$$

The ACVF and ACF provide a useful measure of the degree of dependence among the values of a time series at different times and for this reason play an important role when we consider the prediction of future values of the series in terms of the past and present values. They can be estimated from observations of $X_1, \ldots, X_n$ by computing the *sample* ACVF and ACF as described in Section 1.4.1.

The role of the autocorrelation function in prediction is illustrated by the following simple example. Suppose that $\{X_t\}$ is a stationary Gaussian time series (see Definition A.3.2) and that we have observed X_n. We would like to find the function of X_n that gives us the best predictor of X_{n+h}, the value of the series after another h time units have elapsed. To define the problem we must first say what we mean by "best." A natural and computationally convenient definition is to specify our required predictor to be the function of X_n with minimum mean squared error. In this illustration, and indeed throughout the remainder of this book, we shall use this as our criterion for "best." Now by Proposition A.3.1 the conditional distribution of X_{n+h} given that $X_n = x_n$ is

$$\mathrm{N}\big(\mu + \rho(h)(x_n - \mu), \sigma^2\big(1 - \rho(h)^2\big)\big),$$

where μ and σ^2 are the mean and variance of $\{X_t\}$. It was shown in Problem 1.1 that the value of the constant c that minimizes $E(X_{n+h} - c)^2$ is $c = E(X_{n+h})$ and that the function m of X_n that minimizes $E(X_{n+h} - m(X_n))^2$ is the conditional mean

$$m(X_n) = E(X_{n+h}|X_n) = \mu + \rho(h)(X_n - \mu). \tag{2.1.1}$$

The corresponding mean squared error is

$$E(X_{n+h} - m(X_n))^2 = \sigma^2\big(1 - \rho(h)^2\big). \tag{2.1.2}$$

This calculation shows that at least for stationary Gaussian time series, prediction of X_{n+h} in terms of X_n is more accurate as $|\rho(h)|$ becomes closer to 1, and in the limit as $\rho \to \pm 1$ the best predictor approaches $\mu \pm (X_n - \mu)$ and the corresponding mean squared error approaches 0.

In the preceding calculation the assumption of joint normality of X_{n+h} and X_n played a crucial role. For time series with nonnormal joint distributions the corresponding calculations are in general much more complicated. However, if instead of looking for the best function of X_n for predicting X_{n+h}, we look for the best **linear predictor**, i.e., the best predictor of the form $\ell(X_n) = aX_n + b$, then our problem becomes that of finding a and b to minimize $E(X_{n+h} - aX_n - b)^2$. An elementary calculation (Problem 2.1), shows that the best predictor of this form is

$$\ell(X_n) = \mu + \rho(h)(X_n - \mu) \tag{2.1.3}$$

with corresponding mean squared error

$$E(X_{n+h} - \ell(X_n))^2 = \sigma^2(1 - \rho(h)^2). \tag{2.1.4}$$

Comparison with (2.1.1) and (2.1.3) shows that for Gaussian processes, $\ell(X_n)$ and $m(X_n)$ are the same. In general, of course, $m(X_n)$ will give smaller mean squared error than $\ell(X_n)$, since it is the best of a larger class of predictors (see Problem 1.8). However, the fact that the best linear predictor depends only on the mean and ACF of the series $\{X_t\}$ means that it can be calculated without more detailed knowledge of the joint distributions. This is extremely important in practice because of the difficulty of estimating all of the joint distributions and because of the difficulty of computing the required conditional expectations even if the distributions were known.

As we shall see later in this chapter, similar conclusions apply when we consider the more general problem of predicting X_{n+h} as a function not only of X_n, but also of $X_{n-1}, X_{n-2}, \ldots$. Before pursuing this question we need to examine in more detail the properties of the autocovariance and autocorrelation functions of a stationary time series.

Basic Properties of $\gamma(\cdot)$:

$$\gamma(0) \geq 0,$$

$$|\gamma(h)| \leq \gamma(0) \text{ for all } h,$$

and $\gamma(\cdot)$ is even, i.e.,

$$\gamma(h) = \gamma(-h) \text{ for all } h.$$

Proof The first property is simply the statement that $\text{Var}(X_t) \geq 0$, the second is an immediate consequence of the fact that correlations are less than or equal to 1 in absolute value (or the Cauchy–Schwarz inequality), and the third is established by observing that

$$\gamma(h) = \text{Cov}(X_{t+h}, X_t) = \text{Cov}(X_t, X_{t+h}) = \gamma(-h). \qquad \blacksquare$$

Autocovariance functions have another fundamental property, namely that of nonnegative definiteness.

Definition 2.1.1 A real-valued function κ defined on the integers is **nonnegative definite** if

$$\sum_{i,j=1}^{n} a_i \kappa(i - j) a_j \geq 0 \tag{2.1.5}$$

for all positive integers n and vectors $\mathbf{a} = (a_1, \ldots, a_n)'$ with real-valued components a_i.

Theorem 2.1.1 *A real-valued function defined on the integers is the autocovariance function of a stationary time series if and only if it is even and nonnegative definite.*

Proof To show that the autocovariance function $\gamma(\cdot)$ of any stationary time series $\{X_t\}$ is nonnegative definite, let $\mathbf{a}$ be any $n \times 1$ vector with real components $a_1, \ldots, a_n$ and let $\mathbf{X}_n = (X_n, \ldots, X_1)'$. Then by equation (A.2.5) and the nonnegativity of variances,

$$\text{Var}(\mathbf{a}'\mathbf{X}_n) = \mathbf{a}'\Gamma_n\mathbf{a} = \sum_{i,j=1}^{n} a_i\gamma(i-j)a_j \geq 0,$$

where Γ_n is the covariance matrix of the random vector $\mathbf{X}_n$. The last inequality, however, is precisely the statement that $\gamma(\cdot)$ is nonnegative definite. The converse result, that there exists a stationary time series with autocovariance function κ if κ is even, real-valued, and nonnegative definite, is more difficult to establish (see TSTM, Theorem 1.5.1 for a proof). A slightly stronger statement can be made, namely, that under the specified conditions there exists a stationary *Gaussian* time series $\{X_t\}$ with mean 0 and autocovariance function $\kappa(\cdot)$. ■

Remark 1. An autocorrelation function $\rho(\cdot)$ has all the properties of an autocovariance function and satisfies the additional condition $\rho(0) = 1$. In particular, we can say that $\rho(\cdot)$ is the autocorrelation function of a stationary process if and only if $\rho(\cdot)$ is an ACVF with $\rho(0) = 1$. □

Remark 2. To verify that a given function is nonnegative definite it is often simpler to find a stationary process that has the given function as its ACVF than to verify the conditions (2.1.5) directly. For example, the function $\kappa(h) = \cos(\omega h)$ is nonnegative definite, since (see Problem 2.2) it is the ACVF of the stationary process

$$X_t = A\cos(\omega t) + B\sin(\omega t),$$

where A and B are uncorrelated random variables, both with mean 0 and variance 1. Another illustration is provided by the following example. □

Example 2.1.1 We shall show now that the function defined on the integers by

$$\kappa(h) = \begin{cases} 1, & \text{if } h = 0, \\ \rho, & \text{if } h = \pm 1, \\ 0, & \text{otherwise,} \end{cases}$$

is the ACVF of a stationary time series if and only if $|\rho| \leq \frac{1}{2}$. Inspection of the ACVF of the MA(1) process of Example 1.4.4 shows that κ is the ACVF of such a process if we can find real θ and nonnegative σ^2 such that

$$\sigma^2(1+\theta^2) = 1$$

and

$$\sigma^2\theta = \rho.$$

If $|\rho| \le \frac{1}{2}$, these equations give solutions $\theta = (2\rho)^{-1}\left(1 \pm \sqrt{1-4\rho^2}\right)$ and $\sigma^2 = \left(1+\theta^2\right)^{-1}$. However, if $|\rho| > \frac{1}{2}$, there is no real solution for θ and hence no MA(1) process with ACVF κ. To show that there is no *stationary* process with ACVF κ, we need to show that κ is not nonnegative definite. We shall do this directly from the definition (2.1.5). First, if $\rho > \frac{1}{2}$, $K = [\kappa(i-j)]_{i,j=1}^{n}$, and $\mathbf{a}$ is the n-component vector $\mathbf{a} = (1, -1, 1, -1, \ldots)'$, then

$$\mathbf{a}'K\mathbf{a} = n - 2(n-1)\rho < 0 \text{ for } n > 2\rho/(2\rho-1),$$

showing that $\kappa(\cdot)$ is not nonnegative definite and therefore, by Theorem 2.1.1, is not an autocovariance function. If $\rho < -\frac{1}{2}$, the same argument with $\mathbf{a} = (1, 1, 1, 1, \ldots)'$ again shows that $\kappa(\cdot)$ is not nonnegative definite.

If $\{X_t\}$ is a (weakly) stationary time series, then the vector $(X_1, \ldots, X_n)'$ and the time-shifted vector $(X_{1+h}, \ldots, X_{n+h})'$ have the same mean vectors and covariance matrices for every integer h and positive integer n. A strictly stationary sequence is one in which the joint distributions of these two vectors (and not just the means and covariances) are the same. The precise definition is given below. □

Definition 2.1.2

$\{X_t\}$ is a **strictly stationary time series** if

$$(X_1, \ldots, X_n)' \stackrel{d}{=} (X_{1+h}, \ldots, X_{n+h})'$$

for all all integers h and $n \ge 1$. (Here $\stackrel{d}{=}$ is used to indicate that the two random vectors have the same joint distribution function.)

For reference, we record some of the elementary properties of strictly stationary time series.

Properties of a Strictly Stationary Time Series $\{X_t\}$:

- **a.** The random variables X_t are identically distributed.
- **b.** $(X_t, X_{t+h})' \stackrel{d}{=} (X_1, X_{1+h})'$ for all integers t and h.
- **c.** $\{X_t\}$ is weakly stationary if $E(X_t^2) < \infty$ for all t.
- **d.** Weak stationarity does not imply strict stationarity.
- **e.** An iid sequence is strictly stationary.

Proof Properties (a) and (b) follow at once from Definition 2.1.2. If $EX_t^2 < \infty$, then by (a) and (b) EX_t is independent of t and $\text{Cov}(X_t, X_{t+h}) = \text{Cov}(X_1, X_{1+h})$, which is also

independent of t, proving (c). For (d) see Problem 1.8. If $\{X_t\}$ is an iid sequence of random variables with common distribution function F, then the joint distribution function of $(X_{1+h}, \ldots, X_{n+h})'$ evaluated at $(x_1, \ldots, x_n)'$ is $F(x_1) \cdots F(x_n)$, which is independent of h. ■

One of the simplest ways to construct a time series $\{X_t\}$ that is strictly stationary (and hence stationary if $EX_t^2 < \infty$) is to "filter" an iid sequence of random variables. Let $\{Z_t\}$ be an iid sequence, which by (e) is strictly stationary, and define

$$X_t = g(Z_t, Z_{t-1}, \ldots, Z_{t-q}) \tag{2.1.6}$$

for some real-valued function $g(\cdot, \ldots, \cdot)$. Then $\{X_t\}$ is strictly stationary, since $(Z_{t+h}, \ldots, Z_{t+h-q})' \stackrel{d}{=} (Z_t, \ldots, Z_{t-q})'$ for all integers h. It follows also from the defining equation (2.1.6) that $\{X_t\}$ is ***q*-dependent**, i.e., that X_s and X_t are independent whenever $|t - s| > q$. (An iid sequence is 0-dependent.) In the same way, adopting a second-order viewpoint, we say that a stationary time series is ***q*-correlated** if $\gamma(h) = 0$ whenever $|h| > q$. A white noise sequence is then 0-correlated, while the MA(1) process of Example 1.4.4 is 1-correlated. The moving-average process of order q defined below is q-correlated, and perhaps surprisingly, the converse is also true (Proposition 2.1.1).

The MA(q) Process:

$\{X_t\}$ is a **moving-average process of order *q*** if

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \tag{2.1.7}$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and $\theta_1, \ldots, \theta_q$ are constants.

It is a simple matter to check that (2.1.7) defines a stationary time series that is strictly stationary if $\{Z_t\}$ is iid noise. In the latter case, (2.1.7) is a special case of (2.1.6) with g a linear function.

The importance of MA(q) processes derives from the fact that *every* q-correlated process is an MA(q) process. This is the content of the following proposition, whose proof can be found in TSTM, Section 3.2. The extension of this result to the case $q = \infty$ is essentially Wold's decomposition (see Section 2.6).

Proposition 2.1.1 *If $\{X_t\}$ is a stationary q-correlated time series with mean* 0, *then it can be represented as the* MA(q) *process in* (*2.1.7*).

2.2 Linear Processes

The class of linear time series models, which includes the class of autoregressive moving-average (ARMA) models, provides a general framework for studying stationary processes. In fact, every second-order stationary process is either a linear process or can be transformed to a linear process by subtracting a *deterministic* component. This result is known as Wold's decomposition and is discussed in Section 2.6.

Definition 2.2.1

The time series $\{X_t\}$ is a **linear process** if it has the representation

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \tag{2.2.1}$$

for all t, where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and $\{\psi_j\}$ is a sequence of constants with $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

In terms of the backward shift operator B, (2.2.1) can be written more compactly as

$$X_t = \psi(B)Z_t, \tag{2.2.2}$$

where $\psi(B) = \sum_{j=-\infty}^{\infty} \psi_j B^j$. A linear process is called a **moving average** or **MA(∞)** if $\psi_j = 0$ for all $j < 0$, i.e., if

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}.$$

Remark 1. The condition $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ ensures that the infinite sum in (2.2.1) converges (with probability one), since $E|Z_t| \le \sigma$ and

$$E|X_t| \le \sum_{j=-\infty}^{\infty} \left(|\psi_j| E|Z_{t-j}|\right) \le \left(\sum_{j=-\infty}^{\infty} |\psi_j|\right) \sigma < \infty.$$

It also ensures that $\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty$ and hence (see Appendix C, Example C.1.1) that the series in (2.2.1) converges in mean square, i.e., that X_t is the mean square limit of the partial sums $\sum_{j=-n}^{n} \psi_j Z_{t-j}$. The condition $\sum_{j=-n}^{n} |\psi_j| < \infty$ also ensures convergence in both senses of the more general series (2.2.3) considered in Proposition 2.2.1 below. In Section 10.5 we consider a more general class of linear processes, the fractionally integrated ARMA processes, for which the coefficents are not absolutely summable but only square summable. □

The operator $\psi(B)$ can be thought of as a linear filter, which when applied to the white noise "input" series $\{Z_t\}$ produces the "output" $\{X_t\}$ (see Section 4.3). As established in the following proposition, a linear filter, when applied to any stationary input series, produces a stationary output series.

Proposition 2.2.1 *Let $\{Y_t\}$ be a stationary time series with mean 0 and covariance function γ_Y. If $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, then the time series*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j} = \psi(B)Y_t \tag{2.2.3}$$

is stationary with mean 0 and autocovariance function

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Y(h + k - j). \tag{2.2.4}$$

In the special case where $\{X_t\}$ is a linear process,

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h} \sigma^2. \tag{2.2.5}$$

Proof The argument used in Remark 1, with σ replaced by $\sqrt{\gamma_Y(0)}$, shows that the series in (2.2.3) is convergent. Since $EY_t = 0$, we have

$$E(X_t) = E\left(\sum_{j=-\infty}^{\infty} \psi_j Y_{t-j}\right) = \sum_{j=-\infty}^{\infty} \psi_j E(Y_{t-j}) = 0$$

and

$$\begin{aligned} E(X_{t+h}X_t) &= E\left[\left(\sum_{j=-\infty}^{\infty} \psi_j Y_{t+h-j}\right)\left(\sum_{k=-\infty}^{\infty} \psi_k Y_{t-k}\right)\right] \\ &= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k E(Y_{t+h-j} Y_{t-k}) \\ &= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Y(h - j + k), \end{aligned}$$

which shows that $\{X_t\}$ is stationary with covariance function (2.2.4). (The interchange of summation and expectation operations in the above calculations can be justified by the absolute summability of ψ_j.) Finally, if $\{Y_t\}$ is the white noise sequence $\{Z_t\}$ in (2.2.1), then $\gamma_Y(h - j + k) = \sigma^2$ if $k = j - h$ and 0 otherwise, from which (2.2.5) follows. ■

Remark 2. The absolute convergence of (2.2.3) implies (Problem 2.6) that filters of the form $\alpha(B) = \sum_{j=-\infty}^{\infty} \alpha_j B^j$ and $\beta(B) = \sum_{j=-\infty}^{\infty} \beta_j B^j$ with absolutely summable coefficients can be applied successively to a stationary series $\{Y_t\}$ to generate a new stationary series

$$W_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j},$$

where

$$\psi_j = \sum_{k=-\infty}^{\infty} \alpha_k \beta_{j-k} = \sum_{k=-\infty}^{\infty} \beta_k \alpha_{j-k}. \tag{2.2.6}$$

These relations can be expressed in the equivalent form

$$W_t = \psi(B)Y_t,$$

where

$$\psi(B) = \alpha(B)\beta(B) = \beta(B)\alpha(B), \tag{2.2.7}$$

and the products are defined by (2.2.6) or equivalently by multiplying the series $\sum_{j=-\infty}^{\infty} \alpha_j B^j$ and $\sum_{j=-\infty}^{\infty} \beta_j B^j$ term by term and collecting powers of B. It is clear from (2.2.6) and (2.2.7) that the order of application of the filters $\alpha(B)$ and $\beta(B)$ is immaterial. □

Example 2.2.1 An AR(1) process

In Example 1.4.5, an AR(1) process was defined as a stationary solution $\{X_t\}$ of the equations

$$X_t - \phi X_{t-1} = Z_t, \tag{2.2.8}$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, $|\phi| < 1$, and Z_t is uncorrelated with X_s for each $s < t$. To show that such a solution exists and is the unique stationary solution of (2.2.8), we consider the linear process defined by

$$X_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}. \tag{2.2.9}$$

(The coefficients ϕ^j for $j \geq 0$ are absolutely summable, since $|\phi| < 1$.) It is easy to verify directly that the process (2.2.9) is a solution of (2.2.8), and by Proposition 2.2.1 it is also stationary with mean 0 and ACVF

$$\gamma_X(h) = \sum_{j=0}^{\infty} \phi^j \phi^{j+h} \sigma^2 = \frac{\sigma^2 \phi^h}{1 - \phi^2},$$

for $h \geq 0$.

To show that (2.2.9) is the *only* stationary solution of (2.2.8) let $\{Y_t\}$ be *any* stationary solution. Then, iterating (2.2.8), we obtain

$$\begin{aligned} Y_t &= \phi Y_{t-1} + Z_t \\ &= Z_t + \phi Z_{t-1} + \phi^2 Y_{t-2} \\ &= \cdots \\ &= Z_t + \phi Z_{t-1} + \cdots + \phi^k Z_{t-k} + \phi^{k+1} Y_{t-k-1}. \end{aligned}$$

If $\{Y_t\}$ is stationary, then EY_t^2 is finite and independent of t, so that

$$E(Y_t - \sum_{j=0}^{k} \phi^j Z_{t-j})^2 = \phi^{2k+2} E(Y_{t-k-1})^2$$

$$\to 0 \text{ as } k \to \infty.$$

This implies that Y_t is equal to the mean square limit $\sum_{j=0}^{\infty} \phi^j Z_{t-j}$ and hence that the process defined by (2.2.9) is the unique stationary solution of the equations (2.2.8).

It the case $|\phi| > 1$, the series in (2.2.9) does not converge. However, we can rewrite (2.2.8) in the form

$$X_t = -\phi^{-1} Z_{t+1} + \phi^{-1} X_{t+1}. \tag{2.2.10}$$

Iterating (2.2.10) gives

$$\begin{aligned} X_t &= -\phi^{-1} Z_{t+1} - \phi^{-2} Z_{t+2} + \phi^{-2} X_{t+2} \\ &= \cdots \\ &= -\phi^{-1} Z_{t+1} - \cdots - \phi^{-k-1} Z_{t+k+1} + \phi^{-k-1} X_{t+k+1}, \end{aligned}$$

which shows, by the same arguments used above, that

$$X_t = -\sum_{j=1}^{\infty} \phi^{-j} Z_{t+j} \tag{2.2.11}$$

is the unique stationary solution of (2.2.8). This solution should not be confused with the nonstationary solution $\{X_t\}$ of (2.2.8) obtained when X_0 is any specified random variable that is uncorrelated with $\{Z_t\}$.

The solution (2.2.11) is frequently regarded as unnatural, since X_t as defined by (2.2.11) is correlated with *future* values of Z_s, contrasting with the solution (2.2.9), which has the property that X_t is uncorrelated with Z_s for all $s > t$. It is customary therefore in modeling stationary time series to restrict attention to AR(1) processes with $|\phi| < 1$. Then X_t has the representation (2.2.8) in terms of $\{Z_s, s \le t\}$, and we say that $\{X_t\}$ is a **causal** or **future-independent** function of $\{Z_t\}$, or more concisely that $\{X_t\}$ is a causal autoregressive process. It should be noted that every AR(1) process with $|\phi| > 1$ can be reexpressed as an AR(1) process with $|\phi| < 1$ and a new white noise sequence (Problem 3.8). From a second-order point of view, therefore, nothing is lost by eliminating AR(1) processes with $|\phi| > 1$ from consideration.

If $\phi = \pm 1$, there is no stationary solution of (2.2.8) (see Problem 2.8). □

Remark 3. It is worth remarking that when $|\phi| < 1$ the unique stationary solution (2.2.9) can be found immediately with the aid of (2.2.7). To do this let $\phi(B) = 1 - \phi B$ and $\pi(B) = \sum_{j=0}^{\infty} \phi^j B^j$. Then

$$\psi(B) := \phi(B)\pi(B) = 1.$$

Applying the operator $\pi(B)$ to both sides of (2.2.8), we obtain

$$X_t = \pi(B)Z_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$$

as claimed. □

2.3 Introduction to ARMA Processes

In this section we introduce, through an example, some of the key properties of an important class of linear processes known as ARMA (autoregressive moving average) processes. These are defined by linear difference equations with constant coefficients. As our example we shall consider the ARMA(1,1) process. Higher-order ARMA processes will be discussed in Chapter 3.

Definition 2.3.1

The time series $\{X_t\}$ is an **ARMA(1, 1) process** if it is stationary and satisfies (for every t)

$$X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}, \tag{2.3.1}$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and $\phi + \theta \neq 0$.

Using the backward shift operator B, (2.3.1) can be written more concisely as

$$\phi(B)X_t = \theta(B)Z_t, \tag{2.3.2}$$

where $\phi(B)$ and $\theta(B)$ are the linear filters

$$\phi(B) = 1 - \phi B \text{ and } \theta(B) = 1 + \theta B,$$

respectively.

We first investigate the range of values of ϕ and θ for which a stationary solution of (2.3.1) exists. If $|\phi| < 1$, let $\chi(z)$ denote the power series expansion of $1/\phi(z)$, i.e., $\sum_{j=0}^{\infty} \phi^j z^j$, which has absolutely summable coefficients. Then from (2.2.7) we conclude that $\chi(B)\phi(B) = 1$. Applying $\chi(B)$ to each side of (2.3.2) therefore gives

$$X_t = \chi(B)\theta(B)Z_t = \psi(B)Z_t,$$

where

$$\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j = \left(1 + \phi B + \phi^2 B^2 + \cdots\right)(1 + \theta B).$$

By multiplying out the right-hand side or using (2.2.6), we find that

$$\psi_0 = 1 \text{ and } \psi_j = (\phi + \theta)\phi^{j-1} \text{ for } j \geq 1.$$

As in Example 2.2.1, we conclude that the MA(∞) process

$$X_t = Z_t + (\phi + \theta)\sum_{j=1}^{\infty} \phi^{j-1} Z_{t-j} \tag{2.3.3}$$

is the unique stationary solution of (2.3.1).

Now suppose that $|\phi| > 1$. We first represent $1/\phi(z)$ as a series of powers of z with absolutely summable coefficients by expanding in powers of z^{-1}, giving (Problem 2.7)

$$\frac{1}{\phi(z)} = -\sum_{j=1}^{\infty} \phi^{-j} z^{-j}.$$

Then we can apply the same argument as in the case where $|\phi| < 1$ to obtain the unique stationary solution of (2.3.1). We let $\chi(B) = -\sum_{j=1}^{\infty} \phi^{-j} B^{-j}$ and apply $\chi(B)$ to each side of (2.3.2) to obtain

$$X_t = \chi(B)\theta(B)Z_t = -\theta\phi^{-1} Z_t - (\theta + \phi)\sum_{j=1}^{\infty} \phi^{-j-1} Z_{t+j}. \tag{2.3.4}$$

If $\phi = \pm 1$, there is no stationary solution of (2.3.1). Consequently, there is no such thing as an ARMA(1,1) process with $\phi = \pm 1$ according to our definition.

We can now summarize our findings about the existence and nature of the stationary solutions of the ARMA(1,1) recursions (2.3.2) as follows:

- A stationary solution of the ARMA(1,1) equations exists if and only if $\phi \neq \pm 1$.
- If $|\phi| < 1$, then the unique stationary solution is given by (2.3.3). In this case we say that $\{X_t\}$ is **causal** or a causal function of $\{Z_t\}$, since X_t can be expressed in terms of the current and past values Z_s, $s \leq t$.
- If $|\phi| > 1$, then the unique stationary solution is given by (2.3.4). The solution is **noncausal**, since X_t is then a function of Z_s, $s \geq t$.

Just as causality means that X_t is expressible in terms of Z_s, $s \leq t$, the dual concept of invertibility means that Z_t is expressible in terms of X_s, $s \leq t$. We show now that the ARMA(1,1) process defined by (2.3.1) is invertible if $|\theta| < 1$. To demonstrate this, let $\xi(z)$ denote the power series expansion of $1/\theta(z)$, i.e., $\sum_{j=0}^{\infty}(-\theta)^j z^j$, which has absolutely summable coefficients. From (2.2.7) it therefore follows that $\xi(B)\theta(B) = 1$, and applying $\xi(B)$ to each side of (2.3.2) gives

$$Z_t = \xi(B)\phi(B)X_t = \pi(B)X_t,$$

where

$$\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j = \left(1 - \theta B + (-\theta)^2 B^2 + \cdots\right)(1 - \phi B).$$

By multiplying out the right-hand side or using (2.2.6), we find that

$$Z_t = X_t - (\phi + \theta)\sum_{j=1}^{\infty}(-\theta)^{j-1}X_{t-j}. \tag{2.3.5}$$

Thus the ARMA(1,1) process is **invertible**, since Z_t can be expressed in terms of the present and past values of the process X_s, $s \le t$. An argument like the one used to show noncausality when $|\phi| > 1$ shows that the ARMA(1,1) process is **noninvertible** when $|\theta| > 1$, since then

$$Z_t = -\phi\theta^{-1}X_t + (\theta + \phi)\sum_{j=1}^{\infty}(-\theta)^{-j-1}X_{t+j}. \tag{2.3.6}$$

We summarize these results as follows:

- If $|\theta| < 1$, then the ARMA(1,1) process is **invertible**, and Z_t is expressed in terms of $X_s, s \le t$, by (2.3.5).
- If $|\theta| > 1$, then the ARMA(1,1) process is **noninvertible**, and Z_t is expressed in terms of $X_s, s \ge t$, by (2.3.6).

Remark 1. In the cases $\theta = \pm 1$, the ARMA(1,1) process is invertible in the more general sense that Z_t is a mean square limit of finite linear combinations of X_s, $s \le t$, although it cannot be expressed explicitly as an infinite linear combination of X_s, $s \le t$ (see Section 4.4 of TSTM). In this book the term *invertible* will always be used in the more restricted sense that $Z_t = \sum_{j=0}^{\infty}\pi_j X_{t-j}$, where $\sum_{j=0}^{\infty}|\pi_j| < \infty$. □

Remark 2. If the ARMA(1,1) process $\{X_t\}$ is noncausal or noninvertible with $|\theta| > 1$, then it is possible to find a new white noise sequence $\{W_t\}$ such that $\{X_t\}$ is a causal and noninvertible ARMA(1,1) process relative to $\{W_t\}$ (Problem 4.10). Therefore, from a second-order point of view, nothing is lost by restricting attention to causal and invertible ARMA(1,1) models. This last sentence is also valid for higher-order ARMA models. □

2.4 Properties of the Sample Mean and Autocorrelation Function

A stationary process $\{X_t\}$ is characterized, at least from a second-order point of view, by its mean μ and its autocovariance function $\gamma(\cdot)$. The estimation of μ, $\gamma(\cdot)$, and the autocorrelation function $\rho(\cdot) = \gamma(\cdot)/\gamma(0)$ from observations $X_1, \ldots, X_n$ therefore plays a crucial role in problems of inference and in particular in the problem of constructing an appropriate model for the data. In this section we examine some of the properties of the sample estimates $\bar{x}$ and $\hat{\rho}(\cdot)$ of μ and $\rho(\cdot)$, respectively.

2.4.1 Estimation of μ

The moment estimator of the mean μ of a stationary process is the sample mean

$$\bar{X}_n = n^{-1}(X_1 + X_2 + \cdots + X_n). \tag{2.4.1}$$

It is an unbiased estimator of μ, since

$$E(\bar{X}_n) = n^{-1}(EX_1 + \cdots + EX_n) = \mu.$$

The mean squared error of $\bar{X}_n$ is

$$\begin{aligned} E(\bar{X}_n - \mu)^2 &= \mathrm{Var}(\bar{X}_n) \\ &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j) \\ &= n^{-2}\sum_{i-j=-n}^{n}(n - |i-j|)\gamma(i-j) \\ &= n^{-1}\sum_{h=-n}^{n}\left(1 - \frac{|h|}{n}\right)\gamma(h). \end{aligned} \tag{2.4.2}$$

Now if $\gamma(h) \to 0$ as $h \to \infty$, the right-hand side of (2.4.2) converges to zero, so that $\bar{X}_n$ converges in mean square to μ. If $\sum_{h=-\infty}^{\infty}|\gamma(h)| < \infty$, then (2.4.2) gives $\lim_{n\to\infty} n\mathrm{Var}(\bar{X}_n) = \sum_{|h|<\infty}\gamma(h)$. We record these results in the following proposition.

Proposition 2.4.1 *If $\{X_t\}$ is a stationary time series with mean μ and autocovariance function $\gamma(\cdot)$, then as $n \to \infty$,*

$$\mathrm{Var}(\bar{X}_n) = E(\bar{X}_n - \mu)^2 \to 0 \qquad \text{if } \gamma(n) \to 0,$$

$$nE(\bar{X}_n - \mu)^2 \to \sum_{|h|<\infty}\gamma(h) \qquad \text{if } \sum_{h=-\infty}^{\infty}|\gamma(h)| < \infty.$$

To make inferences about μ using the sample mean $\bar{X}_n$, it is necessary to know the distribution or an approximation to the distribution of $\bar{X}_n$. If the time series is Gaussian (see Definition A.3.2), then by Remark 2 of Section A.3 and (2.4.2),

$$n^{1/2}(\bar{X}_n - \mu) \sim \mathrm{N}\left(0, \sum_{|h|<n}\left(1 - \frac{|h|}{n}\right)\gamma(h)\right).$$

It is easy to construct exact confidence bounds for μ using this result if $\gamma(\cdot)$ is known, and approximate confidence bounds if it is necessary to estimate $\gamma(\cdot)$ from the observations.

For many time series, in particular for linear and ARMA models, $\bar{X}_n$ is approximately normal with mean μ and variance $n^{-1}\sum_{|h|<\infty}\gamma(h)$ for large n (see TSTM, p. 219). An approximate 95% confidence interval for μ is then

$$\left(\bar{X}_n - 1.96v^{1/2}/\sqrt{n},\, \bar{X}_n + 1.96v^{1/2}/\sqrt{n}\right), \tag{2.4.3}$$

where $v = \sum_{|h|<\infty}\gamma(h)$. Of course, v is not generally known, so it must be estimated from the data. The estimator computed in the program ITSM is $\hat{v} = \sum_{|h|<\sqrt{n}}\big(1 - |h|/n\big)\hat{\gamma}(h)$. For ARMA processes this is a good approximation to v for large n.

Example 2.4.1 An AR(1) model

Let $\{X_t\}$ be an AR(1) process with mean μ, defined by the equations

$$X_t - \mu = \phi(X_{t-1} - \mu) + Z_t,$$

where $|\phi| < 1$ and $\{Z_t\} \sim \text{WN}\big(0, \sigma^2\big)$. From Example 2.2.1 we have $\gamma(h) = \phi^{|h|}\sigma^2/(1-\phi^2)$ and hence $v = \big(1+2\sum_{h=1}^{\infty}\phi^h\big)\sigma^2/\big(1-\phi^2\big) = \sigma^2/(1-\phi)^2$. Approximate 95% confidence bounds for μ are therefore given by $\bar{x}_n \pm 1.96\sigma n^{-1/2}/(1-\phi)$. Since ϕ and σ are unknown in practice, they must be replaced in these bounds by estimated values. □

2.4.2 Estimation of $\gamma(\cdot)$ and $\rho(\cdot)$

Recall from Section 1.4.1 that the sample autocovariance and autocorrelation functions are defined by

$$\hat{\gamma}(h) = n^{-1}\sum_{t=1}^{n-|h|}\big(X_{t+|h|} - \bar{X}_n\big)\big(X_t - \bar{X}_n\big) \tag{2.4.4}$$

and

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}. \tag{2.4.5}$$

Both the estimators $\hat{\gamma}(h)$ and $\hat{\rho}(h)$ are biased even if the factor n^{-1} in (2.4.4) is replaced by $(n-h)^{-1}$. Nevertheless, under general assumptions they are nearly unbiased for large sample sizes. The sample ACVF has the desirable property that for each $k \geq 1$ the k-dimensional sample covariance matrix

$$\hat{\Gamma}_k = \begin{bmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \cdots & \hat{\gamma}(k-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \cdots & \hat{\gamma}(k-2) \\ \vdots & \vdots & \cdots & \vdots \\ \hat{\gamma}(k-1) & \hat{\gamma}(k-2) & \cdots & \hat{\gamma}(0) \end{bmatrix} \tag{2.4.6}$$

is nonnegative definite. To see this, first note that if $\hat{\Gamma}_m$ is nonnegative definite, then $\hat{\Gamma}_k$ is nonnegative definite for all $k < m$. So assume $k \geq n$ and write

$$\hat{\Gamma}_k = n^{-1}TT',$$

where T is the $k \times 2k$ matrix

$$T = \begin{bmatrix} 0 & \cdots & 0 & 0 & Y_1 & Y_2 & \cdots & Y_k \\ 0 & \cdots & 0 & Y_1 & Y_2 & \cdots & Y_k & 0 \\ \vdots & & & & & & & \vdots \\ 0 & Y_1 & Y_2 & \cdots & Y_k & 0 \cdots & 0 & \end{bmatrix},$$

$Y_i = X_i - \bar{X}_n$, $i = 1, \ldots, n$, and $Y_i = 0$ for $i = n+1, \ldots, k$. Then for any real $k \times 1$ vector $\mathbf{a}$ we have

$$\mathbf{a}'\hat{\Gamma}_k\mathbf{a} = n^{-1}(\mathbf{a}'T)(T'\mathbf{a}) \geq 0, \tag{2.4.7}$$

and consequently the sample autocovariance matrix $\hat{\Gamma}_k$ and sample autocorrelation matrix

$$\hat{R}_k = \hat{\Gamma}_k/\gamma(0) \tag{2.4.8}$$

are nonnegative definite. Sometimes the factor n^{-1} is replaced by $(n-h)^{-1}$ in the definition of $\hat{\gamma}(h)$, but the resulting covariance and correlation matrices $\hat{\Gamma}_n$ and $\hat{R}_n$ may not then be nonnegative definite. We shall therefore use the definitions (2.4.4) and (2.4.5) of $\hat{\gamma}(h)$ and $\hat{\rho}(h)$.

Remark 1. The matrices $\hat{\Gamma}_k$ and $\hat{R}_k$ are in fact nonsingular if there is at least one nonzero Y_i, or equivalently if $\hat{\gamma}(0) > 0$. To establish this result, suppose that $\hat{\gamma}(0) > 0$ and $\hat{\Gamma}_k$ is singular. Then there is equality in (2.4.7) for some nonzero vector $\mathbf{a}$, implying that $\mathbf{a}'T = 0$ and hence that the rank of T is less than k. Let Y_i be the first nonzero value of $Y_1, Y_2, \ldots, Y_k$, and consider the $k \times k$ submatrix of T consisting of columns $(i+1)$ through $(i+k)$. Since this matrix is lower right triangular with each diagonal element equal to Y_i, its determinant has absolute value $|Y_i|^k \neq 0$. Consequently, the submatrix is nonsingular, and T must have rank k, a contradiction. □

Without further information beyond the observed data $X_1, \ldots, X_n$, it is impossible to give reasonable estimates of $\gamma(h)$ and $\rho(h)$ for $h \geq n$. Even for h slightly smaller than n, the estimates $\hat{\gamma}(h)$ and $\hat{\rho}(h)$ are unreliable, since there are so few pairs (X_{t+h}, X_t) available (only one if $h = n-1$). A useful guide is provided by Box and Jenkins (1976), p. 33, who suggest that n should be at least about 50 and $h \leq n/4$.

The sample ACF plays an important role in the selection of suitable models for the data. We have already seen in Example 1.4.6 and Section 1.6 how the sample ACF can be used to test for iid noise. For systematic inference concerning $\rho(h)$, we need the sampling distribution of the estimator $\hat{\rho}(h)$. Although the distribution of $\hat{\rho}(h)$ is intractable for samples from even the simplest time series models, it can usually be well approximated by a normal distribution for large sample sizes. For linear models and in particular for ARMA models (see Theorem 7.2.2 of TSTM for exact conditions) $\hat{\boldsymbol{\rho}}_k = (\hat{\rho}(1), \ldots, \hat{\rho}(k))'$ is approximately distributed for large n as

$N(\boldsymbol{\rho}_k, n^{-1}W)$, i.e.,

$$\hat{\boldsymbol{\rho}} \approx N\big(\boldsymbol{\rho}, n^{-1}W\big), \tag{2.4.9}$$

where $\boldsymbol{\rho} = (\rho(1), \ldots, \rho(k))'$, and W is the covariance matrix whose (i, j) element is given by **Bartlett's formula**

$$w_{ij} = \sum_{k=-\infty}^{\infty} \big\{\rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) + 2\rho(i)\rho(j)\rho^2(k) - 2\rho(i)\rho(k)\rho(k+j) - 2\rho(j)\rho(k)\rho(k+i)\big\}.$$

Simple algebra shows that

$$w_{ij} = \sum_{k=1}^{\infty} \{\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\} \times \{\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)\}, \tag{2.4.10}$$

which is a more convenient form of w_{ij} for computational purposes.

Example 2.4.2 iid Noise

If $\{X_t\} \sim \text{IID}\big(0, \sigma^2\big)$, then $\rho(h) = 0$ for $|h| > 0$, so from (2.4.10) we obtain

$$w_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

For large n, therefore, $\hat{\rho}(1), \ldots, \hat{\rho}(h)$ are approximately independent and identically distributed normal random variables with mean 0 and variance n^{-1}. This result is the basis for the test that data are generated from iid noise using the sample ACF described in Section 1.6. (See also Example 1.4.6.) □

Example 2.4.3 An MA(1) process

If $\{X_t\}$ is the MA(1) process of Example 1.4.4, i.e., if

$$X_t = Z_t + \theta Z_{t-1}, \quad t = 0, \pm 1, \ldots,$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, then from (2.4.10)

$$w_{ii} = \begin{cases} 1 - 3\rho^2(1) + 4\rho^4(1), & \text{if } i = 1, \\ 1 + 2\rho^2(1), & \text{if } i > 1, \end{cases}$$

is the approximate variance of $n^{-1/2}(\hat{\rho}(i) - \rho(i))$ for large n. In Figure 2.1 we have plotted the sample autocorrelation function $\hat{\rho}(k), k = 0, \ldots, 40$, for 200 observations from the MA(1) model

$$X_t = Z_t - .8Z_{t-1}, \tag{2.4.11}$$

where $\{Z_t\}$ is a sequence of iid N(0, 1) random variables. Here $\rho(1) = -.8/1.64 = -.4878$ and $\rho(h) = 0$ for $h > 1$. The lag-one sample ACF is found to be $\hat{\rho}(1) =$

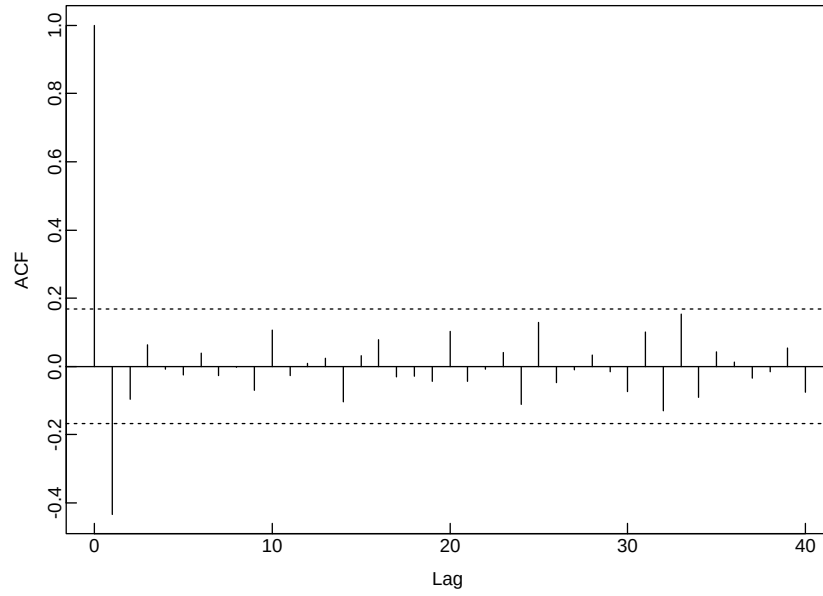

Figure 2-1
The sample autocorrelation function of $n = 200$ observations of the MA(1) process in Example 2.4.3, showing the bounds $\pm 1.96 n^{-1/2}(1 + 2\hat{\rho}^2(1))^{1/2}$.

$-.4333 = -6.128n^{-1/2}$, which would cause us (in the absence of our prior knowledge of $\{X_t\}$) to reject the hypothesis that the data are a sample from an iid noise sequence. The fact that $|\hat{\rho}(h)| \leq 1.96n^{-1/2}$ for $h = 2, \ldots, 40$ strongly suggests that the data are from a model in which observations are uncorrelated past lag 1. In Figure 2.1 we have plotted the bounds $\pm 1.96n^{-1/2}(1+2\rho^2(1))^{1/2}$, indicating the compatibility of the data with the model (2.4.11). Since, however, $\rho(1)$ is not normally known in advance, the autocorrelations $\hat{\rho}(2), \ldots, \hat{\rho}(40)$ would in practice have been compared with the more stringent bounds $\pm 1.96n^{-1/2}$ or with the bounds $\pm 1.96n^{-1/2}(1+2\hat{\rho}^2(1))^{1/2}$ in order to check the hypothesis that the data are generated by a moving-average process of order 1. Finally, it is worth noting that the lag-one correlation $-.4878$ is well inside the 95% confidence bounds for $\rho(1)$ given by $\hat{\rho}(1) \pm 1.96n^{-1/2}(1 - 3\hat{\rho}^2(1) + 4\hat{\rho}^4(1))^{1/2} = -.4333 \pm .1053$. This further supports the compatibility of the data with the model $X_t = Z_t - 0.8Z_{t-1}$. □

Example 2.4.4 An AR(1) process

For the AR(1) process of Example 2.2.1,

$$X_t = \phi X_{t-1} + Z_t,$$

where $\{Z_t\}$ is iid noise and $|\phi| < 1$, we have, from (2.4.10) with $\rho(h) = \phi^{|h|}$,

$$\begin{aligned} w_{ii} &= \sum_{k=1}^{i} \phi^{2i}\left(\phi^{-k} - \phi^{k}\right)^2 + \sum_{k=i+1}^{\infty} \phi^{2k}\left(\phi^{-i} - \phi^{i}\right)^2 \\ &= \left(1 - \phi^{2i}\right)\left(1 + \phi^2\right)\left(1 - \phi^2\right)^{-1} - 2i\phi^{2i}, \end{aligned} \tag{2.4.12}$$

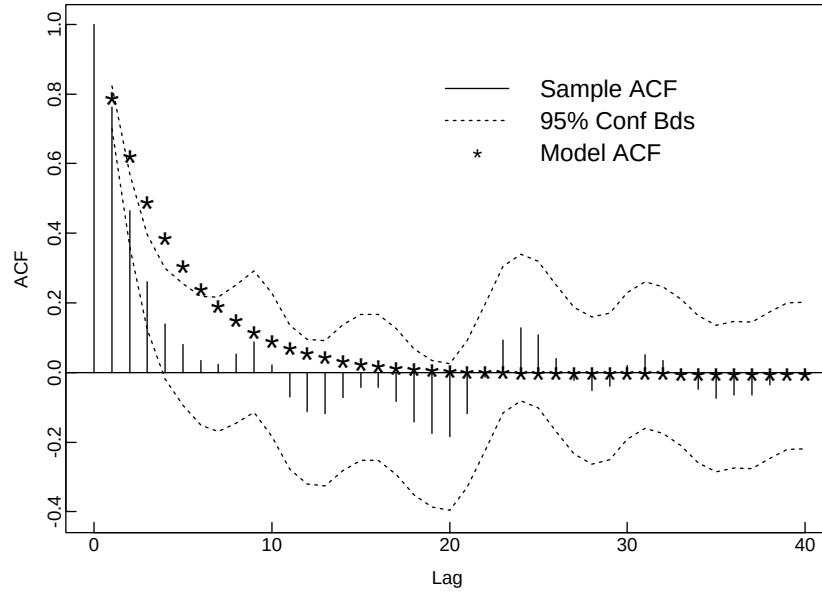

Figure 2-2
The sample autocorrelation function of the Lake Huron residuals of Figure 1.10 showing the bounds $\hat{\rho}(i) \pm 1.96 n^{-1/2} w_{ii}^{1/2}$ and the model ACF $\rho(i) = (.791)^i$.

$i = 1, 2, \ldots$. In Figure 2.2 we have plotted the sample ACF of the Lake Huron residuals $y_1, \ldots, y_{98}$ from Figure 1.10 together with 95% confidence bounds for $\rho(i)$, $i = 1, \ldots, 40$, assuming that data are generated from the AR(1) model

$$Y_t = .791 Y_{t-1} + Z_t \tag{2.4.13}$$

(see equation (1.4.3)). The confidence bounds are computed from $\hat{\rho}(i) \pm 1.96 n^{-1/2} w_{ii}^{1/2}$, where w_{ii} is given in (2.4.12) with $\phi = .791$. The model ACF, $\rho(i) = (.791)^i$, is also plotted in Figure 2.2. Notice that the model ACF lies just outside the confidence bounds at lags 2–6. This suggests some incompatibility of the data with the model (2.4.13). A much better fit to the residuals is provided by the second-order autoregression defined by (1.4.4). □

2.5 Forecasting Stationary Time Series

We now consider the problem of predicting the values $X_{n+h}, h > 0$, of a stationary time series with known mean μ and autocovariance function γ in terms of the values $\{X_n, \ldots, X_1\}$, up to time n. Our goal is to find the *linear combination* of $1, X_n, X_{n-1}, \ldots, X_1$, that forecasts X_{n+h} with minimum mean squared error. The best linear predictor in terms of $1, X_n, \ldots, X_1$ will be denoted by $P_n X_{n+h}$ and clearly has the form

$$P_n X_{n+h} = a_0 + a_1 X_n + \cdots + a_n X_1. \tag{2.5.1}$$

It remains only to determine the coefficients $a_0, a_1, \ldots, a_n$, by finding the values that minimize

$$S(a_0, \ldots, a_n) = E(X_{n+h} - a_0 - a_1 X_n - \cdots - a_n X_1)^2. \tag{2.5.2}$$

(We already know from Problem 1.1 that $P_0 Y = E(Y)$.) Since S is a quadratic function of $a_0, \ldots, a_n$ and is bounded below by zero, it is clear that there is at least one value of $(a_0, \ldots, a_n)$ that minimizes S and that the minimum $(a_0, \ldots, a_n)$ satisfies the equations

$$\frac{\partial S(a_0, \ldots, a_n)}{\partial a_j} = 0, \quad j = 0, \ldots, n. \tag{2.5.3}$$

Evaluation of the derivatives in equations (2.5.3) gives the equivalent equations

$$E\left[X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i}\right] = 0, \tag{2.5.4}$$

$$E\left[(X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n+1-i}) X_{n+1-j}\right] = 0, \quad j = 1, \ldots, n. \tag{2.5.5}$$

These equations can be written more neatly in vector notation as

$$a_0 = \mu\left(1 - \sum_{i=1}^{n} a_i\right) \tag{2.5.6}$$

and

$$\Gamma_n \mathbf{a}_n = \boldsymbol{\gamma}_n(h), \tag{2.5.7}$$

where

$$\mathbf{a}_n = (a_1, \ldots, a_n)', \qquad \Gamma_n = [\gamma(i-j)]_{i,j=1}^{n},$$

and

$$\boldsymbol{\gamma}_n(h) = (\gamma(h), \gamma(h+1), \ldots, \gamma(h+n-1))'.$$

Hence,

$$P_n X_{n+h} = \mu + \sum_{i=1}^{n} a_i (X_{n+1-i} - \mu), \tag{2.5.8}$$

where $\mathbf{a}_n$ satisfies (2.5.7). From (2.5.8) the expected value of the prediction error $X_{n+h} - P_n X_{n+h}$ is zero, and the mean square prediction error is therefore

$$\begin{aligned} E(X_{n+h} - P_n X_{n+h})^2 &= \gamma(0) - 2\sum_{i=1}^{n} a_i \gamma(h+i-1) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \gamma(i-j) a_j \\ &= \gamma(0) - \mathbf{a}_n' \boldsymbol{\gamma}_n(h), \end{aligned} \tag{2.5.9}$$

where the last line follows from (2.5.7).

Remark 1. To show that equations (2.5.4) and (2.5.5) determine $P_n X_{n+h}$ uniquely, let $\{a_j^{(1)}, j = 0, \ldots, n\}$ and $\{a_j^{(2)}, j = 0, \ldots, n\}$ be two solutions and let Z be the difference between the corresponding predictors, i.e.,

$$Z = a_0^{(1)} - a_0^{(2)} + \sum_{j=1}^{n} \left(a_j^{(1)} - a_j^{(2)}\right) X_{n+1-j}.$$

Then

$$Z^2 = Z\left(a_0^{(1)} - a_0^{(2)} + \sum_{j=1}^{n} \left(a_j^{(1)} - a_j^{(2)}\right) X_{n+1-j}\right).$$

But from (2.5.4) and (2.5.5) we have $EZ = 0$ and $E(ZX_{n+1-j}) = 0$ for $j = 1, \ldots, n$. Consequently, $E(Z^2) = 0$ and hence $Z = 0$. □

Properties of $P_n X_{n+h}$:

1. $P_n X_{n+h} = \mu + \sum_{i=1}^{n} a_i (X_{n+1-i} - \mu)$, where $\mathbf{a}_n = (a_1, \ldots, a_n)'$ satisfies (2.5.7).
2. $E(X_{n+h} - P_n X_{n+h})^2 = \gamma(0) - \mathbf{a}_n' \boldsymbol{\gamma}_n(h)$, where $\boldsymbol{\gamma}_n(h) = (\gamma(h), \ldots, \gamma(h+n-1))'$.
3. $E(X_{n+h} - P_n X_{n+h}) = 0$.
4. $E[(X_{n+h} - P_n X_{n+h}) X_j] = 0$, $j = 1, \ldots, n$.

Remark 2. Notice that properties 3 and 4 are exactly equivalent to (2.5.4) and (2.5.5). They can be written more succinctly in the form

$$E[(\text{Error}) \times (\text{PredictorVariable})] = 0. \tag{2.5.10}$$

Equations (2.5.10), one for each predictor variable, therefore uniquely determine $P_n X_{n+h}$. □

Example 2.5.1 One-step prediction of an AR(1) series

Consider now the stationary time series defined in Example 2.2.1 by the equations

$$X_t = \phi X_{t-1} + Z_t, \quad t = 0, \pm 1, \ldots,$$

where $|\phi| < 1$ and $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. From (2.5.7) and (2.5.8), the best linear predictor of X_{n+1} in terms of $\{1, X_n, \ldots, X_1\}$ is (for $n \geq 1$)

$$P_n X_{n+1} = \mathbf{a}_n' \mathbf{X}_n,$$

where $\mathbf{X}_n = (X_n, \ldots, X_1)'$ and

$$\begin{bmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{n-1} \\ \phi & 1 & \phi & \cdots & \phi^{n-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \phi \\ \phi^2 \\ \vdots \\ \phi^n \end{bmatrix}. \tag{2.5.11}$$

A solution of (2.5.11) is clearly

$$\mathbf{a}_n = (\phi, 0, \ldots, 0)',$$

and hence the best linear predictor of X_{n+1} in terms of $\{X_1, \ldots, X_n\}$ is

$$P_n X_{n+1} = \mathbf{a}_n' \mathbf{X}_n = \phi X_n,$$

with mean squared error

$$E(X_{n+1} - P_n X_{n+1})^2 = \gamma(0) - \mathbf{a}_n' \boldsymbol{\gamma}_n(1) = \frac{\sigma^2}{1-\phi^2} - \phi\gamma(1) = \sigma^2.$$

A simpler approach to this problem is to guess, by inspection of the equation defining X_{n+1}, that the best predictor is ϕX_n. Then to verify this conjecture, it suffices to check (2.5.10) for each of the predictor variables $1, X_n, \ldots, X_1$. The prediction error of the predictor ϕX_n is clearly $X_{n+1} - \phi X_n = Z_{n+1}$. But $E(Z_{n+1}Y) = 0$ for $Y = 1$ and for $Y = X_j$, $j = 1, \ldots, n$. Hence, by (2.5.10), ϕX_n is the required best linear predictor in terms of $1, X_1, \ldots, X_n$. □

Prediction of Second-Order Random Variables

Suppose now that Y and $W_n, \ldots, W_1$ are *any* random variables with finite second moments and that the means $\mu = EY$, $\mu_i = EW_i$ and covariances $\mathrm{Cov}(Y, Y)$, $\mathrm{Cov}(Y, W_i)$, and $\mathrm{Cov}(W_i, W_j)$ are all known. It is convenient to introduce the random vector $\mathbf{W} = (W_n, \ldots, W_1)'$, the corresponding vector of means $\boldsymbol{\mu}_W = (\mu_n, \ldots, \mu_1)'$, the vector of covariances

$$\boldsymbol{\gamma} = \mathrm{Cov}(Y, \mathbf{W}) = (\mathrm{Cov}(Y, W_n), \mathrm{Cov}(Y, W_{n-1}), \ldots, \mathrm{Cov}(Y, W_1))',$$

and the covariance matrix

$$\Gamma = \mathrm{Cov}(\mathbf{W}, \mathbf{W}) = \left[\mathrm{Cov}(W_{n+1-i}, W_{n+1-j})\right]_{i,j=1}^{n}.$$

Then by the same arguments used in the calculation of $P_n X_{n+h}$, the best linear predictor of Y in terms of $\{1, W_n, \ldots, W_1\}$ is found to be

$$P(Y|\mathbf{W}) = \mu_Y + \mathbf{a}'(\mathbf{W} - \boldsymbol{\mu}_W), \tag{2.5.12}$$

where $\mathbf{a} = (a_1, \ldots, a_n)'$ is any solution of

$$\Gamma \mathbf{a} = \boldsymbol{\gamma}. \tag{2.5.13}$$

The mean squared error of the predictor is

$$E\left[(Y - P(Y|\mathbf{W}))^2\right] = \mathrm{Var}(Y) - \mathbf{a}'\boldsymbol{\gamma}. \tag{2.5.14}$$

Example 2.5.2 Estimation of a missing value

Consider again the stationary series defined in Example 2.2.1 by the equations

$$X_t = \phi X_{t-1} + Z_t, \quad t = 0, \pm 1, \ldots,$$

where $|\phi| < 1$ and $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$. Suppose that we observe the series at times 1 and 3 and wish to use these observations to find the linear combination of 1, X_1, and X_3 that estimates X_2 with minimum mean squared error. The solution to this problem can be obtained directly from (2.5.12) and (2.5.13) by setting $Y = X_2$ and $\mathbf{W} = (X_1, X_3)'$. This gives the equations

$$\begin{bmatrix} 1 & \phi^2 \\ \phi^2 & 1 \end{bmatrix} \mathbf{a} = \begin{bmatrix} \phi \\ \phi \end{bmatrix},$$

with solution

$$\mathbf{a} = \frac{1}{1+\phi^2} \begin{bmatrix} \phi \\ \phi \end{bmatrix}.$$

The best estimator of X_2 is thus

$$P(X_2|\mathbf{W}) = \frac{\phi}{1+\phi^2}(X_1 + X_3),$$

with mean squared error

$$E[(X_2 - P(X_2|\mathbf{W}))^2] = \frac{\sigma^2}{1-\phi^2} - \mathbf{a}' \begin{bmatrix} \frac{\phi\sigma^2}{1-\phi^2} \\ \frac{\phi\sigma^2}{1-\phi^2} \end{bmatrix} = \frac{\sigma^2}{1+\phi^2}. \qquad \square$$

The Prediction Operator $P(\cdot|\mathbf{W})$

For any given $\mathbf{W} = (W_n, \ldots, W_1)'$ and Y with finite second moments, we have seen how to compute the best linear predictor $P(Y|\mathbf{W})$ of Y in terms of 1, $W_n, \ldots, W_1$ from (2.5.12) and (2.5.13). The function $P(\cdot|\mathbf{W})$, which converts Y into $P(Y|\mathbf{W})$, is called a **prediction operator**. (The operator P_n defined by equations (2.5.7) and (2.5.8) is an example with $\mathbf{W} = (X_n, X_{n-1}, \ldots, X_1)'$.) Prediction operators have a number of useful properties that can sometimes be used to simplify the calculation of best linear predictors. We list some of these below.

Properties of the Prediction Operator $P(\cdot|\mathbf{W})$:

Suppose that $EU^2 < \infty$, $EV^2 < \infty$, $\Gamma = \text{cov}(\mathbf{W}, \mathbf{W})$, and $\beta, \alpha_1, \ldots, \alpha_n$ are constants.

1. $P(U|\mathbf{W}) = EU + \mathbf{a}'(\mathbf{W} - E\mathbf{W})$, where $\Gamma\mathbf{a} = \text{cov}(U, \mathbf{W})$.
2. $E[(U - P(U|\mathbf{W}))\mathbf{W}] = \mathbf{0}$ and $E[U - P(U|\mathbf{W})] = 0$.
3. $E[(U - P(U|\mathbf{W}))^2] = \text{var}(U) - \mathbf{a}'\text{cov}(U, \mathbf{W})$.
4. $P(\alpha_1 U + \alpha_2 V + \beta|\mathbf{W}) = \alpha_1 P(U|\mathbf{W}) + \alpha_2 P(V|\mathbf{W}) + \beta$.
5. $P\left(\sum_{i=1}^n \alpha_i W_i + \beta|\mathbf{W}\right) = \sum_{i=1}^n \alpha_i W_i + \beta$.
6. $P(U|\mathbf{W}) = EU$ if $\text{cov}(U, \mathbf{W}) = \mathbf{0}$.
7. $P(U|\mathbf{W}) = P(P(U|\mathbf{W}, \mathbf{V})|\mathbf{W})$ if $\mathbf{V}$ is a random vector such that the components of $E(\mathbf{V}\mathbf{V}')$ are all finite.

Example 2.5.3 One-step prediction of an AR(p) series

Suppose now that $\{X_t\}$ is a stationary time series satisfying the equations

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t, \quad t = 0, \pm 1, \ldots,$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and Z_t is uncorrelated with X_s for each $s < t$. Then if $n > p$, we can apply the prediction operator P_n to each side of the defining equations, using properties (4), (5), and (6) to get

$$P_n X_{n+1} = \phi_1 X_n + \cdots + \phi_p X_{n+1-p}. \qquad \square$$

Example 2.5.4 An AR(1) series with nonzero mean

The time series $\{Y_t\}$ is said to be an AR(1) process with mean μ if $\{X_t = Y_t - \mu\}$ is a zero-mean AR(1) process. Defining $\{X_t\}$ as in Example 2.5.1 and letting $Y_t = X_t + \mu$, we see that Y_t satisfies the equation

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + Z_t. \tag{2.5.15}$$

If $P_n Y_{n+h}$ is the best linear predictor of Y_{n+h} in terms of $\{1, Y_n, \ldots, Y_1\}$, then application of P_n to (2.5.15) with $t = n + 1, n + 2, \ldots$ gives the recursions

$$P_n Y_{n+h} - \mu = \phi(P_n Y_{n+h-1} - \mu), \quad h = 1, 2, \ldots.$$

Noting that $P_n Y_n = Y_n$, we can solve these equations recursively for $P_n Y_{n+h}$, $h = 1, 2, \ldots$, to obtain

$$P_n Y_{n+h} = \mu + \phi^h (Y_n - \mu). \tag{2.5.16}$$

The corresponding mean squared error is (from (2.5.14))

$$E(Y_{n+h} - P_n Y_{n+h})^2 = \gamma(0)[1 - \mathbf{a}_n' \boldsymbol{\rho}_n(h)]. \tag{2.5.17}$$

From Example 2.2.1 we know that $\gamma(0) = \sigma^2/(1-\phi^2)$ and $\rho(h) = \phi^h, h \geq 0$. Hence, substituting $\mathbf{a}_n = (\phi^h, 0, \ldots, 0)'$ (from (2.5.16)) into (2.5.17) gives

$$E(Y_{n+h} - P_n Y_{n+h})^2 = \sigma^2(1-\phi^{2h})/(1-\phi^2). \tag{2.5.18}$$

□

Remark 3. In general, if $\{Y_t\}$ is a stationary time series with mean μ and if $\{X_t\}$ is the zero-mean series defined by $X_t = Y_t - \mu$, then since the collection of all linear combinations of $1, Y_n, \ldots, Y_1$ is the same as the collection of all linear combinations of $1, X_n, \ldots, X_1$, the linear predictor of any random variable W in terms of $1, Y_n, \ldots, Y_1$ is the same as the linear predictor in terms of $1, X_n, \ldots, X_1$. Denoting this predictor by $P_n W$ and applying P_n to the equation $Y_{n+h} = X_{n+h} + \mu$ gives

$$P_n Y_{n+h} = \mu + P_n X_{n+h}. \tag{2.5.19}$$

Thus the best linear predictor of Y_{n+h} can be determined by finding the best linear predictor of X_{n+h} and then adding μ. Note from (2.5.8) that since $E(X_t) = 0$, $P_n X_{n+h}$ is the same as the best linear predictor of X_{n+h} in terms of $X_n, \ldots, X_1$ only. □

2.5.1 The Durbin–Levinson Algorithm

In view of Remark 3 above, we can restrict attention from now on to zero-mean stationary time series, making the necessary adjustments for the mean if we wish to predict a stationary series with nonzero mean. If $\{X_t\}$ is a zero-mean stationary series with autocovariance function $\gamma(\cdot)$, then in principle the equations (2.5.12) and (2.5.13) completely solve the problem of determining the best linear predictor $P_n X_{n+h}$ of X_{n+h} in terms of $\{X_n, \ldots, X_1\}$. However, the direct approach requires the determination of a solution of a system of n linear equations, which for large n may be difficult and time-consuming. In cases where the process is defined by a system of linear equations (as in Examples 2.5.2 and 2.5.3) we have seen how the linearity of P_n can be used to great advantage. For more general stationary processes it would be helpful if the one-step predictor $P_n X_{n+1}$ based on n previous observations could be used to simplify the calculation of $P_{n+1} X_{n+2}$, the one-step predictor based on $n+1$ previous observations. Prediction algorithms that utilize this idea are said to be **recursive**. Two important examples are the Durbin–Levinson algorithm, discussed in this section, and the innovations algorithm, discussed in Section 2.5.2 below.

We know from (2.5.12) and (2.5.13) that if the matrix Γ_n is nonsingular, then

$$P_n X_{n+1} = \boldsymbol{\phi}_n' \mathbf{X}_n = \phi_{n1} X_n + \cdots + \phi_{nn} X_1,$$

where

$$\boldsymbol{\phi}_n = \Gamma_n^{-1} \boldsymbol{\gamma}_n,$$

$\gamma_n = (\gamma(1), \ldots, \gamma(n))'$, and the corresponding mean squared error is

$$v_n := E(X_{n+1} - P_n X_{n+1})^2 = \gamma(0) - \phi_n' \gamma_n.$$

A useful sufficient condition for nonsingularity of *all* the autocovariance matrices $\Gamma_1, \Gamma_2, \ldots$ is $\gamma(0) > 0$ and $\gamma(h) \to 0$ as $h \to \infty$. (For a proof of this result see TSTM, Proposition 5.1.1.)

The Durbin–Levinson Algorithm:

The coefficients $\phi_{n1}, \ldots, \phi_{nn}$ can be computed recursively from the equations

$$\phi_{nn} = \left[\gamma(n) - \sum_{j=1}^{n-1} \phi_{n-1,j}\gamma(n-j)\right] v_{n-1}^{-1}, \tag{2.5.20}$$

$$\begin{bmatrix} \phi_{n1} \\ \vdots \\ \phi_{n,n-1} \end{bmatrix} = \begin{bmatrix} \phi_{n-1,1} \\ \vdots \\ \phi_{n-1,n-1} \end{bmatrix} - \phi_{nn} \begin{bmatrix} \phi_{n-1,n-1} \\ \vdots \\ \phi_{n-1,1} \end{bmatrix} \tag{2.5.21}$$

and

$$v_n = v_{n-1}\left[1 - \phi_{nn}^2\right], \tag{2.5.22}$$

where $\phi_{11} = \gamma(1)/\gamma(0)$ and $v_0 = \gamma(0)$.

Proof The definition of ϕ_{11} ensures that the equation

$$R_n \phi_n = \rho_n \tag{2.5.23}$$

(where $\rho_n = (\rho(1), \ldots, \rho(n))'$) is satisfied for $n = 1$. The first step in the proof is to show that ϕ_n, defined recursively by (2.5.20) and (2.5.21), satisfies (2.5.23) for all n. Suppose this is true for $n = k$. Then, partitioning R_{k+1} and defining

$$\rho_k^{(r)} := (\rho(k), \rho(k-1), \ldots, \rho(1))'$$

and

$$\phi_k^{(r)} := (\phi_{kk}, \phi_{k,k-1}, \ldots, \phi_{k1})',$$

we see that the recursions imply

$$\begin{aligned} R_{k+1}\phi_{k+1} &= \begin{bmatrix} R_k & \rho_k^{(r)} \\ \rho_k^{(r)\prime} & 1 \end{bmatrix} \begin{bmatrix} \phi_k - \phi_{k+1,k+1}\phi_k^{(r)} \\ \phi_{k+1,k+1} \end{bmatrix} \\ &= \begin{bmatrix} \rho_k - \phi_{k+1,k+1}\rho_k^{(r)} + \phi_{k+1,k+1}\rho_k^{(r)} \\ \rho_k^{(r)\prime}\phi_k - \phi_{k+1,k+1}\rho_k^{(r)\prime}\phi_k^{(r)} + \phi_{k+1,k+1} \end{bmatrix} \\ &= \rho_{k+1}, \end{aligned}$$

as required. Here we have used the fact that if $R_k\phi_k = \rho_k$, then $R_k\phi_k^{(r)} = \rho_k^{(r)}$. This is easily checked by writing out the component equations in reverse order. Since (2.5.23) is satisfied for $n = 1$, it follows by induction that the coefficient vectors ϕ_n defined recursively by (2.5.20) and (2.5.21) satisfy (2.5.23) for all n.

It remains only to establish that the mean squared errors

$$v_n := E(X_{n+1} - \phi_n' \mathbf{X}_n)^2$$

satisfy $v_0 = \gamma(0)$ and (2.5.22). The fact that $v_0 = \gamma(0)$ is an immediate consequence of the definition $P_0 X_1 := E(X_1) = 0$. Since we have shown that $\phi_n' \mathbf{X}_n$ is the best linear predictor of X_{n+1}, we can write, from (2.5.9) and (2.5.21),

$$v_n = \gamma(0) - \phi_n' \gamma_n = \gamma(0) - \phi_{n-1}' \gamma_{n-1} + \phi_{nn} \phi_{n-1}^{(r)\prime} \gamma_{n-1} - \phi_{nn} \gamma(n).$$

Applying (2.5.9) again gives

$$v_n = v_{n-1} + \phi_{nn} \left(\phi_{n-1}^{(r)\prime} \gamma_{n-1} - \gamma(n) \right),$$

and hence, by (2.5.20),

$$v_n = v_{n-1} - \phi_{nn}^2 (\gamma(0) - \phi_{n-1}' \gamma_{n-1}) = v_{n-1} \left(1 - \phi_{nn}^2\right). \qquad \blacksquare$$

Remark 4. Under the conditions of the proposition, the function defined by $\alpha(0) = 1$ and $\alpha(n) = \phi_{nn}$, $n = 1, 2, \ldots$, is known as the **partial autocorrelation function** (PACF) of $\{X_t\}$. It will be discussed further in Section 3.2. Of particular interest is equation (2.5.22), which shows the relation between $\alpha(n)$ and the reduction in the one-step mean squared error as the number of predictors is increased from $n - 1$ to n. $\square$

2.5.2 The Innovations Algorithm

The recursive algorithm to be discussed in this section is applicable to *all* series with finite second moments, regardless of whether they are stationary or not. Its application, however, can be simplified in certain special cases.

Suppose then that $\{X_t\}$ is a zero-mean series with $E|X_t|^2 < \infty$ for each t and

$$E(X_i X_j) = \kappa(i, j). \tag{2.5.24}$$

It will be convenient to introduce the following notation for the best one-step predictors and their mean squared errors:

$$\hat{X}_n = \begin{cases} 0, & \text{if } n = 1, \\ P_{n-1} X_n, & \text{if } n = 2, 3, \ldots, \end{cases}$$

and

$$v_n = E(X_{n+1} - P_n X_{n+1})^2.$$

We shall also introduce the **innovations**, or one-step prediction errors,

$$U_n = X_n - \hat{X}_n.$$

In terms of the vectors $\mathbf{U}_n = (U_1, \ldots, U_n)'$ and $\mathbf{X}_n = (X_1, \ldots, X_n)'$ the last equations can be written as

$$\mathbf{U}_n = A_n \mathbf{X}_n, \tag{2.5.25}$$

where A_n has the form

$$A_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{11} & 1 & 0 & \cdots & 0 \\ a_{22} & a_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ a_{n-1,n-1} & a_{n-1,n-2} & a_{n-1,n-3} & \cdots & 1 \end{bmatrix}.$$

(If $\{X_t\}$ is stationary, then $a_{ij} = -a_j$ with a_j as in (2.5.7) with $h = 1$.) This implies that A_n is nonsingular, with inverse C_n of the form

$$C_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \theta_{11} & 1 & 0 & \cdots & 0 \\ \theta_{22} & \theta_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 1 \end{bmatrix}.$$

The vector of one-step predictors $\hat{\mathbf{X}}_n := (X_1, P_1 X_2, \ldots, P_{n-1} X_n)'$ can therefore be expressed as

$$\hat{\mathbf{X}}_n = \mathbf{X}_n - \mathbf{U}_n = C_n \mathbf{U}_n - \mathbf{U}_n = \boldsymbol{\Theta}_n \left(\mathbf{X}_n - \hat{\mathbf{X}}_n \right), \tag{2.5.26}$$

where

$$\boldsymbol{\Theta}_n = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \theta_{11} & 0 & 0 & \cdots & 0 \\ \theta_{22} & \theta_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \theta_{n-1,n-1} & \theta_{n-1,n-2} & \theta_{n-1,n-3} & \cdots & 0 \end{bmatrix}$$

and $\mathbf{X}_n$ itself satisfies

$$\mathbf{X}_n = C_n \left(\mathbf{X}_n - \hat{\mathbf{X}}_n \right). \tag{2.5.27}$$

Equation (2.5.26) can be rewritten as

$$\hat{X}_{n+1} = \begin{cases} 0, & \text{if } n = 0, \\ \displaystyle\sum_{j=1}^{n} \theta_{nj} \left(X_{n+1-j} - \hat{X}_{n+1-j} \right), & \text{if } n = 1, 2, \ldots, \end{cases} \tag{2.5.28}$$

from which the one-step predictors $\hat{X}_1$, $\hat{X}_2$, ... can be computed recursively once the coefficients θ_{ij} have been determined. The following algorithm generates these

coefficients and the mean squared errors $v_i = E\left(X_{i+1} - \hat{X}_{i+1}\right)^2$, starting from the covariances $\kappa(i, j)$.

The Innovations Algorithm:

The coefficients $\theta_{n1}, \ldots, \theta_{nn}$ can be computed recursively from the equations

$$v_0 = \kappa(1, 1),$$

$$\theta_{n,n-k} = v_k^{-1}\left(\kappa(n+1, k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}v_j\right), \quad 0 \le k < n,$$

and

$$v_n = \kappa(n+1, n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 v_j.$$

(It is a trivial matter to solve first for v_0, then successively for θ_{11}, v_1; θ_{22}, θ_{21}, v_2; θ_{33}, θ_{32}, θ_{31}, v_3; ....)

Proof See TSTM, Proposition 5.2.2. ■

Remark 5. While the Durbin–Levinson recursion gives the coefficients of $X_n, \ldots, X_1$ in the representation $\hat{X}_{n+1} = \sum_{j=1}^{n} \phi_{nj} X_{n+1-j}$, the innovations algorithm gives the coefficients of $\left(X_n - \hat{X}_n\right), \ldots, \left(X_1 - \hat{X}_1\right)$, in the expansion $\hat{X}_{n+1} = \sum_{j=1}^{n} \theta_{nj}\left(X_{n+1-j} - \hat{X}_{n+1-j}\right)$. The latter expansion has a number of advantages deriving from the fact that the innovations are uncorrelated (see Problem 2.20). It can also be greatly simplified in the case of ARMA(p, q) series, as we shall see in Section 3.3. An immediate consequence of (2.5.28) is the innovations representation of X_{n+1} itself. Thus (defining $\theta_{n0} := 1$),

$$X_{n+1} = X_{n+1} - \hat{X}_{n+1} + \hat{X}_{n+1} = \sum_{j=0}^{n} \theta_{nj}\left(X_{n+1-j} - \hat{X}_{n+1-j}\right), \quad n = 0, 1, 2, \ldots. \quad \square$$

Example 2.5.5 Recursive prediction of an MA(1)

If $\{X_t\}$ is the time series defined by

$$X_t = Z_t + \theta Z_{t-1}, \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right),$$

then $\kappa(i, j) = 0$ for $|i - j| > 1$, $\kappa(i, i) = \sigma^2\left(1 + \theta^2\right)$, and $\kappa(i, i+1) = \theta\sigma^2$. Application of the innovations algorithm leads at once to the recursions

$$\theta_{nj} = 0, 2 \le j \le n,$$

$$\theta_{n1} = v_{n-1}^{-1}\theta\sigma^2,$$

$$v_0 = (1 + \theta^2)\sigma^2,$$

and

$$v_n = \left[1 + \theta^2 - v_{n-1}^{-1}\theta^2\sigma^2\right]\sigma^2.$$

For the particular case

$$X_t = Z_t - 0.9Z_{t-1}, \quad \{Z_t\} \sim \text{WN}(0, 1),$$

the mean squared errors v_n of $\hat{X}_{n+1}$ and coefficients θ_{nj}, $1 \le j \le n$, in the innovations representation

$$\hat{X}_{n+1} = \sum_{j=1}^{n} \theta_{nj}\left(X_{n+1-j} - \hat{X}_{n+1-j}\right) = \theta_{n1}\left(X_n - \hat{X}_n\right)$$

are found from the recursions to be as follows:

$$\begin{aligned}
&v_0 = 1.8100,\\
&\theta_{11} = -.4972, \quad v_1 = 1.3625,\\
&\theta_{21} = -.6606, \quad \theta_{22} = 0, \quad v_2 = 1.2155,\\
&\theta_{31} = -.7404, \quad \theta_{32} = 0, \quad \theta_{33} = 0, \quad v_3 = 1.1436,\\
&\theta_{41} = -.7870, \quad \theta_{42} = 0, \quad \theta_{43} = 0, \quad \theta_{44} = 0, \quad v_4 = 1.1017.
\end{aligned}$$

If we apply the Durbin–Levinson algorithm to the same problem, we find that the mean squared errors v_n of $\hat{X}_{n+1}$ and coefficients ϕ_{nj}, $1 \le j \le n$, in the representation

$$\hat{X}_{n+1} = \sum_{j=1}^{n} \phi_{nj} X_{n+1-j}$$

are as follows:

$$\begin{aligned}
&v_0 = 1.8100,\\
&\phi_{11} = -.4972, \quad v_1 = 1.3625,\\
&\phi_{21} = -.6606, \quad \phi_{22} = -.3285, \quad v_2 = 1.2155,\\
&\phi_{31} = -.7404, \quad \phi_{32} = -.4892, \quad \phi_{33} = -.2433, \quad v_3 = 1.1436,\\
&\phi_{41} = -.7870, \quad \phi_{42} = -.5828, \quad \phi_{43} = -.3850, \quad \phi_{44} = -.1914, \quad v_4 = 1.1017.
\end{aligned}$$

Notice that as n increases, v_n approaches the white noise variance and θ_{n1} approaches θ. These results hold for any MA(1) process with $|\theta| < 1$. The innovations algorithm is particularly well suited to forecasting MA(q) processes, since for them $\theta_{nj} = 0$ for $n - j > q$. For AR(p) processes the Durbin–Levinson algorithm is usually more convenient, since $\phi_{nj} = 0$ for $n - j > p$. □

Recursive Calculation of the h-Step Predictors

For h-step prediction we use the result

$$P_n(X_{n+k} - P_{n+k-1}X_{n+k}) = 0, \quad k \ge 1. \tag{2.5.29}$$

This follows from (2.5.10) and the fact that

$$E[(X_{n+k} - P_{n+k-1}X_{n+k} - 0)X_{n+j-1}] = 0, \quad j = 1, \ldots, n.$$

Hence,

$$\begin{aligned} P_n X_{n+h} &= P_n P_{n+h-1} X_{n+h} \\ &= P_n \hat{X}_{n+h} \\ &= P_n \left(\sum_{j=1}^{n+h-1} \theta_{n+h-1,j} \left(X_{n+h-j} - \hat{X}_{n+h-j} \right) \right). \end{aligned}$$

Applying (2.5.29) again and using the linearity of P_n we find that

$$P_n X_{n+h} = \sum_{j=h}^{n+h-1} \theta_{n+h-1,j} \left(X_{n+h-j} - \hat{X}_{n+h-j} \right), \tag{2.5.30}$$

where the coefficients θ_{nj} are determined as before by the innovations algorithm. Moreover, the mean squared error can be expressed as

$$\begin{aligned} E(X_{n+h} - P_n X_{n+h})^2 &= E X_{n+h}^2 - E(P_n X_{n+h})^2 \\ &= \kappa(n+h, n+h) - \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}^2 v_{n+h-j-1}. \end{aligned} \tag{2.5.31}$$

2.5.3 Prediction of a Stationary Process in Terms of Infinitely Many Past Values

It is often useful, when many past observations $X_m, \ldots, X_0, X_1, \ldots, X_n$ $(m < 0)$ are available, to evaluate the best linear predictor of X_{n+h} in terms of $1, X_m, \ldots, X_0, \ldots, X_n$. This predictor, which we shall denote by $P_{m,n} X_{n+h}$, can easily be evaluated by the methods described above. If $|m|$ is large, this predictor can be approximated by the sometimes more easily calculated mean square limit

$$\tilde{P}_n X_{n+h} = \lim_{m \to -\infty} P_{m,n} X_{n+h}.$$

We shall refer to $\tilde{P}_n$ as the **prediction operator based on the infinite past, $\{X_t, -\infty < t \le n\}$**. Analogously we shall refer to P_n as the **prediction operator based on the finite past**, $\{X_1, \ldots, X_n\}$. (Mean square convergence of random variables is discussed in Appendix C.)

Determination of $\tilde{P}_n X_{n+h}$

Like $P_n X_{n+h}$, the best linear predictor $\tilde{P}_n X_{n+h}$ when $\{X_n\}$ is a zero-mean stationary process with autocovariance function $\gamma(\cdot)$ is characterized by the equations

$$E\left[\left(X_{n+h} - \tilde{P}_n X_{n+h} \right) X_{n+1-i} \right] = 0, \quad i = 1, 2, \ldots.$$

If we can find a solution to these equations, it will necessarily be the uniquely defined predictor $\tilde{P}_n X_{n+h}$. An approach to this problem that is often effective is to *assume*

that $\tilde{P}_n X_{n+h}$ can be expressed in the form

$$\tilde{P}_n X_{n+h} = \sum_{j=1}^{\infty} \alpha_j X_{n+1-j},$$

in which case the preceding equations reduce to

$$E\left[\left(X_{n+h} - \sum_{j=1}^{\infty} \alpha_j X_{n+1-j}\right) X_{n+1-i}\right] = 0, \quad i = 1, 2, \ldots,$$

or equivalently,

$$\sum_{j=1}^{\infty} \gamma(i-j)\alpha_j = \gamma(h+i-1), \quad i = 1, 2, \ldots.$$

This is an infinite set of linear equations for the unknown coefficients α_i that determine $\tilde{P}_n X_{n+h}$, *provided that the resulting series converges*.

Properties of $\tilde{P}_n$:

Suppose that $EU^2 < \infty$, $EV^2 < \infty$, a, b, and c are constants, and $\Gamma = \text{Cov}(\mathbf{W}, \mathbf{W})$.

1. $E[(U - \tilde{P}_n(U))X_j] = 0$, $j \le n$.
2. $\tilde{P}_n(aU + bV + c) = a\tilde{P}_n(U) + b\tilde{P}_n(V) + c$.
3. $\tilde{P}_n(U) = U$ if U is a limit of linear combinations of X_j, $j \le n$.
4. $\tilde{P}_n(U) = EU$ if $\text{Cov}(U, X_j) = 0$ for all $j \le n$.

These properties can sometimes be used to simplify the calculation of $\tilde{P}_n X_{n+h}$, notably when the process $\{X_t\}$ is an ARMA process.

Example 2.5.7 Consider the causal invertible ARMA(1,1) process $\{X_t\}$ defined by

$$X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right).$$

We know from (2.3.3) and (2.3.5) that we have the representations

$$X_{n+1} = Z_{n+1} + (\phi + \theta) \sum_{j=1}^{\infty} \phi^{j-1} Z_{n+1-j}$$

and

$$Z_{n+1} = X_{n+1} - (\phi + \theta) \sum_{j=1}^{\infty} (-\theta)^{j-1} X_{n+1-j}.$$

Applying the operator $\tilde{P}_n$ to the second equation and using the properties of $\tilde{P}_n$ gives

$$\tilde{P}_n X_{n+1} = (\phi + \theta) \sum_{j=1}^{\infty} (-\theta)^{j-1} X_{n+1-j}.$$

Applying the operator $\tilde{P}_n$ to the first equation and using the properties of $\tilde{P}_n$ gives

$$\tilde{P}_n X_{n+1} = (\phi + \theta) \sum_{j=1}^{\infty} \phi^{j-1} Z_{n+1-j}.$$

Hence,

$$X_{n+1} - \tilde{P}_n X_{n+1} = Z_{n+1},$$

and so the mean squared error of the predictor $\tilde{P}_n X_{n+1}$ is $EZ_{n+1}^2 = \sigma^2$. □

2.6 The Wold Decomposition

Consider the stationary process

$$X_t = A\cos(\omega t) + B\sin(\omega t),$$

where $\omega \in (0, \pi)$ is constant and A, B are uncorrelated random variables with mean 0 and variance σ^2. Notice that

$$X_n = (2\cos\omega)X_{n-1} - X_{n-2} = \tilde{P}_{n-1} X_n, \quad n = 0, \pm 1, \ldots,$$

so that $X_n - \tilde{P}_{n-1} X_n = 0$ for all n. Processes with the latter property are said to be **deterministic**.

The Wold Decomposition:

If $\{X_t\}$ is a nondeterministic stationary time series, then

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} + V_t, \tag{2.6.1}$$

where

1. $\psi_0 = 1$ and $\sum_{j=0}^{\infty} \psi_j^2 < \infty$,
2. $\{Z_t\} \sim \text{WN}\left(0, \sigma^2\right)$,
3. $\text{Cov}(Z_s, V_t) = 0$ for all s and t,
4. $Z_t = \tilde{P}_t Z_t$ for all t,
5. $V_t = \tilde{P}_s V_t$ for all s and t, and
6. $\{V_t\}$ is deterministic.

Here as in Section 2.5, $\tilde{P}_t Y$ denotes the best predictor of Y in terms of linear combinations, or limits of linear combinations of $1, X_s, -\infty < s \le t$. The sequences $\{Z_t\}$, $\{\psi_j\}$, and $\{V_t\}$ are unique and can be written explicitly as $Z_t = X_t - \tilde{P}_{t-1}X_t$, $\psi_j = E(X_t Z_{t-j})/E(Z_t^2)$, and $V_t = X_t - \sum_{j=0}^{\infty} \psi_j Z_{t-j}$. (See TSTM, p. 188.) For most of the zero-mean stationary time series dealt with in this book (in particular for all ARMA processes) the deterministic component V_t is 0 for all t, and the series is then said to be **purely nondeterministic**.

Example 2.6.1 If $X_t = U_t + Y$, where $\{U_t\} \sim \text{WN}(0, \nu^2)$, $E(U_t Y) = 0$ for all t, and Y has mean 0 and variance τ^2, then $\tilde{P}_{t-1}X_t = Y$, since Y is the mean square limit as $s \to \infty$ of $[X_{t-1} + \cdots + X_{t-s}]/s$, and $E[(X_t - Y)X_s] = 0$ for all $s \le t-1$. Hence the sequences in the Wold decomposition of $\{X_t\}$ are given by $Z_t = U_t$, $\psi_0 = 1$, $\psi_j = 0$ for $j > 0$, and $V_t = Y$. □

Problems

2.1. Suppose that $X_1, X_2, \ldots$, is a stationary time series with mean μ and ACF $\rho(\cdot)$. Show that the best predictor of X_{n+h} of the form $aX_n + b$ is obtained by choosing $a = \rho(h)$ and $b = \mu(1 - \rho(h))$.

2.2. Show that the process

$$X_t = A\cos(\omega t) + B\sin(\omega t), \quad t = 0, \pm 1, \ldots$$

(where A and B are uncorrelated random variables with mean 0 and variance 1 and ω is a fixed frequency in the interval $[0, \pi]$), is stationary and find its mean and autocovariance function. Deduce that the function $\kappa(h) = \cos(\omega h), h = 0, \pm 1, \ldots$, is nonnegative definite.

2.3. a. Find the ACVF of the time series $X_t = Z_t + .3Z_{t-1} - .4Z_{t-2}$, where $\{Z_t\} \sim \text{WN}(0, 1)$.

b. Find the ACVF of the time series $Y_t = \tilde{Z}_t - 1.2\tilde{Z}_{t-1} - 1.6\tilde{Z}_{t-2}$, where $\{\tilde{Z}_t\} \sim \text{WN}(0, .25)$. Compare with the answer found in (a).

2.4. It is clear that the function $\kappa(h) = 1, h = 0, \pm 1, \ldots$, is an autocovariance function, since it is the autocovariance function of the process $X_t = Z, t = 0, \pm 1, \ldots$, where Z is a random variable with mean 0 and variance 1. By identifying appropriate sequences of random variables, show that the following

functions are also autocovariance functions:

(a) $\kappa(h) = (-1)^{|h|}$

(b) $\kappa(h) = 1 + \cos\left(\frac{\pi h}{2}\right) + \cos\left(\frac{\pi h}{4}\right)$

(c) $\kappa(h) = \begin{cases} 1, & \text{if } h = 0, \\ 0.4, & \text{if } h = \pm 1, \\ 0, & \text{otherwise.} \end{cases}$

2.5. Suppose that $\{X_t, t = 0, \pm 1, \ldots\}$ is stationary and that $|\theta| < 1$. Show that for each fixed n the sequence

$$S_m = \sum_{j=1}^{m} \theta^j X_{n-j}$$

is convergent absolutely and in mean square (see Appendix C) as $m \to \infty$.

2.6. Verify equations (2.2.6).

2.7. Show, using the geometric series $1/(1-x) = \sum_{j=0}^{\infty} x^j$ for $|x| < 1$, that $1/(1-\phi z) = -\sum_{j=1}^{\infty} \phi^{-j} z^{-j}$ for $|\phi| > 1$ and $|z| \geq 1$.

2.8. Show that the autoregressive equations

$$X_t = \phi_1 X_{t-1} + Z_t, \quad t = 0, \pm 1, \ldots,$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and $|\phi| = 1$, have *no stationary solution*. HINT: Suppose there does exist a stationary solution $\{X_t\}$ and use the autoregressive equation to derive an expression for the variance of $X_t - \phi_1^{n+1} X_{t-n-1}$ that contradicts the stationarity assumption.

2.9. Let $\{Y_t\}$ be the AR(1) plus noise time series defined by

$$Y_t = X_t + W_t,$$

where $\{W_t\} \sim \text{WN}(0, \sigma_w^2)$, $\{X_t\}$ is the AR(1) process of Example 2.2.1, i.e.,

$$X_t - \phi X_{t-1} = Z_t, \{Z_t\} \sim \text{WN}\left(0, \sigma_z^2\right),$$

and $E(W_s Z_t) = 0$ for all s and t.

a. Show that $\{Y_t\}$ is stationary and find its autocovariance function.

b. Show that the time series $U_t := Y_t - \phi Y_{t-1}$ is 1-correlated and hence, by Proposition 2.1.1, is an MA(1) process.

c. Conclude from (b) that $\{Y_t\}$ is an ARMA(1,1) process and express the three parameters of this model in terms of ϕ, σ_w^2, and σ_z^2.

2.10. Use the program ITSM to compute the coefficients ψ_j and π_j, $j = 1, \ldots, 5$, in the expansions

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$$

and

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}$$

for the ARMA(1,1) process defined by the equations

$$X_t - 0.5X_{t-1} = Z_t + 0.5Z_{t-1}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right).$$

(Select `File>Project>New>Univariate`, then `Model>Specify`. In the resulting dialog box enter 1 for the AR and MA orders, specify $\phi(1) = \theta(1) = 0.5$, and click `OK`. Finally, select `Model>AR/MA Infinity>Default lag` and the values of ψ_j and π_j will appear on the screen.) Check the results with those obtained in Section 2.3.

2.11. Suppose that in a sample of size 100 from an AR(1) process with mean μ, $\phi = .6$, and $\sigma^2 = 2$ we obtain $\bar{x}_{100} = .271$. Construct an approximate 95% confidence interval for μ. Are the data compatible with the hypothesis that $\mu = 0$?

2.12. Suppose that in a sample of size 100 from an MA(1) process with mean μ, $\theta = -.6$, and $\sigma^2 = 1$ we obtain $\bar{x}_{100} = .157$. Construct an approximate 95% confidence interval for μ. Are the data compatible with the hypothesis that $\mu = 0$?

2.13. Suppose that in a sample of size 100, we obtain $\hat{\rho}(1) = .438$ and $\hat{\rho}(2) = .145$.

a. Assuming that the data were generated from an AR(1) model, construct approximate 95% confidence intervals for both $\rho(1)$ and $\rho(2)$. Based on these two confidence intervals, are the data consistent with an AR(1) model with $\phi = .8$?

b. Assuming that the data were generated from an MA(1) model, construct approximate 95% confidence intervals for both $\rho(1)$ and $\rho(2)$. Based on these two confidence intervals, are the data consistent with an MA(1) model with $\theta = .6$?

2.14. Let $\{X_t\}$ be the process defined in Problem 2.2.

a. Find $P_1 X_2$ and its mean squared error.

b. Find $P_2 X_3$ and its mean squared error.

c. Find $\tilde{P}_n X_{n+1}$ and its mean squared error.

2.15. Suppose that $\{X_t, t = 0, \pm 1, \ldots\}$ is a stationary process satisfying the equations

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t,$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and Z_t is uncorrelated with X_s for each $s < t$. Show that the best linear predictor $P_n X_{n+1}$ of X_{n+1} in terms of $1, X_1, \ldots, X_n$, assuming $n > p$, is

$$P_n X_{n+1} = \phi_1 X_n + \cdots + \phi_p X_{n+1-p}.$$

What is the mean squared error of $P_n X_{n+1}$?

2.16. Use the program ITSM to plot the sample ACF and PACF up to lag 40 of the sunspot series $D_t, t = 1, 100$, contained in the ITSM file SUNSPOTS.TSM. (Open the project SUNSPOTS.TSM and click on the second yellow button at the top of the screen to see the graphs. Repeated clicking on this button will toggle between graphs of the sample ACF, sample PACF, and both. To see the numerical values, right-click on the graph and select `Info`.) Fit an AR(2) model to the mean-corrected data by selecting `Model>Estimation>Preliminary` and click `Yes` to subtract the sample mean from the data. In the dialog box that follows, enter 2 for the AR order and make sure that the MA order is zero and that the `Yule-Walker` algorithm is selected *without* AICC minimization. Click `OK` and you will obtain a model of the form

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t, \qquad \text{where } \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right),$$

for the mean-corrected series $X_t = D_t - 46.93$. Record the values of the estimated parameters ϕ_1, ϕ_2, and σ^2. Compare the model and sample ACF and PACF by selecting the third yellow button at the top of the screen. Print the graphs by right-clicking and selecting `Print`.

2.17. Without exiting from ITSM, use the model found in the preceding problem to compute forecasts of the next ten values of the sunspot series. (Select `Forecasting>ARMA`, make sure that the number of forecasts is set to 10 and the box `Add the mean to the forecasts` is checked, and then click `OK`. You will see a graph of the original data with the ten forecasts appended. Right-click on the graph and then on `Info` to get the numerical values of the forecasts. Print the graph as described in Problem 2.16.) The details of the calculations will be taken up in Chapter 3 when we discuss ARMA models in detail.

2.18. Let $\{X_t\}$ be the stationary process defined by the equations

$$X_t = Z_t - \theta Z_{t-1}, \quad t = 0, \pm 1, \ldots,$$

where $|\theta| < 1$ and $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. Show that the best linear predictor $\tilde{P}_n X_{n+1}$ of X_{n+1} based on $\{X_j, -\infty < j \le n\}$ is

$$\tilde{P}_n X_{n+1} = -\sum_{j=1}^{\infty} \theta^j X_{n+1-j}.$$

What is the mean squared error of the predictor $\tilde{P}_n X_{n+1}$?

2.19. If $\{X_t\}$ is defined as in Problem 2.18 and $\theta = 1$, find the best linear predictor $P_n X_{n+1}$ of X_{n+1} in terms of $X_1, \ldots, X_n$. What is the corresponding mean squared error?

2.20. In the innovations algorithm, show that for each $n \ge 2$, the innovation $X_n - \hat{X}_n$ is uncorrelated with $X_1, \ldots, X_{n-1}$. Conclude that $X_n - \hat{X}_n$ is uncorrelated with the innovations $X_1 - \hat{X}_1, \ldots, X_{n-1} - \hat{X}_{n-1}$.

2.21. Let X_1, X_2, X_4, X_5 be observations from the MA(1) model

$$X_t = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right).$$

a. Find the best linear estimate of the missing value X_3 in terms of X_1 and X_2.

b. Find the best linear estimate of the missing value X_3 in terms of X_4 and X_5.

c. Find the best linear estimate of the missing value X_3 in terms of X_1, X_2, X_4, and X_5.

d. Compute the mean squared errors for each of the estimates in (a), (b), and (c).

2.22. Repeat parts (a)–(d) of Problem 2.21 assuming now that the observations X_1, X_2, X_4, X_5 are from the causal AR(1) model

$$X_t = \phi X_{t-1} + Z_t, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right).$$

3 ARMA Models

3.1 ARMA(p, q) Processes
3.2 The ACF and PACF of an ARMA(p, q) Process
3.3 Forecasting ARMA Processes

In this chapter we introduce an important parametric family of stationary time series, the autoregressive moving-average, or ARMA, processes. For a large class of autocovariance functions $\gamma(\cdot)$ it is possible to find an ARMA process $\{X_t\}$ with ACVF $\gamma_X(\cdot)$ such that $\gamma(\cdot)$ is well approximated by $\gamma_X(\cdot)$. In particular, for any positive integer K, there exists an ARMA process $\{X_t\}$ such that $\gamma_X(h) = \gamma(h)$ for $h = 0, 1, \ldots, K$. For this (and other) reasons, the family of ARMA processes plays a key role in the modeling of time series data. The linear structure of ARMA processes also leads to a substantial simplification of the general methods for linear prediction discussed earlier in Section 2.5.

3.1 ARMA(p, q) Processes

In Section 2.3 we introduced an ARMA(1,1) process and discussed some of its key properties. These included existence and uniqueness of stationary solutions of the defining equations and the concepts of causality and invertibility. In this section we extend these notions to the general ARMA(p, q) process.

Definition 3.1.1

$\{X_t\}$ is an **ARMA(p, q) process** if $\{X_t\}$ is stationary and if for every t,

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \qquad (3.1.1)$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and the polynomials $(1 - \phi_1 z - \ldots - \phi_p z^p)$ and $(1 + \theta_1 z + \ldots + \theta_q z^q)$ have no common factors.

The process $\{X_t\}$ is said to be an **ARMA(p, q) process with mean μ** if $\{X_t - \mu\}$ is an ARMA(p, q) process.

It is convenient to use the more concise form of (3.1.1)

$$\phi(B)X_t = \theta(B)Z_t, \tag{3.1.2}$$

where $\phi(\cdot)$ and $\theta(\cdot)$ are the pth and qth-degree polynomials

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$

and

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q,$$

and B is the backward shift operator ($B^j X_t = X_{t-j},\ B^j Z_t = Z_{t-j},\ j = 0, \pm 1, \ldots$). The time series $\{X_t\}$ is said to be an **autoregressive process of order p** (or AR(p)) if $\theta(z) \equiv 1$, and a **moving-average process of order q** (or MA(q)) if $\phi(z) \equiv 1$.

An important part of Definition 3.1.1 is the requirement that $\{X_t\}$ be stationary. In Section 2.3 we showed, for the ARMA(1,1) equations (2.3.1), that a stationary solution exists (and is unique) if and only if $\phi_1 \neq \pm 1$. The latter is equivalent to the condition that the autoregressive polynomial $\phi(z) = 1 - \phi_1 z \neq 0$ for $z = \pm 1$. The analogous condition for the general ARMA(p, q) process is $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0$ for all complex z with $|z| = 1$. (Complex z is used here, since the zeros of a polynomial of degree $p > 1$ may be either real or complex. The region defined by the set of complex z such that $|z| = 1$ is referred to as the unit circle.) If $\phi(z) \neq 0$ for all z on the unit circle, then there exists $\delta > 0$ such that

$$\frac{1}{\phi(z)} = \sum_{j=-\infty}^{\infty} \chi_j z^j \quad \text{for } 1 - \delta < |z| < 1 + \delta,$$

and $\sum_{j=-\infty}^{\infty} |\chi_j| < \infty$. We can then define $1/\phi(B)$ as the linear filter with absolutely summable coefficients

$$\frac{1}{\phi(B)} = \sum_{j=-\infty}^{\infty} \chi_j B^j.$$

Applying the operator $\chi(B) := 1/\phi(B)$ to both sides of (3.1.2), we obtain

$$X_t = \chi(B)\phi(B)X_t = \chi(B)\theta(B)Z_t = \psi(B)Z_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \tag{3.1.3}$$

where $\psi(z) = \chi(z)\theta(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$. Using the argument given in Section 2.3 for the ARMA(1,1) process, it follows that $\psi(B)Z_t$ is the unique stationary solution of (3.1.1).

Existence and Uniqueness:

A stationary solution $\{X_t\}$ of equations (3.1.1) exists (and is also the unique stationary solution) if and only if

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0 \quad \text{for all } |z| = 1. \tag{3.1.4}$$

In Section 2.3 we saw that the ARMA(1,1) process is causal, i.e., that X_t can be expressed in terms of Z_s, $s \leq t$, if and only if $|\phi_1| < 1$. For a general ARMA(p, q) process the analogous condition is that $\phi(z) \neq 0$ for $|z| \leq 1$, i.e., the zeros of the autoregressive polynomial must all be greater than 1 in absolute value.

Causality:

An ARMA(p, q) process $\{X_t\}$ is **causal**, or a **causal function of $\{Z_t\}$**, if there exist constants $\{\psi_j\}$ such that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ and

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} \text{ for all } t. \tag{3.1.5}$$

Causality is equivalent to the condition

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0 \text{ for all } |z| \leq 1. \tag{3.1.6}$$

The proof of the equivalence between causality and (3.1.6) follows from elementary properties of power series. From (3.1.3) we see that $\{X_t\}$ is causal if and only if $\chi(z) := 1/\phi(z) = \sum_{j=0}^{\infty} \chi_j z^j$ (assuming that $\phi(z)$ and $\theta(z)$ have no common factors). But this, in turn, is equivalent to (3.1.6).

The sequence $\{\psi_j\}$ in (3.1.5) is determined by the relation $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z)$, or equivalently by the identity

$$\left(1 - \phi_1 z - \cdots - \phi_p z^p\right)(\psi_0 + \psi_1 z + \cdots) = 1 + \theta_1 z + \cdots + \theta_q z^q.$$

Equating coefficients of z^j, $j = 0, 1, \ldots$, we find that

$$\begin{aligned} 1 &= \psi_0, \\ \theta_1 &= \psi_1 - \psi_0\phi_1, \\ \theta_2 &= \psi_2 - \psi_1\phi_1 - \psi_0\phi_2, \\ &\vdots \end{aligned}$$

or equivalently,

$$\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = \theta_j, \quad j = 0, 1, \ldots, \tag{3.1.7}$$

where $\theta_0 := 1$, $\theta_j := 0$ for $j > q$, and $\psi_j := 0$ for $j < 0$.

Invertibility, which allows Z_t to be expressed in terms of X_s, $s \le t$, has a similar characterization in terms of the moving-average polynomial.

Invertibility:

An ARMA(p, q) process $\{X_t\}$ is **invertible** if there exist constants $\{\pi_j\}$ such that $\sum_{j=0}^{\infty} |\pi_j| < \infty$ and

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j} \text{ for all } t.$$

Invertibility is equivalent to the condition

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q \neq 0 \text{ for all } |z| \le 1.$$

Interchanging the roles of the AR and MA polynomials, we find from (3.1.7) that the sequence $\{\pi_j\}$ is determined by the equations

$$\pi_j + \sum_{k=1}^{q} \theta_k \pi_{j-k} = -\phi_j, \quad j = 0, 1, \ldots, \tag{3.1.8}$$

where $\phi_0 := -1$, $\phi_j := 0$ for $j > p$, and $\pi_j := 0$ for $j < 0$.

Example 3.1.1 An ARMA(1,1) process

Consider the ARMA(1,1) process $\{X_t\}$ satisfying the equations

$$X_t - .5X_{t-1} = Z_t + .4Z_{t-1}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right). \tag{3.1.9}$$

Since the autoregressive polynomial $\phi(z) = 1 - .5z$ has a zero at $z = 2$, which is located outside the unit circle, we conclude from (3.1.4) and (3.1.6) that there exists a unique ARMA process satisfying (3.1.9) that is also causal. The coefficients $\{\psi_j\}$ in the MA(∞) representation of $\{X_t\}$ are found directly from (3.1.7):

$$\begin{aligned}
&\psi_0 = 1,\\
&\psi_1 = .4 + .5,\\
&\psi_2 = .5(.4 + .5),\\
&\psi_j = .5^{j-1}(.4 + .5), \quad j = 1, 2, \ldots.
\end{aligned}$$

The MA polynomial $\theta(z) = 1 + .4z$ has a zero at $z = -1/.4 = -2.5$, which is also located outside the unit circle. This implies that $\{X_t\}$ is invertible with coefficients

$\{\pi_j\}$ given by (see (3.1.8))

$$\pi_0 = 1,$$
$$\pi_1 = -(.4 + .5),$$
$$\pi_2 = -(.4 + .5)(-.4),$$
$$\pi_j = -(.4 + .5)(-.4)^{j-1}, \quad j = 1, 2, \ldots.$$

(A direct derivation of these formulas for $\{\psi_j\}$ and $\{\pi_j\}$ was given in Section 2.3 without appealing to the recursions (3.1.7) and (3.1.8).) □

Example 3.1.2 An AR(2) process

Let $\{X_t\}$ be the AR(2) process

$$X_t = .7X_{t-1} - .1X_{t-2} + Z_t, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right).$$

The autoregressive polynomial for this process has the factorization $\phi(z) = 1 - .7z + .1z^2 = (1 - .5z)(1 - .2z)$, and is therefore zero at $z = 2$ and $z = 5$. Since these zeros lie outside the unit circle, we conclude that $\{X_t\}$ is a causal AR(2) process with coefficients $\{\psi_j\}$ given by

$$\psi_0 = 1,$$
$$\psi_1 = .7,$$
$$\psi_2 = .7^2 - .1,$$
$$\psi_j = .7\psi_{j-1} - .1\psi_{j-2}, \quad j = 2, 3, \ldots.$$

While it is a simple matter to calculate ψ_j numerically for any j, it is possible also to give an explicit solution of these difference equations using the theory of linear difference equations (see TSTM, Section 3.6). □

The option `Model>Specify` of the program ITSM allows the entry of any causal ARMA(p, q) model with $p < 28$ and $q < 28$. This option contains a causality check and will immediately let you know if the entered model is noncausal. (A causal model can be obtained by setting all the AR coefficients equal to .001.) Once a causal model has been entered, the coefficients ψ_j in the MA(∞) representation of the process can be computed by selecting `Model>AR/MA Infinity`. This option will also compute the AR(∞) coefficients π_j, provided that the model is invertible.

Example 3.1.3 An ARMA(2,1) process

Consider the ARMA(2,1) process defined by the equations

$$X_t - .75X_{t-1} + .5625X_{t-2} = Z_t + 1.25Z_{t-1}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right).$$

The AR polynomial $\phi(z) = 1 - .75z + .5625z^2$ has zeros at $z = 2\big(1 \pm i\sqrt{3}\big)/3$, which lie outside the unit circle. The process is therefore causal. On the other hand, the MA polynomial $\theta(z) = 1 + 1.25z$ has a zero at $z = -.8$, and hence $\{X_t\}$ is not invertible. □

Remark 1. It should be noted that causality and invertibility are properties not of $\{X_t\}$ alone, but rather of the relationship between the two processes $\{X_t\}$ and $\{Z_t\}$ appearing in the defining ARMA equations (3.1.1). □

Remark 2. If $\{X_t\}$ is an ARMA process defined by $\phi(B)X_t = \theta(B)Z_t$, where $\theta(z) \neq 0$ if $|z| = 1$, then it is always possible (see TSTM, p. 127) to find polynomials $\tilde{\phi}(z)$ and $\tilde{\theta}(z)$ and a white noise sequence $\{W_t\}$ such that $\tilde{\phi}(B)X_t = \tilde{\theta}(B)W_t$ and $\tilde{\theta}(z)$ and $\tilde{\phi}(z)$ are nonzero for $|z| \leq 1$. However, if the original white noise sequence $\{Z_t\}$ is iid, then the new white noise sequence will not be iid unless $\{Z_t\}$ is Gaussian. □

In view of the preceding remark, we will focus our attention principally on causal and invertible ARMA processes.

3.2 The ACF and PACF of an ARMA(p, q) Process

In this section we discuss three methods for computing the autocovariance function $\gamma(\cdot)$ of a causal ARMA process $\{X_t\}$. The autocorrelation function is readily found from the ACVF on dividing by $\gamma(0)$. The partial autocorrelation function (PACF) is also found from the function $\gamma(\cdot)$.

3.2.1 Calculation of the ACVF

First we determine the ACVF $\gamma(\cdot)$ of the causal ARMA(p, q) process defined by

$$\phi(B)X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right), \tag{3.2.1}$$

where $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ and $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$. The causality assumption implies that

$$X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \tag{3.2.2}$$

where $\sum_{j=0}^{\infty} \psi_j z^j = \theta(z)/\phi(z),\ |z| \leq 1$. The calculation of the sequence $\{\psi_j\}$ was discussed in Section 3.1.

First Method. From Proposition 2.2.1 and the representation (3.2.2), we obtain

$$\gamma(h) = E(X_{t+h}X_t) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+|h|}. \tag{3.2.3}$$

Example 3.2.1 The ARMA(1,1) process

Substituting from (2.3.3) into (3.2.3), we find that the ACVF of the process defined by

$$X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right), \tag{3.2.4}$$

with $|\phi| < 1$ is given by

$$\begin{aligned}
\gamma(0) &= \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 \\
&= \sigma^2 \left[1 + (\theta+\phi)^2 \sum_{j=0}^{\infty} \phi^{2j}\right] \\
&= \sigma^2 \left[1 + \frac{(\theta+\phi)^2}{1-\phi^2}\right], \\
\gamma(1) &= \sigma^2 \sum_{j=0}^{\infty} \psi_{j+1}\psi_j \\
&= \sigma^2 \left[\theta + \phi + (\theta+\phi)^2\phi \sum_{j=0}^{\infty} \phi^{2j}\right] \\
&= \sigma^2 \left[\theta + \phi + \frac{(\theta+\phi)^2\phi}{1-\phi^2}\right],
\end{aligned}$$

and

$$\gamma(h) = \phi^{h-1}\gamma(1), \quad h \geq 2. \qquad \square$$

Example 3.2.2 The MA(q) process

For the process

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right),$$

equation (3.2.3) immediately gives the result

$$\gamma(h) = \begin{cases} \sigma^2 \displaystyle\sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|}, & \text{if } |h| \leq q, \\ 0, & \text{if } |h| > q, \end{cases}$$

where θ_0 is defined to be 1. The ACVF of the MA(q) process thus has the distinctive feature of vanishing at lags greater than q. Data for which the sample ACVF is small for lags greater than q therefore suggest that an appropriate model might be a

moving average of order q (or less). Recall from Proposition 2.1.1 that *every* zero-mean stationary process with correlations vanishing at lags greater than q can be represented as a moving-average process of order q or less. □

Second Method. If we multiply each side of the equations

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

by X_{t-k}, $k = 0, 1, 2, \ldots$, and take expectations on each side, we find that

$$\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = \sigma^2 \sum_{j=0}^{\infty} \theta_{k+j}\psi_j, \quad 0 \le k < m, \tag{3.2.5}$$

and

$$\gamma(k) - \phi_1\gamma(k-1) - \cdots - \phi_p\gamma(k-p) = 0, \quad k \ge m, \tag{3.2.6}$$

where $m = \max(p, q+1)$, $\psi_j := 0$ for $j < 0$, $\theta_0 := 1$, and $\theta_j := 0$ for $j \notin \{0, \ldots, q\}$. In calculating the right-hand side of (3.2.5) we have made use of the expansion (3.2.2). Equations (3.2.6) are a set of homogeneous linear difference equations with constant coefficients, for which the solution is well known (see, e.g., TSTM, Section 3.6) to be of the form

$$\gamma(h) = \alpha_1\xi_1^{-h} + \alpha_2\xi_2^{-h} + \cdots + \alpha_p\xi_p^{-h}, \quad h \ge m - p, \tag{3.2.7}$$

where $\xi_1, \ldots, \xi_p$ are the roots (assumed to be distinct) of the equation $\phi(z) = 0$, and $\alpha_1, \ldots, \alpha_p$ are arbitrary constants. (For further details, and for the treatment of the case where the roots are not distinct, see TSTM, Section 3.6.) Of course, we are looking for the solution of (3.2.6) that also satisfies (3.2.5). We therefore substitute the solution (3.2.7) into (3.2.5) to obtain a set of m linear equations that then uniquely determine the constants $\alpha_1, \ldots, \alpha_p$ and the $m - p$ autocovariances $\gamma(h)$, $0 \le h < m - p$.

Example 3.2.3 The ARMA(1,1) process

For the causal ARMA(1,1) process defined in Example 3.2.1, equations (3.2.5) are

$$\gamma(0) - \phi\gamma(1) = \sigma^2(1 + \theta(\theta + \phi)) \tag{3.2.8}$$

and

$$\gamma(1) - \phi\gamma(0) = \sigma^2\theta. \tag{3.2.9}$$

Equation (3.2.6) takes the form

$$\gamma(k) - \phi\gamma(k-1) = 0, \quad k \ge 2. \tag{3.2.10}$$

The solution of (3.2.10) is

$$\gamma(h) = \alpha\phi^h, \quad h \ge 1.$$

Substituting this expression for $\gamma(h)$ into the two preceding equations (3.2.8) and (3.2.9) gives two linear equations for α and the unknown autocovariance $\gamma(0)$. These equations are easily solved, giving the autocovariances already found for this process in Example 3.2.1. □

Example 3.2.4 The general AR(2) process

For the causal AR(2) process defined by

$$\left(1-\xi_1^{-1}B\right)\left(1-\xi_2^{-1}B\right)X_t = Z_t, \quad |\xi_1|, |\xi_2| > 1, \xi_1 \neq \xi_2,$$

we easily find from (3.2.7) and (3.2.5) using the relations

$$\phi_1 = \xi_1^{-1} + \xi_2^{-1}$$

and

$$\phi_2 = -\xi_1^{-1}\xi_2^{-1}$$

that

$$\gamma(h) = \frac{\sigma^2\xi_1^2\xi_2^2}{(\xi_1\xi_2-1)(\xi_2-\xi_1)}\left[(\xi_1^2-1)^{-1}\xi_1^{1-h} - (\xi_2^2-1)^{-1}\xi_2^{1-h}\right]. \tag{3.2.11}$$

Figures 3.1–3.4 illustrate some of the possible forms of $\gamma(\cdot)$ for different values of ξ_1 and ξ_2. Notice that in the case of complex conjugate roots $\xi_1 = re^{i\theta}$ and $\xi_2 = re^{-i\theta}$, $0 < \theta < \pi$, we can write (3.2.11) in the more illuminating form

$$\gamma(h) = \frac{\sigma^2 r^4 \cdot r^{-h}\sin(h\theta+\psi)}{(r^2-1)(r^4-2r^2\cos 2\theta+1)\sin\theta}, \tag{3.2.12}$$

where

$$\tan\psi = \frac{r^2+1}{r^2-1}\tan\theta \tag{3.2.13}$$

and $\cos\psi$ has the same sign as $\cos\theta$. Thus in this case $\gamma(\cdot)$ has the form of a damped sinusoidal function with damping factor r^{-1} and period $2\pi/\theta$. If the roots are close to the unit circle, then r is close to 1, the damping is slow, and we obtain a nearly sinusoidal autocovariance function. □

Third Method. The autocovariances can also be found by solving the first $p+1$ equations of (3.2.5) and (3.2.6) for $\gamma(0)\ldots,\gamma(p)$ and then using the subsequent equations to solve successively for $\gamma(p+1), \gamma(p+2), \ldots$. This is an especially convenient method for numerical determination of the autocovariances $\gamma(h)$ and is used in the option `Model>ACF/PACF>Model` of the program ITSM.

Example 3.2.5 Consider again the causal ARMA(1,1) process of Example 3.2.1. To apply the third method we simply solve (3.2.8) and (3.2.9) for $\gamma(0)$ and $\gamma(1)$. Then $\gamma(2), \gamma(3), \ldots$

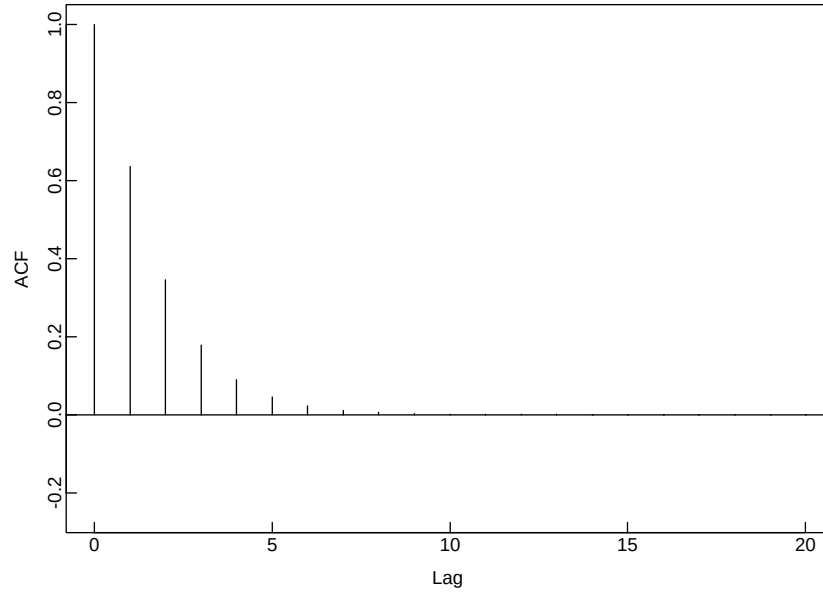

Figure 3-1
The model ACF of the AR(2) series of Example 3.2.4 with $\xi_1 = 2$ and $\xi_2 = 5$.

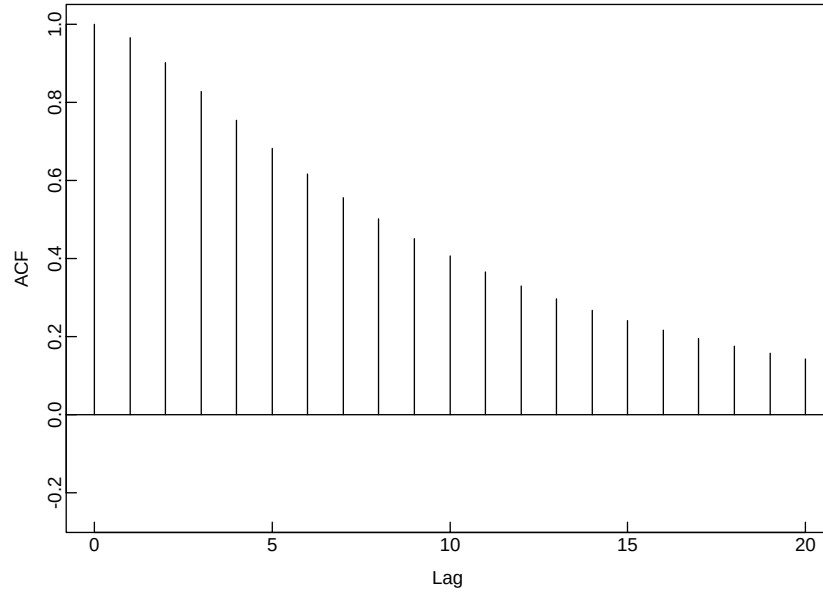

Figure 3-2
The model ACF of the AR(2) series of Example 3.2.4 with $\xi_1 = 10/9$ and $\xi_2 = 2$.

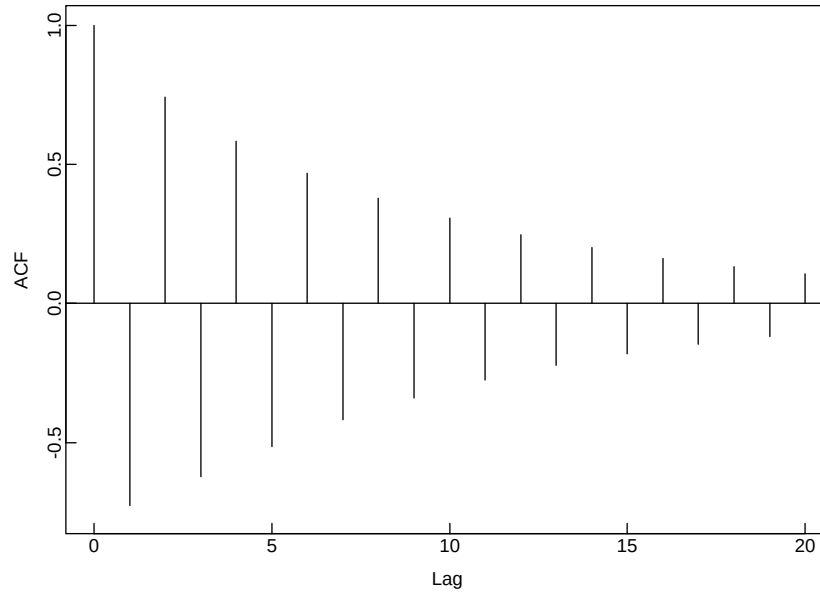

Figure 3-3
The model ACF of the AR(2) series of Example 3.2.4 with $\xi_1 = -10/9$ and $\xi_2 = 2$.

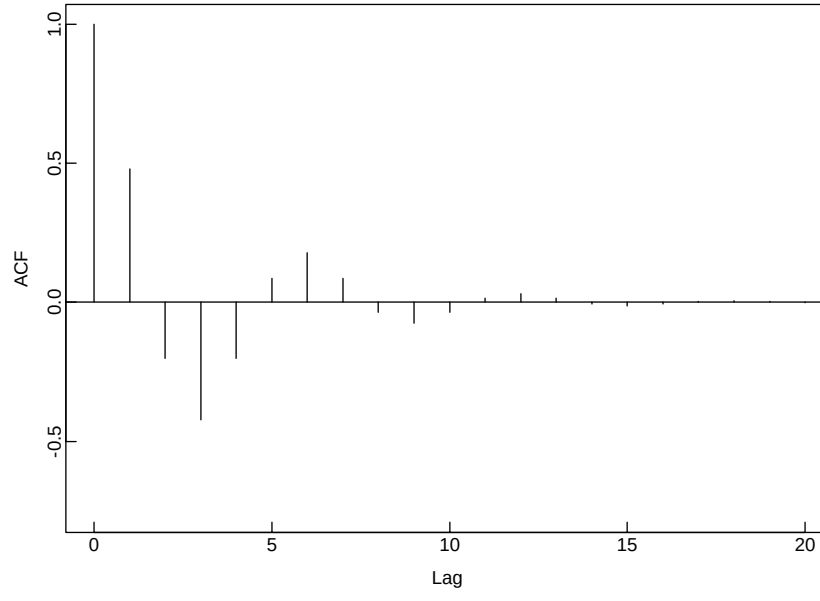

Figure 3-4
The model ACF of the AR(2) series of Example 3.2.4 with $\xi_1 = 2(1 + i\sqrt{3})/3$ and $\xi_2 = 2(1 - i\sqrt{3})/3$.

can be found successively from (3.2.10). It is easy to check that this procedure gives the same results as those obtained in Examples 3.2.1 and 3.2.3. □

3.2.2 The Autocorrelation Function

Recall that the ACF of an ARMA process $\{X_t\}$ is the function $\rho(\cdot)$ found immediately from the ACVF $\gamma(\cdot)$ as

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}.$$

Likewise, for any set of observations $\{x_1, \ldots, x_n\}$, the sample ACF $\hat{\rho}(\cdot)$ is computed as

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

The Sample ACF of an MA(q) Series. Given observations $\{x_1, \ldots, x_n\}$ of a time series, one approach to the fitting of a model to the data is to match the sample ACF of the data with the ACF of the model. In particular, if the sample ACF $\hat{\rho}(h)$ is significantly different from zero for $0 \leq h \leq q$ and negligible for $h > q$, Example 3.2.2 suggests that an MA(q) model might provide a good representation of the data. In order to apply this criterion we need to take into account the random variation expected in the sample autocorrelation function before we can classify ACF values as "negligible." To resolve this problem we can use Bartlett's formula (Section 2.4), which implies that for a large sample of size n from an MA(q) process, the sample ACF values at lags greater than q are approximately normally distributed with means 0 and variances $w_{hh}/n = \big(1+2\rho^2(1)+\cdots+2\rho^2(q)\big)/n$. This means that if the sample is from an MA(q) process and if $h > q$, then $\hat{\rho}(h)$ should fall between the bounds $\pm 1.96\sqrt{w_{hh}/n}$ with probability approximately 0.95. In practice we frequently use the more stringent values $\pm 1.96/\sqrt{n}$ as the bounds between which sample autocovariances are considered "negligible." A more effective and systematic approach to the problem of model selection, which also applies to ARMA(p, q) models with $p > 0$ and $q > 0$, will be discussed in Section 5.5.

3.2.3 The Partial Autocorrelation Function

The **partial autocorrelation function (PACF)** of an ARMA process $\{X_t\}$ is the function $\alpha(\cdot)$ defined by the equations

$$\alpha(0) = 1$$

and

$$\alpha(h) = \phi_{hh}, \quad h \geq 1,$$

where ϕ_{hh} is the last component of

$$\boldsymbol{\phi}_h = \Gamma_h^{-1}\boldsymbol{\gamma}_h, \tag{3.2.14}$$

$\Gamma_h = [\gamma(i-j)]_{i,j=1}^h$, and $\boldsymbol{\gamma}_h = [\gamma(1), \gamma(2), \ldots, \gamma(h)]'$.

For any set of observations $\{x_1, \ldots, x_n\}$ with $x_i \neq x_j$ for some i and j, the **sample PACF** $\hat{\alpha}(h)$ is given by

$$\hat{\alpha}(0) = 1$$

and

$$\hat{\alpha}(h) = \hat{\phi}_{hh}, \quad h \geq 1,$$

where $\hat{\phi}_{hh}$ is the last component of

$$\hat{\boldsymbol{\phi}}_h = \hat{\Gamma}_h^{-1}\hat{\boldsymbol{\gamma}}_h. \tag{3.2.15}$$

We show in the next example that the PACF of a causal AR(p) process is zero for lags greater than p. Both sample and model partial autocorrelation functions can be computed numerically using the program ITSM. Algebraic calculation of the PACF is quite complicated except when q is zero or p and q are both small.

It can be shown (TSTM, p. 171) that ϕ_{hh} is the correlation between the prediction errors $X_h - P(X_h|X_1, \ldots, X_{h-1})$ and $X_0 - P(X_0|X_1, \ldots, X_{h-1})$.

Example 3.2.6 The PACF of an AR(p) process

For the causal AR(p) process defined by

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right),$$

we know (Problem 2.15) that for $h \geq p$ the best linear predictor of X_{h+1} in terms of $1, X_1, \ldots, X_h$ is

$$\hat{X}_{h+1} = \phi_1 X_h + \phi_2 X_{h-1} + \cdots + \phi_p X_{h+1-p}.$$

Since the coefficient ϕ_{hh} of X_1 is ϕ_p if $h = p$ and 0 if $h > p$, we conclude that the PACF $\alpha(\cdot)$ of the process $\{X_t\}$ has the properties

$$\alpha(p) = \phi_p$$

and

$$\alpha(h) = 0 \text{ for } h > p.$$

For $h < p$ the values of $\alpha(h)$ can easily be computed from (3.2.14). For any specified ARMA model the PACF can be evaluated numerically using the option `Model>ACF/PACF>Model` of the program ITSM. □

Example 3.2.7 The PACF of an MA(1) process

For the MA(1) process, it can be shown from (3.2.14) (see Problem 3.12) that the PACF at lag h is

$$\alpha(h) = \phi_{hh} = -(-\theta)^h / \left(1 + \theta^2 + \cdots + \theta^{2h}\right). \qquad \square$$

The Sample PACF of an AR(p) Series. If $\{X_t\}$ is an AR(p) series, then the sample PACF based on observations $\{x_1, \ldots, x_n\}$ should reflect (with sampling variation) the properties of the PACF itself. In particular, if the sample PACF $\hat{\alpha}(h)$ is significantly different from zero for $0 \le h \le p$ and negligible for $h > p$, Example 3.2.6 suggests that an AR(p) model might provide a good representation of the data. To decide what is meant by "negligible" we can use the result that for an AR(p) process the sample PACF values at lags greater than p are approximately independent $N(0, 1/n)$ random variables. This means that roughly 95% of the sample PACF values beyond lag p should fall within the bounds $\pm 1.96/\sqrt{n}$. If we observe a sample PACF satisfying $|\hat{\alpha}(h)| > 1.96/\sqrt{n}$ for $0 \le h \le p$ and $|\hat{\alpha}(h)| < 1.96/\sqrt{n}$ for $h > p$, this suggests an AR(p) model for the data. For a more systematic approach to model selection, see Section 5.5.

3.2.4 Examples

Example 3.2.8 The time series plotted in Figure 3.5 consists of 57 consecutive daily *overshorts* from an underground gasoline tank at a filling station in Colorado. If y_t is the measured amount of fuel in the tank at the end of the tth day and a_t is the measured amount sold minus the amount delivered during the course of the tth day, then the overshort at the end of day t is defined as $x_t = y_t - y_{t-1} + a_t$. Due to the error in measuring the current amount of fuel in the tank, the amount sold, and the amount delivered to the station, we view y_t, a_t, and x_t as observed values from some set of random variables Y_t, A_t, and X_t for $t = 1, \ldots, 57$. (In the absence of any measurement error and any leak in the tank, each x_t would be zero.) The data and their ACF are plotted in Figures 3.5 and 3.6. To check the plausibility of an MA(1) model, the bounds $\pm 1.96\left(1 + 2\hat{\rho}^2(1)\right)^{1/2}/n^{1/2}$ are also plotted in Figure 3.6. Since $\hat{\rho}(h)$ is well within these bounds for $h > 1$, the data appear to be compatible with the model

$$X_t = \mu + Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right). \tag{3.2.16}$$

The mean μ may be estimated by the sample mean $\bar{x}_{57} = -4.035$, and the parameters θ, σ^2 may be estimated by equating the sample ACVF with the model ACVF at lags 0 and 1, and solving the resulting equations for θ and σ^2. This estimation procedure is known as the method of moments, and in this case gives the equations

$$(1 + \theta^2)\sigma^2 = \hat{\gamma}(0) = 3415.72,$$
$$\theta\sigma^2 = \hat{\gamma}(1) = -1719.95.$$

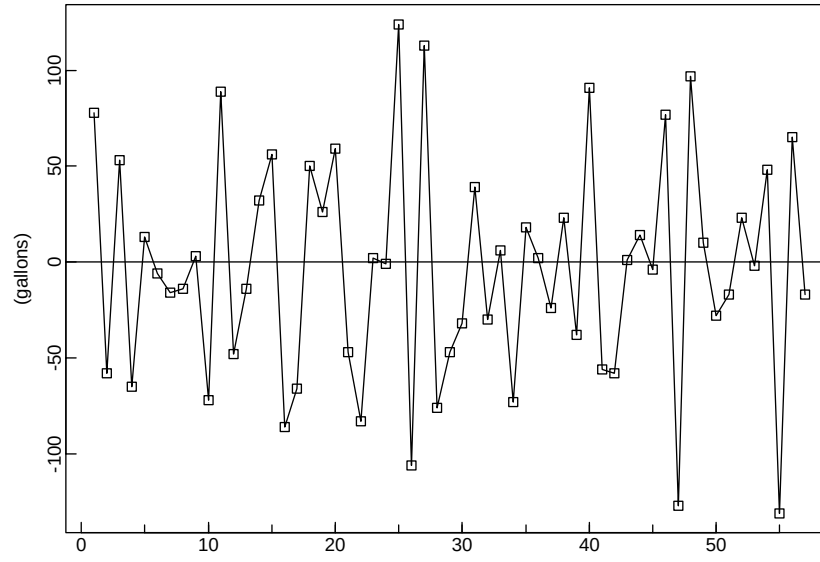

Figure 3-5
Time series of the overshorts in Example 3.2.8.

Using the approximate solution $\theta = -1$ and $\sigma^2 = 1708$, we obtain the noninvertible MA(1) model

$$X_t = -4.035 + Z_t - Z_{t-1}, \quad \{Z_t\} \sim \text{WN}(0, 1708).$$

Typically, in time series modeling we have little or no knowledge of the underlying physical mechanism generating the data, and the choice of a suitable class of models

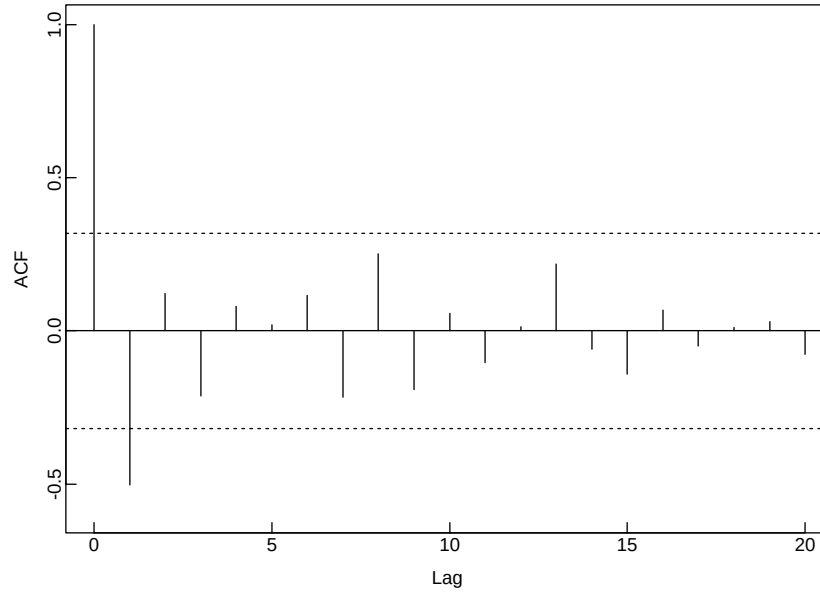

Figure 3-6
The sample ACF of the data in Figure 3.5 showing the bounds $\pm 1.96 n^{-1/2}\left(1 + 2\hat{\rho}^2(1)\right)^{1/2}$ assuming an MA(1) model for the data.

is entirely data driven. For the time series of overshorts, the data, through the graph of the ACF, lead us to the MA(1) model. Alternatively, we can attempt to model the mechanism generating the time series of overshorts using a *structural model*. As we will see, the structural model formulation leads us again to the MA(1) model. In the structural model setup, write Y_t, the observed amount of fuel in the tank at time t, as

$$Y_t = y_t^* + U_t, \tag{3.2.17}$$

where y_t^* is the true (or actual) amount of fuel in the tank at time t (not to be confused with y_t above) and U_t is the resulting *measurement error*. The variable y_t^* is an idealized quantity that in principle cannot be observed even with the most sophisticated measurement devices. Similarly, we assume that

$$A_t = a_t^* + V_t, \tag{3.2.18}$$

where a_t^* is the actual amount of fuel sold minus the actual amount delivered during day t, and V_t is the associated measurement error. We further assume that $\{U_t\} \sim \text{WN}(0, \sigma_U^2)$, $\{V_t\} \sim \text{WN}(0, \sigma_V^2)$, and that the two sequences $\{U_t\}$ and $\{V_t\}$ are uncorrelated with one another ($E(U_t V_s) = 0$ for all s and t). If the change of level per day due to leakage is μ gallons ($\mu < 0$ indicates leakage), then

$$y_t^* = \mu + y_{t-1}^* - a_t^*. \tag{3.2.19}$$

This equation relates the actual amounts of fuel in the tank at the end of days t and $t-1$, adjusted for the actual amounts that have been sold and delivered during the day. Using (3.2.17)–(3.2.19), the model for the time series of overshorts is given by

$$X_t = Y_t - Y_{t-1} + A_t = \mu + U_t - U_{t-1} + V_t.$$

This model is stationary and 1-correlated, since

$$EX_t = E(\mu + U_t - U_{t-1} + V_t) = \mu$$

and

$$\begin{aligned}\gamma(h) &= E[(X_{t+h} - \mu)(X_t - \mu)] \\ &= E[(U_{t+h} - U_{t+h-1} + V_{t+h})(U_t - U_{t-1} + V_t)] \\ &= \begin{cases} 2\sigma_U^2 + \sigma_V^2, & \text{if } h = 0, \\ -\sigma_U^2, & \text{if } |h| = 1, \\ 0, & \text{otherwise.} \end{cases}\end{aligned}$$

It follows from Proposition 2.1.1 that $\{X_t\}$ is the MA(1) model (3.2.16) with

$$\rho(1) = \frac{\theta_1}{1+\theta_1^2} = \frac{-\sigma_U^2}{2\sigma_U^2 + \sigma_V^2}.$$

From this equation we see that the measurement error associated with the adjustment $\{A_t\}$ is zero (i.e., $\sigma_V^2 = 0$) if and only if $\rho(1) = -.5$ or, equivalently, if and only

if $\theta_1 = -1$. From the analysis above, the moment estimator of θ_1 for the overshort data is in fact -1, so that we conclude that there is relatively little measurement error associated with the amount of fuel sold and delivered.

We shall return to a more general discussion of structural models in Chapter 8. □

Example 3.2.9 The sunspot numbers

Figure 3.7 shows the sample PACF of the sunspot numbers $S_1, \ldots, S_{100}$ (for the years $1770 - 1869$) as obtained from ITSM by opening the project SUNSPOTS.TSM and clicking on the second yellow button at the top of the screen. The graph also shows the bounds $\pm 1.96/\sqrt{100}$. The fact that all of the PACF values beyond lag 2 fall within the bounds suggests the possible suitability of an AR(2) model for the mean-corrected data set $X_t = S_t - 46.93$. One simple way to estimate the parameters ϕ_1, ϕ_2, and σ^2 of such a model is to require that the ACVF of the model at lags 0, 1, and 2 should match the sample ACVF at those lags. Substituting the sample ACVF values

$$\hat{\gamma}(0) = 1382.2, \quad \hat{\gamma}(1) = 1114.4, \quad \hat{\gamma}(2) = 591.73,$$

for $\gamma(0)$, $\gamma(1)$, and $\gamma(2)$ in the first three equations of (3.2.5) and (3.2.6) and solving for ϕ_1, ϕ_2, and σ^2 gives the fitted model

$$X_t - 1.318X_{t-1} + 0.634X_{t-2} = Z_t, \quad \{Z_t\} \sim \text{WN}(0, 289.2). \tag{3.2.20}$$

(This method of model fitting is called Yule–Walker estimation and will be discussed more fully in Section 5.1.1.) □

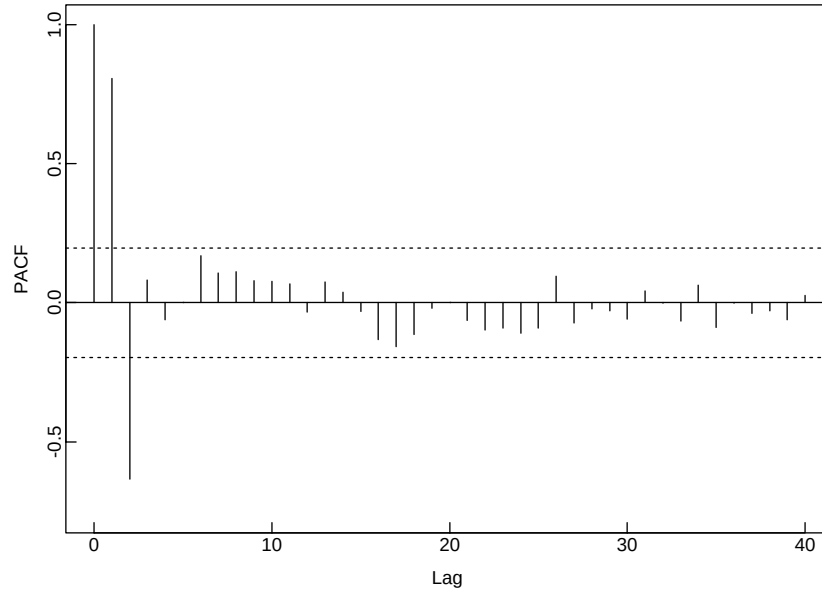

Figure 3-7
The sample PACF of the sunspot numbers with the bounds $\pm 1.96/\sqrt{100}$.

3.3 Forecasting ARMA Processes

The innovations algorithm (see Section 2.5.2) provided us with a recursive method for forecasting second-order zero-mean processes that are not necessarily stationary. For the causal ARMA process

$$\phi(B)X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right),$$

it is possible to simplify the application of the algorithm drastically. The idea is to apply it not to the process $\{X_t\}$ itself, but to the transformed process (cf. Ansley, 1979)

$$\begin{cases} W_t = \sigma^{-1}X_t, & t = 1, \ldots, m, \\ W_t = \sigma^{-1}\phi(B)X_t, & t > m, \end{cases} \tag{3.3.1}$$

where

$$m = \max(p, q). \tag{3.3.2}$$

For notational convenience we define $\theta_0 := 1$ and $\theta_j := 0$ for $j > q$. We shall also assume that $p \geq 1$ and $q \geq 1$. (There is no loss of generality in these assumptions, since in the analysis that follows we may take any of the coefficients ϕ_i and θ_i to be zero.)

The autocovariance function $\gamma_X(\cdot)$ of $\{X_t\}$ can easily be computed using any of the methods described in Section 3.2.1. The autocovariances $\kappa(i, j) = E(W_i W_j)$, $i, j \geq 1$, are then found from

$$\kappa(i, j) = \begin{cases} \sigma^{-2}\gamma_X(i-j), & 1 \leq i,\ j \leq m \\ \sigma^{-2}\left[\gamma_X(i-j) - \sum_{r=1}^{p}\phi_r\gamma_X(r-|i-j|)\right], & \min(i, j) \leq m < \max(i, j) \leq 2m, \\ \sum_{r=0}^{q}\theta_r\theta_{r+|i-j|}, & \min(i, j) > m, \\ 0, & \text{otherwise.} \end{cases} \tag{3.3.3}$$

Applying the innovations algorithm to the process $\{W_t\}$ we obtain

$$\begin{cases} \hat{W}_{n+1} = \sum_{j=1}^{n}\theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j}), & 1 \leq n < m, \\ \hat{W}_{n+1} = \sum_{j=1}^{q}\theta_{nj}(W_{n+1-j} - \hat{W}_{n+1-j}), & n \geq m, \end{cases} \tag{3.3.4}$$

where the coefficients θ_{nj} and the mean squared errors $r_n = E\left(W_{n+1} - \hat{W}_{n+1}\right)^2$ are found recursively from the innovations algorithm with κ defined as in (3.3.3). The notable feature of the predictors (3.3.4) is the vanishing of θ_{nj} when both $n \geq m$ and

$j > q$. This is a consequence of the innovations algorithm and the fact that $\kappa(r, s) = 0$ if $r > m$ and $|r - s| > q$.

Observe now that equations (3.3.1) allow each $X_n, n \geq 1$, to be written as a linear combination of W_j, $1 \leq j \leq n$, and, conversely, each W_n, $n \geq 1$, to be written as a linear combination of X_j, $1 \leq j \leq n$. This means that the best linear predictor of any random variable Y in terms of $\{1, X_1, \ldots, X_n\}$ is the same as the best linear predictor of Y in terms of $\{1, W_1, \ldots, W_n\}$. We shall denote this predictor by $P_n Y$. In particular, the one-step predictors of W_{n+1} and X_{n+1} are given by

$$\hat{W}_{n+1} = P_n W_{n+1}$$

and

$$\hat{X}_{n+1} = P_n X_{n+1}.$$

Using the linearity of P_n and equations (3.3.1) we see that

$$\begin{cases} \hat{W}_t = \sigma^{-1}\hat{X}_t, & t = 1, \ldots, m, \\ \hat{W}_t = \sigma^{-1}\left[\hat{X}_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}\right], & t > m, \end{cases} \tag{3.3.5}$$

which, together with (3.3.1), shows that

$$X_t - \hat{X}_t = \sigma\left[W_t - \hat{W}_t\right] \quad \text{for all } t \geq 1. \tag{3.3.6}$$

Replacing $\left(W_j - \hat{W}_j\right)$ by $\sigma^{-1}\left(X_j - \hat{X}_j\right)$ in (3.3.3) and then substituting into (3.3.4), we finally obtain

$$\hat{X}_{n+1} = \begin{cases} \sum_{j=1}^{n} \theta_{nj}\left(X_{n+1-j} - \hat{X}_{n+1-j}\right), & 1 \leq n < m, \\ \phi_1 X_n + \cdots + \phi_p X_{n+1-p} + \sum_{j=1}^{q} \theta_{nj}\left(X_{n+1-j} - \hat{X}_{n+1-j}\right), & n \geq m, \end{cases} \tag{3.3.7}$$

and

$$E\left(X_{n+1} - \hat{X}_{n+1}\right)^2 = \sigma^2 E\left(W_{n+1} - \hat{W}_{n+1}\right)^2 = \sigma^2 r_n, \tag{3.3.8}$$

where θ_{nj} and r_n are found from the innovations algorithm with κ as in (3.3.3). Equations (3.3.7) determine the one-step predictors $\hat{X}_2, \hat{X}_3, \ldots$ recursively.

Remark 1. It can be shown (see TSTM, Problem 5.6) that if $\{X_t\}$ is invertible, then as $n \to \infty$,

$$E\left(X_n - \hat{X}_n - Z_n\right)^2 \to 0,$$

$$\theta_{nj} \to \theta_j, \ j = 1, \ldots, q,$$

and

$$r_n \to 1.$$

Algebraic calculation of the coefficients θ_{nj} and r_n is not feasible except for very simple models, such as those considered in the following examples. However, numerical implementation of the recursions is quite straightforward and is used to compute predictors in the program ITSM. □

Example 3.3.1 Prediction of an AR(p) process

Applying (3.3.7) to the ARMA(p, 1) process with $\theta_1 = 0$, we easily find that

$$\hat{X}_{n+1} = \phi_1 X_n + \cdots + \phi_p X_{n+1-p}, \quad n \geq p. \qquad \square$$

Example 3.3.2 Prediction of an MA(q) process

Applying (3.3.7) to the ARMA(1, q) process with $\phi_1 = 0$ gives

$$\hat{X}_{n+1} = \sum_{j=1}^{\min(n,q)} \theta_{nj} \left(X_{n+1-j} - \hat{X}_{n+1-j} \right), \quad n \geq 1,$$

where the coefficients θ_{nj} are found by applying the innovations algorithm to the covariances $\kappa(i, j)$ defined in (3.3.3). Since in this case the processes $\{X_t\}$ and $\{\sigma^{-1} W_t\}$ are identical, these covariances are simply

$$\kappa(i, j) = \sigma^{-2} \gamma_X(i - j) = \sum_{r=0}^{q-|i-j|} \theta_r \theta_{r+|i-j|}. \qquad \square$$

Example 3.3.3 Prediction of an ARMA(1,1) process

If

$$X_t - \phi X_{t-1} = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right),$$

and $|\phi| < 1$, then equations (3.3.7) reduce to the single equation

$$\hat{X}_{n+1} = \phi X_n + \theta_{n1}(X_n - \hat{X}_n), \quad n \geq 1.$$

To compute θ_{n1} we first use Example 3.2.1 to find that $\gamma_X(0) = \sigma^2 \left(1 + 2\theta\phi + \theta^2\right) / \left(1 - \phi^2\right)$. Substituting in (3.3.3) then gives, for $i, j \geq 1$,

$$\kappa(i, j) = \begin{cases} \left(1 + 2\theta\phi + \theta^2\right) / \left(1 - \phi^2\right), & i = j = 1, \\ 1 + \theta^2, & i = j \geq 2, \\ \theta, & |i - j| = 1, i \geq 1, \\ 0, & \text{otherwise.} \end{cases}$$

With these values of $\kappa(i, j)$, the recursions of the innovations algorithm reduce to

$$\begin{aligned} r_0 &= \left(1 + 2\theta\phi + \theta^2\right) / \left(1 - \phi^2\right), \\ \theta_{n1} &= \theta / r_{n-1}, \\ r_n &= 1 + \theta^2 - \theta^2 / r_{n-1}, \end{aligned} \tag{3.3.9}$$

which can be solved quite explicitly (see Problem 3.13). □

Example 3.3.4 Numerical prediction of an ARMA(2,3) process

In this example we illustrate the steps involved in numerical prediction of an ARMA(2,3) process. Of course, these steps are shown for illustration only. The calculations are all carried out automatically by ITSM in the course of computing predictors for any specified data set and ARMA model. The process we shall consider is the ARMA process defined by the equations

$$X_t - X_{t-1} + 0.24X_{t-2} = Z_t + 0.4Z_{t-1} + 0.2Z_{t-2} + 0.1Z_{t-3}, \tag{3.3.10}$$

where $\{Z_t\} \sim \text{WN}(0, 1)$. Ten values of $X_1, \ldots, X_{10}$ simulated by the program ITSM are shown in Table 3.1. (These were produced using the option `Model>Specify` to specify the order and parameters of the model and then `Model>Simulate` to generate the series from the specified model.)

The first step is to compute the covariances $\gamma_X(h)$, $h = 0, 1, 2$, which are easily found from equations (3.2.5) with $k = 0, 1, 2$ to be

$$\gamma_X(0) = 7.17133, \quad \gamma_X(1) = 6.44139, \quad \text{and} \quad \gamma_X(2) = 5.0603.$$

From (3.3.3) we find that the symmetric matrix $K = [\kappa(i, j)]_{i,j=1,2,\ldots}$ is given by

$$K = \begin{bmatrix} 7.1713 & & & & & & & & \\ 6.4414 & 7.1713 & & & & & & & \\ 5.0603 & 6.4414 & 7.1713 & & & & & & \\ 0.10 & 0.34 & 0.816 & 1.21 & & & & & \\ 0 & 0.10 & 0.34 & 0.50 & 1.21 & & & & \\ 0 & 0 & 0.10 & 0.24 & 0.50 & 1.21 & & & \\ \cdot & 0 & 0 & 0.10 & 0.24 & 0.50 & 1.21 & & \\ \cdot & \cdot & 0 & 0 & 0.10 & 0.24 & 0.50 & 1.21 & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

The next step is to solve the recursions of the innovations algorithm for θ_{nj} and r_n using these values for $\kappa(i, j)$. Then

$$\hat{X}_{n+1} = \begin{cases} \displaystyle\sum_{j=1}^{n} \theta_{nj} \left(X_{n+1-j} - \hat{X}_{n+1-j}\right), & n = 1, 2, \\ \\ \displaystyle X_n - 0.24X_{n-1} + \sum_{j=1}^{3} \theta_{nj} \left(X_{n+1-j} - \hat{X}_{n+1-j}\right), & n = 3, 4, \ldots, \end{cases}$$

and

$$E\left(X_{n+1} - \hat{X}_{n+1}\right)^2 = \sigma^2 r_n = r_n.$$

The results are shown in Table 3.1. □

h-Step Prediction of an ARMA(p, q) Process

As in Section 2.5, we use $P_n Y$ to denote the best linear predictor of Y in terms of $X_1, \ldots, X_n$ (which, as pointed out after (3.3.4), is the same as the best linear predictor of Y in terms of $W_1, \ldots, W_n$). Then from (2.5.30) we have

$$P_n W_{n+h} = \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}\left(W_{n+h-j} - \hat{W}_{n+h-j}\right) = \sigma^2 \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}\left(X_{n+h-j} - \hat{X}_{n+h-j}\right).$$

Using this result and applying the operator P_n to each side of equations (3.3.1), we conclude that the h-step predictors $P_n X_{n+h}$ satisfy

$$P_n X_{n+h} = \begin{cases} \displaystyle\sum_{j=h}^{n+h-1} \theta_{n+h-1,j}\left(X_{n+h-j} - \hat{X}_{n+h-j}\right), & 1 \le h \le m-n, \\ \displaystyle\sum_{i=1}^{p} \phi_i P_n X_{n+h-i} + \sum_{j=h}^{n+h-1} \theta_{n+h-1,j}\left(X_{n+h-j} - \hat{X}_{n+h-j}\right), & h > m-n. \end{cases} \tag{3.3.11}$$

If, as is almost always the case, $n > m = \max(p, q)$, then for all $h \ge 1$,

$$P_n X_{n+h} = \sum_{i=1}^{p} \phi_i P_n X_{n+h-i} + \sum_{j=h}^{q} \theta_{n+h-1,j}\left(X_{n+h-j} - \hat{X}_{n+h-j}\right). \tag{3.3.12}$$

Once the predictors $\hat{X}_1, \ldots \hat{X}_n$ have been computed from (3.3.7), it is a straightforward calculation, with n fixed, to determine the predictors $P_n X_{n+1}$, $P_n X_{n+2}$, $P_n X_{n+3}, \ldots$

Table 3.1 $\hat{X}_{n+1}$ for the ARMA(2,3) Process of Example 3.3.4.

n	X_{n+1}	r_n	θ_{n1}	θ_{n2}	θ_{n3}	$\hat{X}_{n+1}$
0	1.704	7.1713	0			
1	0.527	1.3856	0.8982	1.5305		
2	1.041	1.0057	1.3685	0.7056	−0.1710	
3	0.942	1.0019	0.4008	0.1806	0.0139	1.2428
4	0.555	1.0019	0.3998	0.2020	0.0732	0.7443
5	−1.002	1.0005	0.3992	0.1995	0.0994	0.3138
6	−0.585	1.0000	0.4000	0.1997	0.0998	−1.7293
7	0.010	1.0000	0.4000	0.2000	0.0998	−0.1688
8	−0.638	1.0000	0.4000	0.2000	0.0999	0.3193
9	0.525	1.0000	0.4000	0.2000	0.1000	−0.8731
10	1.0000	0.4000	0.2000	0.1000	1.0638	
11	1.0000	0.4000	0.2000	0.1000		
12	1.0000	0.4000	0.2000	0.1000		

recursively from (3.3.12) (or (3.3.11) if $n \leq m$). The calculations are performed automatically in the `Forecasting>ARMA` option of the program ITSM.

Example 3.3.5 *h*-step prediction of an ARMA(2,3) process

To compute h-step predictors, $h = 1, \ldots, 10$, for the data of Example 3.3.4 and the model (3.3.10), open the project E334.TSM in ITSM and enter the model using the option `Model>Specify`. Then select `Forecasting>ARMA` and specify 10 for the number of forecasts required. You will notice that the white noise variance is automatically set by ITSM to an estimate based on the sample. To retain the model value of 1, you must reset the white noise variance to this value. Then click `OK` and you will see a graph of the original series with the ten predicted values appended. If you right-click on the graph and select `Info`, you will see the numerical results shown in the following table as well as prediction bounds based on the assumption that the series is Gaussian. (Prediction bounds are discussed in the last paragraph of this chapter.) The mean squared errors are calculated as described below. Notice how the predictors converge fairly rapidly to the mean of the process (i.e., zero) as the lead time h increases. Correspondingly, the one-step mean squared error increases from the white noise variance (i.e., 1) at $h = 1$ to the variance of X_t (i.e., 7.1713), which is virtually reached at $h = 10$. □

The Mean Squared Error of $P_n X_{n+h}$

The mean squared error of $P_n X_{n+h}$ is easily computed by ITSM from the formula

$$\sigma_n^2(h) := E(X_{n+h} - P_n X_{n+h})^2 = \sum_{j=0}^{h-1} \left(\sum_{r=0}^{j} \chi_r \theta_{n+h-r-1,j-r} \right)^2 v_{n+h-j-1}, \quad (3.3.13)$$

Table 3.2 *h*-step predictors for the ARMA(2,3) Series of Example 3.3.4.

h	$P_{10}X_{10+h}$	$\sqrt{MSE}$
1	1.0638	1.0000
2	1.1217	1.7205
3	1.0062	2.1931
4	0.7370	2.4643
5	0.4955	2.5902
6	0.3186	2.6434
7	0.1997	2.6648
8	0.1232	2.6730
9	0.0753	2.6761
10	0.0457	2.6773

where the coefficients χ_j are computed recursively from the equations $\chi_0 = 1$ and

$$\chi_j = \sum_{k=1}^{\min(p,\,j)} \phi_k \chi_{j-k}, \quad j = 1, 2, \ldots. \tag{3.3.14}$$

Example 3.3.6 *h*-step prediction of an ARMA(2,3) process

We now illustrate the use of (3.3.12) and (3.3.13) for the h-step predictors and their mean squared errors by manually reproducing the output of ITSM shown in Table 3.2. From (3.3.12) and Table 3.1 we obtain

$$\begin{aligned} P_{10}X_{12} &= \sum_{i=1}^{2} \phi_i P_{10}X_{12-i} + \sum_{j=2}^{3} \theta_{11,j}\left(X_{12-j} - \hat{X}_{12-j}\right) \\ &= \phi_1 \hat{X}_{11} + \phi_2 X_{10} + 0.2\left(X_{10} - \hat{X}_{10}\right) + 0.1\left(X_9 - \hat{X}_9\right) \\ &= 1.1217 \end{aligned}$$

and

$$\begin{aligned} P_{10}X_{13} &= \sum_{i=1}^{2} \phi_i P_{10}X_{13-i} + \sum_{j=3}^{3} \theta_{12,j}\left(X_{13-j} - \hat{X}_{13-j}\right) \\ &= \phi_1 P_{10}X_{12} + \phi_2 \hat{X}_{11} + 0.1\left(X_{10} - \hat{X}_{10}\right) \\ &= 1.0062. \end{aligned}$$

For $k > 13$, $P_{10}X_k$ is easily found recursively from

$$P_{10}X_k = \phi_1 P_{10}X_{k-1} + \phi_2 P_{10}X_{k-2}.$$

To find the mean squared errors we use (3.3.13) with $\chi_0 = 1$, $\chi_1 = \phi_1 = 1$, and $\chi_2 = \phi_1\chi_1 + \phi_2 = 0.76$. Using the values of θ_{nj} and $v_j (= r_j)$ in Table 3.1, we obtain

$$\sigma_{10}^2(2) = E(X_{12} - P_{10}X_{12})^2 = 2.960$$

and

$$\sigma_{10}^2(3) = E(X_{13} - P_{10}X_{13})^2 = 4.810,$$

in accordance with the results shown in Table 3.2. □

Large-Sample Approximations

Assuming as usual that the ARMA(p, q) process defined by $\phi(B)X_t = \theta(B)Z_t$, $\{Z_t\} \sim \text{WN}\left(0, \sigma^2\right)$, is causal and invertible, we have the representations

$$X_{n+h} = \sum_{j=0}^{\infty} \psi_j Z_{n+h-j} \tag{3.3.15}$$

and

$$Z_{n+h} = X_{n+h} + \sum_{j=1}^{\infty} \pi_j X_{n+h-j}, \tag{3.3.16}$$

where $\{\psi_j\}$ and $\{\pi_j\}$ are uniquely determined by equations (3.1.7) and (3.1.8), respectively. Let $\tilde{P}_n Y$ denote the best (i.e., minimum mean squared error) approximation to Y that is a linear combination or limit of linear combinations of X_t, $-\infty < t \le n$, or equivalently (by (3.3.15) and (3.3.16)) of Z_t, $-\infty < t \le n$. The properties of the operator $\tilde{P}_n$ were discussed in Section 2.5.3. Applying $\tilde{P}_n$ to each side of equations (3.3.15) and (3.3.16) gives

$$\tilde{P}_n X_{n+h} = \sum_{j=h}^{\infty} \psi_j Z_{n+h-j} \tag{3.3.17}$$

and

$$\tilde{P}_n X_{n+h} = -\sum_{j=1}^{\infty} \pi_j \tilde{P}_n X_{n+h-j}. \tag{3.3.18}$$

For $h = 1$ the jth term on the right of (3.3.18) is just X_{n+1-j}. Once $\tilde{P}_n X_{n+1}$ has been evaluated, $\tilde{P}_n X_{n+2}$ can then be computed from (3.3.18). The predictors $\tilde{P}_n X_{n+3}$, $\tilde{P}_n X_{n+4}, \ldots$ can then be computed successively in the same way. Subtracting (3.3.17) from (3.3.15) gives the h-step prediction error as

$$X_{n+h} - \tilde{P}_n X_{n+h} = \sum_{j=0}^{h-1} \psi_j Z_{n+h-j},$$

from which we see that the mean squared error is

$$\tilde{\sigma}^2(h) = \sigma^2 \sum_{j=0}^{h-1} \psi_j^2. \tag{3.3.19}$$

The predictors obtained in this way have the form

$$\tilde{P}_n X_{n+h} = \sum_{j=0}^{\infty} c_j X_{n-j}. \tag{3.3.20}$$

In practice, of course, we have only observations $X_1, \ldots, X_n$ available, so we must truncate the series (3.3.20) after n terms. The resulting predictor is a useful approximation to $P_n X_{n+h}$ if n is large and the coefficients c_j converge to zero rapidly as j increases. It can be shown that the mean squared error (3.3.19) of $\tilde{P}_n X_{n+h}$ can also be obtained by letting $n \to \infty$ in the expression (3.3.13) for the mean squared error of $P_n X_{n+h}$, so that $\tilde{\sigma}^2(h)$ is an easily calculated approximation to $\sigma_n^2(h)$ for large n.

Prediction Bounds for Gaussian Processes

If the ARMA process $\{X_t\}$ is driven by Gaussian white noise (i.e., if $\{Z_t\} \sim \text{IID N}(0, \sigma^2)$), then for each $h \geq 1$ the prediction error $X_{n+h} - P_n X_{n+h}$ is normally distributed with mean 0 and variance $\sigma_n^2(h)$ given by (3.3.19).

Consequently, if $\Phi_{1-\alpha/2}$ denotes the $(1-\alpha/2)$ quantile of the standard normal distribution function, it follows that X_{n+h} lies between the bounds $P_n X_{n+h} \pm \Phi_{1-\alpha/2}\sigma_n(h)$ with probability $(1-\alpha)$. These bounds are therefore called $(1-\alpha)$ prediction bounds for X_{n+h}.

Problems

3.1. Determine which of the following ARMA processes are causal and which of them are invertible. (In each case $\{Z_t\}$ denotes white noise.)

a. $X_t + 0.2X_{t-1} - 0.48X_{t-2} = Z_t$.

b. $X_t + 1.9X_{t-1} + 0.88X_{t-2} = Z_t + 0.2Z_{t-1} + 0.7Z_{t-2}$.

c. $X_t + 0.6X_{t-1} = Z_t + 1.2Z_{t-1}$.

d. $X_t + 1.8X_{t-1} + 0.81X_{t-2} = Z_t$.

e. $X_t + 1.6X_{t-1} = Z_t - 0.4Z_{t-1} + 0.04Z_{t-2}$.

3.2. For those processes in Problem 3.1 that are causal, compute and graph their ACF and PACF using the program ITSM.

3.3. For those processes in Problem 3.1 that are causal, compute the first six coefficients $\psi_0, \psi_1, \ldots, \psi_5$ in the causal representation $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ of $\{X_t\}$.

3.4. Compute the ACF and PACF of the AR(2) process

$$X_t = .8X_{t-2} + Z_t, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right).$$

3.5. Let $\{Y_t\}$ be the ARMA plus noise time series defined by

$$Y_t = X_t + W_t,$$

where $\{W_t\} \sim \text{WN}\left(0, \sigma_w^2\right)$, $\{X_t\}$ is the ARMA(p, q) process satisfying

$$\phi(B)X_t = \theta(B)Z_t, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma_z^2\right),$$

and $E(W_s Z_t) = 0$ for all s and t.

a. Show that $\{Y_t\}$ is stationary and find its autocovariance in terms of σ_W^2 and the ACVF of $\{X_t\}$.

b. Show that the process $U_t := \phi(B)Y_t$ is r-correlated, where $r = \max(p, q)$ and hence, by Proposition 2.1.1, is an MA(r) process. Conclude that $\{Y_t\}$ is an ARMA(p, r) process.

3.6. Show that the two MA(1) processes

$$X_t = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right)$$

$$Y_t = \tilde{Z}_t + \frac{1}{\theta}\tilde{Z}_{t-1}, \quad \{\tilde{Z}_t\} \sim \text{WN}\left(0, \sigma^2\theta^2\right),$$

where $0 < |\theta| < 1$, have the same autocovariance functions.

3.7. Suppose that $\{X_t\}$ is the noninvertible MA(1) process

$$X_t = Z_t + \theta Z_{t-1}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right),$$

where $|\theta| > 1$. Define a new process $\{W_t\}$ as

$$W_t = \sum_{j=0}^{\infty} (-\theta)^{-j} X_{t-j}$$

and show that $\{W_t\} \sim \text{WN}\left(0, \sigma_W^2\right)$. Express σ_W^2 in terms of θ and σ^2 and show that $\{X_t\}$ has the *invertible* representation (in terms of $\{W_t\}$)

$$X_t = W_t + \frac{1}{\theta} W_{t-1}.$$

3.8. Let $\{X_t\}$ denote the unique stationary solution of the autoregressive equations

$$X_t = \phi X_{t-1} + Z_t, \quad t = 0, \pm 1, \ldots,$$

where $\{Z_t\} \sim \text{WN}\left(0, \sigma^2\right)$ and $|\phi| > 1$. Then X_t is given by the expression (2.2.11). Define the new sequence

$$W_t = X_t - \frac{1}{\phi} X_{t-1},$$

show that $\{W_t\} \sim \text{WN}\left(0, \sigma_W^2\right)$, and express σ_W^2 in terms of σ^2 and ϕ. These calculations show that $\{X_t\}$ is the (unique stationary) solution of the *causal* AR equations

$$X_t = \frac{1}{\phi} X_{t-1} + W_t, \quad t = 0, \pm 1, \ldots.$$

3.9. a. Calculate the autocovariance function $\gamma(\cdot)$ of the stationary time series

$$Y_t = \mu + Z_t + \theta_1 Z_{t-1} + \theta_{12} Z_{t-12}, \quad \{Z_t\} \sim \text{WN}\left(0, \sigma^2\right).$$

b. Use the program ITSM to compute the sample mean and sample autocovariances $\hat{\gamma}(h)$, $0 \le h \le 20$, of $\{\nabla\nabla_{12} X_t\}$, where $\{X_t, t = 1, \ldots, 72\}$ is the accidental deaths series DEATHS.TSM of Example 1.1.3.

c. By equating $\hat{\gamma}(1)$, $\hat{\gamma}(11)$, and $\hat{\gamma}(12)$ from part (b) to $\gamma(1)$, $\gamma(11)$, and $\gamma(12)$, respectively, from part (a), find a model of the form defined in (a) to represent $\{\nabla\nabla_{12} X_t\}$.

3.10. By matching the autocovariances and sample autocovariances at lags 0 and 1, fit a model of the form

$$X_t - \mu = \phi(X_{t-1} - \mu) + Z_t, \quad \{Z_t\} \sim \text{WN}(0, \sigma^2),$$

to the data STRIKES.TSM of Example 1.1.6. Use the fitted model to compute the best predictor of the number of strikes in 1981. Estimate the mean squared error of your predictor and construct 95% prediction bounds for the number of strikes in 1981 assuming that $\{Z_t\} \sim \text{iid N}(0, \sigma^2)$.

3.11. Show that the value at lag 2 of the partial ACF of the MA(1) process

$$X_t = Z_t + \theta Z_{t-1}, \quad t = 0, \pm 1, \ldots,$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, is

$$\alpha(2) = -\theta^2/\left(1 + \theta^2 + \theta^4\right).$$

3.12. For the MA(1) process of Problem 3.11, the best linear predictor of X_{n+1} based on $X_1, \ldots, X_n$ is

$$\hat{X}_{n+1} = \phi_{n,1}X_n + \cdots + \phi_{n,n}X_1,$$

where $\boldsymbol{\phi}_n = \left(\phi_{n1}, \ldots, \phi_{nn}\right)'$ satisfies $R_n\boldsymbol{\phi}_n = \boldsymbol{\rho}_n$ (equation (2.5.23)). By substituting the appropriate correlations into R_n and $\boldsymbol{\rho}_n$ and solving the resulting equations (starting with the last and working up), show that for $1 \le j < n$, $\phi_{n,n-j} = (-\theta)^{-j}\left(1 + \theta^2 + \cdots + \theta^{2j}\right)\phi_{nn}$ and hence that the PACF $\alpha(n) := \phi_{nn} = -(-\theta)^n/\left(1 + \theta^2 + \cdots + \theta^{2n}\right)$.

3.13. The coefficients θ_{nj} and one-step mean squared errors $v_n = r_n\sigma^2$ for the general causal ARMA(1,1) process in Example 3.3.3 can be found as follows:

a. Show that if $y_n := r_n/(r_n - 1)$, then the last of equations (3.3.9) can be rewritten in the form

$$y_n = \theta^{-2}y_{n-1} + 1, \quad n \ge 1.$$

b. Deduce that $y_n = \theta^{-2n}y_0 + \sum_{j=1}^{n}\theta^{-2(j-1)}$ and hence determine r_n and θ_{n1}, $n = 1, 2, \ldots$.

c. Evaluate the limits as $n \to \infty$ of r_n and θ_{n1} in the two cases $|\theta| < 1$ and $|\theta| \ge 1$.