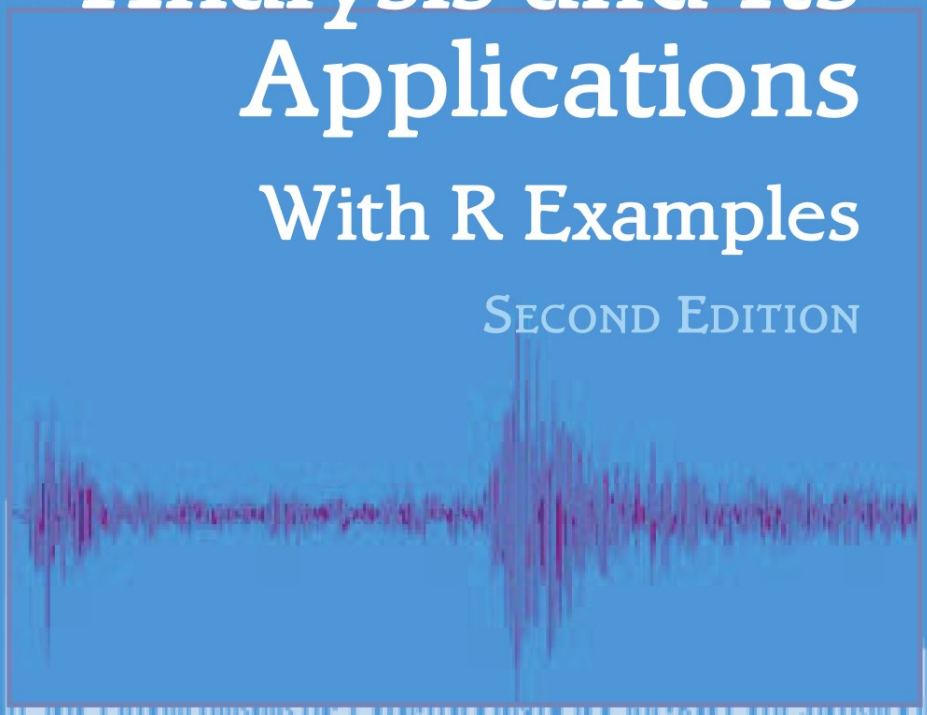
SPRINGER TEXTS IN STATISTICS

Time Series Analysis and Its Applications

With R Examples

SECOND EDITION



Robert H. Shumway
David S. Stoffer

Springer

Springer Texts in Statistics

Advisors:
George Casella Stephen Fienberg Ingram Olkin

Springer Texts in Statistics

Alfred: Elements of Statistics for the Life and Social Sciences
Berger: An Introduction to Probability and Stochastic Processes
Bilodeau and Brenner: Theory of Multivariate Statistics
Blom: Probability and Statistics: Theory and Applications
Brockwell and Davis: Introduction to Times Series and Forecasting, Second Edition
Carmona: Statistical Analysis of Financial Data in S-Plus
Chow and Teicher: Probability Theory: Independence, Interchangeability, Martingales, Third Edition
Christensen: Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data—Nonparametric Regression and Response Surface Maximization, Second Edition
Christensen: Log-Linear Models and Logistic Regression, Second Edition
Christensen: Plane Answers to Complex Questions: The Theory of Linear Models, Third Edition
Creighton: A First Course in Probability Models and Statistical Inference
Davis: Statistical Methods for the Analysis of Repeated Measurements
Dean and Voss: Design and Analysis of Experiments
du Toit, Steyn, and Stumpf: Graphical Exploratory Data Analysis
Durrett: Essentials of Stochastic Processes
Edwards: Introduction to Graphical Modelling, Second Edition
Finkelstein and Levin: Statistics for Lawyers
Flury: A First Course in Multivariate Statistics
Gut: Probability: A Graduate Course
Heiberger and Holland: Statistical Analysis and Data Display: An Intermediate Course with Examples in S-PLUS, R, and SAS
Jobson: Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
Jobson: Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods
Kalbfleisch: Probability and Statistical Inference, Volume I: Probability, Second Edition
Kalbfleisch: Probability and Statistical Inference, Volume II: Statistical Inference, Second Edition
Karr: Probability
Keyfitz: Applied Mathematical Demography, Second Edition
Kiefer: Introduction to Statistical Inference
Kokoska and Nevison: Statistical Tables and Formulae
Kulkarni: Modeling, Analysis, Design, and Control of Stochastic Systems
Lange: Applied Probability
Lange: Optimization
Lehmann: Elements of Large-Sample Theory

(continued after index)

Robert H. Shumway
David S. Stoffer

Time Series Analysis and Its Applications

With R Examples

Second Edition

With 160 Illustrations

Springer

Robert H. Shumway
Department of Statistics
University of California, Davis
Davis, CA 95616
USA
rshumway@ucdavis.edu
or
shumway@wald.ucdavis.edu

David S. Stoffer
Department of Statistics
University of Pittsburgh
Pittsburgh, PA 15260
USA
stoffer@pitt.edu

Library of Congress Control Number: 2005935284

ISBN-10: 0-387-29317-5
ISBN-13: 978-0387-29317-2

Printed on acid-free paper.

Printed in the United States of America. (MVY)

9 8 7 6 5 4 3 2 1

springer.com

To my wife, Ruth, for her support and joie de vivre, and to the memory of my thesis adviser, Solomon Kullback.
R.H.S.

To my family, who constantly remind me what is important.
D.S.S.

Preface to the Second Edition

The second edition marks a substantial change to the first edition. Perhaps the most significant change is the introduction of examples based on the freeware R package. The package, which runs on most operating systems, can be downloaded from The Comprehensive R Archive Network (CRAN) at `http://cran.r-project.org/` or any one of its mirrors. Readers who have experience with the S-PLUS® package will have no problem working with R. For novices, R installs some help manuals, and CRAN supplies links to contributed tutorials such as *R for Beginners*. In our examples, we assume the reader has downloaded and installed R and has downloaded the necessary data files. The data files can be downloaded from the website for the text, `http://www.stat.pitt.edu/stoffer/tsa2/` or any one of its mirrors. We will also provide additional code and other information of interest on the text's website. Most of the material that would be given in an introductory course on time series analysis has associated R code. Although examples are given in R, the material is not R-dependent. In courses we have given using a preliminary version of the new edition of the text, students were allowed to use any package of preference. Although most students used R (or S-PLUS), a number of them completed the course successfully using other programs such as ASTSA, MATLAB®, SAS®, and SPSS®.

Another substantial change from the first edition is that the material has been divided into smaller chapters. The introductory material is now contained in the first two chapters. The first chapter discusses the characteristics of time series, introducing the fundamental concepts of time plot, models for dependent data, auto- and cross-correlation, and their estimation. The second chapter provides a background in regression techniques for time series data. This chapter also includes the related topics of smoothing and exploratory data analysis for preprocessing nonstationary series.

In the first edition, we covered ARIMA and other special topics in the time domain in one chapter. In this edition, univariate ARIMA modeling is presented in its own chapter, Chapter 3. The material on additional time domain topics has been expanded, and moved to its own chapter, Chapter 5. The additional topics include long memory models, GARCH processes, threshold models, regression with autocorrelated errors, lagged regression, transfer function modeling, and multivariate ARMAX models. In this edition, we have removed the discussion on reduced rank models and contemporaneous models from the multivariate ARMAX section. The coverage of GARCH models has been considerably expanded in this edition. The coverage of long memory models has been consolidated, presenting time domain and frequency domain approaches in the same section. For this reason, the chapter is presented after the chapter on spectral analysis.

The chapter on spectral analysis and filtering, Chapter 4, has been expanded to include various types of spectral estimators. In particular, kernel based estimators and spectral window estimators have been included in the dis-

cussion. The chapter now includes a section on wavelets that was in another chapter in the first edition. The reader will also notice a change in notation from the previous edition.

In the first edition, topics were supplemented by theoretical sections at the end of the chapter. In this edition, we have put the theoretical topics in appendices at the end of the text. In particular, Appendix A can be used to supplement the material in the first chapter; it covers some fundamental topics in large sample theory for dependent data. The material in Appendix B includes theoretical material that expands the presentation of time domain topics, and this appendix may be used to supplement the coverage of the chapter on time series regression and the chapter on ARIMA models. Finally, Appendix C contains a theoretical basis for spectral analysis.

The remaining two chapters on state-space and dynamic linear models, Chapter 6, and on additional statistical methods in the frequency domain, Chapter 7, are comparable to the their first edition counterparts. We do mention that the section on multivariate ARMAX, which used to be in the state-space chapter, has been moved to Chapter 5. We have also removed spectral domain canonical correlation analysis and the discussion on wavelets (now in Chapter 4) that were previously in Chapter 7. The material on stochastic volatility models, now in Chapter 6, has been expanded. R programs for some Chapter 6 examples are available on the website for the text; these programs include code for the Kalman filter and smoother, maximum likelihood estimation, the EM algorithm, and fitting stochastic volatility models.

In the previous edition, we set off important definitions by highlighting phrases corresponding to the definition. We believe this practice made it difficult for readers to find important information. In this edition, we have set off definitions as numbered definitions that are presented in italics with the concept being defined in bold letters.

We thank John Kimmel, Executive Editor, Statistics, for his guidance in the preparation and production of this edition of the text. We are particularly grateful to Don Percival and Mike Keim at the University of Washington, for numerous suggestions that led to substantial improvement to the presentation. We also thank the many students and other readers who took the time to mention typographical errors and other corrections to the first edition. In particular, we appreciate the efforts of Jeongeun Kim, Sangdae Han, and Mark Gamalo at the University of Pittsburgh, and Joshua Kerr and Bo Zhou at the University of California, for providing comments on portions of the draft of this edition. Finally, we acknowledge the support of the National Science Foundation.

Robert H. Shumway
Davis, CA
David S. Stoffer
Pittsburgh, PA
August 2005

Preface to the First Edition

The goals of this book are to develop an appreciation for the richness and versatility of modern time series analysis as a tool for analyzing data, and still maintain a commitment to theoretical integrity, as exemplified by the seminal works of Brillinger (1981) and Hannan (1970) and the texts by Brockwell and Davis (1991) and Fuller (1995). The advent of more powerful computing, especially in the last three years, has provided both real data and new software that can take one considerably beyond the fitting of simple time domain models, such as have been elegantly described in the landmark work of Box and Jenkins (see Box et al., 1994). This book is designed to be useful as a text for courses in time series on several different levels and as a reference work for practitioners facing the analysis of time-correlated data in the physical, biological, and social sciences.

We believe the book will be useful as a text at both the undergraduate and graduate levels. An undergraduate course can be accessible to students with a background in regression analysis and might include Sections 1.1-1.8, 2.1-2.9, and 3.1-3.8. Similar courses have been taught at the University of California (Berkeley and Davis) in the past using the earlier book on applied time series analysis by Shumway (1988). Such a course is taken by undergraduate students in mathematics, economics, and statistics and attracts graduate students from the agricultural, biological, and environmental sciences. At the master's degree level, it can be useful to students in mathematics, environmental science, economics, statistics, and engineering by adding Sections 1.9, 2.10-2.14, 3.9, 3.10, 4.1-4.5, to those proposed above. Finally, a two-semester upper-level graduate course for mathematics, statistics and engineering graduate students can be crafted by adding selected theoretical sections from the last sections of Chapters 1, 2, and 3 for mathematics and statistics students and some advanced applications from Chapters 4 and 5. For the upper-level graduate course, we should mention that we are striving for a less rigorous level of coverage than that which is attained by Brockwell and Davis (1991), the classic entry at this level.

A useful feature of the presentation is the inclusion of data illustrating the richness of potential applications to medicine and in the biological, physical, and social sciences. We include data analysis in both the text examples and in the problem sets. All data sets are posted on the World Wide Web at the following URLs: `http://www.stat.ucdavis.edu/~shumway/tsa.html` and `http://www.stat.pitt.edu/~stoffer/tsa.html`, making them easily accessible to students and general researchers. In addition, an exploratory data analysis program written by McQuarrie and Shumway (1994) can be downloaded (as Freeware) from these websites to provide easy access to all of the techniques required for courses through the master's level.

Advances in modern computing have made multivariate techniques in the time and frequency domain, anticipated by the theoretical developments in Brillinger (1981) and Hannan (1970), routinely accessible using higher level

languages, such as MATLAB and S-PLUS. Extremely large data sets driven by periodic phenomena, such as the functional magnetic resonance imaging series or the earthquake and explosion data, can now be handled using extensions to time series of classical methods, like multivariate regression, analysis of variance, principal components, factor analysis, and discriminant or cluster analysis. Chapters 4 and 5 illustrate some of the immense potential that methods have for analyzing high-dimensional data sets.

The many practical data sets are the results of collaborations with research workers in the medical, physical, and biological sciences. Some deserve special mention as a result of the pervasive use we have made of them in the text. The predominance of applications in seismology and geophysics is joint work of the first author with Dr. Robert R. Blandford of the Center for Monitoring Research and Dr. Zoltan Der of Ensco, Inc. We have also made extensive use of the El Niño and Recruitment series contributed by Dr. Roy Mendelssohn of the National Marine Fisheries Service. In addition, Professor Nancy Day of the University of Pittsburgh provided the data used in Chapter 4 in a longitudinal analysis of the effects of prenatal smoking on growth, as well as some of the categorical sleep-state data posted on the World Wide Web. A large magnetic imaging data set that was developed during joint research on pain perception with Dr. Elizabeth Disbrow of the University of San Francisco Medical Center forms the basis for illustrating a number of multivariate techniques in Chapter 5. We are especially indebted to Professor Allan D.R. McQuarrie of the University of North Dakota, who incorporated subroutines in Shumway (1988) into ASTSA for Windows.

Finally, we are grateful to John Kimmel, Executive Editor, Statistics, for his patience, enthusiasm, and encouragement in guiding the preparation and production of this book. Three anonymous reviewers made numerous helpful comments, and Dr. Rahman Azari and Dr. Mitchell Watnik of the University of California, Davis, Division of Statistics, read portions of the draft. Any remaining errors are solely our responsibility.

Robert H. Shumway
Davis, CA
David S. Stoffer
Pittsburgh, PA
August 1999

Contents

Chapter 1

Characteristics of Time Series

1.1 Introduction

The analysis of experimental data that have been observed at different points in time leads to new and unique problems in statistical modeling and inference. The obvious correlation introduced by the sampling of adjacent points in time can severely restrict the applicability of the many conventional statistical methods traditionally dependent on the assumption that these adjacent observations are independent and identically distributed. The systematic approach by which one goes about answering the mathematical and statistical questions posed by these time correlations is commonly referred to as time series analysis.

The impact of time series analysis on scientific applications can be partially documented by producing an abbreviated listing of the diverse fields in which important time series problems may arise. For example, many familiar time series occur in the field of economics, where we are continually exposed to daily stock market quotations or monthly unemployment figures. Social scientists follow populations series, such as birthrates or school enrollments. An epidemiologist might be interested in the number of influenza cases observed over some time period. In medicine, blood pressure measurements traced over time could be useful for evaluating drugs used in treating hypertension. Functional magnetic resonance imaging of brain-wave time series patterns might be used to study how the brain reacts to certain stimuli under various experimental conditions.

Many of the most intensive and sophisticated applications of time series methods have been to problems in the physical and environmental sciences. This fact accounts for the basic engineering flavor permeating the language of

time series analysis. One of the earliest recorded series is the monthly sunspot numbers studied by Schuster (1906). More modern investigations may center on whether a warming is present in global temperature measurements or whether levels of pollution may influence daily mortality in Los Angeles. The modeling of speech series is an important problem related to the efficient transmission of voice recordings. Common features in a time series characteristic known as the power spectrum are used to help computers recognize and translate speech. Geophysical time series such those produced by yearly depositions of various kinds can provide long-range proxies for temperature and rainfall. Seismic recordings can aid in mapping fault lines or in distinguishing between earthquakes and nuclear explosions.

The above series are only examples of experimental databases that can be used to illustrate the process by which classical statistical methodology can be applied in the correlated time series framework. In our view, the first step in any time series investigation always involves careful scrutiny of the recorded data plotted over time. This scrutiny often suggests the method of analysis as well as statistics that will be of use in summarizing the information in the data. Before looking more closely at the particular statistical methods, it is appropriate to mention that two separate, but not necessarily mutually exclusive, approaches to time series analysis exist, commonly identified as the time domain approach and the frequency domain approach.

The time domain approach is generally motivated by the presumption that correlation between adjacent points in time is best explained in terms of a dependence of the current value on past values. The time domain approach focuses on modeling some future value of a time series as a parametric function of the current and past values. In this scenario, we begin with linear regressions of the present value of a time series on its own past values and on the past values of other series. This modeling leads one to use the results of the time domain approach as a forecasting tool and is particularly popular with economists for this reason.

One approach, advocated in the landmark work of Box and Jenkins (1970; see also Box et al., 1994), develops a systematic class of models called autoregressive integrated moving average (ARIMA) models to handle time-correlated modeling and forecasting. The approach includes a provision for treating more than one input series through multivariate ARIMA or through transfer function modeling. The defining feature of these models is that they are multiplicative models, meaning that the observed data are assumed to result from products of factors involving differential or difference equation operators responding to a white noise input.

A more recent approach to the same problem uses additive models more familiar to statisticians. In this approach, the observed data are assumed to result from sums of series, each with a specified time series structure; for example, in economics, assume a series is generated as the sum of trend, a seasonal effect, and error. The state-space model that results is then treated by making judicious use of the celebrated Kalman filters and smoothers, developed origi-

nally for estimation and control in space applications. Two relatively complete presentations from this point of view are in Harvey (1991) and Kitagawa and Gersch (1996). Time series regression is introduced in Chapter 2, and ARIMA and related time domain models are studied in Chapter 3, with the emphasis on classical, statistical, univariate linear regression. Special topics on time domain analysis are covered in Chapter 5; these topics include modern treatments of, for example, time series with long memory and GARCH models for the analysis of volatility. The state-space model, Kalman filtering and smoothing, and related topics are developed in Chapter 5.

Conversely, the frequency domain approach assumes the primary characteristics of interest in time series analyses relate to periodic or systematic sinusoidal variations found naturally in most data. These periodic variations are often caused by biological, physical, or environmental phenomena of interest. A series of periodic shocks may influence certain areas of the brain; wind may affect vibrations on an airplane wing; sea surface temperatures caused by El Niño oscillations may affect the number of fish in the ocean. The study of periodicity extends to economics and social sciences, where one may be interested in yearly periodicities in such series as monthly unemployment or monthly birth rates.

In spectral analysis, the partition of the various kinds of periodic variation in a time series is accomplished by evaluating separately the variance associated with each periodicity of interest. This variance profile over frequency is called the power spectrum. In our view, no schism divides time domain and frequency domain methodology, although cliques are often formed that advocate primarily one or the other of the approaches to analyzing data. In many cases, the two approaches may produce similar answers for long series, but the comparative performance over short samples is better done in the time domain. In some cases, the frequency domain formulation simply provides a convenient means for carrying out what is conceptually a time domain calculation. Hopefully, this book will demonstrate that the best path to analyzing many data sets is to use the two approaches in a complementary fashion. Expositions emphasizing primarily the frequency domain approach can be found in Bloomfield (1976), Priestley (1981), or Jenkins and Watts (1968). On a more advanced level, Hannan (1970), Brillinger (1981), Brockwell and Davis (1991), and Fuller (1995) are available as theoretical sources. Our coverage of the frequency domain is given in Chapters 4 and 7.

The objective of this book is to provide a unified and reasonably complete exposition of statistical methods used in time series analysis, giving serious consideration to both the time and frequency domain approaches. Because a myriad of possible methods for analyzing any particular experimental series can exist, we have integrated real data from a number of subject fields into the exposition and have suggested methods for analyzing these data.

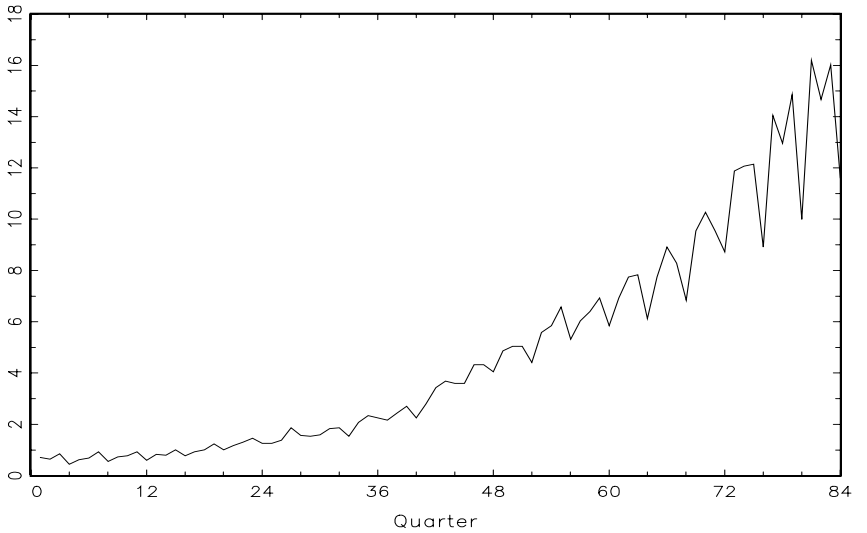


Figure 1.1 Johnson & Johnson quarterly earnings per share, 84 quarters, 1960-I to 1980-IV.

1.2 The Nature of Time Series Data

Some of the problems and questions of interest to the prospective time series analyst can best be exposed by considering real experimental data taken from different subject areas. The following cases illustrate some of the common kinds of experimental time series data as well as some of the statistical questions that might be asked about such data.

Example 1.1 Johnson & Johnson Quarterly Earnings

Figure 1.1 shows quarterly earnings per share for the U.S. company Johnson & Johnson, furnished by Professor Paul Griffin (personal communication) of the Graduate School of Management, University of California, Davis. There are 84 quarters (21 years) measured from the first quarter of 1960 to the last quarter of 1980. Modeling such series begins by observing the primary patterns in the time history. In this case, note the gradually increasing underlying trend and the rather regular variation superimposed on the trend that seems to repeat over quarters. Methods for analyzing data such as these are explored in Chapter 2 (see Problem 2.1) using regression techniques, and in Chapter 6, §6.5, using structural equation modeling.

To plot the data using the R statistical package, suppose you saved the data as `jj.dat` in the directory `mydata`. Then use the following steps to read in the data and plot the time series (the `>` below are prompts, you

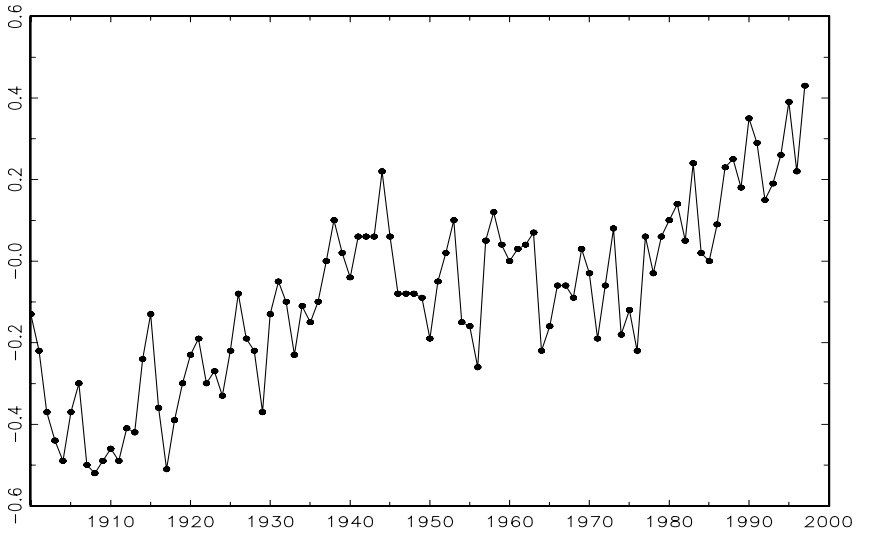


Figure 1.2 Yearly average global temperature deviations (1900–1997) in degrees centigrade.

do not type them):

```
> jj = scan("/mydata/jj.dat")   # yes forward slash
> jj=ts(jj,start=1960, frequency=4)
> plot(jj,  ylab="Quarterly Earnings per Share")
```

You can replace `scan` with `read.table` in this example.

Example 1.2 Global Warming

Consider a global temperature series record, discussed in Jones (1994) and Parker et al. (1994, 1995). The data in Figure 1.2 are a combination of land-air average temperature anomalies (from 1961-1990 average), measured in degrees centigrade, for the years 1900-1997. We note an apparent upward trend in the series that has been used as an argument for the global warming hypothesis. Note also the leveling off at about 1935 and then another rather sharp upward trend at about 1970. The question of interest for global warming proponents and opponents is whether the overall trend is natural or whether it is caused by some human-induced interface. Problem 2.8 examines 634 years of glacial sediment data that might be taken as a long-term temperature proxy. Such percentage changes in temperature do not seem to be unusual over a time period of 100 years. Again, the question of trend is of more interest than particular periodicities.

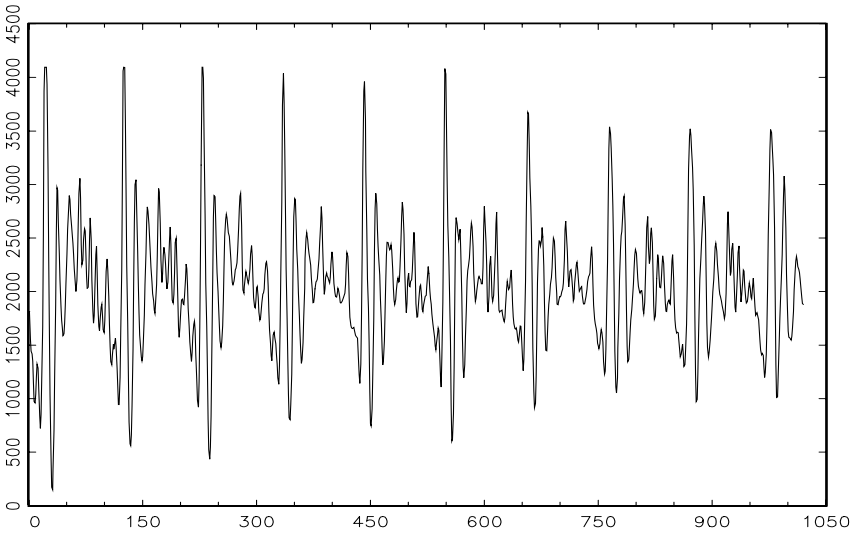


Figure 1.3 Speech recording of the syllable $aaa\cdots hhh$ sampled at 10,000 points per second with $n = 1020$ points.

Example 1.3 Speech Data

More involved questions develop in applications to the physical sciences. Figure 1.3 shows a small .1 second (1000 point) sample of recorded speech for the phrase $aaa\cdots hhh$, and we note the repetitive nature of the signal and the rather regular periodicities. One current problem of great interest is computer recognition of speech, which would require converting this particular signal into the recorded phrase $aaa\cdots hhh$. Spectral analysis can be used in this context to produce a signature of this phrase that can be compared with signatures of various library syllables to look for a match. One can immediately notice the rather regular repetition of small wavelets. The separation between the packets is known as the pitch period and represents the response of the vocal tract filter to a periodic sequence of pulses stimulated by the opening and closing of the glottis.

Example 1.4 New York Stock Exchange

As an example of financial time series data, Figure 1.4 shows the daily returns (or percent change) of the New York Stock Exchange (NYSE) from February 2, 1984 to December 31, 1991. It is easy to spot the crash of October 19, 1987 in the figure. The data shown in Figure 1.4 are typical of return data. The mean of the series appears to be stable

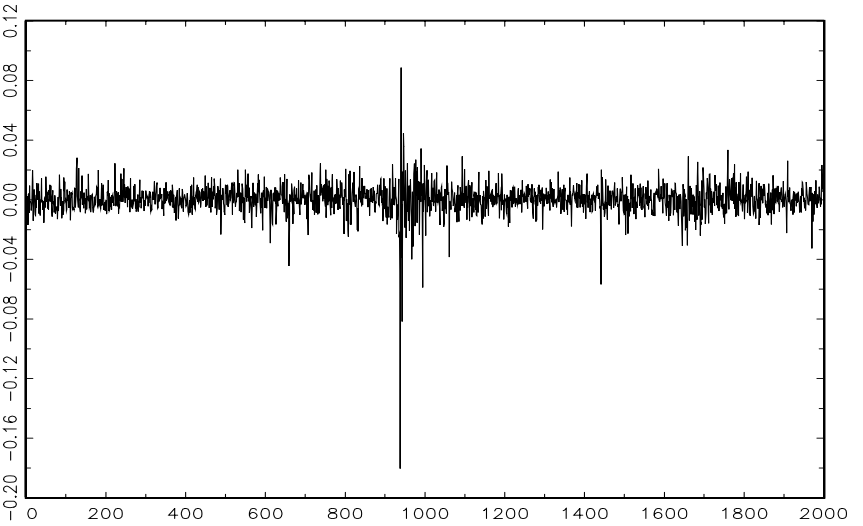


Figure 1.4 Returns of the NYSE. The data are daily value weighted market returns from February 2, 1984 to December 31, 1991 (2000 trading days). The crash of October 19, 1987 occurs at $t = 938$.

with an average return of approximately zero, however, the volatility (or variability) of data changes over time. In fact, the data show volatility clustering; that is, highly volatile periods tend to be clustered together. A problem in the analysis of these type of financial data is to forecast the volatility of future returns. Models such as ARCH and GARCH models (Engle, 1982; Bollerslev, 1986) and stochastic volatility models (Harvey, Ruiz and Shephard, 1994) have been developed to handle these problems. We will discuss these models and the analysis of financial data in Chapters 5 and 6.

Example 1.5 El Niño and Fish Population

We may also be interested in analyzing several time series at once. Figure 1.5 shows monthly values of an environmental series called the Southern Oscillation Index (SOI) and associated Recruitment (number of new fish) furnished by Dr. Roy Mendelssohn of the Pacific Environmental Fisheries Group (personal communication). Both series are for a period of 453 months ranging over the years 1950-1987. The SOI measures changes in air pressure, related to sea surface temperatures in the central Pacific. The central Pacific Ocean warms every three to seven years due to the El Niño effect, which has been blamed, in particular, for the 1997 floods in the midwestern portions of the U.S. Both series in Figure 1.5 tend to exhibit repetitive behavior, with regularly repeating cycles that are easily

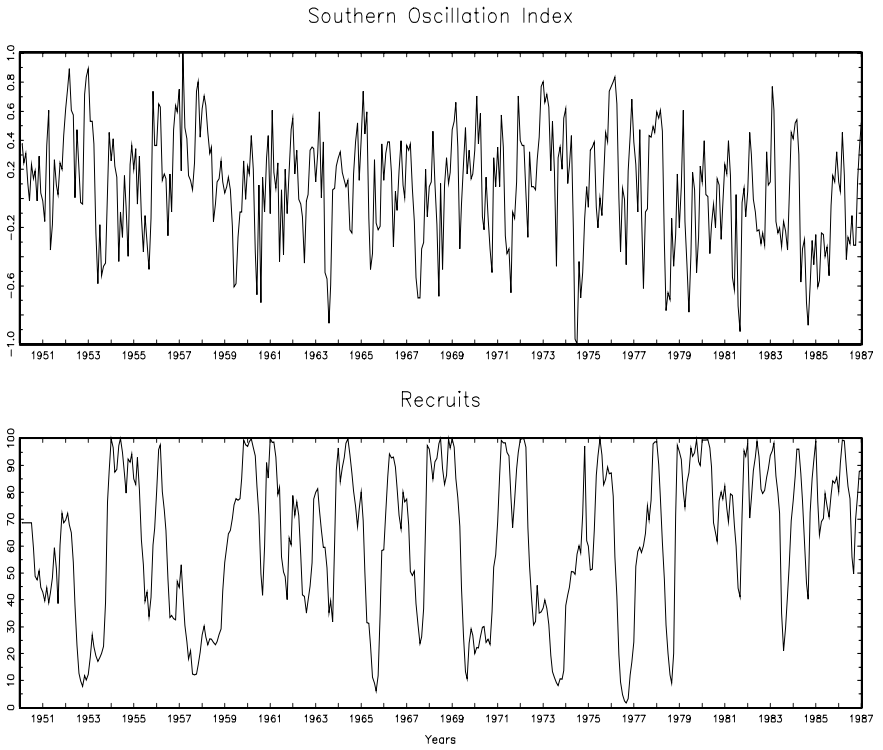


Figure 1.5 Monthly SOI and Recruitment (Estimated new fish), 1950-1987.

visible. This periodic behavior is of interest because underlying processes of interest may be regular and the rate or frequency of oscillation characterizing the behavior of the underlying series would help to identify them. One can also remark that the cycles of the SOI are repeating at a faster rate than those of the Recruitment series. The Recruitment series also shows several kinds of oscillations, a faster frequency that seems to repeat about every 12 months and a slower frequency that seems to repeat about every 50 months. The study of the kinds of cycles and their strengths is the subject of Chapter 4. The two series also tend to be somewhat related; it is easy to imagine that somehow the fish population is dependent on the SOI. Perhaps, even a lagged relation exists, with the SOI signaling changes in the fish population. This possibility suggests trying some version of regression analysis as a procedure for relating the two series. Transfer function modeling, as considered in Chapter 5, can be applied in this case to obtain a model relating Recruitment to its own past and the past values of the SOI Index.

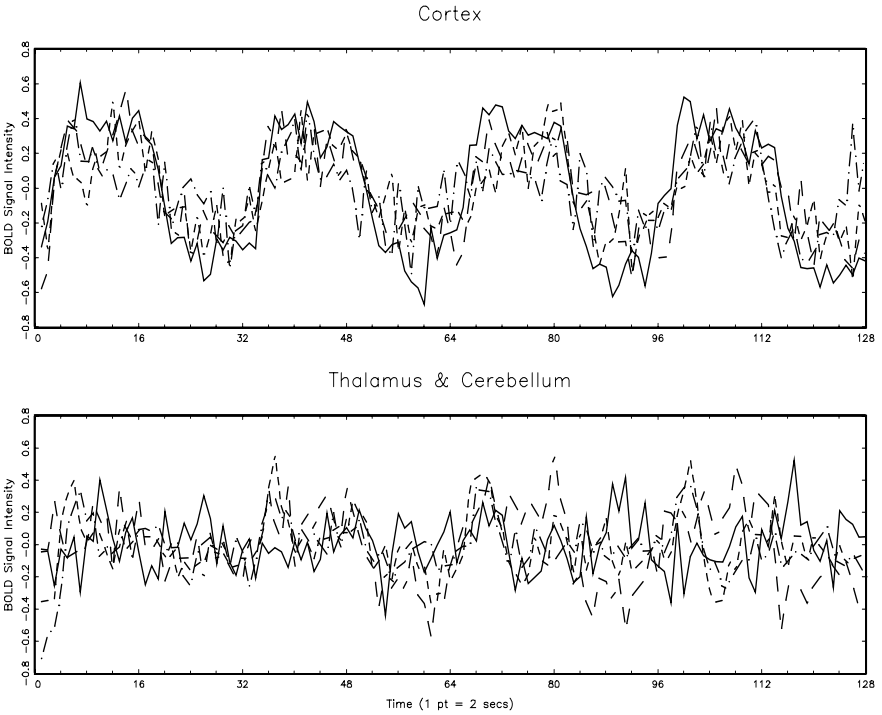


Figure 1.6 fMRI data from various locations in the cortex, thalamus, and cerebellum; $n = 128$ points, one observation taken every 2 seconds.

Example 1.6 fMRI Imaging

A fundamental problem in classical statistics occurs when we are given a collection of independent series or vectors of series, generated under varying experimental conditions or treatment configurations. Such a set of series is shown in Figure 1.6, where we observe data collected from various locations in the brain via functional magnetic resonance imaging (fMRI). In this example, five subjects were given periodic brushing on the hand. The stimulus was applied for 32 seconds and then stopped for 32 seconds; thus, the signal period is 64 seconds. The sampling rate was one observation every 2 seconds for 256 seconds ($n = 128$). For this example, we averaged the results over subjects (these were evoked responses, and all subjects were in phase). The series shown in Figure 1.6 are consecutive measures of blood oxygenation-level dependent (BOLD) signal intensity, which measures areas of activation in the brain. Notice that the periodicities appear strongly in the motor cortex series and less strongly in the thalamus and cerebellum. The fact that one has series from different areas of the brain suggests testing whether the areas are

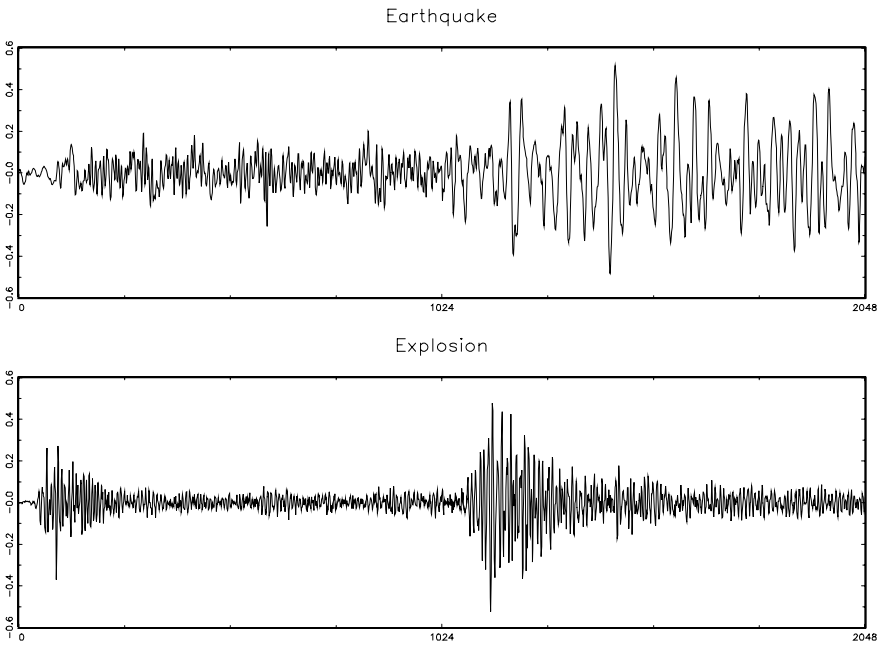


Figure 1.7 Arrival phases from an earthquake (top) and explosion (bottom) at 40 points per second.

responding differently to the brush stimulus. Analysis of variance techniques accomplish this in classical statistics, and we show in Chapter 7 how these classical techniques extend to the time series case, leading to a spectral analysis of variance.

The data are in a file called `fmri.dat`, which consists of nine columns; the first column represents time, whereas the second through ninth columns represent the BOLD signals at eight locations. Assuming the data are located in the directory `mydata`, use the following commands in R to plot the data as in this example.

```
> fmri = read.table("/mydata/fmri.dat")
> par(mfrow=c(2,1))   # sets up the graphics
> ts.plot(fmri[,2:5], lty=c(1,4), ylab="BOLD")
> ts.plot(fmri[,6:9], lty=c(1,4), ylab="BOLD")
```

Example 1.7 Earthquakes and Explosions

As a final example, the series in Figure 1.7 represent two phases or arrivals along the surface, denoted by P ($t = 1, \ldots, 1024$) and S ($t =$

1025, . . . , 2048), at a seismic recording station. The recording instruments in Scandinavia are observing earthquakes and mining explosions with one of each shown in Figure 1.7. The general problem of interest is in distinguishing or discriminating between waveforms generated by earthquakes and those generated by explosions. Features that may be important are the rough amplitude ratios of the first phase P to the second phase S, which tend to be smaller for earthquakes than for explosions. In the case of the two events in Figure 1.7, the ratio of maximum amplitudes appears to be somewhat less than .5 for the earthquake and about 1 for the explosion. Otherwise, note a subtle difference exists in the periodic nature of the S phase for the earthquake. We can again think about spectral analysis of variance for testing the equality of the periodic components of earthquakes and explosions. We would also like to be able to classify future P and S components from events of unknown origin, leading to the time series discriminant analysis developed in Chapter 7.

The data are in the file `eq5exp6.dat` as one column with 4096 entries, the first 2048 observations correspond to an earthquake and the next 2048 observations correspond to an explosion. To read and plot the data as in this example, use the following commands in R:

```
> x = matrix(scan("/mydata/eq5exp6.dat"), ncol=2)
> par(mfrow=c(2,1))
> plot.ts(x[,1], main="Earthquake", ylab="EQ5")
> plot.ts(x[,2], main="Explosion", ylab="EXP6")
```

1.3 Time Series Statistical Models

The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for sample data, like that encountered in the previous section. In order to provide a statistical setting for describing the character of data that seemingly fluctuate in a random fashion over time, we assume a time series can be defined as a collection of random variables indexed according to the order they are obtained in time. For example, we may consider a time series as a sequence of random variables, $x_1, x_2, x_3, \ldots$, where the random variable x_1 denotes the value taken by the series at the first time point, the variable x_2 denotes the value for the second time period, x_3 denotes the value for the third time period, and so on. In general, a collection of random variables, $\{x_t\}$, indexed by t is referred to as a stochastic process. In this text, t will typically be discrete and vary over the integers $t = 0, \pm1, \pm2, ...$, or some subset of the integers. The observed values of a stochastic process are referred to as a realization of the stochastic process. *Because it will be clear from the context of our discussions, we use the term time series whether we are referring generically to the process or to a particular realization and make no notational distinction between the two concepts.*

It is conventional to display a sample time series graphically by plotting the values of the random variables on the vertical axis, or ordinate, with the time scale as the abscissa. It is usually convenient to connect the values at adjacent time periods to reconstruct visually some original hypothetical continuous time series that might have produced these values as a discrete sample. Many of the series discussed in the previous section, for example, could have been observed at any continuous point in time and are conceptually more properly treated as continuous time series. The approximation of these series by discrete time parameter series sampled at equally spaced points in time is simply an acknowledgment that sampled data will, for the most part, be discrete because of restrictions inherent in the method of collection. Furthermore, the analysis techniques are then feasible using computers, which are limited to digital computations. Theoretical developments also rest on the idea that a continuous parameter time series should be specified in terms of finite-dimensional distribution functions defined over a finite number of points in time. This is not to say that the selection of the sampling interval or rate is not an extremely important consideration. The appearance of data can be changed completely by adopting an insufficient sampling rate. We have all seen wagon wheels in movies appear to be turning backwards because of the insufficient number of frames sampled by the camera. This phenomenon leads to a distortion called aliasing.

The fundamental visual characteristic distinguishing the different series shown in Examples 1.1–1.7 is their differing degrees of smoothness. One possible explanation for this smoothness is that it is being induced by the supposition that adjacent points in time are correlated, so the value of the series at time t, say, x_t, depends in some way on the past values $x_{t-1}, x_{t-2}, \ldots$. This model expresses a fundamental way in which we might think about generating realistic-looking time series. To begin to develop an approach to using collections of random variables to model time series, consider Example 1.8.

Example 1.8 White Noise

A simple kind of generated series might be a collection of uncorrelated random variables, w_t, with mean 0 and finite variance σ_w^2. The time series generated from uncorrelated variables is used as a model for noise in engineering applications, where it is called *white noise*; we shall sometimes denote this process as $w_t \sim wn(0, \sigma_w^2)$. The designation white originates from the analogy with white light and indicates that all possible periodic oscillations are present with equal strength.

We will, at times, also require the noise to be iid random variables with mean 0 and variance σ_w^2. We shall distinguish this case by saying white independent noise, or by writing $w_t \sim \text{iid}(0, \sigma_w^2)$. A particularly useful white noise series is Gaussian white noise, wherein the w_t are independent normal random variables, with mean 0 and variance σ_w^2; or more succinctly, $w_t \sim \text{iid N}(0, \sigma_w^2)$. Figure 1.8 shows in the upper panel a collection of 500 such random variables, with $\sigma_w^2 = 1$, plotted in the order in

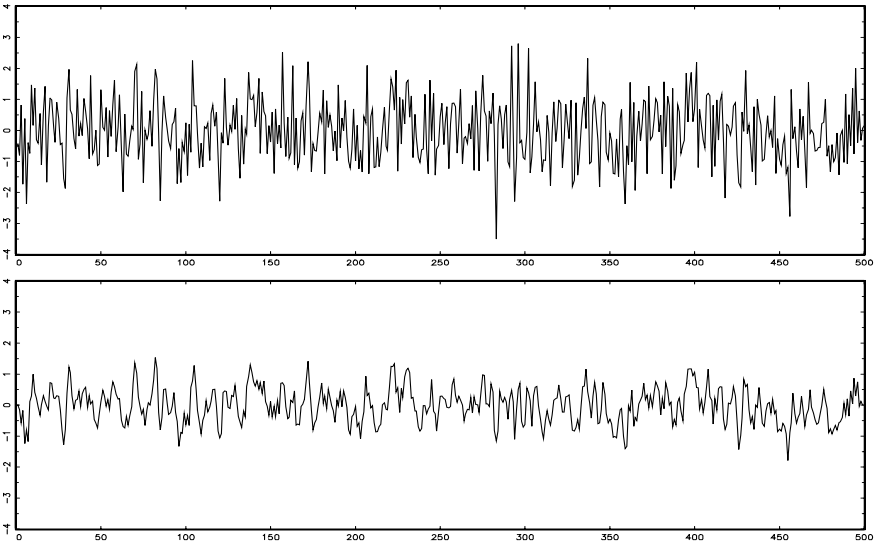


Figure 1.8 Gaussian white noise series (top) and three-point moving average of the Gaussian white noise series (bottom).

which they were drawn. The resulting series bears a slight resemblance to the explosion in Figure 1.7 but is not smooth enough to serve as a plausible model for any of the other experimental series. The plot tends to show visually a mixture of many different kinds of oscillations in the white noise series.

If the stochastic behavior of all time series could be explained in terms of the white noise model, classical statistical methods would suffice. Two ways of introducing serial correlation and more smoothness into time series models are given in Examples 1.9 and 1.10.

Example 1.9 Moving Averages

We might replace the white noise series w_t by a moving average that smoothes the series. For example, consider replacing w_t in Example 1.8 by an average of its current value and its immediate neighbors in the past and future. That is, let

$$v_t = \frac{1}{3}\big(w_{t-1} + w_t + w_{t+1}\big), \tag{1.1}$$

which leads to the series shown in the lower panel of Figure 1.8. Inspecting the series shows a smoother version of the first series, reflecting the fact that the slower oscillations are more apparent and some of the faster

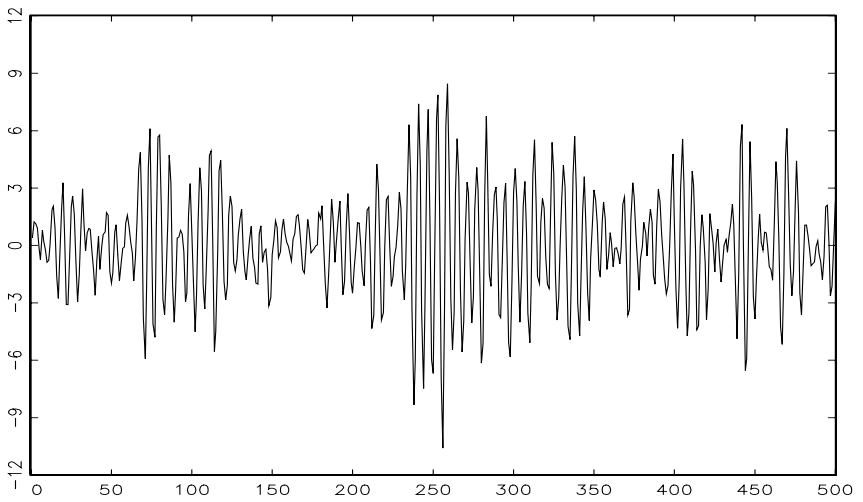


Figure 1.9 Autoregressive series generated from model (1.2).

oscillations are taken out. We begin to notice a similarity to the SOI in Figure 1.5, or perhaps, to some of the fMRI series in Figure 1.6.

To reproduce Figure 1.8 in R use the following commands:[1]

```
> w = rnorm(500,0,1)  # 500 N(0,1) variates
> v = filter(w, sides=2, rep(1,3)/3)  # moving average
> par(mfrow=c(2,1))
> plot.ts(w)
> plot.ts(v)
```

The speech series in Figure 1.3 and the Recruitment series in Figure 1.5, as well as some of the MRI series in Figure 1.6, differ from the moving average series because one particular kind of oscillatory behavior seems to predominate, producing a sinusoidal type of behavior. A number of methods exist for generating series with this quasi-periodic behavior; we illustrate a popular one based on the autoregressive model considered in Chapter 3.

Example 1.10 Autoregressions

Suppose we consider the white noise series w_t of Example 1.8 as input and calculate the output using the second-order equation

$$x_t = x_{t-1} - .90x_{t-2} + w_t \tag{1.2}$$

successively for $t = 1, 2, \ldots, 500$. Equation (1.2) represents a regression or prediction of the current value x_t of a time series as a function of

[1] A linear combination of values in a time series such as in (1.1) is referred to, generically, as a filtered series; hence the command `filter`.

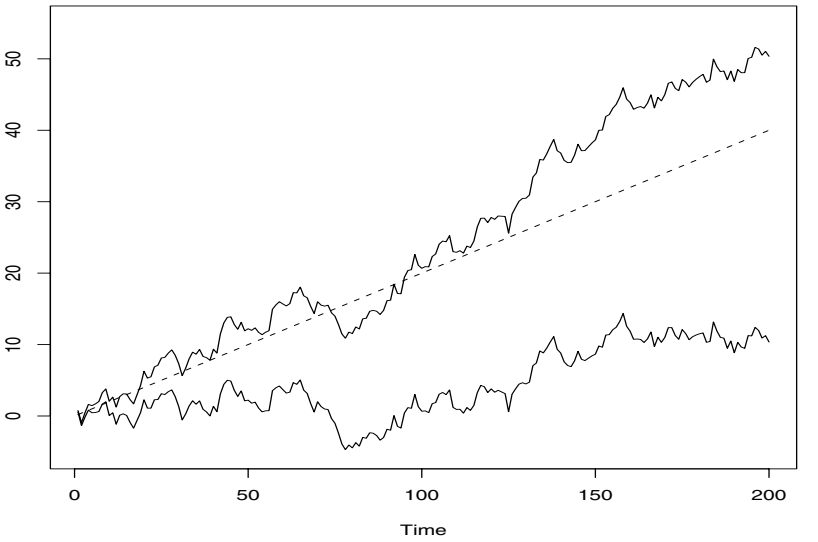


Figure 1.10 Random walk, $\sigma_w = 1$, with drift $\delta = .2$ (upper jagged line), without drift, $\delta = 0$ (lower jagged line), and a straight line with slope .2 (dashed line).

the past two values of the series, and, hence, the term autoregression is suggested for this model. A problem with startup values exists here because (1.2) also depends on the initial conditions x_0 and x_{-1}, but, for now, we assume that we are given these values and generate the succeeding values by substituting into (1.2). The resulting output series is shown in Figure 1.9, and we note the periodic behavior of the series, which is similar to that displayed by the speech series in Figure 1.3. The autoregressive model above and its generalizations can be used as an underlying model for many observed series and will be studied in detail in Chapter 3.

One way to simulate and plot data from the model (1.2) in R is to use the following commands (another way is to use `arima.sim`).

```
> w = rnorm(550,0,1)  # 50 extra to avoid startup problems
> x = filter(w, filter=c(1,-.9), method="recursive")
> plot.ts(x[51:550])
```

Example 1.11 Random Walk

A model for analyzing trend is the random walk with drift model given by

$$x_t = \delta + x_{t-1} + w_t \tag{1.3}$$

for $t = 1, 2, \ldots$, with initial condition $x_0 = 0$, and where w_t is white noise. The constant δ is called the drift, and when $\delta = 0$, (1.3) is called simply

a random walk. The term random walk comes from the fact that, when $\delta = 0$, the value of the time series at time t is the value of the series at time $t-1$ plus a completely random movement determined by w_t. Note that we may rewrite (1.3) as a cumulative sum of white noise variates. That is,

$$x_t = \delta\, t + \sum_{j=1}^{t} w_j \tag{1.4}$$

for $t = 1, 2, \ldots$; either use induction, or plug (1.4) into (1.3) to verify this statement. Figure 1.10 shows 200 observations generated from the model with $\delta = 0$ and .2, and with $\sigma_w = 1$. For comparison, we also superimposed the straight line $.2t$ on the graph.

To reproduce Figure 1.10 in R:

```
> set.seed(154)
> w = rnorm(200,0,1); x = cumsum(w)
> wd = w +.2;  xd = cumsum(wd)
> plot.ts(xd, ylim=c(-5,55))
> lines(x)
> lines(.2*(1:200), lty="dashed")
```

Example 1.12 Signal in Noise

Many realistic models for generating time series assume an underlying signal with some consistent periodic variation, contaminated by adding a random noise. For example, it is easy to detect the regular cycle fMRI series displayed on the top of Figure 1.6. Consider the model

$$x_t = 2\cos(2\pi t/50 + .6\pi) + w_t \tag{1.5}$$

for $t = 1, 2, \ldots, 500$, where the first term is regarded as the signal, shown in the upper panel of Figure 1.11. We note that a sinusoidal waveform can be written as

$$A\cos(2\pi\omega t + \phi), \tag{1.6}$$

where A is the amplitude, ω is the frequency of oscillation, and ϕ is a phase shift. In (1.5), $A = 2$, $\omega = 1/50$ (one cycle every 50 time points), and $\phi = .6\pi$.

An additive noise term was taken to be white noise with $\sigma_w = 1$ (middle panel) and $\sigma_w = 5$ (bottom panel), drawn from a normal distribution. Adding the two together obscures the signal, as shown in the lower panels of Figure 1.11. Of course, the degree to which the signal is obscured depends on the amplitude of the signal and the size of σ_w. The ratio of the amplitude of the signal to σ_w (or some function of the ratio) is sometimes called the signal-to-noise ratio (SNR); the larger the SNR, the easier it is to detect the signal. Note that the signal is easily discernible

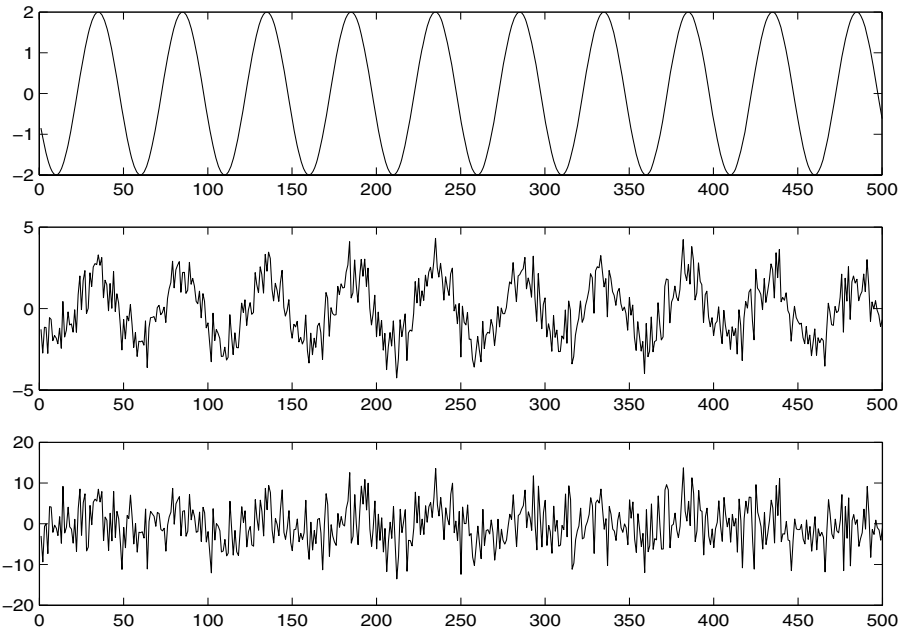


Figure 1.11 Cosine wave with period 50 points (top panel) compared with the cosine wave contaminated with additive white Gaussian noise, $\sigma_w = 1$ (middle panel) and $\sigma_w = 5$ (bottom panel); see (1.5).

in the middle panel of Figure 1.11, whereas the signal is obscured in the bottom panel. Typically, we will not observe the signal, but the signal obscured by noise.

To reproduce Figure 1.11 in R, use the following commands:

```
> t = 1:500
> c = 2*cos(2*pi*t/50 + .6*pi)
> w = rnorm(500,0,1)
> par(mfrow=c(3,1))
> plot.ts(c)
> plot.ts(c + w)
> plot.ts(c + 5*w)
```

In Chapter 4, we will study the use of spectral analysis as a possible technique for detecting regular or periodic signals, such as the one described in Example 1.12. In general, we would emphasize the importance of simple additive models such as given above in the form

$$x_t = s_t + v_t, \tag{1.7}$$

where s_t denotes some unknown signal and v_t denotes a time series that may be white or correlated over time. The problems of detecting a signal and then

in estimating or extracting the waveform of s_t are of great interest in many areas of engineering and the physical and biological sciences. In economics, the underlying signal may be a trend or it may be a seasonal component of a series. Models such as (1.7), where the signal has an autoregressive structure, form the motivation for the state-space model of Chapter 6.

In the above examples, we have tried to motivate the use of various combinations of random variables emulating real time series data. Smoothness characteristics of observed time series were introduced by combining the random variables in various ways. Averaging independent random variables over adjacent time points, as in Example 1.9, or looking at the output of difference equations that respond to white noise inputs, as in Example 1.10, are common ways of generating correlated data. In the next section, we introduce various theoretical measures used for describing how time series behave. As is usual in statistics, the complete description involves the multivariate distribution function of the jointly sampled values $x_1, x_2, \ldots, x_n$, whereas more economical descriptions can be had in terms of the mean and autocorrelation functions. Because correlation is an essential feature of time series analysis, the most useful descriptive measures are those expressed in terms of covariance and correlation functions.

1.4 Measures of Dependence: Autocorrelation and Cross-Correlation

A complete description of a time series, observed as a collection of n random variables at arbitrary integer time points $t_1, t_2, \ldots, t_n$, for any positive integer n, is provided by the joint distribution function, evaluated as the probability that the values of the series are jointly less than the n constants, $c_1, c_2, \ldots, c_n$, i.e.,

$$F(c_1, c_2, \ldots, c_n) = P\big(x_{t_1} \leq c_1, x_{t_2} \leq c_2, \ldots, x_{t_n} \leq c_n\big). \tag{1.8}$$

Unfortunately, the multidimensional distribution function cannot usually be written easily unless the random variables are jointly normal, in which case, expression (1.8) comes from the usual multivariate normal distribution (see Anderson, 1984, or Johnson and Wichern, 1992). A particular case in which the multidimensional distribution function is easy would be for independent and identically distributed standard normal random variables, for which the joint distribution function can be expressed as the product of the marginals, say,

$$F(c_1, c_2, \ldots, c_n) = \prod_{t=1}^{n} \Phi(c_t), \tag{1.9}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\big\{-\frac{z^2}{2}\big\}\, dz \tag{1.10}$$

is the cumulative distribution function of the standard normal.

Although the multidimensional distribution function describes the data completely, it is an unwieldy tool for displaying and analyzing time series data. The distribution function (1.8) must be evaluated as a function of n arguments, so any plotting of the corresponding multivariate density functions is virtually impossible. The one-dimensional distribution functions

$$F_t(x) = P\{x_t \leq x\}$$

or the corresponding one-dimensional density functions

$$f_t(x) = \frac{\partial F_t(x)}{\partial x},$$

when they exist, are often informative for determining whether a particular coordinate of the time series has a well-known density function, like the normal (Gaussian) distribution.

Definition 1.1 *The* **mean function** *is defined as*

$$\mu_{xt} = E(x_t) = \int_{-\infty}^{\infty} x f_t(x)\, dx, \tag{1.11}$$

provided it exists, where E denotes the usual expected value operator. When no confusion exists about which time series we are referring to, we will drop a subscript and write μ_{xt} as μ_t.

The important thing to realize about μ_t is that it is a theoretical mean for the series at one particular time point, where the mean is taken over all possible events that could have produced x_t.

Example 1.13 Mean Function of a Moving Average Series

If w_t denotes a white noise series, then $\mu_{wt} = E(w_t) = 0$ for all t. The top series in Figure 1.8 reflects this, as the series clearly fluctuates around a mean value of zero. Smoothing the series as in Example 1.9 does not change the mean because we can write

$$\mu_{vt} = E(v_t) = \frac{1}{3}[E(w_{t-1}) + E(w_t) + E(w_{t+1})] = 0.$$

Example 1.14 Mean Function of a Random Walk with Drift

Consider the random walk with drift model given in (1.4),

$$x_t = \delta\, t + \sum_{j=1}^{t} w_j, \qquad t = 1, 2, \ldots\ .$$

As in the previous example, because $E(w_t) = 0$ for all t, and δ is a constant, we have

$$\mu_{xt} = E(x_t) = \delta\, t + \sum_{j=1}^{t} E(w_j) = \delta\, t$$

which is a straight line with slope δ. A realization of a random walk with drift can be compared to its mean function in Figure 1.10.

Example 1.15 Mean Function of Signal Plus Noise

A great many practical applications depend on assuming the observed data have been generated by a fixed signal waveform superimposed on a zero-mean noise process, leading to an additive signal model of the form (1.5). It is clear, because the signal in (1.5) is a fixed function of time, we will have

$$\begin{aligned} \mu_{xt} \;=\; E(x_t) &= E\big[2\cos(2\pi t/50 + .6\pi) + w_t\big] \\ &= 2\cos(2\pi t/50 + .6\pi) + E(w_t) \\ &= 2\cos(2\pi t/50 + .6\pi), \end{aligned}$$

and the mean function is just the cosine wave.

The lack of independence between two adjacent values x_s and x_t can be assessed numerically, as in classical statistics, using the notions of covariance and correlation. Assuming the variance of x_t is finite, we have the following definition.

Definition 1.2 *The* **autocovariance function** *is defined as the second moment product*

$$\gamma_x(s,t) = E[(x_s - \mu_s)(x_t - \mu_t)], \tag{1.12}$$

for all s and t. When no possible confusion exists about which time series we are referring to, we will drop the subscript and write $\gamma_x(s,t)$ as $\gamma(s,t)$.

Note that $\gamma_x(s,t) = \gamma_x(t,s)$ for all time points s and t. The autocovariance measures the *linear* dependence between two points on the same series observed at different times. Very smooth series exhibit autocovariance functions that stay large even when the t and s are far apart, whereas choppy series tend to have autocovariance functions that are nearly zero for large separations. The autocovariance (1.12) is the average cross-product relative to the joint density $F(x_s, x_t)$. Recall from classical statistics that if $\gamma_x(s,t) = 0$, x_s and x_t are not linearly related, but there still may be some dependence structure between them. If, however, x_s and x_t are bivariate normal, $\gamma_x(s,t) = 0$ ensures their independence. It is clear that, for $s = t$, the autocovariance reduces to the (assumed finite) variance, because

$$\gamma_x(t,t) = E[(x_t - \mu_t)^2]. \tag{1.13}$$

Example 1.16 Autocovariance of White Noise

The white noise series w_t, shown in the top panel of Figure 1.8, has $E(w_t) = 0$ and

$$\gamma_w(s,t) = E(w_s w_t) = \begin{cases} \sigma_w^2, & s = t \\ 0, & s \neq t \end{cases}$$

where, in this example, $\sigma_w^2 = 1$. Noting that w_s and w_t are uncorrelated for $s \neq t$, we would have $E(w_s w_t) = E(w_s)E(w_t) = 0$ because the mean values of the white noise variates are zero.

Example 1.17 Autocovariance of a Moving Average

Consider applying a three-point moving average to the white noise series w_t of the previous example, as in Example 1.9 ($\sigma_w^2 = 1$). Because v_t in (1.1) has mean zero, we have

$$\begin{aligned} \gamma_v(s,t) &= E[(v_s - 0)(v_t - 0)] \\ &= \frac{1}{9} E[(w_{s-1} + w_s + w_{s+1})(w_{t-1} + w_t + w_{t+1})]. \end{aligned}$$

It is convenient to calculate it as a function of the separation, $s - t = h$, say, for $h = 0, \pm 1, \pm 2, \ldots$. For example, with $h = 0$,

$$\begin{aligned} \gamma_v(t,t) &= \frac{1}{9} E[(w_{t-1} + w_t + w_{t+1})(w_{t-1} + w_t + w_{t+1})] \\ &= \frac{1}{9} [E(w_{t-1}w_{t-1}) + E(w_t w_t) + E(w_{t+1}w_{t+1})] \\ &= \frac{3}{9}. \end{aligned}$$

When $h = 1$,

$$\begin{aligned} \gamma_v(t+1,t) &= \frac{1}{9} E[(w_t + w_{t+1} + w_{t+2})(w_{t-1} + w_t + w_{t+1})] \\ &= \frac{1}{9} [E(w_t w_t) + E(w_{t+1}w_{t+1})] \\ &= \frac{2}{9}, \end{aligned}$$

using the fact that we may drop terms with unequal subscripts. Similar computations give $\gamma_v(t-1,t) = 2/9$, $\gamma_v(t+2,t) = \gamma_v(t-2,t) = 1/9$, and 0 for larger separations. We summarize the values for all s and t as

$$\gamma_v(s,t) = \begin{cases} 3/9, & s = t \\ 2/9, & |s-t| = 1 \\ 1/9, & |s-t| = 2 \\ 0, & |s-t| \geq 3. \end{cases} \tag{1.14}$$

Example 1.17 shows clearly that the smoothing operation introduces a covariance function that decreases as the separation between the two time points increases and disappears completely when the time points are separated by three or more time points. This particular autocovariance is interesting because it only depends on the time separation or lag and not on the absolute location of the points along the series. We shall see later that this dependence suggests a mathematical model for the concept of weak stationarity.

Example 1.18 Autocovariance of a Random Walk

For the random walk model, $x_t = \sum_{j=1}^{t} w_j$, we have

$$\gamma_x(s,t) = \mathrm{cov}(x_s, x_t) = \mathrm{cov}\left(\sum_{j=1}^{s} w_j, \sum_{k=1}^{t} w_k\right) = \min\{s,t\}\,\sigma_w^2,$$

because the w_t are uncorrelated random variables. Note that, as opposed to the previous examples, the autocovariance function of a random walk depends on the particular time values s and t, and not on the time separation or lag. Also, notice that the variance of the random walk, $\mathrm{var}(x_t) = \gamma_x(t,t) = t\,\sigma_w^2$, increases without bound as time t increases. The effect of this variance increase can be seen in Figure 1.10 as the processes starting to move away from their mean functions $\delta\,t$ (note, $\delta = 0$ and .2 in that example).

As in classical statistics, it is more convenient to deal with a measure of association between -1 and 1, and this leads to the following definition.

Definition 1.3 *The* **autocorrelation function (ACF)** *is defined as*

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}. \tag{1.15}$$

The ACF measures the linear predictability of the series at time t, say, x_t, using only the value x_s. We can show easily that $-1 \le \rho(s,t) \le 1$ using the Cauchy–Schwarz inequality.[2] If we can predict x_t *perfectly* from x_s through a linear relationship, $x_t = \beta_0 + \beta_1 x_s$, then the correlation will be 1 when $\beta_1 > 0$, and -1 when $\beta_1 < 0$. Hence, we have a rough measure of the ability to forecast the series at time t from the value at time s.

Often, we would like to measure the predictability of another series y_t from the series x_s. Assuming both series have finite variances, we have

Definition 1.4 *The* **cross-covariance function** *between two series x_t and y_t is*

$$\gamma_{xy}(s,t) = E[(x_s - \mu_{xs})(y_t - \mu_{yt})]. \tag{1.16}$$

[2]Note, the Cauchy–Schwarz inequality implies $|\gamma(s,t)|^2 \le \gamma(s,s)\gamma(t,t)$.

The scaled version of the cross-covariance function is called

Definition 1.5 *The* **cross-correlation function (CCF)**

$$\rho_{xy}(s,t) = \frac{\gamma_{xy}(s,t)}{\sqrt{\gamma_x(s,s)\gamma_y(t,t)}}. \tag{1.17}$$

We may easily extend the above ideas to the case of more than two series, say, $x_{t1}, x_{t2}, \ldots, x_{tr}$; that is, multivariate time series with r components. For example, the extension of (1.12) in this case is

$$\gamma_{jk}(s,t) = E[(x_{sj} - \mu_{sj})(x_{tk} - \mu_{tk})] \qquad j,k = 1,2,\ldots,r. \tag{1.18}$$

In the definitions above, the autocovariance and cross-covariance functions may change as one moves along the series because the values depend on both s and t, the locations of the points in time. In Example 1.17, the autocovariance function depends on the separation of x_s and x_t, say, $h = |s - t|$, and not on where the points are located in time. As long as the points are separated by h units, the location of the two points does not matter. This notion, called weak stationarity, when the mean is constant, is fundamental in allowing us to analyze sample time series data when only a single series is available.

1.5 Stationary Time Series

The preceding definitions of the mean and autocovariance functions are completely general. Although we have not made any special assumptions about the behavior of the time series, many of the preceding examples have hinted that a sort of regularity may exist over time in the behavior of a time series. We introduce the notion of regularity using a concept called stationarity.

Definition 1.6 *A* **strictly stationary** *time series is one for which the probabilistic behavior of every collection of values*

$$\{x_{t_1}, x_{t_2}, \ldots, x_{t_k}\}$$

is identical to that of the time shifted set

$$\{x_{t_1+h}, x_{t_2+h}, \ldots, x_{t_k+h}\}.$$

That is,

$$P\{x_{t_1} \leq c_1, \ldots, x_{t_k} \leq c_k\} = P\{x_{t_1+h} \leq c_1, \ldots, x_{t_k+h} \leq c_k\} \tag{1.19}$$

for all $k = 1, 2, ...$, all time points $t_1, t_2, \ldots, t_k$, all numbers $c_1, c_2, \ldots, c_k$, and all time shifts $h = 0, \pm 1, \pm 2, ...$.

If a time series is strictly stationary, then all of the multivariate distribution functions for subsets of variables must agree with their counterparts in the shifted set for all values of the shift parameter h. For example, when $k = 1$, (1.19) implies that

$$P\{x_s \leq c\} = P\{x_t \leq c\} \tag{1.20}$$

for any time points s and t. This statement implies, e.g., that the probability the value of a time series sampled hourly is negative at 1 AM is the same as at 10 AM. In addition, if the mean function, μ_t, of the series x_t exists, (1.20) implies that $\mu_s = \mu_t$ for all s and t, and hence μ_t must be constant. Note, e.g., that a random walk process with drift is not strictly stationary because its mean function changes with time (see Example 1.14).

When $k = 2$, we can write (1.19) as

$$P\{x_s \leq c_1, x_t \leq c_2\} = P\{x_{s+h} \leq c_1, x_{t+h} \leq c_2\} \tag{1.21}$$

for any time points s and t and shift h. Thus, if the variance function of the process exists, (1.21) implies that the autocovariance function of the series x_t satisfies

$$\gamma(s,t) = \gamma(s+h, t+h)$$

for all s and t and h. We may interpret this result by saying the autocovariance function of the process depends only on the time difference between s and t, and not on the actual times.

The version of stationarity in (1.19) is too strong for most applications. Moreover, it is difficult to assess strict stationarity from a single data set. Rather than impose conditions on all possible distributions of a time series, we will use a milder version that imposes conditions only on the first two moments of the series. We now have the following definition.

Definition 1.7 *A* **weakly stationary** *time series, x_t, is a finite variance process such that*

(i) the mean value function, μ_t, defined in (1.11) is constant and does not depend on time t, and

(ii) the covariance function, $\gamma(s,t)$, defined in (1.12) depends on s and t only through their difference $|s-t|$.

Henceforth, we will use the term **stationary** *to mean weak stationarity; if a process is stationary in the strict sense, we will use the term strictly stationary.*

It should be clear from the discussion of strict stationarity following Definition 1.6 that a strictly stationary, finite variance, time series is also stationary. The converse is not true unless there are further conditions. One important case where stationarity implies strict stationarity is if the time series is Gaussian [meaning all finite distributions, (1.19), of the series are Gaussian]. We will make this concept more precise at the end of this section.

Because the mean function, $E(x_t) = \mu_t$, of a stationary time series is independent of time t, we will write

$$\mu_t = \mu. \tag{1.22}$$

Also, because the covariance function of a stationary time series, $\gamma(s,t)$, depends on s and t only through their difference $|s-t|$, we may simplify the notation. Let $s = t+h$, where h represents the time shift or lag, then

$$\begin{aligned} \gamma(t+h,t) &= E[(x_{t+h}-\mu)(x_t-\mu)] \\ &= E[(x_h-\mu)(x_0-\mu)] \\ &= \gamma(h,0) \end{aligned} \tag{1.23}$$

does not depend on the time argument t; we have assumed that $\mathrm{var}(x_t) = \gamma(0,0) < \infty$. Henceforth, for convenience, we will drop the second argument of $\gamma(h,0)$.

Definition 1.8 *The* **autocovariance function of a stationary time series** *will be written as*

$$\gamma(h) = E[(x_{t+h}-\mu)(x_t-\mu)]. \tag{1.24}$$

Definition 1.9 *The* **autocorrelation function (ACF) of a stationary time series** *will be written using (1.15) as*

$$\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\gamma(t,t)}} = \frac{\gamma(h)}{\gamma(0)}. \tag{1.25}$$

The Cauchy–Schwarz inequality shows again that $-1 \le \rho(h) \le 1$ for all h, enabling one to assess the relative importance of a given autocorrelation value by comparing with the extreme values -1 and 1.

Example 1.19 Stationarity of White Noise

The autocovariance function of the white noise series of Examples 1.8 and 1.16 is easily evaluated as

$$\gamma_w(h) = E(w_{t+h}w_t) = \begin{cases} \sigma_w^2, & h = 0 \\ 0, & h \neq 0, \end{cases}$$

where, in these examples, $\sigma_w^2 = 1$. This means that the series is weakly stationary or stationary. If the white noise variates are also normally distributed or Gaussian, the series is also strictly stationary, as can be seen by evaluating (1.19) using the relationship (1.9).

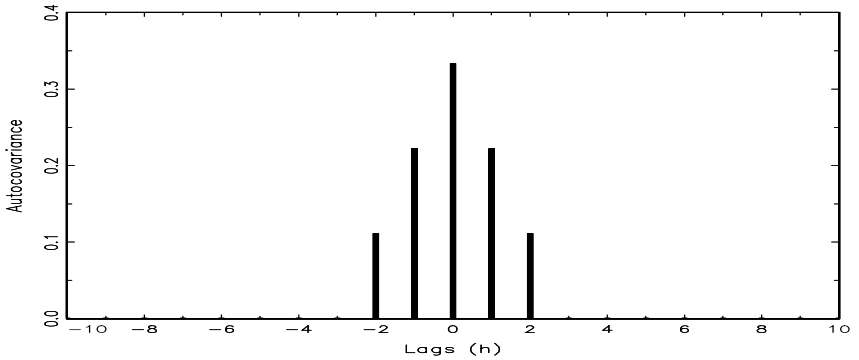


Figure 1.12 Autocovariance function of a three-point moving average.

Example 1.20 Stationarity of a Moving Average

The three-point moving average process used in Examples 1.9 and 1.17 is stationary because we may write the autocovariance function obtained in (1.14) as

$$\gamma_v(h) = \begin{cases} 3/9, & h = 0 \\ 2/9, & h = \pm 1 \\ 1/9, & h = \pm 2 \\ 0, & |h| \geq 3. \end{cases}$$

Figure 1.12 shows a plot of the autocovariance as a function of lag h. Interestingly, the autocovariance is symmetric and decays as a function of lag.

The autocovariance function of a stationary process has several useful properties. First, the value at $h = 0$, namely

$$\gamma(0) = E[(x_t - \mu)^2] \tag{1.26}$$

is the variance of the time series; note that the Cauchy–Schwarz inequality implies

$$|\gamma(h)| \leq \gamma(0).$$

A final useful property, noted in the previous example, is that autocovariance function of a stationary series is symmetric around the origin, that is,

$$\gamma(h) = \gamma(-h) \tag{1.27}$$

for all h. This property follows because shifting the series by h means that

$$\begin{aligned} \gamma(h) &= \gamma(t + h - t) \\ &= E[(x_{t+h} - \mu)(x_t - \mu)] \\ &= E[(x_t - \mu)(x_{t+h} - \mu)] \\ &= \gamma(t - (t + h)) \\ &= \gamma(-h), \end{aligned}$$

which shows how to use the notation as well as proving the result.

When several series are available, a notion of stationarity still applies with additional conditions.

Definition 1.10 *Two time series, say, x_t and y_t, are said to be* **jointly stationary** *if they are each stationary, and the cross-covariance function*

$$\gamma_{xy}(h) = E[(x_{t+h} - \mu_x)(y_t - \mu_y)] \tag{1.28}$$

is a function only of lag h.

Definition 1.11 *The* **cross-correlation function (CCF)** *of jointly stationary time series x_t and y_t is defined as*

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}. \tag{1.29}$$

Again, we have the result $-1 \leq \rho_{xy}(h) \leq 1$ which enables comparison with the extreme values -1 and 1 when looking at the relation between x_{t+h} and y_t. The cross-correlation function satisfies

$$\rho_{xy}(h) = \rho_{yx}(-h), \tag{1.30}$$

which can be shown by manipulations similar to those used to show (1.27).

Example 1.21 Joint Stationarity

Consider the two series, x_t and y_t, formed from the sum and difference of two successive values of a white noise process, say,

$$x_t = w_t + w_{t-1}$$

and

$$y_t = w_t - w_{t-1},$$

where w_t are independent random variables with zero means and variance σ_w^2. It is easy to show that $\gamma_x(0) = \gamma_y(0) = 2\sigma_w^2$ and $\gamma_x(1) = \gamma_x(-1) = \sigma_w^2, \gamma_y(1) = \gamma_y(-1) = -\sigma_w^2$. Also,

$$\begin{aligned} \gamma_{xy}(1) &= E[(x_{t+1} - 0)(y_t - 0)] \\ &= E[(w_{t+1} + w_t)(w_t - w_{t-1})] \\ &= \sigma_w^2 \end{aligned}$$

because only one product is nonzero. Similarly, $\gamma_{xy}(0) = 0, \gamma_{xy}(-1) = -\sigma_w^2$. We obtain, using (1.29),

$$\rho_{xy}(h) = \begin{cases} 0, & h = 0 \\ 1/2, & h = 1 \\ -1/2, & h = -1 \\ 0, & |h| \geq 2. \end{cases}$$

Clearly, the autocovariance and cross-covariance functions depend only on the lag separation, h, so the series are jointly stationary.

Example 1.22 Prediction Using Cross-Correlation

As a simple example of cross-correlation, consider the problem of determining possible leading or lagging relations between two series x_t and y_t. If the model

$$y_t = Ax_{t-\ell} + w_t$$

holds, the series x_t is said to lead y_t for $\ell > 0$ and is said to lag y_t for $\ell < 0$. Hence, the analysis of leading and lagging relations might be important in predicting the value of y_t from x_t. Assuming, for convenience, that x_t and y_t have zero means, and the noise w_t is uncorrelated with the x_t series, the cross-covariance function can be computed as

$$\begin{aligned} \gamma_{yx}(h) &= E(y_{t+h}x_t) \\ &= AE(x_{t+h-\ell}x_t) + E(w_{t+h}x_t) \\ &= A\gamma_x(h-\ell). \end{aligned}$$

The cross-covariance function will look like the autocovariance of the input series x_t, with a peak on the positive side if x_t leads y_t and a peak on the negative side if x_t lags y_t.

The concept of weak stationarity forms the basis for much of the analysis performed with time series. The fundamental properties of the mean and autocovariance functions (1.22) and (1.24) are satisfied by many theoretical models that appear to generate plausible sample realizations. In Examples 1.9 and 1.10, two series were generated that produced stationary looking realizations, and in Example 1.20, we showed that the series in Example 1.9 was, in fact, weakly stationary. Both examples are special cases of the so-called linear process.

Definition 1.12 *A* **linear process**, x_t, *is defined to be a linear combination of white noise variates* w_t, *and is given by*

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j} \tag{1.31}$$

where the coefficients satisfy

$$\sum_{j=-\infty}^{\infty} |\psi_j| < \infty. \tag{1.32}$$

For the linear process (see Problem 1.11), we may show that the autocovariance function is given by

$$\gamma(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_{j+h}\psi_j \tag{1.33}$$

for $h \geq 0$; recall that $\gamma(-h) = \gamma(h)$. This method exhibits the autocovariance function of the process in terms of the lagged products of the coefficients. Note that, for Example 1.9, we have $\psi_0 = \psi_{-1} = \psi_1 = 1/3$ and the result in Example 1.20 comes out immediately. The autoregressive series in Example 1.10 can also be put in this form, as can the general autoregressive moving average processes considered in Chapter 3.

Finally, as previously mentioned, an important case in which a weakly stationary series is also strictly stationary is the normal or Gaussian series.

Definition 1.13 *A process, $\{x_t\}$, is said to be a* **Gaussian process** *if the k-dimensional vectors $\boldsymbol{x} = (x_{t_1}, x_{t_2}, \ldots, x_{t_k})'$, for every collection of time points $t_1, t_2, \ldots, t_k$, and every positive integer k, have a multivariate normal distribution.*

Defining the $k \times 1$ mean vector $E(\boldsymbol{x}) \equiv \boldsymbol{\mu} = (\mu_{t_1}, \mu_{t_2}, \ldots, \mu_{t_k})'$ and the $k \times k$ covariance matrix as $\text{cov}(\boldsymbol{x}) \equiv \Gamma = \{\gamma(t_i, t_j);\ i, j = 1, \ldots, k\}$, the multivariate normal density function can be written as

$$f(\boldsymbol{x}) = (2\pi)^{-n/2}|\Gamma|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\Gamma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right\}, \tag{1.34}$$

where $|\cdot|$ denotes the determinant. This distribution forms the basis for solving problems involving statistical inference for time series. If a Gaussian time series, $\{x_t\}$, is weakly stationary, then $\mu_t = \mu$ and $\gamma(t_i, t_j) = \gamma(|t_i - t_j|)$, so that the vector $\boldsymbol{\mu}$ and the matrix Γ are independent of time. These facts imply that all the finite distributions, (1.34), of the series $\{x_t\}$ depend only on time lag and not on the actual times, and hence the series must be strictly stationary. We use the multivariate normal density in the form given above as well as in a modified version, applicable to complex random variables in the sequel.

1.6 Estimation of Correlation

Although the theoretical autocorrelation and cross-correlation functions are useful for describing the properties of certain hypothesized models, most of the analyses must be performed using sampled data. This limitation means the sampled points $x_1, x_2, \ldots, x_n$ only are available for estimating the mean, autocovariance, and autocorrelation functions. From the point of view of classical statistics, this poses a problem because we will typically not have iid copies of x_t that are available for estimating the covariance and correlation functions. In the usual situation with only one realization, however, the assumption of stationarity becomes critical. Somehow, we must use averages over this single realization to estimate the population means and covariance functions.

Accordingly, if a time series is stationary, the mean function, (1.22), $\mu_t = \mu$ is constant so that we can estimate it by the sample mean,

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t. \tag{1.35}$$

The theoretical autocovariance function, (1.24), is estimated by the sample autocovariance function defined as follows.

Definition 1.14 *The* **sample autocovariance function** *is defined as*

$$\widehat{\gamma}(h) = n^{-1}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x}), \tag{1.36}$$

with $\widehat{\gamma}(-h) = \widehat{\gamma}(h)$ *for* $h = 0, 1, \ldots, n-1$.

The sum in (1.36) runs over a restricted range because x_{t+h} is not available for $t + h > n$. The estimator in (1.36) is generally preferred to the one that would be obtained by dividing by $n-h$ because (1.36) is a non-negative definite function. This means that if we let $\widehat{\Gamma} = \{\widehat{\gamma}(i-j);\ i, j = 1, ..., n\}$ be the $n \times n$ sample covariance matrix of the data $\boldsymbol{x} = (x_1, \ldots, x_n)'$, then $\widehat{\Gamma}$ is a non-negative definite matrix. So, if we let $\boldsymbol{a} = (a_1, \ldots, a_n)'$ be an $n \times 1$ vector of constants, then $\widehat{\mathrm{var}}(\boldsymbol{a}'\boldsymbol{x}) = \boldsymbol{a}'\widehat{\Gamma}\boldsymbol{a} \geq 0$. Thus, the non-negative definite property ensures sample variances of linear combinations of the variates x_t will always be non-negative. Note that neither dividing by n nor $n-h$ in (1.36) yields an unbiased estimate of $\gamma(h)$.

Definition 1.15 *The* **sample autocorrelation function** *is defined, analogously to (1.25), as*

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}. \tag{1.37}$$

The sample autocorrelation function has a sampling distribution that allows us to assess whether the data comes from a completely random or white series or whether correlations are statistically significant at some lags. Precise details are given in Theorem A.7 in Appendix A. We have

Property P1.1: Large Sample Distribution of the ACF
Under general conditions, if x_t *is white noise, then for n large, the sample ACF,* $\widehat{\rho}_x(h)$*, for* $h = 1, 2, \ldots, H$*, where H is fixed but arbitrary, is approximately normally distsributed with zero mean and standard deviation given by*

$$\sigma_{\hat{\rho}_x(h)} = \frac{1}{\sqrt{n}}. \tag{1.38}$$

Based on the above result, we obtain a rough method of assessing whether peaks in $\widehat{\rho}(h)$ are significant by determining whether the observed peak is

outside the interval $\pm 2/\sqrt{n}$ (or plus/minus two standard errors); for a white noise sequence, approximately 95% of the sample ACFs should be within these limits. The applications of this property develop because many statistical modeling procedures depend on reducing a time series to a white noise series by various kinds of transformations. After such a procedure is applied, the plotted ACFs of the residuals should then lie roughly within the limits given above.

Definition 1.16 *The estimators for the cross-covariance function, $\gamma_{xy}(h)$, as given in (1.28) and the cross-correlation, $\rho_{xy}(h)$, in (1.29), are given, respectively, by the* **sample cross-covariance function**

$$\widehat{\gamma}_{xy}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(y_t - \bar{y}), \tag{1.39}$$

where $\widehat{\gamma}_{xy}(-h) = \widehat{\gamma}_{yx}(h)$ determines the function for negative lags, and the **sample cross-correlation function**

$$\widehat{\rho}_{xy}(h) = \frac{\widehat{\gamma}_{xy}(h)}{\sqrt{\widehat{\gamma}_x(0)\widehat{\gamma}_y(0)}}. \tag{1.40}$$

The sample cross-correlation function can be examined graphically as a function of lag h to search for leading or lagging relations in the data using the property mentioned in Example 1.22 for the theoretical cross-covariance function. Because $-1 \leq \widehat{\rho}_{xy}(h) \leq 1$, the practical importance of peaks can be assessed by comparing their magnitudes with their theoretical maximum values. Furthermore, for x_t and y_t independent linear processes of the form (1.31), we have

Property P1.2: Large Sample Distribution of the Cross-Correlation Under Independence
The large sample distribution of $\widehat{\rho}_{xy}(h)$ is normal with mean zero and

$$\sigma_{\hat{\rho}_{xy}} = \frac{1}{\sqrt{n}} \tag{1.41}$$

if at least one of the processes is white independent noise (see Theorem A.8 in Appendix A).

Example 1.23 A Simulated Time Series

To give an example of the procedure for calculating numerically the autocovariance and cross-covariance functions, consider a contrived set of data generated by tossing a fair coin, letting $x_t = 1$ when a head is obtained and $x_t = -1$ when a tail is obtained. Construct y_t as

$$y_t = 5 + x_t - .7x_{t-1}. \tag{1.42}$$

Table 1.1 Sample Realization of the Contrived Series y_t.

t	1	2	3	4	5	6	7	8	9	10
Coin	H	H	T	H	T	T	T	H	T	H
x_t	1	1	-1	1	-1	-1	-1	1	-1	1
y_t	6.7	5.3	3.3	6.7	3.3	4.7	4.7	6.7	3.3	6.7
$y_t - \bar{y}$	1.56	.16	-1.84	1.56	-1.84	$-.44$	$-.44$	1.56	-1.84	1.56

Table 1.1 shows sample realizations of the appropriate processes with $x_0 = -1$ and $n = 10$.

The sample autocorrelation for the series y_t can be calculated using (1.36) and (1.37) for $h = 0, 1, 2, \ldots$. It is not necessary to calculate for negative values because of the symmetry. For example, for $h = 3$, the autocorrelation becomes the ratio of

$$\begin{aligned}
\widehat{\gamma}_y(3) &= 10^{-1}\sum_{t=1}^{7}(y_{t+3} - \bar{y})(y_t - \bar{y}) \\
&= 10^{-1}\Big[(1.56)(1.56) + (-1.84)(.16) + (-.44)(-1.84) \\
&\quad + (-.44)(1.56) + (1.56)(-1.84) + (-1.84)(-.44) \\
&\quad + (1.56)(-.44)\Big] \\
&= -.04848
\end{aligned}$$

to

$$\widehat{\gamma}_y(0) = \frac{1}{10}[(1.56)^2 + (.16)^2 + \cdots + (1.56)^2] = 2.0304$$

so that

$$\widehat{\rho}_y(3) = \frac{-.04848}{2.0304} = -.02388.$$

The theoretical ACF can be obtained from the model (1.42) using the fact that the mean of x_t is zero and the variance of x_t is one. It can be shown that

$$\rho_y(1) = \frac{-.7}{1 + .7^2} = -.47$$

and $\rho_y(h) = 0$ for $|h| > 1$ (Problem 1.23). Table 1.2 compares the theoretical ACF with sample ACFs for a realization where $n = 10$ and another realization where $n = 100$; we note the increased variability in the smaller size sample.

Table 1.2 Theoretical and Sample ACFs for $n = 10$ and $n = 100$

h	$\rho_y(h)$	$\widehat{\rho}_y(h)$ $n = 10$	$\widehat{\rho}_y(h)$ $n = 100$
0	1.00	1.00	1.00
±1	−.47	−.55	−.45
±2	.00	.17	−.12
±3	.00	−.02	.14
±4	.00	.15	.01
±5	.00	−.46	−.01

Example 1.24 ACF of Speech Signal

Computing the sample ACF as in the previous example can be thought of as matching the time series h units in the future, say, x_{t+h} against itself, x_t. Figure 1.13 shows the ACF of the speech series of Figure 1.3. The original series appears to contain a sequence of repeating short signals. The ACF confirms this behavior, showing repeating peaks spaced at about 106-109 points. Autocorrelation functions of the short signals appear, spaced at the intervals mentioned above. The distance between the repeating signals is known as the pitch period and is a fundamental parameter of interest in systems that encode and decipher speech. Because the series is sampled at 10,000 points per second, the pitch period appears to be between .0106 and .0109 seconds.

To compute the sample ACF in R, use

```
> speech = scan("/mydata/speech.dat")
> acf(speech,250)
```

Example 1.25 Correlation Analysis of SOI and Recruitment Data

The autocorrelation and cross-correlation functions are also useful for analyzing the joint behavior of two stationary series whose behavior may be related in some unspecified way. In Example 1.5 (see Figure 1.5), we have considered simultaneous monthly readings of the SOI and the number of new fish (Recruitment) computed from a model. Figure 1.14 shows the autocorrelation and cross-correlation functions (ACFs and CCF) for these two series. Both of the ACFs exhibit periodicities corresponding to the correlation between values separated by 12 units. Observations 12 months or one year apart are strongly positively correlated, as are observations at multiples such as $24, 36, 48, \ldots$ Observations separated by six months are negatively correlated, showing that positive excursions tend to be associated with negative excursions six months removed. This appearance is rather characteristic of the pattern that would be produced by

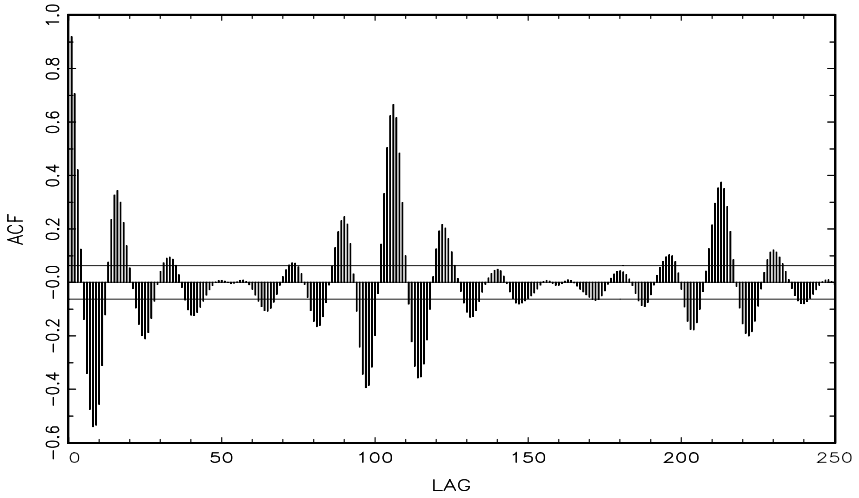


Figure 1.13 ACF of the speech series.

a sinusoidal component with a period of 12 months. The cross-correlation function peaks at $h = -6$, showing that the SOI measured at time $t - 6$ months is associated with the Recruitment series at time t. We could say the SOI leads the Recruitment series by six months. The sign of the ACF is negative, leading to the conclusion that the two series move in different directions, i.e., increases in SOI lead to decreases in Recruitment and vice versa. Again, note the periodicity of 12 months in the CCF. The flat lines shown on the plots indicate $\pm 2/\sqrt{453}$, so that upper values would be exceeded about 2.5% of the time if the noise were white [see (1.38) and (1.41)].

To reproduce Figure 1.14 in R, use the following commands.

```
> soi=scan("/mydata/soi.dat")
> rec=scan("/mydata/recruit.dat")
> par(mfrow=c(3,1))
> acf(soi, 50)
> acf(rec, 50)
> ccf(soi, rec, 50)
```

1.7 Vector-Valued and Multidimensional Series

We frequently encounter situations in which the relationships between a number of jointly measured time series are of interest. For example, in the previous sections, we considered discovering the relationships between the SOI and Recruitment series. Hence, it will be useful to consider the notion of a vector time

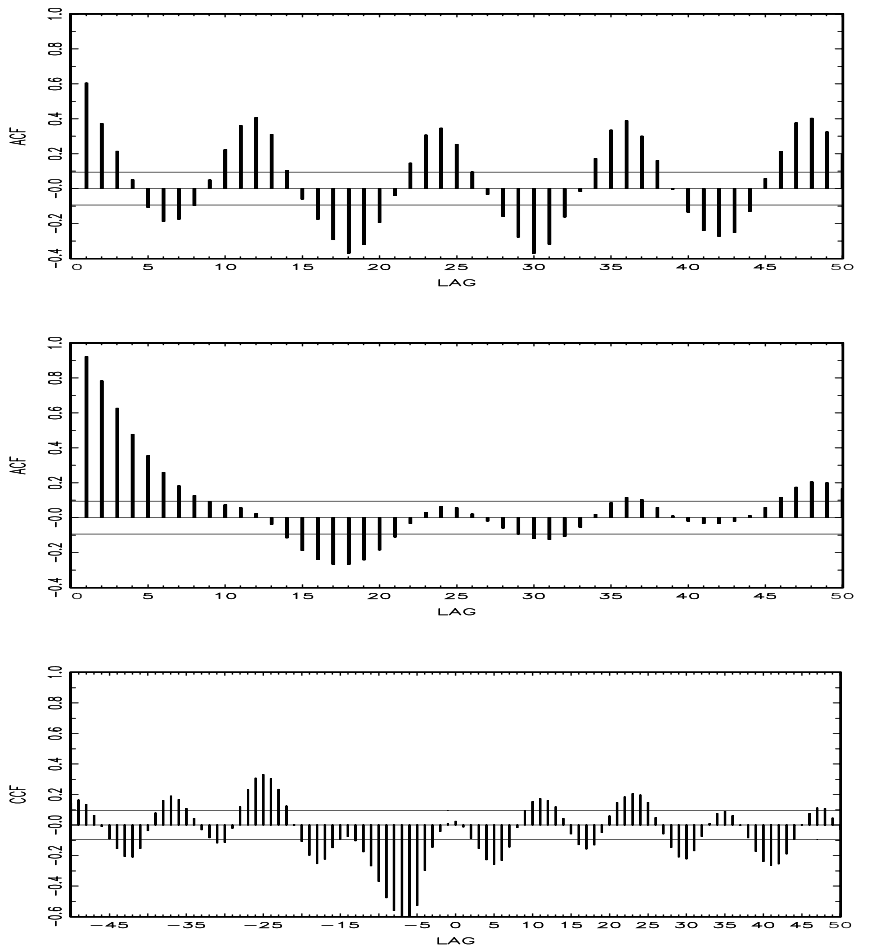


Figure 1.14 Sample ACFs of the SOI series (top) and of the Recruitment series (middle), and the sample CCF of the two series (bottom); negative lags indicate SOI leads Recruitment.

series $\boldsymbol{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tp})'$, which contains as its components p univariate time series. We denote the $p \times 1$ column vector of the observed series as $\boldsymbol{x}_t$. The row vector $\boldsymbol{x}_t'$ is its transpose. For the stationary case, the $p \times 1$ mean vector

$$\boldsymbol{\mu} = E(\boldsymbol{x}_t) \tag{1.43}$$

of the form $\boldsymbol{\mu} = (\mu_{t1}, \mu_{t2}, \ldots, \mu_{tp})'$ and the $p \times p$ autocovariance matrix

$$\Gamma(h) = E[(\boldsymbol{x}_{t+h} - \boldsymbol{\mu})(\boldsymbol{x}_t - \boldsymbol{\mu})'] \tag{1.44}$$

can be defined, where the elements of the matrix $\Gamma(h)$ are the cross-covariance

functions

$$\gamma_{ij}(h) = E[(x_{t+h,i} - \mu_i)(x_{tj} - \mu_j)] \tag{1.45}$$

for $i, j = 1, \ldots, p$. Because $\gamma_{ij}(h) = \gamma_{ji}(-h)$, it follows that

$$\Gamma(-h) = \Gamma'(h). \tag{1.46}$$

Now, the sample autocovariance matrix of the vector series $\boldsymbol{x}_t$ is the $p \times p$ matrix of sample cross-covariances, defined as

$$\widehat{\Gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (\boldsymbol{x}_{t+h} - \bar{\boldsymbol{x}})(\boldsymbol{x}_t - \bar{\boldsymbol{x}})', \tag{1.47}$$

where

$$\bar{\boldsymbol{x}} = n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_t \tag{1.48}$$

denotes the $p \times 1$ sample mean vector. The symmetry property of the theoretical autocovariance (1.46) extends to the sample autocovariance (1.47), which is defined for negative values by taking

$$\widehat{\Gamma}(-h) = \widehat{\Gamma}(h)'. \tag{1.49}$$

In many applied problems, an observed series may be indexed by more than time alone. For example, the position in space of an experimental unit might be described by two coordinates, say, s_1 and s_2. We may proceed in these cases by defining a multidimensional process $x_{\boldsymbol{s}}$ as a function of the $r \times 1$ vector $\boldsymbol{s} = (s_1, s_2, \ldots, s_r)'$ where s_i denotes the coordinate of the i^{th} index.

Example 1.26 Soil Surface Temperatures

As an example, the two-dimensional ($r = 2$) temperature series x_{s_1,s_2} in Figure 1.15 is indexed by a row number s_1 and a column number s_2 that represent positions on a 64×36 spatial grid set out on an agricultural field. The value of the temperature measured at row s_1 and column s_2, is denoted by $x_{\boldsymbol{s}} = x_{s1,s2}$. We can note from the two-dimensional plot that a distinct change occurs in the character of the two-dimensional surface starting at about row 40, where the oscillations along the row axis become fairly stable and periodic. For example, averaging over the 36 columns, we may compute an average value for each s_1 as in Figure 1.16. It is clear that the noise present in the first part of the two-dimensional series is nicely averaged out, and we see a clear and consistent temperature signal.

The autocovariance function of a stationary multidimensional process, $x_{\boldsymbol{s}}$, can be defined as a function of the multidimensional lag vector, say, $\boldsymbol{h} = (h_1, h_2, \ldots, h_r)'$, as

$$\gamma(\boldsymbol{h}) = E[(x_{\boldsymbol{s}+\boldsymbol{h}} - \mu)(x_{\boldsymbol{s}} - \mu)], \tag{1.50}$$

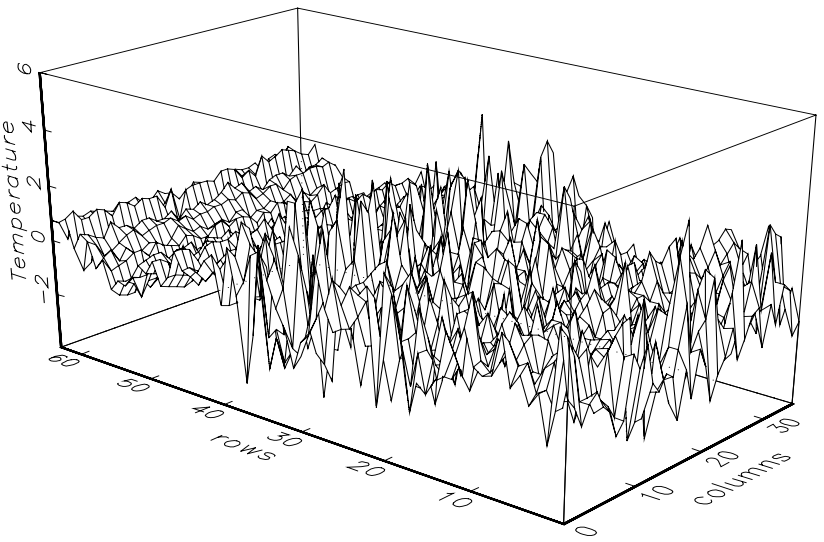


Figure 1.15 Two-dimensional time series of temperature measurements taken on a rectangular field (64×36 with 17-foot spacing). Data are from Bazza et al. (1988).

where

$$\mu = E(x_{\boldsymbol{s}}) \tag{1.51}$$

does not depend on the spatial coordinate $\boldsymbol{s}$. For the two dimensional temperature process, (1.50) becomes

$$\gamma(h_1, h_2) = E[(x_{s_1+h_1, s_2+h_2} - \mu)(x_{s_1,s_2} - \mu)], \tag{1.52}$$

which is a function of lag, both in the row (h_1) and column (h_2) directions.

The multidimensional sample autocovariance function is defined as

$$\widehat{\gamma}(\boldsymbol{h}) = (S_1 S_2 \cdots S_r)^{-1} \sum_{s_1} \sum_{s_2} \cdots \sum_{s_r} (x_{\boldsymbol{s}+\boldsymbol{h}} - \bar{x})(x_{\boldsymbol{s}} - \bar{x}), \tag{1.53}$$

where $\boldsymbol{s} = (s_1, s_2, \ldots, s_r)'$ and the range of summation for each argument is $1 \le s_i \le S_i - h_i$, for $i = 1, \ldots, r$. The mean is computed over the r-dimensional array, that is,

$$\bar{x} = (S_1 S_2 \cdots S_r)^{-1} \sum_{s_1} \sum_{s_2} \cdots \sum_{s_r} x_{s_1, s_2, \cdots, s_r}, \tag{1.54}$$

where the arguments s_i are summed over $1 \le s_i \le S_i$. The multidimensional sample autocorrelation function follows, as usual, by taking the scaled ratio

$$\widehat{\rho}(\boldsymbol{h}) = \frac{\widehat{\gamma}(\boldsymbol{h})}{\widehat{\gamma}(\boldsymbol{0})}. \tag{1.55}$$

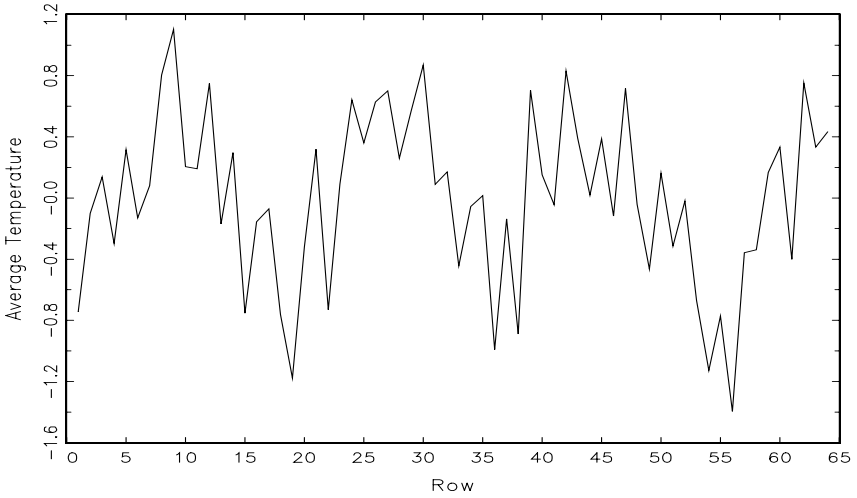


Figure 1.16 Row averages of the two-dimensional soil temperature profile. $\bar{x}_{s_1} = \sum_{s_2} x_{s_1,s_2}/36$.

Example 1.27 Sample ACF of the Soil Temperature Series

The autocorrelation function of the two-dimensional temperature process can be written in the form

$$\widehat{\rho}(h_1, h_2) = \frac{\widehat{\gamma}(h_1, h_2)}{\widehat{\gamma}(0, 0)},$$

where

$$\widehat{\gamma}(h_1, h_2) = (S_1 S_2)^{-1} \sum_{s_1} \sum_{s_2} (x_{s_1+h_1, s_2+h_2} - \bar{x})(x_{s_1,s_2} - \bar{x})$$

Figure 1.17 shows the autocorrelation function for the temperature data, and we note the systematic periodic variation that appears along the rows. The autocovariance over columns seems to be strongest for $h_1 = 0$, implying columns may form replicates of some underlying process that has a periodicity over the rows. This idea can be investigated by examining the mean series over columns as shown in Figure 1.16.

The sampling requirements for multidimensional processes are rather severe because values must be available over some uniform grid in order to compute the ACF. In some areas of application, such as in soil science, we may prefer to sample a limited number of rows or *transects* and hope these are essentially replicates of the basic underlying phenomenon of interest. One-dimensional methods can then be applied. When observations are irregular in time space, modifications to the estimators need to be made. Systematic approaches to the

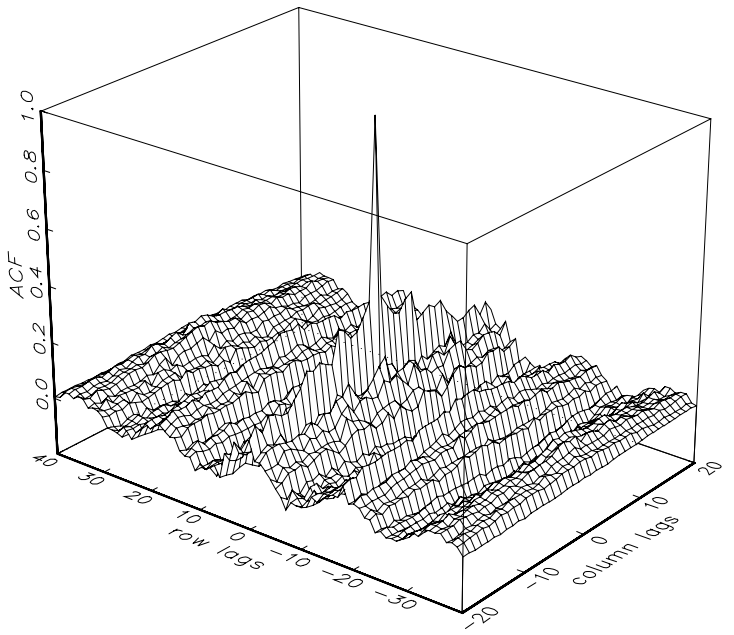


Figure 1.17 Two-dimensional autocorrelation function for the soil temperature data.

problems introduced by irregularly spaced observations have been developed by Journel and Huijbregts (1978) or Cressie (1993). We shall not pursue such methods in detail here, but it is worth noting that the introduction of the variogram

$$2V_x(\boldsymbol{h}) = \mathrm{var}\{x_{\boldsymbol{s}+\boldsymbol{h}} - x_{\boldsymbol{s}}\} \tag{1.56}$$

and its sample estimator

$$2\widehat{V}_x(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})} \sum_{\boldsymbol{s}} (x_{\boldsymbol{s}+\boldsymbol{h}} - x_{\boldsymbol{s}})^2 \tag{1.57}$$

play key roles, where $N(\boldsymbol{h})$ denotes both the number of points located within $\boldsymbol{h}$, and the sum runs over the points in the neighborhood. Clearly, substantial indexing difficulties will develop from estimators of the kind, and often it will be difficult to find non-negative definite estimators for the covariance function. Problem 1.26 investigates the relation between the variogram and the autocovariance function in the stationary case.

Problems

Section 1.2

1.1 To compare the earthquake and explosion signals, plot the data displayed in Figure 1.7 on the same graph using different colors or different line types and comment on the results.

1.2 Consider a signal plus noise model of the general form $x_t = s_t + w_t$, where w_t is Gaussian white noise with $\sigma_w^2 = 1$. Simulate and plot $n = 200$ observations from each of the following two models (*Save the data generated here for use in Problem 1.21*):

(a) $x_t = s_t + w_t$, for $t = 1, ..., 200$, where

$$s_t = \begin{cases} 0, & t = 1, ..., 100 \\ 10 \exp\{-\frac{(t-100)}{20}\} \cos(2\pi t/4), & t = 101, ..., 200. \end{cases}$$

(b) $x_t = s_t + w_t$, for $t = 1, ..., 200$, where

$$s_t = \begin{cases} 0, & t = 1, ..., 100 \\ 10 \exp\{-\frac{(t-100)}{200}\} \cos(2\pi t/4), & t = 101, ..., 200. \end{cases}$$

(c) Compare the general appearance of the series (a) and (b) with the earthquake series and the explosion series shown in Figure 1.7. In addition, plot (or sketch) and compare the signal modulators (a) $\exp\{-t/20\}$ and (b) $\exp\{-t/200\}$, for $t = 1, 2, ..., 100$.

Section 1.3

1.3 (a) Generate $n = 100$ observations from the autoregression

$$x_t = -.9x_{t-2} + w_t$$

with $\sigma_w = 1$, using the method described in Example 1.10. Next, apply the moving average filter

$$v_t = (x_t + x_{t-1} + x_{t-2} + x_{t-3})/4$$

to x_t, the data you generated. Now plot x_t as a line and superimpose v_t as a dashed line. Comment on the behavior of x_t and how applying the moving average filter changes that behavior.

(b) Repeat (a) but with

$$x_t = \cos(2\pi t/4).$$

(c) Repeat (b) but with added $\mathrm{N}(0, 1)$ noise,

$$x_t = \cos(2\pi t/4) + w_t.$$

(d) Compare and contrast (a)–(c).

Section 1.4

1.4 Show that the autocovariance function can be written as

$$\gamma(s,t) = E[(x_s - \mu_s)(x_t - \mu_t)] = E(x_s x_t) - \mu_s \mu_t,$$

where $E[x_t] = \mu_t$.

1.5 For the two series, x_t, in Problem 1.2 (a) and (b):

(a) compute and sketch the mean functions $\mu_x(t)$; for $t = 1, \ldots, 200$.

(b) calculate the autocovariance functions, $\gamma_x(s,t)$, for $s, t = 1, \ldots, 200$.

Section 1.5

1.6 Consider the time series

$$x_t = \beta_1 + \beta_2 t + w_t,$$

where β_1 and β_2 are known constants and w_t is a white noise process with variance σ_w^2.

(a) Determine whether x_t is stationary.

(b) Show that the process $y_t = x_t - x_{t-1}$ is stationary.

(c) Show that the mean of the moving average

$$v_t = \frac{1}{2q+1} \sum_{j=-q}^{q} x_{t-j}$$

is $\beta_1 + \beta_2 t$, and give a simplified expression for the autocovariance function.

1.7 For a moving average process of the form

$$x_t = w_{t-1} + 2w_t + w_{t+1},$$

where w_t are independent with zero means and variance σ_w^2, determine the autocovariance and autocorrelation functions as a function of lag $h = s - t$ and plot.

1.8 Consider the randow walk with drift model

$$x_t = \delta + x_{t-1} + w_t,$$

for $t = 1, 2, \ldots$, with $x_0 = 0$, where w_t is white noise with variance σ_w^2.

(a) Show that the model can be written as $x_t = \delta t + \sum_{k=1}^{t} w_k$.

(b) Find the mean function and the autocovariance function of x_t

(c) Show $\rho_x(t-1,t) = \sqrt{\frac{t-1}{t}} \to 1$ as $t \to \infty$. What is the implication of this result?

(d) Show that the series is not stationary.

(e) Suggest a transformation to make the series stationary, and prove that the transformed series is stationary. (Hint: See Problem 1.6b.)

1.9 A time series with a periodic component can be constructed from

$$x_t = U_1 \sin(2\pi\omega_0 t) + U_2 \cos(2\pi\omega_0 t),$$

where U_1 and U_2 are independent random variables with zero means and $E(U_1^2) = E(U_2^2) = \sigma^2$. The constant ω_0 determines the period or time it takes the process to make one complete cycle. Show that this series is weakly stationary with autocovariance function

$$\gamma(h) = \sigma^2 \cos(2\pi\omega_0 h).$$

1.10 Suppose we would like to predict a single stationary series x_t with zero mean and autocorrelation function $\gamma(h)$ at some time in the future, say, $t+\ell$, for $\ell > 0$.

(a) If we predict using only x_t and some scale multiplier A, show that the mean-square prediction error

$$MSE(A) = E[(x_{t+\ell} - Ax_t)^2]$$

is minimized by the value

$$A = \rho(\ell).$$

(b) Show that the minimum mean-square prediction error is

$$MSE(A) = \gamma(0)[1 - \rho^2(\ell)].$$

(c) Show that if $x_{t+\ell} = Ax_t$, then $\rho(\ell) = 1$ if $A > 0$, and $\rho(\ell) = -1$ if $A < 0$.

1.11 Consider the linear process defined in (1.31).

(a) Verify that the autocovariance function of the process is given by (1.33). Use the result to verify your answer to Problem 1.7.

(b) Show that x_t exists as a limit in mean square (see Appendix A) if (1.32) holds.

1.12 For two weakly stationary series x_t and y_t, verify (1.30).

1.13 Consider the two series

$$x_t = w_t$$
$$y_t = w_t - \theta w_{t-1} + u_t,$$

where w_t and u_t are independent white noise series with variances σ_w^2 and σ_u^2, respectively, and θ is an unspecified constant.

(a) Express the ACF, $\rho_y(h)$, for $h = 0, \pm 1, \pm 2, \ldots$ of the series y_t as a function of σ_w^2, σ_u^2, and θ.

(b) Determine the CCF, $\rho_{xy}(h)$ relating x_t and y_t.

(c) Show that x_t and y_t are jointly stationary.

1.14 Let x_t be a stationary normal process with mean μ_x and autocovariance function $\gamma(h)$. Define the nonlinear time series

$$y_t = \exp\{x_t\}.$$

(a) Express the mean function $E(y_t)$ in terms of μ_x and $\gamma(0)$. The moment generating function of a normal random variable x with mean μ and variance σ^2 is

$$M_x(\lambda) = E[\exp\{\lambda x\}] = \exp\left\{\mu\lambda + \frac{1}{2}\sigma^2\lambda^2\right\}.$$

(b) Determine the autocovariance function of y_t. The sum of the two normal random variables $x_{t+h} + x_t$ is still a normal random variable.

1.15 Let w_t, for $t = 0, \pm 1, \pm 2, \ldots$ be a normal white noise process, and consider the series

$$x_t = w_t w_{t-1}.$$

Determine the mean and autocovariance function of x_t, and state whether it is stationary.

1.16 Consider the series

$$x_t = \sin(2\pi U t),$$

$t = 1, 2, \ldots$, where U has a uniform distribution on the interval $(0, 1)$.

(a) Prove x_t is weakly stationary.

(b) Prove x_t is not strictly stationary. [Hint: consider the joint bivariate cdf (1.19) at the points $t = 1, s = 2$ with $h = 1$, and find values of c_t, c_s where strict stationarity does not hold.]

1.17 Suppose we have the linear process x_t generated by

$$x_t = w_t - \theta w_{t-1},$$

$t = 0, 1, 2, \ldots$, where $\{w_t\}$ is independent and identically distributed with characteristic function $\phi_w(\cdot)$, and θ is a fixed constant.

(a) Express the joint characteristic function of $x_1, x_2, \ldots, x_n$, say,

$$\phi_{x_1,x_2,\ldots,x_n}(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

in terms of $\phi_w(\cdot)$.

(b) Deduce from (a) that x_t is strictly stationary.

1.18 Suppose that x_t is a linear process of the form (1.31) satisfying the absolute summability condition (1.32). Prove

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty.$$

Section 1.6

1.19 (a) Simulate a series of $n = 500$ Gaussian white noise observations as in Example 1.8 and compute the sample ACF, $\widehat{\rho}(h)$, to lag 20. Compare the sample ACF you obtain to the actual ACF, $\rho(h)$. [Recall Example 1.19.]

(b) Repeat part (a) using only $n = 50$. How does changing n affect the results?

1.20 (a) Simulate a series of $n = 500$ moving average observations as in Example 1.9 and compute the sample ACF, $\widehat{\rho}(h)$, to lag 20. Compare the sample ACF you obtain to the actual ACF, $\rho(h)$. [Recall Example 1.20.]

(b) Repeat part (a) using only $n = 50$. How does changing n affect the results?

1.21 Although the model in Problem 1.2 is not stationary (Why?), the sample ACF can be informative. For the data you generated in that problem, calculate and plot the sample ACF, and then comment.

1.22 Simulate a series of $n = 500$ observations from the signal-plus-noise model presented in Example 1.12 with $\sigma_w^2 = 1$. Compute the sample ACF to lag 100 of the data you generated and comment.

1.23 For the time series y_t described in Example 1.23, verify the stated result that $\rho_y(1) = -.47$ and $\rho_y(h) = 0$ for $h > 1$.

1.24 A real-valued function $g(t)$, defined on the integers, is non-negative definite if and only if

$$\sum_{s=1}^{n}\sum_{t=1}^{n} a_s g(s-t) a_t \geq 0$$

for all positive integers n and for all vectors $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)'$. For the matrix $G = \{g(s-t), s, t = 1, 2, \ldots, n\}$, this implies that $\boldsymbol{a}'G\boldsymbol{a} \geq 0$ for all vectors $\boldsymbol{a}$.

(a) Prove that $\gamma(h)$, the autocovariance function of a stationary process, is a non-negative definite function.

(b) Verify that the sample autocovariance $\widehat{\gamma}(h)$ is a non-negative definite function.

Section 1.7

1.25 Consider a collection of time series $x_{1t}, x_{2t}, \ldots, x_{Nt}$ that are observing some common signal μ_t observed in noise processes $e_{1t}, e_{2t}, \ldots, e_{Nt}$, with a model for the j-th observed series given by

$$x_{jt} = \mu_t + e_{jt}.$$

Suppose the noise series have zero means and are uncorrelated for different j. The common autocovariance functions of all series are given by $\gamma_e(s,t)$. Define the sample mean

$$\bar{x}_t = \frac{1}{N}\sum_{j=1}^{N} x_{jt}.$$

(a) Show that $E[\bar{x}_t] = \mu_t$.

(b) Show that $E[(\bar{x}_t - \mu)^2)] = N^{-1}\gamma_e(t,t)$.

(c) How can we use the results in estimating the common signal?

1.26 A concept used in geostatistics, see Journel and Huijbregts (1978) or Cressie (1993), is that of the variogram, defined for a spatial process $x_{\boldsymbol{s}}$, $\boldsymbol{s} = (s_1, s_2)$, for $s_1, s_2 = 0, \pm 1, \pm 2, ...$, as

$$V_x(\boldsymbol{h}) = \frac{1}{2}E[(x_{\boldsymbol{s}+\boldsymbol{h}} - x_{\boldsymbol{s}})^2],$$

where $\boldsymbol{h} = (h_1, h_2)$, for $h_1, h_2 = 0, \pm 1, \pm 2, ...$ Show that, for a stationary process, the variogram and autocovariance functions can be related through

$$V_x(\boldsymbol{h}) = \gamma(\boldsymbol{0}) - \gamma(\boldsymbol{h}),$$

where $\gamma(\boldsymbol{h})$ is the usual lag $\boldsymbol{h}$ covariance function and $\boldsymbol{0} = (0,0)$. Note the easy extension to any spatial dimension.

The following problems require the supplemental material given in Appendix A

1.27 Suppose $x_t = \beta_0 + \beta_1 t$, where β_0 and β_1 are constants. Prove as $n \to \infty$, $\widehat{\rho}_x(h) \to 1$ for fixed h, where $\widehat{\rho}_x(h)$ is the ACF (1.37).

1.28 (a) Suppose x_t is a weakly stationary time series with mean zero and with absolutely summable autocovariance function, $\gamma(h)$, such that

$$\sum_{h=-\infty}^{\infty} \gamma(h) = 0.$$

Prove that $\sqrt{n}\ \bar{x} \xrightarrow{p} 0$, where $\bar{x}$ is the sample mean (1.35).

(b) Give an example of a process that satisfies the conditions of part (a). What is special about this process?

1.29 Let x_t be a linear process of the form (A.44)–(A.45). If we define

$$\tilde{\gamma}(h) = n^{-1} \sum_{t=1}^{n} (x_{t+h} - \mu_x)(x_t - \mu_x),$$

show that

$$n^{1/2}\big(\tilde{\gamma}(h) - \widehat{\gamma}(h)\big) = o_p(1).$$

Hint: The Markov Inequality

$$P\{|x| \geq \epsilon\} < \frac{E|x|}{\epsilon}$$

can be helpful for the cross-product terms.

1.30 For a linear process of the form

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j},$$

where $\{w_t\}$ satisfies the conditions of Theorem A.7 and $|\phi| < 1$, show that

$$\sqrt{n}\frac{(\widehat{\rho}_x(1) - \rho_x(1))}{\sqrt{1 - \rho_x^2(1)}} \xrightarrow{d} N(0, 1),$$

and construct a 95% confidence interval for ϕ when $\widehat{\rho}_x(1) = .64$ and $n = 100$.

1.31 Let $\{x_t;\ t = 0, \pm 1, \pm 2, \ldots\}$ be iid $(0, \sigma^2)$.

(a) For $h \geq 1$ and $k \geq 1$, show that $x_t x_{t+h}$ and $x_s x_{s+k}$ are uncorrelated for all $s \neq t$.

(b) For fixed $h \geq 1$, show that the $h \times 1$ vector

$$\sigma^{-2} n^{-1/2} \sum_{t=1}^{n} (x_t x_{t+1}, \ldots, x_t x_{t+h})' \xrightarrow{d} (z_1, \ldots, z_h)'$$

where $z_1, \ldots, z_h$ are iid $\mathrm{N}(0,1)$ random variables. [Note: the sequence $\{x_t x_{t+h};\ t = 1, 2, \ldots\}$ is h-dependent and white noise $(0, \sigma^4)$. Also, recall the Cramér-Wold device.]

(c) Show, for each $h \geq 1$,

$$n^{-1/2}\left[\sum_{t=1}^{n} x_t x_{t+h} - \sum_{t=1}^{n-h}(x_t - \bar{x})(x_{t+h} - \bar{x})\right] \xrightarrow{p} 0 \qquad \text{as } n \to \infty$$

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

(d) Noting that $n^{-1}\sum_{t=1}^{n} x_t^2 \xrightarrow{p} \sigma^2$, conclude that

$$n^{1/2}\left[\widehat{\rho}(1), \ldots, \widehat{\rho}(h)\right]' \xrightarrow{d} (z_1, \ldots, z_h)'$$

where $\widehat{\rho}(h)$ is the sample ACF of the data $x_1, \ldots, x_n$.

Chapter 2

Time Series Regression and Exploratory Data Analysis

2.1 Introduction

The linear model and its applications are at least as dominant in the time series context as in classical statistics. Regression models are important for time domain models discussed in Chapters 3, 5, and 6, and in the frequency domain models considered in Chapters 4 and 7. The primary ideas depend on being able to express a response series, say x_t, as a linear combination of inputs, say $z_{t1}, z_{t2}, \ldots, z_{tq}$. Estimating the coefficients $\beta_1, \beta_2, \ldots, \beta_q$ in the linear combinations by least squares provides a method for modeling x_t in terms of the inputs.

In the time domain applications of Chapter 3, for example, we will express x_t as a linear combination of previous values $x_{t-1}, x_{t-2}, \ldots, x_p$, of the currently observed series. The outputs x_t may also depend on lagged values of another series, say $y_{t-1}, y_{t-2}, \ldots, y_{t-q}$, that have influence. It is easy to see that forecasting becomes an option when prediction models can be formulated in this form. Time series smoothing and filtering can be expressed in terms of local regression models. Polynomials and regression splines also provide important techniques for smoothing.

If one admits sines and cosines as inputs, the frequency domain ideas that lead to the periodogram and spectrum of Chapter 4 follow from a regression model. Extensions to filters of infinite extent can be handled using regression in the frequency domain. In particular, many regression problems in the frequency domain can be carried out as a function of the periodic components of the input and output series, providing useful scientific intuition into fields like acoustics, oceanographics, engineering, biomedicine, and geophysics.

The above considerations motivate us to include a separate chapter on re-

gression and some of its applications that is written on an elementary level and is formulated in terms of time series. The assumption of linearity, stationarity, and homogeneity of variances over time is critical in the regression context, and therefore we include some material on transformations and other techniques useful in exploratory data analysis.

2.2 Classical Regression in the Time Series Context

We begin our discussion of linear regression in the time series context by assuming some output or dependent time series, say, x_t, for $t = 1, \ldots, n$, is being influenced by a collection of possible inputs or independent series, say, $z_{t1}, z_{t2}, \ldots, z_{tq}$, where we first regard the inputs as fixed and known. This assumption, necessary for applying conventional linear regression, will be relaxed later on. We express this relation through the linear regression model

$$x_t = \beta_1 z_{t1} + \beta_2 z_{t2} + \cdots + \beta_q z_{tq} + w_t, \tag{2.1}$$

where $\beta_1, \beta_2, \ldots, \beta_q$ are unknown fixed regression coefficients, and $\{w_t\}$ is a random error or noise process consisting of independent and identically distributed (iid) normal variables with mean zero and variance σ_w^2; we will relax the iid assumption later. A more general setting within which to embed mean square estimation and linear regression is given in Appendix B, where we introduce Hilbert spaces and the Projection Theorem.

Example 2.1 Estimating a Trend

Consider the global temperature data, say x_t, shown in Figure 1.2. As discussed in Example 1.2, there is an apparent upward trend in the series that has been used to argue the global warming hypothesis. We might use simple linear regression to estimate that trend by fitting the model

$$x_t = \beta_1 + \beta_2 t + w_t, \quad t = 1900, 1901, \ldots, 1997.$$

This is in the form of the regression model (2.1) when we make the identification $q = 2$, $z_{t1} = 1, z_{t2} = t$. Note that we are making the assumption that the errors, w_t, are an iid normal sequence, which may not be true. We will address this problem further in §2.3; the problem of autocorrelated errors is discussed in detail in §5.5. Also note that we could have used, e.g., $t = 0, \ldots, 97$, without affecting the interpretation of the slope coefficient, β_2; only the intercept, β_1, would be affected.

Using simple linear regression, we obtained the estimated coefficients $\widehat{\beta}_1 = -12.186$, and $\widehat{\beta}_2 = .006$ (with a standard error of .0005) yielding a significant estimated increase of .6 degrees centigrade per 100 years. We

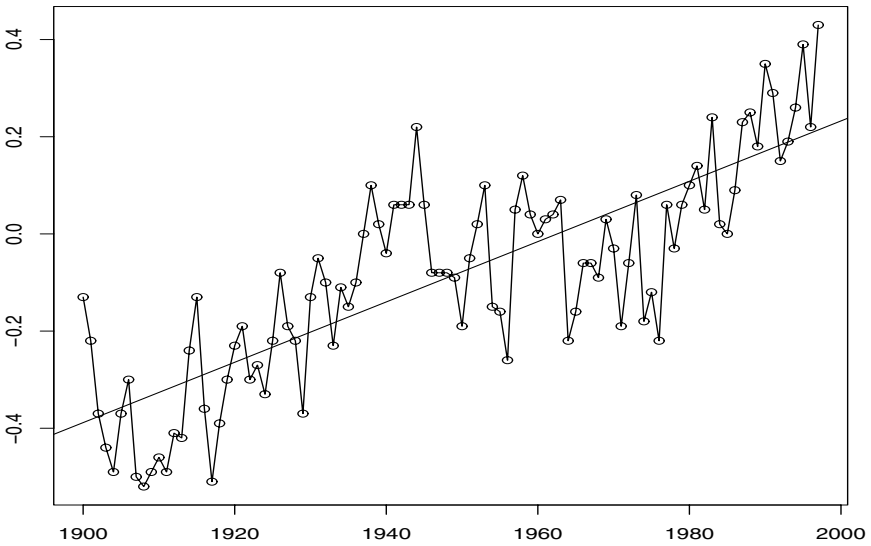


Figure 2.1 Global temperature deviations shown in Figure 1.2 with fitted linear trend line.

discuss the precise way in which the solution was accomplished below. Finally, Figure 2.1 shows the global temperature data, say x_t, with the estimated trend, say $\widehat{x}_t = -12.186 + .006t$, superimposed. It is apparent that the estimated trend line obtained via simple linear regression does not quite capture the trend of the data and better models will be needed.

To perform this analysis in R, we note that the data file `globtemp.dat` has 142 observations starting from the year 1856. We are only using the final 98 observations corresponding to the years 1900 to 1997.

```
> gtemp = scan("/mydata/globtemp.dat")
> x = gtemp[45:142]
> t = 1900:1997
> fit=lm(x~t)    # regress x on t
> summary(fit)   # regression output
> plot(t,x, type="o", xlab="year", ylab="temp deviation")
> abline(fit)    # add regression line to the plot
```

The linear model described by (2.1) above can be conveniently written in a more general notation by defining the column vectors $\boldsymbol{z}_t = (z_{t1}, z_{t2}, \ldots, z_{tq})'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$, where $'$ denotes transpose, so (2.1) can be written in the alternate form

$$x_t = \boldsymbol{\beta}' \boldsymbol{z}_t + w_t. \tag{2.2}$$

where $w_t \sim \text{iid}(0, \sigma_w^2)$. It is natural to consider estimating the unknown coef-

ficient vector $\boldsymbol{\beta}$ by minimizing the residual sum of squares

$$RSS = \sum_{t=1}^{n} (x_t - \boldsymbol{\beta}' \boldsymbol{z}_t)^2, \tag{2.3}$$

with respect to $\beta_1, \beta_2, \ldots, \beta_q$. Minimizing RSS yields the ordinary least squares estimator. This minimization can be accomplished by differentiating (2.3) with respect to the vector $\boldsymbol{\beta}$ or by using the properties of projections. In the notation above, this procedure gives the normal equations

$$\left(\sum_{t=1}^{n} \boldsymbol{z}_t \boldsymbol{z}_t' \right) \widehat{\boldsymbol{\beta}} = \sum_{t=1}^{n} \boldsymbol{z}_t x_t. \tag{2.4}$$

A further simplification of notation results from defining the matrix $Z = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n)'$ as the $n \times q$ matrix composed of the n samples of the input variables and the observed $n \times 1$ vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$. This identification yields

$$(Z'Z)\, \hat{\boldsymbol{\beta}} = Z' \boldsymbol{x} \tag{2.5}$$

and the solution

$$\widehat{\boldsymbol{\beta}} = (Z'Z)^{-1} Z' \boldsymbol{x} \tag{2.6}$$

when the matrix $Z'Z$ is of rank q. The minimized residual sum of squares (2.3) has the equivalent matrix forms

$$\begin{aligned} RSS &= (\boldsymbol{x} - Z\widehat{\boldsymbol{\beta}})'(\boldsymbol{x} - Z\widehat{\boldsymbol{\beta}}) \\ &= \boldsymbol{x}'\boldsymbol{x} - \widehat{\boldsymbol{\beta}}' Z' \boldsymbol{x} \\ &= \boldsymbol{x}'\boldsymbol{x} - \boldsymbol{x}' Z (Z'Z)^{-1} Z' \boldsymbol{x}, \end{aligned} \tag{2.7}$$

to give some useful versions for later reference. The ordinary least squares estimators are unbiased, i.e., $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and have the smallest variance within the class of linear unbiased estimators.

If the errors w_t are normally distributed (Gaussian), $\widehat{\boldsymbol{\beta}}$ is also the maximum likelihood estimator for $\boldsymbol{\beta}$ and is normally distributed with

$$\begin{aligned} \mathrm{cov}(\widehat{\boldsymbol{\beta}}) &= \sigma_w^2 \left(\sum_{t=1}^{n} \boldsymbol{z}_t \boldsymbol{z}_t' \right)^{-1} \\ &= \sigma_w^2 (Z'Z)^{-1} \\ &= \sigma_w^2 C, \end{aligned} \tag{2.8}$$

where

$$C = (Z'Z)^{-1} \tag{2.9}$$

is a convenient notation for later equations. An unbiased estimator for the variance σ_w^2 is

$$s_w^2 = \frac{RSS}{n - q}, \tag{2.10}$$

Table 2.1 Analysis of Variance for Regression

Source	df	Sum of Squares	Mean Square
$z_{t,q_1+1}, \ldots, z_{t,q}$	$q - q_1$	$SS_{reg} = RSS_1 - RSS$	$MS_{reg} = SS_{reg}/(q - q_1)$
Error	$n - q$	RSS	$s_w^2 = RSS/(n - q)$
Total	$n - q_1$	RSS_1	

contrasted with the maximum likelihood estimator $\widehat{\sigma}_w^2 = RSS/n$, which has the divisor n. Under the normal assumption, s_w^2 is distributed proportionally to a chi-squared random variable with $n - q$ degrees of freedom, denoted by χ_{n-q}^2, and independently of $\widehat{\beta}$. It follows that

$$t_{n-q} = \frac{(\widehat{\beta}_i - \beta_i)}{s_w \sqrt{c_{ii}}} \tag{2.11}$$

has the t-distribution with $n-q$ degrees of freedom; c_{ii} denotes the i^{th} diagonal element of C, as defined in (2.9).

Various competing models are of interest to isolate or select the best subset of independent variables. Suppose a proposed model specifies that only a subset $q_1 < q$ independent variables, say, $\boldsymbol{z}_{1t} = (z_{t1}, z_{t2}, \ldots, z_{tq_1})'$ is influencing the dependent variable x_t, so the model

$$x_t = \boldsymbol{\beta}_1' \boldsymbol{z}_{1t} + w_t \tag{2.12}$$

becomes the null hypothesis, where $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_{q_1})'$ is a subset of coefficients of the original q variables. We can test the reduced model (2.12) against the full model (2.2) by comparing the residual sums of squares under the two models using the F-statistic

$$F_{q-q_1,n-q} = \frac{RSS_1 - RSS}{RSS} \frac{n-q}{q-q_1}, \tag{2.13}$$

which has the central F-distribution with $q - q_1$ and $n - q$ degrees of freedom when (2.12) is the correct model. The statistic, which follows from applying the likelihood ratio criterion, has the improvement per number of parameters added in the numerator compared with the error sum of squares under the full model in the denominator. The information involved in the test procedure is often summarized in an Analysis of Variance (ANOVA) table as given in Table 2.1 for this particular case. The difference in the numerator is often called the regression sum of squares

In terms of Table 2.1, it is conventional to write the F-statistic (2.13) as the ratio of the two mean squares, obtaining

$$F_{q-q_1,n-q} = \frac{MS_{reg}}{s_w^2}. \tag{2.14}$$

A special case of interest is $q_1 = 1$ and $z_{1t} = 1$, so the model in (2.12) becomes

$$x_t = \beta_1 + w_t,$$

and we may measure the proportion of variation accounted for by the other variables using

$$R^2_{xz} = \frac{RSS_0 - RSS}{RSS_0}, \tag{2.15}$$

where the residual sum of squares under the reduced model

$$RSS_0 = \sum_{t=1}^{n} (x_t - \bar{x})^2, \tag{2.16}$$

in this case is just the sum of squared deviations from the mean $\bar{x}$. The measure R^2_{xz} is also the squared multiple correlation between x_t and the variables $z_{t2}, z_{t3}, \ldots, z_{tq}$.

The techniques discussed in the previous paragraph can be used to test various models against one another using the F test given in (2.13), (2.14), and the ANOVA table. These tests have been used in the past in a stepwise manner, where variables are added or deleted when the values from the F-test either exceed or fail to exceed some predetermined levels. The procedure, called stepwise multiple regression, is useful in arriving at a set of useful variables. An alternative is to focus on a procedure for model selection that does not proceed sequentially, but simply evaluates each model on its own merits. Suppose we consider a regression model with k coefficients and denote the maximum likelihood estimator for the variance as

$$\widehat{\sigma}^2_k = \frac{RSS_k}{n}, \tag{2.17}$$

where RSS_k denotes the residual sum of squares under the model with k regression coefficients. Then, Akaike (1969, 1973, 1974) suggested measuring the goodness of fit for this particular model by balancing the error of the fit against the number of parameters in the model; we define

Definition 2.1 Akaike's Information Criterion (AIC)

$$\text{AIC} = \ln\ \widehat{\sigma}^2_k + \frac{n + 2k}{n}, \tag{2.18}$$

where $\widehat{\sigma}^2_k$ is given by (2.17) and k is the number of parameters in the model.

The value of k yielding the minimum AIC specifies the best model. The idea is roughly that minimizing $\widehat{\sigma}^2_k$ would be a reasonable objective, except that it decreases monotonically as k increases. Therefore, we ought penalize the error variance by a term proportional to the number of parameters. The choice for the penalty term given by (2.18) is not the only one, and a considerable

literature is available advocating different penalty terms. A corrected form, suggested by Sugiura (1978), and expanded by Hurvich and Tsai (1989), can be based on small-sample distributional results for the linear regression model (details are provided in Problems 2.4 and 2.5). The corrected form is defined as

Definition 2.2 AIC, Bias Corrected (AICc)

$$\text{AICc} = \ln\ \widehat{\sigma}_k^2 + \frac{n+k}{n-k-2}, \tag{2.19}$$

where $\widehat{\sigma}_k^2$ is given by (2.17), k is the number of parameters in the model, and n is the sample size.

We may also derive a correction term based on Bayesian arguments, as in Schwarz (1978), which leads to

Definition 2.3 Schwarz's Information Criterion (SIC)

$$\text{SIC} = \ln\ \widehat{\sigma}_k^2 + \frac{k \ln n}{n}, \tag{2.20}$$

using the same notation as in Definition 2.2.

SIC is also called the Bayesian Information Criterion (BIC) (see also Rissanen, 1978, for an approach yielding the same statistic based on a minimum description length argument). Various simulation studies have tended to verify that SIC does well at getting the correct order in large samples, whereas AICc tends to be superior in smaller samples where the relative number of parameters is large (see McQuarrie and Tsai, 1998, for detailed comparisons). In fitting regression models, two measures that have been used in the past are adjusted R-squared, which is essentially s_w^2, and Mallows C_p, Mallows (1973), which we do not consider in this context.

Example 2.2 Pollution, Temperature and Mortality

The data shown in Figure 2.2 are extracted series from a study by Shumway et al. (1988) of the possible effects of temperature and pollution on daily mortality in Los Angeles County. Note the strong seasonal components in all of the series, corresponding to winter-summer variations and the downward trend in the cardiovascular mortality over the 10-year period.

A scatterplot matrix, shown in Figure 2.3, indicates a possible linear relation between mortality and the pollutant particulates and a possible relation to temperature. Note the curvilinear shape of the temperature mortality curve, indicating that higher temperatures as well as lower temperatures are associated with increases in cardiovascular mortality.

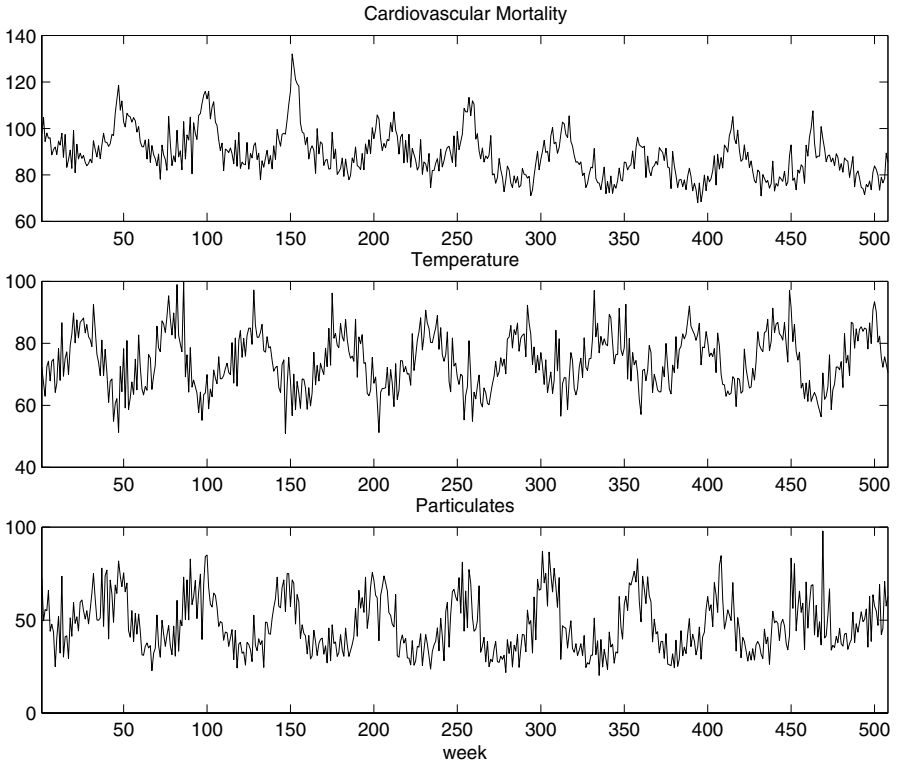


Figure 2.2 Average daily cardiovascular mortality (top), temperature (middle) and particulate pollution (bottom) in Los Angeles County. There are 508 six-day smoothed averages obtained by filtering daily values over the 10 year period 1970-1979.

Based on the scatterplot matrix, we entertain, tentatively, four models where M_t denotes cardiovascular mortality, T_t denotes temperature and P_t denotes the particulate levels. They are

$$M_t = \beta_0 + \beta_1 t + w_t \tag{2.21}$$
$$M_t = \beta_0 + \beta_1 t + \beta_2(T_t - T_\cdot) + w_t \tag{2.22}$$
$$M_t = \beta_0 + \beta_1 t + \beta_2(T_t - T_\cdot) + \beta_3(T_t - T_\cdot)^2 + w_t \tag{2.23}$$
$$M_t = \beta_0 + \beta_1 t + \beta_2(T_t - T_\cdot) + \beta_3(T_t - T_\cdot)^2 + \beta_4 P_t + w_t \tag{2.24}$$

where we adjust temperature for its mean, $T_\cdot = 74.6$, to avoid scaling problems. It is clear that (2.21) is a trend only model, (2.22) is linear temperature, (2.23) is curvilinear temperature and (2.24) is curvilinear temperature and pollution. We summarize some the statistics given for this particular case in Table 2.2. The values of R^2 were computed by

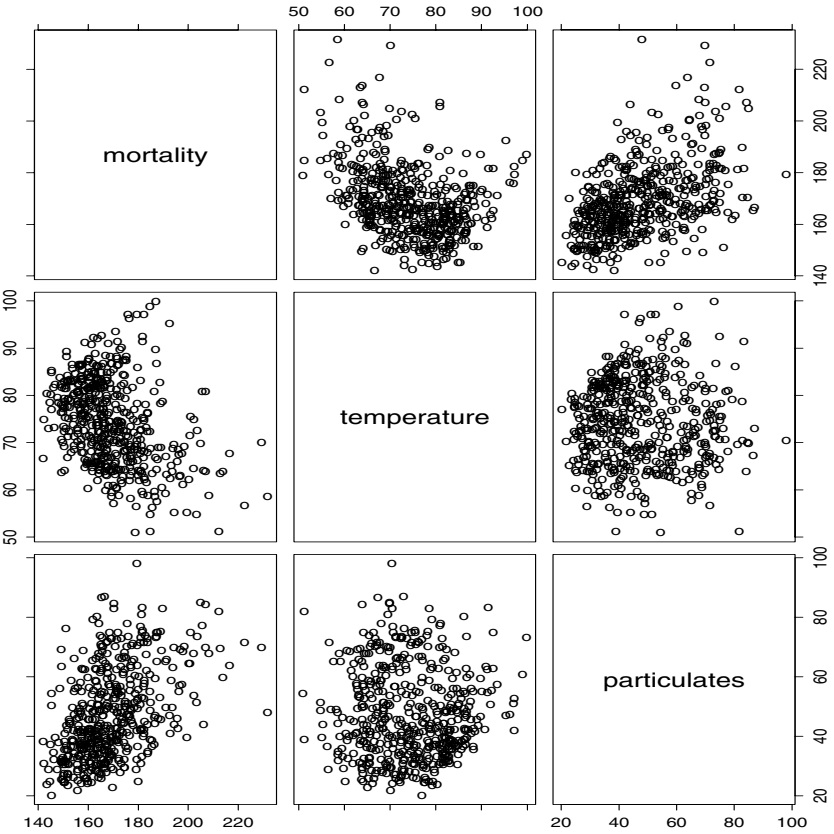


Figure 2.3 Scatterplot matrix showing plausible relations between mortality, temperature, and pollution.

Table 2.2 Summary Statistics for Mortality Models

Model	RSS (2.3)	s_w^2 (2.10)	R^2 (2.15)	AICc (2.19)
(2.21)	40,020	79.09	.21	5.38
(2.22)	31,413	62.20	.38	5.14
(2.23)	27,985	55.52	.45	5.03
(2.24)	20,509	40.77	.60	4.72

noting that $RSS_0 = 50,687$ using (2.16).

We note that each model does substantially better than the one before it and that the model including both temperature, temperature squared and particulates does the best, accounting for some 60% of the variability

and with the best value for AICc. Note that one can compare any two models using the residual sums of squares and (2.13). Hence, a model with only trend could be compared to the full model using $q = 5, q_1 = 2, n = 508$, so

$$F_{3,503} = \frac{(40,020 - 20,509)}{20,509} \frac{503}{3} = 160,$$

which exceeds $F_{3,\infty}(.001) = 5.42$. We obtain the best prediction model,

$$\begin{aligned} \widehat{M}_t &= 81.59 - .027_{(.002)}t - .473_{(.032)}(T_t - 74.6) \\ &\quad + .023_{(.003)}(T_t - 74.6)^2 + .255_{(.019)}P_t, \end{aligned}$$

for mortality, where the standard errors, computed from (2.8)-(2.10), are given in parentheses. As expected, a negative trend is present in time as well as a negative coefficient for adjusted temperature. The quadratic effect of temperature can clearly be seen in the scatterplots of Figure 2.3. Pollution weights positively and can be interpreted as the incremental contribution to daily deaths per unit of particulate pollution. It would still be essential to check the residuals $\widehat{w}_t = M_t - \widehat{M}_t$ for autocorrelation, but we defer this question to the section on correlated least squares, in which the incorporation of time correlation changes the estimated standard errors.

To display the scatterplot matrix, perform the final regression and compute AIC in R, use the following commands:

```
> mort = scan("/mydata/cmort.dat")
> temp = scan("/mydata/temp.dat")
> part = scan("/mydata/part.dat")
> temp = temp - mean(temp)
> temp2 = temp^2
> t = 1:length(mort)
> fit = lm(mort~t + temp + temp2 + part)
> summary(fit)   # Results
> AIC(fit)/508   # R gives n*AIC
> pairs(cbind(mort, temp, part)) # scatterplot matrix
```

2.3 Exploratory Data Analysis

In general, it is necessary for time series data to be stationary, so averaging lagged products over time, as in the previous section, will be a sensible thing to do. With time series data, it is the dependence between the values of the series that is important to measure; we must, at least, be able to estimate autocorrelations with precision. It would be difficult to measure that dependence if the dependence structure is not regular or is changing at every time point.

Hence, to achieve any meaningful statistical analysis of time series data, it will be crucial that, if nothing else, the mean and the autocovariance functions satisfy the conditions of stationarity (for at least some reasonable stretch of time) stated in Definition 1.7. Often, this is not the case, and we will mention some methods in this section for playing down the effects of nonstationarity so the stationary properties of the series may be studied.

A number of our examples came from clearly nonstationary series. The Johnson & Johnson series in Figure 1.1 has a mean that increases exponentially over time, and the increase in the magnitude of the fluctuations around this trend causes changes in the covariance function; the variance of the process, for example, clearly increases as one progresses over the length of the series. Also, the global temperature series shown in Figure 1.2 contains some evidence of a trend over time; human-induced global warming advocates seize on this as empirical evidence to advance their hypothesis that temperatures are increasing.

Perhaps the easiest form of nonstationarity to work with is the trend stationary model wherein the process has stationary behavior around a trend. We may write this type of model as

$$x_t = \mu_t + y_t \tag{2.25}$$

where x_t are the observations, μ_t denotes the trend, and y_t is a stationary process. Quite often, strong trend, μ_t, will obscure the behavior of the stationary process, y_t, as we shall see in numerous examples in Chapter 3. Hence, there is some advantage to removing the trend as a first step in an exploratory analysis of such time series. The steps involved are to obtain a reasonable estimate of the trend component, say $\widehat{\mu}_t$, and then work with the residuals

$$\widehat{y}_t = x_t - \widehat{\mu}_t. \tag{2.26}$$

Consider the following example.

Example 2.3 Detrending Global Temperature

Here we suppose the model is of the form of (2.25),

$$x_t = \mu_t + y_t,$$

where, as we suggested in the analysis of the global temperature data presented in Example 2.1, a straight line might be a reasonable model for the trend, i.e.,

$$\mu_t = \beta_1 + \beta_2\, t.$$

In that example, we estimated the trend using ordinary least squares[1]

[1]Because the error term, y_t, is not assumed to be iid, the reader may feel that weighted least squares is called for in this case. The problem is, we do not know the behavior of y_t, and that is precisely what we are trying to assess at this stage. A notable result by Grenander and Rosenblatt (1957, Ch 7), however, is that under mild conditions on y_t, for polynomial regression or periodic regression, asymptotically, ordinary least squares is equivalent to weighted least squares.

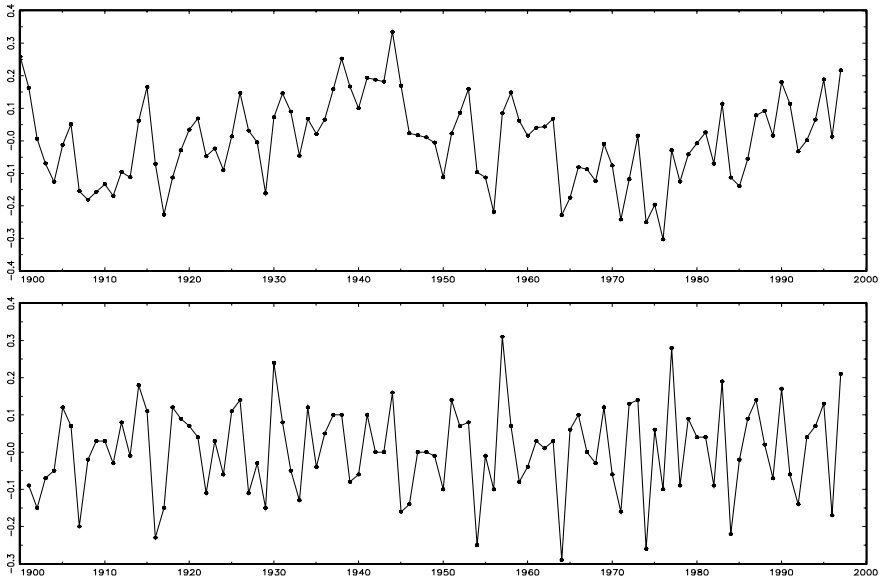


Figure 2.4 Detrended (top) and differenced (bottom) global temperature series. The original data are shown in Figures 1.2 and 2.1.

and found

$$\widehat{\mu}_t = -12.186 + .006\,t.$$

Figure 2.1 shows the data with the estimated trend line superimposed. To obtain the detrended series we simply subtract $\widehat{\mu}_t$ from the observations, x_t, to obtain the detrended series

$$\widehat{y}_t = x_t + 12.186 - .006\,t.$$

The top graph of Figure 2.4 shows the detrended series. Figure 2.5 shows the ACF of the original data (top panel) as well as the ACF of the detrended data (middle panel).

To detrend in R, assuming the data are in `gtemp`:

```
> x = gtemp[45:142]  # use only 1900 to 1997
> t = 1900:1997
> fit = lm(x~t)      # detrended series in fit$resid
> plot(t, fit$resid, type="o", ylab="detrended gtemp")
```

In Example 1.11 and the corresponding Figure 1.10 we saw that a random walk might also be a good model for trend. That is, rather than modeling trend as fixed (as in Example 2.3), we might model trend as a stochastic component using the random walk with drift model,

$$\mu_t = \delta + \mu_{t-1} + w_t, \tag{2.27}$$

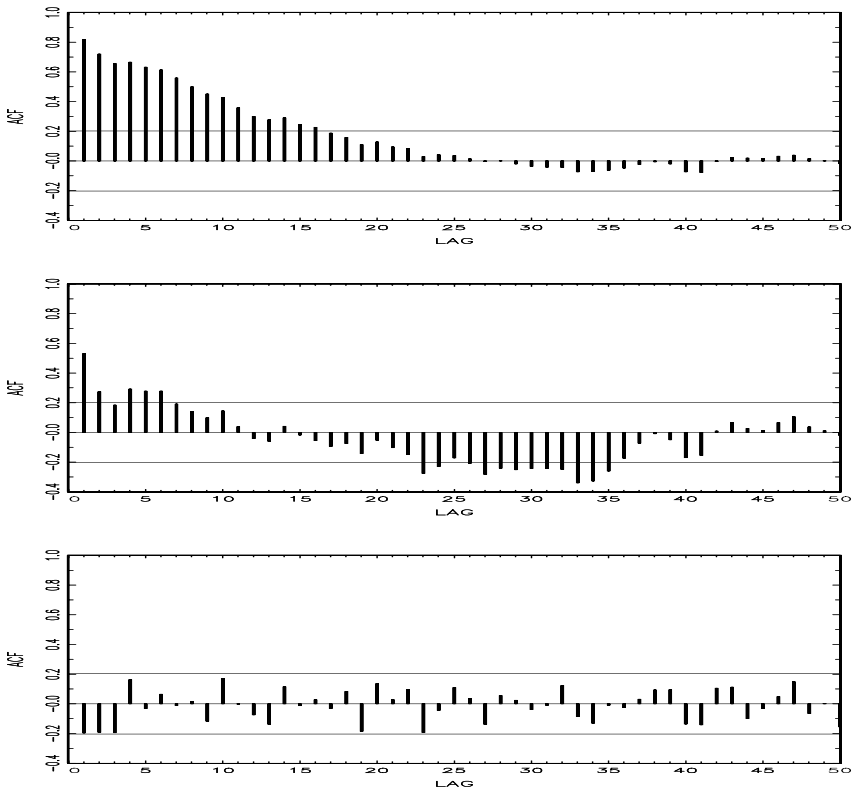


Figure 2.5 Sample ACFs of the global temperature (top), and of the detrended (middle) and the differenced (bottom) series.

where w_t is white noise and is independent of y_t. If the appropriate model is (2.25), then differencing the data, x_t, yields a stationary process; that is,

$$\begin{aligned} x_t - x_{t-1} &= (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) \\ &= \delta + w_t + y_t - y_{t-1}. \end{aligned} \tag{2.28}$$

We leave it as an exercise (Problem 2.7) to show (2.28) is stationary.[2]

One advantage of differencing over detrending to remove trend is that no parameters are estimated in the differencing operation. One disadvantage, however, is that differencing does not yield an estimate of the stationary process y_t as can be seen in (2.28). If an estimate of y_t is essential, then detrending may be more appropriate. If the goal is to coerce the data to stationarity, then

[2]The key to establishing the stationarity of these types of processes is to recall that if $U = \sum_{j=1}^{m} a_j X_j$ and $V = \sum_{k=1}^{r} b_k Y_k$ are linear combinations of random variables $\{X_j\}$ and $\{Y_k\}$, respectively, then $\mathrm{cov}(U, V) = \sum_{j=1}^{m} \sum_{k=1}^{r} a_j b_k \mathrm{cov}(X_j, Y_k)$.

differencing may be more appropriate. Differencing is also a viable tool if the trend is fixed, as in Example 2.3. That is, e.g., if $\mu_t = \beta_1 + \beta_2 t$ in the model (2.25), differencing the data produces stationarity (see Problem 2.6):

$$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \beta_2 + y_t - y_{t-1}.$$

Because differencing plays a central role in time series analysis, it receives its own notation. The first difference is denoted as

$$\nabla x_t = x_t - x_{t-1}. \tag{2.29}$$

As we have seen, the first difference eliminates a linear trend. A second difference, that is, the difference of (2.29), can eliminate a quadratic trend, and so on. In order to define higher differences, we need a variation in notation that we use, for the first time here, and often in our discussion of ARIMA models in Chapter 3.

Definition 2.4 *We define the* **backshift operator** *by*

$$Bx_t = x_{t-1}$$

and extend it to powers $B^2 x_t = B(Bx_t) = Bx_{t-1} = x_{t-2}$*, and so on. Thus,*

$$B^k x_t = x_{t-k}. \tag{2.30}$$

It is clear that we may then rewrite (2.29) as

$$\nabla x_t = (1 - B)x_t, \tag{2.31}$$

and we may extend the notion further. For example, the second difference becomes

$$\begin{aligned} \nabla^2 x_t &= (1 - B)^2 x_t = (1 - 2B + B^2)x_t \\ &= x_t - 2x_{t-1} + x_{t-2} \end{aligned}$$

by the linearity of the operator. To check, just take the difference of the first difference $\nabla(\nabla x_t) = \nabla(x_t - x_{t-1}) = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2})$.

Definition 2.5 Differences of order d *are defined as*

$$\nabla^d = (1 - B)^d, \tag{2.32}$$

where we may expand the operator $(1-B)^d$ *algebraically to evaluate for higher integer values of* d*. When* $d = 1$*, we drop it from the notation.*

The first difference (2.29) is an example of a linear filter applied to eliminate a trend. Other filters, formed by averaging values near x_t, can produce adjusted series that eliminate other kinds of unwanted fluctuations, as in Chapter 3. The differencing technique is an important component of the ARIMA model of Box and Jenkins (1970) (see also Box et al., 1994), to be discussed in Chapter 3.

Example 2.4 Differencing Global Temperature

The first difference of the global temperature series, also shown in Figure 2.4, does not contain the long middle cycle we observe in the detrended series. The ACF of this series is also shown in Figure 2.5. In this case it appears that the differenced process may be white noise, which implies that the global temperature series is a random walk. Finally, notice that removing trend by detrending (i.e., regression techniques) produces different results than removing trend by differencing.

Continuing from Example 2.3, to difference and plot the data in R:

```
> x = gtemp[44:142]  # start at 1899
> plot(1900:1997, diff(x), type="o", xlab="year")
```

An alternative to differencing is a less-severe operation that still assumes stationarity of the underlying time series. This alternative, called fractional differencing, extends the notion of the difference operator (2.32) to fractional powers $-.5 < d < .5$, which still define stationary processes. Granger and Joyeux (1980) and Hosking (1981) introduced long memory time series, which corresponds to the case when $0 < d < .5$. This model is often used for environmental time series arising in hydrology. We will discuss long memory processes in more detail in §5.2.

Often, obvious aberrations are present that can contribute nonstationary as well as nonlinear behavior in observed time series. In such cases, transformations may be useful to equalize the variability over the length of a single series. A particularly useful transformation is

$$y_t = \ln x_t, \tag{2.33}$$

which tends to suppress larger fluctuations that occur over portions of the series where the underlying values are larger. Other possibilities are power transformations in the Box–Cox family of the form

$$y_t = \begin{cases} (x_t^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \ln x_t, & \lambda = 0. \end{cases} \tag{2.34}$$

Methods for choosing the power λ are available (see Johnson and Wichern, 1992, §4.7) but we do not pursue them here. Often, transformations are also used to improve the approximation to normality or to improve linearity in predicting the value of one series from another.

Example 2.5 Paleoclimatic Glacial Varves

Melting glaciers deposit yearly layers of sand and silt during the spring melting seasons, which can be reconstructed yearly over a period ranging from the time deglaciation began in New England (about 12,600 years

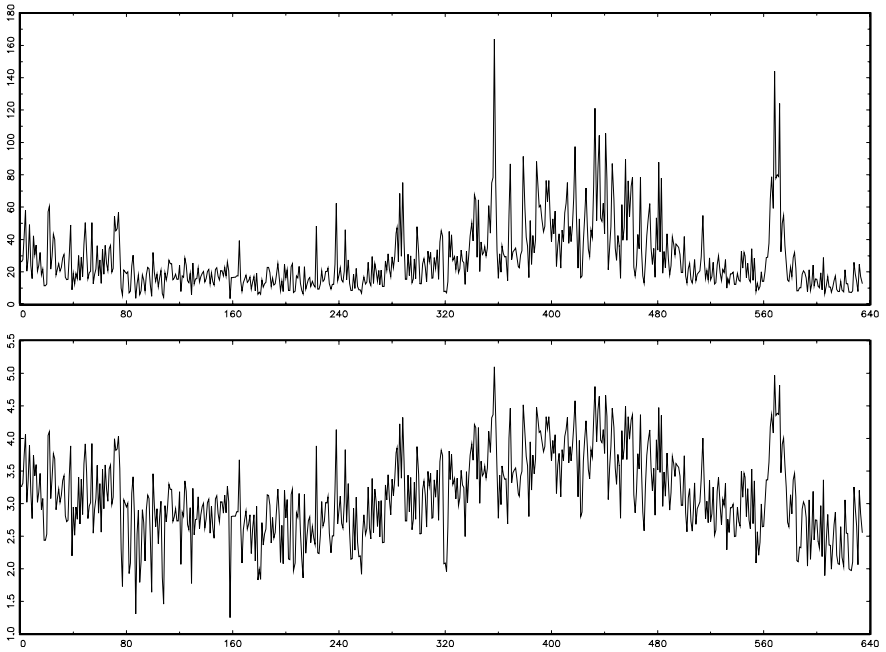


Figure 2.6 Glacial varve thicknesses (top) from Massachusetts for $n = 634$ years compared with log transformed thicknesses (bottom).

ago) to the time it ended (about 6,000 years ago). Such sedimentary deposits, called varves, can be used as proxies for paleoclimatic parameters, such as temperature, because, in a warm year, more sand and silt are deposited from the receding glacier. Figure 2.6 shows the thicknesses of the yearly varves collected from one location in Massachusetts for 634 years, beginning 11,834 years ago. For further information, see Shumway and Verosub (1992). Because the variation in thicknesses increases in proportion to the amount deposited, a logarithmic transformation could remove the nonstationarity observable in the variance as a function of time. Figure 2.6 shows the original and transformed varves, and it is clear that this improvement has occurred. We may also plot the histogram of the original and transformed data, as in Problem 2.8, to argue that the approximation to normality is improved. The ordinary first differences (2.31) are also computed in Problem 2.8, and we note that the first differences have a significant negative correlation at lag $h = 1$. Later, in Chapter 5, we will show that perhaps the varve series has long memory and will propose using fractional differencing.

Next, we consider another preliminary data processing technique that is

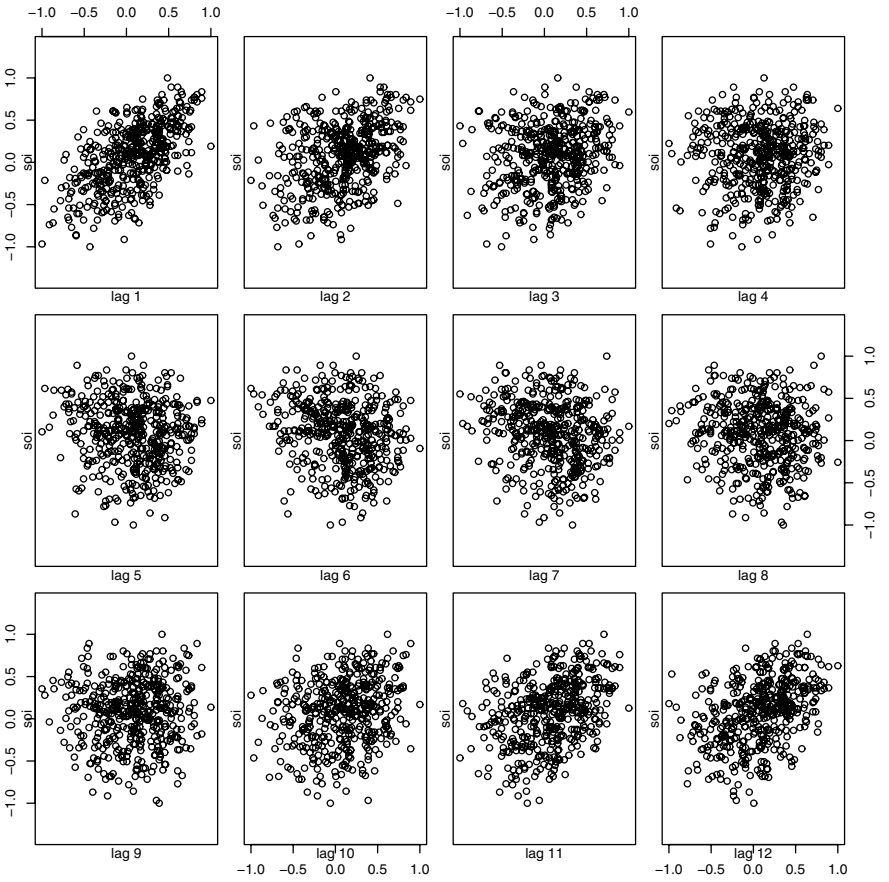


Figure 2.7 Scatterplot matrix relating current SOI values (x_t) to past SOI values (x_{t-h}) at lags $h = 1, 2, ..., 12$.

used for the purpose of visualizing the relations between series at different lags, namely, scatterplot matrices. In the definition of the ACF, we are essentially interested in relations between x_t and x_{t-h}; the autocorrelation function tells us whether a substantial linear relation exists between the series and its own lagged values. The ACF gives a profile of the linear correlation at all possible lags and shows which values of h lead to the best predictability. The restriction of this idea to linear predictability, however, may mask a possible nonlinear relation between current values, x_t, and past values, x_{t-h}. To check for nonlinear relations of this form, it is convenient to display a lagged scatterplot matrix, as in Figure 2.7, that displays values of x_t on the vertical axis plotted against x_{t-h} on the horizontal axis for the SOI x_t. Similarly, we might want to look at values of one series y_t plotted against another series at various

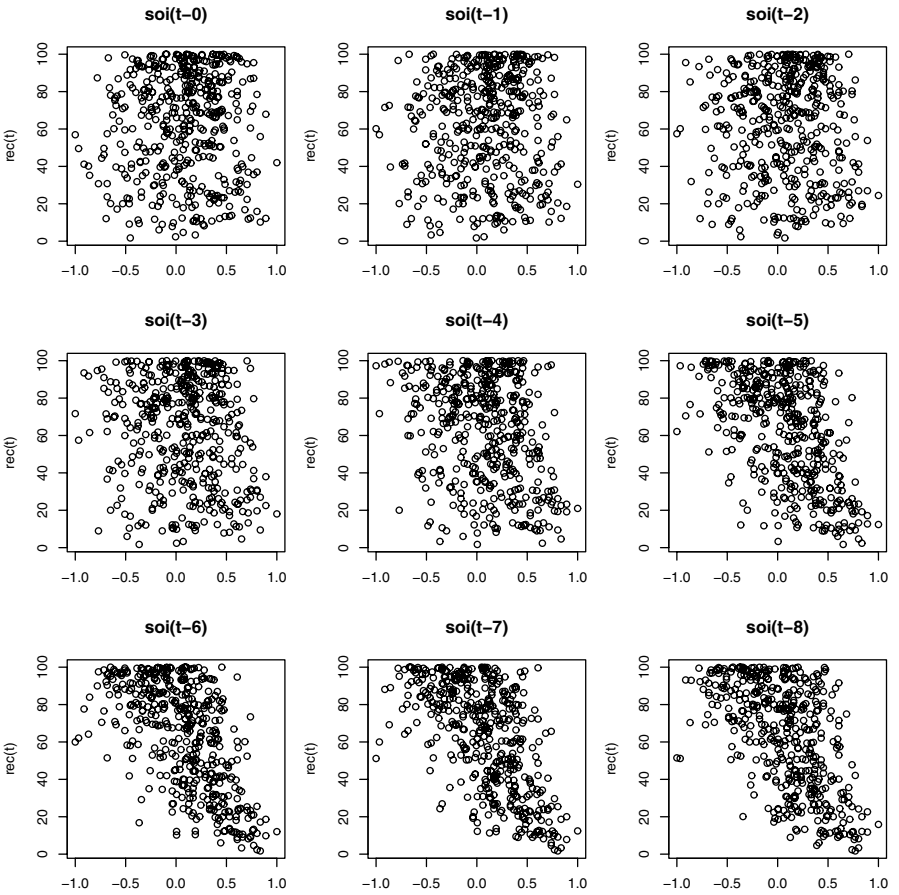


Figure 2.8 Scatterplot matrix of the Recruitment series, y_t, on the vertical axis plotted against the SOI series, x_{t-h}, on the horizontal axis at lags $h = 0, 1, \ldots, 8$.

lags, x_{t-h}, to look for possible nonlinear relations between the two series. Because, for example, we might wish to predict the Recruitment series, say, y_t, from current or past values of the SOI series, x_{t-h}, for $h = 0, 1, 2, ...$ it would be worthwhile to examine the scatterplot matrix. Figure 2.8 shows the lagged scatterplot of the Recruitment series y_t on the vertical axis plotted against the SOI index x_{t-h} on the horizontal axis.

Example 2.6 Scatterplot Matrices, SOI, and Recruitment Series

Consider the possibility of looking for nonlinear functional relations at lags in the SOI series, x_{t-h}, for $h = 0, 1, 2, ...$, and the Recruitment series, y_t. Noting first the top panel in Figure 2.7, we see strong posi-

tive and linear relations at lags $h = 1, 2, 11, 12$, that is, between x_t and $x_{t-1}, x_{t-2}, x_{t-11}, x_{t-12}$, and a negative linear relation at lags $h = 6, 7$. These results match up well with peaks noticed in the ACF in Figure 1.14. Figure 2.8 shows linearity in relating Recruitment, y_t, with the SOI series at $x_{t-5}, x_{t-6}, x_{t-7}, x_{t-8}$, indicating the SOI series tends to lead the Recruitment series and the coefficients are negative, implying that increases in the SOI lead to decreases in the Recruitment, and vice versa. Some possible nonlinear behavior shows as the relation tends to flatten out at both extremes, indicating a logistic type transformation may be useful.

To reproduce Figure 2.7 in R assuming the data are in `soi` and `rec` as before:

```
> lag.plot(soi, lags=12, layout=c(3,4), diag=F)
```

Reproducing Figure 2.8 in R is not as easy, but here is how the figure was generated:

```
> soi=ts(soi)    # make the series
> rec=ts(rec)    # time series objects
> par(mfrow=c(3,3), mar=c(2.5, 4, 4, 1)) # set up plot area
> for(h in 0:8){            # loop through lags 0-8
>  plot(lag(soi,-h),rec, main=paste("soi(t-",h,")",sep=""),
+  ylab="rec(t)",xlab="")
> }
```

As a final exploratory tool, we discuss assessing periodic behavior in time series data using regression analysis and the periodogram; this material may be thought of as an introduction to spectral analysis, which we discuss in detail in Chapter 4. In Example 1.12, we briefly discussed the problem of identifying cyclic or periodic signals in time series. A number of the time series we have seen so far exhibit periodic behavior. For example, the data from the pollution study example shown in Figure 2.2 exhibit strong yearly cycles. Also, the Johnson & Johnson data shown in Figure 1.1 make one cycle every year (four quarters) on top of an increasing trend and the speech data in Figure 1.2 is highly repetitive. The monthly SOI and Recruitment series in Figure 1.6 show strong yearly cycles, but hidden in the series are clues to the El Niño cycle.

Example 2.7 Using Regression to Discover a Signal in Noise

Recall, in Example 1.12 we generated $n = 500$ observations from the model

$$x_t = A\cos(2\pi\omega t + \phi) + w_t, \tag{2.35}$$

where $\omega = 1/50$, $A = 2$, $\phi = .6\pi$, and $\sigma_w = 5$; the data are shown on the bottom panel of Figure 1.11. At this point we assume the frequency of oscillation $\omega = 1/50$ is known, but A and ϕ are unknown parameters. In

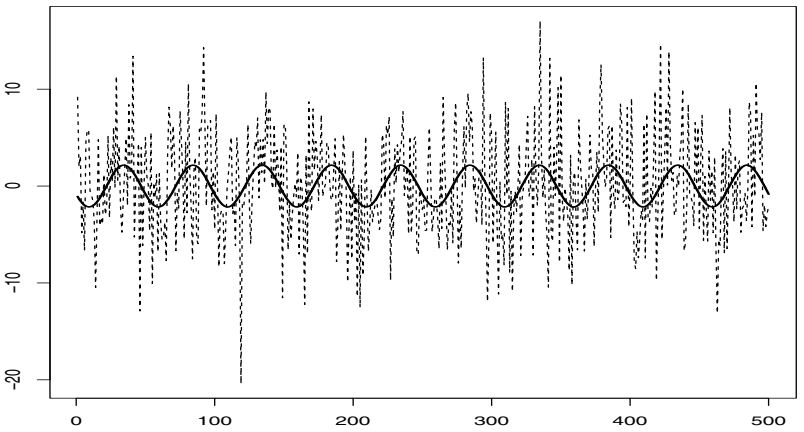


Figure 2.9 Data generated by (2.35) [dashed line] with the fitted [solid] line, (2.37), superimposed.

this case the parameters appear in (2.35) in a nonlinear way, so we use a trigonometric identity and write

$$\begin{aligned} A\cos(2\pi\omega t+\phi) &= A\cos(\phi)\cos(2\pi\omega t) - A\sin(\phi)\sin(2\pi\omega t) \\ &= \beta_1\cos(2\pi\omega t) + \beta_2\sin(2\pi\omega t), \end{aligned}$$

where $\beta_1 = A\cos(\phi)$ and $\beta_2 = -A\sin(\phi)$. Now the model (2.35) can be written in the usual linear regression form given by (no intercept term is needed here)

$$x_t = \beta_1\cos(2\pi t/50) + \beta_2\sin(2\pi t/50) + w_t. \tag{2.36}$$

Using linear regression on the generated data, the fitted model is

$$\widehat{x}_t = -.84_{(.32)}\cos(2\pi t/50) - 1.99_{(.32)}\sin(2\pi t/50) \tag{2.37}$$

with $\widehat{\sigma}_w = 5.08$, where the values in parentheses are the standard errors. We note the actual values of the coefficients for this example are $\beta_1 = 2\cos(.6\pi) = -.62$ and $\beta_2 = -2\sin(.6\pi) = -1.90$. Because the parameter estimates are significant and close to the actual values, it is clear that we are able to detect the signal in the noise using regression, even though the signal appears to be obscured by the noise in the bottom panel of Figure 1.11. Figure 2.9 shows data generated by (2.35) with the fitted line, (2.37), superimposed.

Example 2.8 Using the Periodogram to Discover a Signal in Noise

The analysis in Example 2.7 may seem like cheating because we assumed we knew the value of the frequency parameter ω. If we do not know ω,

we could try to fit the model (2.35) using nonlinear regression with ω as a parameter. Another method is to try various values of ω in a systematic way. Using the regression results of §2.2 (also, see Problem 4.10), we can show the estimated regression coefficients in Example 2.7 take on the special form[3] given by

$$\widehat{\beta}_1 = \frac{\sum_{t=1}^{n} x_t \cos(2\pi t/50)}{\sum_{t=1}^{n} \cos^2(2\pi t/50)} = \frac{2}{n}\sum_{t=1}^{n} x_t \cos(2\pi t/50); \tag{2.38}$$

$$\widehat{\beta}_2 = \frac{\sum_{t=1}^{n} x_t \sin(2\pi t/50)}{\sum_{t=1}^{n} \sin^2(2\pi t/50)} = \frac{2}{n}\sum_{t=1}^{n} x_t \sin(2\pi t/50). \tag{2.39}$$

This suggests looking at all possible regression parameter estimates, say

$$\widehat{\beta}_1(j/n) = \frac{2}{n}\sum_{t=1}^{n} x_t \cos(2\pi t\, j/n); \tag{2.40}$$

$$\widehat{\beta}_2(j/n) = \frac{2}{n}\sum_{t=1}^{n} x_t \sin(2\pi t\, j/n), \tag{2.41}$$

where, $n = 500$ and $j = 1, \ldots, \frac{n}{2} - 1$, and inspecting the results for large values. For the endpoints, $j = 0, n/2$, we have $\widehat{\beta}_1(0) = n^{-1}\sum_{t=1}^{n} x_t$, $\widehat{\beta}_1(\frac{n}{2}) = n^{-1}\sum_{t=1}^{n}(-1)^t x_t$ and $\widehat{\beta}_2(0) = \widehat{\beta}_2(\frac{n}{2}) = 0$.

For this particular example, the values calculated in (2.38) and (2.39) are $\widehat{\beta}_1(10/500)$ and $\widehat{\beta}_2(10/500)$. By doing this, we have regressed a series, x_t, of length n using n regression parameters, so that we will have a perfect fit. The point, however, is that if the data contain any cyclic behavior we are likely to catch it by performing these saturated regressions.

Next, note that the regression coefficients $\widehat{\beta}_1(j/n)$ and $\widehat{\beta}_2(j/n)$, for each j, are essentially measuring the correlation of the data with a sinusoid oscillating at j cycles in n time points.[4] Hence, an appropriate measure of the presence of a frequency of oscillation of j cycles in n time points in the data would be

$$P(j/n) = \widehat{\beta}_1^2(j/n) + \widehat{\beta}_2^2(j/n), \tag{2.42}$$

which is basically a measure of squared correlation. The quantity (2.42) is sometimes called the periodogram, but we will call $P(j/n)$ the scaled periodogram and we will investigate its properties in Chapter 4. Figure 2.10 shows the scaled periodogram for the data generated by (2.35), and it

[3] In the notation of §2.2, the estimates are $\sum_{t=1}^{n} x_t z_t \,/\, \sum_{t=1}^{n} z_t^2$. Here, $z_t = \cos(2\pi t/50)$ or $z_t = \sin(2\pi t/50)$.

[4] In the notation of §2.2, the regression coefficients (2.40) and (2.41) are of the form $\sum_t x_t z_t / \sum_t z_t^2$ whereas sample correlations are of the form $\sum_t x_t z_t \big/ \left(\sum_t x_t^2 \sum_t z_t^2\right)^{1/2}$.

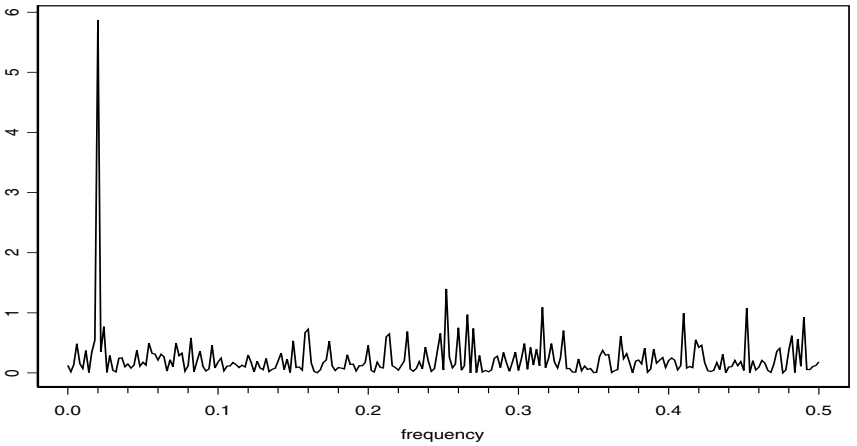


Figure 2.10 The scaled periodogram, (2.42), of the 500 observations generated by (2.35). The data are displayed in Figures 1.11 and 2.9.

easily discovers the periodic component with frequency $\omega = .02 = 10/500$ even though it is difficult to visually notice that component in Figure 1.11 due to the noise.

Finally, we mention that it is not necessary to run a large regression

$$x_t = \sum_{j=0}^{n/2} \beta_1(j/n)\cos(2\pi tj/n) + \beta_2(j/n)\sin(2\pi tj/n) \tag{2.43}$$

to obtain the values of $\beta_1(j/n)$ and $\beta_2(j/n)$ [with $\beta_2(0) = \beta_2(1/2) = 0$] because they can be computed quickly if n (assumed even here) is a highly composite integer. There is no error in (2.43) because there are n observations and n parameters; the regression fit will be perfect. The discrete Fourier transform (DFT) is a complex-valued weighted average of the data given by

$$d(j/n) = n^{-1/2}\sum_{t=1}^{n} x_t \exp(-2\pi itj/n), \tag{2.44}$$

and values j/n are called the Fourier or fundamental frequencies. Because of a large number of redundancies in the calculation, (2.44) may be computed quickly using the fast Fourier transform (FFT), which is available in many computing packages such as Matlab, S-PLUS and R. We note that[5]

$$|d(j/n)|^2 = \frac{1}{n}\left(\sum_{t=1}^{n} x_t\cos(2\pi tj/n)\right)^2 + \frac{1}{n}\left(\sum_{t=1}^{n} x_t\sin(2\pi tj/n)\right)^2 \tag{2.45}$$

[5] $e^{-i\alpha} = \cos(\alpha) - i\sin(\alpha)$ and if $z = a - ib$, then $|z|^2 = z\overline{z} = (a - ib)(a + ib) = a^2 + b^2$.

and it is this quantity that is called the periodogram; we will write

$$I(j/n) = |d(j/n)|^2.$$

So, we may calculate the scaled periodogram, (2.42), using the periodogram as

$$P(j/n) = \frac{4}{n} I(j/n). \tag{2.46}$$

We will discuss this approach in more detail and provide examples with data in Chapter 4.

A figure similar to Figure 2.10 can be created in R using the following commands[6]:

```
> t = 1:500
> x = 2*cos(2*pi*t/50 + .6*pi) + rnorm(500,0,5)
> I = abs(fft(x)/sqrt(500))^2   # the periodogram
> P = (4/500)*I                 # the scaled periodogram
> f = 0:250/500
> plot(f, P[1:251], type="l", xlab="frequency", ylab=" ")
> abline(v=seq(0,.5,.02), lty="dotted")
```

Example 2.9 The Periodogram as a Matchmaker

Another way of understanding the results of the previous example is to consider the problem of matching the data with sinusoids oscillating at various frequency. For example, Figure 2.11 shows $n = 100$ observations (as a solid line) generated by the model

$$x_t = \cos(2\pi t\,[2/100]) + w_t, \tag{2.47}$$

where w_t is Gaussian white noise with $\sigma_w = 1$. Superimposed on x_t are cosines oscillating at frequency 1/100, 2/100, and 3/100 (shown as dashed lines). Also included in the figure are correlations of x_t with the particular cosine, $\cos(2\pi tj/100)$, for $j = 1, 2, 3$. Note that the data match up well with the cosine oscillating at 2 cycles every 100 points (with a correlation of .57), whereas the data do not match up well with the other two cosines. For example, in the top panel of Figure 2.11, there is a decreasing trend in the data until observation 25, and then the data start an increasing trend to observation 50, whereas the cosine making one cycle (1/100) continues to decrease until observation 50.

[6]Different packages scale the FFT differently; consult the documentation. R calculates (2.44) without scaling by $n^{-1/2}$.

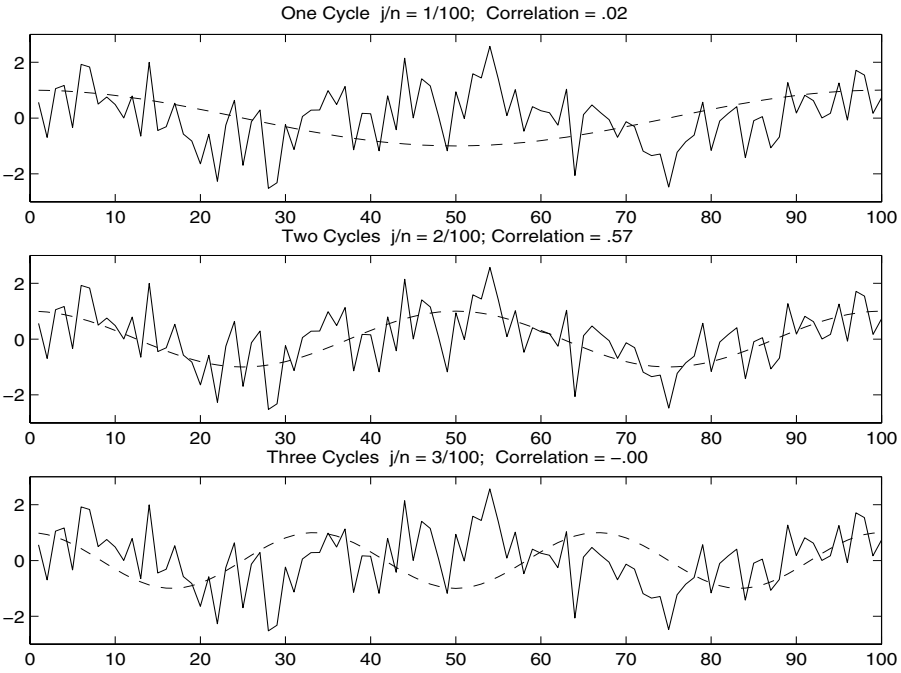


Figure 2.11 Data generated by (2.47) represented as a solid line with cosines oscillating at various frequencies superimposed (dashed lines). The correlation indicates the degree to which the two series line up.

2.4 Smoothing in the Time Series Context

In §1.4, we introduced the concept of smoothing a time series, and in Example 1.9, we discussed using a moving average to smooth white noise. This method is useful in discovering certain traits in a time series, such as long-term trend and seasonal components. In particular, if x_t represents the observations, then

$$m_t = \sum_{j=-k}^{k} a_j x_{t-j}, \tag{2.48}$$

where $a_j = a_{-j} \geq 0$ and $\sum_{j=-k}^{k} a_j = 1$ is a symmetric moving average of the data.

Example 2.10 Moving Average Smoother

For example, Figure 2.12 shows the weekly mortality series discussed in Example 2.2, a five-point moving average (which is essentially a monthly average with $k = 2$) that helps bring out the seasonal component and a 53-point moving average (which is essentially a yearly average with $k =$

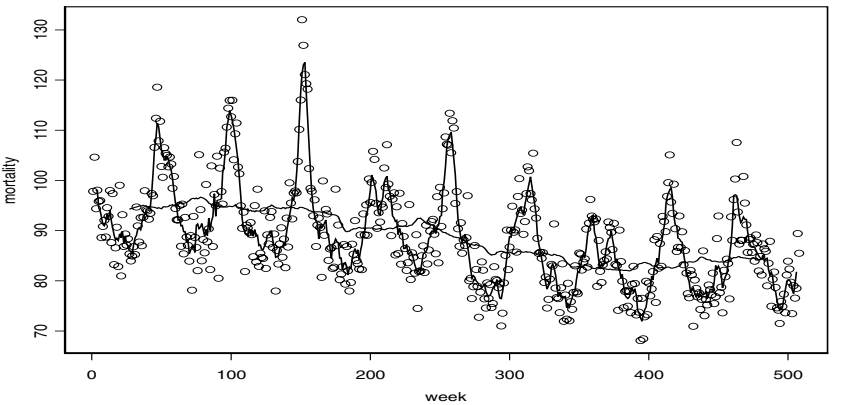


Figure 2.12 The weekly cardiovascular mortality series discussed in Example 2.2 smoothed using a five-week moving average and a 53-week moving average.

26) that helps bring out the (negative) trend in cardiovascular mortality. In both cases, the weights, $a_{-k}, \ldots, a_0, \ldots, a_k$, we used were all the same, and equal to $1/(2k+1)$.[7]

To reproduce Figure 2.12 in R assuming the mortality series is in `mort`:

```
> t = 1:length(mort)
> ma5 = filter(mort, sides=2, rep(1,5)/5)
> ma53 = filter(mort, sides=2, rep(1,53)/53)
> plot(t, mort, xlab="week", ylab="mortality")
> lines(ma5)
> lines(ma53)
```

Many other techniques are available for smoothing times series data based on methods from scatterplot smoothers. The general setup for a time plot is

$$x_t = f_t + y_t, \tag{2.49}$$

where f_t is some smooth function of time, and y_t is a stationary process. We may think of the moving average smoother m_t, given in (2.48), as an estimator of f_t. An obvious choice for f_t in (2.49) is polynomial regression

$$f_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p. \tag{2.50}$$

We have seen the results of a linear fit on the global temperature data in Example 2.1. For periodic data, one might employ periodic regression

$$\begin{aligned} f_t &= \alpha_0 + \alpha_1 \cos(2\pi\omega_1 t) + \beta_1 \sin(2\pi\omega_1 t) \\ &\quad + \cdots + \alpha_p \cos(2\pi\omega_p t) + \beta_p \sin(2\pi\omega_p t), \end{aligned} \tag{2.51}$$

[7]Sometimes, the end weights, a_{-k} and a_k are set equal to half the value of the other weights.

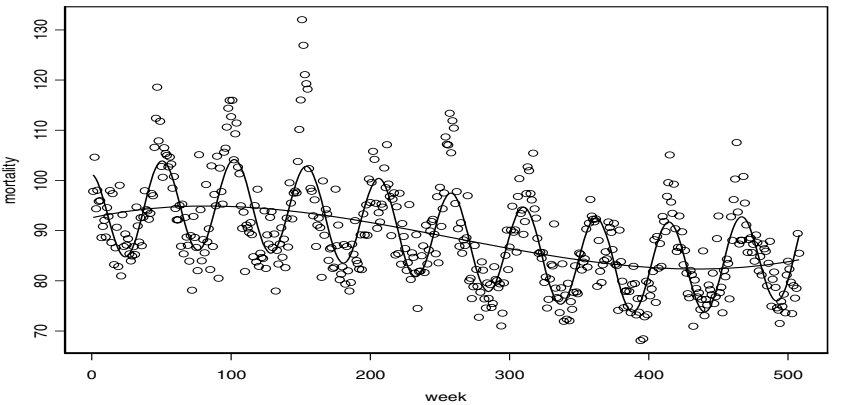


Figure 2.13 The weekly cardiovascular mortality series with a cubic trend and cubic trend plus periodic regression.

where $\omega_1, \ldots, \omega_p$ are distinct, specified frequencies. In addition, one might consider combining (2.50) and (2.51). These smoothers can be applied using classical linear regression.

Example 2.11 Polynomial and Periodic Regression Smoothers

Figure 2.13 shows the weekly mortality series with an estimated (via ordinary least squares) cubic smoother

$$\widehat{f}_t = \widehat{\beta}_0 + \widehat{\beta}_1 t + \widehat{\beta}_2 t^2 + \widehat{\beta}_3 t^3$$

superimposed to emphasize the trend, and an estimated (via ordinary least squares) cubic smoother plus a periodic regression

$$\widehat{f}_t = \widehat{\beta}_0 + \widehat{\beta}_1 t + \widehat{\beta}_2 t^2 + \widehat{\beta}_3 t^3 + \widehat{\alpha}_1 \cos(2\pi t/52) + \widehat{\alpha}_2 \sin(2\pi t/52)$$

superimposed to emphasize trend and seasonality.

The R commands for this example are:

```
> t = 1:length(mort)
> t2 = t^2
> t3 = t^3
> c = cos(2*pi*t/52)
> s = sin(2*pi*t/52)
> fit1 = lm(mort~t + t2 + t3)
> fit2 = lm(mort~t + t2 + t3 + c + s)
> plot(t, mort)
> lines(fit1$fit)
> lines(fit2$fit)
```

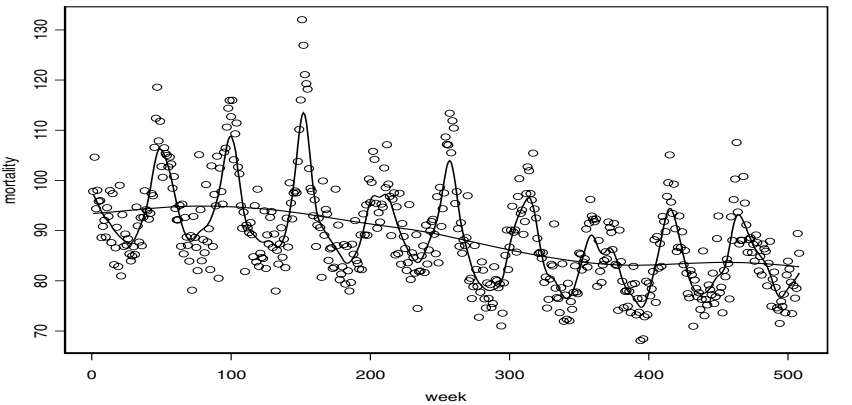


Figure 2.14 Kernel smoothers of the mortality data.

Modern regression techniques can be used to fit general smoothers to the pairs of points (t, x_t) where the estimate of f_t is smooth. Many of the techniques can easily be applied to time series data using the R or S-PLUS statistical packages; see Venables and Ripley (1994, Chapter 10) for details on applying these methods in S-PLUS (R is similar). A problem with the techniques used in Example 2.11 is that they assume f_t is the same function over the range of time, t; we might say that the technique is global. The moving average smoothers in Example 2.10 fit the data better because the technique is local; that is, moving average smoothers allow for the possibility that f_t is a different function over time. We describe some other local methods in the following examples.

Example 2.12 Kernel Smoothing

Kernel smoothing is a moving average smoother that uses a weight function, or kernel, to average the observations. Figure 2.14 shows kernel smoothing of the mortality series, where f_t in (2.49) is estimated by

$$\widehat{f}_t = \sum_{i=1}^{n} w_t(i) x_t, \tag{2.52}$$

where

$$w_t(i) = K\left(\frac{t-i}{b}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{t-j}{b}\right). \tag{2.53}$$

This estimator is called the Naradaya–Watson estimator (Watson, 1966). In (2.53), $K(\cdot)$ is a kernel function; typically, the normal kernel, $K(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$, is used. To implement this in R, use the `ksmooth` function. The wider the bandwidth, b, the smoother the result. In Figure 2.14, the values of b for this example were $b = 10$ (roughly weighted

monthly averages; that is, $b/2$ is the inner quartile range of the kernel) for the seasonal component, and $b = 104$ (roughly weighted yearly averages) for the trend component.

Figure 2.14 can be reproduced in R (or S-PLUS) as follows; we assume `t` and `mort` are available from the previous example:

```
> plot(t, mort)
> lines(ksmooth(t, mort, "normal", bandwidth=5))
> lines(ksmooth(t, mort, "normal", bandwidth=104))
```

Example 2.13 Nearest Neighbor and Locally Weighted Regression

Another approach to smoothing a time plot is nearest neighbor regression. The technique is based on k-nearest neighbors linear regression, wherein one uses the data $\{x_{t-k/2}, \ldots, x_t, \ldots, x_{t+k/2}\}$ to predict x_t using linear regression; the result is $\widehat{f}_t$. For example, Figure 2.15 shows cardiovascular mortality and the nearest neighbor method using the R (or S-PLUS) smoother `supsmu`. We used $k = n/2$ to estimate the trend and $k = n/100$ to estimate the seasonal component. In general, `supsmu` uses a variable window for smoothing (see Friedman, 1984), but it can be used for correlated data by fixing the smoothing window, as was done here.

Lowess is a method of smoothing that is rather complex, but the basic idea is close to nearest neighbor regression. Figure 2.15 shows smoothing of mortality using the R or S-PLUS function `lowess` (see Cleveland, 1979). First, a certain proportion of nearest neighbors to x_t are included in a weighting scheme; values closer to x_t in time get more weight. Then, a robust weighted regression is used to predict x_t and obtain the smoothed estimate of f_t. The larger the fraction of nearest neighbors included, the smoother the estimate $\widehat{f}_t$ will be. In Figure 2.15, the smoother uses about two-thirds of the data to obtain an estimate of the trend component, and the seasonal component uses 2% of the data.

Figure 2.15 can be reproduced in R or S-PLUS as follows (assuming `t` and `mort` are available from the previous example):

```
> par(mfrow=c(2,1))
> plot(t, mort, main="nearest neighbor")
> lines(supsmu(t, mort, span=.5))
> lines(supsmu(t, mort, span=.01))
> plot(t, mort, main="lowess")
> lines(lowess(t, mort, .02))
> lines(lowess(t, mort, 2/3))
```

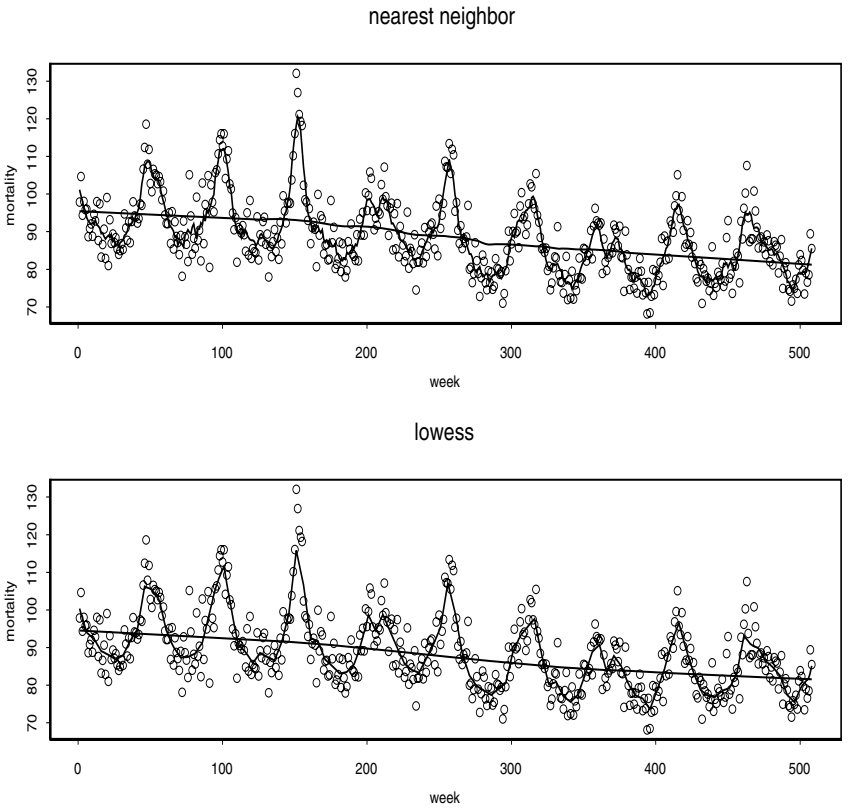


Figure 2.15 Nearest neightbor (`supsmu`) and locally weighted least squares (`lowess`) smoothers of the mortality data.

Example 2.14 Smoothing Splines

An extension of polynomial regression is to first divide time $t = 1, \ldots, n$, into k intervals, $[t_0 = 1, t_1]$, $[t_1 + 1, t_2]$, $\ldots$, $[t_{k-1} + 1, t_k = n]$. The values $t_0, t_1, \ldots, t_k$ are called *knots*. Then, in each interval, one fits a regression of the form (2.50); typically, $p = 3$, and this is called cubic splines.

A related method is smoothing splines, which minimizes a compromise between the fit and the degree of smoothness given by

$$\sum_{t=1}^{n} [x_t - f_t]^2 + \lambda \int \left(f_t^{''}\right)^2 dt, \tag{2.54}$$

where f_t is a cubic spline with a knot at each t. The degree of smoothness is controlled by $\lambda > 0$. Figure 2.16 shows smoothing splines on mortality using $\lambda = 10^{-7}$ for the seasonal component, and $\lambda = 0.1$ for the trend.

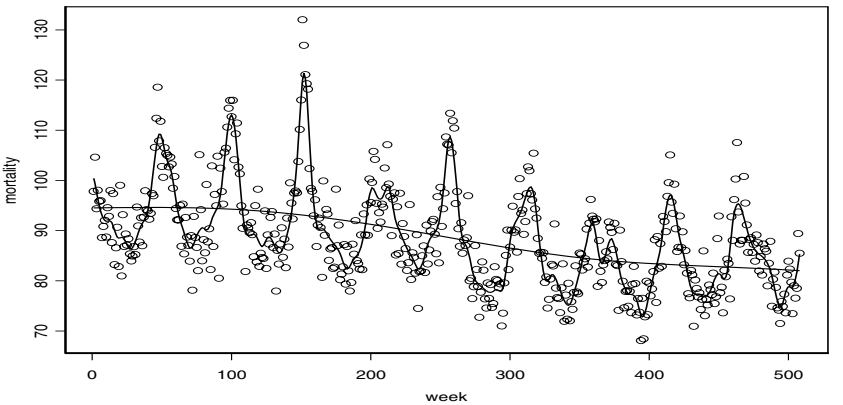


Figure 2.16 Smoothing splines fit to the mortality data.

Figure 2.16 can be reproduced in R or S-PLUS as follows (assuming `t` and `mort` are available from the previous example):

```
> plot(t, mort)
> lines(smooth.spline(t, mort, spar=.0000001))
> lines(smooth.spline(t, mort, spar=.1))
```

Example 2.15 Smoothing One Series as a Function of Another

In addition to smoothing time plots, smoothing techniques can be applied to smoothing a time series as a function of another time series. In this example, we smooth the scatterplot of two contemporaneously measured time series, mortality as a function of temperature. In Example 2.2, we discovered a nonlinear relationship between mortality and temperature. Continuing along these lines, Figure 2.17 shows scatterplots of mortality, M_t, and temperature, T_t, along with M_t is smoothed as a function of T_t using lowess and using smoothing splines. In both cases, mortality increases at extreme temperatures, but in an asymmetric way; mortality is higher at colder temperatures than at hotter temperatures. The minimum mortality rate seems to occur at approximately 80° F.

Figure 2.17 can be reproduced in R or S-PLUS as follows (assuming `mort` and `temp` contain the mortality and temperature data):

```
> par(mfrow=c(2,1))
> plot(temp, mort, main="lowess")
> lines(lowess(temp,mort))
> plot(temp, mort, main="smoothing splines")
> lines(smooth.spline(temp,mort))
```

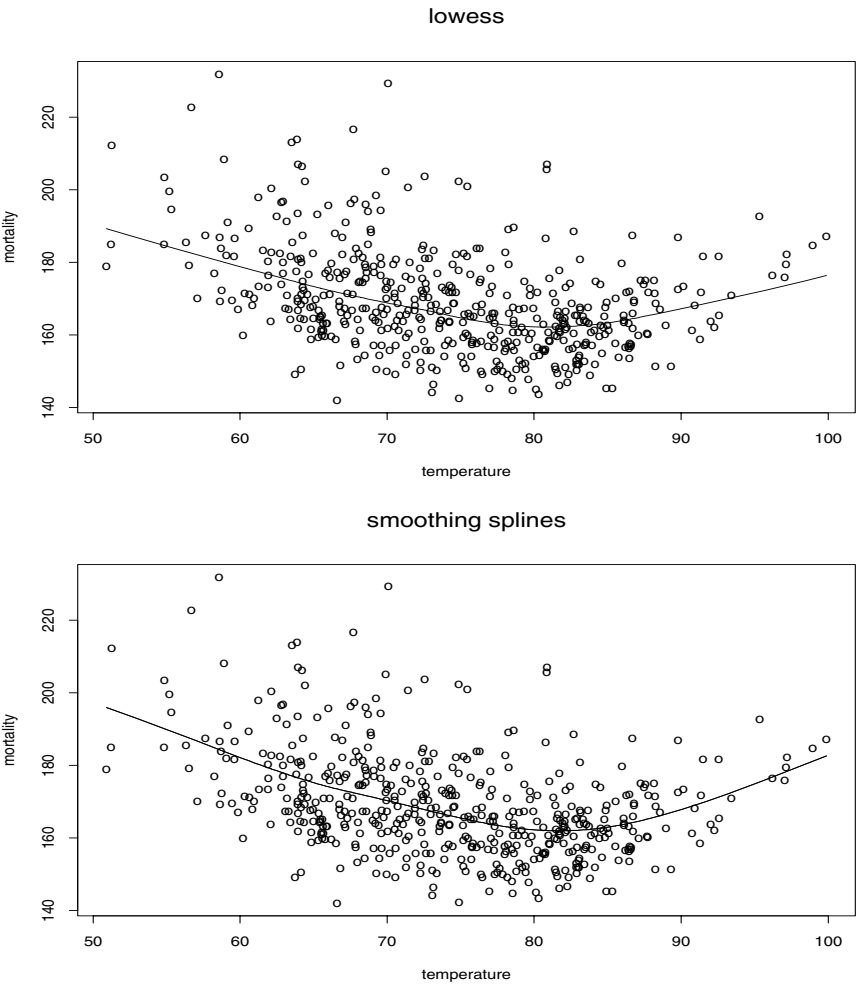


Figure 2.17 Smoothers of mortality as a function of temperature using lowess and smoothing splines.

As a final word of caution, the methods mentioned above do not particularly take into account the fact that the data are serially correlated, and most of the techniques mentioned have been designed for independent observations. That is, for example, the smoothers shown in Figure 2.17 are calculated under the false assumption that the pairs (M_t, T_t), for $t = 1, \ldots, 508$, are iid pairs of observations. In addition, the degree of smoothness used in the previous examples were chosen arbitrarily to bring out what might be considered obvious features in the data set.

Problems

Section 2.2

2.1 For the Johnson & Johnson data, say y_t, for $t = 1, \ldots, 84$, shown in Figure 1.1, let $x_t = \ln(y_t)$.

(a) Fit the regression model

$$x_t = \beta t + \alpha_1 Q_1(t) + \alpha_2 Q_2(t) + \alpha_3 Q_3(t) + \alpha_4 Q_4(t) + w_t$$

where $Q_i(t) = 1$ if time t corresponds to quarter $i = 1, 2, 3, 4$, and zero otherwise. The $Q_i(t)$'s are called indicator variables. We will assume for now that w_t is a Gaussian white noise sequence. What is the interpretation of the parameters β, α_1, α_2, α_3, and α_4? [Note: In R, to regress x on z without an intercept, use `lm(x~0+z)`; an easy way to generate $Q_1(t)$ is `Q1=rep(c(1,0,0,0),21)`.]

(b) What happens if you include an intercept term in the model in (a)?

(c) Graph the data, x_t, and superimpose the fitted values, say $\widehat{x}_t$, on the graph. Examine the residuals, $x_t - \widehat{x}_t$, and state your conclusions. Does it appear that the model fits the data well?

2.2 For the mortality data examined in Example 2.2:

(a) Add another component to the regression in (2.24) that accounts for the particulate count four weeks prior; that is, add P_{t-4} to the regression in (2.24). State your conclusion. [Note: In R, make sure the data are time series objects by using the `ts()` command, e.g., `mort=ts(mort)`. Center the temperature series and let `t = ts(1:length(mort))`. Then use `ts.intersect(mort, t, temp, temp^2, part, lag(part,-4))` to combine the series into a time series matrix object with six columns and regress the first column on the other columns.]

(b) Draw a scatterplot matrix of M_t, T_t, P_t and P_{t-4} and then calculate the pairwise correlations between the series. Compare the relationship between M_t and P_t versus M_t and P_{t-4}.

2.3 Generate a random walk with drift, (1.4), of length $n = 500$ with $\delta = .1$ and $\sigma_w = 1$. Call the data x_t for $t = 1, \ldots, 500$. Fit the regression $x_t = \beta t + w_t$ using least squares. Plot the data, the mean function (i.e., $\mu_t = .1\, t$) and the fitted line, $\widehat{x}_t = \widehat{\beta}\, t$, on the same graph. Discuss your results.

2.4 *Kullback-Leibler Information.* Given the random vector $\boldsymbol{y}$, we define the information for discriminating between two densities in the same family,

indexed by a parameter $\boldsymbol{\theta}$, say $f(\boldsymbol{y};\boldsymbol{\theta}_1)$ and $f(\boldsymbol{y};\boldsymbol{\theta}_2)$, as

$$I(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) = \frac{1}{n} E_1 \ln \frac{f(\boldsymbol{y};\boldsymbol{\theta}_1)}{f(\boldsymbol{y};\boldsymbol{\theta}_2)}, \tag{2.55}$$

where E_1 denotes expectation with respect to the density determined by $\boldsymbol{\theta}_1$. For the Gaussian regression model, the parameters are $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$. Show that we obtain

$$I(\boldsymbol{\theta}_1;\boldsymbol{\theta}_2) = \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - \ln\frac{\sigma_1^2}{\sigma_2^2} - 1\right) + \frac{1}{2}\frac{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)' Z' Z (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)}{n\sigma_2^2} \tag{2.56}$$

in that case.

2.5 *Model Selection.* Both selection criteria (2.18) and (2.19) are derived from information theoretic arguments, based on the well-known Kullback–Leibler discrimination information numbers (see Kullback and Leibler, 1951, Kullback, 1978). We give an argument due to Hurvich and Tsai (1989). We think of the measure (2.56) as measuring the discrepancy between the two densities, characterized by the parameter values $\boldsymbol{\theta}_1' = (\boldsymbol{\beta}_1', \sigma_1^2)'$ and $\boldsymbol{\theta}_2' = (\boldsymbol{\beta}_2', \sigma_2^2)'$. Now, if the true value of the parameter vector is $\boldsymbol{\theta}_1$, we argue that the best model would be one that minimizes the discrepancy between the theoretical value and the sample, say $I(\boldsymbol{\theta}_1;\widehat{\boldsymbol{\theta}})$. Because $\boldsymbol{\theta}_1$ will not be known, Hurvich and Tsai (1989) considered finding an unbiased estimator for $E_1[I(\boldsymbol{\beta}_1, \sigma_1^2; \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)]$, where

$$I(\boldsymbol{\beta}_1, \sigma_1^2; \hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \frac{1}{2}\left(\frac{\sigma_1^2}{\hat{\sigma}^2} - \ln\frac{\sigma_1^2}{\hat{\sigma}^2} - 1\right) + \frac{1}{2}\frac{(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}})' Z' Z (\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}})}{n\hat{\sigma}^2}$$

and $\boldsymbol{\beta}$ is a $k \times 1$ regression vector. Show that

$$E_1[I(\boldsymbol{\beta}_1, \sigma_1^2; \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)] = \frac{1}{2}\left(-\ln\sigma_1^2 + E_1 \ln\widehat{\sigma}^2 + \frac{n+k}{n-k-2} - 1\right), \tag{2.57}$$

using the distributional properties of the regression coefficients and error variance. An unbiased estimator for $E_1 \log\widehat{\sigma}^2$ is $\log\widehat{\sigma}^2$. Hence, we have shown that the expectation of the above discrimination information is as claimed. As models with differing dimensions k are considered, only the second and third terms in (2.57) will vary and we only need unbiased estimators for those two terms. This gives the form of AICc quoted in (2.19) in the chapter. You will need the two distributional results

$$\frac{n\widehat{\sigma}^2}{\sigma_1^2} \sim \chi^2_{n-k}$$

and

$$\frac{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_1)' Z' Z (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_1)}{\sigma_1^2} \sim \chi^2_k$$

The two quantities are distributed independently as chi-squared distributions with the indicated degrees of freedom. If $x \sim \chi^2_n$, $E(1/x) = 1/(n-2)$.

Section 2.3

2.6 Consider a process consisting of a linear trend with an additive noise term consisting of independent random variables w_t with zero means and variances σ^2_w, that is,

$$x_t = \beta_0 + \beta_1 t + w_t,$$

where β_0, β_1 are fixed constants.

(a) Prove x_t is nonstationary.

(b) Prove that the first difference series $\nabla x_t = x_t - x_{t-1}$ is stationary by finding its mean and autocovariance function.

(c) Repeat part (b) if w_t is replaced by a general stationary process, say y_t, with mean function μ_y and autocovariance function $\gamma_y(h)$.

2.7 Show (2.28) is stationary.

2.8 The glacial varve record plotted in Figure 2.6 exhibits some nonstationarity that can be improved by transforming to logarithms and some additional nonstationarity that can be corrected by differencing the logarithms.

(a) Verify that the untransformed glacial varves has intervals over which $\widehat{\gamma}(0)$ changes by computing the zero-lag autocovariance over two different intervals. Argue that the transformation $y_t = \ln x_t$ stabilizes the variance over the series. Plot the histograms of x_t and y_t to see whether the approximation to normality is improved by transforming the data.

(b) Examine the sample ACF, $\widehat{\rho}_y(h)$, of y_t and comment. Do any time intervals, of the order 100 years, exist where one can observe behavior comparable to that observed in the global temperature records in Figure 1.2?

(c) Compute the first difference $u_t = y_t - y_{t-1}$ of the log transformed varve records, and examine its time plot and autocorrelation function, $\widehat{\rho}_u(h)$, and argue that a first difference produces a reasonably stationary series. Can you think of a practical interpretation for u_t?

(d) Based on the sample ACF of the differenced transformed series computed in (c), argue that a generalization of the model given by Example 1.23 might be reasonable. Assume

$$u_t = \mu_u + w_t - \theta w_{t-1}$$

is stationary when the inputs w_t are assumed independent with mean 0 and variance σ_w^2. Show that

$$\gamma_u(h) = \begin{cases} \sigma_w^2(1+\theta^2) & \text{if } h = 0 \\ -\theta\ \sigma_w^2 & \text{if } h = \pm 1 \\ 0 & \text{if } |h| \geq 1. \end{cases}$$

Using the sample ACF and the printed autocovariance $\widehat{\gamma}_u(0)$, derive estimators for θ and σ^2. This is an application of the method of moments from classical statistics, where estimators of the parameters are derived by equating sample moments to theoretical moments.

2.9 Consider the two time series representing average wholesale U.S. gas and oil prices over 180 months, beginning in July 1973 and ending in December 1987. Analyze the data using some of the techniques in this chapter with the idea that we should be looking at how changes in oil prices influence changes in gas prices. For further reading, see Liu (1991). In particular,

(a) Plot the raw data, and look at the autocorrelation functions to argue that the untransformed data series are nonstationary.

(b) It is often argued in economics that price changes are important, in particular, the percentage change in prices from one month to the next. On this basis, argue that a transformation of the form $y_t = \ln x_t - \ln x_{t-1}$ might be applied to the data, where x_t is the oil or gas price series.

(c) Use lagged multiple scatterplots and the autocorrelation and cross-correlation functions of the transformed oil and gas price series to investigate the properties of these series. Is it possible to guess whether gas prices are raised more quickly in response to increasing oil prices than they are decreased when oil prices are decreased? Use the cross-correlation function over the first 100 months compared with the cross-correlation function over the last 80 months. Do you think that it might be possible to predict log percentage changes in gas prices from log percentage changes in oil prices? Plot the two series on the same scale.

2.10 In this problem, we will explore the periodic nature of S_t, the SOI series displayed in Figure 1.5.

(a) Detrend the series by fitting a regression of S_t on time t. Is there a significant trend in the sea surface temperature? Comment.

(b) Calculate the periodogram for the detrended series obtained in part (a). Identify the frequencies of the two main peaks (with an obvious one at the frequency of one cycle every 12 months). What is the probable El Niño cycle indicated by the minor peak?

Section 2.4

2.11 For the data plotted in Figure 1.5, let S_t denote the SOI index series, and let R_t denote the Recruitment series.

(a) Draw a lag plot similar to the one in Figure 2.7 for R_t and comment.

(b) Reexamine the scatterplot matrix of R_t versus S_{t-h} shown in Figure 2.8 and the CCF of the two series shown in Figure 1.14, and fit the regression

$$\begin{aligned} R_t &= \alpha + \beta_0 S_t + \beta_1 S_{t-1} + \beta_2 S_{t-2} + \beta_3 S_{t-3} + \beta_4 S_{t-4} \\ &\quad + \beta_5 S_{t-5} + \beta_6 S_{t-6} + \beta_7 S_{t-7} + \beta_8 S_{t-8} + w_t. \end{aligned}$$

Compare the magnitudes and signs of the coefficients $\beta_0, \ldots, \beta_8$ with the scatterplots in Figure 2.8 and with the CCF in Figure 1.14.

(c) Use some of the smoothing techniques described in §2.4 to discover whether a trend exists in the Recruitment series, R_t, and to explore the periodic behavior of the data.

(d) In Example 2.6, some nonlinear behavior exists between the current value of Recruitment and past values of the SOI index. Use the smoothing techniques described in §2.4 to explore this possibility, concentrating on the scatterplot of R_t versus S_{t-6}.

2.12 Use a smoothing technique described in §2.4 to estimate the trend in the global temperature series displayed in Figure 1.2. Use the entire data set (see Example 2.1 for details).

Chapter 3

ARIMA Models

3.1 Introduction

In Chapters 1 and 2, we introduced autocorrelation and cross-correlation functions (ACFs and CCFs) as tools for clarifying relations that may occur within and between time series at various lags. In addition, we explained how to build linear models based on classical regression theory for exploiting the associations indicated by large values of the ACF or CCF. The time domain, or regression, methods of this chapter are appropriate when we are dealing with possibly nonstationary, shorter time series; these series are the rule rather than the exception in many applications. In addition, if the emphasis is on forecasting future values, then the problem is easily treated as a regression problem. This chapter develops a number of regression techniques for time series that are all related to classical ordinary and weighted or correlated least squares.

Classical regression is often insufficient for explaining all of the interesting dynamics of a time series. For example, the ACF of the residuals of the simple linear regression fit to the global temperature data (see Example 2.3 of Chapter 2) reveals additional structure in the data that the regression did not capture. Instead, the introduction of correlation as a phenomenon that may be generated through lagged linear relations leads to proposing the autoregressive (AR) and autoregressive moving average (ARMA) models. Adding nonstationary models to the mix leads to the autoregressive integrated moving average (ARIMA) model popularized in the landmark work by Box and Jenkins (1970). The Box–Jenkins method for identifying a plausible ARIMA model is given in this chapter along with techniques for parameter estimation and forecasting for these models. A partial theoretical justification of the use of ARMA models is discussed in Appendix B, §B.4.

3.2 Autoregressive Moving Average Models

The classical regression model of Chapter 2 was developed for the static case, namely, we only allow the dependent variable to be influenced by current values of the independent variables. In the time series case, it is desirable to allow the dependent variable to be influenced by the past values of the independent variables and possibly by its own past values. If the present can be plausibly modeled in terms of only the past values of the independent inputs, we have the enticing prospect that forecasting will be possible.

Introduction to Autoregressive Models

Autoregressive models are based on the idea that the current value of the series, x_t, can be explained as a function of p past values, $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$, where p determines the number of steps into the past needed to forecast the current value. As a typical case, recall Example 1.10 in which data were generated using the model

$$x_t = x_{t-1} - .90x_{t-2} + w_t,$$

where w_t is white Gaussian noise with $\sigma_w^2 = 1$. We have now assumed the current value is a particular *linear* function of past values. The regularity that persists in Figure 1.9 gives an indication that forecasting for such a model might be a distinct possibility, say, through some version such as

$$x_{n+1}^n = x_n - .90x_{n-1},$$

where the quantity on the left-hand side denotes the forecast at the next period $n+1$ based on the observed data, $x_1, x_2, \ldots, x_n$. We will make this notion more precise in our discussion of forecasting (§3.5).

The extent to which it might be possible to forecast a real data series from its own past values can be assessed by looking at the autocorrelation function and the lagged scatterplot matrices discussed in Chapter 2. For example, the lagged scatterplot matrix for the Southern Oscillation Index (SOI), shown in Figure 2.7, gives a distinct indication that lags 1 and 2, for example, are linearly associated with the current value. The ACF shown in Figure 1.14 shows relatively large positive values at lags 1, 2, 12, 24, and 36 and large negative values at 18, 30, and 42. We note also the possible relation between the SOI and Recruitment series indicated in the scatterplot matrix shown in Figure 2.8. We will indicate in later sections on transfer function and vector AR modeling how to handle the dependence on values taken by other series.

The preceding discussion motivates the following definition.

Definition 3.1 *An* **autoregressive model of order p**, *abbreviated* **AR(p)**, *is of the form*

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t, \tag{3.1}$$

where x_t is stationary, $\phi_1, \phi_2, \ldots, \phi_p$ are constants ($\phi_p \neq 0$). Unless otherwise stated, we assume that w_t is a Gaussian white noise series with mean zero and

variance σ_w^2. The mean of x_t in (3.1) is zero. If the mean, μ, of x_t is not zero, replace x_t by $x_t - \mu$ in (3.1), i.e.,

$$x_t - \mu = \phi_1(x_{t-1} - \mu) + \phi_2(x_{t-2} - \mu) + \cdots + \phi_p(x_{t-p} - \mu) + w_t,$$

or write

$$x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t, \tag{3.2}$$

where $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$.

We note that (3.2) is similar to the regression model of §2.2, and hence the term auto (or self) regression. Some technical difficulties, however, develop from applying that model because the regressors, $x_{t-1}, \ldots, x_{t-p}$, are random components, whereas $\boldsymbol{z}_t$ was assumed to be fixed. A useful form follows by using the backshift operator (2.30) to write the AR(p) model, (3.1), as

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)x_t = w_t, \tag{3.3}$$

or even more concisely as

$$\phi(B)x_t = w_t. \tag{3.4}$$

The properties of $\phi(B)$ are important in solving (3.4) for x_t. This leads to the following definition.

Definition 3.2 *The* **autoregressive operator** *is defined to be*

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{3.5}$$

We initiate the investigation of AR models by considering the first-order model, AR(1), given by $x_t = \phi x_{t-1} + w_t$. Iterating backwards k times, we get

$$\begin{aligned} x_t &= \phi x_{t-1} + w_t = \phi(\phi x_{t-2} + w_{t-1}) + w_t \\ &= \phi^2 x_{t-2} + \phi w_{t-1} + w_t \\ &\vdots \\ &= \phi^k x_{t-k} + \sum_{j=0}^{k-1} \phi^j w_{t-j}. \end{aligned}$$

This method suggests that, by continuing to iterate backwards, and provided that $|\phi| < 1$ and x_t is stationary, we can represent an AR(1) model as a linear process given by[1]

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}. \tag{3.6}$$

[1]Note that $\lim_{k\to\infty} E\left(x_t - \sum_{j=0}^{k-1} \phi^j w_{t-j}\right)^2 = \lim_{k\to\infty} \phi^{2k} E\left(x_{t-k}^2\right) = 0$, so (3.6) exists in the mean square sense (see Appendix A for a definition).

The AR(1) process defined by (3.6) is stationary with mean

$$E(x_t) = \sum_{j=0}^{\infty} \phi^j E(w_{t-j}) = 0,$$

and autocovariance function,

$$\begin{aligned} \gamma(h) &= \operatorname{cov}(x_{t+h}, x_t) = E\left[\left(\sum_{j=0}^{\infty} \phi^j w_{t+h-j}\right)\left(\sum_{k=0}^{\infty} \phi^k w_{t-k}\right)\right] \\ &= \sigma_w^2 \sum_{j=0}^{\infty} \phi^j \phi^{j+h} = \sigma_w^2 \phi^h \sum_{j=0}^{\infty} \phi^{2j} = \frac{\sigma_w^2 \phi^h}{1-\phi^2}, \quad h \geq 0. \end{aligned} \tag{3.7}$$

Recall that $\gamma(h) = \gamma(-h)$, so we will only exhibit the autocovariance function for $h \geq 0$. From (3.7), the ACF of an AR(1) is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^h, \quad h \geq 0, \tag{3.8}$$

and $\rho(h)$ satisfies the recursion

$$\rho(h) = \phi\, \rho(h-1), \quad h = 1, 2, \ldots . \tag{3.9}$$

We will discuss the ACF of a general AR(p) model in §3.4.

Example 3.1 The Sample Path of an AR(1) Process

Figure 3.1 shows a time plot of two AR(1) processes, one with $\phi = .9$ and one with $\phi = -.9$; in both cases, $\sigma_w^2 = 1$. In the first case, $\rho(h) = .9^h$, for $h \geq 0$, so observations close together in time are positively correlated with each other. This result means that observations at contiguous time points will tend to be close in value to each other; this fact shows up in the top of Figure 3.1 as a very smooth sample path for x_t. Now, contrast this to the case in which $\phi = -.9$, so that $\rho(h) = (-.9)^h$, for $h \geq 0$. This result means that observations at contiguous time points are negatively correlated but observations two time points apart are positively correlated. This fact shows up in the bottom of Figure 3.1, where, for example, if an observation, x_t, is positive, the next observation, x_{t+1}, is typically negative, and the next observation, x_{t+2}, is typically positive. Thus, in this case, the sample path is very choppy.

A figure similar to Figure 3.1 can be created in R using the following commands:

```
> par(mfrow=c(2,1))
> plot(arima.sim(list(order=c(1,0,0), ar=.9), n=100),
+  ylab="x",main=(expression("AR(1)  "*phi*" = +.9")))
> plot(arima.sim(list(order=c(1,0,0), ar=-.9), n=100),
+  ylab="x",main=(expression("AR(1)  "*phi*" = -.9")))
```

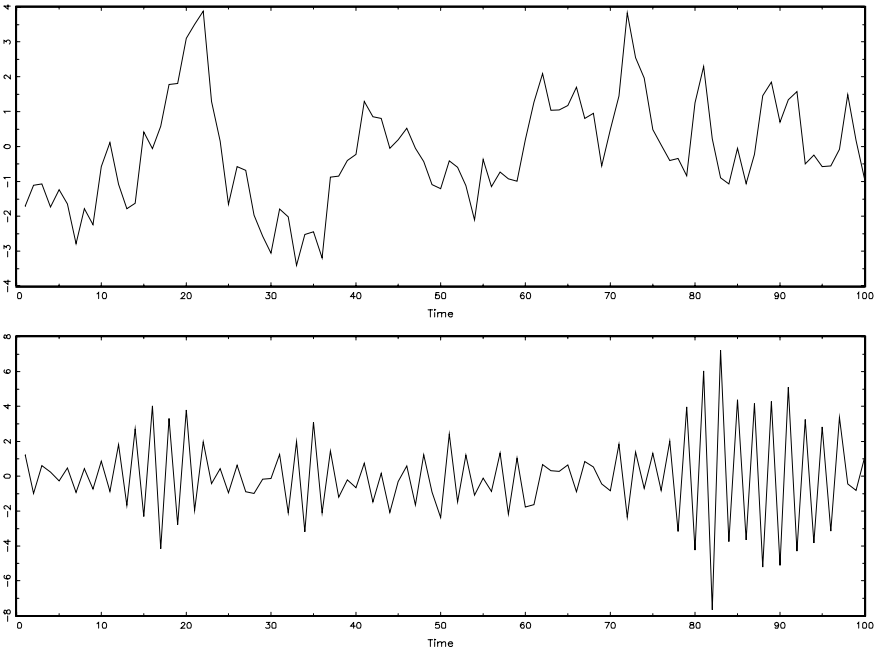


Figure 3.1 Simulated AR(1) models: $\phi = .9$ (top); $\phi = -.9$ (bottom).

Example 3.2 Explosive AR Models and Causality

In Example 1.18, it was discovered that the random walk $x_t = x_{t-1} + w_t$ is not stationary. We might wonder whether there is a stationary AR(1) process with $|\phi| > 1$. Such processes are called explosive because the values of the time series quickly become large in magnitude. Clearly, because $|\phi|^j$ increases without bound as $j \to \infty$, $\sum_{j=0}^{k-1} \phi^j w_{t-j}$ will not converge (in mean square) as $k \to \infty$, so the intuition used to get (3.6) will not work directly. We can, however, modify that argument to obtain a stationary model as follows. Write $x_{t+1} = \phi x_t + w_{t+1}$, in which case,

$$\begin{aligned} x_t &= \phi^{-1}x_{t+1} - \phi^{-1}w_{t+1} = \phi^{-1}\left(\phi^{-1}x_{t+2} - \phi^{-1}w_{t+2}\right) - \phi^{-1}w_{t+1} \\ &\vdots \\ &= \phi^{-k}x_{t+k} - \sum_{j=1}^{k-1} \phi^{-j}w_{t+j}, \end{aligned} \tag{3.10}$$

by iterating forward k steps. Because $|\phi|^{-1} < 1$, this result suggests the

stationary future dependent AR(1) model

$$x_t = -\sum_{j=1}^{\infty} \phi^{-j} w_{t+j}.$$

The reader can verify that this is stationary and of the AR(1) form $x_t = \phi x_{t-1} + w_t$. Unfortunately, this model is useless because it requires us to know the future to be able to predict the future. When a process does not depend on the future, such as the AR(1) when $|\phi| < 1$, we will say the process is causal. In the explosive case of this example, the process is stationary, but it is also future dependent, and not causal.

The technique of iterating backwards to get an idea of the stationary solution of AR models works well when $p = 1$, but not for larger orders. A general technique is that of matching coefficients. Consider the AR(1) model in operator form

$$\phi(B)x_t = w_t, \tag{3.11}$$

where $\phi(B) = 1 - \phi B$, and $|\phi| < 1$. Also, write the model in equation (3.6) using operator form as

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi(B)w_t, \tag{3.12}$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ and $\psi_j = \phi^j$. Suppose we did not know that $\psi_j = \phi^j$. We could substitute $\psi(B)w_t$ from (3.12) for x_t in (3.11) to obtain

$$\phi(B)\psi(B)w_t = w_t. \tag{3.13}$$

The coefficients of B on the left-hand side of (3.13) must be equal to those on right-hand side of (3.13), which means

$$(1 - \phi B)(1 + \psi_1 B + \psi_2 B^2 + \cdots + \psi_j B^j + \cdots) = 1. \tag{3.14}$$

Reorganizing the coefficients in (3.14),

$$1 + (\psi_1 - \phi)B + (\psi_2 - \psi_1\phi)B^2 + \cdots + (\psi_j - \psi_{j-1}\phi)B^j + \cdots = 1,$$

we see that for each $j = 1, 2, \ldots$, the coefficient of B^j on the left must be zero because it is zero on the right. The coefficient of B on the left is $(\psi_1 - \phi)$, and equating this to zero, $\psi_1 - \phi = 0$, leads to $\psi_1 = \phi$. Continuing, the coefficient of B^2 is $(\psi_2 - \psi_1\phi)$, so $\psi_2 = \phi^2$. In general,

$$\psi_j = \psi_{j-1}\phi,$$

with $\psi_0 = 1$, which leads to the general solution $\psi_j = \phi^j$.

Another way to think about the operations we just performed is to consider the AR(1) model in operator form, $\phi(B)x_t = w_t$. Now multiply both sides by $\phi^{-1}(B)$ (assuming the inverse operator exists) to get

$$\phi^{-1}(B)\phi(B)x_t = \phi^{-1}(B)w_t,$$

or

$$x_t = \phi^{-1}(B)w_t.$$

We know already that

$$\phi^{-1}(B) = 1 + \phi B + \phi^2 B^2 + \cdots + \phi^j B^j + \cdots,$$

that is, $\phi^{-1}(B)$ is $\psi(B)$ in (3.12). Thus, we notice that working with operators is like working with polynomials. That is, consider the polynomial $\phi(z) = 1 - \phi z$, where z is a complex number and $|\phi| < 1$. Then,

$$\phi^{-1}(z) = \frac{1}{(1-\phi z)} = 1 + \phi z + \phi^2 z^2 + \cdots + \phi^j z^j + \cdots, \quad |z| \leq 1,$$

and the coefficients of B^j in $\phi^{-1}(B)$ are the same as the coefficients of z^j in $\phi^{-1}(z)$. In other words, we may treat the backshift operator, B, as a complex number, z. These results will be generalized in our discussion of ARMA models. We will find the polynomials corresponding to the operators useful in exploring the general properties of ARMA models.

Introduction to Moving Average Models

As an alternative to the autoregressive representation in which the x_t on the left-hand side of the equation are assumed to be combined linearly, the moving average model of order q, abbreviated as MA(q), assumes the white noise w_t on the right-hand side of the defining equation are combined linearly to form the observed data.

Definition 3.3 *The* **moving average model of order q**, *or* **MA(q)** *model, is defined to be*

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q} \tag{3.15}$$

where there are q lags in the moving average and $\theta_1, \theta_2, \ldots, \theta_q$ *(*$\theta_q \neq 0$*) are parameters.*[2] *The noise* w_t *is assumed to be Gaussian white noise.*

The system is the same as the infinite moving average defined as the linear process (3.12), where $\psi_0 = 1$, $\psi_j = \theta_j$, for $j = 1, \ldots, q$, and $\psi_j = 0$ for other values. We may also write the MA(q) process in the equivalent form

$$x_t = \theta(B)w_t, \tag{3.16}$$

using the following definition.

[2]Some texts and software packages write the MA model with negative coefficients; that is, $x_t = w_t - \theta_1 w_{t-1} - \theta_2 w_{t-2} - \cdots - \theta_q w_{t-q}$.

Definition 3.4 *The* **moving average operator** *is*

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \tag{3.17}$$

Unlike the autoregressive process, the moving average process is stationary for any values of the parameters $\theta_1, \ldots, \theta_q$; details of this result are provided in §3.4.

Example 3.3 Autocorrelation and Sample Path of an MA(1) Process

Consider the MA(1) model $x_t = w_t + \theta w_{t-1}$. Then,

$$\gamma(h) = \begin{cases} (1+\theta^2)\sigma_w^2, & h = 0 \\ \theta\sigma_w^2, & h = 1 \\ 0, & h > 1, \end{cases}$$

and the autocorrelation function is

$$\rho(h) = \begin{cases} \frac{\theta}{(1+\theta^2)}, & h = 1 \\ 0, & h > 1. \end{cases}$$

Note $|\rho(1)| \leq 1/2$ for all values of θ (Problem 3.1). Also, x_t is correlated with x_{t-1}, but not with $x_{t-2}, x_{t-3}, \ldots$. Contrast this with the case of the AR(1) model in which the correlation between x_t and x_{t-k} is never zero. When $\theta = .5$, for example, x_t and x_{t-1} are positively correlated, and $\rho(1) = .4$. When $\theta = -.5$, x_t and x_{t-1} are negatively correlated, $\rho(1) = -.4$. Figure 3.2 shows a time plot of these two processes with $\sigma_w^2 = 1$. The series in Figure 3.2 where $\theta = .5$ is smoother than the series in Figure 3.2, where $\theta = -.5$.

A figure similar to Figure 3.2 can be created in R as follows:

```
> par(mfrow=c(2,1))
> plot(arima.sim(list(order=c(0,0,1), ma=.5), n=100),
+  ylab="x",main=(expression("MA(1)  "*theta*" = +.5")))
> plot(arima.sim(list(order=c(0,0,1), ma=-.5), n=100),
+  ylab="x",main=(expression("MA(1)  "*theta*" = -.5")))
```

Example 3.4 Non-uniqueness of MA Models and Invertibility

Using Example 3.3, we note that for an MA(1) model, $\rho(h)$ is the same for θ and $\frac{1}{\theta}$; try 5 and $\frac{1}{5}$, for example. In addition, the pair $\sigma_w^2 = 1$ and $\theta = 5$ yield the same autocovariance function as the pair $\sigma_w^2 = 25$ and $\theta = 1/5$, namely,

$$\gamma(h) = \begin{cases} 26, & h = 0 \\ 5, & h = 1 \\ 0, & h > 1. \end{cases}$$

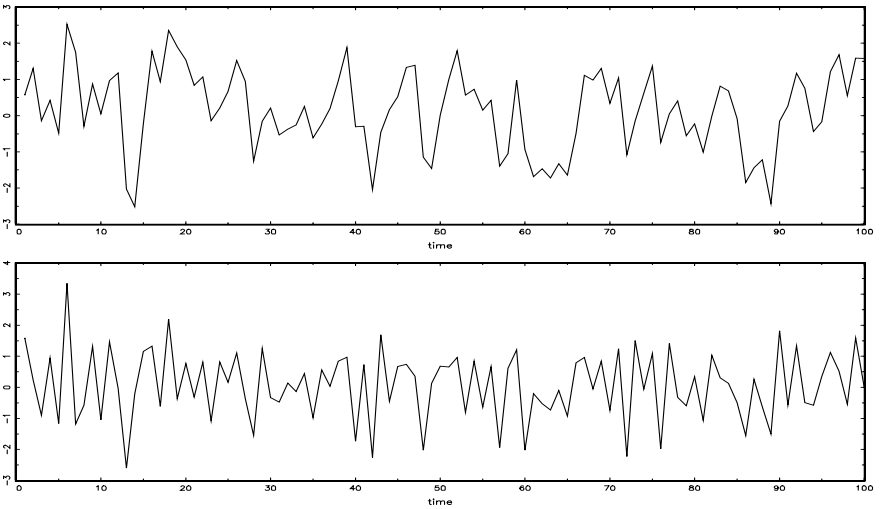


Figure 3.2 Simulated MA(1) models: $\theta = .5$ (top); $\theta = -.5$ (bottom).

Thus, the MA(1) processes

$$x_t = w_t + \frac{1}{5}w_{t-1}, \quad w_t \sim \text{iid } \text{N}(0, 25)$$

and

$$x_t = v_t + 5v_{t-1}, \quad v_t \sim \text{iid } \text{N}(0, 1)$$

are the same because of normality (i.e., all finite distributions are the same). We can only observe the time series x_t and not the noise, w_t or v_t, so we cannot distinguish between the models. Hence, we will have to choose only one of them. For convenience, by mimicking the criterion of causality for AR models, we will choose the model with an infinite AR representation. Such a process is called an invertible process.

To discover which model is the invertible model, we can reverse the roles of x_t and w_t (because we are mimicking the AR case) and write the MA(1) model as $w_t = -\theta w_{t-1} + x_t$. Following the steps that led to (3.6), if $|\theta| < 1$, then $w_t = \sum_{j=0}^{\infty}(-\theta)^j x_{t-j}$, which is the desired infinite AR representation of the model. Hence, given a choice, we will choose the model with $\sigma_w^2 = 25$ and $\theta = 1/5$ because it is invertible.

As in the AR case, the polynomial, $\theta(z)$, corresponding to the moving average operators, $\theta(B)$, will be useful in exploring general properties of MA processes. For example, following the steps of equations (3.11)–(3.14), we can write the MA(1) model as $x_t = \theta(B)w_t$, where $\theta(B) = 1 + \theta B$. If $|\theta| < 1$, then we can write the model as $\pi(B)x_t = w_t$, where $\pi(B) = \theta^{-1}(B)$. Let

$\theta(z) = 1 + \theta z$, for $|z| \leq 1$, then $\pi(z) = \theta^{-1}(z) = 1/(1 + \theta z) = \sum_{j=0}^{\infty}(-\theta)^j z^j$, and we determine that $\pi(B) = \sum_{j=0}^{\infty}(-\theta)^j B^j$.

Autoregressive Moving Average Models

We now proceed with the general development of autoregressive, moving average, and mixed autoregressive moving average (ARMA), models for stationary time series.

Definition 3.5 *A time series* $\{x_t;\ t = 0, \pm 1, \pm 2, \ldots\}$ *is* **ARMA(p, q)** *if it is stationary and*

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}, \tag{3.18}$$

with $\phi_p \neq 0$, $\theta_q \neq 0$, *and* $\sigma_w^2 > 0$. *The parameters* p *and* q *are called the autoregressive and the moving average orders, respectively. If* x_t *has a nonzero mean* μ, *we set* $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$ *and write the model as*

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}. \tag{3.19}$$

Unless stated otherwise, $\{w_t;\ t = 0, \pm 1, \pm 2, \ldots\}$ *is a Gaussian white noise sequence.*

As previously noted, when $q = 0$, the model is called an autoregressive model of order p, AR(p), and when $p = 0$, the model is called a moving average model of order q, MA(q). To aid in the investigation of ARMA models, it will be useful to write them using the AR operator, (3.5), and the MA operator, (3.17). In particular, the ARMA(p, q) model in (3.18) can then be written in concise form as

$$\phi(B)x_t = \theta(B)w_t. \tag{3.20}$$

Before we discuss the conditions under which (3.18) is causal and invertible, we point out a potential problem with the ARMA model.

Example 3.5 Parameter Redundancy

Consider a white noise process $x_t = w_t$. Equivalently, we can write this as $.5x_{t-1} = .5w_{t-1}$ by shifting back one unit of time and multiplying by .5. Now, subtract the two representations to obtain

$$x_t - .5x_{t-1} = w_t - .5w_{t-1},$$

or

$$x_t = .5x_{t-1} - .5w_{t-1} + w_t, \tag{3.21}$$

which looks like an ARMA(1, 1) model. Of course, x_t is still white noise; nothing has changed in this regard [i.e., $x_t = w_t$ is the solution to (3.21)],

but we have hidden the fact that x_t is white noise because of the parameter redundancy or over-parameterization. Write the parameter redundant model in operator form as $\phi(B)x_t = \theta(B)w_t$, or

$$(1 - .5B)x_t = (1 - .5B)w_t.$$

Apply the operator $\phi(B)^{-1} = (1 - .5B)^{-1}$ to both sides to obtain

$$x_t = (1 - .5B)^{-1}(1 - .5B)x_t = (1 - .5B)^{-1}(1 - .5B)w_t = w_t,$$

which is the original model. We can easily detect the problem of over-parameterization with the use of the operators or their associated polynomials. That is, write the AR polynomial $\phi(z) = (1 - .5z)$, the MA polynomial $\theta(z) = (1 - .5z)$, and note that both polynomials have a common factor, namely $(1 - .5z)$. This common factor immediately identifies the parameter redundancy. Discarding the common factor in each leaves $\phi(z) = 1$ and $\theta(z) = 1$, from which we conclude $\phi(B) = 1$ and $\theta(B) = 1$, and we deduce that the model is actually white noise. The consideration of parameter redundancy will be crucial when we discuss estimation for general ARMA models. As this example points out, we might fit an ARMA$(1, 1)$ model to white noise data and find that the parameter estimates are significant. If we were unaware of parameter redundancy, we might claim the data are correlated when in fact they are not (Problem 3.19).

Examples 3.2, 3.4, and 3.5 point to a number of problems with the general definition of ARMA(p, q) models, as given by (3.18), or, equivalently, by (3.20). To summarize, we have seen the following problems:

(i) parameter redundant models,

(ii) stationary AR models that depend on the future, and

(iii) MA models that are not unique.

To overcome these problems, we will require some additional restrictions on the model parameters. First, we make the following definitions.

Definition 3.6 *The* **AR and MA polynomials** *are defined as*

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p, \quad \phi_p \neq 0, \tag{3.22}$$

and

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q, \quad \theta_q \neq 0, \tag{3.23}$$

respectively, where z is a complex number.

To address the first problem, we will henceforth refer to an ARMA(p, q) model to mean that it is in its simplest form. That is, in addition to the original definition given in equation (3.18), we will also require that $\phi(z)$ and $\theta(z)$ have no common factors. So, the process, $x_t = .5x_{t-1} - .5w_{t-1} + w_t$, discussed in Example 3.5 is not referred to as an ARMA$(1, 1)$ process because, in its reduced form, x_t is white noise.

To address the problem of future-dependent models, we formally introduce the concept of causality.

Definition 3.7 *An ARMA(p, q) model, $\phi(B)x_t = \theta(B)w_t$, is said to be* **causal**, *if the time series $\{x_t;\ t = 0, \pm1, \pm2, \ldots\}$ can be written as a one-sided linear process:*

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi(B)w_t, \tag{3.24}$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$, and $\sum_{j=0}^{\infty} |\psi_j| < \infty$; we set $\psi_0 = 1$.

In Example 3.2, the AR(1) process, $x_t = \phi x_{t-1} + w_t$, is causal only when $|\phi| < 1$. Equivalently, the process is causal only when the root of $\phi(z) = 1 - \phi z$ is bigger than one in absolute value. That is, the root, say, z_0, of $\phi(z)$ is $z_0 = 1/\phi$ (because $\phi(z_0) = 0$) and $|z_0| > 1$ because $|\phi| < 1$. In general, we have the following property.

Property P3.1: Causality of an ARMA(p, q) Process
An ARMA(p, q) model is causal if and only if $\phi(z) \neq 0$ for $|z| \leq 1$. The coefficients of the linear process given in (3.24) can be determined by solving

$$\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\theta(z)}{\phi(z)}, \quad |z| \leq 1.$$

Another way to phrase Property P3.1 is that *an ARMA process is causal only when the roots of $\phi(z)$ lie outside the unit circle*; that is, $\phi(z) = 0$ only when $|z| > 1$. Finally, to address the problem of uniqueness discussed in Example 3.4, we choose the model that allows an infinite autoregressive representation.

Definition 3.8 *An ARMA(p, q) model, $\phi(B)x_t = \theta(B)w_t$, is said to be* **invertible**, *if the time series $\{x_t;\ t = 0, \pm1, \pm2, \ldots\}$ can be written as*

$$\pi(B)x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} = w_t, \tag{3.25}$$

where $\pi(B) = \sum_{j=0}^{\infty} \pi_j B^j$, and $\sum_{j=0}^{\infty} |\pi_j| < \infty$; we set $\pi_0 = 1$.

Analogous to Property P3.1, we have the following property.

Property P3.2: Invertibility of an ARMA(p, q) Process
An ARMA(p, q) model is invertible if and only if $\theta(z) \neq 0$ *for* $|z| \leq 1$. *The coefficients* π_j *of* $\pi(B)$ *given in (3.25) can be determined by solving*

$$\pi(z) = \sum_{j=0}^{\infty} \pi_j z^j = \frac{\phi(z)}{\theta(z)}, \quad |z| \leq 1.$$

Another way to phrase Property P3.2 is that *an ARMA process is invertible only when the roots of* $\theta(z)$ *lie outside the unit circle*; that is, $\theta(z) = 0$ only when $|z| > 1$. The proof of Property P3.1 is given in Appendix B (the proof of Property P3.2 is similar and, hence, is not provided). The following examples illustrate these concepts.

Example 3.6 Parameter Redundancy, Causality, and Invertibility

Consider the process

$$x_t = .4x_{t-1} + .45x_{t-2} + w_t + w_{t-1} + .25w_{t-2},$$

or, in operator form,

$$(1 - .4B - .45B^2)x_t = (1 + B + .25B^2)w_t.$$

At first, x_t appears to be an ARMA$(2, 2)$ process. But, the associated polynomials

$$\phi(z) = 1 - .4z - .45z^2 = (1 + .5z)(1 - .9z)$$

$$\theta(z) = (1 + z + .25z^2) = (1 + .5z)^2$$

have a common factor that can be canceled. After cancellation, the polynomials become $\phi(z) = (1 - .9z)$ and $\theta(z) = (1 + .5z)$, so the model is an ARMA$(1, 1)$ model, $(1 - .9B)x_t = (1 + .5B)w_t$, or

$$x_t = .9x_{t-1} + .5w_{t-1} + w_t. \tag{3.26}$$

The model is causal because $\phi(z) = (1 - .9z) = 0$ when $z = 10/9$, which is outside the unit circle. The model is also invertible because the root of $\theta(z) = (1 + .5z)$ is $z = -2$, which is outside the unit circle.

To write the model as a linear process, we can obtain the ψ-weights using Property P3.1:

$$\begin{aligned} \psi(z) &= \frac{\theta(z)}{\phi(z)} = \frac{(1 + .5z)}{(1 - .9z)} \\ &= (1 + .5z)(1 + .9z + .9^2z^2 + .9^3z^3 + \cdots) \quad |z| \leq 1. \end{aligned}$$

The coefficient of z^j in $\psi(z)$ is $\psi_j = (.5 + .9).9^{j-1}$, for $j \geq 1$, so (3.26) can be written as

$$x_t = w_t + 1.4 \sum_{j=1}^{\infty} .9^{j-1} w_{t-j}.$$

Similarly, to find the invertible representation using Property P3.2:

$$\pi(z) = \frac{\phi(z)}{\theta(z)} = (1 - .9z)(1 - .5z + .5^2 z^2 - .5^3 z^3 + \cdots) \quad |z| \leq 1.$$

In this case, the π-weights are given by $\pi_j = (-1)^j(.9 + .5).5^{j-1}$, for $j \geq 1$, and hence, we can also write (3.26) as

$$x_t = 1.4 \sum_{j=1}^{\infty} (-.5)^{j-1} x_{t-j} + w_t.$$

Example 3.7 Causal Conditions for an AR(2) Process

For an AR(1) model, $(1 - \phi B)x_t = w_t$, to be causal, the root of $\phi(z) = 1 - \phi z$ must lie outside of the unit circle. In this case, the root (or zero) occurs at $z_0 = 1/\phi$, i.e., $\phi(z_0) = 0$, so it is easy to go from the causal requirement on the root, that is, $|1/\phi| > 1$, to a requirement on the parameter, that is, $|\phi| < 1$. It is not so easy to establish this relationship for higher order models.

For example, the AR(2) model, $(1 - \phi_1 B - \phi_2 B^2)x_t = w_t$, is causal when the two roots of $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$ lie outside of the unit circle. Using the quadratic formula, this requirement can be written as

$$\left| \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2} \right| > 1.$$

The roots of $\phi(z)$ may be real and distinct, real and equal, or a complex conjugate pair. If we denote those roots by z_1 and z_2, we can write $\phi(z) = (1 - z_1^{-1} z)(1 - z_2^{-1} z)$; note that $\phi(z_1) = \phi(z_2) = 0$. The model can be written in operator form as $(1 - z_1^{-1} B)(1 - z_2^{-1} B)x_t = w_t$. From this representation, it follows that $\phi_1 = (z_1^{-1} + z_2^{-1})$ and $\phi_2 = -(z_1 z_2)^{-1}$. This relationship can be used to establish the following equivalent condition for causality:

$$\phi_1 + \phi_2 < 1, \quad \phi_2 - \phi_1 < 1, \quad \text{and} \quad |\phi_2| < 1. \tag{3.27}$$

This causality condition specifies a triangular region in the parameter space. We leave the details of the equivalence to the reader (Problem 3.4).

3.3 Difference Equations

The study of the behavior of ARMA processes and their ACFs is greatly enhanced by a basic knowledge of difference equations, simply because they are difference equations. This topic is also useful in the study of time domain models and stochastic processes in general. We will give a brief and heuristic account of the topic along with some examples of the usefulness of the theory. For details, the reader is referred to Mickens (1987).

Suppose we have a sequence of numbers $u_0, u_1, u_2, \ldots$ such that

$$u_n - \alpha u_{n-1} = 0, \quad \alpha \neq 0, \quad n = 1, 2, \ldots \ . \tag{3.28}$$

For example, recall (3.9) in which we showed that the ACF of an AR(1) process is a sequence, $\rho(h)$, satisfying

$$\rho(h) - \phi\rho(h-1) = 0, \quad h = 1, 2, \ldots \ .$$

Equation (3.28) represents a homogeneous difference equation of order 1. To solve the equation, we write:

$$\begin{aligned} u_1 &= \alpha u_0 \\ u_2 &= \alpha u_1 = \alpha^2 u_0 \\ &\vdots \\ u_n &= \alpha u_{n-1} = \alpha^n u_0. \end{aligned}$$

Given an initial condition $u_0 = c$, we may solve (3.28), namely, $u_n = \alpha^n c$.

In operator notation, (3.28) can be written as $(1 - \alpha B)u_n = 0$. The polynomial associated with (3.28) is $\alpha(z) = 1 - \alpha z$, and the root, say, z_0, of this polynomial is $z_0 = 1/\alpha$; that is $\alpha(z_0) = 0$. We know the solution to (3.28), with initial condition $u_0 = c$, is

$$u_n = \alpha^n c = \left(z_0^{-1}\right)^n c.$$

That is, the solution to the difference equation (3.28) depends only on the initial condition and the inverse of the root to the associated polynomial $\alpha(z)$.

Now suppose that the sequence satisfies

$$u_n - \alpha_1 u_{n-1} - \alpha_2 u_{n-2} = 0, \quad \alpha_2 \neq 0, \quad n = 2, 3, \ldots \tag{3.29}$$

This equation is a homogeneous difference equation of order 2. The corresponding polynomial is

$$\alpha(z) = 1 - \alpha_1 z - \alpha_2 z^2,$$

which has two roots, say, z_1 and z_2; that is, $\alpha(z_1) = \alpha(z_2) = 0$. We will consider two cases. First suppose $z_1 \neq z_2$. Then the general solution to (3.29) is

$$u_n = c_1 z_1^{-n} + c_2 z_2^{-n}, \tag{3.30}$$

where c_1 and c_2 depend on the initial conditions. This claim can be verified by direct substitution of (3.30) into (3.29):

$$\begin{aligned} c_1 z_1^{-n} + c_2 z_2^{-n} \; &- \alpha_1 \left(c_1 z_1^{-(n-1)} + c_2 z_2^{-(n-1)} \right) - \alpha_2 \left(c_1 z_1^{-(n-2)} + c_2 z_2^{-(n-2)} \right) \\ &= c_1 z_1^{-n} \left(1 - \alpha_1 z_1 - \alpha_2 z_1^2 \right) + c_2 z_2^{-n} \left(1 - \alpha_1 z_2 - \alpha_2 z_2^2 \right) \\ &= c_1 z_1^{-n} \alpha(z_1) + c_2 z_2^{-n} \alpha(z_2) \\ &= 0. \end{aligned}$$

Given two initial conditions u_0 and u_1, we may solve for c_1 and c_2:

$$\begin{aligned} u_0 &= c_1 + c_2 \\ u_1 &= c_1 z_1^{-1} + c_2 z_2^{-1}, \end{aligned}$$

where z_1 and z_2 can be solved for in terms of α_1 and α_2 using the quadratic formula, for example.

When the roots are equal, $z_1 = z_2 \; (= z_0)$, the general solution to (3.29) is

$$u_n = z_0^{-n}(c_1 + c_2 n). \tag{3.31}$$

This claim can also be verified by direct substitution of (3.31) into (3.29):

$$\begin{aligned} z_0^{-n}(c_1 + c_2 n) \; &- \alpha_1 \left(z_0^{-(n-1)}[c_1 + c_2(n-1)] \right) - \alpha_2 \left(z_0^{-(n-2)}[c_1 + c_2(n-2)] \right) \\ &= z_0^{-n}(c_1 + c_2 n) \left(1 - \alpha_1 z_0 - \alpha_2 z_0^2 \right) + c_2 z_0^{-n+1} \left(\alpha_1 + 2\alpha_2 z_0 \right) \\ &= c_2 z_0^{-n+1} \left(\alpha_1 + 2\alpha_2 z_0 \right). \end{aligned}$$

To show that $(\alpha_1 + 2\alpha_2 z_0) = 0$, write $1 - \alpha_1 z - \alpha_2 z^2 = (1 - z_0^{-1} z)^2$, and take derivatives with respect to z on both sides of the equation to obtain $(\alpha_1 + 2\alpha_2 z) = 2z_0^{-1}(1 - z_0^{-1} z)$. Thus, $(\alpha_1 + 2\alpha_2 z_0) = 2z_0^{-1}(1 - z_0^{-1} z_0) = 0$, as was to be shown. Finally, given two initial conditions, u_0 and u_1, we can solve for c_1 and c_2:

$$\begin{aligned} u_0 &= c_1 \\ u_1 &= (c_1 + c_2) z_0^{-1}. \end{aligned}$$

To summarize these results, in the case of distinct roots, the solution to the homogeneous difference equation of degree two was

$$\begin{aligned} u_n \;=\; & z_1^{-n} \times (\text{a polynomial in } n \text{ of degree } m_1 - 1) \\ & + z_2^{-n} \times (\text{a polynomial in } n \text{ of degree } m_2 - 1), \end{aligned}$$

where m_1 is the multiplicity of the root z_1 and m_2 is the multiplicity of the root z_2. In this example, of course, $m_1 = m_2 = 1$, and we called the polynomials of degree zero c_1 and c_2, respectively. In the case of the repeated root, the solution was

$$u_n = z_0^{-n} \times (\text{a polynomial in } n \text{ of degree } m_0 - 1),$$

where m_0 is the multiplicity of the root z_0; that is, $m_0 = 2$. In this case, we wrote the polynomial of degree one as $c_1 + c_2 n$. In both cases, we solved for c_1 and c_2 given two initial conditions, u_0 and u_1.

Example 3.8 The ACF of an AR(2) Process

Suppose $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$ is a causal AR(2) process. Multiply each side of the model by x_{t-h} for $h > 0$, and take expectation:

$$E(x_t x_{t-h}) = \phi_1 E(x_{t-1} x_{t-h}) + \phi_2 E(x_{t-2} x_{t-h}) + E(w_t x_{t-h}).$$

The result is

$$\gamma(h) = \phi_1 \gamma(h-1) + \phi_2 \gamma(h-2), \quad h = 1, 2, \ldots . \tag{3.32}$$

In (3.32), we used the fact that $E(x_t) = 0$ and for $h > 0$,

$$E(w_t x_{t-h}) = E\big(w_t \sum_{j=0}^{\infty} \psi_j w_{t-h-j}\big) = 0.$$

Divide (3.32) through by $\gamma(0)$ to obtain the difference equation for the ACF of the process:

$$\rho(h) - \phi_1 \rho(h-1) - \phi_2 \rho(h-2) = 0, \quad h = 1, 2, \ldots . \tag{3.33}$$

The initial conditions are $\rho(0) = 1$ and $\rho(-1) = \phi_1/(1 - \phi_2)$, which is obtained by evaluating (3.33) for $h = 1$ and noting that $\rho(1) = \rho(-1)$.

Using the results for the homogeneous difference equation of order two, let z_1 and z_2 be the roots of the associated polynomial, $\phi(z) = 1 - \phi_1 z - \phi_2 z^2$. Because the model is causal, we know the roots are outside the unit circle: $|z_1| > 1$ and $|z_2| > 1$. Now, consider the solution for three cases:

(i) When z_1 and z_2 are real and distinct, then

$$\rho(h) = c_1 z_1^{-h} + c_2 z_2^{-h},$$

so $\rho(h) \to 0$ exponentially fast as $h \to \infty$.

(ii) When $z_1 = z_2 (= z_0)$ are real and equal, then

$$\rho(h) = z_0^{-h}(c_1 + c_2 h),$$

so $\rho(h) \to 0$ exponentially fast as $h \to \infty$.

(iii) When $z_1 = \bar{z}_2$ are a complex conjugate pair, then $c_2 = \bar{c}_1$ (because $\rho(h)$ is real), and

$$\rho(h) = c_1 z_1^{-h} + \bar{c}_1 \bar{z}_1^{-h}.$$

Write c_1 and z_1 in polar coordinates, for example, $z_1 = |z_1| e^{i\theta}$, where θ is the angle whose tangent is the ratio of the imaginary

part and the real part of z_1 (sometimes called $\arg(z_1)$; the range of θ is $[-\pi, \pi]$). Then, using the fact that $e^{i\alpha} + e^{-i\alpha} = 2\cos(\alpha)$, the solution has the form

$$\rho(h) = a|z_1|^{-h}\cos(h\theta + b),$$

where a and b are determined by the initial conditions. Again, $\rho(h)$ dampens to zero exponentially fast as $h \to \infty$, but it does so in a sinusoidal fashion. The implication of this result is shown in the next example.

Example 3.9 The Sample Path of an AR(2) with Complex Roots

Figure 3.3 shows $n = 144$ observations from the AR(2) model

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t,$$

with $\sigma_w^2 = 1$, and with complex roots chosen so the process exhibits pseudo-cyclic behavior at the rate of one cycle every 12 time points. The autoregressive polynomial for this model is $\phi(z) = 1 - 1.5z + .75z^2$. The roots of $\phi(z)$ are $1 \pm i/\sqrt{3}$, and $\theta = \tan^{-1}(1/\sqrt{3}) = 2\pi/12$ radians per unit time. To convert the angle to cycles per unit time, divide by 2π to get $1/12$ cycles per unit time. The ACF for this model is shown in §3.4, Figure 3.4.

To reproduce Figure 3.3 in R:

```
> set.seed(5)
> ar2 = arima.sim(list(order = c(2,0,0), ar =c(1.5,-.75)),
+    n = 144)
> plot.ts(ar2, axes=F); box(); axis(2)
> axis(1, seq(0,144,24))
> abline(v=seq(0,144,12), lty="dotted")
```

To calculate and display the ACF for this model in R:

```
> acf = ARMAacf(ar=c(1.5,-.75), ma=0, 50)
> plot(acf, type="h", xlab="lag")
> abline(h=0)
```

We now exhibit the solution for the general homogeneous difference equation of order p:

$$u_n - \alpha_1 u_{n-1} - \cdots - \alpha_p u_{n-p} = 0, \quad \alpha_p \neq 0, \quad n = p, p+1, \ldots . \tag{3.34}$$

The associated polynomial is

$$\alpha(z) = 1 - \alpha_1 z - \cdots - \alpha_p z^p.$$

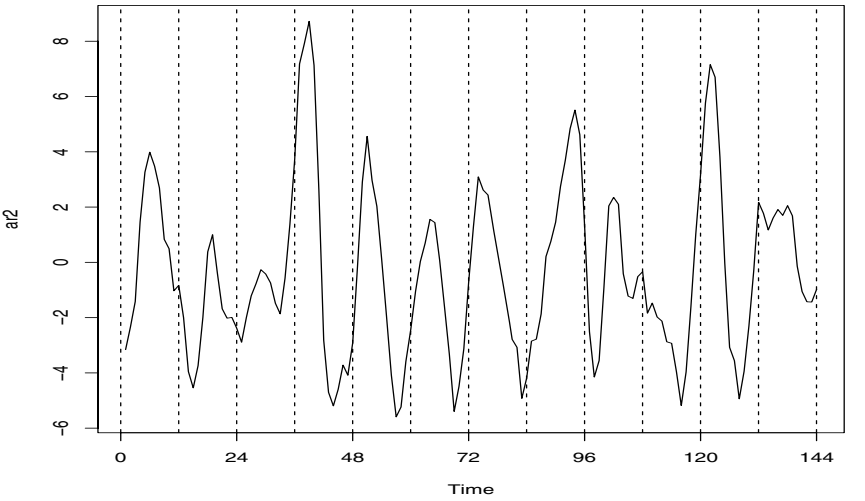


Figure 3.3 Simulated AR(2) model, $n = 144$ with $\phi_1 = 1.5$ and $\phi_2 = -.75$.

Suppose $\alpha(z)$ has r distinct roots, z_1 with multiplicity m_1, z_2 with multiplicity m_2, ..., and z_r with multiplicity m_r, such that $m_1 + m_2 + \cdots + m_r = p$. The general solution to the difference equation (3.34) is

$$u_n = z_1^{-n} P_1(n) + z_2^{-n} P_2(n) + \cdots + z_r^{-n} P_r(n), \tag{3.35}$$

where $P_j(n)$, for $j = 1, 2, \ldots, r$, is a polynomial in n, of degree $m_j - 1$. Given p initial conditions $u_0, \ldots, u_{p-1}$, we can solve for the $P_j(n)$ explicitly.

Example 3.10 Determining the ψ-weights for a Causal ARMA(p, q)

For a causal ARMA(p, q) model, $\phi(B)x_t = \theta(B)w_t$, where the zeros of $\phi(z)$ are outside the unit circle, recall that we may write

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j},$$

where the ψ-weights are determined using Property P3.1.

For the pure MA(q) model, $\psi_0 = 1$, $\psi_j = \theta_j$, for $j = 1, \ldots, q$, and $\psi_j = 0$, otherwise. For the general case of ARMA(p, q) models, the task of solving for the ψ-weights is much more complicated, as was demonstrated in Example 3.6. The use of the theory of homogeneous difference equations can help here. To solve for the ψ-weights in general, we must match the coefficients in $\psi(z)\phi(z) = \theta(z)$:

$$(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots)(1 - \phi_1 z - \phi_2 z^2 - \cdots) = (1 + \theta_1 z + \theta_2 z^2 + \cdots).$$

The first few values are

$$\begin{aligned} \psi_0 &= 1 \\ \psi_1 - \phi_1\psi_0 &= \theta_1 \\ \psi_2 - \phi_1\psi_1 - \phi_2\psi_0 &= \theta_2 \\ \psi_3 - \phi_1\psi_2 - \phi_2\psi_1 - \phi_3\psi_0 &= \theta_3 \\ &\vdots \end{aligned}$$

where we would take $\phi_j = 0$ for $j > p$, and $\theta_j = 0$ for $j > q$. The ψ-weights satisfy the homogeneous difference equation given by

$$\psi_j - \sum_{k=1}^{p} \phi_k \psi_{j-k} = 0, \quad j \geq \max(p, q+1), \tag{3.36}$$

with initial conditions

$$\psi_j - \sum_{k=1}^{j} \phi_k \psi_{j-k} = \theta_j, \quad 0 \leq j \leq \max(p, q+1). \tag{3.37}$$

The general solution depends on the roots of the AR polynomial $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$, as seen from (3.36). The specific solution will, of course, depend on the initial conditions.

Consider the ARMA process given in (3.26), $x_t = .9x_{t-1} + .5w_{t-1} + w_t$. Because $\max(p, q+1) = 2$, using (3.37), we have $\psi_0 = 1$ and $\psi_1 = .9 + .5 = 1.4$. By (3.36), for $j = 2, 3, \ldots$, the ψ-weights satisfy $\psi_j - .9\psi_{j-1} = 0$. The general solution is $\psi_j = c\,.9^j$. To find the specific solution, use the initial condition $\psi_1 = 1.4$, so $1.4 = .9c$ or $c = 1.4/.9$. Finally, $\psi_j = 1.4(.9)^{j-1}$, for $j \geq 1$, as we saw in Example 3.6.

To view, for example, the first 50 ψ-weights in R, use:

```
> ARMAtoMA(ar=.9, ma=.5, 50)         # for a list
> plot(ARMAtoMA(ar=.9, ma=.5, 50))   # for a graph
```

3.4 Autocorrelation and Partial Autocorrelation Functions

We begin by exhibiting the ACF of an MA(q) process, $x_t = \theta(B)w_t$, where $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$. Because x_t is a finite linear combination of white noise terms, the process is stationary with mean

$$E(x_t) = \sum_{j=0}^{q} \theta_j E(w_{t-j}) = 0,$$

where we have written $\theta_0 = 1$, and with autocovariance function

$$\begin{aligned}\gamma(h) = \text{cov}\,(x_{t+h}, x_t) &= E\left[\left(\sum_{j=0}^{q}\theta_j w_{t+h-j}\right)\left(\sum_{k=0}^{q}\theta_k w_{t-k}\right)\right] \\ &= \begin{cases}\sigma_w^2\sum_{j=0}^{q-h}\theta_j\theta_{j+h}, & 0 \le h \le q \\ 0, & h > q.\end{cases}\end{aligned} \tag{3.38}$$

Recall that $\gamma(h) = \gamma(-h)$, so we will only display the values for $h \ge 0$. The cutting off of $\gamma(h)$ after q lags is the signature of the MA(q) model. Dividing (3.38) by $\gamma(0)$ yields the ACF of an MA(q):

$$\rho(h) = \begin{cases}\frac{\sum_{j=0}^{q-h}\theta_j\theta_{j+h}}{1+\theta_1^2+\cdots+\theta_q^2}, & 1 \le h \le q \\ 0, & h > q.\end{cases} \tag{3.39}$$

For a causal ARMA(p, q) model, $\phi(B)x_t = \theta(B)w_t$, where the zeros of $\phi(z)$ are outside the unit circle, write

$$x_t = \sum_{j=0}^{\infty}\psi_j w_{t-j}.$$

It follows immediately that $E(x_t) = 0$. Also, the autocovariance function of x_t can be written as:

$$\gamma(h) = \text{cov}(x_{t+h}, x_t) = \sigma_w^2\sum_{j=0}^{\infty}\psi_j\psi_{j+h}, \quad h \ge 0. \tag{3.40}$$

We could then use (3.36) and (3.37) to solve for the ψ-weights. In turn, we could solve for $\gamma(h)$, and the ACF $\rho(h) = \gamma(h)/\gamma(0)$. As in Example 3.8, it is also possible to obtain a homogeneous difference equation directly in terms of $\gamma(h)$. First, we write

$$\begin{aligned}\gamma(h) &= \text{cov}(x_{t+h}, x_t) = E\left[\left(\sum_{j=1}^{p}\phi_j x_{t+h-j} + \sum_{j=0}^{q}\theta_j w_{t+h-j}\right)x_t\right] \\ &= \sum_{j=1}^{p}\phi_j\gamma(h-j) + \sigma_w^2\sum_{j=h}^{q}\theta_j\psi_{j-h}, \quad h \ge 0,\end{aligned} \tag{3.41}$$

where we have used the fact that $x_t = \sum_{k=0}^{\infty}\psi_k w_{t-k}$ and for $h \ge 0$,

$$E(w_{t+h-j}x_t) = E\left[w_{t+h-j}\left(\sum_{k=0}^{\infty}\psi_k w_{t-k}\right)\right] = \psi_{j-h}\sigma_w^2.$$

From (3.41), we can write a general homogeneous equation for the ACF of a causal ARMA process:

$$\gamma(h) - \phi_1\gamma(h-1) - \cdots - \phi_p\gamma(h-p) = 0, \quad h \geq \max(p, q+1), \tag{3.42}$$

with initial conditions

$$\gamma(h) - \sum_{j=1}^{p} \phi_j\gamma(h-j) = \sigma_w^2 \sum_{j=h}^{q} \theta_j\psi_{j-h}, \quad 0 \leq h < \max(p, q+1). \tag{3.43}$$

Dividing (3.42) and (3.43) through by $\gamma(0)$ will allow us to solve for the ACF, $\rho(h) = \gamma(h)/\gamma(0)$.

Example 3.11 The ACF of an ARMA$(1,1)$

Consider the causal ARMA$(1,1)$ process $x_t = \phi x_{t-1} + \theta w_{t-1} + w_t$, where $|\phi| < 1$. Based on (3.42), the autocovariance function satisfies

$$\gamma(h) - \phi\gamma(h-1) = 0, \quad h = 2, 3, \ldots,$$

so the general solution is $\gamma(h) = c\phi^h$, for $h = 1, 2, \ldots$. To obtain the initial conditions, we use (3.43):

$$\begin{aligned} \gamma(0) &= \phi\gamma(1) + \sigma_w^2[1 + \theta\phi + \theta^2] \\ \gamma(1) &= \phi\gamma(0) + \sigma_w^2\theta. \end{aligned}$$

Solving for $\gamma(0)$ and $\gamma(1)$, we obtain:

$$\begin{aligned} \gamma(0) &= \sigma_w^2 \frac{1 + 2\theta\phi + \theta^2}{1 - \phi^2} \\ \gamma(1) &= \sigma_w^2 \frac{(1 + \theta\phi)(\phi + \theta)}{1 - \phi^2}. \end{aligned}$$

To solve for c, note that $\gamma(1) = c\phi$, in which case $c = \gamma(1)/\phi$. Hence, the specific solution is

$$\gamma(h) = \sigma_w^2 \frac{(1 + \theta\phi)(\phi + \theta)}{1 - \phi^2}\phi^{h-1}.$$

Finally, dividing through by $\gamma(0)$ yields the ACF

$$\rho(h) = \frac{(1 + \theta\phi)(\phi + \theta)}{1 + 2\theta\phi + \theta^2}\phi^{h-1}, \quad h \geq 1. \tag{3.44}$$

Example 3.12 The ACF of an AR(p)

For a causal AR(p), it follows immediately from (3.42) that

$$\rho(h) - \phi_1\rho(h-1) - \cdots - \phi_p\rho(h-p) = 0, \quad h \geq p. \tag{3.45}$$

Let $z_1, \ldots, z_r$ denote the roots of $\phi(z)$, each with multiplicity $m_1, \ldots, m_r$, respectively, where $m_1 + \cdots + m_r = p$. Then, from (3.35), the general solution is

$$\rho(h) = z_1^{-h}P_1(h) + z_2^{-h}P_2(h) + \cdots + z_r^{-h}P_r(h), \quad h \geq p, \tag{3.46}$$

where $P_j(h)$ is a polynomial in h of degree $m_j - 1$.

Recall that for a causal model, all of the roots are outside the unit circle, $|z_i| > 1$, for $i = 1, \ldots, r$. If all the roots are real, then $\rho(h)$ dampens exponentially fast to zero as $h \to \infty$. If some of the roots are complex, then they will be in conjugate pairs and $\rho(h)$ will dampen, in a sinusoidal fashion, exponentially fast to zero as $h \to \infty$. In the case of complex roots, the time series will appear to be cyclic in nature. This, of course, is also true for ARMA models in which the AR part has complex roots.

The Partial Autocorrelation Function (PACF)

We have seen in (3.39), for MA(q) models, the ACF will be zero for lags greater than q. Moreover, because $\theta_q \neq 0$, the ACF will not be zero at lag q. Thus, the ACF provides a considerable amount of information about the order of the dependence when the process is a moving average process. If the process, however, is ARMA or AR, the ACF alone tells us little about the orders of dependence. Hence, it is worthwhile pursuing a function that will behave like the ACF of MA models, but for AR models, namely, the partial autocorrelation function (PACF).

To motivate the idea, consider a causal AR(1) model, $x_t = \phi x_{t-1} + w_t$. Then,

$$\begin{aligned} \gamma(2) = \operatorname{cov}(x_t, x_{t-2}) &= \operatorname{cov}(\phi x_{t-1} + w_t, x_{t-2}) \\ &= \operatorname{cov}(\phi^2 x_{t-2} + \phi w_{t-1} + w_t, x_{t-2}) = \phi^2\gamma(0). \end{aligned}$$

This result follows from causality because x_{t-2} involves $\{w_{t-2}, w_{t-3}, \ldots\}$, which are all uncorrelated with w_t and w_{t-1}. The correlation between x_t and x_{t-2} is not zero, as it would be for an MA(1), because x_t is dependent on x_{t-2} through x_{t-1}. Suppose we break this chain of dependence by removing (or partialing out) x_{t-1}. That is, we consider the correlation between $x_t - \phi x_{t-1}$ and $x_{t-2} - \phi x_{t-1}$, because it is the correlation between x_t and x_{t-2} with the linear dependence of each on x_{t-1} removed. In this way, we have broken the dependence chain between x_t and x_{t-2}. In fact,

$$\operatorname{cov}(x_t - \phi x_{t-1}, x_{t-2} - \phi x_{t-1}) = \operatorname{cov}(w_t, x_{t-2} - \phi x_{t-1}) = 0.$$

To formally define the PACF for mean-zero stationary time series, let x_h^{h-1} denote the regression of x_h on $\{x_{h-1}, x_{h-2}, \ldots, x_1\}$, which we write as[3]

$$x_h^{h-1} = \beta_1 x_{h-1} + \beta_2 x_{h-2} + \cdots + \beta_{h-1} x_1. \tag{3.47}$$

No intercept term is needed in (3.47) because the mean of x_t is zero. In addition, let x_0^{h-1} denote the regression of x_0 on $\{x_1, x_2, \ldots, x_{h-1}\}$, then

$$x_0^{h-1} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{h-1} x_{h-1}. \tag{3.48}$$

The coefficients, $\beta_1, \ldots, \beta_{h-1}$ are the same in (3.47) and (3.48); we will explain this result in the next section.

Definition 3.9 *The* **partial autocorrelation function (PACF)** *of a stationary process, x_t, denoted ϕ_{hh}, for $h = 1, 2, \ldots$, is*

$$\phi_{11} = \text{corr}(x_1, x_0) = \rho(1) \tag{3.49}$$

and

$$\phi_{hh} = \text{corr}(x_h - x_h^{h-1},\, x_0 - x_0^{h-1}), \quad h \geq 2. \tag{3.50}$$

Both $(x_h - x_h^{h-1})$ and $(x_0 - x_0^{h-1})$ are uncorrelated with $\{x_1, x_2, \ldots, x_{h-1}\}$. By stationarity, the PACF, ϕ_{hh}, is the correlation between x_t and x_{t-h} with the linear dependence of $\{x_{t-1}, \ldots, x_{t-(h-1)}\}$, on each, removed. If the process x_t is Gaussian, then $\phi_{hh} = \text{corr}(x_t, x_{t-h} |\, x_{t-1}, \ldots, x_{t-(h-1)})$. That is, ϕ_{hh} is the correlation coefficient between x_t and x_{t-h} in the bivariate distribution of (x_t, x_{t-h}) conditional on $\{x_{t-1}, \ldots, x_{t-(h-1)}\}$.

Example 3.13 The PACF of a Causal AR(1)

Consider the PACF of the AR(1) process given by $x_t = \phi x_{t-1} + w_t$, with $|\phi| < 1$. By definition, $\phi_{11} = \rho(1) = \phi$. To calculate ϕ_{22}, consider the regression of x_2 on x_1, say, $x_2^1 = \beta x_1$. We choose β to minimize

$$E(x_2 - \beta x_1)^2 = \gamma(0) - 2\beta\gamma(1) + \beta^2\gamma(0).$$

Taking derivatives and setting the result equal to zero, we have $\beta = \gamma(1)/\gamma(0) = \rho(1) = \phi$. Thus, $x_2^1 = \phi x_1$. Next, consider the regression of x_0 on x_1, say $x_0^1 = \beta x_1$. We choose β to minimize

$$E(x_0 - \beta x_1)^2 = \gamma(0) - 2\beta\gamma(1) + \beta^2\gamma(0).$$

This is the same equation as before, so $\beta = \phi$ and $x_0^1 = \phi x_1$. Hence, $\phi_{22} = \text{corr}(x_2 - \phi x_1, x_0 - \phi x_1)$. But, note

$$\text{cov}(x_2 - \phi x_1, x_0 - \phi x_1) = \gamma(2) - 2\phi\gamma(1) + \phi^2\gamma(0) = 0$$

because $\gamma(h) = \gamma(0)\phi^h$. Thus, $\phi_{22} = 0$. In the next example, we will see that in this case $\phi_{hh} = 0$, for all $h > 1$.

[3]The term regression here refers to regression in the population sense. That is, x_h^{h-1} is the linear combination of $\{x_{h-1}, x_{h-2}, \ldots, x_1\}$ that minimizes $E(x_h - \sum_{j=1}^{h-1} \alpha_j x_j)^2$.

Example 3.14 The PACF of a Causal AR(p)

Let $x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + w_t$, where the roots of $\phi(z)$ are outside the unit circle. In particular, $x_h = \sum_{j=1}^{p} \phi_j x_{h-j} + w_h$. When $h > p$, the regression of x_h on $x_{h-1}, \ldots, x_1$, is

$$x_h^{h-1} = \sum_{j=1}^{p} \phi_j x_{h-j}.$$

We have not proved this obvious result yet, but we will prove it in the next section. Thus, when $h > p$,

$$\begin{aligned} \phi_{hh} &= \operatorname{corr}(x_h - x_h^{h-1}, x_0 - x_0^{h-1}) \\ &= \operatorname{corr}(w_h, x_0 - x_0^{h-1}) = 0, \end{aligned}$$

because, by causality, $x_0 - x_0^{h-1}$ depends only on $\{w_{h-1}, w_{h-2}, \ldots\}$; recall equation (3.48). When $h \leq p$, ϕ_{pp} is not zero, and $\phi_{11}, \ldots, \phi_{p-1,p-1}$ are not necessarily zero. Figure 3.4 shows the ACF and the PACF of the AR(2) model presented in Example 3.9.

To reproduce Figure 3.4 in R, use the following commands:

```
> acf = ARMAacf(ar=c(1.5,-.75), ma=0, 24)
> pacf = ARMAacf(ar=c(1.5,-.75), ma=0, 24, pacf=T)
> par(mfrow=c(1,2))
> plot(acf, type="h", xlab="lag")
> abline(h=0)
> plot(pacf, type="h", xlab="lag")
> abline(h=0)
```

Example 3.15 The PACF of an Invertible MA(q)

For an invertible MA(q), we can write $x_t = -\sum_{j=1}^{\infty} \pi_j x_{t-j} + w_t$. Moreover, no finite representation exists. From this result, it should be apparent that the PACF will never cut off, as in the case of an AR(p).

For an MA(1), $x_t = w_t + \theta w_{t-1}$, with $|\theta| < 1$, calculations similar to Example 3.13 will yield $\phi_{22} = -\theta^2/(1 + \theta^2 + \theta^4)$. For the MA(1) in general, we can show that

$$\phi_{hh} = -\frac{(-\theta)^h(1 - \theta^2)}{1 - \theta^{2(h+1)}}, \quad h \geq 1.$$

In the next section, we will discuss methods of calculating the PACF. The PACF for MA models behaves much like the ACF for AR models. Also, the PACF for AR models behaves much like the ACF for MA models. Because an invertible ARMA model has an infinite AR representation, the PACF will not cut off. We may summarize these results in Table 3.1.

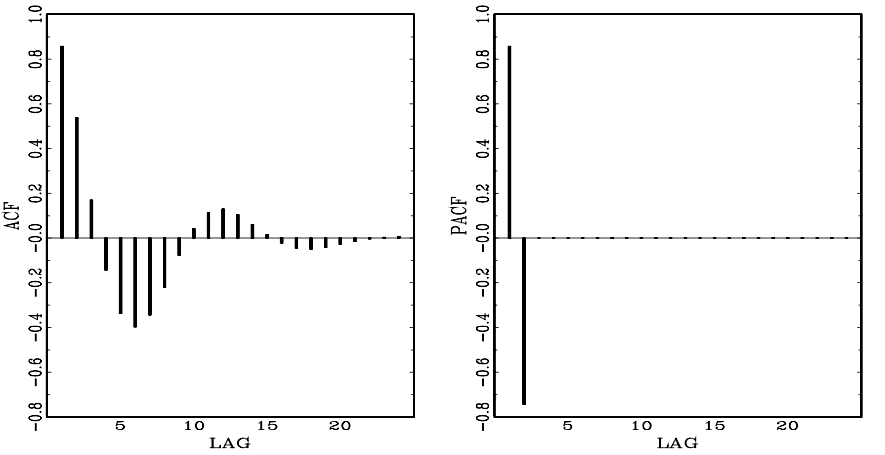


Figure 3.4 The ACF and PACF, to lag 24, of an AR(2) model with $\phi_1 = 1.5$ and $\phi_2 = -.75$.

Table 3.1 Behavior of the ACF and PACF for Causal and Invertible ARMA Models

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off	Cuts off after lag q	Tails off
PACF	Cuts off after lag p	Tails off	Tails off

Example 3.16 Preliminary Analysis of the Recruitment Series

We consider the problem of modeling the Recruitment series (number of new fish) shown in Figure 1.5. There are 453 months of observed recruitment ranging over the years 1950-1987. The ACF and the PACF given in Figure 3.5 are consistent with the behavior of an AR(2). The ACF has cycles corresponding roughly to a 12-month period, and the PACF has large values for $h = 1, 2$ and then is essentially zero for higher order lags. Based on Table 3.1, these results suggest that a second-order ($p = 2$) autoregressive model might provide a good fit. Although we will discuss estimation in detail in §3.6, we ran a regression (see §2.2) using the data triplets $\{(y; z_1, z_2) : (x_3; x_2, x_1), (x_4; x_3, x_2), \ldots, (x_{453}; x_{452}, x_{451})\}$ to fit a model of the form

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$$

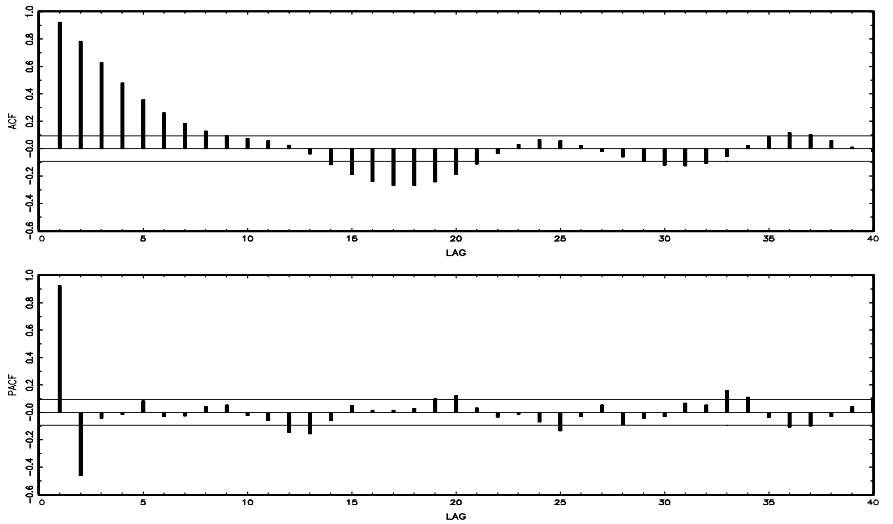


Figure 3.5 ACF and PACF of the Recruitment series.

for $t = 3, 4, \ldots, 453$. The values of the estimates were $\widehat{\phi}_0 = 6.74(1.11)$, $\widehat{\phi}_1 = 1.35(.04), \widehat{\phi}_2 = -.46(.04)$, and $\widehat{\sigma}_w^2 = 90.31$, where the estimated standard errors are in parentheses.

To reproduce this analysis and the ACF and PACF in Figure 3.5 in R:

```
> rec = scan("/mydata/recruit.dat")
> par(mfrow=c(2,1))
> acf(rec, 48)
> pacf(rec, 48)
> fit=ar.ols(rec,aic=F,order.max=2,demean=F,intercept=T)
> fit         # estimates
> fit$asy.se  # standard errors
```

3.5 Forecasting

In forecasting, the goal is to predict future values of a time series, x_{n+m}, $m = 1, 2, \ldots$, based on the data collected to the present, $\boldsymbol{x} = \{x_n, x_{n-1}, \ldots, x_1\}$. Throughout this section, we will assume x_t is stationary and the model parameters are known. The problem of forecasting when the model parameters are unknown will be discussed in the next section; also, see Problem 3.25. The minimum mean square error predictor of x_{n+m} is

$$x_{n+m}^n = E(x_{n+m} \mid x_n, x_{n-1}, \ldots, x_1)$$

because the conditional expectation minimizes the mean square error

$$E\left[x_{n+m} - g(\boldsymbol{x})\right]^2, \tag{3.51}$$

where $g(\boldsymbol{x})$ is a function of the observations $\boldsymbol{x}$; see Problem 3.13.

First, we will restrict attention to predictors that are linear functions of the data, that is, predictors of the form

$$x_{n+m}^n = \alpha_0 + \sum_{k=1}^{n} \alpha_k x_k, \tag{3.52}$$

where $\alpha_0, \alpha_1, \ldots, \alpha_n$ are real numbers. Linear predictors of the form (3.52) that minimize the mean square prediction error (3.51) are called best linear predictors (BLPs). As we shall see, linear prediction depends only on the second-order moments of the process, which are easy to estimate from the data. Much of the material in this section is enhanced by the theoretical material presented in Appendix B. For example, Theorem B.3 states that if the process is Gaussian, minimum mean square error predictors and best linear predictors are the same. The following property, which is based on the projection theorem, Theorem B.1 of Appendix B, is a key result.

Property P3.3: Best Linear Prediction for Stationary Processes
Given data $x_1, \ldots, x_n$, the best linear predictor, $x_{n+m}^n = \alpha_0 + \sum_{k=1}^n \alpha_k x_k$, of x_{n+m}, for $m \geq 1$, is found by solving

$$E\left[\left(x_{n+m} - x_{n+m}^n\right) x_k\right] = 0, \quad k = 0, 1, \ldots, n, \tag{3.53}$$

where $x_0 = 1$.

The equations specified in (3.53) are called the prediction equations, and they are used to solve for the coefficients $\{\alpha_0, \alpha_1, \ldots, \alpha_n\}$. If $E(x_t) = \mu$, the first equation ($k = 0$) of (3.53) implies

$$E(x_{n+m}^n) = E(x_{n+m}) = \mu.$$

Thus, taking expectation in (3.52), we have

$$\mu = \alpha_0 + \sum_{k=1}^{n} \alpha_k \mu \qquad \text{or} \qquad \alpha_0 = \mu\left(1 - \sum_{k=1}^{n} \alpha_k\right).$$

Hence, the form of the BLP is

$$x_{n+m}^n = \mu + \sum_{k=1}^{n} \alpha_k (x_k - \mu).$$

Thus, until we discuss estimation, there is no loss of generality in considering the case that $\mu = 0$, in which case, $\alpha_0 = 0$.

Consider, first, one-step-ahead prediction. That is, given $\{x_1, \ldots, x_n\}$, we wish to forecast the value of the time series at the next time point, x_{n+1}. The BLP of x_{n+1} is

$$x_{n+1}^n = \phi_{n1}x_n + \phi_{n2}x_{n-1} + \cdots + \phi_{nn}x_1, \tag{3.54}$$

where, for purposes that will become clear shortly, we have written α_k in (3.52), as $\phi_{n,n+1-k}$ in (3.54), for $k = 1, \ldots, n$. Using Property P3.3, the coefficients $\{\phi_{n1}, \phi_{n2}, \ldots, \phi_{nn}\}$ satisfy

$$E\left[\left(x_{n+1} - \sum_{j=1}^{n} \phi_{nj}x_{n+1-j}\right) x_{n+1-k}\right] = 0, \quad k = 1, \ldots, n,$$

or

$$\sum_{j=1}^{n} \phi_{nj}\gamma(k-j) = \gamma(k), \quad k = 1, \ldots, n. \tag{3.55}$$

The prediction equations (3.55) can be written in matrix notation as

$$\Gamma_n \boldsymbol{\phi}_n = \boldsymbol{\gamma}_n, \tag{3.56}$$

where $\Gamma_n = \{\gamma(k-j)\}_{j,k=1}^n$ is an $n \times n$ matrix, $\boldsymbol{\phi}_n = (\phi_{n1}, \ldots, \phi_{nn})'$ is an $n \times 1$ vector, and $\boldsymbol{\gamma}_n = (\gamma(1), \ldots, \gamma(n))'$ is an $n \times 1$ vector.

The matrix Γ_n is nonnegative definite. If Γ_n is singular, there are many solutions to (3.56), but, by the projection theorem (Theorem B.1), x_{n+1}^n is unique. If Γ_n is nonsingular, the elements of $\boldsymbol{\phi}_n$ are unique, and are given by

$$\boldsymbol{\phi}_n = \Gamma_n^{-1}\boldsymbol{\gamma}_n. \tag{3.57}$$

For ARMA models, the fact that $\sigma_w^2 > 0$ and $\gamma(h) \to 0$ as $h \to \infty$ is enough to ensure that Γ_n is positive definite (Problem 3.11). It is sometimes convenient to write the one-step-ahead forecast in vector notation

$$x_{n+1}^n = \boldsymbol{\phi}_n' \boldsymbol{x}, \tag{3.58}$$

where $\boldsymbol{x} = (x_n, x_{n-1}, \ldots, x_1)'$.

The mean square one-step-ahead prediction error is

$$P_{n+1}^n = E(x_{n+1} - x_{n+1}^n)^2 = \gamma(0) - \boldsymbol{\gamma}_n'\Gamma_n^{-1}\boldsymbol{\gamma}_n. \tag{3.59}$$

To verify (3.59) using (3.57) and (3.58),

$$\begin{aligned} E(x_{n+1} - x_{n+1}^n)^2 &= E(x_{n+1} - \boldsymbol{\phi}_n'\boldsymbol{x})^2 = E(x_{n+1} - \boldsymbol{\gamma}_n'\Gamma_n^{-1}\boldsymbol{x})^2 \\ &= E(x_{n+1}^2 - 2\boldsymbol{\gamma}_n'\Gamma_n^{-1}\boldsymbol{x}x_{n+1} + \boldsymbol{\gamma}_n'\Gamma_n^{-1}\boldsymbol{x}\boldsymbol{x}'\Gamma_n^{-1}\boldsymbol{\gamma}_n) \\ &= \gamma(0) - 2\boldsymbol{\gamma}_n'\Gamma_n^{-1}\boldsymbol{\gamma}_n + \boldsymbol{\gamma}_n'\Gamma_n^{-1}\Gamma_n\Gamma_n^{-1}\boldsymbol{\gamma}_n \\ &= \gamma(0) - \boldsymbol{\gamma}_n'\Gamma_n^{-1}\boldsymbol{\gamma}_n. \end{aligned}$$

Example 3.17 Prediction for an AR(2)

Suppose we have a causal AR(2) process $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$, and one observation x_1. Then, using equation (3.57), the one-step-ahead prediction of x_2 based on x_1 is

$$x_2^1 = \phi_{11} x_1 = \frac{\gamma(1)}{\gamma(0)} x_1 = \rho(1) x_1.$$

Now, suppose we want the one-step-ahead prediction of x_3 based on two observations x_1 and x_2. We could use (3.57) again and solve

$$x_3^2 = \phi_{21} x_2 + \phi_{22} x_1 = (\gamma(1), \gamma(2)) \begin{pmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{pmatrix}^{-1} \begin{pmatrix} x_2 \\ x_1 \end{pmatrix},$$

but, it should be apparent from the model that $x_3^2 = \phi_1 x_2 + \phi_2 x_1$. Because $\phi_1 x_2 + \phi_2 x_1$ satisfies the prediction equations (3.53),

$$E\{[x_3 - (\phi_1 x_2 + \phi_2 x_1)]x_1\} = E(w_3 x_1) = 0,$$

$$E\{[x_3 - (\phi_1 x_2 + \phi_2 x_1)]x_2\} = E(w_3 x_2) = 0,$$

it follows that, indeed, $x_3^2 = \phi_1 x_2 + \phi_2 x_1$, and by the uniqueness of the coefficients in this case, that $\phi_{21} = \phi_1$ and $\phi_{22} = \phi_2$. Continuing in this way, it is easy to verify that, for $n \geq 2$,

$$x_{n+1}^n = \phi_1 x_n + \phi_2 x_{n-1}.$$

That is, $\phi_{n1} = \phi_1, \phi_{n2} = \phi_2$, and $\phi_{nj} = 0$, for $j = 3, 4, \ldots, n$.

From Example 3.17, it should be clear (Problem 3.38) that, if the time series is a causal AR(p) process, then, for $n \geq p$,

$$x_{n+1}^n = \phi_1 x_n + \phi_2 x_{n-1} + \cdots + \phi_p x_{n-p+1}. \tag{3.60}$$

For ARMA models in general, the prediction equations will not be as simple as the pure AR case. In addition, for n large, the use of (3.57) is prohibitive because it requires the inversion of a large matrix. There are, however, iterative solutions that do not require any matrix inversion. In particular, we mention the recursive solution due to Levinson (1947) and Durbin (1960).

Property P3.4: The Durbin–Levinson Algorithm
Equations (3.57) and (3.59) can be solved iteratively as follows:

$$\phi_{00} = 0, \quad P_1^0 = \gamma(0). \tag{3.61}$$

For $n \geq 1$,

$$\phi_{nn} = \frac{\rho(n) - \sum_{k=1}^{n-1} \phi_{n-1,k}\ \rho(n-k)}{1 - \sum_{k=1}^{n-1} \phi_{n-1,k}\ \rho(k)}, \quad P_{n+1}^n = P_n^{n-1}(1 - \phi_{nn}^2), \tag{3.62}$$

where, for $n \geq 2$,

$$\phi_{nk} = \phi_{n-1,k} - \phi_{nn}\phi_{n-1,n-k}, \quad k = 1, 2, \ldots, n-1. \tag{3.63}$$

The proof of Property P3.4 is left as an exercise; see Problem 3.12.

Example 3.18 Using the Durbin–Levinson Algorithm

To use the algorithm, start with $\phi_{00} = 0$, $P_1^0 = \gamma(0)$. Then, for $n = 1$,

$$\phi_{11} = \rho(1) \quad \text{and} \quad P_2^1 = \gamma(0)[1 - \phi_{11}^2].$$

For $n = 2$,

$$\begin{aligned} \phi_{22} &= \frac{\rho(2) - \phi_{11}\,\rho(1)}{1 - \phi_{11}\,\rho(1)} = \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2} \\ \phi_{21} &= \phi_{11} - \phi_{22}\phi_{11} = \rho(1)[1 - \phi_{22}] \\ P_3^2 &= \gamma(0)[1 - \phi_{11}^2][1 - \phi_{22}^2]. \end{aligned}$$

For $n = 3$,

$$\phi_{33} = \frac{\rho(3) - \phi_{21}\,\rho(2) - \phi_{22}\,\rho(1)}{1 - \phi_{21}\,\rho(1) - \phi_{22}\,\rho(2)},$$

and so on.

An important consequence of the Durbin–Levinson algorithm is (see Problem 3.12) as follows.

Property P3.5: Iterative Solution for the PACF
The PACF of a stationary process x_t, *can be obtained iteratively via (3.62) as* ϕ_{nn}, *for* $n = 1, 2, \ldots$.

Example 3.19 The PACF of an AR(2)

From Example 3.14, we know that for an AR(2), $\phi_{hh} = 0$ for $h > 2$, but we will use the results of Example 3.17 and Property P3.5 to calculate the first three values of the PACF. Recall (Example 3.8) that for an AR(2), $\rho(1) = \phi_1/(1 - \phi_2)$, and in general $\rho(h) - \phi_1\rho(h-1) - \phi_2\rho(h-2) = 0$, for $h \geq 2$. Then,

$$\begin{aligned} \phi_{11} &= \rho(1) = \frac{\phi_1}{1 - \phi_2} \\ \phi_{22} &= \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2} = \frac{\left[\phi_1\left(\frac{\phi_1}{1-\phi_2}\right) + \phi_2\right] - \left(\frac{\phi_1}{1-\phi_2}\right)^2}{1 - \left(\frac{\phi_1}{1-\phi_2}\right)^2} = \phi_2 \\ \phi_{21} &= \phi_1 \\ \phi_{33} &= \frac{\rho(3) - \phi_1\rho(2) - \phi_2\rho(1)}{1 - \phi_1\rho(1) - \phi_2\rho(2)} = 0. \end{aligned}$$

So far, we have concentrated on one-step-ahead prediction, but Property P3.3 allows us to calculate the BLP of x_{n+m} for any $m \geq 1$. Given data, $\{x_1, \ldots, x_n\}$, the m-step-ahead predictor is

$$x_{n+m}^n = \phi_{n1}^{(m)} x_n + \phi_{n2}^{(m)} x_{n-1} + \cdots + \phi_{nn}^{(m)} x_1, \tag{3.64}$$

where $\{\phi_{n1}^{(m)}, \phi_{n2}^{(m)}, \ldots, \phi_{nn}^{(m)}\}$ satisfy the prediction equations,

$$\sum_{j=1}^{n} \phi_{nj}^{(m)} E(x_{n+1-j} x_{n+1-k}) = E(x_{n+m} x_{n+1-k}), \quad k = 1, \ldots, n,$$

or

$$\sum_{j=1}^{n} \phi_{nj}^{(m)} \gamma(k-j) = \gamma(m+k-1), \quad k = 1, \ldots, n. \tag{3.65}$$

The prediction equations can again be written in matrix notation as

$$\Gamma_n \boldsymbol{\phi}_n^{(m)} = \boldsymbol{\gamma}_n^{(m)}, \tag{3.66}$$

where $\boldsymbol{\gamma}_n^{(m)} = (\gamma(m), \ldots, \gamma(m+n-1))'$, and $\boldsymbol{\phi}_n^{(m)} = (\phi_{n1}^{(m)}, \ldots, \phi_{nn}^{(m)})'$ are $n \times 1$ vectors.

The mean square m-step-ahead prediction error is

$$P_{n+m}^n = E\left(x_{n+m} - x_{n+m}^n\right)^2 = \gamma(0) - \boldsymbol{\gamma}_n^{(m)'} \Gamma_n^{-1} \boldsymbol{\gamma}_n^{(m)}. \tag{3.67}$$

Another useful algorithm for calculating forecasts was given by Brockwell and Davis (1991, Chapter 5). This algorithm follows directly from applying the projection theorem (Theorem B.1) to the innovations, $x_t - x_t^{t-1}$, for $t = 1, \ldots, n$, using the fact that the innovations $x_t - x_t^{t-1}$ and $x_s - x_s^{s-1}$ are uncorrelated for $s \neq t$ (see Problem 3.39). We present the case in which x_t is a mean-zero stationary time series.

Property P3.6: The Innovations Algorithm
The one-step-ahead predictors, x_{t+1}^t, and their mean-squared errors, P_{t+1}^t, can be calculated iteratively as

$$x_1^0 = 0, \quad P_1^0 = \gamma(0)$$

$$x_{t+1}^t = \sum_{j=1}^{t} \theta_{tj}(x_{t+1-j} - x_{t+1-j}^{t-j}), \quad t = 1, 2, \ldots \tag{3.68}$$

$$P_{t+1}^t = \gamma(0) - \sum_{j=0}^{t-1} \theta_{t,t-j}^2 P_{j+1}^j \quad t = 1, 2, \ldots, \tag{3.69}$$

where, for $j = 0, 1, \ldots, t-1$,

$$\theta_{t,t-j} = \left(\gamma(t-j) - \sum_{k=0}^{j-1} \theta_{j,j-k} \theta_{t,t-k} P_{k+1}^k\right) \left(P_{j+1}^j\right)^{-1}. \tag{3.70}$$

Given data $x_1, \ldots, x_n$, the innovations algorithm can be calculated successively for $t = 1$, then $t = 2$ and so on, in which case the calculation of x_{n+1}^n and P_{n+1}^n is made at the final step $t = n$. The m-step-ahead predictor and its mean-square error based on the innovations algorithm (Problem 3.39) are given by

$$x_{n+m}^n = \sum_{j=m}^{n+m-1} \theta_{n+m-1,j}(x_{n+m-j} - x_{n+m-j}^{n+m-j-1}), \tag{3.71}$$

$$P_{n+m}^n = \gamma(0) - \sum_{j=m}^{n+m-1} \theta_{n+m-1,j}^2 P_{n+m-j}^n, \tag{3.72}$$

where the $\theta_{n+m-1,j}$ are obtained by continued iteration of (3.70).

Example 3.20 Prediction for an MA(1)

The innovations algorithm lends itself well to prediction for moving average processes. Consider an MA(1) model, $x_t = w_t + \theta w_{t-1}$. Recall that $\gamma(0) = (1 + \theta^2)\sigma_w^2$, $\gamma(1) = \theta\sigma_w^2$, and $\gamma(h) = 0$ for $h > 1$. Then, using Property P3.6, we have

$$\begin{aligned} \theta_{n1} &= \theta\sigma_w^2/P_n^{n-1} \\ \theta_{nj} &= 0, \quad j = 2, \ldots, n \\ P_1^0 &= (1 + \theta^2)\sigma_w^2 \\ P_{n+1}^n &= (1 + \theta^2 - \theta\theta_{n1})\sigma_w^2. \end{aligned}$$

Finally, from (3.68), the one-step-ahead predictor is

$$x_{n+1}^n = \theta\left(x_n - x_n^{n-1}\right)\sigma_w^2/P_n^{n-1}.$$

Forecasting ARMA Processes

The general prediction equations (3.53) provide little insight into forecasting for ARMA models in general. There are a number of different ways to express these forecasts, and each aids in understanding the special structure of ARMA prediction. Throughout, we assume x_t is a causal and invertible ARMA(p, q) process, $\phi(B)x_t = \theta(B)w_t$, where $w_t \sim$ iid N$(0, \sigma_w^2)$. In the non-zero mean case, $E(x_t) = \mu$, simply replace x_t with $x_t - \mu$ in the model. First, we consider two types of forecasts. We write x_{n+m}^n to mean the minimum mean square error predictor of x_{n+m} based on the data $\{x_n, \ldots, x_1\}$, that is,

$$x_{n+m}^n = E(x_{n+m} \mid x_n, \ldots, x_1).$$

For ARMA models, it is easier to calculate the predictor of x_{n+m}, assuming we have the complete history of the process $\{x_n, x_{n-1}, \ldots\}$. We will denote the predictor of x_{n+m} *based on the infinite past* as

$$\widetilde{x}_{n+m} = E(x_{n+m} \mid x_n, x_{n-1}, \ldots).$$

The idea here is that, for large samples, $\widetilde{x}_{n+m}$ will provide a good approximation to x^n_{n+m}.

Now, write x_{n+m} in its causal and invertible forms:

$$x_{n+m} = \sum_{j=0}^{\infty} \psi_j w_{n+m-j}, \quad \psi_0 = 1 \tag{3.73}$$

$$w_{n+m} = \sum_{j=0}^{\infty} \pi_j x_{n+m-j}, \quad \pi_0 = 1. \tag{3.74}$$

Then, taking conditional expectations in (3.73), we have

$$\widetilde{x}_{n+m} = \sum_{j=0}^{\infty} \psi_j \widetilde{w}_{n+m-j} = \sum_{j=m}^{\infty} \psi_j w_{n+m-j}, \tag{3.75}$$

because, by (3.74),

$$\widetilde{w}_t \equiv E(w_t | x_n, x_{n-1}, \ldots) = \begin{cases} 0, & t > n \\ \\ w_t, & t \leq n. \end{cases}$$

Similarly, taking conditional expectations in (3.74), we have

$$0 = \widetilde{x}_{n+m} + \sum_{j=1}^{\infty} \pi_j \widetilde{x}_{n+m-j},$$

or

$$\widetilde{x}_{n+m} = -\sum_{j=1}^{m-1} \pi_j \widetilde{x}_{n+m-j} - \sum_{j=m}^{\infty} \pi_j x_{n+m-j}, \tag{3.76}$$

using the fact $E(x_t \,\big|\, x_n, x_{n-1}, \ldots) = x_t$, for $t \leq n$. Prediction is accomplished recursively using (3.76), starting with the one-step-ahead predictor, $m = 1$, and then continuing for $m = 2, 3, \ldots$. Using (3.75), we can write

$$x_{n+m} - \widetilde{x}_{n+m} = \sum_{j=0}^{m-1} \psi_j w_{n+m-j},$$

so the mean square prediction error can be written as

$$P^n_{n+m} = E(x_{n+m} - \widetilde{x}_{n+m})^2 = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2. \tag{3.77}$$

Also, we note, for a fixed sample size, n, the prediction errors are correlated. That is, for $k \geq 1$,

$$E\{(x_{n+m} - \widetilde{x}_{n+m})(x_{n+m+k} - \widetilde{x}_{n+m+k})\} = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j \psi_{j+k}. \tag{3.78}$$

Example 3.21 Long-Range Forecasts

Consider forecasting an ARMA process with mean μ. From the zero-mean case in (3.75) we can deduce that the m-step-ahead forecast can be written as

$$\widetilde{x}_{n+m} = \mu + \sum_{j=m}^{\infty} \psi_j w_{n+m-j}. \tag{3.79}$$

Noting that the ψ-weights dampen to zero exponentially fast, it is clear that

$$\widetilde{x}_{n+m} \to \mu$$

exponentially fast (in the mean square sense) as $m \to \infty$. Moreover, by (3.77), the mean square prediction error

$$P^n_{n+m} \to \sigma_w^2 \sum_{j=0}^{\infty} \psi_j^2, \tag{3.80}$$

exponentially fast as $m \to \infty$.

It should be clear from (3.79) and (3.80) that ARMA forecasts quickly settle to the mean with a constant prediction error as the forecast horizon, m, grows. This effect can be seen in Figure 3.6 where the recruitment series is forecast for 24 months; see Example 3.23.

When n is small, the general prediction equations (3.53) can be used easily. When n is large, we would use (3.76) by truncating, because only the data $x_1, x_2, \ldots, x_n$ are available. In this case, we can truncate (3.76) by setting $\sum_{j=n+m}^{\infty} \pi_j x_{n+m-j} = 0$. The truncated predictor is then written as

$$\widetilde{x}^n_{n+m} = -\sum_{j=1}^{m-1} \pi_j \widetilde{x}^n_{n+m-j} - \sum_{j=m}^{n+m-1} \pi_j x_{n+m-j}, \tag{3.81}$$

which is also calculated recursively, $m = 1, 2, \ldots$. The mean square prediction error, in this case, is approximated using (3.77).

For AR(p) models, and when $n > p$, equation (3.60) yields the exact predictor, x^n_{n+m}, of x_{n+m}, and there is no need for approximations. That is, for $n > p$, $\widetilde{x}^n_{n+m} = \widetilde{x}_{n+m} = x^n_{n+m}$. Also, in this case, the one-step-ahead prediction error is $E(x_{n+1} - x^n_{n+1})^2 = \sigma_w^2$. For general ARMA($p, q$) models, the truncated predictors (Problem 3.15) for $m = 1, 2, \ldots$, are

$$\widetilde{x}^n_{n+m} = \phi_1 \widetilde{x}^n_{n+m-1} + \cdots + \phi_p \widetilde{x}^n_{n+m-p} + \theta_1 \widetilde{w}^n_{n+m-1} + \cdots + \theta_q \widetilde{w}^n_{n+m-q}, \tag{3.82}$$

where $\widetilde{x}^n_t = x_t$ for $1 \le t \le n$ and $\widetilde{x}^n_t = 0$ for $t \le 0$. The truncated prediction errors are given by: $\widetilde{w}^n_t = 0$ for $t \le 0$ or $t > n$, and $\widetilde{w}^n_t = \phi(B)\widetilde{x}^n_t - \theta_1 \widetilde{w}^n_{t-1} - \cdots - \theta_q \widetilde{w}^n_{t-q}$ for $1 \le t \le n$.

Example 3.22 Forecasting an ARMA$(1,1)$ Series

Given data $x_1, \ldots, x_n$, for forecasting purposes, write the model as

$$x_{n+1} = \phi x_n + w_{n+1} + \theta w_n.$$

Then, based on (3.82), the one-step-ahead truncated forecast is

$$\widetilde{x}^n_{n+1} = \phi x_n + 0 + \theta \widetilde{w}^n_n.$$

For $m \geq 2$, we have

$$\widetilde{x}^n_{n+m} = \phi \widetilde{x}^n_{n+m-1},$$

which can be calculated recursively, $m = 2, 3, \ldots$.

To calculate $\widetilde{w}^n_n$, which is needed to initialize the successive forecasts, the model can be written as $w_t = x_t - \phi x_{t-1} - \theta w_{t-1}$ for $t = 1, \ldots, n$. For truncated forecasting, using (3.82), put $\widetilde{w}^n_0 = 0$, $\widetilde{w}^n_1 = x_1$, and then iterate the errors forward in time

$$\widetilde{w}^n_t = x_t - \phi x_{t-1} - \theta \widetilde{w}^n_{t-1}, \quad t = 2, \ldots, n.$$

The approximate forecast variance is computed from (3.77) using the ψ-weights determined as in Example 3.10. In particular, the ψ-weights satisfy $\psi_j = (\phi + \theta)\phi^{j-1}$, for $j \geq 1$. This result gives

$$\begin{aligned} P^n_{n+m} &= \sigma^2_w \left[1 + (\phi + \theta)^2 \sum_{j=1}^{m-1} \phi^{2(j-1)} \right] \\ &= \sigma^2_w \left[1 + \frac{(\phi + \theta)^2 (1 - \phi^{2(m-1)})}{(1 - \phi^2)} \right]. \end{aligned}$$

To assess the precision of the forecasts, prediction intervals are typically calculated along with the forecasts. In general, $(1 - \alpha)$ prediction intervals are of the form

$$x^n_{n+m} \pm c_{\frac{\alpha}{2}} \sqrt{P^n_{n+m}}, \tag{3.83}$$

where $c_{\alpha/2}$ is chosen to get the desired degree of confidence. For example, if the process is Gaussian, then choosing $c_{\alpha/2} = 2$ will yield an approximate 95% prediction interval for x_{n+m}. If we are interested in establishing prediction intervals over more than one time period, then $c_{\alpha/2}$ should be adjusted appropriately, for example, by using Bonferroni's inequality [see (4.55) in Chapter 4 or Johnson and Wichern, 1992, Chapter 5].

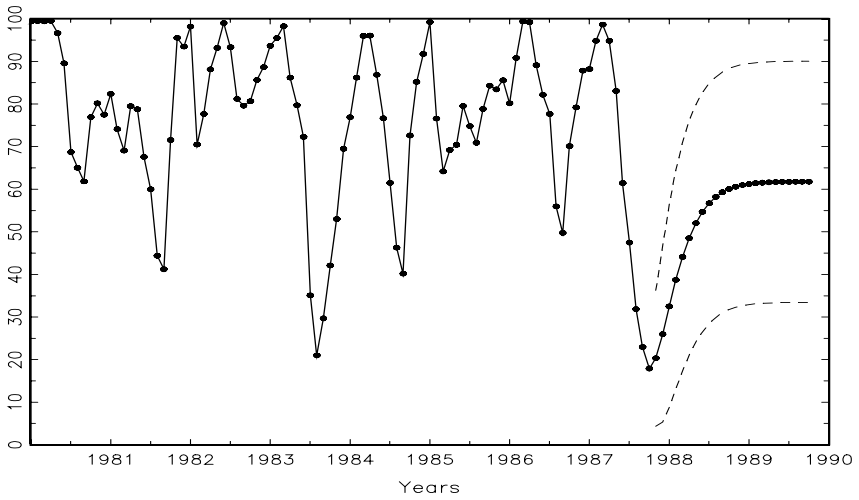


Figure 3.6 Twenty-four month forecasts for the Recruitment series. The actual data shown are from January 1980 to September 1987, and then forecasts plus and minus one standard error are displayed.

Example 3.23 Forecasting the Recruitment Series

Using the parameter estimates as the actual parameter values, Figure 3.6 shows the result of forecasting the Recruitment series given in Example 3.16 over a 24-month horizon, $m = 1, 2, \ldots, 24$. The actual forecasts are calculated as

$$x^n_{n+m} = 6.74 + 1.35x^n_{n+m-1} - .46x^n_{n+m-2}$$

for $n = 453$ and $m = 1, 2, \ldots, 12$. Recall that $x^s_t = x_t$ when $t \leq s$. The forecasts errors P^n_{n+m} are calculated using (3.77). Recall that $\widehat{\sigma}^2_w = 90.31$, and using (3.36) from Example 3.10, we have $\psi_j = 1.35\psi_{j-1} - .46\psi_{j-2}$ for $j \geq 2$, where $\psi_0 = 1$ and $\psi_1 = 1.35$. Thus, for $n = 453$,

$$\begin{aligned} P^n_{n+1} &= 90.31, \\ P^n_{n+2} &= 90.31(1 + 1.35^2), \\ P^n_{n+3} &= 90.31(1 + 1.35^2 + [1.35^2 - .46]^2), \end{aligned}$$

and so on.

Note how the forecast levels off quickly and the prediction intervals are wide, even though in this case the forecast limits are only based on one standard error; that is, $x^n_{n+m} \pm \sqrt{P^n_{n+m}}$. We will revisit this problem, including appropriate R commands, in Example 3.26.

We complete this section with a brief discussion of backcasting. In backcasting, we want to predict x_{1-m}, $m = 1, 2, \ldots$, based on the data $\{x_1, \ldots, x_n\}$.

Write the backcast as

$$x_{1-m}^n = \sum_{j=1}^{n} \alpha_j x_j. \tag{3.84}$$

Analogous to (3.65), the prediction equations (assuming $\mu = 0$) are

$$\sum_{j=1}^{n} \alpha_j E(x_j x_k) = E(x_{1-m} x_k), \quad k = 1, \ldots, n, \tag{3.85}$$

or

$$\sum_{j=1}^{n} \alpha_j \gamma(k-j) = \gamma(m+k-1), \quad k = 1, \ldots, n. \tag{3.86}$$

These equations are precisely the prediction equations for forward prediction. That is, $\alpha_j \equiv \phi_{nj}^{(m)}$, for $j = 1, \ldots, n$, where the $\phi_{nj}^{(m)}$ are given by (3.66). Finally, the backcasts are given by

$$x_{1-m}^n = \phi_{n1}^{(m)} x_1 + \cdots + \phi_{nn}^{(m)} x_n, \quad m = 1, 2, \ldots. \tag{3.87}$$

Example 3.24 Backcasting an ARMA$(1,1)$

Consider a causal and invertible ARMA(1,1) process, $x_t = \phi x_{t-1} + \theta w_{t-1} + w_t$; we will call this the forward model. We have just seen that best linear prediction backward in time is the same as best linear prediction forward in time for stationary models. Because we are assuming ARMA models are Gaussian, we also have that minimum mean square error prediction backward in time is the same as forward in time for ARMA models. Thus, the process can equivalently be generated by the backward model $x_t = \phi x_{t+1} + \theta v_{t+1} + v_t$, where $\{v_t\}$ is a Gaussian[4] white noise process with variance σ_w^2. We may write $x_t = \sum_{j=0}^{\infty} \psi_j v_{t+j}$, where $\psi_0 = 1$; this means that x_t is uncorrelated with $\{v_{t-1}, v_{t-2}, \ldots\}$, in analogy to the forward model.

Given data $\{x_1, \ldots, x_n\}$, truncate $v_n^n = E(v_n \mid x_1, \ldots, x_n)$ to zero. That is, put $\widetilde{v}_n^n = 0$, as an initial approximation, and then generate the errors backward

$$\widetilde{v}_t^n = x_t - \phi x_{t+1} + \theta \widetilde{v}_{t+1}^n, \quad t = (n-1), (n-2), \ldots, 1.$$

Then,

$$\widetilde{x}_0^n = \phi x_1 + \theta \widetilde{v}_1^n + \widetilde{v}_0^n = \phi x_1 + \theta \widetilde{v}_1^n,$$

because $\widetilde{v}_t^n = 0$ for $t \leq 0$. Continuing, the general truncated backcasts are given by

$$\widetilde{x}_{1-m}^n = \phi \widetilde{x}_{2-m}^n, \quad m = 2, 3, \ldots \ .$$

[4] In the stationary Gaussian case, (a) the distribution of $\{x_{n+1}, x_n, \ldots, x_1\}$ is the same as (b) the distribution of $\{x_0, x_1 \ldots, x_n\}$. In forecasting we use (a) to obtain $E(x_{n+1}|x_n, \ldots, x_1)$; in backcasting we use (b) to obtain $E(x_0|x_1, \ldots, x_n)$. Because (a) and (b) are the same, the two problems are equivalent.

3.6 Estimation

Throughout this section, we assume we have n observations, $x_1, \ldots, x_n$, from a causal and invertible Gaussian ARMA(p, q) process in which, initially, the order parameters, p and q, are known. Our goal is to estimate the parameters, $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$, and σ_w^2. We will discuss the problem of determining p and q later in this section.

We begin with method of moments estimators. The idea behind these estimators is that of equating population moments to sample moments and then solving for the parameters in terms of the sample moments. We immediately see that, if $E(x_t) = \mu$, then the method of moments estimator of μ is the sample average, $\bar{x}$. Thus, while discussing method of moments, we will assume $\mu = 0$. Although the method of moments can produce good estimators, they can sometimes lead to suboptimal estimators. We first consider the case in which the method leads to optimal (efficient) estimators, that is, AR(p) models.

When the process is AR(p),

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t,$$

the first $p + 1$ equations of (3.42) and (3.43), $h = 0, 1, \ldots, p$, lead to the following:

Definition 3.10 *The* **Yule–Walker equations** *are given by*

$$\begin{aligned} \gamma(h) &= \phi_1\gamma(h-1) + \cdots + \phi_p\gamma(h-p), \quad h = 1, 2, \ldots, p, && (3.88)\\ \sigma_w^2 &= \gamma(0) - \phi_1\gamma(1) - \cdots - \phi_p\gamma(p). && (3.89) \end{aligned}$$

In matrix notation, the Yule–Walker equations are

$$\Gamma_p\boldsymbol{\phi} = \boldsymbol{\gamma}_p, \quad \sigma_w^2 = \gamma(0) - \boldsymbol{\phi}'\boldsymbol{\gamma}_p, \tag{3.90}$$

where $\Gamma_p = \{\gamma(k-j)\}_{j,k=1}^p$ is a $p \times p$ matrix, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ is a $p \times 1$ vector, and $\boldsymbol{\gamma}_p = (\gamma(1), \ldots, \gamma(p))'$ is a $p \times 1$ vector. Using the method of moments, we replace $\gamma(h)$ in (3.90) by $\widehat{\gamma}(h)$ [see equation (1.36)] and solve

$$\widehat{\boldsymbol{\phi}} = \widehat{\Gamma}_p^{-1}\widehat{\boldsymbol{\gamma}}_p, \quad \widehat{\sigma}_w^2 = \widehat{\gamma}(0) - \widehat{\boldsymbol{\gamma}}_p'\widehat{\Gamma}_p^{-1}\widehat{\boldsymbol{\gamma}}_p. \tag{3.91}$$

These estimators are typically called the Yule–Walker estimators. For calculation purposes, it is sometimes more convenient to work with the sample ACF. By factoring $\widehat{\gamma}(0)$ in (3.91), we can write the Yule–Walker estimates as

$$\widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{R}}_p^{-1}\widehat{\boldsymbol{\rho}}_p, \quad \widehat{\sigma}_w^2 = \widehat{\gamma}(0)\left[1 - \widehat{\boldsymbol{\rho}}_p'\widehat{\boldsymbol{R}}_p^{-1}\widehat{\boldsymbol{\rho}}_p\right], \tag{3.92}$$

where $\widehat{R}_p = \{\widehat{\rho}(k-j)\}_{j,k=1}^p$ is a $p \times p$ matrix and $\widehat{\boldsymbol{\rho}}_p = (\widehat{\rho}(1), \ldots, \widehat{\rho}(p))'$ is a $p \times 1$ vector.

For AR(p) models, if the sample size is large, the Yule–Walker estimators are approximately normally distributed, and $\widehat{\sigma}_w^2$ is close to the true value of σ_w^2. We state these results in Property P3.7. For details, see Appendix B, §B.3.

Property P3.7: Large Sample Results for Yule–Walker Estimators
The asymptotic ($n \to \infty$) behavior of the Yule–Walker estimators in the case of causal AR(p) processes is as follows:

$$\sqrt{n}\left(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \sigma_w^2 \Gamma_p^{-1}\right), \qquad \widehat{\sigma}_w^2 \xrightarrow{p} \sigma_w^2. \tag{3.93}$$

The Durbin–Levinson algorithm, (3.61)-(3.63), can be used to calculate $\widehat{\boldsymbol{\phi}}$ without inverting $\widehat{\Gamma}_p$ or $\widehat{R}_p$, by replacing $\gamma(h)$ by $\widehat{\gamma}(h)$ in the algorithm. In running the algorithm, we will iteratively calculate the $h \times 1$ vector, $\widehat{\boldsymbol{\phi}}_h = (\widehat{\phi}_{h1}, \ldots, \widehat{\phi}_{hh})'$, for $h = 1, 2, \ldots$. Thus, in addition to obtaining the desired forecasts, the Durbin–Levinson algorithm yields $\widehat{\phi}_{hh}$, the sample PACF. Using (3.93), we can show the following property.

Property P3.8: Large Sample Distribution of the PACF
For a causal AR(p) process, asymptotically ($n \to \infty$),

$$\sqrt{n}\,\widehat{\phi}_{hh} \xrightarrow{d} \mathrm{N}(0, 1), \quad \text{for} \quad h > p. \tag{3.94}$$

Example 3.25 Yule–Walker Estimation for an AR(2) Process

The data shown in Figure 3.3 were $n = 144$ simulated observations from the AR(2) model

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t,$$

where $w_t \sim$ iid N$(0, 1)$. For this data, $\widehat{\gamma}(0) = 8.434$, $\widehat{\rho}(1) = .834$, and $\widehat{\rho}(2) = .476$. Thus,

$$\widehat{\boldsymbol{\phi}} = \begin{pmatrix} \widehat{\phi}_1 \\ \widehat{\phi}_2 \end{pmatrix} = \begin{bmatrix} 1 & .834 \\ .834 & 1 \end{bmatrix}^{-1} \begin{pmatrix} .834 \\ .476 \end{pmatrix} = \begin{pmatrix} 1.439 \\ -.725 \end{pmatrix}$$

and

$$\widehat{\sigma}_w^2 = 8.434\left[1 - (.834, .476)\begin{pmatrix} 1.439 \\ -.725 \end{pmatrix}\right] = 1.215.$$

By Property P3.7, the asymptotic variance–covariance matrix of $\widehat{\boldsymbol{\phi}}$,

$$\frac{1}{144}\frac{1.215}{8.434}\begin{bmatrix} 1 & .834 \\ .834 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} .057^2 & -.003 \\ -.003 & .057^2 \end{bmatrix},$$

can be used to get confidence regions for, or make inferences about $\widehat{\boldsymbol{\phi}}$ and its components. For example, an approximate 95% confidence interval

for ϕ_2 is $-.725 \pm 2(.057)$, or $(-.839, -.611)$, which contains the true value of $\phi_2 = -.75$.

For this data, the first three sample partial autocorrelations were $\widehat{\phi}_{11} = \widehat{\rho}(1) = .834$, $\widehat{\phi}_{22} = \widehat{\phi}_2 = -.725$, and $\widehat{\phi}_{33} = -.075$. According to Property P3.8, the asymptotic standard error of $\widehat{\phi}_{33}$ is $1/\sqrt{144} = .083$, and the observed value, $-.075$, is less than one standard deviation from $\phi_{33} = 0$.

Example 3.26 Yule–Walker Estimation of the Recruitment Series

In Example 3.16 we fit an AR(2) model to the recruitment series using regression. Below are the results of fitting the same model using Yule-Walker estimation in R (assuming the data are in `rec`), which are nearly identical to the values in Example 3.16.

```
> rec.yw = ar.yw(rec, order=2)
> rec.yw$x.mean
   [1] 62.26278                  # mean estimate
> rec.yw$ar
   [1]  1.3315874 -.4445447      # phi1 and phi2 estimates
> sqrt(diag(rec.yw$asy.var.coef))
   [1] .04222637 .04222637       # their standard errors
> rec.yw$var.pred
   [1] 94.79912                  # error variance estimate
```

To obtain the 24 month ahead predictions and their standard errors, and then plot the results as in Example 3.23, use the R commands:

```
> rec.pr = predict(rec.yw, n.ahead=24)
> U = rec.pr$pred + rec.pr$se
> L = rec.pr$pred - rec.pr$se
> month = 360:453
> plot(month, rec[month], type="o", xlim=c(360,480),
+   ylab="recruits")
> lines(rec.pr$pred, col="red", type="o")
> lines(U, col="blue", lty="dashed")
> lines(L, col="blue", lty="dashed")
```

In the case of AR(p) models, the Yule–Walker estimators given in (3.92) are optimal in the sense that the asymptotic distribution, (3.93), is the best asymptotic normal distribution. This is because, given initial conditions, AR(p) models are linear models, and the Yule–Walker estimators are essentially least squares estimators. If we use method of moments for MA or ARMA models, we will not get optimal estimators because such processes are nonlinear in the parameters.

Example 3.27 Method of Moments Estimation for an MA(1) Process

Consider the time series

$$x_t = w_t + \theta w_{t-1},$$

where $|\theta| < 1$. The model can then be written as

$$x_t = \sum_{j=1}^{\infty} (-\theta)^j x_{t-j} + w_t,$$

which is nonlinear in θ. The first two population autocovariances are $\gamma(0) = \sigma_w^2(1+\theta^2)$ and $\gamma(1) = \sigma_w^2\theta$, so the estimate of θ is found by solving:

$$\widehat{\rho}(1) = \frac{\widehat{\gamma}(1)}{\widehat{\gamma}(0)} = \frac{\widehat{\theta}}{1+\widehat{\theta}^2}.$$

Two solutions exist, so we would pick the invertible one. If $|\widehat{\rho}(1)| \leq \frac{1}{2}$, the solutions are real, otherwise, a real solution does not exist. Even though $|\rho(1)| < \frac{1}{2}$ for an invertible MA(1), it may happen that $|\widehat{\rho}(1)| \geq \frac{1}{2}$ because it is an estimator. When $|\widehat{\rho}(1)| < \frac{1}{2}$, the invertible estimate is

$$\widehat{\theta} = \frac{1 - \sqrt{1 - 4\widehat{\rho}(1)^2}}{2\widehat{\rho}(1)}.$$

It can be shown[5] that

$$\widehat{\theta} \sim \text{AN}\left(\theta,\ \frac{1+\theta^2+4\theta^4+\theta^6+\theta^8}{n(1-\theta^2)^2}\right).$$

The maximum likelihood estimator (which we discuss next) of θ, in this case, has an asymptotic variance of $(1-\theta^2)/n$. When $\theta = .5$, for example, the ratio of the asymptotic variance of the method of moments estimator to the maximum likelihood estimator of θ is about 3.5. That is, for large samples, the variance of the method of moments estimator is about 3.5 times larger than the variance of the MLE of θ when $\theta = .5$.

Maximum Likelihood and Least Squares Estimation

To fix ideas, we first focus on the causal AR(1) case. Let

$$x_t = \mu + \phi(x_{t-1} - \mu) + w_t$$

where $|\phi| < 1$ and $w_t \sim$ iid $\text{N}(0, \sigma_w^2)$. Given data $x_1, x_2, \ldots, x_n$, we seek the likelihood

$$L(\mu, \phi, \sigma_w^2) = f_{\mu,\phi,\sigma_w^2}(x_1, x_2, \ldots, x_n).$$

[5]The result follows by using the delta method and Theorem A.7 given in Appendix A. See the proof of Theorem A.7 for details on the delta method.

In the case of an AR(1), we may write the likelihood as

$$L(\mu, \phi, \sigma_w^2) = f(x_1) f(x_2 \mid x_1) \cdots f(x_n \mid x_{n-1}),$$

where we have dropped the parameters in the densities, $f(\cdot)$, to ease the notation. Because $x_t \mid x_{t-1} \sim \mathrm{N}\left(\mu + \phi(x_{t-1} - \mu), \sigma_w^2\right)$, we have

$$f(x_t \mid x_{t-1}) = f_w[(x_t - \mu) - \phi(x_{t-1} - \mu)],$$

where $f_w(\cdot)$ is the density of w_t, that is, the normal density with mean zero and variance σ_w^2. We may then write the likelihood as

$$L(\mu, \phi, \sigma_w) = f(x_1) \prod_{t=2}^{n} f_w \left[(x_t - \mu) - \phi(x_{t-1} - \mu)\right].$$

To find $f(x_1)$, we can use the causal representation

$$x_1 = \mu + \sum_{j=0}^{\infty} \phi^j w_{1-j}$$

to see that x_1 is normal, with mean μ and variance $\sigma_w^2/(1 - \phi^2)$. Finally, for an AR(1), the likelihood is

$$L(\mu, \phi, \sigma_w^2) = (2\pi\sigma_w^2)^{-n/2}(1 - \phi^2)^{1/2} \exp\left[-\frac{S(\mu, \phi)}{2\sigma_w^2}\right], \tag{3.95}$$

where

$$S(\mu, \phi) = (1 - \phi^2)(x_1 - \mu)^2 + \sum_{t=2}^{n} \left[(x_t - \mu) - \phi(x_{t-1} - \mu)\right]^2. \tag{3.96}$$

Typically, $S(\mu, \phi)$ is called the unconditional sum of squares. We could have also considered the estimation of μ and ϕ using unconditional least squares, that is, estimation by minimizing $S(\mu, \phi)$.

Taking the partial derivative of the log of (3.95) with respect to σ_w^2 and setting the result equal to zero, we see that for any given values of μ and ϕ in the parameter space, $\sigma_w^2 = n^{-1}S(\mu, \phi)$ maximizes the likelihood. Thus, the maximum likelihood estimate of σ_w^2 is

$$\widehat{\sigma}_w^2 = n^{-1} S(\widehat{\mu}, \widehat{\phi}), \tag{3.97}$$

where $\widehat{\mu}$ and $\widehat{\phi}$ are the MLEs of μ and ϕ, respectively. If we replace n in (3.97) by $n - 2$, we would obtain the unconditional least squares estimate of σ_w^2.

If, in (3.95), we take logs, replace σ_w^2 by $\widehat{\sigma}_w^2$, and ignore constants, $\widehat{\mu}$ and $\widehat{\phi}$ are the values that minimize the criterion function

$$l(\mu, \phi) = \ln\left[n^{-1}S(\mu, \phi)\right] - n^{-1}\ln(1 - \phi^2). \tag{3.98}$$

That is, $l(\mu, \phi) \propto -2 \ln L(\mu, \phi, \widehat{\sigma}_w^2)$.[6] Because (3.96) and (3.98) are complicated functions of the parameters, the minimization of $l(\mu, \phi)$ or $S(\mu, \phi)$ is accomplished numerically. In the case of AR models, we have the advantage that, conditional on initial values, they are linear models. That is, we can drop the term in the likelihood that causes the nonlinearity. Conditioning on x_1, the conditional likelihood becomes

$$\begin{aligned} L(\mu, \phi, \sigma_w^2 | x_1) &= \prod_{t=2}^{n} f_w \left[(x_t - \mu) - \phi(x_{t-1} - \mu) \right] \\ &= (2\pi\sigma_w^2)^{-(n-1)/2} \exp\left[-\frac{S_c(\mu, \phi)}{2\sigma_w^2} \right], \end{aligned} \tag{3.99}$$

where the conditional sum of squares is

$$S_c(\mu, \phi) = \sum_{t=2}^{n} \left[(x_t - \mu) - \phi(x_{t-1} - \mu) \right]^2. \tag{3.100}$$

The conditional MLE of σ_w^2 is

$$\widehat{\sigma}_w^2 = S_c(\widehat{\mu}, \widehat{\phi})/(n-1), \tag{3.101}$$

and $\widehat{\mu}$ and $\widehat{\phi}$ are the values that minimize the conditional sum of squares, $S_c(\mu, \phi)$. Letting $\alpha = \mu(1 - \phi)$, the conditional sum of squares can be written as

$$S_c(\mu, \phi) = \sum_{t=2}^{n} \left[x_t - (\alpha + \phi x_{t-1}) \right]^2. \tag{3.102}$$

The problem is now the linear regression problem stated in §2.2. Following the results from least squares estimation, we have $\widehat{\alpha} = \bar{x}_{(2)} - \widehat{\phi}\bar{x}_{(1)}$, where $\bar{x}_{(1)} = (n-1)^{-1} \sum_{t=1}^{n-1} x_t$, and $\bar{x}_{(2)} = (n-1)^{-1} \sum_{t=2}^{n} x_t$, and the conditional estimates are then

$$\widehat{\mu} = \frac{\bar{x}_{(2)} - \widehat{\phi}\bar{x}_{(1)}}{1 - \widehat{\phi}} \tag{3.103}$$

$$\widehat{\phi} = \frac{\sum_{t=2}^{n} (x_t - \bar{x}_{(2)})(x_{t-1} - \bar{x}_{(1)})}{\sum_{t=2}^{n} (x_{t-1} - \bar{x}_{(1)})^2}. \tag{3.104}$$

From (3.103) and (3.104), we see that $\widehat{\mu} \approx \bar{x}$ and $\widehat{\phi} \approx \widehat{\rho}(1)$. That is, the Yule–Walker estimators and the conditional least squares estimators are approximately the same. The only difference is the inclusion or exclusion of terms involving the end points, x_1 and x_n. We can also adjust the estimate of σ_w^2 in (3.101) to be equivalent to the least squares estimator, that is, divide $S_c(\widehat{\mu}, \widehat{\phi})$ by $(n-3)$ instead of $(n-1)$ in (3.101).

For general AR(p) models, maximum likelihood estimation, unconditional least squares, and conditional least squares follow analogously to the AR(1)

[6] The criterion function is sometimes called the profile likelihood.

example. For general ARMA models, it is difficult to write the likelihood as an explicit function of the parameters. Instead, it is advantageous to write the likelihood in terms of the innovations, or one-step-ahead prediction errors, $x_t - x_t^{t-1}$. This will also be useful in Chapter 6 when we study state-space models.

Suppose x_t is a causal ARMA(p,q) process with $w_t \sim$ iid N$(0,\sigma_w^2)$. Let $\boldsymbol{\beta} = (\mu, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$ be the $(p+q+1)\times 1$ vector of the model parameters. The likelihood can be written as

$$L(\boldsymbol{\beta}, \sigma_w^2) = \prod_{t=1}^{n} f(x_t \mid x_{t-1}, \ldots, x_1).$$

The conditional distribution of x_t given $x_{t-1}, \ldots, x_1$ is Gaussian with mean x_t^{t-1} and variance P_t^{t-1}. In addition, for ARMA models, we may write $P_t^{t-1} = \sigma_w^2 r_t^{t-1}$ where r_t^{t-1} does not depend on σ_w^2 (this can readily be seen from Proposition P3.4 by noting $P_1^0 = \gamma(0) = \sigma_w^2 \sum_{j=0}^{\infty} \psi_j^2$).

The likelihood of the data can now be written as

$$L(\boldsymbol{\beta}, \sigma_w^2) = (2\pi\sigma_w^2)^{-n/2} \left[r_1^0(\boldsymbol{\beta}) r_2^1(\boldsymbol{\beta}) \cdots r_n^{n-1}(\boldsymbol{\beta})\right]^{-1/2} \exp\left[-\frac{S(\boldsymbol{\beta})}{2\sigma_w^2}\right], \quad (3.105)$$

where

$$S(\boldsymbol{\beta}) = \sum_{t=1}^{n} \left[\frac{(x_t - x_t^{t-1}(\boldsymbol{\beta}))^2}{r_t^{t-1}(\boldsymbol{\beta})}\right]. \quad (3.106)$$

Both x_t^{t-1} and r_t^{t-1} are functions of $\boldsymbol{\beta}$, and we make that fact explicit in (3.105)-(3.106). Given values for $\boldsymbol{\beta}$ and σ_w^2, the likelihood may be evaluated using the techniques of §3.5. Maximum likelihood estimation would now proceed by maximizing (3.105) with respect to $\boldsymbol{\beta}$ and σ_w^2. As in the AR(1) example, we have

$$\widehat{\sigma}_w^2 = n^{-1} S(\widehat{\boldsymbol{\beta}}), \quad (3.107)$$

where $\widehat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ that minimizes the criterion function

$$l(\boldsymbol{\beta}) = \ln\left[n^{-1} S(\boldsymbol{\beta})\right] + n^{-1} \sum_{t=1}^{n} \ln r_t^{t-1}(\boldsymbol{\beta}). \quad (3.108)$$

For example, for the AR(1) model previously discussed, the generic $l(\boldsymbol{\beta})$ in (3.108) is $l(\mu,\phi)$ in (3.98), and the generic $S(\boldsymbol{\beta})$ in (3.106) is $S(\mu,\phi)$ given in (3.96). From (3.96) and (3.98) we see $x_1^0 = \mu$, and $x_t^{t-1} = \mu + \phi(x_{t-1} - \mu)$ for $t = 2, \ldots, n$. Also $r_1^0 = (1-\phi^2)$, and $r_t^{t-1} = 1$ for $t = 2, \ldots, n$.

Unconditional least squares would be performed by minimizing (3.106) with respect to $\boldsymbol{\beta}$. Conditional least squares estimation would involve minimizing (3.106) with respect to $\boldsymbol{\beta}$ but where, to ease the computational burden, the predictions and their errors are obtained by conditioning on initial values of the data. In general, numerical optimization routines are used to obtain the actual estimates and their standard errors.

Example 3.28 The Newton–Raphson and Scoring Algorithms

Two common numerical optimization routines for accomplishing maximum likelihood estimation are Newton–Raphson and scoring. We will give a brief account of the mathematical ideas here. The actual implementation of these algorithms is much more complicated than our discussion might imply. For details, the reader is referred to any of the *Numerical Recipes* books, for example, Press et al. (1993).

Let $l(\boldsymbol{\beta})$ be a criterion function of k parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ that we wish to minimize with respect to $\boldsymbol{\beta}$. For example, consider the likelihood function given by (3.98) or by (3.108). Suppose $l(\widehat{\boldsymbol{\beta}})$ is the extremum that we are interested in finding, and $\widehat{\boldsymbol{\beta}}$ is found by solving $\partial l(\boldsymbol{\beta})/\partial\beta_j = 0$, for $j = 1, \ldots, k$. Let $l^{(1)}(\boldsymbol{\beta})$ denote the $k \times 1$ vector of partials

$$l^{(1)}(\boldsymbol{\beta}) = \left(\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_k} \right)'.$$

Note, $l^{(1)}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, the $k \times 1$ zero vector. Let $l^{(2)}(\boldsymbol{\beta})$ denote the $k \times k$ matrix of second-order partials

$$l^{(2)}(\boldsymbol{\beta}) = \left\{ -\frac{\partial l^2(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} \right\}_{i,j=1}^{k},$$

and assume $l^{(2)}(\boldsymbol{\beta})$ is nonsingular. Let $\boldsymbol{\beta}_{(0)}$ be an initial estimator of $\boldsymbol{\beta}$. Then, using a Taylor expansion, we have the following approximation:

$$\mathbf{0} = l^{(1)}(\widehat{\boldsymbol{\beta}}) \approx l^{(1)}(\boldsymbol{\beta}_{(0)}) - l^{(2)}(\boldsymbol{\beta}_{(0)}) \left[\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right].$$

Setting the right-hand side equal to zero and solving for $\widehat{\boldsymbol{\beta}}$ (call the solution $\boldsymbol{\beta}_{(1)}$), we get

$$\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(0)} + \left[l^{(2)}(\boldsymbol{\beta}_{(0)}) \right]^{-1} l^{(1)}(\boldsymbol{\beta}_{(0)}).$$

The Newton–Raphson algorithm proceeds by iterating this result, replacing $\boldsymbol{\beta}_{(0)}$ by $\boldsymbol{\beta}_{(1)}$ to get $\boldsymbol{\beta}_{(2)}$, and so on, until convergence. Under a set of appropriate conditions, the sequence of estimators, $\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)}, \ldots$, will converge to $\widehat{\boldsymbol{\beta}}$, the MLE of $\boldsymbol{\beta}$.

For maximum likelihood estimation, the criterion function used is $l(\boldsymbol{\beta})$ given by (3.108); $l^{(1)}(\boldsymbol{\beta})$ is called the score vector, and $l^{(2)}(\boldsymbol{\beta})$ is called the Hessian. In the method of scoring, we replace $l^{(2)}(\boldsymbol{\beta})$ by $E[l^{(2)}(\boldsymbol{\beta})]$, the information matrix. Under appropriate conditions, the inverse of the information matrix is the asymptotic variance–covariance matrix of the estimator $\widehat{\boldsymbol{\beta}}$. This is sometimes approximated by the inverse of the Hessian at $\widehat{\boldsymbol{\beta}}$. If the derivatives are difficult to obtain, it is possible to use quasi-maximum likelihood estimation where numerical techniques are used to approximate the derivatives.

Example 3.29 MLE for the Recruitment Series

So far, we have fit an AR(2) model to the recruitment series using ordinary least squares (Example 3.16) and using Yule–Walker (Example 3.26). The following is an R session used to fit an AR(2) model via maximum likelihood estimation to the recruitment series; these results can be compared to the results in Examples 3.16 and 3.26. As before, we assume the data have been read into R as `rec`.

```
> rec.mle = ar.mle(rec, order=2)
> rec.mle$x.mean
   [1] 62.26153
> rec.mle$ar
   [1]  1.3512809 -.4612736
> sqrt(diag(rec.mle$asy.var.coef))
   [1] .04099159 .04099159
> rec.mle$var.pred
   [1] 89.33597
```

We now discuss least squares for ARMA(p, q) models via Gauss–Newton. For general and complete details of the Gauss–Newton procedure, the reader is referred to Fuller (1995). Let x_t be a causal and invertible Gaussian ARMA(p, q) process. Write $\boldsymbol{\beta} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$, and for the ease of discussion, we will put $\mu = 0$. We write the model in terms of the errors

$$w_t(\boldsymbol{\beta}) = x_t - \sum_{j=1}^{p} \phi_j x_{t-j} - \sum_{k=1}^{q} \theta_k w_{t-k}(\boldsymbol{\beta}), \tag{3.109}$$

emphasizing the dependence of the errors on the parameters.

For conditional least squares, we approximate the residual sum of squares by conditioning on $x_1, \ldots, x_p$ $(p > 0)$ and $w_p = w_{p-1} = w_{p-2} = \cdots = w_{1-q} = 0$ $(q > 0)$, in which case we may evaluate (3.109) for $t = p+1, p+2, \ldots, n$. Using this conditioning argument, the conditional error sum of squares is

$$S_c(\boldsymbol{\beta}) = \sum_{t=p+1}^{n} w_t^2(\boldsymbol{\beta}).$$

Minimizing $S_c(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ yields the conditional least squares estimates. If $q = 0$, the problem is linear regression, and no iterative technique is needed to minimize $S_c(\phi_1, \ldots, \phi_p)$. If $q > 0$, the problem becomes nonlinear regression, and we will have to rely on numerical optimization.

When n is large, conditioning on a few initial values will have little influence on the final parameter estimates. In the case of small to moderate sample sizes, one may wish to rely on unconditional least squares. The unconditional least squares problem is to choose $\boldsymbol{\beta}$ to minimize the unconditional sum of squares, which we have generically denoted by $S(\boldsymbol{\beta})$ in this section. The unconditional

sum of squares can be written in various ways, and one useful form in the case of ARMA(p, q) models is derived in Box et al. (1994, Appendix A7.3). They showed (see Problem 3.18) the unconditional sum of squares can be written as

$$S(\boldsymbol{\beta}) = \sum_{t=-\infty}^{n} \widehat{w}_t^2(\boldsymbol{\beta}),$$

where $\widehat{w}_t(\boldsymbol{\beta}) = E(w_t \mid x_1, \ldots, x_n)$. When $t \leq 0$, the $\widehat{w}_t(\boldsymbol{\beta})$ are obtained by backcasting. As a practical matter, we approximate $S(\boldsymbol{\beta})$ by starting the sum at $t = -M + 1$, where M is chosen large enough to guarantee $\sum_{t=-\infty}^{-M} \widehat{w}_t^2(\boldsymbol{\beta}) \approx 0$. In the case of unconditional least squares estimation, a numerical optimization technique is needed even when $q = 0$.

To employ Gauss–Newton, let $\boldsymbol{\beta}_{(0)} = (\phi_1^{(0)}, \ldots, \phi_p^{(0)}, \theta_1^{(0)}, \ldots, \theta_q^{(0)})'$ be an initial estimate of $\boldsymbol{\beta}$. For example, we could obtain $\boldsymbol{\beta}_{(0)}$ by method of moments. The first-order Taylor expansion of $w_t(\boldsymbol{\beta})$ is

$$w_t(\boldsymbol{\beta}) \approx w_t(\boldsymbol{\beta}_{(0)}) - \left(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}\right)' \boldsymbol{z}_t(\boldsymbol{\beta}_{(0)}), \tag{3.110}$$

where

$$\boldsymbol{z}_t(\boldsymbol{\beta}_{(0)}) = \left(-\frac{\partial w_t(\boldsymbol{\beta}_{(0)})}{\partial \beta_1}, \ldots, -\frac{\partial w_t(\boldsymbol{\beta}_{(0)})}{\partial \beta_{p+q}}\right)', \quad t = 1, \ldots, n.$$

The linear approximation of $S_c(\boldsymbol{\beta})$ is

$$Q(\boldsymbol{\beta}) = \sum_{t=p+1}^{n} \left[w_t(\boldsymbol{\beta}_{(0)}) - \left(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)}\right)' \boldsymbol{z}_t(\boldsymbol{\beta}_{(0)})\right]^2 \tag{3.111}$$

and this is the quantity that we will minimize. For approximate unconditional least squares, we would start the sum in (3.111) at $t = -M + 1$, for a large value of M, and work with the backcasted values.

Using the results of ordinary least squares (§2.2), we know

$$\widehat{(\boldsymbol{\beta} - \boldsymbol{\beta}_{(0)})} = \left(n^{-1} \sum_{t=p+1}^{n} \boldsymbol{z}_t(\boldsymbol{\beta}_{(0)}) \boldsymbol{z}_t'(\boldsymbol{\beta}_{(0)})\right)^{-1} \left(n^{-1} \sum_{t=p+1}^{n} \boldsymbol{z}_t(\boldsymbol{\beta}_{(0)}) w_t(\boldsymbol{\beta}_{(0)})\right) \tag{3.112}$$

minimizes $Q(\boldsymbol{\beta})$. From (3.112), we write the one-step Gauss–Newton estimate as

$$\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(0)} + \Delta(\boldsymbol{\beta}_{(0)}), \tag{3.113}$$

where $\Delta(\boldsymbol{\beta}_{(0)})$ denotes the right-hand side of (3.112). Gauss–Newton estimation is accomplished by replacing $\boldsymbol{\beta}_{(0)}$ by $\boldsymbol{\beta}_{(1)}$ in (3.113). This process is repeated by calculating, at iteration $j = 2, 3, \ldots$,

$$\boldsymbol{\beta}_{(j)} = \boldsymbol{\beta}_{(j-1)} + \Delta(\boldsymbol{\beta}_{(j-1)})$$

until convergence.

Example 3.30 Gauss–Newton for an MA(1)

Consider an invertible MA(1) process, $x_t = w_t + \theta w_{t-1}$. Write the truncated errors as

$$w_t(\theta) = x_t - \theta w_{t-1}(\theta), \quad t = 1, \ldots, n, \tag{3.114}$$

where we condition on $w_0(\theta) = 0$. Taking derivatives,

$$-\frac{\partial w_t(\theta)}{\partial \theta} = w_{t-1}(\theta) + \theta \frac{\partial w_{t-1}(\theta)}{\partial \theta}, \quad t = 1, \ldots, n, \tag{3.115}$$

where $\partial w_0(\theta)/\partial\theta = 0$. Using the notation of (3.110), we can also write (3.115) as

$$z_t(\theta) = w_{t-1}(\theta) - \theta z_{t-1}(\theta), \quad t = 1, \ldots, n, \tag{3.116}$$

where $z_0(\theta) = 0$.

Let $\theta_{(0)}$ be an initial estimate of θ, for example, the estimate given in Example 3.27. Then, the Gauss–Newton procedure for conditional least squares is given by

$$\theta_{(j+1)} = \theta_{(j)} + \frac{\sum_{t=1}^{n} z_t(\theta_{(j)}) w_t(\theta_{(j)})}{\sum_{t=1}^{n} z_t^2(\theta_{(j)})}, \quad j = 0, 1, 2, \ldots, \tag{3.117}$$

where the values in (3.117) are calculated recursively using (3.114) and (3.116). The calculations are stopped when $|\theta_{(j+1)} - \theta_{(j)}|$, or $|Q(\theta_{(j+1)}) - Q(\theta_{(j)})|$, are smaller than some preset amount.

Example 3.31 Fitting the Glacial Varve Series

Consider the series of glacial varve thicknesses from Massachusetts for $n = 634$ years, as analyzed in Example 2.5 and in Problem 1.8, where it was argued that a first-order moving average model might fit the logarithmically transformed and differenced varve series, say,

$$\nabla[\ln(x_t)] = \ln(x_t) - \ln(x_{t-1}) = \ln\left(\frac{x_t}{x_{t-1}}\right),$$

which can be interpreted as being proportional to the percentage change in the thickness.

The sample ACF and PACF, shown in Figure 3.7, confirm the tendency of $\nabla[\ln(x_t)]$ to behave as a first-order moving average process as the ACF has only a significant peak at lag one and the PACF decreases exponentially. Using Table 3.1, this sample behavior fits that of the MA(1) very well.

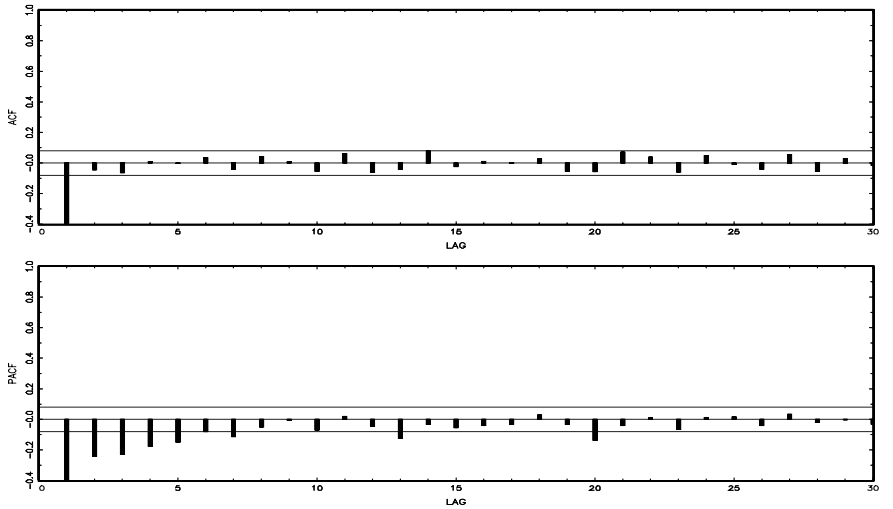


Figure 3.7 ACF and PACF of transformed glacial varves.

Nine iterations of the Gauss–Newton procedure, (3.117), starting with $\widehat{\theta}_0 = -.1$ yielded the values

$$-.442, -.624, -.717, -.750, -.763, -.768, -.771, -.772, -.772$$

for $\theta_{(1)}, \ldots, \theta_{(9)}$, and a final estimated error variance $\widehat{\sigma}_w^2 = .236$. Using the final value of $\widehat{\theta} = \theta_{(9)} = -.772$ and the vectors z_t of partial derivatives in (3.116) leads to a standard error of .025 and a t-value of $-.772/.025 = -30.88$ with 632 degrees of freedom (one is lost in differencing).

In the general case of causal and invertible ARMA(p, q) models, maximum likelihood estimation and conditional and unconditional least squares estimation (and Yule–Walker estimation in the case of AR models) all lead to optimal estimators. The proof of this general result can be found in a number of texts on theoretical time series analysis (for example, Brockwell and Davis, 1991, or Hannan, 1970, to mention a few). We will denote the ARMA coefficient parameters by $\boldsymbol{\beta} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)'$.

Property P3.9: Large Sample Distribution of the Estimators
Under appropriate conditions, for causal and invertible ARMA processes, the maximum likelihood, the unconditional least squares, and the conditional least squares estimators, each initialized by the method of moments estimator, all provide optimal estimators of σ_w^2 and $\boldsymbol{\beta}$, in the sense that $\widehat{\sigma}_w^2$ is consistent, and the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ is the best asymptotic normal distribution. In particular, as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \sigma_w^2\, \boldsymbol{\Gamma}_{p,q}^{-1}\right). \tag{3.118}$$

In (3.118), the variance–covariance matrix of the estimator $\widehat{\boldsymbol{\beta}}$ is the inverse of the information matrix. In this case, the $(p+q)\times(p+q)$ matrix $\Gamma_{p,q}$, has the form

$$\Gamma_{p,q} = \begin{pmatrix} \Gamma_{\phi\phi} & \Gamma_{\phi\theta} \\ \Gamma_{\theta\phi} & \Gamma_{\theta\theta} \end{pmatrix}. \tag{3.119}$$

The $p\times p$ matrix $\Gamma_{\phi\phi}$ is given by (3.90), that is, the ij-th element of $\Gamma_{\phi\phi}$, for $i,j=1,\ldots,p$, is $\gamma_x(i-j)$ from an AR(p) process, $\phi(B)x_t = w_t$. Similarly, $\Gamma_{\theta\theta}$ is a $q\times q$ matrix with the ij-th element, for $i,j=1,\ldots,q$, equal to $\gamma_y(i-j)$ from an AR(q) process, $\theta(B)y_t = w_t$. The $p\times q$ matrix $\Gamma_{\phi\theta} = \{\gamma_{xy}(i-j)\}$, for $i=1,\ldots,p$; $j=1,\ldots,q$; that is, the ij-th element is the cross-covariance between the two AR processes given by $\phi(B)x_t = w_t$ and $\theta(B)y_t = w_t$. Finally, $\Gamma_{\theta\phi} = \Gamma'_{\phi\theta}$ is $q\times p$. Further discussion of Property P3.9, including a proof for the case of least squares estimators for AR(p) processes, can be found in Appendix B, §B.3.

Example 3.32 Some Specific Asymptotic Distributions

The following are some specific cases of Property P3.9.

AR(1): $\gamma_x(0) = \sigma_w^2/(1-\phi^2)$, so $\sigma_w^2\Gamma_{1,0}^{-1} = (1-\phi^2)$. Thus,

$$\widehat{\phi} \sim \text{AN}\left[\phi, n^{-1}(1-\phi^2)\right]. \tag{3.120}$$

AR(2): The reader can verify that

$$\gamma_x(0) = \left(\frac{1-\phi_2}{1+\phi_2}\right)\frac{\sigma_w^2}{(1-\phi_2)^2-\phi_1^2}$$

and $\gamma_x(1) = \phi_1\gamma_x(0) + \phi_2\gamma_x(1)$. From these facts, we can compute $\Gamma_{2,0}^{-1}$. In particular, we have

$$\begin{pmatrix} \widehat{\phi}_1 \\ \widehat{\phi}_2 \end{pmatrix} \sim \text{AN}\left[\begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix},\ n^{-1}\begin{pmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ \text{sym} & 1-\phi_2^2 \end{pmatrix}\right]. \tag{3.121}$$

MA(1): In this case, write $\theta(B)y_t = w_t$, or $y_t + \theta y_{t-1} = w_t$. Then, analogous to the AR(1) case, $\gamma_y(0) = \sigma_w^2/(1-\theta^2)$, so $\sigma_w^2\Gamma_{0,1}^{-1} = (1-\theta^2)$. Thus,

$$\widehat{\theta} \sim \text{AN}\left[\theta, n^{-1}(1-\theta^2)\right]. \tag{3.122}$$

MA(2): Write $y_t + \theta_1 y_{t-1} + \theta_2 y_{t-2} = w_t$, so , analogous to the AR(2) case, we have

$$\begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{pmatrix} \sim \text{AN}\left[\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix},\ n^{-1}\begin{pmatrix} 1-\theta_2^2 & \theta_1(1+\theta_2) \\ \text{sym} & 1-\theta_2^2 \end{pmatrix}\right]. \tag{3.123}$$

ARMA(1,1): To calculate $\Gamma_{\phi\theta}$, we must find $\gamma_{xy}(0)$, where $x_t - \phi x_{t-1} = w_t$ and $y_t + \theta y_{t-1} = w_t$. We have

$$\begin{aligned}\gamma_{xy}(0) &= \operatorname{cov}(x_t, y_t) = \operatorname{cov}(\phi x_{t-1} + w_t, -\theta y_{t-1} + w_t) \\ &= -\phi\theta\gamma_{xy}(0) + \sigma_w^2.\end{aligned}$$

Solving, we find, $\gamma_{xy}(0) = \sigma_w^2/(1+\phi\theta)$. Thus,

$$\begin{pmatrix}\widehat{\phi}\\ \widehat{\theta}\end{pmatrix} \sim \text{AN}\left[\begin{pmatrix}\phi\\ \theta\end{pmatrix},\ n^{-1}\begin{bmatrix}(1-\phi^2)^{-1} & (1+\phi\theta)^{-1}\\ \text{sym} & (1-\theta^2)^{-1}\end{bmatrix}^{-1}\right]. \tag{3.124}$$

The reader might wonder, for example, why the asymptotic distributions of $\widehat{\phi}$ from an AR(1) [equation (3.120)] and $\widehat{\theta}$ from an MA(1) [equation (3.122)] are of the same form. It is possible to explain this unexpected result heuristically using the intuition of linear regression. That is, for the normal regression model presented in §2.2 with no intercept term, $x_t = \beta z_t + w_t$, we know $\widehat{\beta}$ is normally distributed with mean β, and from (2.8),

$$\operatorname{var}\left\{\sqrt{n}\left(\widehat{\beta}-\beta\right)\right\} = n\sigma_w^2\left(\sum_{t=1}^n z_t^2\right)^{-1} = \sigma_w^2\left(n^{-1}\sum_{t=1}^n z_t^2\right)^{-1}.$$

For the causal AR(1) model given by $x_t = \phi x_{t-1} + w_t$, the intuition of regression tells us to expect that, for n large,

$$\sqrt{n}\left(\widehat{\phi}-\phi\right)$$

is approximately normal with mean zero and with variance given by

$$\sigma_w^2\left(n^{-1}\sum_{t=2}^n x_{t-1}^2\right)^{-1}.$$

Now, $n^{-1}\sum_{t=2}^n x_{t-1}^2$ is the sample variance (recall that the mean of x_t is zero) of the x_t, so as n becomes large we would expect it to approach $\operatorname{var}(x_t) = \gamma(0) = \sigma_w^2/(1-\phi^2)$. Thus, the large sample variance of $\sqrt{n}\left(\widehat{\phi}-\phi\right)$ is

$$\sigma_w^2\gamma_x(0)^{-1} = \sigma_w^2\left(\frac{\sigma_w^2}{1-\phi^2}\right)^{-1} = (1-\phi^2);$$

that is, (3.120) holds.

In the case of an MA(1), we may use the discussion of Example 3.30 to write an approximate regression model for the MA(1). That is, consider the approximation (3.116) as the regression model

$$z_t(\widehat{\theta}) = -\theta z_{t-1}(\widehat{\theta}) + w_{t-1},$$

where now, $z_{t-1}(\widehat{\theta})$ as defined in Example 3.30, plays the role of the regressor. Continuing with the analogy, we would expect the asymptotic distribution of $\sqrt{n}\left(\widehat{\theta}-\theta\right)$ to be normal, with mean zero, and approximate variance

$$\sigma_w^2\left(n^{-1}\sum_{t=2}^{n} z_{t-1}^2(\widehat{\theta})\right)^{-1}.$$

As in the AR(1) case, $n^{-1}\sum_{t=2}^{n} z_{t-1}^2(\widehat{\theta})$ is the sample variance of the $z_t(\widehat{\theta})$ so, for large n, this should be $\mathrm{var}\{z_t(\theta)\} = \gamma_z(0)$, say. But note, as seen from (3.116), $z_t(\theta)$ is approximately an AR(1) process with parameter $-\theta$. Thus,

$$\sigma_w^2\gamma_z(0)^{-1} = \sigma_w^2\left(\frac{\sigma_w^2}{1-(-\theta)^2}\right)^{-1} = (1-\theta^2),$$

which agrees with (3.122). Finally, the asymptotic distributions of the AR parameters estimates and the MA parameter estimates are of the same form because in the MA case, the "regressors" are the differential processes $z_t(\theta)$ that have AR structure, and it is this structure that determines the asymptotic variance of the estimators. For a rigorous account of this approach for the general case, see Fuller (1995, Theorem 5.5.4).

In Example 3.31, the estimated standard error of $\widehat{\theta}$ was .025. In the example, this value was calculated as the square root of

$$s_w^2\left(n^{-1}\sum_{t=2}^{n} z_{t-1}^2(\widehat{\theta})\right)^{-1},$$

where $n = 633$, $s_w^2 = .236$, and $\widehat{\theta} = -.772$. Using (3.122), we could have also calculated this value using the asymptotic approximation, the square root of $(1-.772^2)/633$, which is also .025.

The asymptotic behavior of the parameter estimators gives us an additional insight into the problem of fitting ARMA models to data. For example, suppose a time series follows an AR(1) process and we decide to fit an AR(2) to the data. Does any problem occur in doing this? More generally, why not simply fit large-order AR models to make sure that we capture the dynamics of the process? After all, if the process is truly an AR(1), the other autoregressive parameters will not be significant. The answer is that if we overfit, we will lose efficiency. For example, if we fit an AR(1) to an AR(1) process, for large n, $\mathrm{var}(\widehat{\phi}_1) \approx n^{-1}(1-\phi_1^2)$. But if we fit an AR(2) to the AR(1) process, for large n, $\mathrm{var}(\widehat{\phi}_1) \approx n^{-1}(1-\phi_2^2) = n^{-1}$ because $\phi_2 = 0$. Thus, the variance of ϕ_1 has been inflated, making the estimator less precise. We do want to mention that overfitting can be used as a diagnostic tool. For example, if we fit an AR(2) model to the data and are satisfied with that model, then adding one more parameter and fitting an AR(3) should lead to approximately the same model as in the AR(2) fit. We will discuss model diagnostics in more detail in §3.8.

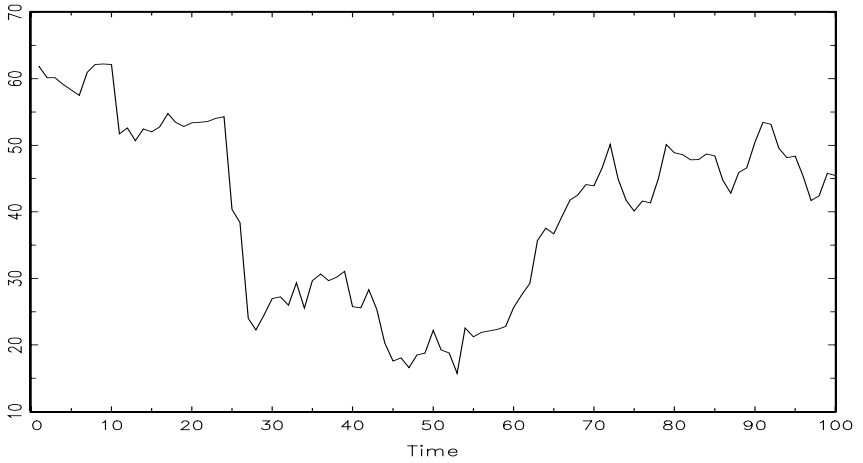


Figure 3.8 One hundred observations generated from the AR(1) model in Example 3.33.

If n is small, or if the parameters are close to the boundaries, the asymptotic approximations can be quite poor. The bootstrap can be helpful in this case; for a broad treatment of the bootstrap, see Efron and Tibshirani (1994). We discuss the case of an AR(1) here and leave the general discussion for Chapter 6. For now, we give a simple example of the bootstrap for an AR(1) process.

Example 3.33 Bootstrapping an AR(1)

We consider an AR(1) model with a regression coefficient near the boundary of causality and an error process that is symmetric but not normal. Specifically, consider the stationary and causal model

$$x_t = \mu + \phi(x_{t-1} - \mu) + w_t, \tag{3.125}$$

where $\mu = 50$, $\phi = .95$, and w_t are iid double exponential with location zero, and scale parameter $\beta = 2$. The density of w_t is given by

$$f_{w_t}(w) = \frac{1}{2\beta} \exp\{-|w|/\beta\} \quad -\infty < w < \infty.$$

In this example, $E(w_t) = 0$ and $\text{var}(w_t) = 2\beta^2 = 8$. Figure 3.8 shows $n = 100$ simulated observations from this process. This particular realization is interesting; the data look like they were generated from a nonstationary process with three different mean levels. In fact, the data were generated from a well-behaved, albeit non-normal, stationary and causal model. To show the advantages of the bootstrap, we will act as if we do not know the actual error distribution and we will proceed as if it were normal; of

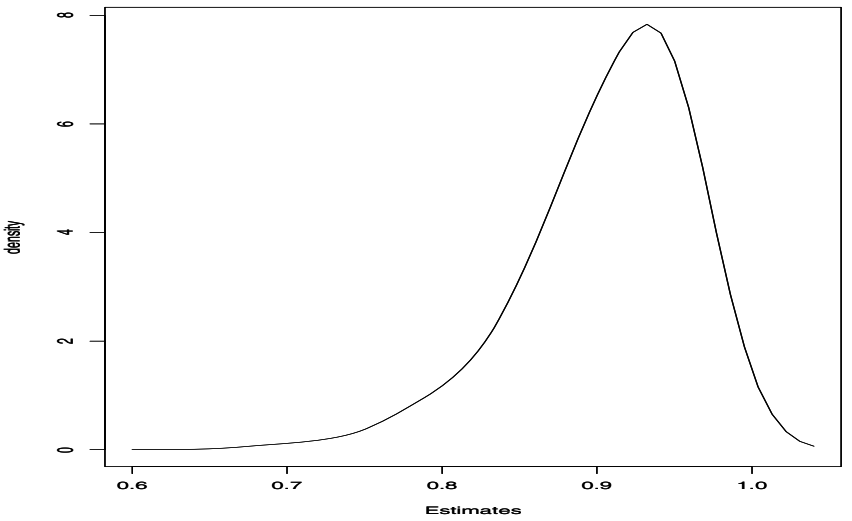


Figure 3.9 Finite sample density of the Yule–Walker estimate of ϕ in Example 3.33.

course, this means, for example, that the normal based MLE of ϕ will not be the actual MLE because the data are not normal.

Using the data shown in Figure 3.8, we obtained the Yule–Walker estimates $\widehat{\mu} = 40.048$, $\widehat{\phi} = .957$, and $s_w^2 = 15.302$, where s_w^2 is the estimate of $\mathrm{var}(w_t)$. Based on Property P3.9, we would say that $\widehat{\phi}$ is approximately normal with mean ϕ (which we supposedly do not know) and variance $(1 - \phi^2)/100$, which we would approximate by $(1 - .957^2)/100 = .029^2$.

To assess the finite sample distribution of $\widehat{\phi}$ when $n = 100$, we simulated 1000 realizations of this AR(1) process and estimated the parameters via Yule–Walker. The finite sampling density of the Yule–Walker estimate of ϕ, based on the 1000 repeated simulations, is shown in Figure 3.9. Clearly the sampling distribution is not close to normality for this sample size. The mean of the distribution shown in Figure 3.9 is .907, and the variance of the distribution is $.052^2$; these values are considerably different than the asymptotic values. Some of the quantiles of the finite sample distribution are .81 (5%), .84 (10%), .88 (25%), .92 (50%), .95 (75%), .96 (90%), and .97 (95%).

Before discussing the bootstrap, we first investigate the sample innovation process, $x_t - x_t^{t-1}$, with corresponding variances P_t^{t-1}. For the AR(1) model in this example,

$$x_t^{t-1} = \mu + \phi(x_{t-1} - \mu), \qquad t = 2, \ldots, 100.$$

From this, it follows that

$$P_t^{t-1} = E(x_t - x_t^{t-1})^2 = \sigma_w^2, \qquad t = 2, \ldots, 100.$$

When $t = 1$, we have

$$x_1^0 = \mu \quad \text{and} \quad P_1^0 = \sigma_w^2/(1-\phi^2).$$

Thus, the innovations have zero mean but different variances; in order that all of the innovations have the same variance, σ_w^2, we will write them as

$$\begin{aligned} \epsilon_1 &= (x_1 - \mu)\sqrt{(1-\phi^2)} \\ \epsilon_t &= (x_t - \mu) - \phi(x_{t-1} - \mu), \quad \text{for} \quad t = 2, \ldots, 100. \end{aligned} \tag{3.126}$$

From these equations, we can write the model in terms of the innovations ϵ_t as

$$\begin{aligned} x_1 &= \mu + \epsilon_1/\sqrt{(1-\phi^2)} \\ x_t &= \mu + \phi(x_{t-1} - \mu) + \epsilon_t \quad \text{for} \quad t = 2, \ldots, 100. \end{aligned} \tag{3.127}$$

Next, replace the parameters with their estimates in (3.126), that is, $n = 100$, $\widehat{\mu} = 40.048$, and $\widehat{\phi} = .957$, and denote the resulting sample innovations as $\{\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_{100}\}$. To obtain one bootstrap sample, first randomly sample, with replacement, $n = 100$ values from the set of sample innovations; call the sampled values $\{\epsilon_1^*, \ldots, \epsilon_{100}^*\}$. Now, generate a bootstrapped data set sequentially by setting

$$\begin{aligned} x_1^* &= 40.048 + \epsilon_1^*/\sqrt{(1-.957^2)} \\ x_t^* &= 40.048 + .957(x_{t-1}^* - 40.048) + \epsilon_t^*, \quad t = 2, \ldots, n. \end{aligned} \tag{3.128}$$

Next, estimate the parameters as if the data were x_t^*. Call these estimates $\widehat{\mu}(1)$, $\widehat{\phi}(1)$, and $s_w^2(1)$. Repeat this process a large number, B, of times, generating a collection of bootstrapped parameter estimates, $\{\widehat{\mu}(b), \widehat{\phi}(b), s_w^2(b), b = 1, \ldots, B\}$. We can then approximate the finite sample distribution of an estimator from the bootstrapped parameter values. For example, we can approximate the distribution of $\widehat{\phi} - \phi$ by the empirical distribution of $\widehat{\phi}(b) - \widehat{\phi}$, for $b = 1, \ldots, B$.

Figure 3.10 shows the bootstrap histogram of 200 bootstrapped estimates of ϕ using the data shown in Figure 3.8. In particular, the mean of the distribution of $\widehat{\phi}(b)$ is .918 with a variance of $.046^2$. Some quantiles of this distribution are .83 (5%), .85 (10%), .90 (25%), .93 (50%), .95 (75%), .97 (90%), and .98 (95%). Clearly, the bootstrap distribution of $\widehat{\phi}$ is closer to the distribution of $\widehat{\phi}$ shown in Figure 3.9 than to the asymptotic (normal) approximation.

To perform a similar bootstrap exercise in R, use the following commands. We note that the R estimation procedure is conditional on the first observation, so the first residual is not returned. To get around this problem, we simply fix the first observation and bootstrap the remaining data. The simulated data are available in the file `ar1boot.dat`.[7]

```
> x = scan("/mydata/ar1boot.dat")
> m = mean(x)              # estimate of mu
> fit = ar.yw(x, order=1)
> phi = fit$ar             # estimate of phi
> nboot = 200              # number of bootstrap replicates
> resids = fit$resid
> resids = resids[2:100]  # the first resid is NA
> x.star = x               # initialize x.star
> phi.star = matrix(0, nboot, 1)
> for (i in 1:nboot) {
+  resid.star = sample(resids)
+   for (t in 1:99){
+   x.star[t+1] = m + phi*(x.star[t]-m) + resid.star[t]
+   }
+  phi.star[i] = ar.yw(x.star, order=1)$ar
+  }
```

Now, 200 bootstrapped estimates are available in `phi.star`, and various methods can be used to evaluate the estimates. For example, to obtain a histogram of the estimates, `hist(phi.star)` can be used . Also consider the statistics `mean(phi.star)`, `sd(phi.star)`, for the mean and standard deviation, and `quantile(phi.star, probs = seq(0, 1, .25))` for some quantiles. Other interesting graphics are `boxplot(phi.star)` for a boxplot and `stem(phi.star)` for a stem-and-leaf diagram.

3.7 Integrated Models for Nonstationary Data

In Chapters 1 and 2, we saw that if x_t is a random walk, $x_t = x_{t-1} + w_t$, then by differencing x_t, we find that $\nabla x_t = w_t$ is stationary. In many situations, time series can be thought of as being composed of two components, a nonstationary trend component and a zero-mean stationary component. For example, in §2.2 we considered the model

$$x_t = \mu_t + y_t, \tag{3.129}$$

[7] If you want to simulate your own data, use the following commands:

```
> e = rexp(150, rate = .5); u = runif(150,-1,1); de = e*sign(u)
> x = 50 + arima.sim(n = 100, list(ar =.95), innov = de, n.start = 50)
```

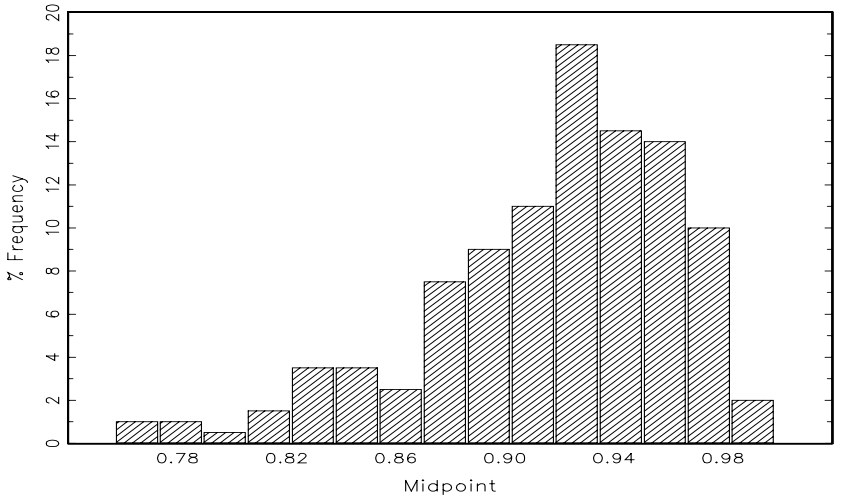


Figure 3.10 Bootstrap histogram of $\widehat{\phi}$ based on 200 bootstraps.

where $\mu_t = \beta_0 + \beta_1 t$ and y_t is stationary. Differencing such a process will lead to a stationary process:

$$\nabla x_t = x_t - x_{t-1} = \beta_1 + y_t - y_{t-1} = \beta_1 + \nabla y_t.$$

Another model that leads to first differencing is the case in which μ_t in (3.129) is stochastic and slowly varying according to a random walk. That is, in (3.129)

$$\mu_t = \mu_{t-1} + v_t$$

where v_t is stationary. In this case,

$$\nabla x_t = v_t + \nabla y_t,$$

is stationary. If μ_t in (3.129) is a k-th order polynomial, $\mu_t = \sum_{j=0}^{k} \beta_j t^j$, then (Problem 3.26) the differenced series $\nabla^k y_t$ is stationary. Stochastic trend models can also lead to higher order differencing. For example, suppose in (3.129)

$$\mu_t = \mu_{t-1} + v_t \quad \text{and} \quad v_t = v_{t-1} + e_t,$$

where e_t is stationary. Then, $\nabla x_t = v_t + \nabla y_t$ is not stationary, but

$$\nabla^2 x_t = e_t + \nabla^2 y_t$$

is stationary.

The integrated ARMA, or ARIMA model, is a broadening of the class of ARMA models to include differencing.

Definition 3.11 *A process, x_t is said to be* **ARIMA(p,d,q)** *if*

$$\nabla^d x_t = (1-B)^d x_t$$

is ARMA(p,q). In general, we will write the model as

$$\phi(B)(1-B)^d x_t = \theta(B) w_t. \tag{3.130}$$

If $E(\nabla^d x_t) = \mu$, we write the model as

$$\phi(B)(1-B)^d x_t = \alpha + \theta(B) w_t,$$

where $\alpha = \mu(1-\phi_1 - \cdots - \phi_p)$.

Example 3.34 IMA$(1,1)$ and EWMA

The ARIMA(0,1,1), or IMA(1,1) model is of interest because many economic time series can be successfully modeled this way. In addition, the model leads to a frequently used, and abused, forecasting method called exponentially weighted moving averages (EWMA). We will write the model as

$$x_t = x_{t-1} + w_t - \lambda w_{t-1} \tag{3.131}$$

because this model formulation is easier to work with here, and it leads to the standard representation for EWMA. When $|\lambda| < 1$, the model has an invertible representation,

$$x_t = \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} x_{t-j} + w_t. \tag{3.132}$$

Verification of (3.132) is left to the reader (Problem 3.27). From (3.132), we have that the one-step-ahead prediction, using the notation of §3.5, is

$$\begin{aligned} \tilde{x}_{n+1} &= \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} x_{n+1-j} \\ &= (1-\lambda)x_n + \lambda \sum_{j=1}^{\infty} (1-\lambda)\lambda^{j-1} x_{n-j} \\ &= (1-\lambda)x_n + \lambda \tilde{x}_n. \end{aligned} \tag{3.133}$$

Based on (3.133), the truncated forecasts are obtained by setting $\tilde{x}_1^0 = 0$, and then updating as follows:

$$\tilde{x}_{n+1}^n = (1-\lambda)x_n + \lambda \tilde{x}_n^{n-1}, \quad n \geq 1. \tag{3.134}$$

From (3.134), we see that the new forecast is a linear combination of the old forecast and the new observation. In EWMA, the parameter λ is called the smoothing constant and is restricted to be between zero and one. Larger values of λ lead to smoother forecasts. This method of forecasting is popular because it is easy to use; we need only retain the previous forecast value and the current observation to forecast the next time period. Unfortunately, as previously suggested, the method is often abused because some forecasters do not verify that the observations follow an IMA$(1, 1)$ process, and often arbitrarily pick values of λ.

Finally, the model for the glacial varve series in Example 3.31 is an IMA$(1, 1)$ on the logarithms of the data. Recall that the fitted model there was $\ln x_t = \ln x_{t-1} + w_t - .772w_{t-1}$ and $\mathrm{var}(w_t) = .236$.

3.8 Building ARIMA Models

There are a few basic steps to fitting ARIMA models to time series data. These steps involve plotting the data, possibly transforming the data, identifying the dependence orders of the model, parameter estimation, diagnostics, and model choice. First, as with any data analysis, we should construct a time plot of the data, and inspect the graph for any anomalies. If, for example, the variability in the data grows with time, it will be necessary to transform the data to stabilize the variance. In such cases, the Box–Cox class of power transformations, equation (2.34), could be employed. Also, the particular application might suggest an appropriate transformation. For example, suppose a process evolves as a fairly small and stable percent change, such as an investment. For example, we might have

$$x_t = (1 + p_t)x_{t-1},$$

where x_t is the value of the investment at time t and p_t is the percentage change from period $t - 1$ to t, which may be negative. Taking logs we have

$$\ln(x_t) = \ln(1 + p_t) + \ln(x_{t-1}),$$

or

$$\nabla[\ln(x_t)] = \ln(1 + p_t).$$

If the percent change p_t stays relatively small in magnitude, then $\ln(1+p_t) \approx p_t$ and, thus,

$$\nabla[\ln(x_t)] \approx p_t,$$

will be a relatively stable process. Frequently, $\nabla[\ln(x_t)]$ is called the return or growth rate. This general idea was used in Example 3.31, and we will use it again in Example 3.35.

After suitably transforming the data, the next step is to identify preliminary values of the autoregressive order, p, the order of differencing, d, and the

Figure 3.11 Quarterly U.S. GNP from 1947(1) to 2002(3).

moving average order, q. We have already addressed, in part, the problem of selecting d. A time plot of the data will typically suggest whether any differencing is needed. If differencing is called for, then difference the data once, $d = 1$, and inspect the time plot of ∇x_t. If additional differencing is necessary, then try differencing again and inspect a time plot of $\nabla^2 x_t$. Be careful not to overdifference because this may introduce dependence where none exists. For example, $x_t = w_t$ is serially uncorrelated, but $\nabla x_t = w_t - w_{t-1}$ is MA(1). In addition to time plots, the sample ACF can help in indicating whether differencing is needed. Because the polynomial $\phi(z)(1-z)^d$ has a unit root, the sample ACF, $\widehat{\rho}(h)$, will not decay to zero fast as h increases. Thus, a slow decay in $\widehat{\rho}(h)$ is an indication that differencing may be needed.

When preliminary values of d have been settled, the next step is to look at the sample ACF and PACF of $\nabla^d x_t$ for whatever values of d have been chosen. Using Table 3.1 as a guide, preliminary values of p and q are chosen. Recall that, if $p = 0$ and $q > 0$, the ACF cuts off after lag q, and the PACF tails off. If $q = 0$ and $p > 0$, the PACF cuts off after lag p, and the ACF tails off. If $p > 0$ and $q > 0$, both the ACF and PACF will tail off. Because we are dealing with estimates, it will not always be clear whether the sample ACF or PACF is tailing off or cutting off. Also, two models that are seemingly different can actually be very similar. With this in mind, we should not worry about being so precise at this stage of the model fitting. At this stage, a few preliminary values of p, d, and q should be at hand, and we can start estimating the parameters.

Example 3.35 Analysis of GNP Data

In this example, we consider the analysis of quarterly U.S. GNP from 1947(1) to 2002(3), $n = 223$ observations. The data are Real U.S. Gross National Product in billions of chained 1996 dollars and they have been seasonally adjusted. The data were obtained from the Federal Reserve

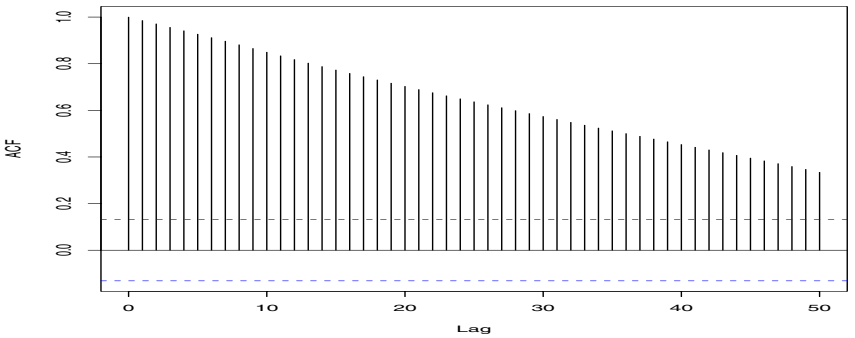


Figure 3.12 Sample ACF of the GNP data.

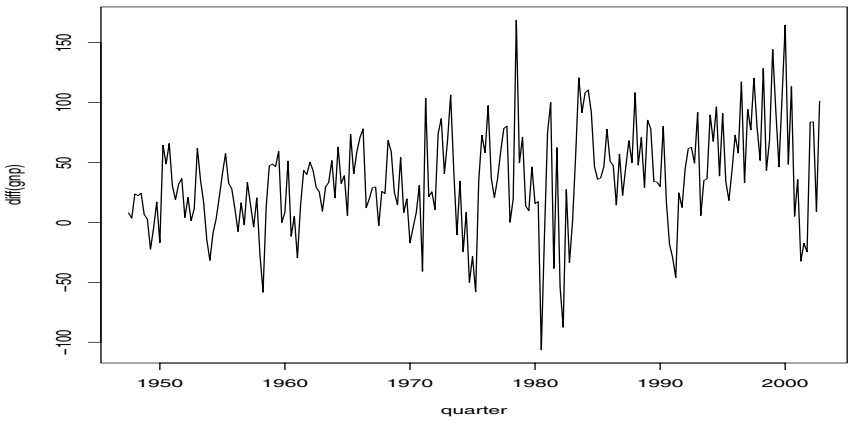


Figure 3.13 First difference of the U.S. GNP data.

Bank of St. Louis (http://research.stlouisfed.org/). Figure 3.11 shows a plot of the data, say, y_t. Because strong trend hides any other effect, it is not clear from Figure 3.11 that the variance is increasing with time. For the purpose of demonstration, the sample ACF of the data is displayed in Figure 3.12. Figure 3.13 shows the first difference of the data, ∇y_t, and now that the trend has been removed we are able to notice that the variability in the second half of the data is larger than in the first half of the data. Also, it appears as though a trend is still present after differencing. The growth rate, say, $x_t = \nabla \ln(y_t)$, is plotted in Figure 3.14, and, appears to be a stable process. Moreover, we may interpret the values of x_t as the percentage quarterly growth of U.S. GNP.

The sample ACF and PACF of the quarterly growth rate are plotted in Figure 3.15. Inspecting the sample ACF and PACF, we might feel that the ACF is cutting off at lag 2 and the PACF is tailing off. This

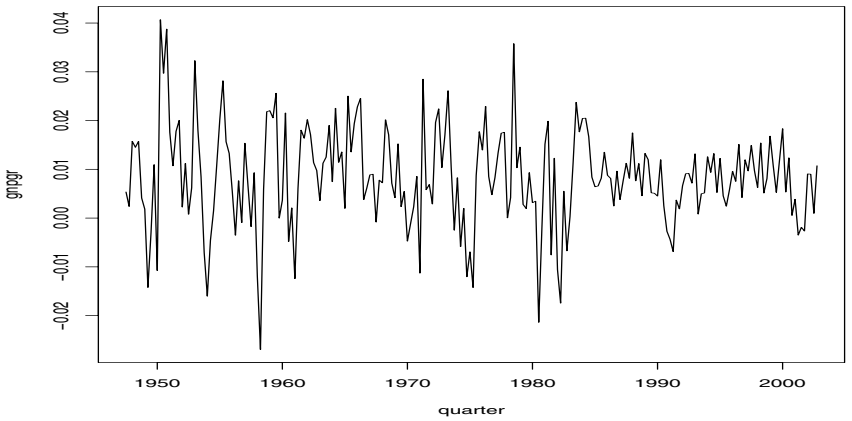


Figure 3.14 U.S. GNP quarterly growth rate.

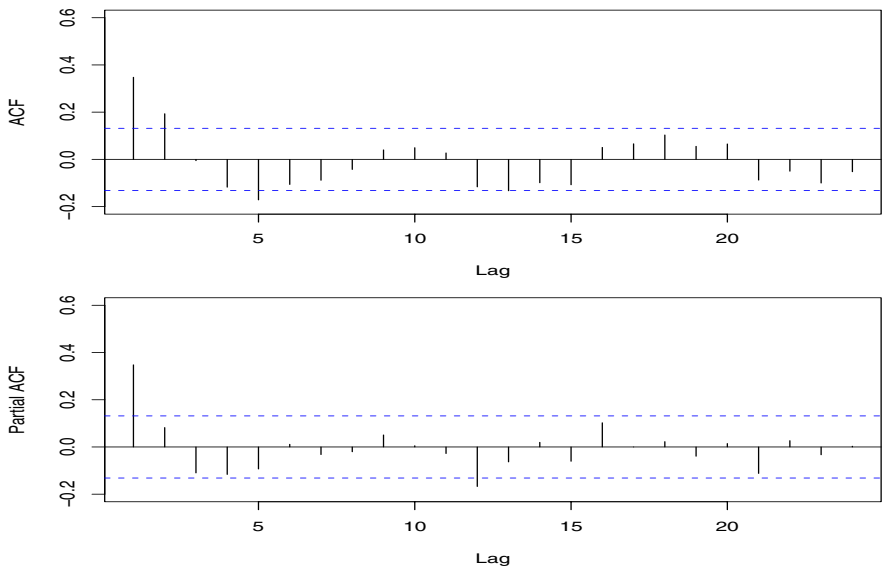


Figure 3.15 Sample ACF and PACF of the GNP quarterly growth rate.

would suggest the GNP growth rate follows an MA(2) process, or log GNP follows an ARIMA$(0, 1, 2)$ model. Rather than focus on one model, we will also suggest that it appears that the ACF is tailing off and the PACF is cutting off at lag 1. This suggests an AR(1) model for the growth rate, or ARIMA$(1, 1, 0)$ for log GNP. As a preliminary analysis, we will fit both models.

Using MLE to fit the MA(2) model for the growth rate, x_t, the estimated

model is

$$x_t = .008_{(.001)} + .303_{(.065)}\widehat{w}_{t-1} + .204_{(.064)}\widehat{w}_{t-2} + \widehat{w}_t, \tag{3.135}$$

where $\widehat{\sigma}_w = .0094$ is based on 219 degrees of freedom. The values in parentheses are the corresponding estimated standard errors. All of the regression coefficients are significant, including the constant. *We make a special note of this because, as a default, some computer packages do not fit a constant in a differenced model.* That is, these packages assume, by default, that there is no drift. In this example, not including a constant leads to the wrong conclusions about the nature of the U.S. economy. Not including a constant assumes the average quarterly growth rate is zero, whereas the U.S. GNP average quarterly growth rate is about 1% (which can be seen easily in Figure 3.14). We leave it to the reader to investigate what happens when the constant is not included.

The estimated AR(1) model is

$$x_t = .005_{(.0006)} + .347_{(.063)}x_{t-1} + \widehat{w}_t, \tag{3.136}$$

where $\widehat{\sigma}_w = .0095$ on 220 degrees of freedom.

We will discuss diagnostics next, but assuming both of these models fit well, how are we to reconcile the apparent differences of the estimated models (3.135) and (3.136)? In fact, the fitted models are nearly the same. To show this, consider an AR(1) model of the form in (3.136) without a constant term; that is,

$$x_t = .35x_{t-1} + w_t,$$

and write it in its causal form, $x_t = \sum_{j=0}^{\infty}\psi_j w_{t-j}$, where we recall $\psi_j = .35^j$. Thus, $\psi_0 = 1, \psi_1 = .350, \psi_2 = .123, \psi_3 = .043, \psi_4 = .015, \psi_5 = .005, \psi_6 = .002, \psi_7 = .001, \psi_8 = 0, \psi_9 = 0, \psi_{10} = 0$, and so forth. Thus,

$$x_t \approx .35w_{t-1} + .12w_{t-2} + w_t,$$

which is similar to the fitted MA(2) model in (3.136).

The analyses and graphics of the example can be performed in R using the following commands. We note that we did not fit integrated models to log GNP, but rather we fit nonintegrated models to the growth rate, x_t. We believe at the time of writing that there is a problem with fitting ARIMA models with a nonzero constant in R. The data are in a file called `gnp96.dat`; the file contains two columns, the first column is the quarter and the second column is the GNP.

```
> gnp96 = read.table("/mydata/gnp96.dat")
> gnp = ts(gnp96[,2], start=1947, frequency=4)
> plot(gnp)
```

```
> acf(gnp, 50)
> gnpgr = diff(log(gnp))  # growth rate
> plot.ts(gnpgr)
> par(mfrow=c(2,1))
> acf(gnpgr, 24)
> pacf(gnpgr, 24)
> # ARIMA fits:
> gnpgr.ar = arima(gnpgr, order = c(1, 0, 0))
> gnpgr.ma = arima(gnpgr, order = c(0, 0, 2))
> # to view the results:
> gnpgr.ar    # potential problem here (see below *)
> gnpgr.ma
> ARMAtoMA(ar=.35, ma=0, 10)  # prints psi-weights
```

*At this time, the R output for the AR fit lists the estimated mean and its standard error, but calls it the intercept. That is, the output says it is giving you $\widehat{\alpha}$ when in fact it's listing $\widehat{\mu}$. In this case, $\widehat{\alpha} = \widehat{\mu}(1 - \widehat{\phi})$.

The next step in model fitting is diagnostics. This investigation includes the analysis of the residuals as well as model comparisons. Again, the first step involves a time plot of the innovations (or residuals), $x_t - \widehat{x}_t^{t-1}$, or of the standardized innovations

$$e_t = \left(x_t - \widehat{x}_t^{t-1}\right) \Big/ \sqrt{\widehat{P}_t^{t-1}}, \tag{3.137}$$

where $\widehat{x}_t^{t-1}$ is the one-step-ahead prediction of x_t based on the fitted model and $\widehat{P}_t^{t-1}$ is the estimated one-step-ahead error variance. If the model fits well, the standardized residuals should behave as an iid sequence with mean zero and variance one. The time plot should be inspected for any obvious departures from this assumption. Unless the time series is Gaussian, it is not enough that the residuals are uncorrelated. For example, it is possible in the non-Gaussian case to have an uncorrelated process for which values contiguous in time are highly dependent. As an example, we mention the family of GARCH models that are discussed in Chapter 5.

Investigation of marginal normality can be accomplished visually by looking at a histogram of the residuals. In addition to this, a normal probability plot or a Q-Q plot can help in identifying departures from normality. See Johnson and Wichern (1992, Chapter 4) for details of this test as well as additional tests for multivariate normality.

There are several tests of randomness, for example the runs test, that could be applied to the residuals. We could also inspect the sample autocorrelations of the residuals, say, $\widehat{\rho}_e(h)$, for any patterns or large values. Recall that, for a white noise sequence, the sample autocorrelations are approximately independently and normally distributed with zero means and variances $1/n$. Hence, a good check on the correlation structure of the residuals is to plot $\widehat{\rho}_e(h)$ versus h along with the error bounds of $\pm 2/\sqrt{n}$. The residuals from a model fit,

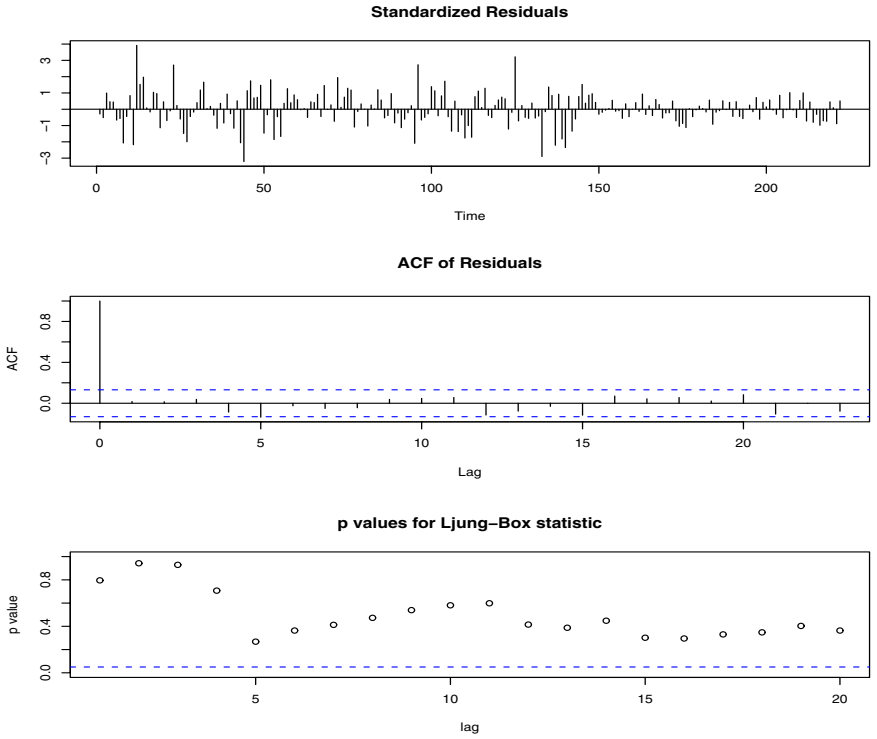


Figure 3.16 Diagnostics of the residuals from MA(2) fit on GNP growth rate.

however, will not quite have the properties of a white noise sequence and the variance of $\widehat{\rho}_e(h)$ can be much less than $1/n$. Details can be found in Box and Pierce (1970) and McLeod (1978). This part of the diagnostics can be viewed as a visual inspection of $\widehat{\rho}_e(h)$ with the main concern being the detection of obvious departures from the independence assumption.

In addition to plotting $\widehat{\rho}_e(h)$, we can perform a general test that takes into consideration the magnitudes of $\widehat{\rho}_e(h)$ as a group. For example, it may be the case that, individually, each $\widehat{\rho}_e(h)$ is small in magnitude, say, each one is just slightly less that $2/\sqrt{n}$ in magnitude, but, collectively, the values are large. The Ljung–Box–Pierce Q-statistic given by

$$Q = n(n+2)\sum_{h=1}^{H} \frac{\widehat{\rho}_e^2(h)}{n-h} \tag{3.138}$$

can be used to perform such a test. The value H in (3.138) is chosen somewhat arbitrarily, typically, $H = 20$. Under the null hypothesis of model adequacy, asymptotically ($n \to \infty$), $Q \sim \chi^2_{H-p-q}$. Thus, we would reject the null hypothesis at level α if the value of Q exceeds the $(1-\alpha)$-quantile of the χ^2_{H-p-q}

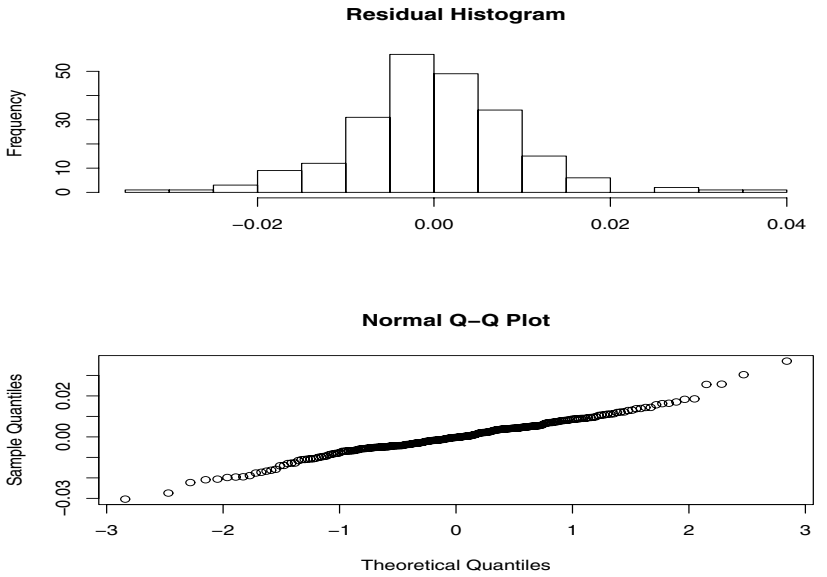


Figure 3.17 Histogram of the residuals (top), and a normal Q-Q plot of the residuals (bottom).

distribution. Details can be found in Box and Pierce (1970), Ljung and Box (1978), and Davies et al. (1977).

Example 3.36 Diagnostics for GNP Growth Rate Example

We will focus on the MA(2) fit from Example 3.35; the analysis of the AR(1) residuals is similar. Figure 3.16 displays a plot of the standardized residuals, the ACF of the residuals (note that R includes the correlation at lag zero which is always one), and the value of the Q-statistic, (3.138), at lags $H = 1$ through $H = 20$. These diagnostics are provided by issuing the command

```
> tsdiag(gnpgr.ma, gof.lag=20)
```

where `gnpgr.ma` was described in the previous example.

Inspection of the time plot of the standardized residuals in Figure 3.16 shows no obvious patterns. Notice that there are outliers, however, with a few values exceeding 3 standard deviations in magnitude. The ACF of the standardized residuals shows no apparent departure from the model assumptions, and the Q-statistic is never significant at the lags shown.

Finally, Figure 3.17 shows a histogram of the residuals (top), and a normal Q-Q plot of the residuals (bottom). Here we see the residuals are somewhat close to normality except for a few extreme values in the tails.

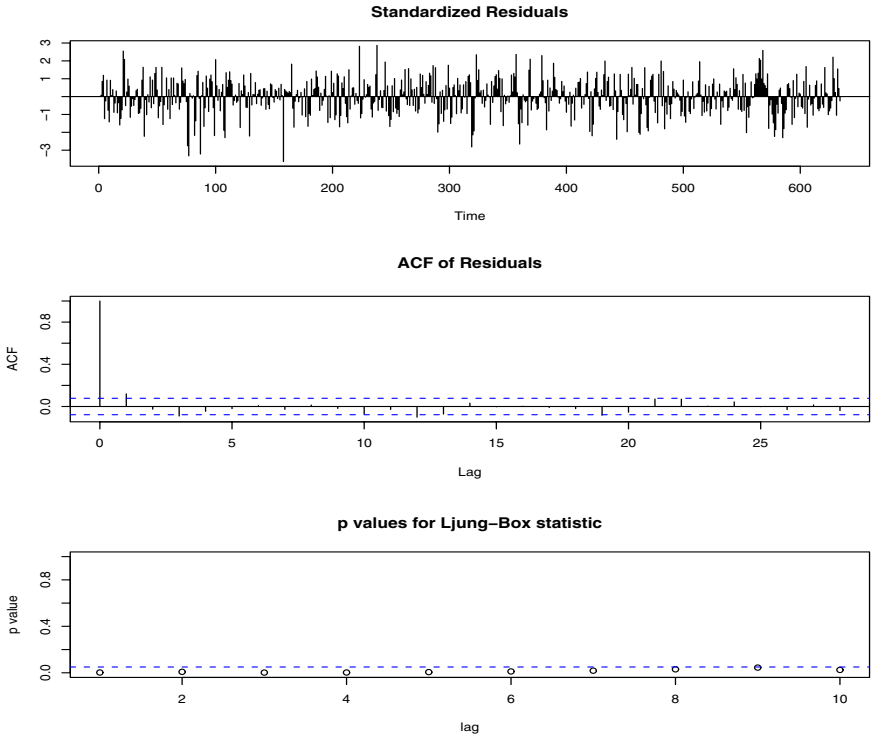


Figure 3.18 Diagnostics for the ARIMA$(0, 1, 1)$ fit to the logged varve data.

Running a Shapiro–Wilk test (Royston, 1982) yields a p-value of .003, which indicates the residuals are not normal. Hence, the model appears to fit well except for the fact that a distribution with heavier tails than the normal distribution should be employed. We discuss some possibilities in Chapters 5 and 6. These diagnostics can be performed in R by issuing the commands:

```
> hist(gnpgr.ma$resid, br=12)
> qqnorm(gnpgr.ma$resid)
> shapiro.test(gnpgr.ma$resid)
```

Example 3.37 Diagnostics for the Glacial Varve Series

In Example 3.31, we fit an ARIMA$(0, 1, 1)$ model to the logarithms of the glacial varve data. Figure 3.18 shows the diagnostics from that fit, and we notice a significant lag 1 correlation. In addition, the Q-statistic is significant for every value of H displayed. Because the ACF of the residuals appear to be tailing off, an AR term is suggested.

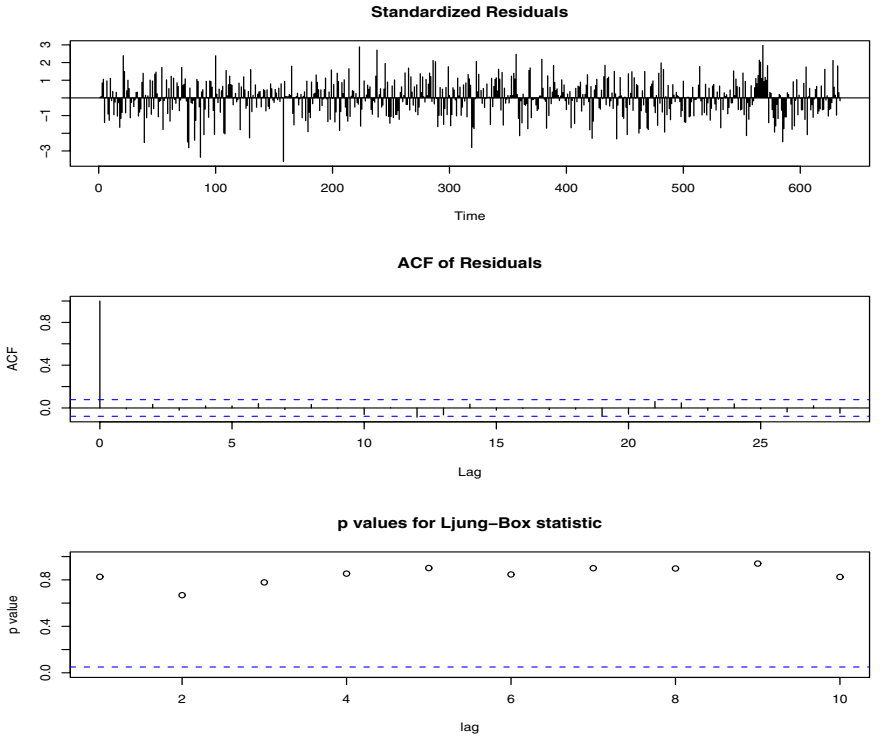


Figure 3.19 Diagnostics for the ARIMA$(1, 1, 1)$ fit to the logged varve data.

Next, we fit an ARIMA$(1, 1, 1)$ to the logged varve data and obtained the estimates $\widehat{\phi} = .23_{(.05)}$, $\widehat{\theta} = -.89_{(.03)}$, and $\widehat{\sigma}_w^2 = .23$. Hence the AR term is significant. Diagnostics for this model are displayed in Figure 3.19, and it appears this model fits the data well.

To implement these analyses in R, use the following commands (we assume the data are in `varve`):

```
> varve.ma = arima(log(varve), order = c(0, 1, 1))
> varve.ma     # to display results
> tsdiag(varve.ma)
> varve.arma = arima(log(varve), order = c(1, 1, 1))
> varve.arma    # to display results
> tsdiag(varve.arma, gof.lag=20)
```

In Example 3.35, we have two competing models, an AR(1) and an MA(2) on the GNP growth rate, that each appear to fit the data well. In addition, we might also consider that an AR(2) or an MA(3) might do better for forecasting. Perhaps combining both models, that is, fitting an ARMA$(1, 2)$ to the GNP

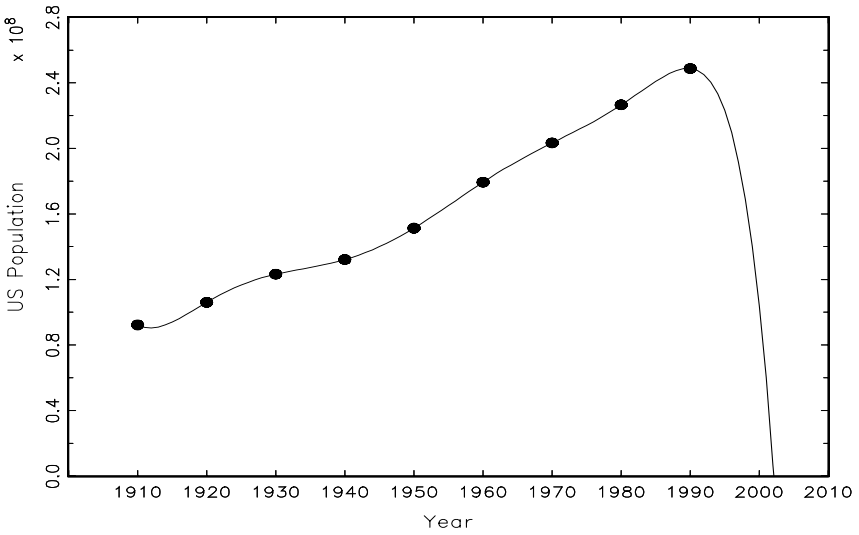


Figure 3.20 A perfect fit and a terrible forecast.

growth rate, would be the best. As previously mentioned, we have to be concerned with overfitting the model; it is not always the case that more is better. Overfitting leads to less-precise estimators, and adding more parameters may fit the data better but may also lead to bad forecasts. This result is illustrated in the following example.

Example 3.38 A Problem with Overfitting

Figure 3.20 shows the U.S. population by official census, every 10 years from 1910 to 1990, as points. If we use these nine observations to predict the future population of the U.S., we can use an eight-degree polynomial so the fit to the nine observations is perfect. The model in this case is

$$x_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_8 t^8 + w_t.$$

The fitted model, which is plotted through the year 2010 as a line, passes through the nine observations. The model predicts that the population of the U.S. will be close to zero in the year 2000, and will cross zero sometime in the year 2002!

The final step of model fitting is model choice or model selection. That is, we must decide which model we will retain for forecasting. The most popular techniques, AIC, AICc, and SIC, were described in §2.2 in the context of regression models. A discussion of AIC based on Kullback–Leibler distance was given in Problems 2.4 and 2.5.

Example 3.39 Model Choice for the U.S. GNP Series

Returning to the analysis of the U.S. GNP data presented in Examples 3.35 and 3.36, recall that two models, an AR(1) and an MA(2), fit the GNP growth rate well. To choose the final model, we compare the AIC, the AICc, and the SIC for both models.

Below are the R commands for the comparison. The effective sample size in this example is 222. We note that R returns AIC[8] as part of the ARIMA fit.

```
> # AIC
> gnpgr.ma$aic
   [1] -1431.929    # MA(2)
> gnpgr.ar$aic
   [1] -1431.221    # AR(1)
> # AICc - see Section 2.2
> log(gnpgr.ma$sigma2)+(222+2)/(222-2-2)
   [1] -8.297199    # MA(2)
> log(gnpgr.ar$sigma2)+(222+1)/(222-1-2)
   [1] -8.294156    # AR(1)
> # SIC or BIC - see Section 2.2
> log(gnpgr.ma$sigma2)+(2*log(222)/222)
   [1] -9.276049    # MA(2)
> log(gnpgr.ar$sigma2)+(1*log(222)/222)
   [1] -9.288084    # AR(1)
```

The AIC and AICc both prefer the MA(2) fit, whereas the SIC (or BIC) prefers the simpler AR(1) model. It is often the case that the SIC will select a model of smaller order than the AIC or AICc. It would not be unreasonable in this case to retain the AR(1) because pure autoregressive models are easier to work with.

3.9 Multiplicative Seasonal ARIMA Models

In this section, we introduce several modifications made to the ARIMA model to account for seasonal and nonstationary behavior. Often, the dependence on the past tends to occur most strongly at multiples of some underlying seasonal lag s. For example, with monthly economic data, there is a strong yearly component occurring at lags that are multiples of $s = 12$, because

[8] R calculates this value as AIC $= -2 \ln L_x(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}_w^2) + 2(p+q)$, where $L_x(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}_w^2)$ is the likelihood of the data evaluated at the MLE; see (3.105). Note that AIC consists of two parts, one measuring model fit and one penalizing for the addition of parameters. Dividing this quantity by n, writing $k = p + q$, and ignoring constants and terms involving initial conditions, we obtain AIC as given in §2.2. Details are provided in Problems 2.4 and 2.5.

of the strong connections of all activity to the calendar year. Data taken quarterly will exhibit the yearly repetitive period at $s = 4$ quarters. Natural phenomena such as temperature also have strong components corresponding to seasons. Hence, the natural variability of many physical, biological, and economic processes tends to match with seasonal fluctuations. Because of this, it is appropriate to introduce autoregressive and moving average polynomials that identify with the seasonal lags. The resulting pure seasonal autoregressive moving average model, say, ARMA$(P, Q)_s$, then takes the form

$$\Phi_P(B^s)x_t = \Theta_Q(B^s)w_t, \tag{3.139}$$

with the following definition.

Definition 3.12 *The operators*

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \tag{3.140}$$

and

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs} \tag{3.141}$$

are the **seasonal autoregressive operator** *and the* **seasonal moving average operator** *of orders P and Q, respectively, with seasonal period s.*

Analogous to the properties of nonseasonal ARMA models, the pure seasonal ARMA$(P, Q)_s$ is causal only when the roots of $\Phi_P(z^s)$ lie outside the unit circle, and it is invertible only when the roots of $\Theta_Q(z^s)$ lie outside the unit circle.

Example 3.40 A Seasonal ARMA Series

A first-order seasonal autoregressive moving average series that might run over months could be written as

$$(1 - \Phi B^{12})x_t = (1 + \Theta B^{12})w_t$$

or

$$x_t = \Phi x_{t-12} + w_t + \Theta w_{t-12}.$$

This model exhibits the series x_t in terms of past lags at the multiple of the yearly seasonal period $s = 12$ months. It is clear from the above form that estimation and forecasting for such a process involves only straightforward modifications of the unit lag case already treated. In particular, the causal condition requires $|\Phi| < 1$, and the invertible condition requires $|\Theta| < 1$.

For the first-order seasonal ($s = 12$) MA model, $x_t = w_t + \Theta w_{t-12}$, it is easy to verify that

$$\gamma(0) \quad = \quad (1 + \Theta^2)\sigma^2$$

Table 3.2 Behavior of the ACF and PACF for Causal and Invertible Pure Seasonal ARMA Models

	AR$(P)_s$	MA$(Q)_s$	ARMA$(P,Q)_s$
ACF*	Tails off at lags ks, $k = 1, 2, \ldots,$	Cuts off after lag Qs	Tails off at lags ks
PACF*	Cuts off after lag Ps	Tails off at lags ks $k = 1, 2, \ldots,$	Tails off at lags ks

*The values at nonseasonal lags $h \neq ks$, for $k = 1, 2, \ldots$, are zero.

$$\begin{aligned} \gamma(\pm 12) &= \Theta\sigma^2 \\ \gamma(h) &= 0, \quad \text{otherwise.} \end{aligned}$$

Thus, the only nonzero correlation, aside from lag zero, is

$$\rho(\pm 12) = \Theta/(1+\Theta^2).$$

For the first-order seasonal ($s = 12$) AR model, using the techniques of the nonseasonal AR(1), we have

$$\begin{aligned} \gamma(0) &= \sigma^2/(1-\Phi^2) \\ \gamma(\pm 12k) &= \sigma^2\Phi^k/(1-\Phi^2) \quad k = 1, 2, \ldots \\ \gamma(h) &= 0, \quad \text{otherwise.} \end{aligned}$$

In this case, the only non-zero correlations are

$$\rho(\pm 12k) = \Phi^k, \quad k = 0, 1, 2, \ldots \ .$$

These results can be verified using the general result that $\gamma(h) = \Phi\gamma(h-12)$, for $h \geq 1$. For example, when $h = 1$, $\gamma(1) = \Phi\gamma(11)$, but when $h = 11$, we have $\gamma(11) = \Phi\gamma(1)$, which implies that $\gamma(1) = \gamma(11) = 0$. In addition to these results, the PACF have the analogous extensions from nonseasonal to seasonal models.

As an initial diagnostic criterion, we can use the properties for the pure seasonal autoregressive and moving average series listed in Table 3.2. These properties may be considered as generalizations of the properties for nonseasonal models that were presented in Table 3.1.

In general, we can combine the seasonal and nonseasonal operators into a multiplicative seasonal autoregressive moving average model, denoted by ARMA$(p,q) \times (P,Q)_s$, and write

$$\Phi_P(B^s)\phi(B)x_t = \Theta_Q(B^s)\theta(B)w_t \tag{3.142}$$

as the overall model. Although the diagnostic properties in Table 3.2 are not strictly true for the overall mixed model, the behavior of the ACF and PACF tends to show rough patterns of the indicated form. In fact, for mixed models, we tend to see a mixture of the facts listed in Tables 3.1 and 3.2. In fitting such models, focusing on the seasonal autoregressive and moving average components first generally leads to more satisfactory results.

Example 3.41 A Mixed Seasonal Model

Consider an ARMA$(0,1) \times (1,0)_{12}$ model

$$x_t = \Phi x_{t-12} + w_t + \theta w_{t-1},$$

where $|\Phi| < 1$ and $|\theta| < 1$. Then, because x_{t-12}, w_t, and w_{t-1} are uncorrelated, and x_t is stationary, $\gamma(0) = \Phi^2\gamma(0) + \sigma_w^2 + \theta^2\sigma_w^2$, or

$$\gamma(0) = \frac{1+\theta^2}{1-\Phi^2}\ \sigma_w^2.$$

In addition, multiplying the model by x_{t-h}, $h > 0$, and taking expectations, we have $\gamma(1) = \Phi\gamma(11) + \theta\sigma_w^2$, and $\gamma(h) = \Phi\gamma(h-12)$, for $h \geq 2$. Thus, the ACF for this model is

$$\begin{aligned} \rho(12h) &= \Phi^h \quad h = 1, 2, \ldots \\ \rho(12h-1) &= \rho(12h+1) = \frac{\theta}{1+\theta^2}\Phi^h \quad h = 0, 1, 2, \ldots, \\ \rho(h) &= 0, \quad \text{otherwise.} \end{aligned}$$

The ACF and PACF for this model, with $\Phi = .8$ and $\theta = -.5$, are shown in Figure 3.21. These type of correlation relationships, although idealized here, are typically seen with seasonal data.

To reproduce Figure 3.21 in R, use the following commands:

```
> phi = c(rep(0,11),.8)
> acf = ARMAacf(ar=phi, ma=-.5, 50)
> pacf = ARMAacf(ar=phi, ma=-.5, 50, pacf=T)
> par(mfrow=c(1,2))
> plot(acf, type="h", xlab="lag")
> abline(h=0)
> plot(pacf, type="h", xlab="lag")
> abline(h=0)
```

Seasonal nonstationarity can occur, for example, when the process is nearly periodic in the season. For example, with average monthly temperatures over the years, each January would be approximately the same, each February would be approximately the same, and so on. In this case, we might think of average monthly temperature x_t as being modeled as

$$x_t = S_t + w_t,$$

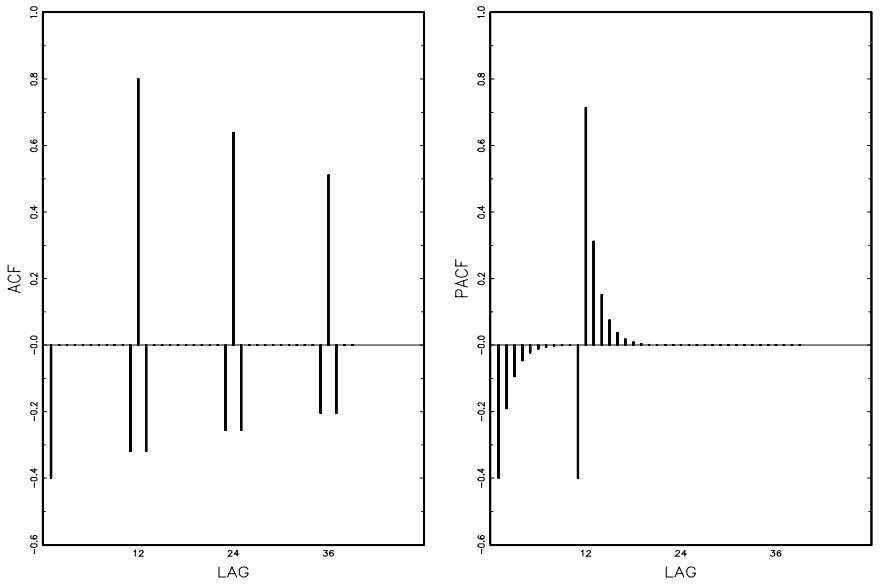


Figure 3.21 ACF and PACF of the mixed seasonal ARMA model $x_t = .8x_{t-12} + w_t - .5w_{t-1}$.

where S_t is a seasonal component that varies slowly from one year to the next, according to a random walk,

$$S_t = S_{t-12} + v_t.$$

In this model, w_t and v_t are uncorrelated white noise processes. The tendency of data to follow this type of model will be exhibited in a sample ACF that is large and decays very slowly at lags $h = 12k$, for $k = 1, 2, \ldots$. If we subtract the effect of successive years from each other, we find that

$$(1 - B^{12})x_t = x_t - x_{t-12} = v_t + w_t - w_{t-12}.$$

This model is a stationary $\text{MA}(1)_{12}$, and its ACF will have a peak only at lag 12. In general, seasonal differencing can be indicated when the ACF decays slowly at multiples of some season s, but is negligible between the periods. Then, a seasonal difference of order D is defined as

$$\nabla_s^D x_t = (1 - B^s)^D x_t, \tag{3.143}$$

where $D = 1, 2, \ldots$ takes integer values. Typically, $D = 1$ is sufficient to obtain seasonal stationarity.

Incorporating these ideas into a general model leads to the following definition.

Definition 3.13 *The multiplicative* **seasonal autoregressive integrated moving average** *model, or* **SARIMA** *model, of Box and Jenkins (1970) is given by*

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \alpha + \Theta_Q(B^s)\theta(B)w_t, \tag{3.144}$$

where w_t is the usual Gaussian white noise process. The general model is denoted as $\mathbf{ARIMA}(\boldsymbol{p,d,q}) \times (\boldsymbol{P,D,Q})_{\boldsymbol{s}}$. *The ordinary autoregressive and moving average components are represented by polynomials $\phi(B)$ and $\theta(B)$ of orders p and q, respectively [see (3.5) and (3.17)], and the seasonal autoregressive and moving average components by $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ [see (3.140) and (3.141)] of orders P and Q and ordinary and seasonal difference components by $\nabla^d = (1-B)^d$ and $\nabla_s^D = (1-B^s)^D$.*

Example 3.42 A SARIMA Model

Consider the following model, which often provides a reasonable representation for seasonal, nonstationary, economic time series. We exhibit the equations for the model, denoted by ARIMA$(0,1,1) \times (0,1,1)_{12}$ in the notation given above, where the seasonal fluctuations occur every 12 months. Then, the model (3.144) becomes

$$(1-B^{12})(1-B)x_t = (1+\Theta B^{12})(1+\theta B)w_t. \tag{3.145}$$

Expanding both sides of (3.145) leads to the representation

$$(1-B-B^{12}+B^{13})x_t = (1+\theta B+\Theta B^{12}+\Theta\theta B^{13})w_t,$$

or in difference equation form

$$x_t = x_{t-1} + x_{t-12} - x_{t-13} + w_t + \theta w_{t-1} + \Theta w_{t-12} + \Theta\theta w_{t-13}.$$

Selecting the appropriate model for a given set of data from all of those represented by the general form (3.144) is a daunting task, and we usually think first in terms of finding difference operators that produce a roughly stationary series and then in terms of finding a set of simple autoregressive moving average or multiplicative seasonal ARMA to fit the resulting residual series. Differencing operations are applied first, and then the residuals are constructed from a series of reduced length. Next, the ACF and the PACF of these residuals are evaluated. Peaks that appear in these functions can often be eliminated by fitting an autoregressive or moving average component in accordance with the general properties of Tables 3.1 and 3.2. In considering whether the model is satisfactory, the diagnostic techniques discussed in §3.8 still apply.

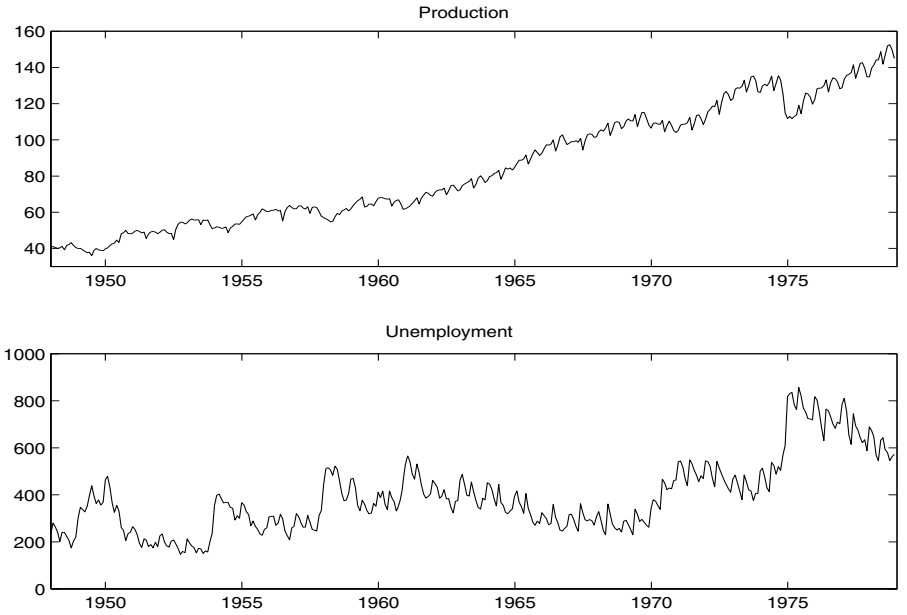


Figure 3.22 Values of the Monthly Federal Reserve Board Production Index and Unemployment (1948-1978, $n = 372$ months).

Example 3.43 Analysis of the Federal Reserve Board Production Index.

A problem of great interest in economics involves first identifying a model within the Box–Jenkins class for a given time series and then producing forecasts based on the model. For example, we might consider applying this methodology to the Federal Reserve Board Production Index shown in Figure 3.22. The ACFs and PACFs for this series are shown in Figure 3.23, and we note the slow decay in the ACF and the peak at lag $h = 1$ in the PACF, indicating nonstationary behavior.

Following the recommended procedure, a first difference was taken, and the ACF and PACF of the first difference

$$\nabla x_t = x_t - x_{t-1}$$

are shown in Figure 3.24. Noting the peaks at $12, 24, 36$, and 48 with relatively slow decay suggested a seasonal difference and Figure 3.25 shows the seasonal difference of the differenced production, say,

$$\nabla_{12}\nabla x_t = (1 - B^{12})(1 - B)x_t.$$

Characteristics of the ACF and PACF of this series tend to show a strong peak at $h = 12$ in the autocorrelation function, with smaller peaks appearing at $h = 24, 36$, combined with peaks at $h = 12, 24, 36, 48$, in the

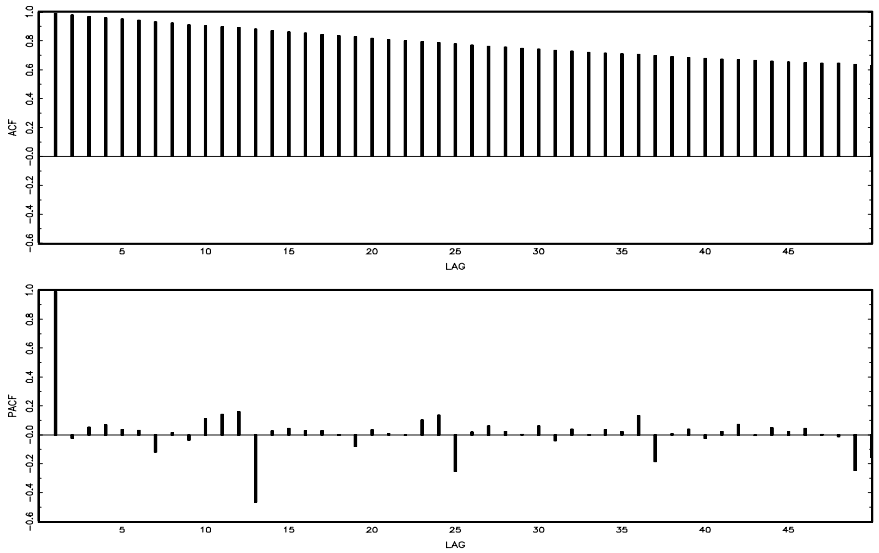


Figure 3.23 ACF and PACF of the production series.

partial autocorrelation function. Using Table 3.2, this suggests either a seasonal moving average of order $Q = 1$, a seasonal autoregression of possible order $P = 2$, or due to the fact that both the ACF and PACF may be tailing off at the seasonal lags, perhaps both components, $P = 2$ and $Q = 1$, are needed.

Inspecting the ACF and the PACF at the within season lags, $h = 1, \ldots, 11$, it appears that both the ACF and PACF are tailing off. Based on Table 3.1, this result indicates that we should consider fitting a model with both $p > 0$ and $q > 0$ for the nonseasonal components. Hence, at first we will consider $p = 1$ and $q = 1$.

Fitting the three models suggested by these observations and computing the AIC for each, we obtain:

$$\begin{aligned} &(i)\ \text{ARIMA}(1,1,1) \times (0,1,1)_{12}, && \text{AIC} = 1162.30 \\ &(ii)\ \text{ARIMA}(1,1,1) \times (2,1,0)_{12}, && \text{AIC} = 1169.04 \\ &(iii)\ \text{ARIMA}(1,1,1) \times (2,1,1)_{12}, && \text{AIC} = 1148.43 \end{aligned}$$

On the basis of the AICs, we prefer the $\text{ARIMA}(1,1,1)\times(2,1,1)_{12}$ model. Figure 3.26 shows the diagnostics for this model, leading to the conclusion that the model is adequate. We note, however, the presence of a few outliers.

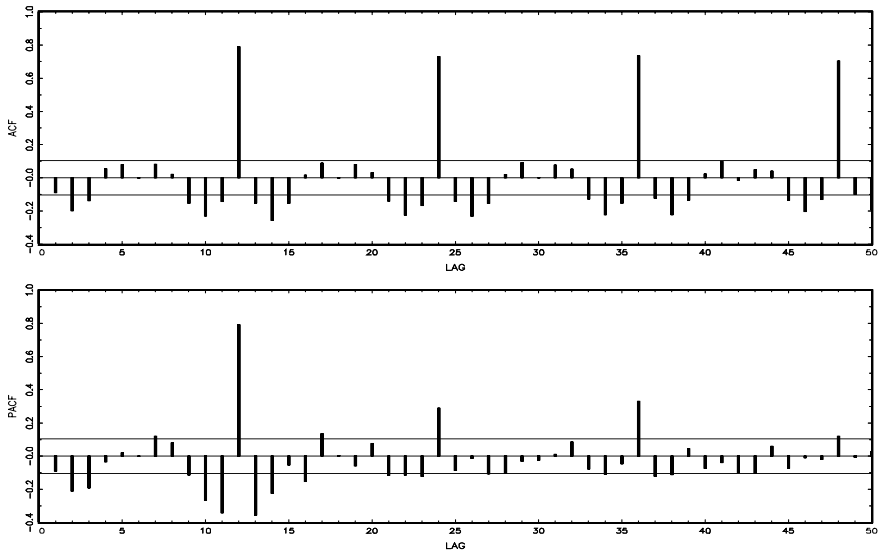


Figure 3.24 ACF and PACF of differenced production, $(1 - B)x_t$.

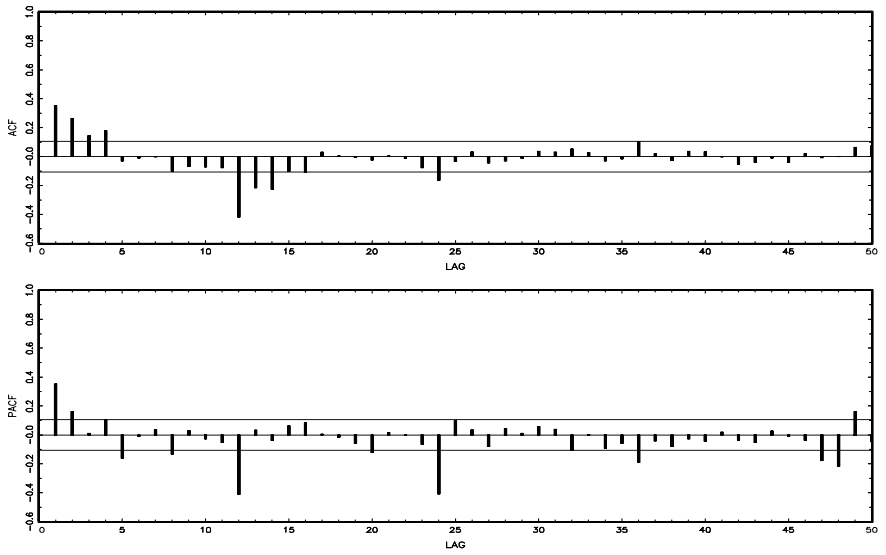


Figure 3.25 ACF and PACF of first differenced and then seasonally differenced production, $(1 - B)(1 - B^{12})x_t$.

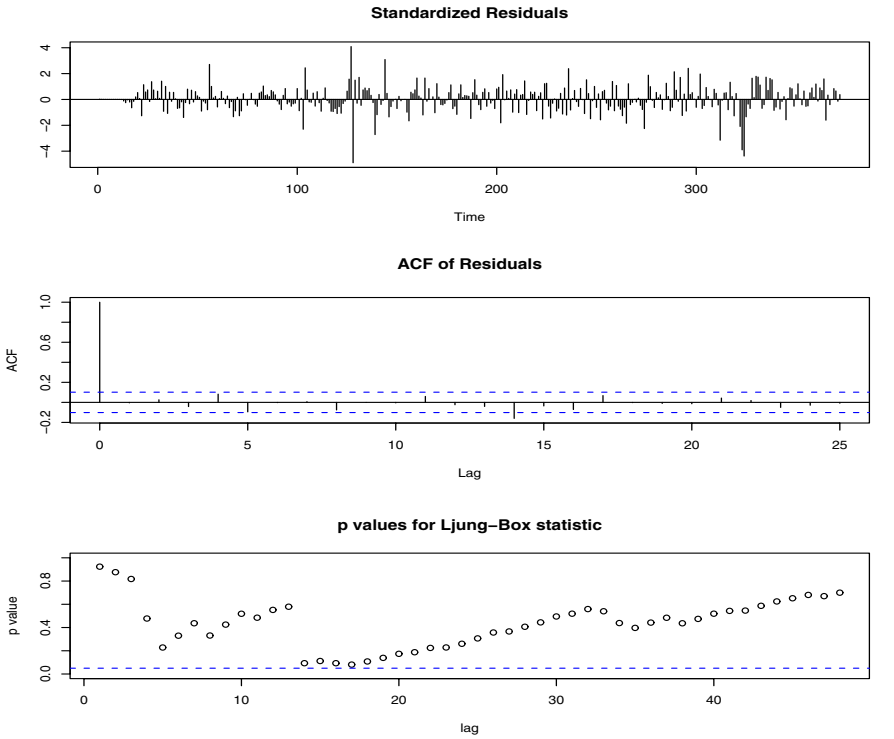


Figure 3.26 Diagnostics for the ARIMA$(1,1,1) \times (2,1,1)_{12}$ fit on the Production data.

The fitted ARIMA$(1,1,1) \times (2,1,1)_{12}$ is

$$(1 + .22_{(.08)}B^{12} + .28_{(.06)}B^{24})(1 - .58_{(.11)}B)\nabla_{12}\nabla\widehat{x}_t$$
$$= (1 - .50_{(.07)}B^{12})(1 - .27_{(.13)}B)\widehat{w}_t$$

with $\widehat{\sigma}_w^2 = 1.35$. Forecasts based on the fitted model for the next 12 months are shown in Figure 3.27.

Finally, we present the R code necessary to reproduce most of the analyses performed in Example 3.43.

```
> prod=scan("/mydata/prod.dat")
> par(mfrow=c(2,1))     # (P)ACF of data
> acf(prod, 48)
> pacf(prod, 48)
> par(mfrow=c(2,1))     # (P)ACF of d1 data
> acf(diff(prod), 48)
> pacf(diff(prod), 48)
```

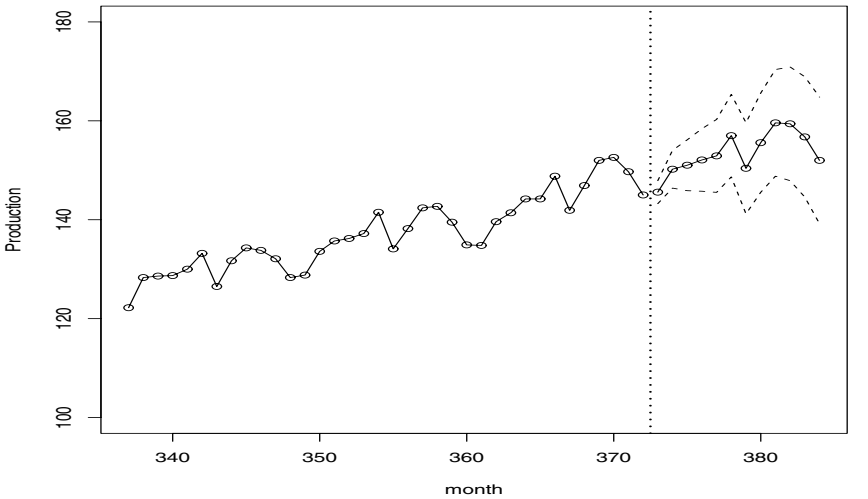


Figure 3.27 Forecasts and limits for production index. The vertical dotted line separates the data from the predictions.

```
> par(mfrow=c(2,1))    # (P)ACF of d1-d12 data
> acf(diff(diff(prod),12), 48)
> pacf(diff(diff(prod),12), 48)

> ### fit model (iii)
> prod.fit3 = arima(prod, order=c(1,1,1),
+   seasonal=list(order=c(2,1,1), period=12))
> prod.fit3    # to view the results
> tsdiag(prod.fit3, gof.lag=48)  # diagnostics

> ### forecasts for the final model
> prod.pr = predict(prod.fit3, n.ahead=12)
> U = prod.pr$pred + 2*prod.pr$se
> L = prod.pr$pred - 2*prod.pr$se
> month=337:372
> plot(month, prod[month], type="o", xlim=c(337,384),
+   ylim=c(100,180), ylab="Production")
> lines(prod.pr$pred, col="red", type="o")
> lines(U, col="blue", lty="dashed")
> lines(L, col="blue", lty="dashed")
> abline(v=372.5,lty="dotted")
```

Problems

Section 3.2

3.1 For an MA(1), $x_t = w_t + \theta w_{t-1}$, show that $|\rho_x(1)| \leq 1/2$ for any number θ. For which values of θ does $\rho_x(1)$ attain its maximum and minimum?

3.2 Let w_t be white noise with variance σ_w^2 and let $|\phi| < 1$ be a constant. Consider the process

$$\begin{aligned} x_1 &= w_1 \\ x_t &= \phi x_{t-1} + w_t \quad t = 2, 3, \ldots \ . \end{aligned}$$

(a) Find the mean and the variance of $\{x_t,\ t = 1, 2, \ldots\}$. Is x_t stationary?

(b) Show

$$\operatorname{corr}(x_t, x_{t-h}) = \phi^h \left[\frac{\operatorname{var}(x_{t-h})}{\operatorname{var}(x_t)} \right]^{1/2}$$

for $h \geq 0$.

(c) Argue that for large t,

$$\operatorname{var}(x_t) \approx \frac{\sigma_w^2}{1 - \phi^2}$$

and

$$\operatorname{corr}(x_t, x_{t-h}) \approx \phi^h, \quad h \geq 0,$$

so in a sense, x_t is "asymptotically stationary."

(d) Comment on how you could use these results to simulate n observations of a stationary Gaussian AR(1) model from simulated iid N(0,1) values.

(e) Now suppose $x_1 = w_1/\sqrt{1 - \phi^2}$. Is this process stationary?

3.3 Identify the following models as ARMA(p, q) models (watch out for parameter redundancy), and determine whether they are causal and/or invertible:

(a) $x_t = .80x_{t-1} - .15x_{t-2} + w_t - .30w_{t-1}$.

(b) $x_t = x_{t-1} - .50x_{t-2} + w_t - w_{t-1}$.

3.4 Verify the causal conditions for an AR(2) model given in (3.27). That is, show that an AR(2) is causal if and only if (3.27) holds.

Section 3.3

3.5 For the AR(2) model given by $x_t = -.9x_{t-2} + w_t$, find the roots of the autoregressive polynomial, and then sketch the ACF, $\rho(h)$.

3.6 For the AR(2) autoregressive series shown below, determine a set of difference equations that can be used to find $\psi_j, j = 0, 1, \ldots$ in the representation (3.24) and the autocorrelation function $\rho(h), h = 0, 1, \ldots$. Solve for the constants in the ACF using the known initial conditions, and plot the first eight values.

(a) $x_t + 1.6x_{t-1} + .64x_{t-2} = w_t$.

(b) $x_t - .40x_{t-1} - .45x_{t-2} = w_t$.

(c) $x_t - 1.2x_{t-1} + .85x_{t-2} = w_t$.

Section 3.4

3.7 Verify the calculations for the autocorrelation function of an ARMA(1, 1) process given in Example 3.11. Compare the form with that of the ACF for the ARMA(1, 0) and the ARMA(0, 1) series. Plot the ACFs of the three series on the same graph for $\phi = .6, \theta = .9$, and comment on the diagnostic capabilities of the ACF in this case.

3.8 Generate $n = 100$ observations from each of the three models discussed in Problem 3.7. Compute the sample ACF for each model and compare it to the theoretical values. Compute the sample PACF for each of the generated series and compare the sample ACFs and PACFs with the general results given in Table 3.1.

Section 3.5

3.9 Let M_t represent the cardiovascular mortality series discussed in Chapter 2, Example 2.2.

(a) Fit an AR(2) to M_t using linear regression as in Example 3.16.

(b) Assuming the fitted model in (a) is the true model, find the forecasts over a four-week horizon, x^n_{n+m}, for $m = 1, 2, 3, 4$, and the corresponding 95% prediction intervals.

3.10 Consider the MA(1) series

$$x_t = w_t + \theta w_{t-1},$$

where w_t is white noise with variance σ^2_w.

(a) Derive the minimum mean square error one-step forecast based on the infinite past, and determine the mean square error of this forecast.

(b) Let $\widetilde{x}_{n+1}^n$ be the truncated one-step-ahead forecast as given in (3.82). Show that

$$E\left[(x_{n+1} - \widetilde{x}_{n+1}^n)^2\right] = \sigma^2(1 + \theta^{2+2n}).$$

Compare the result with (a), and indicate how well the finite approximation works in this case.

3.11 In the context of equation (3.56), show that, if $\gamma(0) > 0$ and $\gamma(h) \to 0$ as $h \to \infty$, then Γ_n is positive definite.

3.12 Suppose x_t is stationary with zero mean and recall the definition of the PACF given by (3.49) and (3.50). That is, let

$$\epsilon_t = x_t - \sum_{i=1}^{h-1} a_i x_{t-i}$$

and

$$\delta_{t-h} = x_{t-h} - \sum_{j=1}^{h-1} b_j x_{t-j}$$

be the two residuals where $\{a_1, \ldots, a_{h-1}\}$ and $\{b_1, \ldots, b_{h-1}\}$ are chosen so that they minimize the mean-squared errors

$$E[\epsilon_t^2] \quad \text{and} \quad E[\delta_{t-h}^2].$$

The PACF at lag h was defined as the cross-correlation between ϵ_t and δ_{t-h}; that is,

$$\phi_{hh} = \frac{E(\epsilon_t \delta_{t-h})}{\sqrt{E(\epsilon_t^2)E(\delta_{t-h}^2)}}.$$

Let R_h be the $h \times h$ matrix with elements $\rho(i-j), i, j = 1, \ldots, h$, and let $\boldsymbol{\rho}_h = (\rho(1), \rho(2), \ldots, \rho(h))'$ be the vector of lagged autocorrelations, $\rho(h) = \mathrm{corr}(x_{t+h}, x_t)$. Let $\widetilde{\boldsymbol{\rho}}_h = (\rho(h), \rho(h-1), \ldots, \rho(1))'$ be the reversed vector. In addition, let x_t^h denote the BLP of x_t given $\{x_{t-1}, \ldots, x_{t-h}\}$:

$$x_t^h = \alpha_{h1} x_{t-1} + \cdots + \alpha_{hh} x_{t-h},$$

as described in Property P3.3. Prove

$$\phi_{hh} = \frac{\rho(h) - \widetilde{\boldsymbol{\rho}}_{h-1}' R_{h-1}^{-1} \boldsymbol{\rho}_h}{1 - \widetilde{\boldsymbol{\rho}}_{h-1}' R_{h-1}^{-1} \widetilde{\boldsymbol{\rho}}_{h-1}} = \alpha_{hh}.$$

In particular, this result proves Property P3.4.

Hint: Divide the prediction equations [see (3.56)] by $\gamma(0)$ and write the matrix equation in the partitioned form as

$$\begin{pmatrix} R_{h-1} & \widetilde{\boldsymbol{\rho}}_{h-1} \\ \widetilde{\boldsymbol{\rho}}'_{h-1} & \rho(0) \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \alpha_{hh} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\rho}_{h-1} \\ \rho(h) \end{pmatrix},$$

where the $h \times 1$ vector of coefficients $\boldsymbol{\alpha} = (\alpha_{h1}, \ldots, \alpha_{hh})'$ is partitioned as $\boldsymbol{\alpha} = (\boldsymbol{\alpha}'_1, \alpha_{hh})'$.

3.13 Suppose we wish to find a prediction function $g(x)$ that minimizes

$$MSE = E[(y - g(x))^2],$$

where x and y are jointly distributed random variables with density function $f(x, y)$.

(a) Show that MSE is minimized by the choice

$$g(x) = E(y \mid x).$$

Hint:

$$MSE = \int \left[\int (y - g(x))^2 f(y|x) dy \right] f(x) dx.$$

(b) Apply the above result to the model

$$y = x^2 + z,$$

where x and z are independent zero-mean normal variables with variance one. Show that $MSE = 1$.

(c) Suppose we restrict our choices for the function $g(x)$ to linear functions of the form

$$g(x) = a + bx$$

and determine a and b to minimize MSE. Show that $a = 1$ and

$$b = \frac{E(xy)}{E(x^2)} = 0$$

and $MSE = 3$. What do you interpret this to mean?

3.14 For an AR(1) model, determine the general form of the m-step-ahead forecast x^t_{t+m} and show

$$E[(x_{t+m} - x^t_{t+m})^2] = \sigma_w^2 \frac{1 - \phi^{2m}}{1 - \phi^2}.$$

3.15 Consider the ARMA(1,1) model discussed in Example 3.6, equation (3.26); that is, $x_t = .9x_{t-1} + .5w_{t-1} + w_t$. Show that truncated prediction as defined in (3.81) is equivalent to truncated prediction using the recursive formula (3.82).

3.16 Verify statement (3.78), that for a fixed sample size, the ARMA prediction errors are correlated.

Section 3.6

3.17 Let M_t represent the cardiovascular mortality series discussed in Chapter 2, Example 2.2. Fit an AR(2) model to the data using linear regression and using Yule–Walker.

(a) Compare the parameter estimates obtained by the two methods.

(b) Compare the estimated standard errors of the coefficients obtained by linear regression with their corresponding asymptotic approximations, as given in Property P3.9.

3.18 Suppose $x_1, \ldots, x_n$ are observations from an AR(1) process with $\mu = 0$.

(a) Show the backcasts can be written as $x_t^n = \phi^{1-t} x_1$, for $t \leq 1$.

(b) In turn, show, for $t \leq 1$, the backcasted errors are $\widehat{w}_t(\phi) = x_t^n - \phi x_{t-1}^n = \phi^{1-t}(1-\phi^2)x_1$.

(c) Use the result of (b) to show $\sum_{t=-\infty}^{1} \widehat{w}_t^2(\phi) = (1-\phi^2)x_1^2$.

(d) Use the result of (c) to verify the unconditional sum of squares, $S(\phi)$, can be written in the innovations form as $\sum_{t=-\infty}^{n} \widehat{w}_t^2(\phi)$.

(e) Find x_t^{t-1} and r_t^{t-1}, and show that $S(\phi)$ can also be written as $\sum_{t=1}^{n}(x_t - x_t^{t-1})^2 \,/\, r_t^{t-1}$.

3.19 Generate $n = 500$ observations from the ARMA model given by

$$x_t = .9x_{t-1} + w_t - .9w_{t-1},$$

with $w_t \sim$ iid N$(0, 1)$. Plot the simulated data, compute the sample ACF and PACF of the simulated data, and fit an ARMA$(1, 1)$ model to the data. What happened and how do you explain the results?

3.20 Generate 10 realizations of length $n = 200$ of a series from an ARMA(1,1) model with $\phi_1 = .90, \theta_1 = .2$ and $\sigma^2 = .25$. Fit the model by nonlinear least squares or maximum likelihood in each case and compare the estimators to the true values.

3.21 Generate $n = 50$ observations from a Gaussian AR(1) model with $\phi = .99$ and $\sigma_w = 1$. Using an estimation technique of your choice, compare the approximate asymptotic distribution of your estimate (the one you would use for inference) with the results of a bootstrap experiment (use $B = 200$).

3.22 Using Example 3.30 as your guide, find the Gauss–Newton procedure for estimating the autoregressive parameter, ϕ, from the AR(1) model, $x_t = \phi x_{t-1} + w_t$, given data $x_1, \ldots, x_n$. Does this procedure produce the unconditional or the conditional estimator? *Hint:* Write the model as $w_t(\phi) = x_t - \phi x_{t-1}$; your solution should work out to be a non-recursive procedure.

3.23 Consider the stationary series generated by

$$x_t = \alpha + \phi x_{t-1} + w_t + \theta w_{t-1},$$

where $E(x_t) = \mu$, $|\theta| < 1, |\phi| < 1$ and the w_t are iid random variables with zero mean and variance σ_w^2.

(a) Determine the mean as a function of α for the above model. Find the autocovariance and ACF of the process x_t, and show that the process is weakly stationary. Is the process strictly stationary?

(b) Prove the limiting distribution as $n \to \infty$ of the sample mean,

$$\bar{x} = n^{-1} \sum_{t=1}^{n} x_t,$$

is normal, and find its limiting mean and variance in terms of α, ϕ, θ, and σ_w^2. (Note: This part uses results from Appendix A.)

3.24 A problem of interest in the analysis of geophysical time series involves a simple model for observed data containing a signal and a reflected version of the signal with unknown amplification factor a and unknown time delay δ. For example, the depth of an earthquake is proportional to the time delay δ for the P wave and its reflected form pP on a seismic record. Assume the signal is white and Gaussian with variance σ_s^2, and consider the generating model

$$x_t = s_t + a s_{t-\delta}.$$

(a) Prove the process x_t is stationary. If $|a| < 1$, show that

$$s_t = \sum_{j=0}^{\infty} (-a)^j x_{t-\delta j}$$

is a mean square convergent representation for the signal s_t, for $t = 1, \pm 1, \pm 2, \ldots$.

(b) If the time delay δ is assumed to be known, suggest an approximate computational method for estimating the parameters a and σ_s^2 using maximum likelihood and the Gauss–Newton method.

(c) If the time delay δ is an unknown integer, specify how we could estimate the parameters including δ. Generate a $n = 500$ point series with $a = .9$, $\sigma_w^2 = 1$ and $\delta = 5$. Estimate the integer time delay δ by searching over $\delta = 3, 4, \ldots, 7$.

3.25 *Forecasting with estimated parameters:* Let $x_1, x_2, \ldots, x_n$ be a sample of size n from a causal AR(1) process, $x_t = \phi x_{t-1} + w_t$. Let $\widehat{\phi}$ be the Yule–Walker estimator of ϕ.

(a) Show $\widehat{\phi} - \phi = O_p(n^{-1/2})$. See Appendix A for the definition of $O_p(\cdot)$.

(b) Let x^n_{n+1} be the one-step-ahead forecast of x_{n+1} given the data $x_1, \ldots, x_n$, based on the known parameter, ϕ, and let $\widehat{x}^n_{n+1}$ be the one-step-ahead forecast when the parameter is replaced by $\widehat{\phi}$. Show $x^n_{n+1} - \widehat{x}^n_{n+1} = O_p(n^{-1/2})$.

Section 3.7

3.26 Suppose

$$y_t = \beta_0 + \beta_1 t + \cdots + \beta_q t^q + x_t, \quad \beta_q \neq 0,$$

where x_t is stationary. First, show that $\nabla^k x_t$ is stationary for any $k = 1, 2, \ldots$, and then show that $\nabla^k y_t$ is not stationary for $k < q$, but is stationary for $k \geq q$.

3.27 Verify that the IMA(1,1) model given in (3.131) can be inverted and written as (3.132).

3.28 For the logarithm of the glacial varve data, say, x_t, presented in Example 3.31, use the first 100 observations and calculate the EWMA, $\widetilde{x}^t_{t+1}$, given in (3.134) for $t = 1, \ldots, 100$, using $\lambda = .25, .50$, and $.75$, and plot the EWMAs and the data superimposed on each other. Comment on the results.

Section 3.8

3.29 In Example 3.36, we presented the diagnostics for the MA(2) fit to the GNP growth rate series. Using that example as a guide, complete the diagnostics for the AR(1) fit.

3.30 Using the gas price series described in Problem 2.9, fit an ARIMA(p, d, q) model to the data, performing all necessary diagnostics. Comment.

3.31 The second column in the data file `globtemp2.dat` are annual global temperature deviations from 1880 to 2004. The data are an update to the Hansen-Lebedeff global temperature data and the URL of the data source is in the file. Fit an ARIMA(p, d, q) model to the data, performing all of the necessary diagnostics. After deciding on an appropriate model, forecast (with limits) the next 10 years. Comment. In R, use `read.table` to load the data file.

3.32 One of the series collected along with particulates, temperature, and mortality described in Example 2.2 is the sulfur dioxide series. Fit an ARIMA(p, d, q) model to the data, performing all of the necessary diagnostics. After deciding on an appropriate model, forecast the data into the future four time periods ahead (about one month) and calculate 95% prediction intervals for each of the four forecasts. Comment.

Section 3.9

3.33 Consider the ARIMA model

$$x_t = w_t + \Theta w_{t-2}.$$

(a) Identify the model using the notation ARIMA$(p, d, q) \times (P, D, Q)_s$.

(b) Show that the series is invertible for $|\Theta| < 1$, and find the coefficients in the representation

$$w_t = \sum_{k=0}^{\infty} \pi_k x_{t-k}.$$

(c) Develop equations for the m-step ahead forecast, $\widetilde{x}_{n+m}$, and its variance based on the infinite past, $x_n, x_{n-1}, \ldots$.

3.34 Sketch the ACF of the seasonal ARIMA$(0, 1) \times (1, 0)_{12}$ model with $\Phi = .8$ and $\theta = .5$.

3.35 Fit a seasonal ARIMA model of your choice to the unemployment data displayed in Figure 3.22. Use the estimated model to forecast the next 12 months.

3.36 Fit a seasonal ARIMA model of your choice to the U.S. Live Birth Series (`birth.dat`). Use the estimated model to forecast the next 12 months.

3.37 Fit an appropriate seasonal ARIMA model to the log-transformed Johnson and Johnson earnings series of Example 1.1. Use the estimated model to forecast the next 4 quarters.

The following problems require the supplemental material given in Appendix B

3.38 Suppose $x_t = \sum_{j=1}^{p} \phi_j x_{t-j} + w_t$, where $\phi_p \neq 0$ and w_t is white noise such that w_t is uncorrelated with $\{x_k; k < t\}$. Use the Projection Theorem to show that, for $n > p$, the BLP of x_{n+1} on $\overline{\text{sp}}\{x_k, k \leq n\}$ is

$$\widehat{x}_{n+1} = \sum_{j=1}^{p} \phi_j x_{n+1-j}.$$

3.39 Use the Projection Theorem to derive the Innovations Algorithm, Property P3.6, equations (3.68)-(3.70). Then, use Theorem B.2 to derive the m-step-ahead forecast results given in (3.71) and (3.72).

3.40 Consider the series $x_t = w_t - w_{t-1}$, where w_t is a white noise process with mean zero and variance σ_w^2. Suppose we consider the problem of predicting x_{n+1}, based on only $x_1, \ldots, x_n$. Use the Projection Theorem to answer the questions below.

(a) Show the best linear predictor is

$$x_{n+1}^n = -\frac{1}{n+1} \sum_{k=1}^{n} k\, x_k.$$

(b) Prove the mean square error is

$$E(x_{n+1} - x_{n+1}^n)^2 = \frac{n+2}{n+1}\, \sigma_w^2.$$

3.41 Use Theorem B.2 and B.3 to verify (3.105).

3.42 Prove Theorem B.2.

3.43 Prove Property P3.2.

Chapter 4

Spectral Analysis and Filtering

4.1 Introduction

The notion that a time series exhibits repetitive or regular behavior over time is of fundamental importance because it distinguishes time series analysis from classical statistics, which assumes complete independence over time. We have seen how dependence over time can be introduced through models that describe in detail the way certain empirical data behaves, even to the extent of producing forecasts based on the models. It is natural that models based on predicting the present as a regression on the past, such as are provided by the celebrated ARIMA or state-space forms, will be attractive to statisticians, who are trained to view nature in terms of linear models. In fact, the difference equations used to represent these kinds of models are simply the discrete versions of linear differential equations that may, in some instances, provide the ideal physical model for a certain phenomenon. An alternate version of the way nature behaves exists, however, and is based on a decomposition of an empirical series into its regular components.

In this chapter, we argue, the concept of regularity of a series can best be expressed in terms of periodic variations of the underlying phenomenon that produced the series, expressed as Fourier frequencies being driven by sines and cosines. Such a possibility was discussed in Chapters 1 and 2. From a regression point of view, we may imagine a system responding to various driving frequencies by producing linear combinations of sine and cosine functions. Expressed in these terms, the time domain approach may be thought of as regression of the present on the past, whereas the frequency domain approach may be considered as regression of the present on periodic sines and cosines. The frequency domain approaches are the focus of this chapter and

Chapter 7. To illustrate the two methods for generating series with a single primary periodic component, consider Figure 1.9, which was generated from a simple second-order autoregressive model, and the middle and bottom panels of Figure 1.11, which were generated by adding a cosine wave with a period of 50 points to white noise. Both series exhibit strong periodic fluctuations, illustrating that both models can generate time series with regular behavior. As discussed in Examples 2.7–2.9, a fundamental objective of spectral analysis is to identify the dominant frequencies in a series and to find an explanation of the system from which the measurements were derived.

Of course, the primary justification for any alternate model must lie in its potential for explaining the behavior of some empirical phenomenon. In this sense, an explanation involving only a few kinds of primary oscillations becomes simpler and more physically meaningful than a collection of parameters estimated for some selected difference equation. It is the tendency of observed data to show periodic kinds of fluctuations that justifies the use of frequency domain methods. Many of the examples in §1.2 are time series representing real phenomena that are driven by periodic components. The speech recording of the syllable *aa...hh* in Figure 1.3 contains a complicated mixture of frequencies related to the opening and closing of the glottis. Figure 1.5 shows the monthly SOI, which we later explain as a combination of two kinds of periodicities, a seasonal periodic component of 12 months and an El Niño component of about three to five years. Of fundamental interest is the return period of the El Niño phenomenon, which can have profound effects on local climate. Also of interest is whether the different periodic components of the new fish population depend on corresponding seasonal and El Niño-type oscillations. We introduce the coherence as a tool for relating the common periodic behavior of two series. Seasonal periodic components are often pervasive in economic time series; this phenomenon can be seen in the quarterly earnings series shown in Figure 1.1. In Figure 1.6, we see the extent to which various parts of the brain will respond to a periodic stimulus generated by having the subject do alternate left and right finger tapping. Figure 1.7 shows series from an earthquake and a nuclear explosion. The relative amounts of energy at various frequencies for the two phases can produce statistics, useful for discriminating between earthquakes and explosions.

In this chapter, we summarize an approach to handling correlation generated in stationary time series that begins by transforming the series to the frequency domain. This simple linear transformation essentially matches sines and cosines of various frequencies against the underlying data and serves two purposes as discussed in Examples 2.7 and 2.8. The periodogram that was introduced in Example 2.8 has its population counterpart called the power spectrum, and its estimation is a main goal of spectral analysis. Another purpose of exploring this topic is statistical convenience resulting from the periodic components being nearly uncorrelated. This property facilitates writing likelihoods based on classical statistical methods

An important part of analyzing data in the frequency domain, as well as

the time domain, is the investigation and exploitation of the properties of the time-invariant linear filter. This special linear transformation is used similarly to linear regression in conventional statistics, and we use many of the same terms in the time series context. We have previously mentioned the coherence as a measure of the relation between two series at a given frequency, and we show later that this coherence also measures the performance of the best linear filter relating the two series. Linear filtering can also be an important step in isolating a signal embedded in noise. For example, the lower panels of Figure 1.11 contain a signal contaminated with an additive noise, whereas the upper panel contains the pure signal. It might also be appropriate to ask whether a linear filter transformation exists that could be applied to the lower panel to produce a series closer to the signal in the upper panel. The use of filtering for reducing noise will also be a part of the presentation in this chapter. We emphasize, throughout, the analogy between filtering techniques and conventional linear regression.

Many frequency scales will often coexist, depending on the nature of the problem. For example, in the Johnson & Johnson data set in Figure 1.1, the predominant frequency of oscillation is one cycle per year (4 quarters), or .25 cycles per observation. The predominant frequency in the SOI and fish populations series in Figure 1.5 is also one cycle per year, but this corresponds to 1 cycle every 12 months, or .083 cycles per observation. For simplicity, we measure frequency, ω, at cycles per time point and discuss the implications of certain frequencies in terms of the problem context. Of descriptive interest is the period of a time series, defined as the number of points in a cycle, i.e.,

$$T = \frac{1}{\omega}. \tag{4.1}$$

Hence, the predominant period of the Johnson & Johnson series is 1/.25 or 4 quarters per cycle, whereas the predominant period of the SOI series is 12 months per cycle.

4.2 Cyclical Behavior and Periodicity

As previously mentioned, we have already encountered the notion of periodicity in numerous examples in Chapters 1 and 2. The general notion of periodicity can be made more precise by introducing some terminology. In order to define the rate at which a series oscillates, we first define a cycle as one complete period of a sine or cosine function defined over a time interval of length 2π. As in (1.5), we consider the periodic process

$$x_t = A\cos(2\pi\omega t + \phi) \tag{4.2}$$

for $t = 0, \pm 1, \pm 2, \ldots$, where ω is a frequency index, defined in cycles per unit time with A determining the height or *amplitude* of the function and ϕ, called

the *phase*, determining the start point of the cosine function. We can introduce random variation in this time series by allowing the amplitude and phase to vary randomly.

As discussed in Example 2.7, for purposes of data analysis, it is easier to use a trigonometric identity[1] and write (4.2) as

$$x_t = U_1 \cos(2\pi\omega t) + U_2 \sin(2\pi\omega t), \tag{4.3}$$

where $U_1 = A\cos\phi$ and $U_2 = -A\sin\phi$ are often taken to be normally distributed random variables. In this case, the amplitude is $A = \sqrt{U_1^2 + U_2^2}$ and the phase is $\phi = \tan^{-1}(-U_2/U_1)$. From these facts we can show that if, and only if, in (4.2), A and ϕ are independent random variables, where A^2 is chi-squared with 2 degrees of freedom, and ϕ is uniformly distributed on $(-\pi, \pi)$, then U_1 and U_2 are independent, standard normal random variables (see Problem 4.2).

The above random process is also a function of its frequency, defined by the parameter ω. The frequency is measured in cycles per unit time, or in cycles per point in the above illustration. For $\omega = 1$, the series makes one cycle per time unit; for $\omega = .50$, the series makes a cycle every two time units; for $\omega = .25$, every four units, and so on. In general, data that occurs at discrete time points will need at least two points to determine a cycle, so the highest frequency of interest is .5 cycles per point. This frequency is called the *folding frequency* and defines the highest frequency that can be seen in discrete sampling. Higher frequencies sampled this way will appear at lower frequencies, called *aliases*; an example is the way a camera samples a rotating wheel on a moving automobile in a movie, in which the wheel appears to be rotating at a different rate. For example, movies are recorded at 24 frames per second. If the camera is filming a wheel that is rotating at the rate of 24 cycles per second (or 24 Hertz), the wheel will appear to stand still (that's about 110 miles per hour in case you were wondering).

Consider a generalization of (4.3) that allows mixtures of periodic series, with multiple frequencies and amplitudes.

$$x_t = \sum_{k=1}^{q} \left[U_{k1} \cos(2\pi\omega_k t) + U_{k2} \sin(2\pi\omega_k t)\right], \tag{4.4}$$

where U_{k1}, U_{k2}, for $k = 1, 2, \ldots, q$, are independent zero-mean random variables with variances σ_k^2, and the ω_k are distinct frequencies. Notice that (4.4) exhibits the process as a sum of independent components, with variance σ_k^2 for frequency ω_k. Using the independence of the Us and a trig identity,[1] it is easy to show (Problem 4.3) that the autocovariance function of the process is

$$\gamma(h) = \sum_{k=1}^{q} \sigma_k^2 \cos(2\pi\omega_k h), \tag{4.5}$$

[1] $\cos(\alpha \pm \beta) = \cos(\alpha)\cos(\beta) \mp \sin(\alpha)\sin(\beta)$.

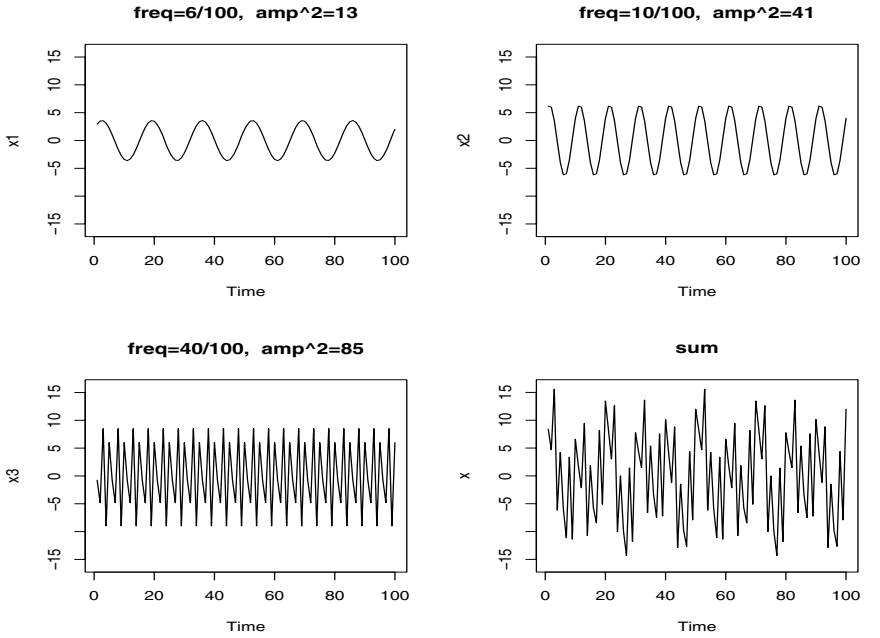


Figure 4.1 Periodic components and their sum as described in Example 4.1.

and we note the autocovariance function is the sum of periodic components with weights proportional to the variances σ_k^2. Hence, x_t is a mean-zero stationary processes with variance

$$\gamma(0) = E(x_t^2) = \sum_{k=1}^{q} \sigma_k^2, \tag{4.6}$$

which exhibits the overall variance as a sum of variances of each of the component parts.

Example 4.1 A Periodic Series

Figure 4.1 shows an example of the mixture (4.4) with $q = 3$ constructed in the following way. First, for $t = 1, \ldots, 100$, we generated three series

$$x_{t1} = 2\cos(2\pi t\, 6/100) + 3\sin(2\pi t\, 6/100)$$
$$x_{t2} = 4\cos(2\pi t\, 10/100) + 5\sin(2\pi t\, 10/100)$$
$$x_{t3} = 6\cos(2\pi t\, 40/100) + 7\sin(2\pi t\, 40/100)$$

These three series are displayed in Figure 4.1 along with the corresponding frequencies and squared amplitudes. For example, the squared amplitude of x_{t1} is $2^2 + 3^2 = 13$. Hence, the maximum and minimum values that x_{t1} will attain are $\pm\sqrt{13} = \pm 3.61$.

Finally, we constructed

$$x_t = x_{t1} + x_{t2} + x_{t3}$$

and this series is also displayed in Figure 4.1. We note that x_t appears to behave as some of the periodic series we saw in Chapters 1 and 2. The systematic sorting out of the essential frequency components in a time series, including their relative contributions, constitutes one of the main objectives of spectral analysis.

The R code to reproduce Figure 4.1 is

```
> t = 1:100
> x1 = 2*cos(2*pi*t*6/100) + 3*sin(2*pi*t*6/100)
> x2 = 4*cos(2*pi*t*10/100) + 5*sin(2*pi*t*10/100)
> x3 = 6*cos(2*pi*t*40/100) + 7*sin(2*pi*t*40/100)
> x = x1 + x2 + x3
> par(mfrow=c(2,2))
> plot.ts(x1, ylim=c(-16,16), main="freq=6/100, amp^2=13")
> plot.ts(x2, ylim=c(-16,16), main="freq=10/100, amp^2=41")
> plot.ts(x3, ylim=c(-16,16), main="freq=40/100, amp^2=85")
> plot.ts(x, ylim=c(-16,16), main="sum")
```

Example 4.2 The Scaled Periodogram for Example 4.1

In §2.3, Example 2.8, we introduced the periodogram as a way to discover the periodic components of a time series. Recall that the scaled periodogram is given by

$$P(j/n) = \left(\frac{2}{n}\sum_{t=1}^{n} x_t \cos(2\pi tj/n)\right)^2 + \left(\frac{2}{n}\sum_{t=1}^{n} x_t \sin(2\pi tj/n)\right)^2 \quad (4.7)$$

and it may regarded as a measure of the squared correlation of the data with sinusoids oscillating at a frequency of $\omega_j = j/n$, or j cycles in n time points. Recall that we are basically computing the regression of the data on the sinusoids varying at the fundamental frequencies, j/n. As discussed in Example 2.8, the periodogram may be computed quickly using the fast Fourier transform (FFT), and there is no need to run repeated regressions.

The scaled periodogram of the data, x_t, simulated in Example 4.1 is shown in Figure 4.2, and it clearly identifies the three components x_{t1}, x_{t2}, and x_{t3} of x_t. Moreover, the heights of the scaled periodogram shown in the figure are

$$P(6/100) = 13, \quad P(10/100) = 41, \quad P(40/100) = 85$$

and $P(j/n) = 0$ otherwise. These are exactly the values of the squared amplitudes of the components generated in Example 4.1. This outcome

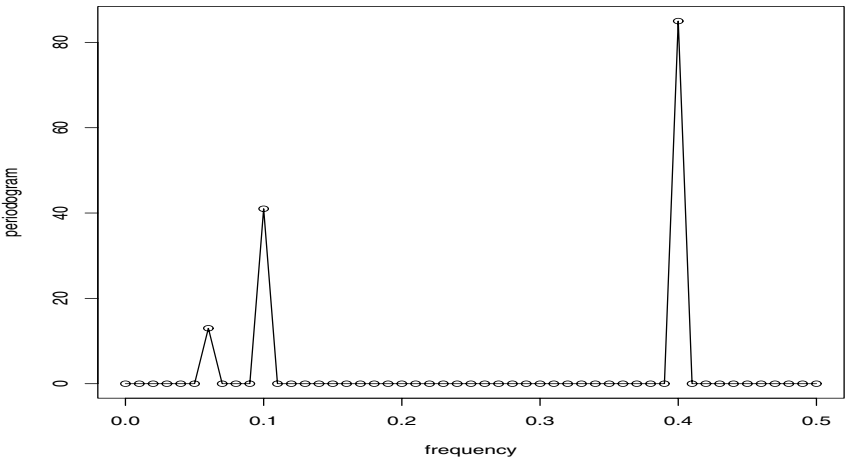


Figure 4.2 Periodogram of the data generated in Example 4.1.

suggests that the periodogram may provide some insight into the variance components, (4.6), of a real set of data.

Assuming the simulated data, `x`, were retained from the previous example, the R code to reproduce Figure 4.2 is

```
> P = abs(2*fft(x)/100)^2
> f = 0:50/100
> plot(f, P[1:51], type="o", xlab="frequency",
+   ylab="periodogram")
```

A curious reader may also wish to plot the entire periodogram over all fundamental frequencies between zero and one. A quick and easy way to do this is to use the command `plot.ts(P)`.

If we consider the data x_t in Example 4.1 as a color (waveform) made up of primary colors x_{t1}, x_{t2}, x_{t3} at various strengths (amplitudes), then we might consider the periodogram as a prism that decomposes the color x_t into its primary colors (spectrum). Hence the term *spectral analysis*.

Another fact that may be of use in understanding the periodogram is that for any time series sample $x_1, \ldots, x_n$, where n is odd, we may write, *exactly*

$$x_t = a_0 + \sum_{j=1}^{(n-1)/2} \left[a_j \cos(2\pi t\, j/n) + b_j \sin(2\pi t\, j/n) \right], \tag{4.8}$$

for $t = 1, \ldots, n$ and suitably chosen coefficients. If n is even, the representation (4.8) can be modified by summing to $(n/2 - 1)$ and adding an additional component given by $a_{n/2} \cos(2\pi t\, 1/2) = a_{n/2}(-1)^t$. The crucial point here is that (4.8) is exact for any sample. Hence (4.4) may be thought of as an

approximation to (4.8), the idea being that many of the coefficients in (4.8) may be close to zero. Recall from Example 2.8, that

$$P(j/n) = a_j^2 + b_j^2, \tag{4.9}$$

so the scaled periodogram indicates which periodic components in (4.8) are large and which components are small. We also saw (4.9) in Example 4.2.

The periodogram, which was introduced in Schuster (1898) and used in Schuster (1906) for studying the periodicities in the sunspot series (shown in Figure 4.31 in the Problems section) is a sample based statistic. In Example 4.2, we discussed the fact that the periodogram may be giving us an idea of the variance components associated with each frequency, as presented in (4.6), of a time series. These variance components, however, are population parameters. The concepts of population parameters and sample statistics, as they relate to spectral analysis of time series can be generalized to cover stationary time series and that is the topic of the next section.

4.3 The Spectral Density

The idea that a time series is composed of periodic components, appearing in proportion to their underlying variances, is fundamental in the spectral representation given in Theorem C.2 of Appendix C. The result is quite technical because it involves stochastic integration; that is, integration with respect to a stochastic process. In nontechnical terms, Theorem C.2 says that (4.4) is approximately true for any stationary time series. In other words, *any stationary time series may be thought of, approximately, as the random superposition of sines and cosines oscillating at various frequencies.*

Given that (4.4) is approximately true for all stationary time series, the next question is whether a meaningful representation for its autocovariance function, like the one displayed in (4.5), also exists. The answer is yes, and this representation is given in Theorem C.1 of Appendix C. The following example will help explain the result.

Example 4.3 A Periodic Stationary Process

Consider a periodic stationary random process given by (4.3), with a fixed frequency ω_0, say,

$$x_t = U_1 \cos(2\pi\omega_0 t) + U_2 \sin(2\pi\omega_0 t),$$

where U_1 and U_2 are independent zero-mean random variables with equal variance σ^2. The number of time periods needed for the above series to

complete one cycle is exactly $1/\omega_0$, and the process makes exactly ω_0 cycles per point for $t = 0, \pm 1, \pm 2, \ldots$. It is easily shown that[2]

$$\begin{aligned}\gamma(h) &= \sigma^2 \cos(2\pi\omega_0 h) = \frac{\sigma^2}{2} e^{-2\pi i\omega_0 h} + \frac{\sigma^2}{2} e^{2\pi i\omega_0 h} \\ &= \int_{-1/2}^{1/2} e^{2\pi i\omega h} dF(\omega)\end{aligned}$$

using a Riemann–Stieltjes integration, where $F(\omega)$ is the function defined by

$$F(\omega) = \begin{cases} 0 & \omega < -\omega_0 \\ \sigma^2/2, & -\omega_0 \le \omega < \omega_0 \\ \sigma^2 & \omega \ge \omega_0 \,. \end{cases}$$

The function $F(\omega)$ behaves like a cumulative distribution function for a discrete random variable, except that $F(\infty) = \sigma^2 = \gamma_x(0)$ instead of one. In fact, $F(\omega)$ is a cumulative distribution function, not of probabilities, but rather of variances associated with the frequency ω_0 in an analysis of variance, with $F(\infty)$ being the total variance of the process x_t. Hence, we term $F(\omega)$ the *spectral distribution function.*

Theorem C.1 in Appendix C states that a representation such as the one given in Example 4.3 always exists for a stationary process. In particular, if x_t is stationary with autocovariance $\gamma(h) = E[(x_{t+h} - \mu)(x_t - \mu)]$, then there exists a unique monotonically increasing function $F(\omega)$, called the spectral distribution function, that is bounded, with $F(-\infty) = F(-1/2) = 0$, and $F(\infty) = F(1/2) = \gamma(0)$ such that

$$\gamma(h) = \int_{-1/2}^{1/2} e^{2\pi i\omega h} \, dF(\omega). \tag{4.10}$$

A more important situation we use repeatedly is the one covered by Theorem C.3, where it is shown that, subject to absolute summability of the autocovariance, the spectral distribution function is absolutely continuous with $dF(\omega) = f(\omega)\, d\omega$, and the representation (4.10) becomes the motivation for the property given below.

Property P4.1: The Spectral Density
If the autocovariance function, $\gamma(h)$, of a stationary process satisfies

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty, \tag{4.11}$$

[2]Some identities may be helpful here: $e^{i\alpha} = \cos(\alpha) + i\sin(\alpha)$, so $\cos(\alpha) = (e^{i\alpha} + e^{-i\alpha})/2$ and $\sin(\alpha) = (e^{i\alpha} - e^{-i\alpha})/2i$.

then it has the representation

$$\gamma(h) = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega h} \, f(\omega) \, d\omega \quad h = 0, \pm 1, \pm 2, \ldots \tag{4.12}$$

as the inverse transform of the spectral density, which has the representation

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h)\mathrm{e}^{-2\pi i\omega h} \quad -1/2 \leq \omega \leq 1/2. \tag{4.13}$$

This spectral density is the analogue of the probability density function; the fact that $\gamma(h)$ is non-negative definite ensures

$$f(\omega) \geq 0$$

for all ω (see Appendix C, Theorem C.3 for details). It follows immediately from (4.12) and (4.13) that

$$f(\omega) = f(-\omega)$$

and

$$f(\omega + 1) = f(\omega),$$

verifying the spectral density is an even function of period one. Because of the evenness, we will typically only plot $f(\omega)$ for $\omega \geq 0$. In addition, putting $h = 0$ in (4.12) yields

$$\gamma(0) = \mathrm{var}(x_t) = \int_{-1/2}^{1/2} f(\omega) \, d\omega,$$

which expresses the total variance as the integrated spectral density over all of the frequencies. We show later on, that a linear filter can isolate the variance in certain frequency intervals or bands.

Analogous to probability theory, $\gamma(h)$ in (4.12) is the characteristic function of the spectral density $f(\omega)$ in (4.13). These facts should make it clear that, when the condition of Property P4.1 is satisfied, the autocovariance function $\gamma(h)$ and the spectral density function $f(\omega)$ contain the same information. That information, however, is expressed in different ways. The autocovariance function expresses information in terms of lags, whereas the spectral density expresses the same information in terms of cycles. Some problems are easier to work with when considering lagged information and we would tend to handle those problems in the time domain. Nevertheless, other problems are easier to work with when considering periodic information and we would tend to handle those problems in the spectral domain.

We also mention, at this point, that we have been focusing on the frequency ω, expressed in cycles per point rather than the more common (in statistics)

alternative $\lambda = 2\pi\omega$ that would give radians per point. Finally, the absolute summability condition, (4.11), is not satisfied by (4.5), the example that we have used to introduce the idea of a spectral representation. The condition, however, is satisfied for ARMA models.

We note that the autocovariance function, $\gamma(h)$, in (4.12) and the spectral density, $f(\omega)$, in (4.13) are Fourier transform pairs. In general, we have the following definition.

Definition 4.1 *For a general function* $\{a_t; t = 0, \pm 1, \pm 2, \ldots\}$ *satisfying the absolute summability condition*

$$\sum_{t=-\infty}^{\infty} |a_t| < \infty, \tag{4.14}$$

we define a **Fourier transform pair** *to be of the form*

$$A(\omega) = \sum_{t=-\infty}^{\infty} a_t \mathrm{e}^{-2\pi i\omega t} \tag{4.15}$$

and

$$a_t = \int_{-1/2}^{1/2} A(\omega)\mathrm{e}^{2\pi i\omega t}\, d\omega. \tag{4.16}$$

The use of (4.12) and (4.13) as Fourier transform pairs is fundamental in the study of stationary discrete time processes. Under the summability condition (4.11), the Fourier transform pair (4.12) and (4.13) will exist and this relation is unique. If $f(\omega)$ and $g(\omega)$ are two spectral densities for which

$$\int_{-1/2}^{1/2} f(\omega)\mathrm{e}^{2\pi i\omega h}\, d\omega = \int_{-1/2}^{1/2} g(\omega)\mathrm{e}^{2\pi i\omega h}\, d\omega \tag{4.17}$$

for all $h = 0, \pm 1, \pm 2, \ldots$, then

$$f(\omega) = g(\omega) \tag{4.18}$$

almost everywhere.

It is illuminating to examine the spectral density for the series that we have looked at in earlier discussions.

Example 4.4 White Noise Series

As a simple example, consider the theoretical power spectrum of a sequence of uncorrelated random variables, w_t, with variance σ_w^2. A simulated set of data is displayed in the top of Figure 1.8. Because the autocovariance function was computed in Example 1.16 as $\gamma_w(h) = \sigma_w^2$ for $h = 0$, and zero, otherwise, it follows from (4.13) that

$$f_w(\omega) = \sigma_w^2$$

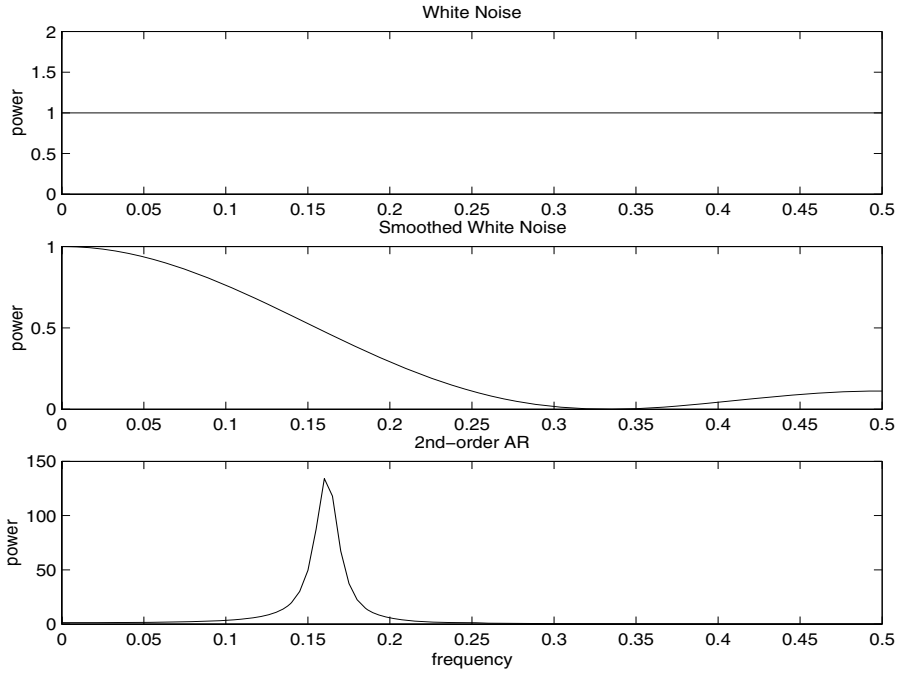


Figure 4.3 Theoretical spectra of white noise (top), smoothed white noise (middle), and a second-order autoregressive process (bottom).

for $-1/2 \le \omega \le 1/2$ with the resulting equal power at all frequencies. This property is seen in the realization, which seems to contain all different frequencies in a roughly equal mix. In fact, the name white noise comes from the analogy to white light, which contains all frequencies in the color spectrum. Figure 4.3 shows a plot of the white noise spectrum for $\sigma_w^2 = 1$.

Example 4.5 A Simple Moving Average

A series that does not have an equal mix of frequencies is the smoothed white noise series shown in the bottom panel of Figure 1.8. Specifically, we construct the three-point moving average series, defined by

$$v_t = \frac{1}{3}(w_{t-1} + w_t + w_{t+1}).$$

It is clear from the sample realization that the series has less of the higher or faster frequencies, and we calculate its power spectrum to verify this observation. We have previously computed the autocovariance of this

process in Example 1.17, obtaining

$$\gamma_v(h) = \frac{\sigma_w^2}{9}\Big(3 - |h|\Big)$$

for $|h| \le 2$ and $\gamma_y(h) = 0$ for $|h| > 2$. Then, using (4.13) gives

$$\begin{aligned} f_v(\omega) &= \sum_{h=-2}^{2} \gamma_y(h)\, \mathrm{e}^{-2\pi i\omega h} \\ &= \frac{\sigma_w^2}{9}\big(\mathrm{e}^{-4\pi i\omega} + \mathrm{e}^{4\pi i\omega}\big) + \frac{2\sigma_w^2}{9}\big(\mathrm{e}^{-2\pi i\omega} + \mathrm{e}^{2\pi\omega}\big) + \frac{3\sigma_w^2}{9} \\ &= \frac{\sigma_w^2}{9}\big[3 + 4\cos(2\pi\omega) + 2\cos(4\pi\omega)\big]. \end{aligned}$$

Plotting the spectrum for $\sigma_w^2 = 1$, as in Figure 4.3, shows the lower frequencies near zero have greater power and the higher or faster frequencies, say, $\omega > .2$, tend to have less power.

Example 4.6 A Second-Order Autoregressive Series

As a final example, we consider the spectrum of an AR(2) series of the form

$$x_t - \phi_1 x_{t-1} - \phi_2 x_{t-2} = w_t,$$

for the special case $\phi_1 = 1$ and $\phi_2 = -.9$. Recall Example 1.10 and Figure 1.9, which shows a sample realization of such a process for $\sigma_w = 1$. We note the data exhibit a strong periodic component that makes a cycle about every six points. First, computing the autocovariance function of the right side and equating it to the autocovariance on the left yields

$$\begin{aligned} \gamma_w(h) &= E[(x_{t+h} - \phi_1 x_{t+h-1} - \phi_2 x_{t+h-2})(x_t - \phi_1 x_{t-1} - \phi_2 x_{t-2})] \\ &= [1 + \phi_1^2 + \phi_2^2]\gamma_x(h) + (\phi_1\phi_2 - \phi_1)[\gamma_x(h+1) + \gamma_x(h-1)] \\ &\quad - \phi_2[\gamma_x(h+2) + \gamma_x(h-2)] \\ &= 2.81\gamma_x(h) - 1.90[\gamma_x(h+1) + \gamma_x(h-1)] \\ &\quad + .90[\gamma_x(h+2) + \gamma_x(h-2)], \end{aligned}$$

where we have substituted the values of $\phi_1 = 1$ and $\phi_2 = -.9$ in the equation. Now, substituting the spectral representation (4.12) for $\gamma_x(h)$ in the above equation yields

$$\begin{aligned} \gamma_w(h) &= \int_{-1/2}^{1/2} [2.81 - 1.90(\mathrm{e}^{2\pi i\omega} + \mathrm{e}^{-2\pi i\omega}) \\ &\qquad + .90(\mathrm{e}^{4\pi i\omega} + \mathrm{e}^{-4\pi i\omega})]\ \mathrm{e}^{2\pi i\omega h} f_x(\omega) d\omega \\ &= \int_{-1/2}^{1/2} [2.81 - 3.80\cos(2\pi\omega) + 1.80\cos(4\pi\omega)]\ \mathrm{e}^{2\pi i\omega h} f_x(\omega)\ d\omega. \end{aligned}$$

If the spectrum of the white noise process is $g_w(\omega)$, the uniqueness of the Fourier transform allows us to identify

$$g_w(\omega) = [2.81 - 3.80\cos(2\pi\omega) + 1.80\cos(4\pi\omega)]\ f_x(\omega).$$

But, as we have already seen, $g_w(\omega) = \sigma_w^2$, from which we deduce that

$$f_x(\omega) = \frac{\sigma_w^2}{2.81 - 3.80\cos(2\pi\omega) + 1.80\cos(4\pi\omega)}$$

is the spectrum of the autoregressive series. Setting $\sigma_w = 1$, Figure 4.3 displays $f_x(\omega)$ and shows a strong power component at about $\omega = .16$ cycles per point or a period between six and seven cycles per point and very little power at other frequencies. In this case, modifying the white noise series by applying the second-order AR operator has concentrated the power or variance of the resulting series in a very narrow frequency band.

The above examples have been given primarily to motivate the use of the power spectrum for describing the theoretical variance fluctuations of a stationary time series. Indeed, the interpretation of the spectral density function as the variance of the time series over a given frequency band gives us the intuitive explanation for its physical meaning. The plot of the function $f(\omega)$ over the frequency argument ω can even be thought of as an analysis of variance, in which the columns or block effects are the frequencies, indexed by ω.

4.4 Periodogram and Discrete Fourier Transform

We are now ready to tie together the periodogram, which is the sample-based concept presented in §4.2, with the spectral density, which is the population-based concept of §4.3.

Definition 4.2 *Given data $x_1, \ldots, x_n$, we define the* **discrete Fourier transform (DFT)** *to be*

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^{n} x_t e^{-2\pi i \omega_j t} \tag{4.19}$$

for $j = 0, 1, \ldots, n-1$, where the frequencies $\omega_j = j/n$ are called the **Fourier** *or* **fundamental frequencies**.

If n is a highly composite integer (i.e., it has many factors), the DFT can be computed by the fast Fourier transform (FFT) introduced in Cooley and Tukey (1965). Also, different packages scale the FFT differently, so it is a good idea to consult the documentation. R computes the DFT defined in (4.19) without the factor $n^{-1/2}$, but with an additional factor of $e^{2\pi i \omega_j}$ that

can be ignored because we will be interested in the squared modulus of the DFT. Sometimes it is helpful to exploit the inversion result for DFTs which shows the linear transformation is one-to-one. For the inverse DFT we have,

$$x_t = n^{-1/2} \sum_{j=0}^{n-1} d(\omega_j) \mathrm{e}^{2\pi i \omega_j t} \tag{4.20}$$

for $t = 1, \ldots, n$. The following example shows how to calculate the DFT and its inverse in R for the data set $\{1, 2, 3, 4\}$; note that R writes a complex number $z = a + ib$ as `a+bi`.

```
> x = 1:4
> dft = fft(x)/sqrt(4)
> dft
   [1]  5+0i -1+1i -1+0i -1-1i
> idft = fft(dft, inverse=T)/sqrt(4)
> idft
   [1] 1+0i 2+0i 3+0i 4+0i
```

We now define the periodogram as the squared modulus[3] of the DFT.

Definition 4.3 *Given data $x_1, \ldots, x_n$, we define the* **periodogram** *to be*

$$I(\omega_j) = |d(\omega_j)|^2 \tag{4.21}$$

for $j = 0, 1, 2, \ldots, n-1$.

Note that $I(0) = n\bar{x}^2$, where $\bar{x}$ is the sample mean. In addition, because $\sum_{t=1}^{n} \exp(-2\pi i \omega_j t) = 0$ for $j \neq 0$,[4] we can write the DFT as

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^{n} (x_t - \bar{x}) e^{-2\pi i \omega_j t} \tag{4.22}$$

for $j \neq 0$. Thus, for $j \neq 0$,

$$\begin{aligned} I(\omega_j) \;=\; |d(\omega_j)|^2 \;&=\; n^{-1} \sum_{t=1}^{n} \sum_{s=1}^{n} (x_t - \bar{x})(x_s - \bar{x}) e^{-2\pi i \omega_j (t-s)} \\ &=\; n^{-1} \sum_{h=-(n-1)}^{n-1} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}) e^{-2\pi i \omega_j h} \\ &=\; \sum_{h=-(n-1)}^{n-1} \widehat{\gamma}(h) e^{-2\pi i \omega_j h} \end{aligned} \tag{4.23}$$

[3] If $z = a + ib$ is a complex number, then $\bar{z} = a - ib$, and $|z|^2 = z\bar{z} = a^2 + b^2$.

[4] Note $\sum_{t=1}^{n} z^t = z \frac{1-z^n}{1-z}$ for $z \neq 1$.

where we have put $h = t - s$, with $\widehat{\gamma}(h)$ as given in (1.36).

Recall, $P(\omega_j) = (4/n)I(\omega_j)$ where $P(\omega_j)$ is the scaled periodogram defined in (4.7). Henceforth we will work with $I(\omega_j)$ instead of $P(\omega_j)$. Note that, in view of (4.23), $I(\omega_j)$ in (4.21) is the sample version of $f(\omega_j)$ given in (4.13). That is, we may think of the periodogram, $I(\omega_j)$, as the "sample spectral density" of x_t.

It is sometimes useful to work with the real and imaginary parts of the DFT individually. To this end, we define the following transforms.

Definition 4.4 *Given data $x_1, \ldots, x_n$, we define the* **cosine transform**

$$d_c(\omega_j) = n^{-1/2} \sum_{t=1}^{n} x_t \cos(2\pi\omega_j t) \tag{4.24}$$

and the **sine transform**

$$d_s(\omega_j) = n^{-1/2} \sum_{t=1}^{n} x_t \sin(2\pi\omega_j t) \tag{4.25}$$

where $\omega_j = j/n$ for $j = 0, 1, \ldots, n-1$.

We note that $d(\omega_j) = d_c(\omega_j) - i\, d_s(\omega_j)$ and hence

$$I(\omega_j) = d_c^2(\omega_j) + d_s^2(\omega_j). \tag{4.26}$$

We have also discussed the fact that spectral analysis can be thought of as an analysis of variance. The next example examines this notion.

Example 4.7 Spectral ANOVA

Let $x_1, \ldots, x_n$ be a sample of size n, where for ease, n is odd. Then, recalling Example 2.8 and the discussion around (4.8) and (4.9),

$$x_t = a_0 + \sum_{j=1}^{m} \left[a_j \cos(2\pi\omega_j t) + b_j \sin(2\pi\omega_j t)\right], \tag{4.27}$$

where $m = (n-1)/2$, is exact for $t = 1, \ldots, n$. In particular, using multiple regression formulas, we have $a_0 = \bar{x}$,

$$a_j = \frac{2}{n} \sum_{t=1}^{n} x_t \cos(2\pi\omega_j t) = \frac{2}{\sqrt{n}} d_c(\omega_j)$$

$$b_j = \frac{2}{n} \sum_{t=1}^{n} x_t \sin(2\pi\omega_j t) = \frac{2}{\sqrt{n}} d_s(\omega_j).$$

Hence, we may write

$$(x_t - \bar{x}) = \frac{2}{\sqrt{n}} \sum_{j=1}^{m} \left[d_c(\omega_j) \cos(2\pi\omega_j t) + d_s(\omega_j) \sin(2\pi\omega_j t)\right]$$

for $t = 1, \ldots, n$. Squaring both sides and summing we have[5]

$$\sum_{t=1}^{n}(x_t - \bar{x})^2 = 2\sum_{j=1}^{m}\left[d_c^2(\omega_j) + d_s^2(\omega_j)\right] = 2\sum_{j=1}^{m} I(\omega_j).$$

Thus, we have partitioned the sum of squares into harmonic components represented by frequency ω_j with the periodogram, $I(\omega_j)$, being the mean square regression. This leads to the ANOVA table:

Source	df	SS	MS
ω_1	2	$2I(\omega_1)$	$I(\omega_1)$
ω_2	2	$2I(\omega_2)$	$I(\omega_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
ω_m	2	$2I(\omega_m)$	$I(\omega_m)$
Total	$n-1$	$\sum_{t=1}^{n}(x_t - \bar{x})^2$	

This decomposition means that if the data contain some strong periodic components, then the periodogram values corresponding to those frequencies (or near those frequencies) will be large. On the other hand, the corresponding values of the periodogram will be small for periodic components not present in the data. The following is an R example to help explain this concept. We consider $n = 5$ observations given by $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 2, x_5 = 1$. Note that the data complete one cycle, but not in a sinusoidal way. Thus, we should expect the $\omega_1 = 1/5$ component to be relatively large but not exhaustive, and the $\omega_2 = 2/5$ component to be small.

```
> x = c(1,2,3,2,1)
> t = 1:5
> c1 = cos(2*pi*t*1/5)
> s1 = sin(2*pi*t*1/5)
> c2 = cos(2*pi*t*2/5)
> s2 = sin(2*pi*t*2/5)
> creg = lm(x~c1+s1+c2+s2)
> anova(creg) # partial output and combined ANOVA shown
                          #               ANOVA
            Df  Sum Sq #   Source  df        SS        MS
   c1        1 1.79443 #
   s1        1 0.94721 # freq=1/5   2  2.74164  1.37082
   c2        1 0.00557 #
   s2        1 0.05279 # freq=2/5   2  0.05836  0.02918
   Residuals 0 0.00000 #
```

[5]Recall $\sum_{t=1}^{n}\cos^2(2\pi\omega_j t) = \sum_{t=1}^{n}\sin^2(2\pi\omega_j t) = n/2$ for $j \neq 0$ or a multiple of n. Also $\sum_{t=1}^{n}\cos(2\pi\omega_j t)\sin(2\pi\omega_k t) = 0$ for any j and k.

```
> abs(fft(x))^2/5        # the periodogram (as a check)
   [1] 16.2000  1.3708  0.02918  0.02918  1.3708
>  #       I(0)  I(1/5)   I(2/5)   I(3/5)  I(4/5)
```

Note that $\bar{x} = 1.8$ so $I(0) = 5 \times 1.8^2 = 16.2$. Also, as a check

$$I(1/5) = [\texttt{SS(c1)} + \texttt{SS(s1)}]/2 = (1.79443 + .94721)/2 = 1.3708,$$

$$I(2/5) = [\texttt{SS(c2)} + \texttt{SS(s2)}]/2 = (.00557 + .05279)/2 = .02918,$$

and $I(j/5) = I(1 - j/5)$, for $j = 3, 4$. Finally, we note that the sum of squares associated with the residuals is zero, indicating an exact fit.

We are now ready to present some large sample properties of the periodogram. First, let μ be the mean of a stationary process x_t with absolutely summable autocovariance function $\gamma(h)$ and spectral density $f(\omega)$. We can use the same argument as in (4.23), replacing $\bar{x}$ by μ in (4.22), to write

$$I(\omega_j) = n^{-1} \sum_{h=-(n-1)}^{n-1} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \mu)(x_t - \mu) e^{-2\pi i \omega_j h} \tag{4.28}$$

where ω_j is a non-zero fundamental frequency. Taking expectation in (4.28) we obtain

$$E\left[I(\omega_j)\right] = \sum_{h=-(n-1)}^{n-1} \left(\frac{n - |h|}{n}\right) \gamma(h) e^{-2\pi i \omega_j h}. \tag{4.29}$$

For any given $\omega \neq 0$, choose a fundamental frequency $\omega_{j:n} \to \omega$ as $n \to \infty$,[6] from which it follows by (4.29) that

$$E\left[I(\omega_{j:n})\right] \to f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i h \omega} \tag{4.30}$$

as $n \to \infty$.[7] In other words, under absolute summability of $\gamma(h)$, the spectral density is the long-term average of the periodogram.

To examine the asymptotic distribution of the periodogram, we note that if x_t is a normal time series, the sine and cosine transforms will also be jointly normal, because they are linear combinations of the jointly normal random variables $x_1, x_2, \ldots, x_n$. In that case, the assumption that the covariance function satisfies the condition

$$\theta = \sum_{h=-\infty}^{\infty} |h||\gamma(h)| \; < \infty \tag{4.31}$$

[6] By this we mean $\omega_{j:n}$ is a frequency of the form j_n/n, where $\{j_n\}$ is a sequence of integers chosen so that $j_n/n \to \omega$ as $n \to \infty$.

[7] From Definition 4.3 we have $I(0) = n\bar{x}^2$, so the analogous result for the case $\omega = 0$ is $E[I(0)] - n\mu^2 = n \operatorname{var}(\bar{x}) \to f(0)$ as $n \to \infty$.

is enough to obtain simple large sample approximations for the variances and covariances. Using the same argument used to develop (4.29) we have

$$\mathrm{cov}[d_c(\omega_j), d_c(\omega_k)] = \sum_{s=1}^{n}\sum_{t=1}^{n} \gamma(s-t)\cos(2\pi\omega_j s)\cos(2\pi\omega_k t), \tag{4.32}$$

$$\mathrm{cov}[d_c(\omega_j), d_s(\omega_k)] = \sum_{s=1}^{n}\sum_{t=1}^{n} \gamma(s-t)\cos(2\pi\omega_j s)\sin(2\pi\omega_k t), \tag{4.33}$$

and

$$\mathrm{cov}[d_s(\omega_j), d_s(\omega_k)] = \sum_{s=1}^{n}\sum_{t=1}^{n} \gamma(s-t)\sin(2\pi\omega_j s)\sin(2\pi\omega_k t), \tag{4.34}$$

where the variance terms are obtained by setting $\omega_j = \omega_k$ in (4.32) and (4.34). In Appendix C, §C.2, we show the terms in (4.32)-(4.34) have interesting properties under assumption (4.31), namely, for $\omega_j, \omega_k \neq 0$ or $1/2$,

$$\mathrm{cov}[d_c(\omega_j), d_c(\omega_k)] = \begin{cases} f(\omega_j)/2 + \epsilon_n, & \omega_j = \omega_k \\ \epsilon_n, & \omega_j \neq \omega_k \end{cases} \tag{4.35}$$

$$\mathrm{cov}[d_s(\omega_j), d_s(\omega_k)] = \begin{cases} f(\omega_j)/2 + \epsilon_n, & \omega_j = \omega_k \\ \epsilon_n, & \omega_j \neq \omega_k \end{cases} \tag{4.36}$$

and

$$\mathrm{cov}[d_c(\omega_j), d_s(\omega_k)] = \epsilon_n, \tag{4.37}$$

where the error term ϵ_n in the approximations can be bounded,

$$|\epsilon_n| \leq \theta/n, \tag{4.38}$$

and θ is given by (4.31). If $\omega_j = \omega_k = 0$ or $1/2$ in (4.35), the multiplier $1/2$ disappears; note that $d_s(0) = d_s(1/2) = 0$, so (4.36) does not apply.

Example 4.8 Covariance of Sines and Cosines for an MA Process

For the three-point moving average series of Example 4.5, the theoretical spectrum is shown in Figure 4.3. For $n = 256$ points, the theoretical covariance matrix of the vector

$$\boldsymbol{d} = (d_c(\omega_{26}), d_s(\omega_{26}), d_c(\omega_{27}), d_s(\omega_{27}))'$$

is

$$\mathrm{cov}(\boldsymbol{d}) = \begin{pmatrix} .3752 & -.0009 & -.0022 & -.0010 \\ -.0009 & .3777 & -.0009 & .0003 \\ -.0022 & -.0009 & .3667 & -.0010 \\ -.0010 & .0003 & -.0010 & .3692 \end{pmatrix}.$$

The diagonal elements can be compared with the theoretical spectral values of .7548 for the spectrum at frequency $\omega_{26} = .102$, and of .7378 for the spectrum at $\omega_{27} = .105$. Hence, the cosine and sine transforms produce nearly uncorrelated variables with variances approximately equal to one half of the theoretical spectrum. For this particular case, the uniform bound is determined from $\theta = 8/9$, yielding $|\epsilon_{256}| \leq .0035$ for the bound on the approximation error.

If $x_t \sim \text{iid}(0, \sigma^2)$, then it follows from (4.31)-(4.37) and the central limit theorem[8] that

$$d_c(\omega_{j:n}) \sim \text{AN}(0, \sigma^2/2) \quad \text{and} \quad d_s(\omega_{j:n}) \sim \text{AN}(0, \sigma^2/2) \tag{4.39}$$

jointly and independently, and independent of $d_c(\omega_{k:n})$ and $d_s(\omega_{k:n})$ provided $\omega_{j:n} \to \omega_1$ and $\omega_{k:n} \to \omega_2$ where $0 < \omega_1 \neq \omega_2 < 1/2$. We note that in this case, $f(\omega) = \sigma^2$. In view of (4.39), it follows immediately that as $n \to \infty$,

$$\frac{2I(\omega_{j:n})}{\sigma^2} \xrightarrow{d} \chi^2_2 \quad \text{and} \quad \frac{2I(\omega_{k:n})}{\sigma^2} \xrightarrow{d} \chi^2_2 \tag{4.40}$$

with $I(\omega_{j:n})$ and $I(\omega_{k:n})$ being asymptotically independent, where χ^2_ν denotes a chi-squared random variable with ν degrees of freedom.

Using the central limit theory of §C.2, it is fairly easy to extend the results of the iid case to the case of a linear process.

Property P4.2: Distribution of the Periodogram Ordinates
If

$$x_t = \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}, \qquad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty \tag{4.41}$$

where $w_t \sim iid(0, \sigma_w^2)$, and (4.31) holds, then for any collection of m distinct frequencies ω_j with $\omega_{j:n} \to \omega_j$

$$\frac{2I(\omega_{j:n})}{f(\omega_j)} \xrightarrow{d} \text{iid } \chi^2_2 \tag{4.42}$$

provided $f(\omega_j) > 0$, for $j = 1, \ldots, m$.

This result is stated more precisely in Theorem C.7 of §C.3. Other approaches to large sample normality of the periodogram ordinates are in terms of cumulants, as in Brillinger (1981), or in terms of mixing conditions, such as in Rosenblatt (1956). We adopt the approach here used by Hannan (1970), Fuller (1995), and Brockwell and Davis (1991).

[8] If $Y_j \sim \text{iid}(0, \sigma^2)$ and $\{a_j\}$ are constants for which $\sum_{j=1}^n a_j^2 / \max_{1 \leq j \leq n} a_j^2 \to \infty$ as $n \to \infty$, then $\sum_{j=1}^n a_j Y_j \sim \text{AN}\left(0, \sigma^2 \sum_{j=1}^n a_j^2\right)$; the notation AN is explained in Definition A.5.

The distributional result (4.42) can be used to derive an approximate confidence interval for the spectrum in the usual way. Let $\chi^2_\nu(\alpha)$ denote the lower α probability tail for the chi-squared distribution with ν degrees of freedom; that is,

$$\Pr\{\chi^2_\nu \le \chi^2_\nu(\alpha)\} = \alpha. \tag{4.43}$$

Then, an approximate $100(1-\alpha)\%$ confidence interval for the spectral density function would be of the form

$$\frac{2\ I(\omega_{j:n})}{\chi^2_2(1-\alpha/2)} \le f(\omega) \le \frac{2\ I(\omega_{j:n})}{\chi^2_2(\alpha/2)} \tag{4.44}$$

Often, nonstationary trends are present that should be eliminated before computing the periodogram. Trends introduce extremely low frequency components in the periodogram that tend to obscure the appearance at higher frequencies. For this reason, it is usually conventional to center the data prior to a spectral analysis using either mean-adjusted data of the form $x_t - \bar{x}$ to eliminate the zero or d-c component or to use detrended data of the form $x_t - \widehat{\beta}_1 - \widehat{\beta}_2 t$ to eliminate the term that will be considered a half cycle by the spectral analysis. Note that higher order polynomial regressions in t or nonparametric smoothing (linear filtering) could be used in cases where the trend is nonlinear.

As previously indicated, it is often convenient to calculate the DFTs, and hence the periodogram, using the fast Fourier transform algorithm. The FFT utilizes a number of redundancies in the calculation of the DFT when n is highly composite; that is, an integer with many factors of $2, 3$, or 5, the best case being when $n = 2^p$ is a factor of 2. Details may be found in Cooley and Tukey (1965). To accommodate this property, we can pad the centered (or detrended) data of length n to the next highly composite integer n' by adding zeros, i.e., setting $x^c_{n+1} = x^c_{n+2} = \cdots = x^c_{n'} = 0$, where x^c_t denotes the centered data. This means that the fundamental frequency ordinates will be $\omega_j = j/n'$ instead of j/n. We illustrate by considering the periodogram of the SOI and Recruitment series, as has been given in Figure 1.5 of Chapter 1. Recall that the series are monthly series and $n = 453$. To find n' in R, use the command `nextn(453)` to see that $n' = 480$ will be used in the spectral analyses by default (use `help(spec.pgram)` to see how to override this default).

Example 4.9 Periodogram of SOI and Recruitment Series

Figure 4.4 shows the periodograms of each series. As previously indicated, the centered data have been padded to a series of length 480. We notice a narrow band peak at the obvious yearly cycle, $\omega = 1/12$. In addition, there is considerable amount of power in a wide band at the lower frequencies that is centered around the four-year cycle $\omega = 1/48$ representing a possible El Niño effect. This wide band activity suggests that the possible El Niño cycle is irregular, but tends to be around four

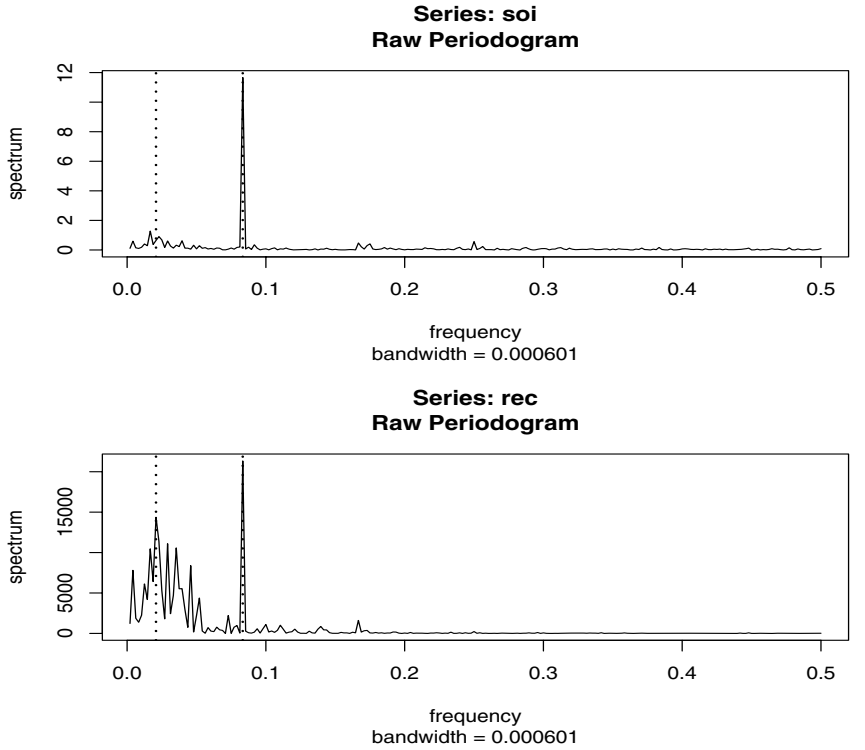


Figure 4.4 Periodogram of SOI and Recruitment, $n = 453$ ($n' = 480$), showing common peaks at $\omega = 1/12 = .083$ and $\omega = 1/48 = .021$ cycles/month.

years on average. We will continue to address this problem as we move to more sophisticated analyses.

Noting $\chi^2_2(.025) = .05$ and $\chi^2_2(.975) = 7.38$, we can obtain approximate 95% confidence intervals for the frequencies of interest. For example, the periodogram of the SOI series is $I_S(1/12) = 11.64$ at the yearly cycle. An approximate 95% confidence interval for the spectrum $f_S(1/12)$ is then

$$[2(11.67)/7.38,\ 2(11.67)/.05] = [3.16,\ 460.81],$$

which is too wide to be of much use. We do notice, however, that the lower value of 3.16 is higher than any other periodogram ordinate, so it is safe to say that this value is significant. On the other hand, an approximate 95% confidence interval for the spectrum at the four-year cycle, $f_S(1/48)$, is

$$[2(.64)/7.38,\ 2(.64)/.05] = [.17,\ 25.47],$$

which again is extremely wide, and with which we are unable to establish

significance of the peak.

We now give the R commands that can be used to reproduce Figure 4.4. To calculate and graph the periodogram, we used the `spec.pgram` command in R. We have set `log="no"` because R will plot the periodogram on a $\log_{10}$ scale by default. Figure 4.4 displays a `bandwidth` and by default, R tapers the data (which we override in the commands below). We will discuss bandwidth and tapering in the next section, so ignore these concepts for the time being.

```
> soi = scan("/mydata/soi.dat")
> rec = scan("/mydata/rec.dat")
> par(mfrow=c(2,1))
> soi.per = spec.pgram(soi, taper=0, log="no")
> abline(v=1/12, lty="dotted")
> abline(v=1/48, lty="dotted")
> rec.per = spec.pgram(rec, taper=0, log="no")
> abline(v=1/12, lty="dotted")
> abline(v=1/48, lty="dotted")
```

The confidence intervals for the SOI series at the yearly cycle, $\omega = 1/12 = 40/480$, and the possible El Niño cycle of four years $\omega = 1/48 = 10/480$ can be computed in R as follows:

```
> soi.per$spec[40] # soi pgram at freq 1/12 = 40/480
   [1] 11.66677
> soi.per$spec[10] # soi pgram at freq 1/48 = 10/480
   [1] 0.6447554
> # -- conf intervals -- # returned value:
> U = qchisq(.025,2)     # 0.05063562
> L = qchisq(.975,2)     # 7.377759
> 2*soi.per$spec[10]/L   # 0.1747835
> 2*soi.per$spec[10]/U   # 25.46648
> 2*soi.per$spec[40]/L   # 3.162688
> 2*soi.per$spec[40]/U   # 460.813
> #-- replace soi with rec above to get recruit values
```

The example above makes it fairly clear the periodogram as an estimator is susceptible to large uncertainties, and we need to find a way to reduce the variance. Not surprisingly, this result follows if we think about the periodogram, $I(\omega_j)$ as an estimator of the spectral density $f(\omega)$ and realize that it is the sum of squares of only two random variables for any sample size. The solution to this dilemma is suggested by the analogy with classical statistics where we look for independent random variables with the same variance and average the squares of these common variance observations. Independence and equality of variance do not hold in the time series case, but the covariance structure of the two adjacent estimators given in Example 4.8 suggests that for neighboring frequencies, these assumptions are approximately true.

4.5 Nonparametric Spectral Estimation

To continue the discussion that ended the previous section, we define a frequency band, $\mathcal{B}$, of $L << n$ contiguous fundamental frequencies centered around $\omega_j = j/n$ that are close to the frequency of interest, ω, as

$$\mathcal{B} = \left\{\omega : \ \omega_j - \frac{m}{n} \leq \omega \leq \omega_j + \frac{m}{n}\right\}, \tag{4.45}$$

where

$$L = 2m + 1 \tag{4.46}$$

is an odd number, chosen such that the spectral values in the interval $\mathcal{B}$,

$$f(\omega_j + k/n), \quad k = -m, \ldots, 0, \ldots, m$$

are approximately equal to $f(\omega)$. This structure can be realized for large sample sizes, as shown formally in §C.2. Values of the spectrum in this band should be relatively constant, as well, for the smoothed spectra defined below to be good estimators.

Using the above band, we may now define an averaged or smoothed periodogram as the average of the periodogram values, say,

$$\bar{f}(\omega) = \frac{1}{L}\sum_{k=-m}^{m} I(\omega_j + k/n), \tag{4.47}$$

as the average over the band $\mathcal{B}$.

Under the assumption that the spectral density is fairly constant in the band $\mathcal{B}$, and in view of (4.42) we can show that under appropriate conditions,[9] for large n, the periodograms in (4.47) are approximately distributed as independent $f(\omega)\chi^2_2/2$ random variables, for $0 < \omega < 1/2$, as long as we keep L fairly small relative to n. This result is discussed formally in §C.2. Thus, under these conditions, $L\bar{f}(\omega)$ is the sum of L approximately independent $f(\omega)\chi^2_2/2$ random variables. It follows that, for large n,

$$\frac{2L\bar{f}(\omega)}{f(\omega)} \dot{\sim} \chi^2_{2L} \tag{4.48}$$

where $\dot{\sim}$ means *approximately distributed as*.

In this scenario, it seems reasonable to call the length of the interval defined by (4.45),

$$B_w = \frac{L}{n} \tag{4.49}$$

the bandwidth. Bandwidth, of course, refers to the width of the frequencey band used in smoothing the periodogram. The concept of the bandwidth, however, becomes more complicated with the introduction of spectral estimators

[9]The conditions, which are sufficient, are that x_t is a linear process, as described in Property P4.2, with $\sum_{j>0}\sqrt{j}\,|\psi_j| < \infty$, and w_t has a finite fourth moment.

that smooth with unequal weights. Note (4.49) implies the degrees of freedom can be expressed as

$$2L = 2B_w n, \tag{4.50}$$

or twice the time-bandwidth product. The result (4.48) can be rearranged to obtain an approximate $100(1-\alpha)\%$ confidence interval of the form

$$\frac{2L\bar{f}(\omega)}{\chi^2_{2L}(1-\alpha/2)} \leq f(\omega) \leq \frac{2L\bar{f}(\omega)}{\chi^2_{2L}(\alpha/2)} \tag{4.51}$$

for the true spectrum, $f(\omega)$.

Many times, the visual impact of a spectral density plot will be improved by plotting the logarithm of the spectrum instead of the spectrum.[10] This phenomenon can occur when regions of the spectrum exist with peaks of interest much smaller than some of the main power components. For the log spectrum, we obtain an interval of the form

$$\left[\ln \bar{f}(\omega) + \ln 2L - \ln \chi^2_{2L}(1-\alpha/2),\ \ln \bar{f}(\omega) + \ln 2L - \ln \chi^2_{2L}(\alpha/2)\right]. \tag{4.52}$$

We can also test hypotheses relating to the equality of spectra using the fact that the distributional result (4.48) implies that the ratio of spectra based on roughly independent samples will have an approximate $F_{2L,2L}$ distribution. The independent estimators can either be from different frequency bands or from different series.

If zeros are appended before computing the spectral estimators, we need to adjust the degrees of freedom and an approximation is to replace $2L$ by $2Ln/n'$. Hence, we define the adjusted degrees of freedom as

$$df = \frac{2Ln}{n'} \tag{4.53}$$

and use it instead of $2L$ in the confidence intervals (4.51) and (4.52). For example, (4.51) becomes

$$\frac{df\,\bar{f}(\omega)}{\chi^2_{df}(1-\alpha/2)} \leq f(\omega) \leq \frac{df\,\bar{f}(\omega)}{\chi^2_{df}(\alpha/2)}. \tag{4.54}$$

A number of assumptions are made in computing the approximate confidence intervals given above, which may not hold in practice. In such cases, it may be reasonable to employ resampling techniques such as one of the parametric bootstraps proposed by Hurvich and Zeger (1987) or a nonparametric local bootstrap proposed by Paparoditis and Politis (1999). To develop the bootstrap distributions, we assume that the contiguous DFTs in a frequency band of the form (4.45) all came from a time series with identical spectrum $f(\omega)$. This, in fact, is exactly the same assumption made in deriving the large-sample theory. We may then simply resample the L DFTs in the band, with

[10] The log transformation is the variance stabilizing transformation in this situation.

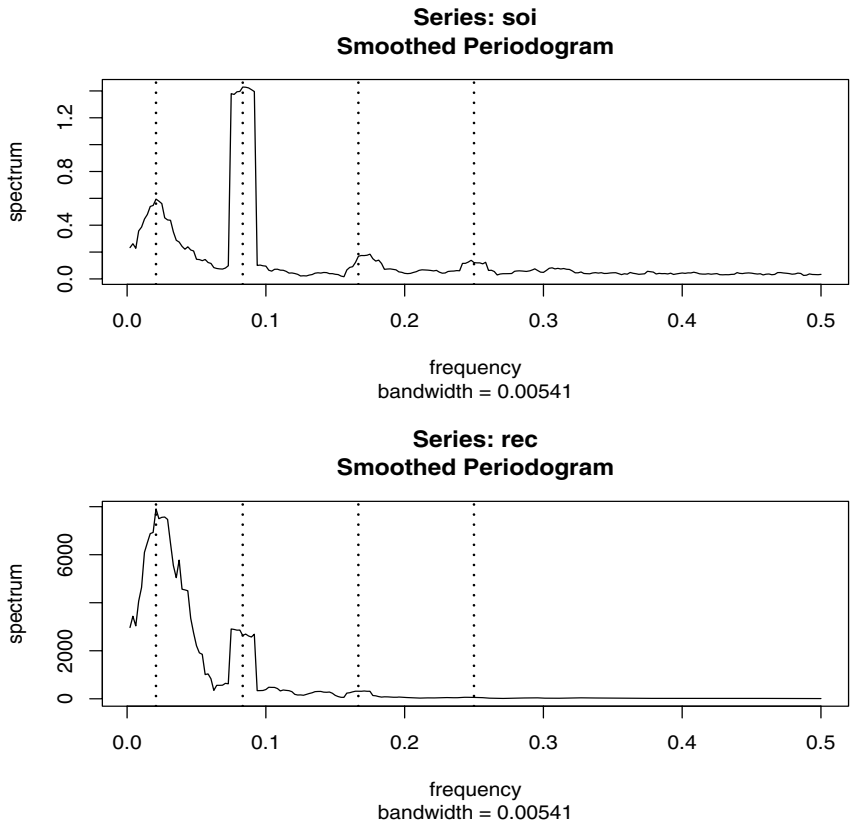


Figure 4.5 The averaged periodogram of the SOI and Recruitment series $n = 453$, $n' = 480$, $L = 9$, $df = 17$, showing common peaks at the four year period, $\omega = 1/48 = .021$ cycles/month, the yearly period, $\omega = 1/12 = .083$ cycles/month and some of its harmonics $\omega = k/12$ for $k = 2, 3$.

replacement, calculating a spectral estimate from each bootstrap sample. The sampling distribution of the bootstrap estimators approximates the distribution of the nonparametric spectral estimator. For further details, including the theoretical properties of such estimators, see Paparoditis and Politis (1999).

Before proceeding further, we pause to consider computing the average periodograms for the SOI and Recruitment series, as shown in Figure 4.5.

Example 4.10 Averaged Periodogram of SOI and Recruitment Series

Generally, it is a good idea to try several bandwidths that seem to be compatible with the general overall shape of the spectrum, as suggested by the periodogram. The SOI and Recruitment series periodograms,

previously computed in Figure 4.4, suggest the power in the lower El Niño frequency needs smoothing to identify the predominant overall period. Trying values of L leads to the choice $L = 9$ as a reasonable value, and the result is displayed in Figure 4.5. In our notation, the bandwidth in this case is $B_w = 9/480 = .01875$ cycles per month for the spectral estimator. This bandwidth means we are assuming a relatively constant spectrum over about $.01875/.5 = 3.75\%$ of the entire frequency interval $(0, 1/2)$. The bandwidth reported in R is taken from Bloomfield (2000), and in the current case amounts to dividing (4.49) by $\sqrt{12}$. An excellent discussion of the concept of bandwidth may be found in Percival and Walden (1993, §6.7). To obtain the bandwidth, $B_w = .01875$, from the one reported by R in Figure 4.5, we can multiply .00541 by $\sqrt{12}$.

The smoothed spectra shown in Figure 4.5 provide a sensible compromise between the noisy version, shown in Figure 4.4, and a more heavily smoothed spectrum, which might lose some of the peaks. An undesirable effect of averaging can be noticed at the yearly cycle, $\omega = 1/12$, where the narrow band peaks that appeared in the periodograms in Figure 4.4 have been flattened and spread out to nearby frequencies. We also notice, and have marked, the appearance of harmonics of the yearly cycle, that is, frequencies of the form $\omega = k/12$ for $k = 1, 2, \ldots$. Harmonics typically occur when a periodic component is present, but not in a sinusoidal fashion.

Figure 4.5 can be reproduced in R using the following commands. The basic call is to the function `spec.pgram`. To compute averaged periodograms, use the Daniell kernel, and specify m, where $L = 2m + 1$ ($L = 9$ and $m = 4$ in this example). We will explain the kernel concept later in this section, specifically just prior to Example 4.11.

```
> par(mfrow=c(2,1))
> k = kernel("daniell",4)
> soi.ave = spec.pgram(soi, k, taper=0, log="no")
> abline(v=1/12, lty="dotted")
> abline(v=2/12, lty="dotted")
> abline(v=3/12, lty="dotted")
> abline(v=1/48, lty="dotted")
> #-- Repeat 5 lines above using rec in place of soi
> soi.ave$bandwidth           # reported bandwidth
   [1] 0.005412659
> soi.ave$bandwidth*sqrt(12)  # Bw
   [1] 0.01875
```

The adjusted degrees of freedom are $df = 2(9)(453)/480 \approx 17$. We can use this value for the 95% confidence intervals, with $\chi^2_{df}(.025) = 7.56$ and $\chi^2_{df}(.975) = 30.17$. Substituting into (4.54) gives the intervals in Table 4.1 for the two frequency bands identified as having the maximum

Table 4.1 Confidence Intervals for the Spectra of the SOI and Recruitment Series

Series	ω	Period	Power	Lower	Upper
SOI	1/48	4 years	.59	.33	1.34
	1/12	1 year	1.43	.80	3.21
Recruits	1/48	4 years	7.91	4.45	17.78
$\times 10^3$	1/12	1 year	2.63	1.48	5.92

power. To examine the two peak power possibilities, we may look at the 95% confidence intervals and see whether the lower limits are substantially larger than adjacent baseline spectral levels. For example, the El Niño frequency of 48 months has lower limits that exceed the values the spectrum would have if there were simply a smooth underlying spectral function without the peaks. The relative distribution of power over frequencies is different, with the SOI index having less power at the lower frequency, relative to the seasonal periods, and the recruit series having relatively more power at the lower or El Niño frequency.

The entries in Table 4.1 for SOI can be obtained in R as follows:

```
> df = soi.ave$df      # df = 16.9875  (returned values)
> U = qchisq(.025,df)  # U = 7.555916
> L = qchisq(.975,df)  # L = 30.17425
> soi.ave$spec[10]     # 0.5942431
> soi.ave$spec[40]     # 1.428959
> # -- intervals --
> df*soi.ave$spec[10]/L   # 0.334547
> df*soi.ave$spec[10]/U   # 1.336000
> df*soi.ave$spec[40]/L   # 0.8044755
> df*soi.ave$spec[40]/U   # 3.212641
> #-- repeat above commands with soi replaced by rec
```

Finally, Figure 4.6 shows the averaged periodograms in Figure 4.5 plotted on a $\log_{10}$ scale. This is the default plot in R, and these graphs can be obtained by removing the statement `log="no"` in the `spec.pgram` call. Notice that the default plot also shows a generic confidence interval of the form (4.52) (with ln replaced by $\log_{10}$) in the upper right-hand corner. To use it, imagine placing the tick mark on the averaged periodogram ordinate of interest; the resulting bar then constitutes an approximate 95% confidence interval for the spectrum at that frequency. Of course, actual intervals may be computed as was done in this example. We note that displaying the estimates on a log scale tends to emphasize the harmonic components.

This example points out the necessity for having some relatively systematic procedure for deciding whether peaks are significant. The question of deciding

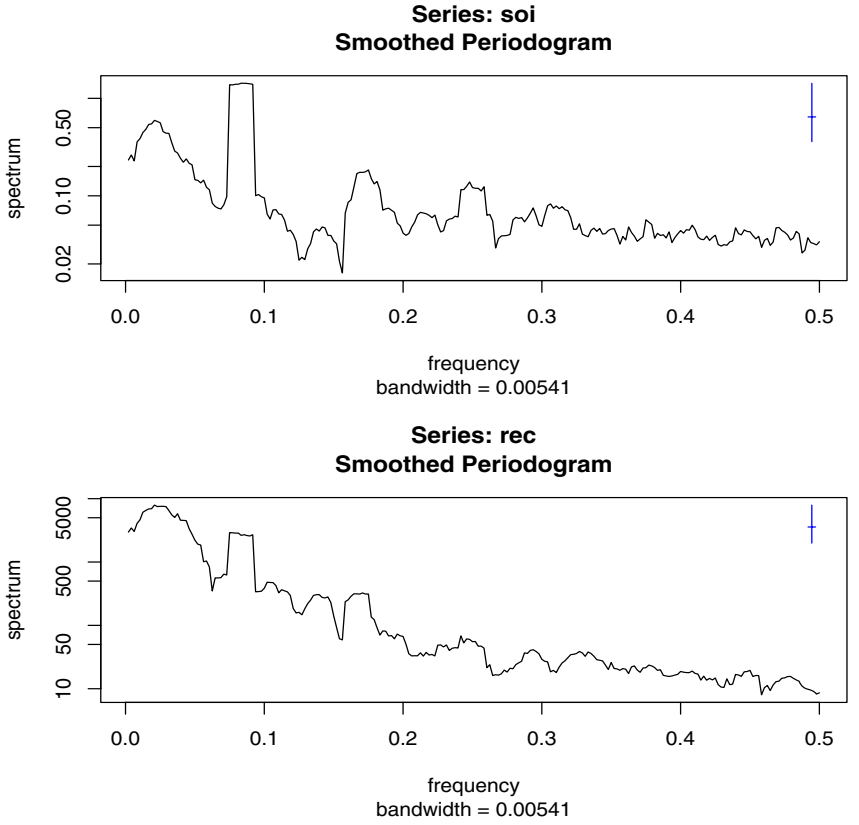


Figure 4.6 Figure 4.5 with the average periodogram ordinates plotted on a $\log_{10}$ scale. The display in the upper right-hand corner represents a generic 95% confidence interval.

whether a single peak is significant usually rests on establishing what we might think of as a baseline level for the spectrum, defined rather loosely as the shape that one would expect to see if no spectral peaks were present. This profile can usually be guessed by looking at the overall shape of the spectrum that includes the peaks; usually, a kind of baseline level will be apparent, with the peaks seeming to emerge from this baseline level. If the lower confidence limit for the spectral value is still greater than the baseline level at some predetermined level of significance, we may claim that frequency value as a statistically significant peak. To maintain an α that is consistent with our stated indifference to the upper limits, we might use a one-sided confidence interval.

An important aspect of interpreting the significance of confidence intervals and tests involving spectra is that typically, more than one frequency will be of interest, so that we will potentially be interested in simultaneous statements

about a whole collection of frequencies. For example, it would be unfair to claim in Table 4.1 the two frequencies of interest as being statistically significant and all other potential candidates as nonsignificant at the overall level of $\alpha = .05$. In this case, we follow the usual statistical approach, noting that if K statements $S_1, S_1, \ldots, S_k$ are made at significance level α, i.e., $P\{S_k\} = 1 - \alpha$, then the overall probability all statements are true satisfies the Bonferroni inequality

$$P\{\text{all } S_k \text{ true}\} \geq 1 - K\alpha. \tag{4.55}$$

For this reason, it is desirable to set the significance level for testing each frequency at α/K if there are K potential frequencies of interest. If, *a priori*, potentially $K = 10$ frequencies are of interest, setting $\alpha = .01$ would give an overall significance level of bound of .10.

The use of the confidence intervals and the necessity for smoothing requires that we make a decision about the bandwidth B_w over which the spectrum will be essentially constant. Taking too broad a band will tend to smooth out valid peaks in the data when the constant variance assumption is not met over the band. Taking too narrow a band will lead to confidence intervals so wide that peaks are no longer statistically significant. Thus, we note that there is a conflict here between variance properties or bandwidth stability, which can be improved by increasing B_w and resolution, which can be improved by decreasing B_w. A common approach is to try a number of different bandwidths and to look qualitatively at the spectral estimators for each case.

To address the problem of resolution, it should be evident that the flattening of the peaks in Figures 4.5 and 4.6 was due to the fact that simple averaging was used in computing $\bar{f}(\omega)$ defined in (4.47). There is no particular reason to use simple averaging, and we might improve the estimator by employing a weighted average, say

$$\widehat{f}(\omega) = \sum_{k=-m}^{m} h_k \, I(\omega_j + k/n), \tag{4.56}$$

using the same definitions as in (4.47) but where now, the weights satisfy

$$h_{-k} = h_k > 0 \quad \text{all } k \quad \text{and} \quad \sum_{k=-m}^{m} h_k = 1.$$

In particular, it seems reasonable that the resolution of the estimator will improve if we use weights that decrease as distance from the center weight h_0 increases; we will return to this idea shortly. To obtain the averaged periodogram, $\bar{f}(\omega)$, in (4.56), set $h_k = L^{-1}$, for all k, where $L = 2m + 1$. The asymptotic theory established for $\bar{f}(\omega)$ still holds for $\widehat{f}(\omega)$ provided that the weights satisfy the additional condition that if $m \to \infty$ as $n \to \infty$ but $m/n \to 0$, then

$$\sum_{k=-m}^{m} h_k^2 \to 0.$$

Under these conditions, as $n \to \infty$,

(i) $E\left(\widehat{f}(\omega)\right) \to f(\omega)$

(ii) $\left(\sum_{k=-m}^{m} h_k^2\right)^{-1} \mathrm{cov}\left(\widehat{f}(\omega), \widehat{f}(\lambda)\right) \to f^2(\omega) \quad \text{for } \omega = \lambda \neq 0, 1/2.$

In (ii), replace $f^2(\omega)$ by 0 if $\omega \neq \lambda$ and by $2f^2(\omega)$ if $\omega = \lambda = 0$ or $1/2$.

We have already seen these results in the case of $\bar{f}(\omega)$, where the weights are constant, $h_k = L^{-1}$, in which case $\sum_{k=-m}^{m} h_k^2 = L^{-1}$. The distributional properties of (4.56) are more difficult now because $\widehat{f}(\omega)$ is a weighted linear combination of asymptotically independent χ^2 random variables. An approximation that seems to work well is to replace L by $\left(\sum_{k=-m}^{m} h_k^2\right)^{-1}$. That is, define

$$L_h = \left(\sum_{k=-m}^{m} h_k^2\right)^{-1} \tag{4.57}$$

and use the approximation[11]

$$\frac{2L_h \widehat{f}(\omega)}{f(\omega)} \dot{\sim} \chi^2_{2L_h}. \tag{4.58}$$

In analogy to (4.49), we will define the bandwidth in this case to be

$$B_w = \frac{L_h}{n}. \tag{4.59}$$

Using the approximation (4.58) we obtain an approximate $100(1-\alpha)\%$ confidence interval of the form

$$\frac{2L_h \widehat{f}(\omega)}{\chi^2_{2L_h}(1-\alpha/2)} \leq f(\omega) \leq \frac{2L_h \widehat{f}(\omega)}{\chi^2_{2L_h}(\alpha/2)} \tag{4.60}$$

for the true spectrum, $f(\omega)$. If the data are padded to n', then replace $2L_h$ in (4.60) with $df = 2L_h n/n'$ as in (4.53).

An easy way to generate the weights in R is by repeated use of the Daniell kernel. For example, with $m = 1$ and $L = 2m + 1 = 3$, the Daniell kernel has weights $\{h_k\} = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$; applying this kernel to a sequence of numbers, $\{u_t\}$, produces

$$\widehat{u}_t = \frac{1}{3}u_{t-1} + \frac{1}{3}u_t + \frac{1}{3}u_{t+1}.$$

We can apply the same kernel again to the $\widehat{u}_t$,

$$\widehat{\widehat{u}}_t = \frac{1}{3}\widehat{u}_{t-1} + \frac{1}{3}\widehat{u}_t + \frac{1}{3}\widehat{u}_{t+1},$$

[11]The approximation proceeds as follows: If $\widehat{f} \dot{\sim} c\chi^2_\nu$, where c is a constant, then $E\widehat{f} \approx c\nu$ and $\mathrm{var}\widehat{f} \approx f^2 \sum_k h_k^2 \approx c^2 2\nu$. Solving, $c \approx f \sum_k h_k^2/2 = f/2L_h$ and $\nu \approx 2\left(\sum_k h_k^2\right)^{-1} = 2L_h$.

which simplifies to

$$\widehat{\widehat{u}}_t = \frac{1}{9}u_{t-2} + \frac{2}{9}u_{t-1} + \frac{3}{9}u_t + \frac{2}{9}u_{t+1} + \frac{1}{9}u_{t+2}.$$

The modified Daniell kernel puts half weights at the end points, so with $m = 1$ the weights are $\{h_k\} = \{\frac{1}{4}, \frac{2}{4}, \frac{1}{4}\}$ and

$$\widehat{u}_t = \frac{1}{4}u_{t-1} + \frac{1}{2}u_t + \frac{1}{4}u_{t+1}.$$

Applying the same kernel again yields

$$\widehat{\widehat{u}}_t = \frac{1}{16}u_{t-2} + \frac{4}{16}u_{t-1} + \frac{6}{16}u_t + \frac{4}{16}u_{t+1} + \frac{1}{16}u_{t+2}.$$

These coefficients can be obtained in R by issuing the `kernel` command. For example, `kernel("modified.daniell",c(1,1))` would produce the coefficients of the last example. It is also possible to use different values of m, e.g., try `kernel("modified.daniell",c(1,2))` or `kernel("daniell",c(1,2))`. The other kernels that are currently available in R are the Dirichlet kernel and the Fejér kernel, which we will discuss shortly.

Example 4.11 Smoothed Periodogram of the SOI and Recruitment Series

In this example, we estimate the spectra of the SOI and Recruitment series using the smoothed periodogram estimate in (4.56). We used a modified Daniell kernel twice, with $m = 3$ both times. This yields $L_h = 1/\sum_{k=-m}^{m} h_k^2 = 9.232$, which is close to the value of $L = 9$ used in Example 4.10. In this case, the bandwidth is $B_w = 9.232/480 = .019$ and the modified degrees of freedom is $df = 2L_h 453/480 = 17.43$. The weights, h_k, can be obtained in R as follows:

```
> kernel("modified.daniell", c(3,3))
coef[-6] = 0.006944  # = coef[ 6]
coef[-5] = 0.027778  # = coef[ 5]
coef[-4] = 0.055556  # = coef[ 4]
coef[-3] = 0.083333  # = coef[ 3]
coef[-2] = 0.111111  # = coef[ 2]
coef[-1] = 0.138889  # = coef[ 1]
coef[ 0] = 0.152778
```

The resulting spectral estimates can be viewed in Figure 4.7 and we notice that the estimates more appealing than those in Figure 4.5. Figure 4.7 was generated in R as follows; we also show how to obtain *df* and B_w.

```
> par(mfrow=c(2,1))
> k = kernel("modified.daniell", c(3,3))
> soi.smo = spec.pgram(soi, k, taper=0, log="no")
```

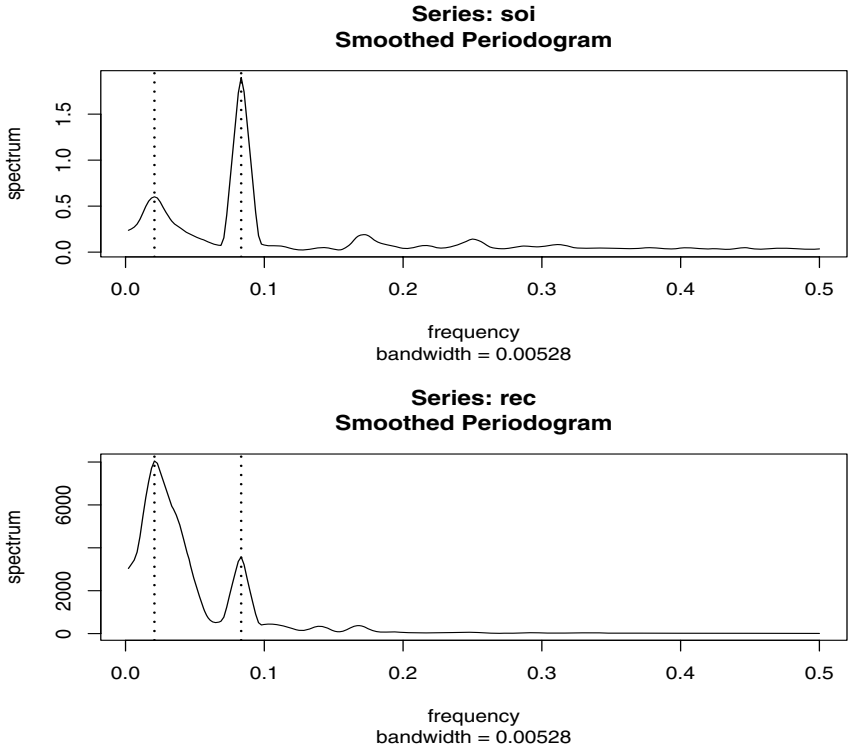


Figure 4.7 Smoothed spectral estimates of the SOI and Recruitment series; see Example 4.11 for details.

```
> abline(v=1/12, lty="dotted")
> abline(v=1/48, lty="dotted")
> #-- Repeat above 3 lines with rec replacing soi
> df = soi.smo2$df             # df=17.42618
> Lh = 1/sum(k[-k$m:k$m]^2)    # Lh=9.232413
> Bw = Lh/480                  # Bw=0.01923419
```

The bandwidth reported by R is .00528, which is approximately $B_w/\sqrt{12}$; type `bandwidth.kernel` to see how R computes bandwidth. Reissuing the `spec.pgram` commands with `log="no"` removed will result in a figure similar to Figure 4.6. Finally, we mention that R uses the modified Daniell kernel by default. For example, an easier way to obtain `soi.smo` is to issue the command:

```
> soi.smo = spectrum(soi, spans=c(7,7), taper=0)
```

Notice that `spans` is a vector of odd integers, given in terms of $L = 2m+1$ instead of m. These values give the widths of modified Daniell smoother to be used to smooth the periodogram.

We are now ready to introduce the concept of *tapering*; this will lead us to the notion of a spectral window. For example, suppose x_t is a mean-zero, stationary process with spectral density $f_x(\omega)$. If we replace the original series by the tapered series

$$y_t = h_t x_t, \tag{4.61}$$

for $t = 1, 2, \ldots, n$, and use the modified DFT

$$d_y(\omega_j) = n^{-1/2} \sum_{t=1}^{n} h_t x_t \mathrm{e}^{-2\pi i \omega_j t}, \tag{4.62}$$

and let $I_y(\omega_j) = |d_y(\omega_j)|^2$, we obtain (see Problem 4.15)

$$E[I_y(\omega_j)] = \int_{-1/2}^{1/2} W_n(\omega_j - \omega)\ f_x(\omega)\ d\omega \tag{4.63}$$

where

$$W_n(\omega) = |H_n(\omega)|^2 \tag{4.64}$$

and

$$H_n(\omega) = n^{-1/2} \sum_{t=1}^{n} h_t \mathrm{e}^{-2\pi i \omega t}. \tag{4.65}$$

The value $W_n(\omega)$ is called a spectral window because, in view of (4.63), it is determining which part of the spectral density $f_x(\omega)$ is being "seen" by the estimator $I_y(\omega_j)$ on average. In the case that $h_t = 1$ for all t, $I_y(\omega_j) = I_x(\omega_j)$ is simply the periodogram of the data and the window is

$$W_n(\omega) = \frac{\sin^2(n\pi\omega)}{n \sin^2(\pi\omega)} \tag{4.66}$$

with $W_n(0) = n$, which is known as the Fejér or modified Bartlett kernel. If we consider the averaged periodogram in (4.47), namely

$$\bar{f}_x(\omega) = \frac{1}{L} \sum_{k=-m}^{m} I_x(\omega_j + k/n),$$

the window, $W_n(\omega)$, in (4.63) will take the form

$$W_n(\omega) = \frac{1}{nL} \sum_{k=-m}^{m} \frac{\sin^2[n\pi(\omega + k/n)]}{\sin^2[\pi(\omega + k/n)]}\,. \tag{4.67}$$

Tapers generally have a shape that enhances the center of the data relative to the extremities, such as a cosine bell of the form

$$h_t = .5\left[1 + \cos\left(\frac{2\pi(t - \bar{t})}{n}\right)\right], \tag{4.68}$$

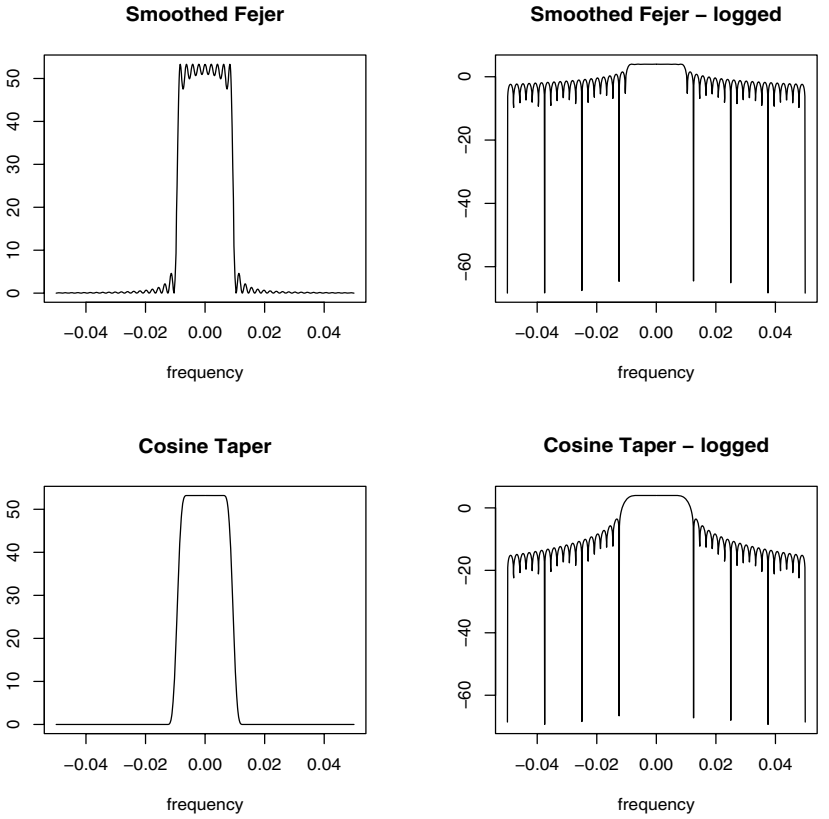


Figure 4.8 Averaged Fejér window (top row) and the corresponding cosine taper window (bottom row) for $L = 9,\ n = 480$.

where $\bar{t} = (n+1)/2$, favored by Blackman and Tukey (1959). In Figure 4.8, we have plotted the shapes of two windows, $W_n(\omega)$, for $n = 480$ and $L = 9$, when (i) $h_t \equiv 1$, in which case, (4.67) applies, and (ii) h_t is the cosine taper in (4.68). In both cases the predicted bandwidth should be $B_w = 9/480 = .01875$ cycles per point, which corresponds to the "width" of the windows shown in Figure 4.8. Both windows produce an integrated average spectrum over this band but the untapered window in the top panels shows considerable ripples over the band and outside the band. The ripples outside the band are called sidelobes and tend to introduce frequencies from outside the interval that may contaminate the desired spectral estimate within the band. For example, a large dynamic range for the values in the spectrum introduces spectra in contiguous frequency intervals several orders of magnitude greater than the value in the interval of interest. This effect is sometimes called leakage. Finally,

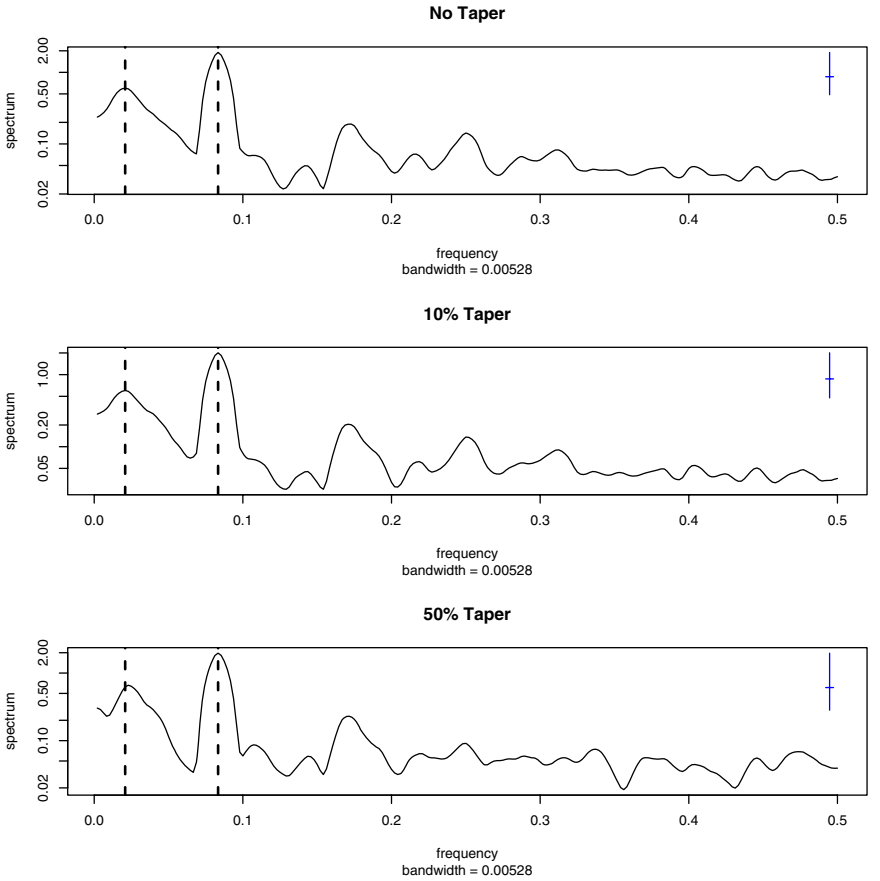


Figure 4.9 Smoothed spectral estimates of the SOI (on a $\log_{10}$ scale) without tapering (top), with 10% tapering (middle) and with 50% or complete tapering (bottom); see Example 4.12 for details.

the logged values in Figure 4.8 emphasize the suppression of the sidelobes in the Fejér kernel when a cosine taper is used.

Example 4.12 The Effect of Tapering the SOI Series

In this example we examine the effect of various tapers on the estimate of the spectrum of the SOI series. The results for the Recruitment series are similar. Figure 4.9 shows three spectral estimates plotted on a $\log_{10}$ scale along with the corresponding approximate 95% confidence intervals in the upper right. The degree of smoothing here is the same as in Example 4.11. The top of Figure 4.9 shows the estimate without any tapering and hence it is the same as the estimated spectrum displayed in the top of Figure 4.7. The middle panel in Figure 4.9 shows the effect of

10% tapering (the R default), which means that the cosine taper is being applied only to the ends of the series, 10% on each side. The bottom panel shows the results with 50% tapering; that is, (4.68) is being applied to the entire set of data.

The three spectral estimates are qualitatively similar, but note that in the fully tapered case, the peak El Niño cycle is at the 42 month (3.5 year) cycle instead of the 48 month (4 year) cycle. Also, notice that the confidence interval bands are increasing as the tapering increases. This occurrence is due to the fact that by tapering we are decreasing the amount of information, and hence the degrees of freedom; details, which are similar to the ideas discussed in (4.57)–(4.58), may be found in Bloomfield (2000, §9.5).

The following R session was used to generate Figure 4.9:

```
> par(mfrow=c(3,1))
> spectrum(soi, spans=c(7,7), taper=0, main="No Taper")
> abline(v=1/12,lty="dashed")
> abline(v=1/48,lty="dashed")
> spectrum(soi, spans=c(7,7), main="10% Taper")
> abline(v=1/12,lty="dashed")
> abline(v=1/48,lty="dashed")
> spectrum(soi, spans=c(7,7), taper=.5, main="50% Taper")
> abline(v=1/12,lty="dashed")
> abline(v=1/48,lty="dashed")
```

Example 4.13 Spectra of P and S Components for Earthquake and Explosion

Figure 4.10 shows the spectra computed separately from the two phases of the earthquake and explosion in Figure 1.7 of Chapter 1. In all cases we used a modified Daniell smoother with $L = 21$ being passed twice, and with 10% tapering. This leads to approximately 54 degrees of freedom. Because the sampling rate is 40 points per second, the folding frequency is 20 cycles per second or 20 Hertz (Hz). The highest frequency shown in the plots is .25 cycles per point or 10 Hz because there is no signal activity at frequencies beyond 10 Hz. A fundamental problem in the analysis of seismic data is discriminating between earthquakes and explosions using the kind of instruments that might be used in monitoring a nuclear test ban treaty. If we plot an ensemble of earthquakes and explosions comparable to Figure 1.7, some gross features appear that may lead to discrimination. The most common differences that we look for are subtle differences between the spectra of the two classes of events. In this case, note the strong frequency components of the P and S components of the explosion are close to the frequency .10 cycles per point or 1 Hz. On the other hand, the spectral content of the earthquakes tends to

occur along a broader frequency band and at lower frequencies for both components. Often, we assume that the ratio of P to S power is in different proportions at different frequencies, and this distinction can form a basis for discriminating between the two classes. In §7.7, we test formally for discrimination using a random effects analysis of variance approach.

Figure 4.10 was generated in R as follows:

```
> x = matrix(scan("/mydata/eq5exp6.dat"), ncol=2)
> eqP = x[1:1024, 1];  eqS = x[1025:2048, 1]
> exP = x[1:1024, 2];  exS = x[1025:2048, 2]
> par(mfrow=c(2,2))
> eqPs=spectrum(eqP, spans=c(21,21),
+   log="no", xlim=c(0,.25), ylim=c(0,.04))
> eqSs=spectrum(eqS, spans=c(21,21),
+   log="no", xlim=c(0,.25), ylim=c(0,.4))
> exPs=spectrum(exP, spans=c(21,21),
+   log="no", xlim=c(0,.25), ylim=c(0,.04))
> exSs=spectrum(exS, spans=c(21,21),
+   log="no", xlim=c(0,.25), ylim=c(0,.4))
> exSs$df
   [1] 53.87862
```

We close this section with a brief discussion of lag window estimators. First, consider the periodogram, $I(\omega_j)$, which was shown in (4.23) to be of the form

$$I(\omega_j) = \sum_{|h|<n} \widehat{\gamma}(h) e^{-2\pi i \omega_j h}.$$

Thus, (4.56) can be written as

$$\begin{aligned} \widehat{f}(\omega) &= \sum_{|k|\le m} h_k \, I(\omega_j + k/n) \\ &= \sum_{|k|\le m} h_k \sum_{|h|<n} \widehat{\gamma}(h) e^{-2\pi i(\omega_j + k/n)h} \\ &= \sum_{|h|<n} g(h/n)\, \widehat{\gamma}(h) e^{-2\pi i \omega_j h}. \end{aligned} \tag{4.69}$$

where $g(h/n) = \sum_{|k|\le m} h_k \exp(-2\pi i k h/n)$. Equation (4.69) suggests estimators of the form

$$\widetilde{f}(\omega) = \sum_{|h|\le r} w(h/r)\, \widehat{\gamma}(h) e^{-2\pi i \omega h} \tag{4.70}$$

where $w(\cdot)$ is a weight function, called the lag window, that satisfies

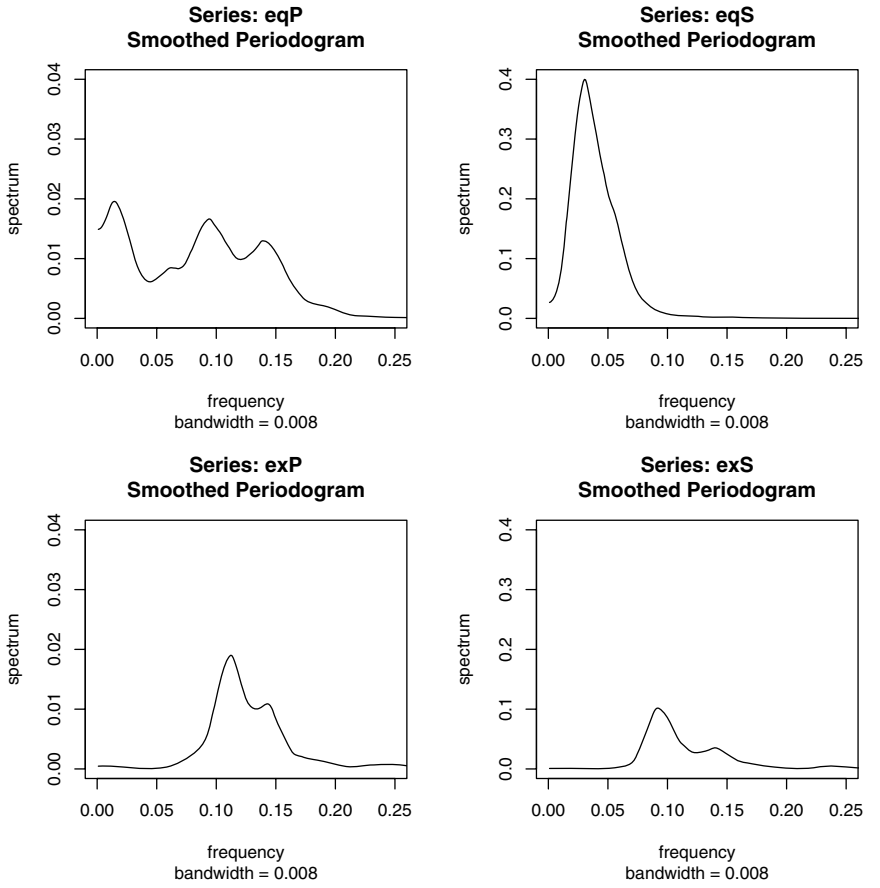


Figure 4.10 Spectral analysis of P and S components of an earthquake and an explosion, $n = 1024$. Each estimate is based on a modified Daniell smoother with $L = 21$ being passed twice, and with 10% tapering. This leads to approximately 54 degrees of freedom. Multiply frequency by 40 to convert to Hertz (cycles/second).

(*i*) $w(0) = 1$

(*ii*) $|w(x)| \leq 1$ and $w(x) = 0$ for $|x| > 1$,

(*iii*) $w(x) = w(-x)$.

Note that if $w(x) = 1$ for $|x| < 1$ and $r = n$, then $\widetilde{f}(\omega_j) = I(\omega_j)$, the periodogram. This result indicates the problem with the periodogram as an estimator of the spectral density is that it gives too much weight to the values of $\widehat{\gamma}(h)$ when h is large, and hence is unreliable [e.g, there is only one pair of

observations used in the estimate $\widehat{\gamma}(n-1)$, and so on]. The smoothing window is defined to be

$$W(\omega) = \sum_{h=-r}^{r} w(h/r)e^{-2\pi i\omega h}, \tag{4.71}$$

and it determines which part of the periodogram will be used to form the estimate of $f(\omega)$. The asymptotic theory for $\widehat{f}(\omega)$ holds for $\widetilde{f}(\omega)$ under the same conditions and provided $r \to \infty$ as $n \to \infty$ but with $r/n \to 0$. We have

$$E\{\widetilde{f}(\omega)\} \to f(\omega); \tag{4.72}$$

$$\frac{n}{r}\text{cov}\left(\widetilde{f}(\omega), \widetilde{f}(\lambda)\right) \to f^2(\omega)\int_{-1}^{1} w^2(x)dx \qquad \omega = \lambda \neq 0, 1/2. \tag{4.73}$$

In (4.73), replace $f^2(\omega)$ by 0 if $\omega \neq \lambda$ and by $2f^2(\omega)$ if $\omega = \lambda = 0$ or $1/2$.

Many authors have developed various windows and Brillinger (2001, Ch 3) and Brockwell and Davis (1991, Ch 10) are good sources of detailed information on this topic. We mention a few.

The rectangular lag window, which gives uniform weight in (4.70),

$$w(x) = 1, \quad |x| \leq 1,$$

corresponds to the Dirichlet smoothing window given by

$$W(\omega) = \frac{\sin(2\pi r + \pi)\omega}{\sin(\pi\omega)}. \tag{4.74}$$

This smoothing window takes on negative values, which may lead to estimates of the spectral density that are negative a various frequencies. Using (4.73) in this case, for large n we have

$$\text{var}\{\widetilde{f}(\omega)\} \approx \frac{2r}{n}f^2(\omega).$$

The Parzen lag window is defined to be

$$w(x) = \begin{cases} 1 - 6x + 6|x|^3 & |x| < 1/2, \\ 2(1-|x|)^3 & 1/2 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

This leads to an approximate smoothing window of

$$W(\omega) = \frac{6}{\pi r^3}\frac{\sin^4(r\omega/4)}{\sin^4(\omega/2)}.$$

For large n, the variance of the estimator is approximately

$$\text{var}\{\widetilde{f}(\omega)\} \approx .539f^2(\omega)/n.$$

The Tukey-Hanning lag window has the form

$$w(x) = \frac{1}{2}(1 + \cos(x)), \quad |x| \leq 1$$

which leads to the smoothing window

$$W(\omega) = \frac{1}{4}D_r(2\pi\omega - \pi/r) + \frac{1}{2}D_r(2\pi\omega) + \frac{1}{4}D_r(2\pi\omega + \pi/r)$$

where $D_r(\omega)$ is the Dirichlet kernel in (4.74). The approximate large sample variance of the estimator is

$$\text{var}\{\widetilde{f}(\omega)\} \approx \frac{3r}{4n}f^2(\omega).$$

The triangular lag window, also known as the Bartlett or Fejér window, given by

$$w(x) = 1 - |x|, \quad |x| \leq 1$$

leads to the Fejér smoothing window:

$$W(\omega) = \frac{\sin^2(\pi r\omega)}{r\sin^2(\pi\omega)}.$$

In this case, (4.73) yields

$$\text{var}\{\widetilde{f}(\omega)\} \approx \frac{2r}{3n}f^2(\omega).$$

The idealized rectangular smoothing window, also called the Daniell window, is given by

$$W(\omega) = \begin{cases} r & |\omega| \leq 1/2r \\ 0 & \text{otherwise} \end{cases},$$

and leads to the sinc lag window, namely

$$w(x) = \frac{\sin(\pi x)}{\pi x}, \quad |x| \leq 1.$$

From (4.73) we have

$$\text{var}\{\widetilde{f}(\omega)\} \approx \frac{r}{n}f^2(\omega).$$

For lag window estimators, the width of the rectangular window that leads to the same asymptotic variance as a given lag window estimator is sometimes called the bandwidth. For example, the bandwidth of the rectangular window is $b_r = 1/r$ and the asymptotic variance is $\frac{1}{nb_r}f^2$. The asymptotic variance of the triangular window is $\frac{2r}{3n}f^2$, so setting $\frac{1}{nb_r}f^2 = \frac{2r}{3n}f^2$ and solving we get $b_r = 3/2r$ as the corresponding bandwidth.

4.6 Multiple Series and Cross-Spectra

The notion of analyzing frequency fluctuations using classical statistical ideas extends to the case in which there are several jointly stationary series, for example, x_t and y_t. In this case, we can introduce the idea of a correlation indexed by frequency, called the coherence. The results in Appendix C, §C.2, imply the covariance function

$$\gamma_{xy}(h) = E[(x_{t+h} - \mu_x)(y_t - \mu_y)]$$

has the representation

$$\gamma_{xy}(h) = \int_{-1/2}^{1/2} f_{xy}(\omega) \mathrm{e}^{2\pi i \omega h} \, d\omega \quad h = 0, \pm 1, \pm 2, ..., \tag{4.75}$$

where the cross-spectrum is defined as the Fourier transform

$$f_{xy}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{xy}(h) \, \mathrm{e}^{-2\pi i \omega h} \quad -1/2 \leq \omega \leq 1/2, \tag{4.76}$$

assuming that the cross-covariance function is absolutely summable, as was the case for the autocovariance. The cross-spectrum is generally a complex-valued function, and it is often written as[12]

$$f_{xy}(\omega) = c_{xy}(\omega) - i q_{xy}(\omega), \tag{4.77}$$

where

$$c_{xy}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{xy}(h) \, \cos(2\pi\omega h) \tag{4.78}$$

and

$$q_{xy}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{xy}(h) \, \sin(2\pi\omega h) \tag{4.79}$$

are defined as the cospectrum and quadspectrum, respectively. Because of the relationship $\gamma_{yx}(h) = \gamma_{xy}(-h)$, it follows, by substituting into (4.76) and rearranging, that

$$f_{yx}(\omega) = \overline{f_{xy}(\omega)}. \tag{4.80}$$

This result, in turn, implies that the cospectrum and quadspectrum satisfy

$$c_{yx}(\omega) = c_{xy}(\omega) \tag{4.81}$$

and

$$q_{yx}(\omega) = -q_{xy}(\omega). \tag{4.82}$$

[12]For this section, it will be useful to recall the facts $e^{-i\alpha} = \cos(\alpha) - i\sin(\alpha)$ and if $z = a + ib$, then $\overline{z} = a - ib$.

An important example of the application of the cross-spectrum is to the problem of predicting an output series y_t from some input series x_t through a linear filter relation such as the three-point moving average considered below. A measure of the strength of such a relation is the squared coherence function, defined as

$$\rho^2_{y\cdot x}(\omega) = \frac{|f_{yx}(\omega)|^2}{f_{xx}(\omega) f_{yy}(\omega)}, \tag{4.83}$$

where $f_{xx}(\omega)$ and $f_{yy}(\omega)$ are the individual spectra of the x_t and y_t series, respectively. Although we consider a more general form of this that applies to multiple inputs later, it is instructive to display the single input case as (4.83) to emphasize the analogy with conventional squared correlation, which takes the form

$$\rho^2_{yx} = \frac{\sigma^2_{yx}}{\sigma^2_x \sigma^2_y},$$

for random variables with variances σ^2_x and σ^2_y and covariance $\sigma_{yx} = \sigma_{xy}$. This motivates the interpretation of squared coherence and the squared correlation between two time series at frequency ω.

Example 4.14 Cross-Spectrum and Coherence of a Process and a Three-Point Moving Average

As a simple example, we compute the cross-spectrum between x_t and the three-point moving average $y_t = (x_{t-1} + x_t + x_{t+1})/3$, where x_t is a stationary input process with spectral density $f_{xx}(\omega)$. First,

$$\begin{aligned}
\gamma_{xy}(h) &= E[x_{t+h} y_t] \\
&= \frac{1}{3} E\left[x_{t+h} \ (x_{t-1} + x_t + x_{t+1})\right] \\
&= \frac{1}{3}\left(\gamma_{xx}(h+1) + \gamma_{xx}(h) + \gamma_{xx}(h-1)\right) \\
&= \frac{1}{3} \int_{-1/2}^{1/2} (\mathrm{e}^{2\pi i\omega} + 1 + \mathrm{e}^{-2\pi i\omega}) \mathrm{e}^{2\pi i\omega h} f_{xx}(\omega) \ d\omega \\
&= \frac{1}{3} \int_{-1/2}^{1/2} [1 + 2\cos(2\pi\omega)] f_{xx}(\omega) \mathrm{e}^{2\pi i\omega h} \ d\omega.
\end{aligned}$$

Using the uniqueness of the Fourier transform, we argue from the spectral representation (4.75) that the above must be the transform of $f_{xy}(\omega)$, implying that

$$f_{xy}(\omega) = \frac{1}{3} \left[1 + \cos(2\pi\omega)\right] \ f_{xx}(\omega)$$

so that the cross-spectrum is real in this case. From Example 4.5, the spectral density of y_t is

$$f_{yy}(\omega) = \frac{1}{9}[3 + 4\cos(2\pi\omega) + 2\cos(4\pi\omega)] f_{xx}(\omega)$$

$$= \frac{1}{9}\,[1 + 2\cos(2\pi\omega)]^2\; f_{xx}(\omega),$$

using the identity $\cos(2\alpha) = 2\cos^2(\alpha) - 1$ in the last step. Substituting into (4.83) yields the squared coherence between x_t and y_t as unity over all frequencies. This is a characteristic inherited by more general linear filters, as will be shown in Problem 4.23. However, if some noise is added to the three-point moving average, the coherence is not unity; these kinds of models will be considered in detail later.

Property P4.3: Spectral Representation of a Vector Stationary Process
If the elements of the $p \times p$ autocovariance function matrix

$$\Gamma(h) = E[(\boldsymbol{x}_{t+h} - \boldsymbol{\mu})(\boldsymbol{x}_t - \boldsymbol{\mu})']$$

of a p-dimensional stationary time series, $\boldsymbol{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tp})'$, has elements satisfying

$$\sum_{h=-\infty}^{\infty} |\gamma_{jk}(h)| < \infty \tag{4.84}$$

for all $j, k = 1, \ldots, p$, then $\Gamma(h)$ has the representation

$$\Gamma(h) = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega h}\; f(\omega)\; d\omega \quad h = 0, \pm 1, \pm 2, ..., \tag{4.85}$$

as the inverse transform of the spectral density matrix, $f(\omega) = \{f_{jk}(\omega)\}$, for $j, k = 1, \ldots, p$, with elements equal to the cross-spectral components. The matrix $f(\omega)$ has the representation

$$f(\omega) = \sum_{h=-\infty}^{\infty} \Gamma(h)\mathrm{e}^{-2\pi i\omega h} \quad -1/2 \le \omega \le 1/2. \tag{4.86}$$

Example 4.15 Spectral Matrix of a Bivariate Process

Consider a jointly stationary bivariate process (x_t, y_t). We arrange the autocovariances in the matrix

$$\Gamma(h) = \begin{pmatrix} \gamma_{xx}(h) & \gamma_{xy}(h) \\ \gamma_{yx}(h) & \gamma_{yy}(h) \end{pmatrix}.$$

The spectral matrix would be given by

$$f(\omega) = \begin{pmatrix} f_{xx}(\omega) & f_{xy}(\omega) \\ f_{yx}(\omega) & f_{yy}(\omega) \end{pmatrix},$$

where the Fourier transform (4.85) and (4.86) relate the autocovariance and spectral matrices.

The extension of spectral estimation to vector series is fairly obvious. For the vector series $\boldsymbol{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tp})'$, we may use the vector of DFTs, say $\boldsymbol{d}(\omega_j) = (d_1(\omega_j), d_2(\omega_j), \ldots, d_p(\omega_j))'$, and estimate the spectral matrix by

$$\bar{f}(\omega) = L^{-1} \sum_{k=-m}^{m} I(\omega_j + k/n) \tag{4.87}$$

where now

$$I(\omega_j) = \boldsymbol{d}(\omega_j)\, \boldsymbol{d}^*(\omega_j) \tag{4.88}$$

is a $p \times p$ complex matrix.[13]

Again, the series may be tapered before the DFT is taken in (4.87) and we can use weighted estimation,

$$\widehat{f}(\omega) = \sum_{k=-m}^{m} h_k\, I(\omega_j + k/n) \tag{4.89}$$

where $\{h_k\}$ are weights as defined in (4.56). The estimate of squared coherence between two series, y_t and x_t is

$$\widehat{\rho}^2_{y\cdot x}(\omega) = \frac{|\widehat{f}_{yx}(\omega)|^2}{\widehat{f}_{xx}(\omega)\widehat{f}_{yy}(\omega)}. \tag{4.90}$$

If the spectral estimates in (4.90) are obtained using equal weights, we will write $\bar{\rho}^2_{y\cdot x}(\omega)$ for the estimate.

Under general conditions, if $\rho^2_{y\cdot x}(\omega) > 0$ then

$$|\widehat{\rho}_{y\cdot x}(\omega)| \sim AN\left(|\rho_{y\cdot x}(\omega)|, \left(1 - \rho^2_{y\cdot x}(\omega)\right)^2 / 2L_h\right) \tag{4.91}$$

where L_h is defined in (4.57); the details of this result may be found in Brockwell and Davis (1991, Ch 11). We may use (4.91) to obtain approximate confidence intervals for the square coherency $\rho^2_{y\cdot x}(\omega)$.

We can test the hypothesis that $\rho^2_{y\cdot x}(\omega) = 0$ if we use $\bar{\rho}^2_{y\cdot x}(\omega)$ for the estimate with $L > 1$,[14] that is,

$$\bar{\rho}^2_{y\cdot x}(\omega) = \frac{|\bar{f}_{yx}(\omega)|^2}{\bar{f}_{xx}(\omega)\bar{f}_{yy}(\omega)}. \tag{4.92}$$

In this case, under the null hypothesis, the statistic

$$F_{2,2L-2} = \frac{\bar{\rho}^2_{y\cdot x}(\omega)}{(1 - \bar{\rho}^2_{y\cdot x}(\omega))}(L-1) \tag{4.93}$$

[13] If Z is a complex matrix, then $Z^* = \overline{Z}'$ denotes the conjugate transpose operation. That is, Z^* is the result of replacing each element of Z by its complex conjugate and transposing the resulting matrix.

[14] If $L = 1$ then $\bar{\rho}^2_{y\cdot x}(\omega) \equiv 1$.

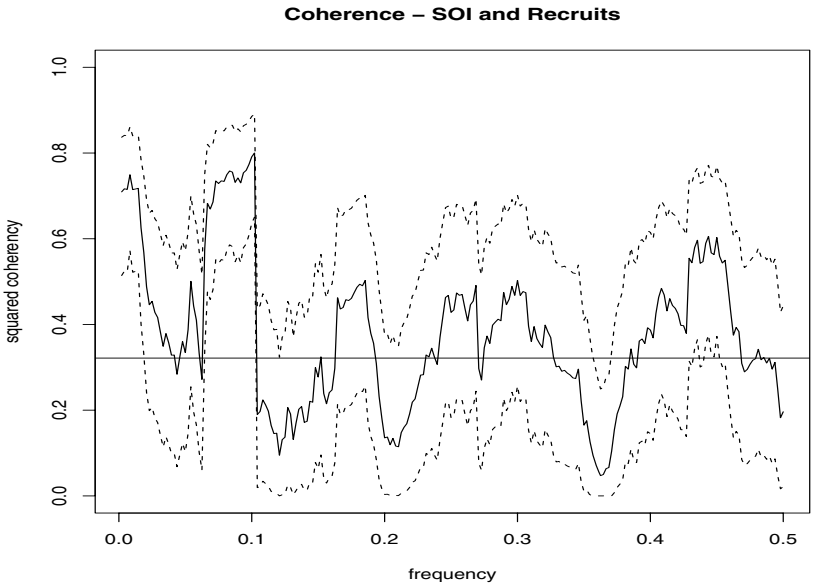


Figure 4.11 Coherence function between the SOI and Recruitment series; $L = 19,\ n = 453,\ n' = 480$, and $\alpha = .001$.

has an approximate F-distribution with 2 and $2L-2$ degrees of freedom. When the series have been extended to length n', we replace $2L-2$ by $df-2$, where df is defined in (4.53). Solving (4.93) for a particular significance level α leads to

$$C_\alpha = \frac{F_{2,2L-2}(\alpha)}{L-1+F_{2,2L-2}(\alpha)} \tag{4.94}$$

as the approximate value that must be exceeded for the original squared coherence to be able to reject $\rho^2_{y\cdot x}(\omega) = 0$ at an *a priori* specified frequency.

Example 4.16 Coherence Between SOI and Recruitment Series

Figure 4.11 shows the squared coherence between the SOI and Recruitment series over a wider band than was used for the spectrum. In this case, we used $L = 19,\ df = 2(19)(453/480) \approx 36$ and $F_{2,df-2}(.001) \approx 8.53$ at the significance level $\alpha = .001$. Hence, we may reject the hypothesis of no coherence for values of $C_{.001} > .32$. We emphasize that this method is crude because, in addition to the fact that the F-statistic is approximate, we are examining the squared coherence across all frequencies with the Bonferroni inequality, (4.55), in mind. Figure 4.11 also exhibits confidence bands as part of the R plotting routine. We emphasize that these bands are only valid for ω where $\rho^2_{y\cdot x}(\omega) > 0$.

In this case, the seasonal frequency and the El Niño frequencies ranging between about 3 and 7 year periods are strongly coherent. Other frequencies are also strongly coherent, although the strong coherence is less impressive because the underlying power spectrum at these higher frequencies is fairly small. Finally, we note that the coherence is persistent at the seasonal harmonic frequencies.

This example may be reproduced using the following R commands.

```
> x = ts(cbind(soi,rec))
> s = spec.pgram(x, kernel("daniell",9), taper=0)
> s$df                        # df = 35.8625
> f = qf(.999, 2, s$df-2)  #  f = 8.529792
> c = f/(18+f)                #  c = 0.3188779
> plot(s, plot.type = "coh", ci.lty = 2)
> abline(h = c)
```

4.7 Linear Filters

Some of the examples of the previous sections have hinted at the possibility the distribution of power or variance in a time series can be modified by making a linear transformation. In this section, we explore that notion further by defining a linear filter and showing how it can be used to extract signals from a time series. The linear filter modifies the spectral characteristics of a time series in a predictable way, and the systematic development of methods for taking advantage of the special properties of linear filters is an important topic in time series analysis.

A linear filter uses a set of specified coefficients a_t, for $t = 0, \pm 1, \pm 2 \ldots$, to transform a stationary input series, x_t, producing an output series, y_t, of the form

$$y_t = \sum_{r=-\infty}^{\infty} a_r x_{t-r}. \tag{4.95}$$

The form (4.95) is also called a convolution in some statistical contexts. The coefficients, collectively called the *impulse response function*, are required to satisfy absolute summability

$$\sum_{t=-\infty}^{\infty} |a_t| < \infty, \tag{4.96}$$

so (4.95) exists as a limit in mean square and the infinite Fourier transform

$$A_{yx}(\omega) = \sum_{t=-\infty}^{\infty} a_t \mathrm{e}^{-2\pi i \omega t}, \tag{4.97}$$

called the *frequency response function*, is well defined. We have already encountered several linear filters, for example, the simple three-point moving average in Example 4.5, which can be put into the form of (4.95) by letting $a_{-1} = a_0 = a_1 = 1/3$ and taking $a_t = 0$ for $|t| \geq 2$.

The importance of the linear filter stems from its ability to enhance certain parts of the spectrum of the input series. To see this, the autocovariance function of the filtered output (4.95) can be derived as

$$\begin{aligned}
\gamma_{yy}(h) &= E[(y_{t+h} - Ey_{t+h})(y_t - Ey_t)] \\
&= E\left[\sum_r \sum_s a_r(x_{t+h-r} - \mu)(x_{t-s} - \mu)a_s\right] \\
&= \sum_r \sum_s a_r \gamma_{xx}(h - r + s)a_s \\
&= \sum_r \sum_s a_r \left[\int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega(h-r+s)} f_{xx}(\omega)d\omega\right] a_s \\
&= \int_{-1/2}^{1/2} \left(\sum_r a_r \mathrm{e}^{-2\pi i\omega r}\right)\left(\sum_s a_s \mathrm{e}^{2\pi i\omega s}\right) \mathrm{e}^{2\pi i\omega h} f_{xx}(\omega)\ d\omega \\
&= \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega h} |A_{yx}(\omega)|^2 f_{xx}(\omega)\ d\omega,
\end{aligned}$$

where we have first replaced $\gamma_{xx}(\cdot)$ by its representation (4.12) and then substituted $A_{yx}(\omega)$ from (4.97). The computation is one we do repeatedly, exploiting the uniqueness of the Fourier transform. Now, because the left-hand side is the Fourier transform of the spectral density of the output, say, $f_{yy}(\omega)$, we get the important filtering property as follows.

Property P4.4: Output Spectrum of a Filtered Stationary Series
The spectrum of the filtered output y_t in (4.95) is related to the spectrum of the input x_t by

$$f_{yy}(\omega) = |A_{yx}(\omega)|^2\ f_{xx}(\omega), \tag{4.98}$$

where the frequency response function $A_{yx}(\omega)$ is defined in (4.97).

The result (4.98) enables us to calculate the exact effect on the spectrum of any given filtering operation. This important property shows the spectrum of the input series is changed by filtering and the effect of the change can be characterized as a frequency-by-frequency multiplication by the squared magnitude of the frequency response function. Again, an obvious analogy to a property of the variance in classical statistics holds, namely, if x is a random variable with variance σ_x^2, then $y = ax$ will have variance $\sigma_y^2 = a^2\sigma_x^2$, so the variance of the linearly transformed random variable is changed by multiplication by a^2 in much the same way as the linearly filtered spectrum is changed in (4.98).

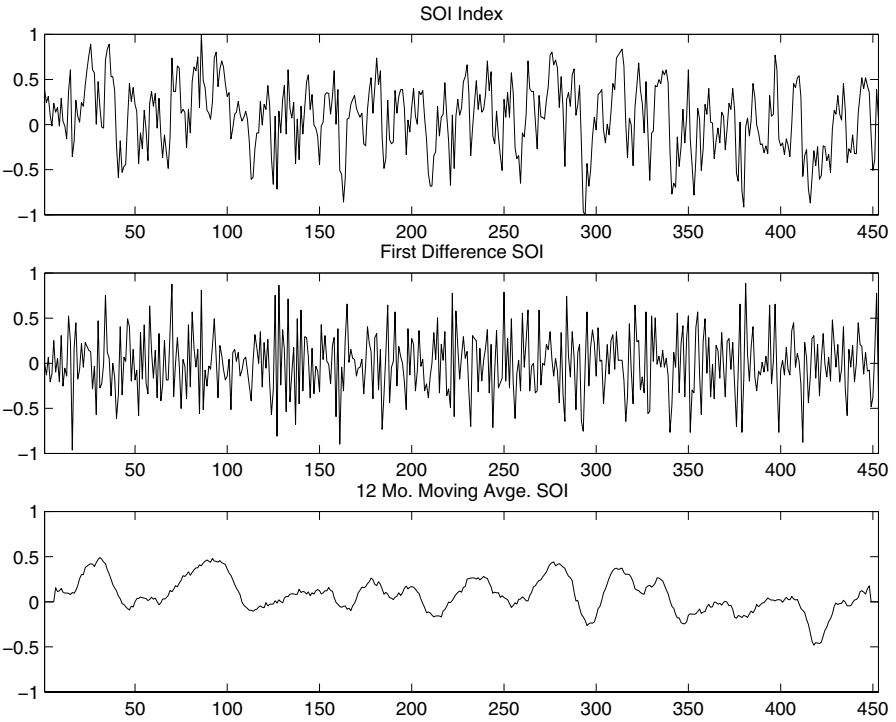


Figure 4.12 SOI series (top) compared with the differenced SOI (middle) and a centered 12-month moving average (bottom).

Example 4.17 First Difference and Moving Average Filters

We illustrate the effect of filtering with two common examples, the first difference filter

$$y_t = \nabla x_t = x_t - x_{t-1}$$

and the symmetric moving average filter

$$y_t = \frac{1}{24}\left(x_{t-6} + x_{t+6}\right) + \frac{1}{12}\sum_{r=-5}^{5} x_{t-r},$$

which is a modified Daniell kernel with $m = 6$. The results of filtering the SOI series using the two filters are shown in the middle and bottom panels of Figure 4.12. Notice that the effect of differencing is to roughen the series because it tends to retain the higher or faster frequencies. The centered moving average smoothes the series because it retains the lower frequencies and tends to attenuate the higher frequencies. In general, differencing is an example of a *high-pass filter* because it retains or passes

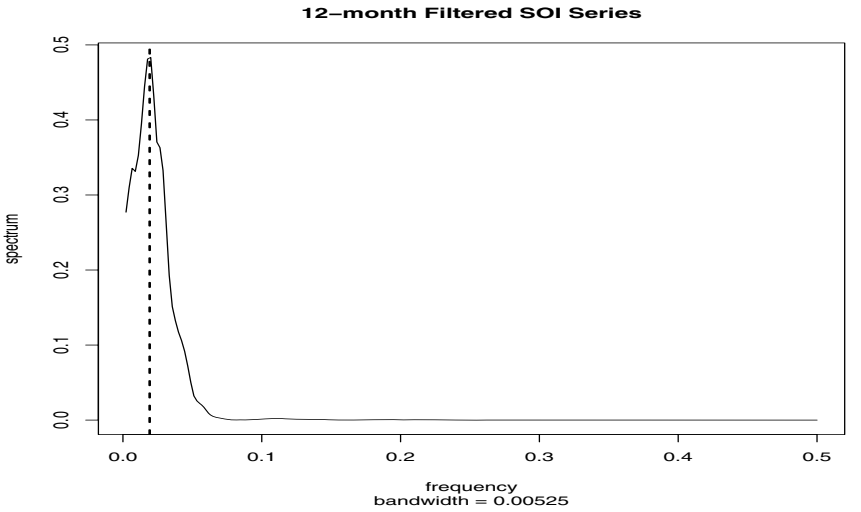


Figure 4.13 Spectral analysis of the SOI series after applying a 12-month moving average filter. The vertical line corresponds to the 52-month cycle.

the higher frequencies, whereas the moving average is a *low-pass filter* because it passes the lower or slower frequencies.

Notice that the slower periods are enhanced in the symmetric moving average and the seasonal or yearly frequencies are attenuated. The filtered series makes about 9 cycles in the length of the data (about one cycle every 52 months) and the moving average filter tends to enhance or extract the signal that is associated with El Niño. Moreover, by the low-pass filtering of the data, we get a better sense of the El Niño effect and its irregularity. Figure 4.13 shows the results of a spectral analysis on the low-pass filtered SOI series. It is clear that all high frequency behavior has been removed and the El Niño cycle is accentuated; the dotted vertical line in the figure corresponds to the 52 months cycle.

Now, having done the filtering, it is essential to determine the exact way in which the filters change the input spectrum. We shall use (4.97) and (4.98) for this purpose. The first difference filter can be written in the form (4.95) by letting $a_0 = 1, a_1 = -1$, and $a_r = 0$ otherwise. This implies that

$$A_{yx}(\omega) = 1 - \mathrm{e}^{-2\pi i\omega},$$

and the squared frequency response becomes

$$\begin{aligned} |A_{yx}(\omega)|^2 &= (1 - \mathrm{e}^{-2\pi i\omega})(1 - \mathrm{e}^{2\pi i\omega}) \\ &= 2[1 - \cos(2\pi\omega)]. \end{aligned} \tag{4.99}$$

The top panel of Figure 4.14 shows that the first difference filter will attenuate the lower frequencies and enhance the higher frequencies be-

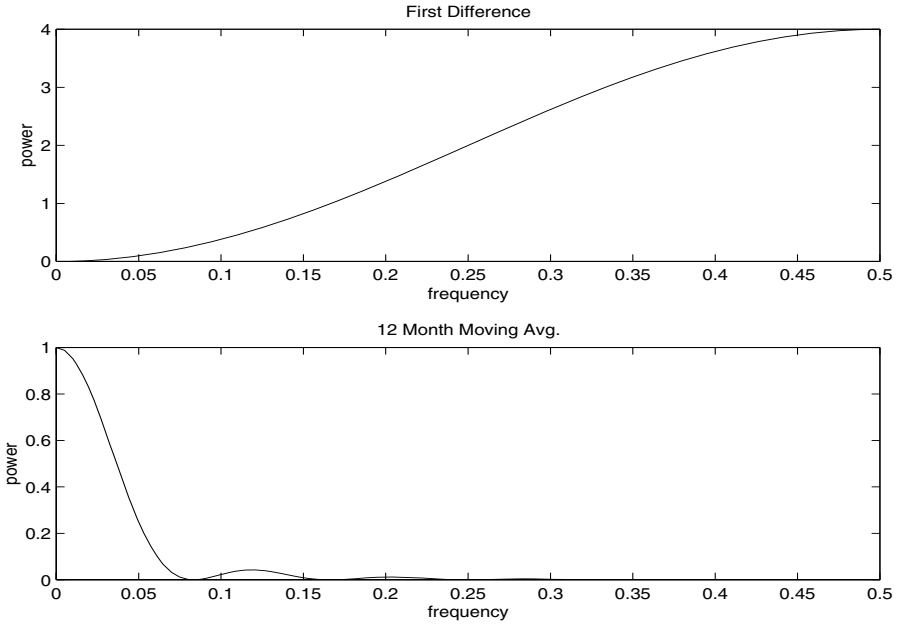


Figure 4.14 Squared frequency response functions of the first difference and 12-month moving average filters.

cause the multiplier of the spectrum, $|A_{yx}(\omega)|^2$, is large for the higher frequencies and small for the lower frequencies. Generally, the slow rise of this kind of filter does not particularly recommend it as a procedure for retaining only the high frequencies.

For the centered 12-month moving average, we can take $a_{-6} = a_6 = 1/24$, $a_k = 1/12$ for $-5 \leq k \leq 5$ and $a_k = 0$ elsewhere. Substituting and recognizing the cosine terms gives

$$A_{yx}(\omega) = \frac{1}{12}\left[1 + \cos(12\pi\omega) + 2\sum_{k=1}^{5}\cos(2\pi\omega k)\right]. \tag{4.100}$$

Plotting the squared frequency response of this function as in Figure 4.14 shows that we can expect this filter to cut most of the frequency content above .05 cycles per point. This corresponds to eliminating periods shorter than $T = 1/.05 = 20$ points. In particular, this drives down the yearly components with periods of $T = 12$ months and enhances the El Niño frequency, which is somewhat lower. The filter is not completely efficient at attenuating high frequencies; some power contributions are left at higher frequencies, as shown in the function $|A_{yx}(\omega)|^2$ and in the filtered series in Figure 4.3.

The following R session shows how to filter the data, perform the spectral

analysis of this example, and plot the squared frequency response curve of the difference filter.

```
> par(mfrow=c(3,1))
> plot.ts(soi)          # the data
> plot.ts(diff(soi))    # first difference
> k = kernel("modified.daniell", 6)  #-- 12 month filter
> soif = kernapply(soi,k)
> plot.ts(soif)
> windows()  # open new graphics device - use x11() in unix
> spectrum(soif, spans=9, log="no")  #-- spectral analysis
> abline(v=1/52, lty="dotted")
> windows()
> w = seq(0,.5, length=1000)    #-- frequency response
> FR = abs(1-exp(2i*pi*w))^2
> plot(w, FR, type="l")
```

The two filters discussed in the previous example were different in that the frequency response function of the first difference was complex-valued, whereas the frequency response of the moving average was purely real. A short derivation similar to that used to verify (4.98) shows, when x_t and y_t are related by the linear filter relation (4.95), the cross-spectrum satisfies

$$f_{yx}(\omega) = A_{yx}(\omega) f_{xx}(\omega),$$

so the frequency response is of the form

$$A_{yx}(\omega) = \frac{f_{yx}(\omega)}{f_{xx}(\omega)} \tag{4.101}$$

$$= \frac{c_{yx}(\omega)}{f_{xx}(\omega)} - i\frac{q_{yx}(\omega)}{f_{xx}(\omega)}, \tag{4.102}$$

where we have used (4.77) to get the last form. Then, we may write (4.102) in polar coordinates as

$$A_{yx}(\omega) = |A_{yx}(\omega)|\ \exp\{-i\ \phi_{yx}(\omega)\}, \tag{4.103}$$

where the amplitude and phase of the filter are defined by

$$|A_{yx}(\omega)| = \frac{\sqrt{c^2_{yx}(\omega) + q^2_{yx}(\omega)}}{f_{xx}(\omega)} \tag{4.104}$$

and

$$\phi_{yx}(\omega) = \tan^{-1}\left(-\frac{q_{yx}(\omega)}{c_{yx}(\omega)}\right). \tag{4.105}$$

A simple interpretation of the phase of a linear filter is that it exhibits time delays as a function of frequency in the same way as the spectrum represents

the variance as a function of frequency. Additional insight can be gained by considering the simple delaying filter

$$y_t = Ax_{t-D},$$

where the series gets replaced by a version, amplified by multiplying by A and delayed by D points. For this case,

$$f_{yx}(\omega) = A\mathrm{e}^{-2\pi i\omega D} f_{xx}(\omega),$$

and the amplitude is $|A|$, and the phase is

$$\phi_{yx}(\omega) = -2\pi\omega D,$$

or just a linear function of frequency ω. For this case, applying a simple time delay causes phase delays that depend on the frequency of the periodic component being delayed. Interpretation is further enhanced by setting $x_t = \cos(2\pi\omega t)$, in which case $y_t = A\cos(2\pi\omega t - 2\pi\omega D)$. Thus, the output series, y_t, has the same period as the input series, x_t, but the amplitude of the output has increased by a factor of $|A|$ and the phase has been changed by a factor of $-2\pi\omega D$.

Example 4.18 Amplitude and Phase of Difference and Moving Average

We consider calculating the amplitude and phase of the two filters discussed in Example 4.17. The case for the moving average is easy because $A_{yx}(\omega)$ given in (4.100) is purely real. So, the amplitude is just $|A_{yx}(\omega)|$ and the phase is $\phi_{yx}(\omega) = 0$. In general, symmetric ($a_t = a_{-t}$) filters have zero phase. The first difference, however, changes this, as we might expect from the example above involving the time delay filter. In this case, the squared amplitude is given in (4.99). To compute the phase, we write

$$\begin{aligned} A_{yx}(\omega) &= 1 - e^{-2\pi i\omega} \\ &= e^{-i\pi\omega}(e^{i\pi\omega} - e^{-i\pi\omega}) \\ &= 2ie^{-i\pi\omega}\sin(\pi\omega) \\ &= 2\sin^2(\pi\omega) + 2i\cos(\pi\omega)\sin(\pi\omega) \\ &= \frac{c_{yx}(\omega)}{f_{xx}(\omega)} - i\frac{q_{yx}(\omega)}{f_{xx}(\omega)}, \end{aligned}$$

so

$$\begin{aligned} \phi_{yx}(\omega) &= \tan^{-1}\left(-\frac{q_{yx}(\omega)}{c_{yx}(\omega)}\right) \\ &= \tan^{-1}\left(\frac{\cos(\pi\omega)}{\sin(\pi\omega)}\right). \end{aligned}$$

Noting that

$$\cos(\pi\omega) = \sin(-\pi\omega + \pi/2)$$

and that

$$\sin(\pi\omega) = \cos(-\pi\omega + \pi/2),$$

we get

$$\phi_{yx}(\omega) = -\pi\omega + \pi/2,$$

and the phase is again a linear function of frequency.

The above tendency of the frequencies to arrive at different times in the filtered version of the series remains as one of two annoying features of the difference type filters. The other weakness is the gentle increase in the frequency response function. If low frequencies are really unimportant and high frequencies are to be preserved, we would like to have a somewhat sharper response than is obvious in Figure 4.14. Similarly, if low frequencies are important and high frequencies are not, the moving average filters are also not very efficient at passing the low frequencies and attenuating the high frequencies. Improvement is possible by using longer filters, obtained by approximations to the infinite inverse Fourier transform. The design of filters will be discussed in §4.10 and §4.11.

We will occasionally use results for multivariate series $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})'$ that are comparable to the simple property shown in (4.98). Consider the matrix filter

$$\boldsymbol{y}_t = \sum_{r=-\infty}^{\infty} A_r \boldsymbol{x}_{t-r}, \tag{4.106}$$

where $\{A_r\}$ denotes a sequence of $q \times p$ matrices such that $\sum_{r=-\infty}^{\infty} ||A_r|| < \infty$, $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})'$ is a $p \times 1$ stationary vector process with mean vector $\boldsymbol{\mu}_x$ and $p \times p$, matrix covariance function $\Gamma_{xx}(h)$ and spectral matrix $f_{xx}(\omega)$, and $\boldsymbol{y}_t$ is the $q \times 1$ vector output process. Then, we can obtain the following property.

Property P4.5: Output Spectral Matrix of a Linearly Filtered Stationary Vector Series

The spectral matrix of the filtered output $\boldsymbol{y}_t$ in (4.106) is related to the spectrum of the input $\boldsymbol{x}_t$ by

$$f_{yy}(\omega) = \mathcal{A}(\omega) f_{xx}(\omega) \mathcal{A}^*(\omega), \tag{4.107}$$

where the matrix frequency response function $\mathcal{A}(\omega)$ is defined by

$$\mathcal{A}(\omega) = \sum_{t=-\infty}^{\infty} A_t \exp(-2\pi i \omega t). \tag{4.108}$$

4.8 Parametric Spectral Estimation

The methods of §4.5 lead to estimators generally referred to as nonparametric spectra because no assumption is made about the parametric form of the spectral density. In Example 4.6, we derived the spectrum of a second-order autoregressive series and we might consider basing a spectral estimator on this function, using the estimated parameters ϕ_1, ϕ_2, and σ_w^2. Then, substituting the parameter estimates into the spectral density $f_x(\omega)$ determined in that example would lead to a parametric estimator for the spectrum. Similarly, we might fit a p-th order autoregression, with the order p determined by one of the model selection criteria, such as AIC, AICc, and SIC, defined in (2.18)-(2.20) for the regression model. Parametric autoregressive spectral estimators will often have superior resolution in problems when several closely spaced narrow spectral peaks are present and are preferred by engineers for a broad variety of problems (see Kay, 1988). The development of autoregressive spectral estimators has been summarized by Parzen (1983).

To be specific, consider the equation determining the order p autoregressive model (2.1), written in the form

$$x_t - \sum_{k=1}^{p} \phi_k x_{t-k} = w_t, \tag{4.109}$$

where w_t is a white noise process with mean zero and variance σ_w^2. Then, note the linear filter Property P4.4, combined with equating the spectra of the left- and right-hand sides of the defining equation above yields

$$|\phi(\mathrm{e}^{-2\pi i\omega})|^2 f_x(\omega) = \sigma_w^2, \tag{4.110}$$

where

$$\phi(\mathrm{e}^{-2\pi i\omega}) = 1 - \sum_{k=1}^{p} \phi_k \mathrm{e}^{-2\pi i\omega k}. \tag{4.111}$$

Then, denoting the maximum likelihood or least squares estimators of the model parameters by $\widehat{\phi}_1, \widehat{\phi}_2, \ldots, \widehat{\phi}_p$ and $\widehat{\sigma}_w^2$, we may substitute them into the form of the spectrum implied by (4.110), obtaining

$$\widehat{f}_x(\omega) = \frac{\widehat{\sigma}_w^2}{|\widehat{\phi}(\mathrm{e}^{-2\pi i\omega})|^2}. \tag{4.112}$$

The asymptotic distribution of the autoregressive spectral estimator has been obtained by Berk (1974) under the conditions $p \to \infty$, $p^3/n \to 0$ as $p,\ n \to \infty$, which may be too severe for most applications. The limiting results imply a confidence interval of the form

$$\frac{\widehat{f}_x(\omega)}{(1 + C z_{\alpha/2})} \le f_x(\omega) \le \frac{\widehat{f}_x(\omega)}{(1 - C z_{\alpha/2})}, \tag{4.113}$$

where

$$C = \sqrt{2p/n} \tag{4.114}$$

and $z_{\alpha/2}$ is ordinate corresponding to the upper $\alpha/2$ probability of the standard normal distribution. If the sampling distribution is to be checked, we suggest applying the bootstrap estimator to get the sampling distribution of $\hat{f}_x(\omega)$ using a procedure similar to the one used for $p = 1$ in Example 3.33. An alternative for higher order autoregressive series is to put the AR(p) in state-space form and use the bootstrap procedure discussed in §6.7.

An interesting fact about rational spectra of the form (4.110) is that any spectral density can be approximated, arbitrarily close by the spectrum of an AR process.

Property P4.6: Approximating a Spectral Density with an AR Spectrum

Let $g(\omega)$ be the spectral density of a stationary process. Then, given $\epsilon > 0$, there is a time series with the representation

$$x_t = \sum_{k=1}^{p} \phi_k x_{t-k} + w_t$$

where w_t is white noise with variance σ_w^2, such that

$$|f_x(\omega) - g(\omega)| < \epsilon \quad \text{all } \omega \in [-1/2, 1/2].$$

Moreover, p is finite and the roots of $\phi(z) = 1 - \sum_{k=1}^{p} \phi_k z^k$ are outside the unit circle.

One drawback of the property is that it does not tell us how large p must be before the approximation is reasonable; in some situations p may be extremely large. Property P4.6 also holds for MA and for ARMA processes in general, and a proof of the result may be found in Fuller (1996, Ch 4). For an ARMA(p, q) process we would have

$$f_x(\omega) = \sigma_w^2 \frac{|\theta(e^{-2\pi i\omega})|^2}{|\phi(e^{-2\pi i\omega})|^2} \tag{4.115}$$

where $\theta(z) = 1 + \sum_{k=1}^{q} \theta_k z^k$. We demonstrate the technique in the following example.

Example 4.19 Autoregressive Spectral Estimator of the SOI Series

Consider obtaining results comparable to the nonparametric estimators shown in Figure 4.5 for the SOI series. Fitting successively higher order models for $p = 1, 2, \ldots, 30$ yields a minimum SIC at $p = 15$ and a minimum AICc at $p = 16$, as shown in Figure 4.15. We can see from Figure 4.15 that SIC is very definite about which model it chooses; that is, the minimum SIC is very distinct. On the other hand, it is not clear

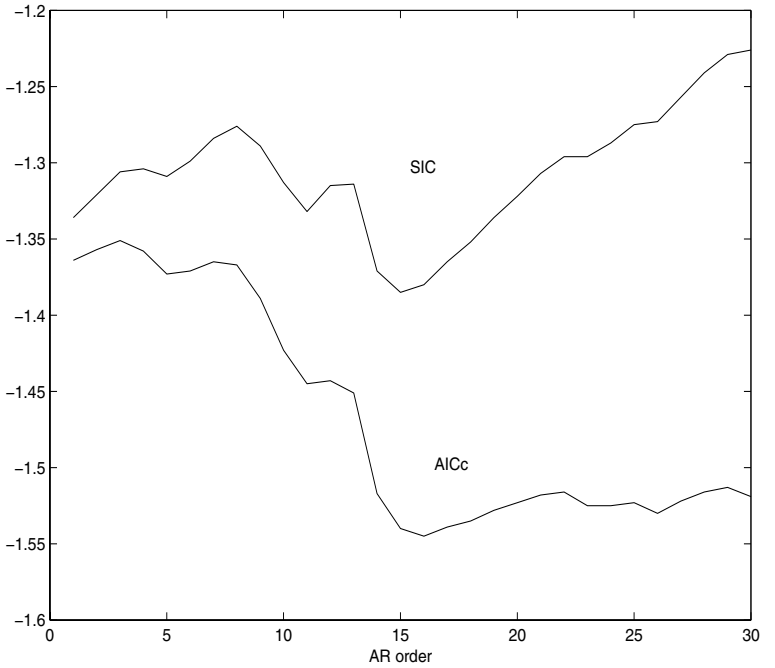


Figure 4.15 Model selection criteria AICc and SIC as a function of order p for autoregressive models fitted to the SOI series.

what is going to happen with AICc; that is, the minimum is not so clear, and there is some concern that AICc will start decreasing after $p = 30$. Minimum AIC selects the $p = 15$ model (but suffers from the same uncertainty as AICc) as will be seen in the R example. The spectra of the two cases are almost identical, as shown in Figure 4.16, and we note the strong peaks at 52 months and 12 months corresponding to the nonparametric estimators obtained in §4.5. In addition, the harmonics of the yearly period is evident in the estimated spectrum.

To perform a similar analysis in R, the command `spec.ar` can be used to fit the best model via AIC and plot the resulting spectrum. A quick way to obtain the AIC values is to run the `ar` command as follows.

```
> spec.ar(soi, log="no")           # plot min AIC spectrum
> abline(v=1/52, lty="dotted")     # locate El Nino period
> abline(v=1/12, lty="dotted")     # locate yearly period
> soi.ar = ar(soi, order.max=30)   # obtain AICs
> plot(0:30, soi.ar$aic, type="l") # plot AICs
> soi.ar                           # results
```

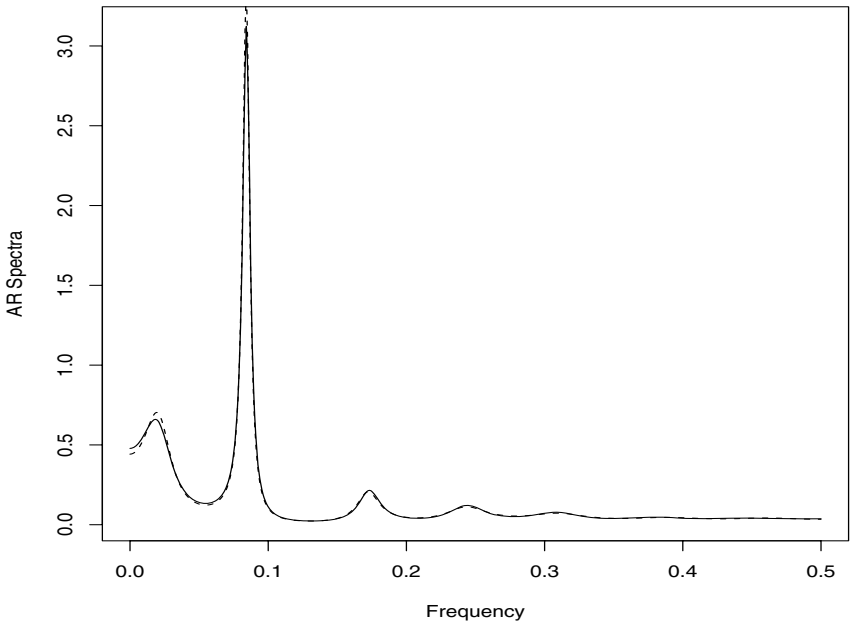


Figure 4.16 Autoregressive spectral estimators for the SOI series using models selected by AIC and SIC ($p = 15$, solid line) and by AICc ($p = 16$, dashed line). The first peak corresponds to the El Niño period of 52 months.

```
Coefficients:
       1        2        3        4        5
  0.4237   0.0803   0.1411   0.0750  -0.0446
       6        7        8        9       10
 -0.0816  -0.0686  -0.0640   0.0159   0.1099
      11       12       13       14       15
  0.1656   0.1482   0.0231  -0.1814  -0.1406

Order selected 15  sigma^2 estimated as  0.07575
```

Use the command `spec.ar(soi, order=16, log="no")` to obtain the AR(16) spectrum.

Finally, it should be mentioned that any parametric spectrum, say $f(\omega;\boldsymbol{\theta})$, depending on the vector parameter $\boldsymbol{\theta}$ can be estimated via the approximate Whittle likelihood, see Whittle (1961), using the approximate properties of the discrete Fourier transform derived in Appendix C. We have that the DFTs, $d(\omega_j)$, are approximately complex normally distributed with mean zero and variance $f(\omega_j;\boldsymbol{\theta})$ and are approximately independent for $\omega_j \neq \omega_k$. This implies

that an approximate log likelihood can be written in the form

$$\ln L(\boldsymbol{x};\boldsymbol{\theta}) \approx -\sum_{0<\omega_j<1/2}\left(\ln f_x(\omega_j;\boldsymbol{\theta}) + \frac{|d(\omega_j)|^2}{f_x(\omega_j;\boldsymbol{\theta})}\right), \tag{4.116}$$

where the sum is sometimes expanded to include the frequencies $\omega_j = 0, 1/2$. If the form with the two additional frequencies is used, the multiplier of the sum will be unity, except for the purely real points at $\omega_j = 0, 1/2$ for which the multiplier is $1/2$. For a discussion of applying the Whittle approximation to the problem of estimating parameters in an ARMA spectrum, see Anderson (1978). Although this yields valid answers, it seems more involved than simply using the time domain methods discussed in Chapter 3. The Whittle likelihood will be useful in fitting long memory models that will be discussed in Chapter 5.

4.9 Dynamic Fourier Analysis and Wavelets

If a time series, x_t, is stationary, its second-order behavior remains the same, regardless of the time t. It makes sense to match a stationary time series with sines and cosines because they, too, behave the same forever. Indeed, based on the Spectral Representation Theorem (Appendix C, §C.1), we may regard a stationary series as the superposition of sines and cosines that oscillate at various frequencies. As seen in this text, however, many time series are not stationary. Typically, the data are coerced into stationarity via transformations, or we restrict attention to parts of the data where stationarity appears to adhere. In some cases, the nonstationarity of a time series is of interest. That is to say, it is the local behavior of the process, and not the global behavior of the process, that is of concern to the investigator. As a case in point, we mention the explosion and earthquake series first presented in Example 1.7 (see Figure 1.7) and subsequently analyzed using Fourier methods in Example 4.13. The following example emphasizes the importance of dynamic (or time-frequency) Fourier analysis.

Example 4.20 Dynamic Fourier Analysis of the Explosion and Earthquake Series

Consider the earthquake and explosion series displayed in Figure 1.7. As a summary of the local behavior of these series, the estimated spectra of the P and S waves in Example 4.13 leave a lot to be desired. Figures 4.17 and 4.18 show the time-frequency analysis of the earthquake and explosion series, respectively. The idea here is to summarize the spectral behavior of the signal as it evolves over time. First, a Fourier analysis is performed on a short section of the data. Then, the section is shifted, and a Fourier analysis is performed on the new section. This process is repeated until the end of the data, and the results are plotted as in Figures 4.17 and 4.18. Specifically, in this example, let x_t, for

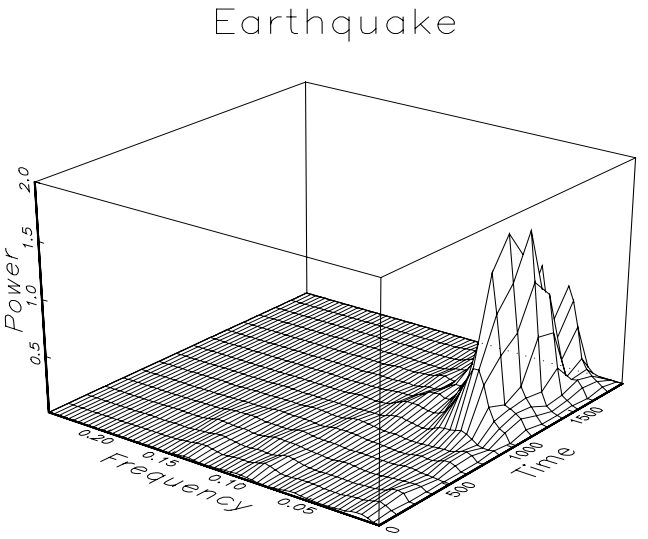


Figure 4.17 Time-frequency plot for the dynamic Fourier analysis of the earthquake series shown in Figure 1.7.

$t = 1, \ldots, 2048$, represent the series of interest. Then, the sections of the data that were analyzed were $\{x_{t_k+1}, \ldots, x_{t_k+256}\}$, for $t_k = 128k$, and $k = 0, 1, \ldots, 14$. Each section was tapered using a cosine bell, and spectral estimation was performed using a triangular set of $L = 5$ weights. The sections overlap each other, however, this practice is not necessary and sometimes not desirable; see Percival and Walden (1993, §6.17) for a further discussion of this problem.

The results of the dynamic analysis are shown as the estimated spectra (for frequencies up to $\omega = .25$) for each starting location (time), $t_k = 128k$, with $k = 0, 1, \ldots, 14$. The S component for the earthquake shows power at the low frequencies only, and the power remains strong for a long time. In contrast, the explosion shows power at higher frequencies than the earthquake, and the power of the signals (P and S waves) does not last as long as in the case of the earthquake.

The following is an R session that corresponds to a similar analysis of the explosion series in this example.

```
> eqexp = matrix(scan("/mydata/eq5exp6.dat"), ncol=2)
> ex = eqexp[,2]  # the explosion series
> ## -- dynamic spectral analysis -- ##
> nobs = length(ex)      # number of observations
```

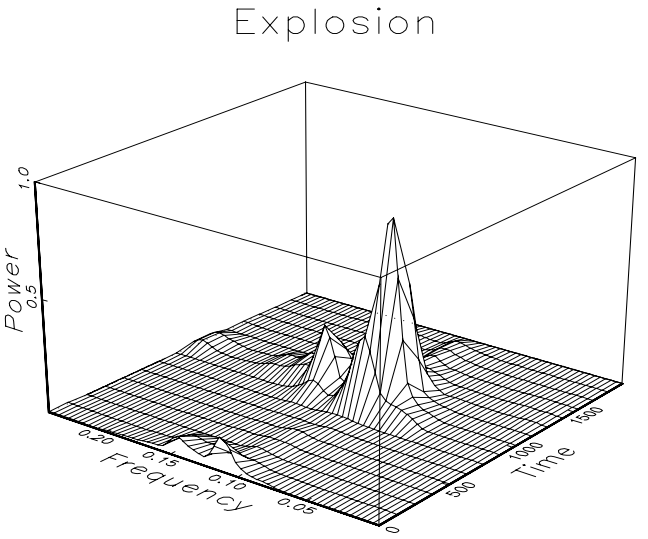


Figure 4.18 Time-frequency plot for the dynamic Fourier analysis of the explosion series shown in Figure 1.7.

```
> wsize = 256            # window size
> overlap = 128          # overlap
> ovr = wsize-overlap
> nseg = floor(nobs/ovr)-1;   # number of segments
> krnl = kernel("daniell", c(1,1))    # kernel
> ex.spec = matrix(0,wsize/2,nseg)
>  for (k in 1:nseg) {
+  a = ovr*(k-1)+1
+  b = wsize+ovr*(k-1)
+  ex.spec[,k]=spectrum(ex[a:b],krnl,taper=.5,plot=F)$spec
+  }
> ## -- plot results -- ##
> x = seq(0, .5, len = nrow(ex.spec))
> y = seq(0, ovr*nseg, len = ncol(ex.spec))
> persp(x, y, ex.spec, zlab="Power", xlab="frequency",
+        ylab="time", ticktype = "detailed", theta=25, d=2)
```

One way to view the time-frequency analysis of Example 4.20 is to consider it as being based on local transforms of the data x_t of the form

$$d_{j,k} = n^{-1/2} \sum_{t=1}^{n} x_t \psi_{j,k}(t), \tag{4.117}$$

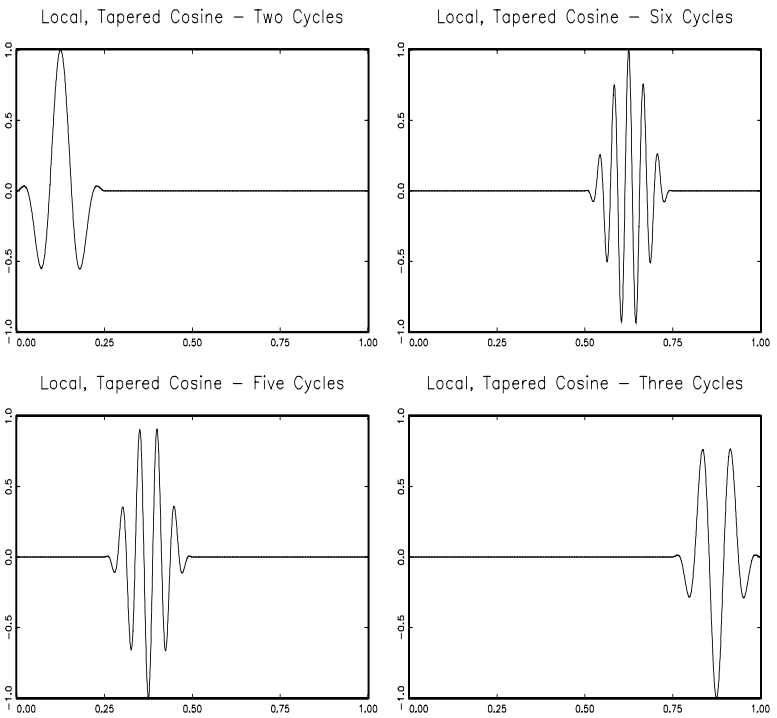


Figure 4.19 Local, tapered cosines at various frequencies.

where

$$\psi_{j,k}(t) = \begin{cases} (n/m)^{1/2} h_t \, e^{-2\pi i t j/m} & t \in [t_k + 1, t_k + m] \\ 0 & \text{otherwise} \end{cases} \tag{4.118}$$

where h_t is a taper and m is some fraction of n. In Example 4.20, $n = 2048$, $m = 256$, $t_k = 128k$, for $k = 0, 1, \ldots, 14$, and h_t was a cosine bell taper over 256 points. In (4.117) and (4.118), j indexes frequency, $\omega_j = j/m$, for $j = 1, 2, \ldots, [m/2]$, and k indexes the location, or time shift, of the transform. In this case, the transforms are based on tapered cosines and sines that have been zeroed out over various regions in time. The key point here is that the transforms are based on *local* sinusoids. Figure 4.19 shows an example of four local, tapered cosine functions at various frequencies. In that figure, the length of the data is considered to be one, and the cosines are localized to a fourth of the data length.

In addition to dynamic Fourier analysis as a method to overcome the restriction of stationarity, researchers have sought various alternative methods. A recent, and successful, alternative is wavelet analysis. A website `http://www.wavelet.org` is devoted to wavelets, which includes information about books, technical papers, software, and links to other sites. In addi-

tion, we mention the monograph on wavelets by Daubechies (1992), the text by Percival and Walden (2000), and we note that many statistical software manufacturers have wavelet modules that sit on top of their base package. In this section, we rely primarily on the S-PLUS wavelets module (with a manual written by Bruce and Gao, 1996), however, we will present some R code where possible. The basic idea of wavelet analysis is to imitate dynamic Fourier analysis, but with functions (wavelets) that may be better suited to capture the local behavior of nonstationary time series.

Wavelets come in families generated by a father wavelet, ϕ, and a mother wavelet, ψ. The father wavelets are used to capture the smooth, low-frequency nature of the data, whereas the mother wavelets are used to capture the detailed, and high-frequency nature of the data. The father wavelet integrates to one, and the mother wavelet integrates to zero

$$\int \phi(t)dt = 1 \quad \text{and} \quad \int \psi(t)dt = 0. \tag{4.119}$$

For a simple example, consider the Haar function,

$$\psi(t) = \begin{cases} 1, & 0 \le t < 1/2, \\ -1, & 1/2 \le t < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{4.120}$$

The father in this case is $\phi(t) = 1$ for $t \in [0, 1)$ and zero otherwise. The Haar functions are useful for demonstrating properties of wavelets, but they do not have good time-frequency localization properties. Figure 4.20 displays two of the more commonly used wavelets that are available with the S-PLUS wavelets module, the *daublet4* and *symmlet8* wavelets, which are described in detail in Daubechies (1992). The number after the name refers to the width and smoothness of the wavelet; for example, the symmlet10 wavelet is wider and smoother than the symmlet8 wavelet. Daublets are one of the first type of continuous orthogonal wavelets with compact support, and symmlets were constructed to be closer to symmetry than daublets. In general, wavelets do not have an analytical form, but instead they are generated using numerical methods.

Figure 4.20 was generated in S-PLUS using the wavelet module as follows:

```
> d4f <- wavelet("d4", mother=F)
> d4m <- wavelet("d4")
> s8f <- wavelet("s8", mother=F)
> s8m <- wavelet("s8")
> par(mfrow=c(2,2))
> plot(d4f)
> plot(d4m)
> plot(s8f)
> plot(s8m)
```

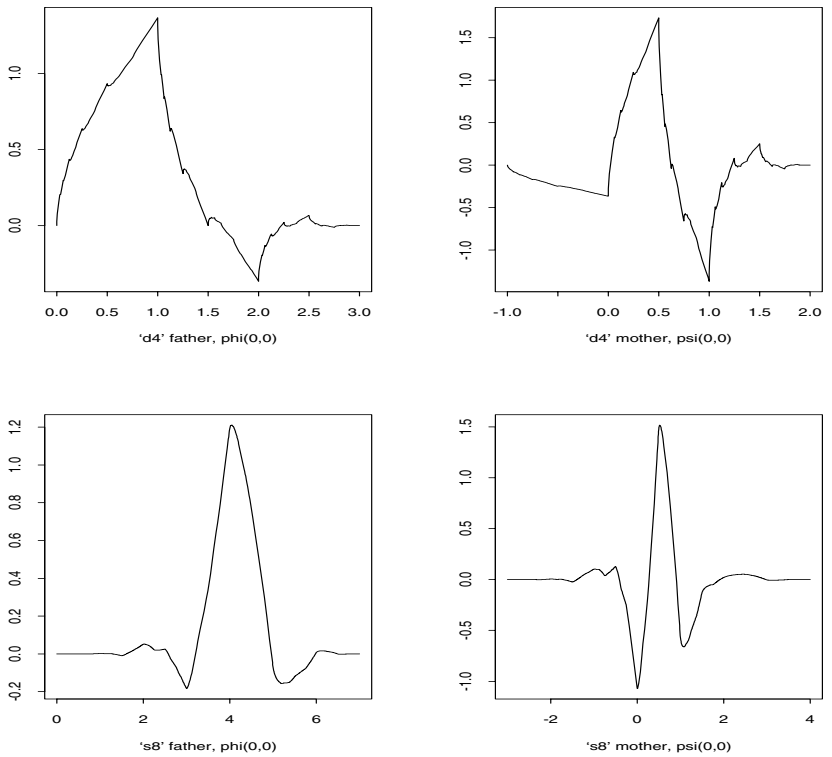


Figure 4.20 Father and mother daublet4 wavelets (top row); father and mother symmlet8 wavelets (bottom row).

It is possible to draw some wavelets in R using the `wavethresh` package. In that package, daublets are called DaubExPhase and symmlets are called DaubLeAsymm. The following R session displays some of the available wavelets (this will not reproduce Figure 4.20) and it assumes the `wavethresh` package has been downloaded into R and then loaded at the start of the session. The `filter.number` determines the width and smoothness of the wavelet.

```
> par(mfrow=c(2,2))
> draw.default(filter.number=2, family="DaubExPhase")
> draw.default(filter.number=4, family="DaubExPhase")
> draw.default(filter.number=4, family="DaubLeAsymm")
> draw.default(filter.number=9, family="DaubLeAsymm")
```

When we depart from periodic functions, such as sines and cosines, the precise meaning of frequency, or cycles per unit time, is lost. When using wavelets, we typically refer to scale rather than frequency. The orthogonal

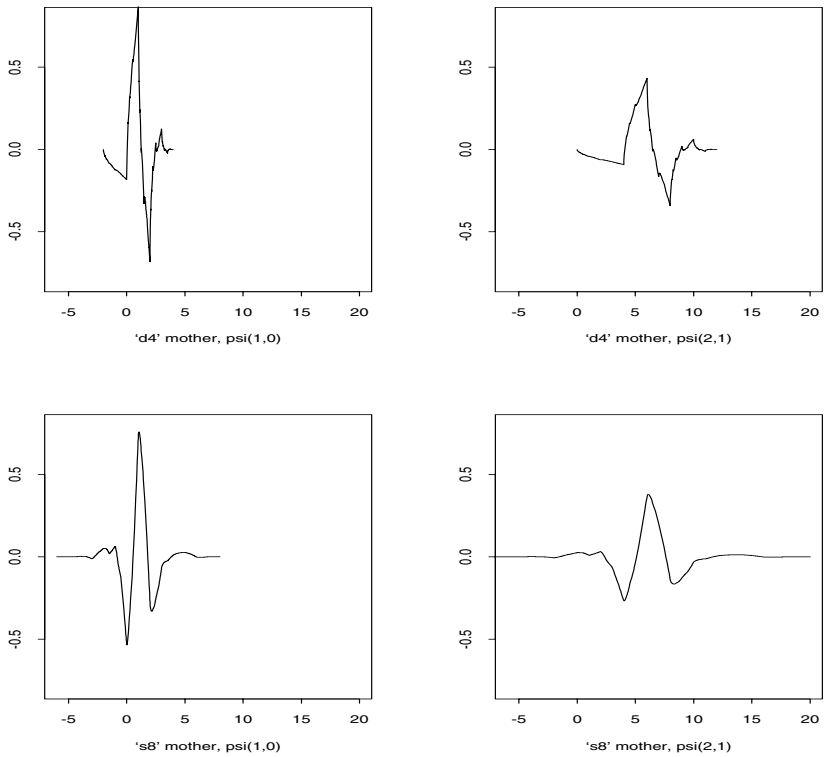


Figure 4.21 Scaled and translated daublet4 wavelets, $\psi_{1,0}(t)$ and $\psi_{2,1}(t)$ (top row); scaled and translated symmlet8 wavelets, $\psi_{1,0}(t)$ and $\psi_{2,1}(t)$ (bottom row).

wavelet decomposition of a time series, x_t, for $t = 1, \ldots, n$ is

$$\begin{aligned} x_t &= \sum_k s_{J,k}\phi_{J,k}(t) + \sum_k d_{J,k}\psi_{J,k}(t) \\ &\quad + \sum_k d_{J-1,k}\psi_{J-1,k}(t) + \cdots + \sum_k d_{1,k}\psi_{1,k}(t), \end{aligned} \tag{4.121}$$

where J is the number of scales, and k ranges from one to the number of coefficients associated with the specified component (see Example 4.21). In (4.121), the wavelet functions $\phi_{J,k}(t), \psi_{J,k}(t), \psi_{J-1,k}(t), \ldots, \psi_{1,k}(t)$ are generated from the father wavelet, $\phi(t)$, and the mother wavelet, $\psi(t)$, by translation (shift) and scaling:

$$\phi_{J,k}(t) = 2^{-J/2}\phi\left(\frac{t - 2^J k}{2^J}\right), \tag{4.122}$$

$$\psi_{j,k}(t) = 2^{-j/2}\psi\left(\frac{t - 2^j k}{2^j}\right), \quad j = 1, \ldots, J. \tag{4.123}$$

The choice of dyadic shifts and scales is arbitrary but convenient. The shift or translation parameter is $2^j k$, and scale parameter is 2^j. The wavelet functions are spread out and shorter for larger values of j (or scale parameter 2^j) and tall and narrow for small values of the scale. Figure 4.21 shows $\psi_{1,0}(t)$ and $\psi_{2,1}(t)$ generated from the daublet4 (top row), and the symmlet8 (bottom row) mother wavelets. We may think of $1/2^j$ (or 1/scale) in wavelet analysis as being the analogue of frequency ($\omega_j = j/n$) in Fourier analysis. For example, when $j = 1$, the scale parameter of 2 is akin to the Nyquist frequency of 1/2, and when $j = 6$, the scale parameter of 2^6 is akin to a low frequency ($1/2^6 \approx 0.016$). In other words, larger values of the scale refer to slower, smoother (or coarser) movements of the signal, and smaller values of the scale refer to faster, choppier (or finer) movements of the signal. Figure 4.21 was generated in S-PLUS using the wavelet module as follows:

```
> d4.1 <- wavelet("d4", level=1, shift=0)
> d4.2 <- wavelet("d4", level=2, shift=1)
> s8.1 <- wavelet("s8", level=1, shift=0)
> s8.2 <- wavelet("s8", level=2, shift=1)
> par(mfrow=c(2,2))
> plot(d4.1, ylim=c(-.8,.8), xlim=c(-6,20))
> plot(d4.2, ylim=c(-.8,.8), xlim=c(-6,20))
> plot(s8.1, ylim=c(-.8,.8), xlim=c(-6,20))
> plot(s8.2, ylim=c(-.8,.8), xlim=c(-6,20))
```

The discrete wavelet transform (DWT) of the data x_t are the coefficients $s_{J,k}$ and $d_{j,k}$ for $j = J, J-1, \ldots, 1$, in (4.121). To some degree of approximation, they are given by[15]

$$s_{J,k} = n^{-1/2} \sum_{t=1}^{n} x_t \phi_{J,k}(t), \tag{4.124}$$

$$d_{j,k} = n^{-1/2} \sum_{t=1}^{n} x_t \psi_{j,k}(t) \quad j = J, J-1, \ldots, 1. \tag{4.125}$$

It is the magnitudes of the coefficients that measure the importance of the corresponding wavelet term in describing the behavior of x_t. As in Fourier analysis, the DWT is not computed as shown but is calculated using a fast algorithm. The $s_{J,k}$ are called the smooth coefficients because they represent the smooth behavior of the data. The $d_{j,k}$ are called the detail coefficients because they tend to represent the finer, more high-frequency nature, of the data.

[15]The actual DWT coefficients are defined via a set of filters whose coefficients are close to what you would get by sampling the father and mother wavelets, but not exactly so; see the discussion surrounding Figures 471 and 478 in Percival and Walden (2000).

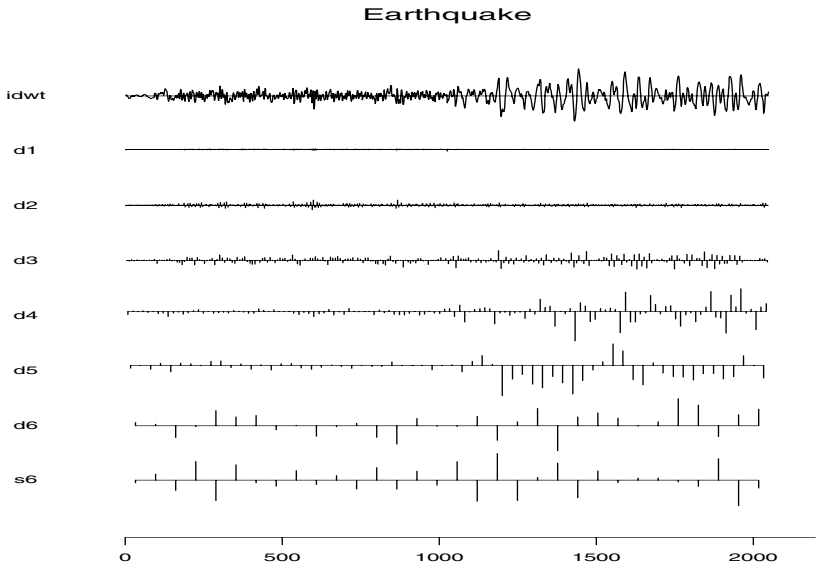


Figure 4.22 Discrete wavelet transform of the earthquake series using the symmlet8 wavelets, and $J = 6$ levels of scale.

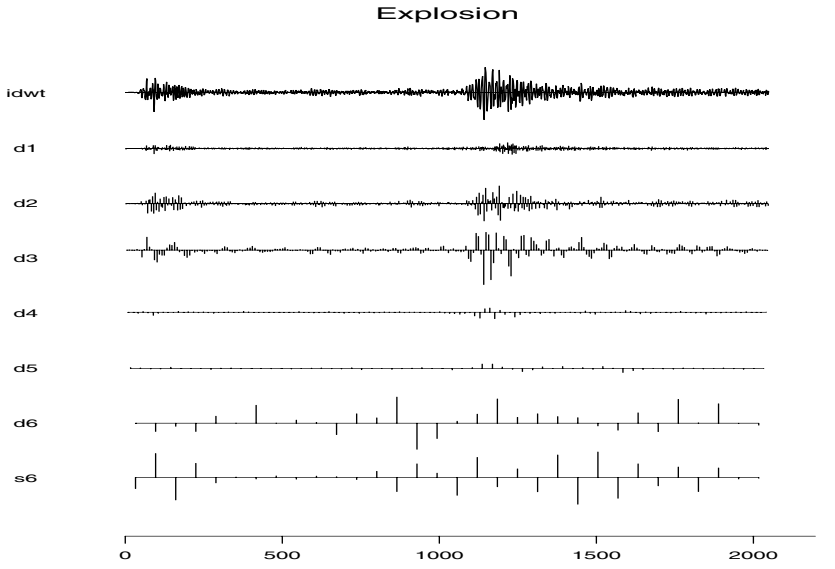


Figure 4.23 Discrete wavelet transform of the explosion series using the symmlet8 wavelets and $J = 6$ levels of scale.

Example 4.21 Wavelet Analysis of the Explosion and Earthquake Series

Figures 4.22 and 4.23 show the DWTs, based on the symmlet8 wavelet basis, for the earthquake and explosion series, respectively. Each series is of length $n = 2^{11} = 2048$, and in this example, the DWTs are calculated using $J = 6$ levels. In this case, $n/2 = 2^{10} = 1024$ values are in $d1 = \{d_{1,k};\ k = 1, \ldots, 2^{10}\}$, $n/2^2 = 2^9 = 512$ values are in $d2 = \{d_{2,k};\ k = 1, \ldots, 2^9\}$, and so on, until finally, $n/2^6 = 2^5 = 32$ values are in $d6$ and in $s6$. The detail values $d_{1,k}, \ldots, d_{6,k}$ are plotted at the same scale, and hence, the relative importance of each value can be seen from the graph. The smooth values $s_{6,k}$ are typically larger than the detail values and plotted on a different scale. The top of Figures 4.22 and 4.23 show the inverse DWT (IDWT) computed from all of the coefficients. The displayed IDWT is a reconstruction of the data, and it reproduces the data except for round-off error.

Comparing the DWTs, the earthquake is best represented by wavelets with larger scale than the explosion. One way to measure the importance of each level, $d1, d2, \ldots, d6, s6$, is to evaluate the proportion of the total power (or energy) explained by each. The total power of a time series x_t, for $t = 1, \ldots, n$, is $TP = \sum_{t=1}^{n} x_t^2$. The total power associated with each level of scale is (recall $n = 2^{11}$),

$$TP_6^s = \sum_{k=1}^{n/2^6} s_{6,k}^2 \quad \text{and} \quad TP_j^d = \sum_{k=1}^{n/2^j} d_{j,k}^2, \quad j = 1, \ldots, 6.$$

Because we are working with an orthogonal basis, we have

$$TP = TP_6^s + \sum_{j=1}^{6} TP_j^d,$$

and the proportion of the total power explained by each level of detail would be the ratios TP_j^d/TP for $j = 1, \ldots, 6$, and for the smooth level, it would be TP_6^s/TP. These values are listed in Table 4.2. From that table nearly 80% of the total power of the earthquake series is explained by the higher scale details $d4$ and $d5$, whereas 90% of the total power is explained by the smaller scale details $d2$ and $d3$ for the explosion.

Figures 4.24 and 4.25 show the time-scale plots based on the DWT of the earthquake series and the explosion series, respectively. These figures are the wavelet analog of the time-frequency plots shown in Figures 4.17 and 4.18. The power axis represents the magnitude of each value d_{jk} or $s_{6,k}$. The time axis matches the time axis in the DWTs shown in Figures 4.22 and 4.23, and the scale axis is plotted as 1/scale, listed from the coarsest scale to the finest scale. On the 1/scale axis, the coarsest scale values,

Table 4.2 Fraction of the Total Power for the DWTs of the Earthquake and the Explosion

Component	Earthquake	Explosion
s6	0.009	0.002
d6	0.043	0.002
d5	0.377	0.007
d4	0.367	0.015
d3	0.160	0.559
d2	0.040	0.349
d1	0.003	0.066

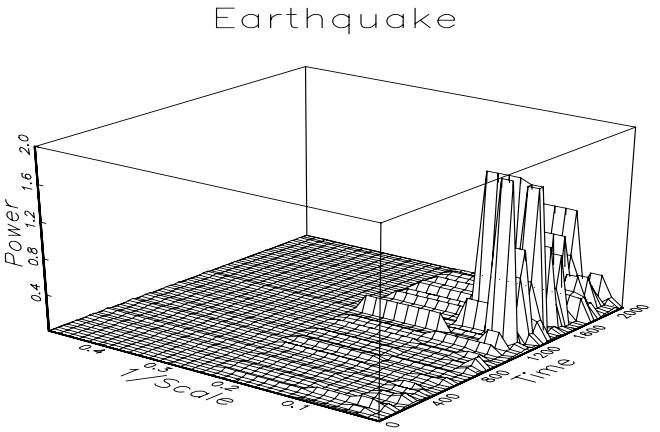


Figure 4.24 Time-scale plot of the earthquake series.

represented by the smooth coefficients $s6$, are plotted over the range $[0, 2^{-6})$, the coarsest detail values, $d6$, are plotted over $[2^{-6}, 2^{-5})$, and so on. In these figures, we did not plot the finest scale values, $d1$, so the finest scale values exhibited in Figures 4.24 and 4.25 are in $d2$, which are plotted over the range $[2^{-2}, 2^{-1})$.

The conclusions drawn from these plots are the same as those drawn from Figures 4.17 and 4.18. That is, the S wave for the earthquake shows power at the high scales (or low 1/scale) only, and the power remains strong for a long time. In contrast, the explosion shows power at smaller scales (or higher 1/scale) than the earthquake, and the power of the signals (P and S waves) do not last as long as in the case of the earthquake.

The analyses of this example were performed using the S-PLUS wavelets

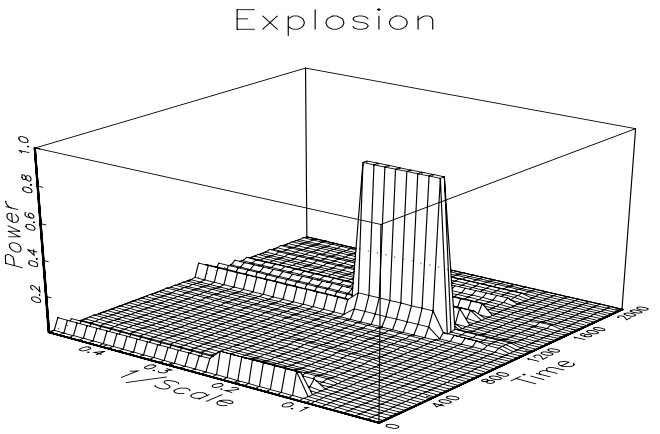


Figure 4.25 Time-scale plot of the explosion series.

module (which must be loaded prior to the analyses) as follows:

```
> eqexp <- matrix(scan("/mydata/eq5exp6.dat"), ncol=2)
> eq <- eqexp[,1]  # the earthquake series
> ex <- eqexp[,2]  # the explosion series
> eq.dwt <- dwt(eq)
> ex.dwt <- dwt(ex)
> plot(eq.dwt)
> plot(ex.dwt)
> # -- energy distributions (Table 4.2) --#
> dotchart(eq.dwt)   # a graphic
> summary(eq.dwt)    # numerical details
> dotchart(ex.dwt)
> summary(ex.dwt)
> #-- time scale plots (Figs 4.24-4.25 but not in 3d) --#
> time.scale.plot(eq.dwt)
> time.scale.plot(ex.dwt)
```

Similar analyses may be performed in R using the `wavethresh` or the `waveslim` packages. We exhibit the analysis for the earthquake series using `waveslim`, assuming it has been downloaded into R and then loaded at the start of the R session.

```
> eq.dwt = dwt(eq, n.levels=6)
> # --  plot the dwt and calculate TP -- #
> TP = matrix(0,7,1)
> par(mfcol=c(7,1), pty="m", mar=c(3,4,2,2))
> for(i in 1:6){
```

```
+  plot.ts(up.sample(eq.dwt[[i]], 2^i), type="h", axes=F,
+          ylab=names(eq.dwt)[i])
+ abline(h=0)
+ axis(side=2)
+ TP[i]=sum(eq.dwt[[i]]^2)
+ }
> plot.ts(up.sample(eq.dwt[[7]], 2^6), type="h", axes=F,
+          ylab=names(eq.dwt)[7])
> abline(h=0)
> axis(side=2)
> axis(side=1)
> TP[7]=sum(eq.dwt[[7]]^2)
> TP/sum(eq^2)              # the energy distribution
```

In the R code, we plotted the wavelet transform on different scales. To plot the ordinates of the wavelet transforms on the same scale, include a command like `ylim=c(-1.5,1.5)` in each `plot.ts()` command.

Wavelets can be used to perform nonparametric smoothing along the lines first discussed in §2.4, but with an emphasis on localized behavior. Although a considerable amount of literature exists on this topic, we will present the basic ideas. For further information, we refer the reader to Donoho and Johnstone (1994, 1995). As in §2.4, we suppose the data x_t can be written in terms of a signal plus noise model as

$$x_t = s_t + \epsilon_t. \tag{4.126}$$

The goal here is to remove the noise from the data, and obtain an estimate of the signal, s_t, without having to specify a parametric form of the signal. The technique based on wavelets is referred to as waveshrink.

The basic idea behind waveshrink is to shrink the wavelet coefficients in the DWT of x_t toward zero in an attempt to denoise the data and then to estimate the signal via (4.121) with the new coefficients. One obvious way to shrink the coefficients toward zero is to simply zero out any coefficient smaller in magnitude than some predetermined value, λ. Such a shrinkage rule is discontinuous and sometimes it is preferable to use a continuous shrinkage function. One such method, termed soft shrinkage, proceeds as follows. If the value of a coefficient is a, we set that coefficient to zero if $|a| \leq \lambda$, and to $\text{sign}(a)(|a| - \lambda)$ if $|a| > \lambda$. The choice of a shrinkage method is based on the goal of the signal extraction. This process entails choosing a value for the shrinkage threshold, λ, and we may wish to use a different threshold value, say, λ_j, for each level of scale $j = 1, \ldots, J$. One particular method that works well if we are interested in a relatively high degree of smoothness in the estimate is to choose $\lambda = \widehat{\sigma}_\epsilon \sqrt{2 \log n}$ for all scale levels, where $\widehat{\sigma}_\epsilon$ is an estimate of the scale of the noise, σ_ϵ. Typically a robust estimate of σ_ϵ is used, e.g., the median of the absolute deviations of the data from the median (MAD). For other thresholding techniques or for a better understanding of waveshrink, see

Donoho and Johnstone (1994, 1995), or the S-PLUS wavelets module manual (Bruce and Gao, 1996, Ch 6).

Example 4.22 Waveshrink Analysis of the Explosion and Earthquake Series

Figure 4.26 shows the results of a waveshrink analysis on the earthquake and explosion series. In this example, soft shrinkage was used with a universal threshold of $\lambda = \widehat{\sigma}_\epsilon \sqrt{2 \log n}$ where $\widehat{\sigma}_\epsilon$ is the MAD. Figure 4.26 displays the data x_t, the estimated signal $\widehat{s}_t$, as well as the residuals $x_t - \widehat{s}_t$. According to this analysis, the earthquake is mostly signal and characterized by prolonged energy, whereas the explosion is comprised of short bursts of energy.

Figure 4.26 was generated in S-PLUS using the wavelets module. For example, the analysis of the earthquake series was performed as follows.

```
> eq.dwt <- dwt(eq)
> eq.shrink <- waveshrink(eq.dwt, shrink.rule="universal",
+   shrink.fun="soft")
```

In R, using the `waveslim` package for the earthquake series, use the following commands.

```
> eq.dwt = dwt(eq, n.levels=6)
> eq.trsh = universal.thresh(eq.dwt, hard=F)
> eq.smo = idwt(eq.trsh)
> par(mfrow=c(3,1))
> plot.ts(eq, ylab="Earthquake", ylim=c(-.5,.5))
> plot.ts(eq.smo,ylab="Smoothed Earthquake",ylim=c(-.5,.5))
> plot.ts(eq-eq.smo, ylab="Noise", ylim=c(-.5,.5))
```

4.10 Lagged Regression Models

One of the intriguing possibilities offered by the coherence analysis of the relation between the SOI and Recruitment series discussed in Example 4.16 would be extending classical regression to the analysis of lagged regression models of the form

$$y_t = \sum_{r=-\infty}^{\infty} \beta_r x_{t-r} + v_t, \tag{4.127}$$

where v_t is a stationary noise process, x_t is the observed input series, and y_t is the observed output series. We are interested in estimating the filter coefficients β_r relating the adjacent lagged values of x_t to the output series y_t.

In the case of SOI and Recruitment series, we might identify the El Niño driving series, SOI, as the input, x_t, and y_t, the Recruitment series, as the

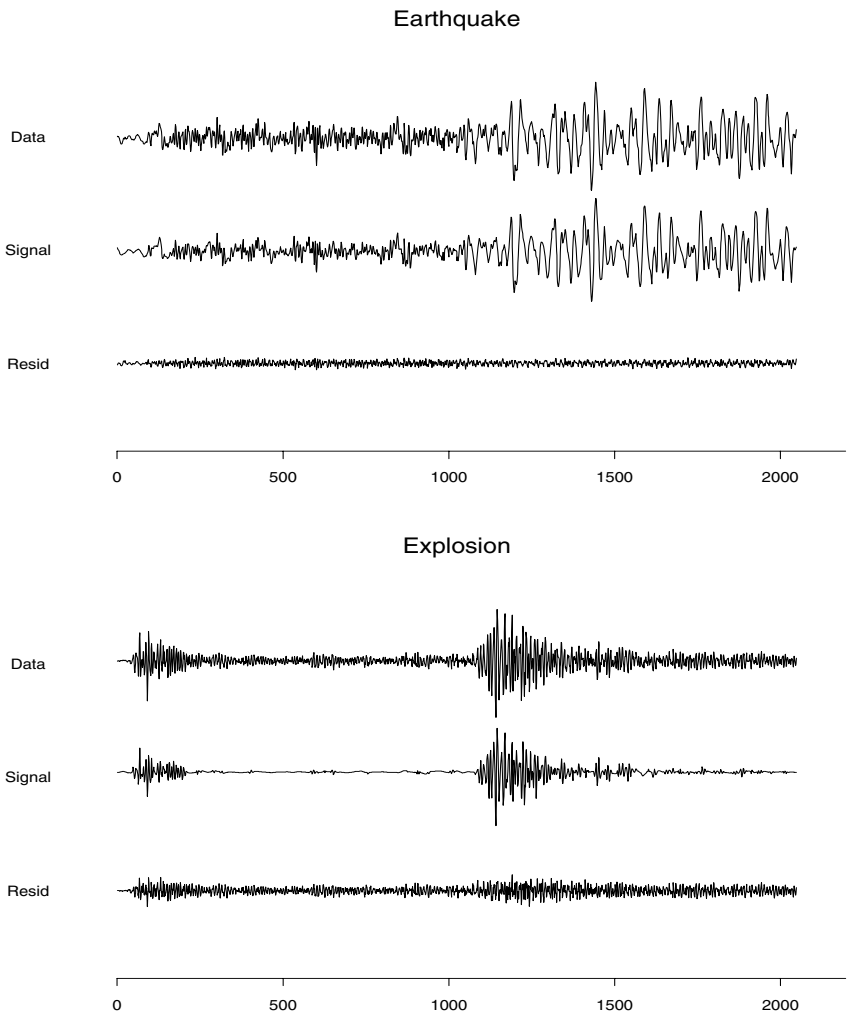


Figure 4.26 Waveshrink estimates of the earthquake signal and of the explosion signal.

output. In general, there will be more than a single possible input series and we may envision a $q \times 1$ vector of driving series. This multivariate input situation is covered in Chapter 7. The model given by (4.127) is useful under several different scenarios, corresponding to different assumptions that can be made about the components.

We assume that the inputs and outputs have zero means and are jointly stationary with the 2×1 vector process $(x_t, y_t)'$ having a spectral matrix of

the form

$$f(\omega) = \begin{pmatrix} f_{xx}(\omega) & f_{xy}(\omega) \\ f_{yx}(\omega) & f_{yy}(\omega) \end{pmatrix}. \tag{4.128}$$

Here, $f_{xy}(\omega)$ is the cross-spectrum relating the input x_t to the output y_t, and $f_{xx}(\omega)$ and $f_{yy}(\omega)$ are the spectra of the input and output series, respectively. Generally, we observe two series, regarded as input and output and search for regression functions $\{\beta_t\}$ relating the inputs to the outputs. We assume all autocovariance functions satisfy the absolute summability conditions of the form (4.31).

Then, minimizing the mean squared error

$$MSE = E\left(y_t - \sum_{r=-\infty}^{\infty} \beta_r x_{t-r}\right)^2 \tag{4.129}$$

leads to the usual orthogonality conditions

$$E\left[\left(y_t - \sum_{r=-\infty}^{\infty} \beta_r x_{t-r}\right) x_{t-s}\right] = 0 \tag{4.130}$$

for all $s = 0, \pm 1, \pm 2, \ldots$. Taking the expectations inside leads to the normal equations

$$\sum_{r=-\infty}^{\infty} \beta_r \ \gamma_{xx}(s-r) = \gamma_{yx}(s) \tag{4.131}$$

for $s = 0, \pm 1, \pm 2, \ldots$. These equations might be solved, with some effort, if the covariance functions were known exactly. If data (x_t, y_t) for $t = 1, ..., n$ are available, we might use a finite approximation to the above equations with $\widehat{\gamma}_{xx}(h)$ and $\widehat{\gamma}_{yx}(h)$ substituted into (4.131). If the regression vectors are essentially zero for $|s| \geq M/2$, and $M < n$, the system (4.131) would be of full rank and the solution would involve inverting an $(M-1) \times (M-1)$ matrix.

A frequency domain approximate solution is easier in this case for two reasons. First, the computations depend on spectra and cross-spectra that can be estimated from sample data using the techniques of §4.6. In addition, no matrices will have to be inverted, although the frequency domain ratio will have to be computed for each frequency. In order to develop the frequency domain solution, substitute the representation (4.85) into the normal equations, using the convention defined in (4.128). The left side of (4.131) can then be written in the form

$$\int_{-1/2}^{1/2} \sum_{r=-\infty}^{\infty} \beta_r \ \mathrm{e}^{2\pi i\omega(s-r)} \ f_{xx}(\omega) \ d\omega = \int_{-1/2}^{1/2} \mathrm{e}^{2\pi i\omega s} B(\omega) f_{xx}(\omega) \ d\omega,$$

where

$$B(\omega) = \sum_{r=-\infty}^{\infty} \beta_r \ \mathrm{e}^{-2\pi i\omega r} \tag{4.132}$$

is the Fourier transform of the regression coefficients β_t. Now, because $\gamma_{yx}(s)$ is the inverse transform of the cross-spectrum $f_{yx}(\omega)$, we might write the system of equations in the frequency domain, using the uniqueness of the Fourier transform, as

$$B(\omega)f_{xx}(\omega) = f_{yx}(\omega), \tag{4.133}$$

which then become the analogs of the usual normal equations. Then, we may take

$$\widehat{B}(\omega_k) = \frac{\widehat{f}_{yx}(\omega_k)}{\widehat{f}_{xx}(\omega_k)} \tag{4.134}$$

as the estimator for the Fourier transform of the regression coefficients, evaluated at some subset of fundamental frequencies $\omega_k = k/M$ with $M << n$. Generally, we assume smoothness of $B(\cdot)$ over intervals of the form $\{\omega_k + \ell/n;\ \ell = -(L-1)/2, \ldots, (L-1)/2\}$. The inverse transform of the function $\widehat{B}(\omega)$ would give $\widehat{\beta}_t$, and we note that the discrete time approximation can be taken as

$$\widehat{\beta}_t = M^{-1} \sum_{k=0}^{M-1} \widehat{B}(\omega_k) \mathrm{e}^{2\pi i \omega_k t} \tag{4.135}$$

for $t = 0, \pm 1, \pm 2, \ldots, \pm(M/2 - 1)$. If we were to use (4.135) to define $\widehat{\beta}_t$ for $|t| \geq M/2$, we would end up with a sequence of coefficients that is periodic with a period of M. In practice we define $\widehat{\beta}_t = 0$ for $|t| \geq M/2$ instead. Problem 4.32 explores the error resulting from this approximation.

Example 4.23 Lagged Regression Results for SOI and Recruitment Series

The high coherence between the SOI and Recruitment series noted in Example 4.16 suggests a lagged regression relation between the two series. A natural direction for the implication in this situation is implied because we feel that the sea surface temperature or SOI should be the input and the Recruitment series should be the output. With this in mind, let x_t be the SOI series and y_t the Recruitment series.

Although we think naturally of the SOI as the input and the Recruitment as the output, two input-output configurations are of interest. With SOI as the input, the model is

$$y_t = \sum_{r=-\infty}^{\infty} a_r x_{t-r} + w_t$$

whereas a model that reverses the two roles would be

$$x_t = \sum_{r=-\infty}^{\infty} b_r y_{t-r} + v_t,$$

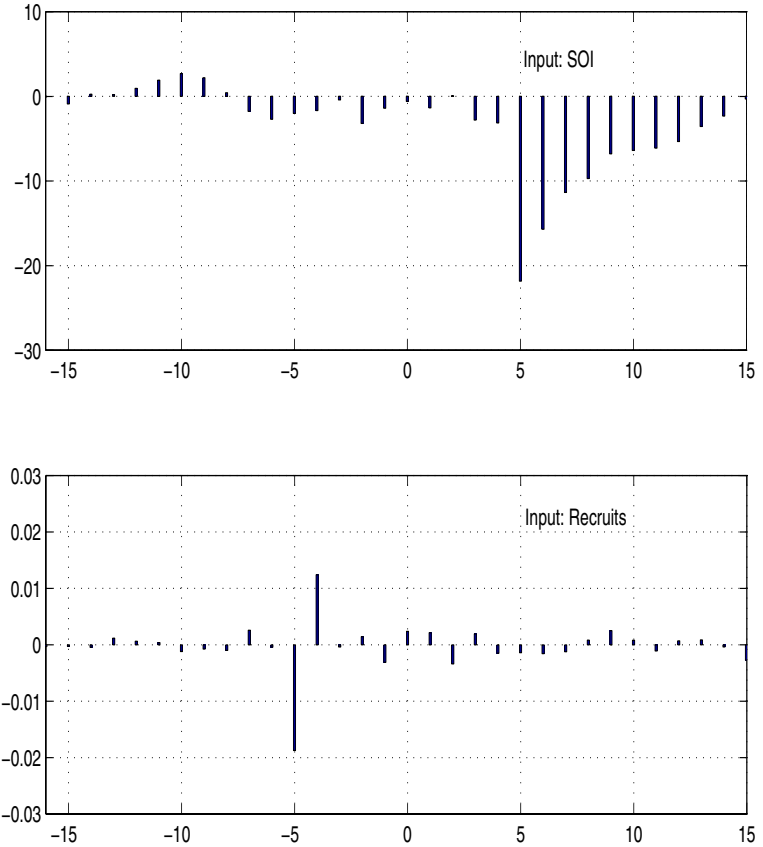


Figure 4.27 Estimated impulse response functions relating SOI to Recruitment (top) and Recruitment to SOI (bottom) $L = 15, M = 32$.

where w_t and v_t are white noise processes. Even though there is no plausible environmental explanation for the second of these two models, displaying both possibilities helps to settle on a parsimonious transfer function model. The two estimated regression or impulse response functions with $M = 32$ and $L = 15$ are shown in Figure 4.27. Note the negative peak at a lag of five points in the first of the two situations where the SOI series is assumed to be the input. The fall-off after lag five seems to be approximately exponential. A possible model for this situation is

$$y_t = -22x_{t-5} - 15x_{t-6} - 11x_{t-7} - 10x_{t-8} - 7x_{t-9} - \ldots + w_t.$$

If we examine the inverse relation, namely, a regression model with the Recruitment series y_t as the input, we get a much simpler model that

seems to depend on only two coefficients, namely,

$$x_t = .012y_{t+4} - .018y_{t+5} + v_t,$$

or, shifting by five points and transposing,

$$y_t = .667y_{t-1} - 56x_{t-5} + \epsilon_t,$$

where ϵ_t is white noise. Using the backshift operator, we may write

$$(1 - .667B)y_t = -56B^5 x_t + \epsilon_t.$$

The analysis of this example was performed using the time series package ASTSA, which is available for download from the website of this text.

The example shows we can get a clean estimator for the transfer functions relating the two series if the coherence $\widehat{\rho}^2_{xy}(\omega)$ is large. The reason is that we can write the minimized mean squared error (4.129) as

$$MSE = E\Big[\big(y_t - \sum_{r=-\infty}^{\infty} \beta_r x_{t-r}\big) y_t\Big] = \gamma_{yy}(0) - \sum_{r=-\infty}^{\infty} \beta_r \gamma_{xy}(-r),$$

using the result about the orthogonality of the data and error term in the Projection theorem. Then, substituting the spectral representations of the autocovariance and cross-covariance functions and identifying the Fourier transform (4.132) in the result leads to

$$\begin{aligned} MSE &= \int_{-1/2}^{1/2} [f_{yy}(\omega) - B(\omega) f_{xy}(\omega)]\, d\omega \\ &= \int_{-1/2}^{1/2} f_{yy}(\omega)[1 - \rho^2_{yx}(\omega)] d\omega, \end{aligned} \tag{4.136}$$

where $\rho^2_{yx}(\omega)$ is just the squared coherence given by (4.83). The similarity of (4.136) to the usual mean square error that results from predicting y from x is obvious. In that case, we would have

$$E(y - \beta x)^2 = \sigma_y^2(1 - \rho^2_{xy})$$

for jointly distributed random variables x and y with zero means, variances σ_x^2 and σ_y^2, and covariance $\sigma_{xy} = \rho_{xy}\sigma_x\sigma_y$. Because the mean squared error in (4.136) satisfies $MSE \geq 0$ with $f_{yy}(\omega)$ a non-negative function, it follows that the coherence satisfies

$$0 \leq \rho^2_{xy}(\omega) \leq 1$$

for all ω. Furthermore, Problem 4.33 shows the squared coherence is one when the output are linearly related by the filter relation (4.127), and there

is no noise, i.e., $v_t = 0$. Hence, the multiple coherence gives a measure of the association or correlation between the input and output series as a function of frequency.

The matter of verifying that the F-distribution claimed for (4.93) will hold when the sample coherence values are substituted for theoretical values still remains. Again, the form of the F-statistic is exactly analogous to the usual t-test for no correlation in a regression context. We give an argument leading to this conclusion later using the results in Appendix C, §C.3. Another question that has not been resolved in this section is the extension to the case of multiple inputs $x_{t1}, x_{t2}, \ldots, x_{tq}$. Often, more than just a single input series is present that can possibly form a lagged predictor of the output series y_t. An example is the cardiovascular mortality series that depended on possibly a number of pollution series and temperature. We discuss this particular extension as a part of the multivariate time series techniques considered in Chapter 7.

4.11 Signal Extraction and Optimum Filtering

A model closely related to regression can be developed by assuming again that

$$y_t = \sum_{r=-\infty}^{\infty} \beta_r x_{t-r} + v_t, \tag{4.137}$$

but where the βs are known and x_t is some unknown random signal that is uncorrelated with the noise process v_t. In this case, we observe only y_t and are interested in an estimator for the signal x_t of the form

$$\widehat{x}_t = \sum_{r=-\infty}^{\infty} a_r y_{t-r}. \tag{4.138}$$

In the frequency domain, it is convenient to make the additional assumptions that the series x_t and v_t are both mean-zero stationary series with spectra $f_{xx}(\omega)$ and $f_{vv}(\omega)$, often referred to as the signal spectrum and noise spectrum, respectively. Often, the special case $\beta_t = \delta_t$, in which δ_t is the Kronecker delta, is of interest because (4.137) reduces to the simple signal plus noise model

$$y_t = x_t + v_t \tag{4.139}$$

in that case. In general, we seek the set of filter coefficients a_t that minimize the mean squared error of estimation, say,

$$MSE = E[(x_t - \sum_{r=-\infty}^{\infty} a_r y_{t-r})^2]. \tag{4.140}$$

This problem was originally solved by Kolmogorov (1941) and by Wiener (1949), who derived the result in 1941 and published it in classified reports during World War II.

We can apply the orthogonality principle to write

$$E[(x_t - \sum_{r=-\infty}^{\infty} a_r y_{t-r}) y_{t-s}] = 0$$

for $s = 0, \pm 1, \pm 2, \ldots$, which leads to

$$\sum_{r=-\infty}^{\infty} a_r \gamma_{yy}(s-r) = \gamma_{xy}(s),$$

to be solved for the filter coefficients. Substituting the spectral representations for the autocovariance functions into the above and identifying the spectral densities through the uniqueness of the Fourier transform produces

$$A(\omega) f_{yy}(\omega) = f_{xy}(\omega), \tag{4.141}$$

where $A(\omega)$ and the optimal filter a_t are Fourier transform pairs, as in Definition 4.1 for $B(\omega)$ and β_t. Now, a special consequence of the model is that (see Problem 4.23)

$$f_{xy}(\omega) = \overline{B(\omega)} f_{xx}(\omega) \tag{4.142}$$

and

$$f_{yy}(\omega) = |B(\omega)|^2 f_{xx}(\omega) + f_{vv}(\omega), \tag{4.143}$$

implying the optimal filter would be Fourier transform of

$$A(\omega) = \frac{\overline{B(\omega)}}{\left(|B(\omega)|^2 + \frac{f_{vv}(\omega)}{f_{xx}(\omega)}\right)}, \tag{4.144}$$

where the second term in the denominator is just the inverse of the signal to noise ratio, say,

$$\mathrm{SNR}(\omega) = \frac{f_{xx}(\omega)}{f_{vv}(\omega)}. \tag{4.145}$$

The result shows the optimum filters can be computed for this model if the signal and noise spectra are both known or if we can assume knowledge of the signal-to-noise ratio $\mathrm{SNR}(\omega)$ as function of frequency. In Chapter 7, we show some methods for estimating these two parameters in conjunction with random effects analysis of variance models, but we assume here that it is possible to specify the signal-to-noise ratio *a priori*. If the signal-to-noise ratio is known, the optimal filter can be computed by the inverse transform of the function $A(\omega)$. It is more likely that the inverse transform will be intractable and a finite filter approximation like that used in the previous section can be applied to the data. In this case, we will have

$$a_t^M = M^{-1} \sum_{k=0}^{M-1} A(\omega_k) \mathrm{e}^{2\pi i \omega_k t} \tag{4.146}$$

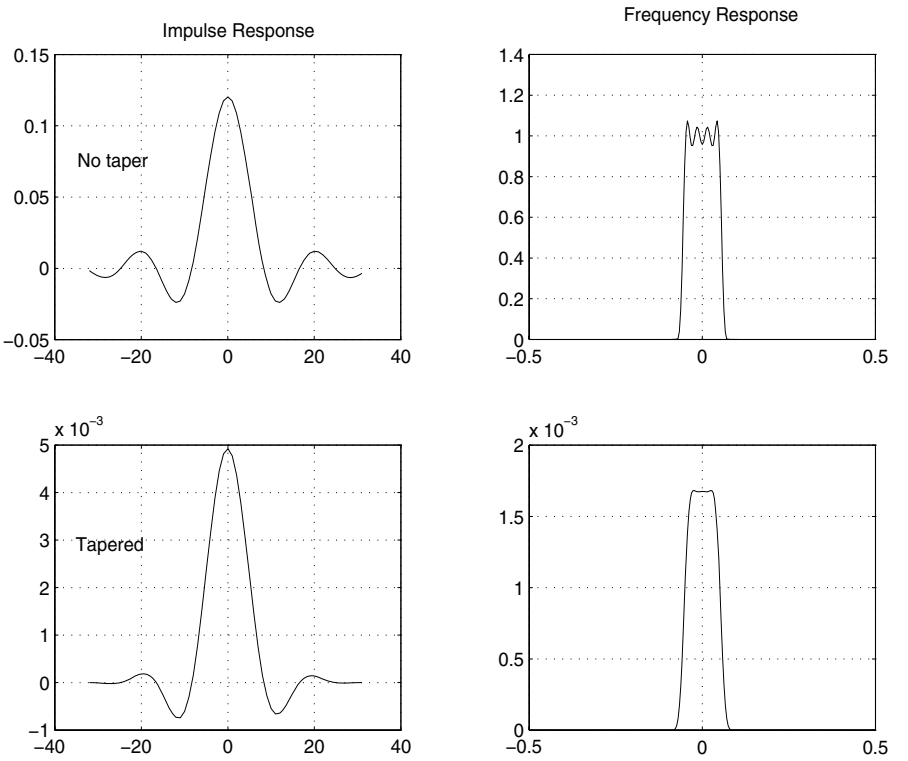


Figure 4.28 Impulse Response and frequency response functions for designed SOI filters. Note the ripples in the top panel frequency response of the untapered filter.

as the estimated filter function. It will often be the case that the form of the specified frequency response will have some rather sharp transitions between regions where the signal-to-noise ratio is high and regions where there is little signal. In these cases, the shape of the frequency response function will have ripples that can introduce frequencies at different amplitudes. An aesthetic solution to this problem is to introduce tapering as was done with spectral estimation in (4.61)-(4.68). We use below the tapered filter $\tilde{a}_t = h_t a_t$ where h_t is the cosine taper given in (4.68). The squared frequency response of the resulting filter will be $|\tilde{A}(\omega)|^2$, where

$$\tilde{A}(\omega) = \sum_{t=-\infty}^{\infty} a_t h_t \mathrm{e}^{-2\pi i \omega t}. \tag{4.147}$$

The results are illustrated in the following example that extracts the El Niño component of the sea surface temperature series.

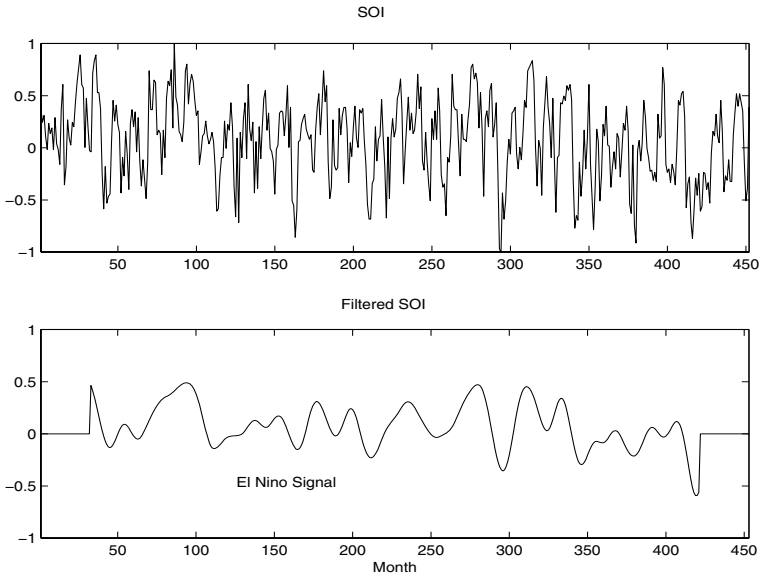


Figure 4.29 Original SOI series (top) compared to filtered version showing the estimated El Niño temperature signal (bottom).

Example 4.24 Estimating the El Niño Signal Using Optimal Filters

Figure 4.5 shows the spectrum of the SOI series, and we note that essentially two components have power, the El Niño frequency of about .02 cycles per month and a yearly frequency of about .08 cycles per month. We assume, for this example, that we wish to preserve the lower frequency as signal and to eliminate the higher order frequencies. In this case, we assume the simple signal plus noise model

$$y_t = x_t + v_t,$$

so that there is no convolving function β_t. Furthermore, the signal-to-noise ratio is assumed to be high to about .06 cycles per month and zero thereafter. The optimal frequency response was assumed to be unity to .05 cycles per point and then to decay linearly to zero in several steps. Figure 4.28 shows the Fourier transform, (4.146), at $M = 64$ frequencies, say, a_t^M and the tapered version $h_t a_t^M$. The estimated squared frequency response, approximated as a long (256 point) transform of the form (4.147), has ripples when tapering is not applied and is relatively smooth for the tapered filter. Figure 4.28 shows both positive and negative frequencies. Figure 4.29 shows the original and filtered SOI index, and we see a smooth extracted signal that conveys the essence of the

underlying El Niño signal. The frequency response of the designed filter can be compared with that of the symmetric 12-month moving average applied to the same series in Example 4.17. The filtered series, shown in Figure 4.3, shows a good deal of higher frequency chatter riding on the smoothed version, which has been introduced by the higher frequencies that leak through in the squared frequency response, as in Figure 4.14.

The analysis of this example was performed using the time series package ASTSA, which is available for download from the website of this text.

The design of finite filters with a specified frequency response requires some experimentation with various target frequency response functions and we have only touched on the methodology here. The filter designed here, sometimes called a low-pass filter reduces the high frequencies and keeps or passes the low frequencies. Alternately, we could design a high-pass filter to keep high frequencies if that is where the signal is located. An example of a simple high-pass filter is the first difference with a frequency response that is shown in Figure 4.14. We can also design band-pass filters that keep frequencies in specified bands. For example, seasonal adjustment filters are often used in economics to reject seasonal frequencies while keeping both high frequencies, lower frequencies, and trend (see, for example, Grether and Nerlove, 1970).

The filters we have discussed here are all symmetric two-sided filters, because the designed frequency response functions were purely real. Alternatively, we may design recursive filters to produce a desired response. An example of a recursive filter is one that replaces the input x_t by the filtered output

$$y_t = \sum_{k=1}^{p} \phi_k y_{t-k} + x_t - \sum_{k=1}^{q} \theta_k x_{t-k}. \tag{4.148}$$

Note the similarity between (4.148) and the ARIMA$(p, 1, q)$ model, in which the white noise component is replaced by the input. Transposing the terms involving y_t and using the basic linear filter result in Property 4.4 leads to

$$f_y(\omega) = \frac{|\theta(\mathrm{e}^{-2\pi i\omega})|^2}{|\phi(\mathrm{e}^{-2\pi i\omega})|^2} f_x(\omega), \tag{4.149}$$

where

$$\phi(\mathrm{e}^{-2\pi i\omega}) = 1 - \sum_{k=1}^{p} \phi_k \mathrm{e}^{-2\pi ik\omega}$$

and

$$\theta(\mathrm{e}^{-2\pi i\omega}) = 1 - \sum_{k=1}^{q} \theta_k \mathrm{e}^{-2\pi ik\omega}.$$

Recursive filters such as those given by (4.149) distort the phases of arriving frequencies, and we do not consider the problem of designing such filters in any detail.

4.12 Spectral Analysis of Multidimensional Series

Multidimensional series of the form $x_{\boldsymbol{s}}$, where $\boldsymbol{s} = (s_1, s_2, \ldots, s_r)'$ is an r-dimensional vector of spatial coordinates or a combination of space and time coordinates, were introduced in §1.7. The example given there, shown in Figure 1.15, was a collection of temperature measurements taking on a rectangular field. This data would form a two-dimensional process, indexed by row and column in space. In that section, the multidimensional autocovariance function of an r-dimensional stationary series was given as $\gamma_x(\boldsymbol{h}) = E[x_{\boldsymbol{s}+\boldsymbol{h}} x_{\boldsymbol{s}}]$, where the multidimensional lag vector is $\boldsymbol{h} = (h_1, h_2, \ldots, h_r)'$.

The multidimensional wavenumber spectrum is given as the Fourier transform of the autocovariance, namely,

$$f_x(\boldsymbol{\omega}) = \sum_{\boldsymbol{h}} \gamma_x(\boldsymbol{h}) \mathrm{e}^{-2\pi i \boldsymbol{\omega}' \boldsymbol{h}}. \tag{4.150}$$

Again, the inverse result

$$\gamma_x(\boldsymbol{h}) = \int_{-1/2}^{1/2} f_x(\boldsymbol{\omega}) e^{2\pi i \boldsymbol{\omega}' \boldsymbol{h}} d\boldsymbol{\omega} \tag{4.151}$$

holds, where the integral is over the multidimensional range of the vector $\boldsymbol{\omega}$. The wavenumber argument is exactly analogous to the frequency argument, and we have the corresponding intuitive interpretation as the cycling rate ω_i per distance traveled s_i in the i-th direction.

Two-dimensional processes occur often in practical applications, and the representations above reduce to

$$f_x(\omega_1, \omega_2) = \sum_{h_1=-\infty}^{\infty} \sum_{h_2=-\infty}^{\infty} \gamma_x(h_1, h_2) \mathrm{e}^{-2\pi i(\omega_1 h_1 + \omega_2 h_2)} \tag{4.152}$$

and

$$\gamma_x(h_1, h_2) = \int_{-1/2}^{1/2} \int_{-1/2}^{1/2} f_x(\omega_1, \omega_2) \mathrm{e}^{2\pi i(\omega_1 h_1 + \omega_2 h_2)} d\omega_1 \, d\omega_2 \tag{4.153}$$

in the case $r = 2$. The notion of linear filtering generalizes easily to the two-dimensional case by defining the impulse response function a_{s_1,s_2} and the spatial filter output as

$$y_{s_1,s_2} = \sum_{u_1} \sum_{u_2} a_{u_1,u_2} x_{s_1-u_1,s_2-u_2}. \tag{4.154}$$

The spectrum of the output of this filter can be derived as

$$f_y(\omega_1, \omega_2) = |A(\omega_1, \omega_2)|^2 f_x(\omega_1, \omega_2), \tag{4.155}$$

where

$$A(\omega_1, \omega_2) = \sum_{u_1} \sum_{u_2} a_{u_1,u_2} \mathrm{e}^{-2\pi i(\omega_1 u_1 + \omega_2 u_2)}. \tag{4.156}$$

These results are analogous to those in the one-dimensional case, described by Property P4.4.

The multidimensional DFT is also a straightforward generalization of the univariate expression. In the two-dimensional case with data on a rectangular grid, $\{x_{s_1,s_2};\ s_1 = 1, ..., n_1,\ s_2 = 1, ..., n_2\}$, we will write, for $-1/2 \le \omega_1, \omega_2 \le 1/2$,

$$d(\omega_1, \omega_2) = (n_1 n_2)^{-1/2} \sum_{s_1=1}^{n_1} \sum_{s_2=1}^{n_2} x_{s_1,s_2} \mathrm{e}^{-2\pi i(\omega_1 s_1 + \omega_2 s_2)} \tag{4.157}$$

as the two-dimensional DFT, where the frequencies ω_1, ω_2 are evaluated at multiples of $(1/n_1, 1/n_2)$ on the spatial frequency scale. The two-dimensional wavenumber spectrum can be estimated by the smoothed sample wavenumber spectrum

$$\bar{f}_x(\omega_1, \omega_2) = (L_1 L_2)^{-1} \sum_{\ell_1,\ell_2} \left| d(\omega_1 + \ell_1/n_1, \omega_2 + \ell_2/n_2) \right|^2, \tag{4.158}$$

where the sum is taken over the grid $\{-m_j \le \ell_j \le m_j;\ j = 1, 2\}$, where $L_1 = 2m_1 + 1$ and $L_2 = 2m_2 + 1$. The statistic

$$\frac{2L_1 L_2 \bar{f}_x(\omega_1, \omega_2)}{f_x(\omega_1, \omega_2)} \dot{\sim} \chi^2_{2L_1 L_2} \tag{4.159}$$

can be used to set confidence intervals or make approximate tests against a fixed assumed spectrum $f_0(\omega_1, \omega_2)$. We may also extend this analysis to weighted estimation and window estimation as discussed in §4.5.

Example 4.25 Wavenumber Spectrum of Soil Surface Temperatures

As an example, consider the periodogram of the two-dimensional temperature series shown in Figure 1.15 and analyzed by Bazza et al. (1988). We recall the spatial coordinates in this case will be (s_1, s_2), which define the spatial coordinates rows and columns so that the frequencies in the two directions will be expressed as cycles per row and cycles per column. Figure 4.30 shows the periodogram of the two-dimensional temperature series, and we note the ridge of strong spectral peaks running over rows at a column frequency of zero. An obvious periodic component appears at frequencies of .0625 and −.0625 cycles per row, which corresponds to 16 rows or about 272 ft. On further investigation of previous irrigation patterns over this field, treatment levels of salt varied periodically over columns. This analysis is extended in Problem 4.17, where we recover the salt treatment profile over rows and compare it to a signal, computed by averaging over columns.

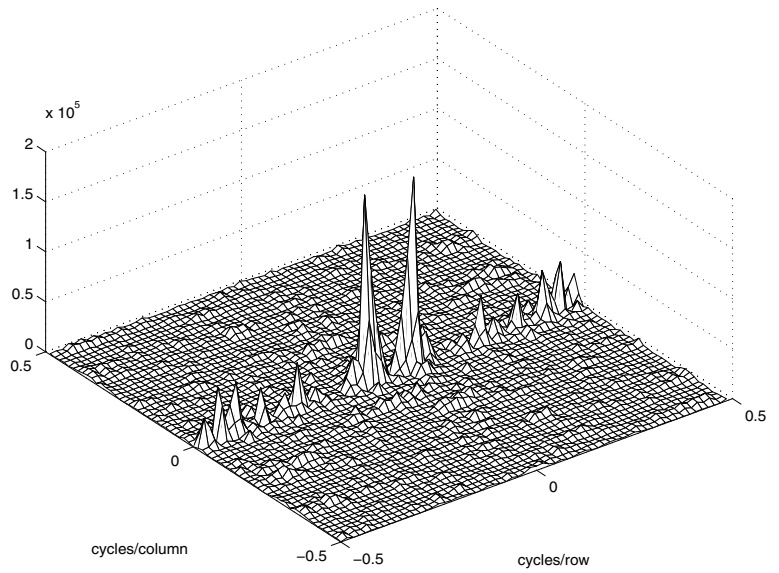


Figure 4.30 Two-dimensional periodogram of soil temperature profile showing peak at .0625 cycles/row. The period is 16 rows, and this corresponds to 16×17 ft = 272 ft.

Another application of two-dimensional spectral analysis of agricultural field trials is given in McBratney and Webster (1981), who used it to detect ridge and furrow patterns in yields. The requirement for regular, equally spaced samples on fairly large grids has tended to limit enthusiasm for strict two-dimensional spectral analysis. An exception is when a propagating signal from a given velocity and azimuth is present so predicting the wavenumber spectrum as a function of velocity and azimuth becomes feasible (see Shumway et al., 1999).

Problems

Section 4.2

4.1 Repeat the simulations and analyses in Examples 4.1 and 4.2 with the following changes:

(a) Change the sample size to $n = 128$ and generate and plot the same series as in Example 4.1:

$$\begin{aligned} x_{t1} &= 2\cos(2\pi t\, 6/100) + 3\sin(2\pi t\, 6/100), \\ x_{t2} &= 4\cos(2\pi t\, 10/100) + 5\sin(2\pi t\, 10/100), \end{aligned}$$

$$\begin{aligned} x_{t3} &= 6\cos(2\pi t\, 40/100) + 7\sin(2\pi t\, 40/100), \\ x_t &= x_{t1} + x_{t2} + x_{t3}. \end{aligned}$$

What is the major difference between these series and the series generated in Example 4.1? (Hint: The answer is *fundamental*. But if your answer is the series are longer, you may be punished severely.)

(b) As in Example 4.2, compute and plot the periodogram of the series, x_t, generated in (a) and comment.

(c) Repeat the analyses of (a) and (b) but with $n = 100$ (as in Example 4.1), and adding noise to x_t; that is

$$x_t = x_{t1} + x_{t2} + x_{t3} + w_t$$

where $w_t \sim$ iid $\mathrm{N}(0, 25)$. That is, you should simulate and plot the data, and then plot the periodogram of x_t and comment.

4.2 With reference to equations (4.2) and (4.3), let $Z_1 = U_1$ and $Z_2 = -U_2$ be independent, standard normal variables. Consider the polar coordinates of the point (Z_1, Z_2), that is,

$$A^2 = Z_1^2 + Z_2^2 \quad \text{and} \quad \phi = \tan^{-1}(Z_2/Z_1).$$

(a) Find the joint density of A^2 and ϕ, and from the result, conclude that A^2 and ϕ are independent random variables, where A^2 is a chi-squared random variable with 2 df, and ϕ is uniformly distributed on $(-\pi, \pi)$.

(b) Going in reverse from polar coordinates to rectangular coordinates, suppose we assume that A^2 and ϕ are independent random variables, where A^2 is chi-squared with 2 df, and ϕ is uniformly distributed on $(-\pi, \pi)$. With $Z_1 = A\cos(\phi)$ and $Z_2 = A\sin(\phi)$, where A is the positive square root of A^2, show that Z_1 and Z_2 are independent, standard normal random variables.

4.3 Verify (4.5).

Section 4.3

4.4 A time series was generated by first drawing the white noise series w_t from a normal distribution with mean zero and variance one. The observed series x_t was generated from

$$x_t = w_t - \theta w_{t-1}, \quad t = 0, \pm 1, \pm 2, \ldots,$$

where θ is a parameter.

(a) Derive the theoretical mean value and autocovariance functions for the series x_t and w_t. Are the series x_t and w_t stationary? Give your reasons.

(b) Give a formula for the power spectrum of x_t, expressed in terms of θ and ω.

4.5 A first-order autoregressive model is generated from the white noise series w_t using the generating equations

$$x_t = \phi x_{t-1} + w_t,$$

where ϕ, for $|\phi| < 1$, is a parameter and the w_t are independent random variables with mean zero and variance σ_w^2.

(a) Show the power spectrum of x_t is given by

$$f_x(\omega) = \frac{\sigma_w^2}{1 + \phi^2 - 2\phi\cos(2\pi\omega)}.$$

(b) Verify the autocovariance function of this process is

$$\gamma_x(h) = \frac{\sigma_w^2\ \phi^{|h|}}{1 - \phi^2},$$

$h = 0, \pm 1, \pm 2, \ldots$, by showing that the inverse transform of $\gamma_x(h)$ is the spectrum derived in part (a).

4.6 In applications, we will often observe series containing a signal that has been delayed by some unknown time D, i.e.,

$$x_t = s_t + As_{t-D} + n_t,$$

where s_t and n_t are stationary and independent with zero means and spectral densities $f_s(\omega)$ and $f_n(\omega)$, respectively. The delayed signal is multiplied by some unknown constant A.

(a) Prove

$$f_x(\omega) = [1 + A^2 + 2A\cos(2\pi\omega D)]f_s(\omega) + f_n(\omega).$$

(b) How could the periodicity expected in the spectrum derived in (a) be used to estimate the delay D? (Hint: Consider the case where $f_n(\omega) = 0$; i.e., there is no noise.)

4.7 Suppose x_t and y_t are stationary zero-mean time series with x_t independent of y_s for all s and t. Consider the product series

$$z_t = x_t y_t.$$

Prove the spectral density for z_t can be written as

$$f_z(\omega) = \int_{-1/2}^{1/2} f_x(\omega - \nu) f_y(\nu)\ d\nu.$$

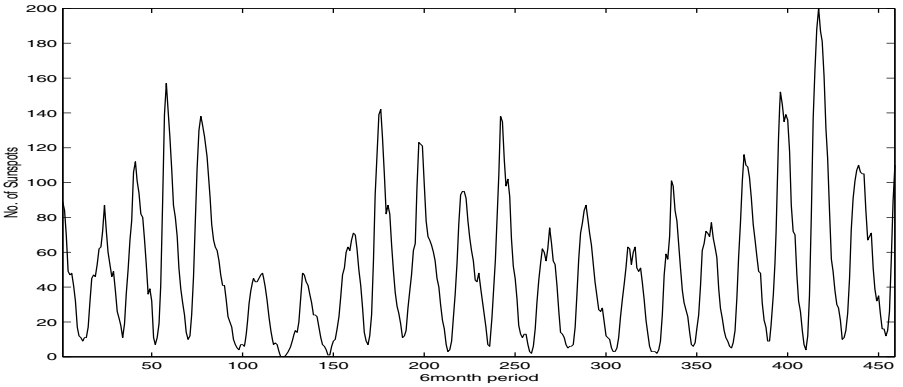


Figure 4.31 Smoothed 12-month sunspot numbers sampled twice per year, $n = 459$.

Section 4.4

4.8 Figure 4.31 shows the biyearly smoothed (12-month moving average) number of sunspots from June 1749 to December 1978 with $n = 459$ points that were taken twice per year. With Example 4.9 as a guide, perform a periodogram analysis of the sunspot data (the data are in the file `sunspots.dat`) identifying the predominant periods and obtaining confidence intervals for the identified periods. Interpret your findings.

4.9 The levels of salt concentration known to have occurred over rows, corresponding to the average temperature levels for the soil science data considered in Figures 1.15 and 1.16, are shown in Figure 4.32. The data are in the file `salt.dat`, which consists of one column of 128 observations; the first 64 observations correspond to the temperature series. Identify the dominant frequencies by performing separate spectral analyses on the two series. Include confidence intervals for the dominant frequencies and interpret your findings.

4.10 Let the observed series x_t be composed of a periodic signal and noise so it can be written as

$$x_t = \beta_1 \cos(2\pi\omega_k t) + \beta_2 \sin(2\pi\omega_k t) + w_t,$$

where w_t is a white noise process with variance σ_w^2. The frequency ω_k is assumed to be known and of the form k/n in this problem. Suppose we consider estimating β_1, β_2 and σ_w^2 by least squares, or equivalently, by maximum likelihood if the w_t are assumed to be Gaussian.

(a) Prove, for a fixed ω_k, the minimum squared error is attained by

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = 2n^{-1/2} \begin{pmatrix} d_c(\omega_k) \\ d_s(\omega_k) \end{pmatrix},$$

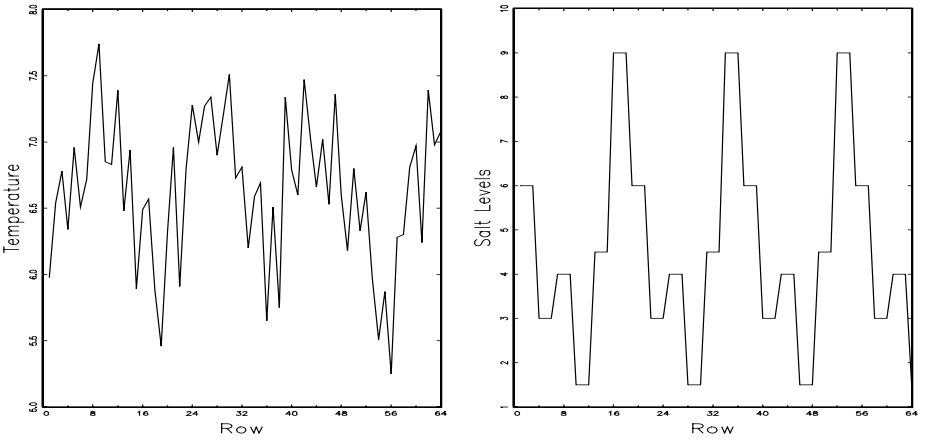


Figure 4.32 Temperature and salt profiles over 64 rows at 17-ft spacing.

where the cosine and sine transforms (4.24) and (4.25) appear on the right-hand side.

(b) Prove that the error sum of squares can be written as

$$SSE = \sum_{t=1}^{n} x_t^2 - 2I_x(\omega_k)$$

so that the value of ω_k that minimizes squared error is the same as the value that maximizes the periodogram $I_x(\omega_k)$ estimator (4.21).

(c) Under the Gaussian assumption and fixed ω_k, show that the F-test of no regression leads to an F-statistic that is a monotone function of $I_x(\omega_k)$.

4.11 Prove the convolution property of the DFT, namely,

$$\sum_{s=1}^{n} a_s x_{t-s} = \sum_{k=0}^{n-1} d_A(\omega_k) d_x(\omega_k) \exp\{2\pi\omega_k t\},$$

for $t = 1, 2, \ldots, n$, where $d_A(\omega_k)$ and $d_x(\omega_k)$ are the discrete Fourier transforms of a_t and x_t, respectively, and we assume that $x_t = x_{t+n}$ is periodic.

Section 4.5

4.12 Repeat Problem 4.8 using a nonparametric spectral estimation procedure. In addition to discussing your findings in detail, comment on your choice of a spectral estimate with regard to smoothing and tapering.

4.13 Repeat Problem 4.9 using a nonparametric spectral estimation procedure. In addition to discussing your findings in detail, comment on your choice of a spectral estimate with regard to smoothing and tapering.

4.14 The periodic behavior of a time series induced by echoes can also be observed in the spectrum of the series; this fact can be seen from the results stated in Problem 4.6(a). Using the notation of that problem, suppose we observe $x_t = s_t + As_{t-D} + n_t$, which implies the spectra satisfy $f_x(\omega) = [1 + A^2 + 2A\cos(2\pi\omega D)]f_s(\omega) + f_n(\omega)$. If the noise is negligible ($f_n(\omega) \approx 0$) then $\log f_x(\omega)$ is approximately the sum of a periodic component, $\log[1+A^2+2A\cos(2\pi\omega D)]$, and $\log f_s(\omega)$. Bogart et al. (1962) proposed treating the detrended log spectrum as a pseudo time series and calculating its spectrum, or *cepstrum*, which should show a peak at a *quefrency* corresponding to $1/D$. The cepstrum can be plotted as a function of quefrency, from which the delaty D can be estimated.

For the speech series presented in Example 1.3, estimate the pitch period using cepstral analysis as follows. The data are in the file `speech.dat`.

(a) Calculate and display the log-periodogram of the data. Is the periodogram periodic, as predicted?

(b) Perform a cepstral (spectral) analysis on the detrended logged periodogram, and use the results to estimate the delay D. How does your answer compare with the analysis of Example 1.24, which was based on the ACF?

4.15 Use Property P4.1 to verify (4.63). Then verify (4.66) and (4.67)

4.16 Consider two time series

$$x_t = w_t - w_{t-1},$$

$$y_t = \frac{1}{2}(w_t + w_{t-1}),$$

formed from the white noise series w_t with variance $\sigma_w^2 = 1$.

(a) Are x_t and y_t jointly stationary? Recall the cross-covariance function must also be a function only of the lag h and cannot depend on time.

(b) Compute the spectra $f_y(\omega)$ and $f_x(\omega)$, and comment on the difference between the two results.

(c) Suppose sample spectral estimators $\bar{f}_y(.10)$ are computed for the series using $L = 3$. Find a and b such that

$$P\left\{a \le \bar{f}_y(.10) \le b\right\} = .90.$$

This expression gives two points that will contain 90% of the sample spectral values. Put 5% of the area in each tail.

Section 4.6

4.17 Analyze the coherency between the temperature and salt data discussed in Problem 4.9. Discuss your findings.

4.18 Consider two processes

$$x_t = w_t \quad \text{and} \quad y_t = \phi x_{t-D} + v_t$$

where w_t and v_t are independent white noise processes with common variance σ^2, ϕ is a constant, and D is a fixed integer delay.

(a) Compute the coherency between x_t and y_t.

(b) Simulate $n = 1024$ normal observations from x_t and y_t for $\phi = .9$, $\sigma^2 = 1$, and $D = 0$. Then estimate and plot the coherency between the simulated series for the following values of L and comment: (i) $L = 1$, (ii) $L = 3$, (iii) $L = 41$, and (iv) $L = 101$.

Section 4.7

4.19 For the processes in Problem 4.18,

(a) Compute the phase between x_t and y_t.

(b) Simulate $n = 1024$ observations from x_t and y_t for $\phi = .9$, $\sigma^2 = 1$, and $D = 1$. Then estimate and plot the phase between the simulated series for the following values of L and comment: (i) $L = 1$, (ii) $L = 3$, (iii) $L = 41$, and (iv) $L = 101$.

4.20 Consider the bivariate time series records containing monthly U.S. production as measured monthly by the Federal Reserve Board Production Index and unemployment as given in Figure 3.22.

(a) Compute the spectrum and the log spectrum for each series, and identify statistically significant peaks. Explain what might be generating the peaks. Compute the coherence, and explain what is meant when a high coherence is observed at a particular frequency.

(b) What would be the effect of applying the filter

$$u_t = x_t - x_{t-1}$$

followed by

$$v_t = u_t - u_{t-12}$$

to the series given above? Plot the predicted frequency responses of the simple difference filter and of the seasonal difference of the first difference.

(c) Apply the filters successively to one of the two series and plot the output. Examine the output after taking a first difference and comment on whether stationarity is a reasonable assumption. Why or why not? Plot after taking the seasonal difference of the first difference. What can be noticed about the output that is consistent with what you have predicted from the frequency response? Verify by computing the spectrum of the output after filtering.

4.21 Determine the theoretical power spectrum of the series formed by combining the white noise series w_t to form

$$y_t = w_{t-2} + 4w_{t-1} + 6w_t + 4w_{t+1} + w_{t+2}.$$

Determine which frequencies are present by plotting the power spectrum.

4.22 Let $x_t = \cos(2\pi\omega t)$, and consider the output

$$y_t = \sum_{k=-\infty}^{\infty} a_k x_{t-k},$$

where $\sum_k |a_k| < \infty$. Show

$$y_t = |A(\omega)| \cos(2\pi\omega t + \phi(\omega)),$$

where $|A(\omega)|$ and $\phi(\omega)$ are the amplitude and phase of the filter, respectively. Interpret the result in terms of the relationship between the input series, x_t, and the output series, y_t.

4.23 Suppose x_t is a stationary series, and we apply two filtering operations in succession, say,

$$y_t = \sum_r a_r x_{t-r},$$

and then

$$z_t = \sum_s b_s y_{t-s}.$$

(a) Show the spectrum of the output is

$$f_z(\omega) = |A(\omega)|^2 |B(\omega)|^2 f_x(\omega),$$

where $A(\omega)$ and $B(\omega)$ are the Fourier transforms of the filter sequences a_t and b_t, respectively.

(b) What would be the effect of applying the filter

$$u_t = x_t - x_{t-1}$$

followed by

$$v_t = u_t - u_{t-12}$$

to a time series?

(c) Plot the predicted frequency responses of the simple difference filter and of the seasonal difference of the first difference. Filters like these are called seasonal adjustment filters in economics because they tend to attenuate frequencies at multiples of the monthly periods. The difference filter tends to attenuate low-frequency trends.

4.24 Suppose we are given a stationary zero-mean series x_t with spectrum $f_x(\omega)$ and then construct the derived series

$$y_t = ay_{t-1} + x_t, \quad t = \pm 1, \pm 2, \ldots .$$

(a) Show how the theoretical $f_y(\omega)$ is related to $f_x(\omega)$.

(b) Plot the function that multiplies $f_x(\omega)$ in part (a) for $a = .1$ and for $a = .8$. This filter is called a recursive filter.

Section 4.8

4.25 Often, the periodicities in the sunspot series are investigated by fitting an autoregressive spectrum of sufficiently high order. The main periodicity is often stated to be in the neighborhood of 11 years. Fit an autoregressive spectral estimator to the sunspot data using a model selection method of your choice. Compare the result with a conventional nonparametric spectral estimator found in Problem 4.8.

4.26 Fit an autoregressive spectral estimator to the Recruitment series and compare it to the results of Example 4.11.

4.27 Suppose a sample time series with $n = 256$ points is available from the first-order autoregressive model. Furthermore, suppose a sample spectrum computed with $L = 3$ yields the estimated value $\bar{f}_x(1/8) = 2.25$. Is this sample value consistent with $\sigma_w^2 = 1, \phi = .5$? Repeat using $L = 11$ if we just happen to obtain the same sample value.

4.28 Suppose we wish to test the noise alone hypothesis $H_0 : x_t = n_t$ against the signal-plus-noise hypothesis $H_1 : x_t = s_t + n_t$, where s_t and n_t are uncorrelated zero-mean stationary processes with spectra $f_s(\omega)$ and $f_n(\omega)$. Suppose that we want the test over a band of $L = 2m + 1$ frequencies of the form $\omega_{j:n} + k/n$, for $k = 0, \pm 1, \pm 2, \ldots, \pm m$ near some fixed frequency ω. Assume that both the signal and noise spectra are approximately constant over the interval.

(a) Prove the approximate likelihood-based test statistic for testing H_0 against H_1 is proportional to

$$T = \sum_k |d_x(\omega_{j:n} + k/n)|^2 \left(\frac{1}{f_n(\omega)} - \frac{1}{f_s(\omega) + f_n(\omega)} \right).$$

(b) Find the approximate distributions of T under H_0 and H_1.

(c) Define the false alarm and signal detection probabilities as $P_F = P\{T > K|H_0\}$ and $P_d = P\{T > k|H_1\}$, respectively. Express these probabilities in terms of the signal-to-noise ratio $f_s(\omega)/f_n(\omega)$ and appropriate chi-squared integrals.

Section 4.9

4.29 Repeat the dynamic Fourier analysis of Example 4.20 on the remaining seven earthquakes and seven explosions in the data file `eq+exp.dat`. Do the conclusions about the difference between earthquakes and explosions stated in the example still seem valid?

4.30 Repeat the wavelet analyses of Examples 4.21 and 4.22 on all earthquake and explosion series in the data file `eq+exp.dat`. Do the conclusions about the difference between earthquakes and explosions stated in Examples 4.21 and 4.22 still seem valid?

4.31 Using Examples 4.20-4.22 as a guide, perform a dynamic Fourier analysis and wavelet analyses (dwt and waveshrink analysis) on the event of unknown origin that took place near the Russian nuclear test facility in Novaya Zemlya. State your conclusion about the nature of the event at Novaya Zemlya.

Section 4.10

4.32 Consider the problem of approximating the filter output

$$y_t = \sum_{k=-\infty}^{\infty} a_k x_{t-k}, \quad \sum_{-\infty}^{\infty} |a_k| < \infty,$$

by

$$y_t^M = \sum_{|k|<M/2} a_k^M x_{t-k}$$

for $t = M/2-1, M/2, \ldots, n-M/2$, where x_t is available for $t = 1, \ldots, n$ and

$$a_t^M = M^{-1} \sum_{k=0}^{M-1} A(\omega_k) \exp\{2\pi i \omega_k t\}$$

with $\omega_k = k/M$. Prove

$$E\{(y_t - y_t^M)^2\} \le 4\gamma_x(0) \Big(\sum_{|k|\ge M/2} |a_k| \Big)^2.$$

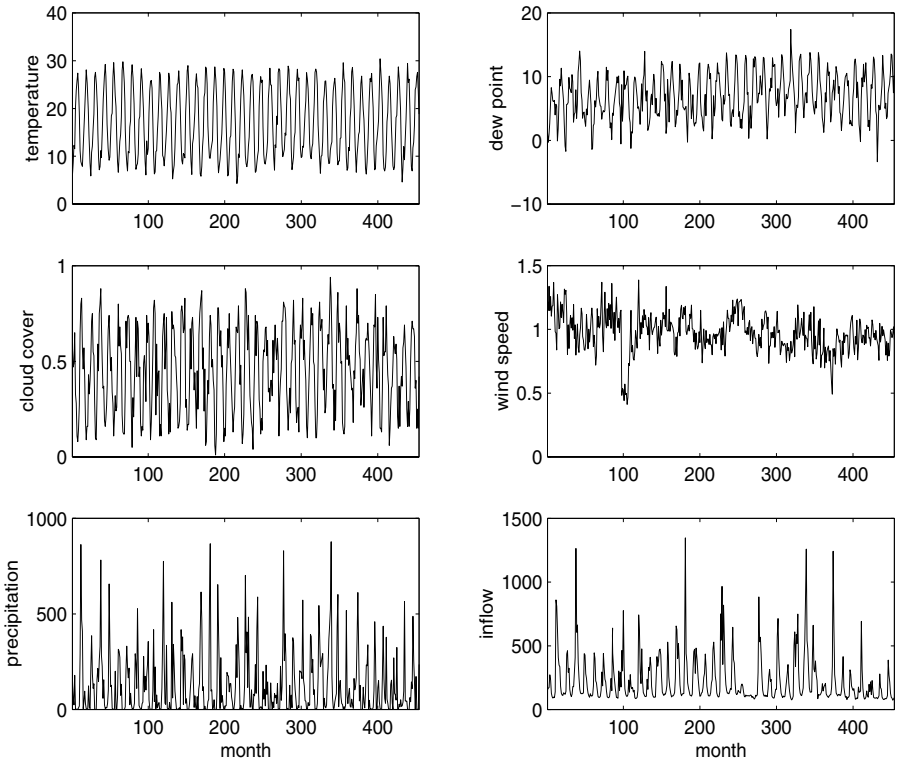


Figure 4.33 Monthly values of weather and inflow at Shasta Lake

4.33 Prove the squared coherence $\rho^2_{y\cdot x}(\omega) = 1$ for all ω when

$$y_t = \sum_{r=-\infty}^{\infty} a_r x_{t-r},$$

that is, when x_t and y_t can be related exactly by a linear filter.

4.34 Figure 4.33 contains 454 months of measured values for the climatic variables air temperature, dew point, cloud cover, wind speed, precipitation, and inflow at Shasta Lake in California. We would like to look at possible relations among the weather factors and between the weather factors and the inflow to Shasta Lake.

(a) Argue the strongest determinant of the inflow series is precipitation using the coherence functions. Use transformed inflow $I_t = \log i_t$, where i_t is inflow, and transformed precipitation $P_t = \sqrt{p}_t$, where

p_t is precipitation. It should be mentioned here that Chapter 6 discusses methods for determining whether inflow might depend jointly on several input series.

(b) Using the estimated impulse response function, argue for the model

$$I_t = \alpha_0 + \frac{\alpha_1}{1 - \phi B} P_t,$$

where the notation is as discussed in Chapter 2. What would be a reasonable value for ϕ? Assume the means are taken out of the series before the analysis begins.

Section 4.11

4.35 Consider the *signal plus noise* model

$$y_t = \sum_{r=-\infty}^{\infty} \beta_r x_{t-r} + v_t,$$

where the signal and noise series, x_t and v_t are both stationary with spectra $f_x(\omega)$ and $f_v(\omega)$, respectively. Assuming that x_t and v_t are independent of each other for all t, verify (4.142) and (4.143).

4.36 Consider the model

$$y_t = x_t + v_t,$$

where

$$x_t = \phi_1 x_{t-1} + w_t,$$

such that v_t is Gaussian white noise and independent of x_t with $\text{var}(v_t) = \sigma_v^2$, and w_t is Gaussian white noise and independent of v_t, with $\text{var}(w_t) = \sigma_w^2$, and $|\phi_1| < 1$ and $Ex_0 = 0$. Prove that the spectrum of the observed series y_t is

$$f_y(\omega) = \frac{\sigma^2 |1 - \theta_1 e^{-2\pi i \omega}|^2}{|1 - \phi_1 e^{-2\pi i \omega}|^2},$$

where

$$\theta_1 = \frac{c \pm \sqrt{c^2 - 4}}{2},$$

$$\sigma^2 = \frac{\sigma_v^2 \phi_1}{\theta_1},$$

and

$$c = \frac{\sigma_w^2 + \sigma_v^2(1 + \phi_1^2)}{\sigma_v^2 \phi_1}.$$

4.37 Consider the same model as in the preceding problem.

(a) Prove the optimal smoothed estimator of the form

$$\widehat{x}_t = \sum_{s=-\infty}^{\infty} a_s y_{t-s}$$

has

$$a_s = \frac{\sigma_w^2}{\sigma^2} \frac{\theta_1^{|s|}}{1-\theta_1^2}.$$

(b) Show the mean square error is given by

$$E\{(x_t - \widehat{x}_t)^2\} = \frac{\sigma_v^2 \sigma_w^2}{\sigma^2(1-\theta_1^2)}.$$

(c) Compare mean square error of the estimator in part (b) with that of the optimal finite estimator of the form

$$\widehat{x}_t = a_1 y_{t-1} + a_2 y_{t-2}$$

when $\sigma_v^2 = .053, \sigma_w^2 = .172$, and $\phi_1 = .9$.

Section 4.12

4.38 Consider the two-dimensional linear filter given as the output (4.154).

(a) Express the two-dimensional autocovariance function of the output, say, $\gamma_y(h_1, h_2)$, in terms of an infinite sum involving the autocovariance function of $x_{\boldsymbol{s}}$ and the filter coefficients a_{s_1,s_2}.

(b) Use the expression derived in (a), combined with (4.153) and (4.156) to derive the spectrum of the filtered output (4.155).

The following problems require the supplemental material given in Appendix C

4.39 Let w_t be a Gaussian white noise series with variance σ_w^2. Prove that the results of Theorem C.4 hold without error for the DFT of w_t.

4.40 Show that condition (4.41) implies (C.19) under the assumption that $w_t \sim wn(0, \sigma_w^2)$.

4.41 Prove Lemma C.4.

4.42 Finish the proof of Theorem C.5.

4.43 For the zero-mean complex random vector $\boldsymbol{z} = \boldsymbol{x}_c - i\boldsymbol{x}_s$, with $\mathrm{cov}(\boldsymbol{z}) = \Sigma = C - iQ$, with $\Sigma = \Sigma^*$, define

$$w = 2\mathrm{Re}(\boldsymbol{a}^*\boldsymbol{z}),$$

where $\boldsymbol{a} = \boldsymbol{a}_c - i\boldsymbol{a}_s$ is an arbitrary non-zero complex vector. Prove

$$\mathrm{cov}(w) = 2\boldsymbol{a}^*\Sigma\boldsymbol{a}.$$

Recall * denotes the complex conjugate transpose.

Chapter 5

Additional Time Domain Topics

5.1 Introduction

In this chapter, we present material that may be considered special or advanced topics in the time domain. Chapter 6 is devoted to one of the most useful and interesting time domain topics, state-space models. So, we do not cover state-space models or related topics—of which there are many—in this chapter. This chapter consists of sections of independent topics that may be read in any order. Most of the sections depend on a basic knowledge of ARMA models, forecasting and estimation, which is the material that is covered in Chapter 3, §3.1-§3.8. A few sections, for example the first section on long memory models, require some knowledge of spectral analysis and related topics covered in Chapter 4. In addition to long memory, we discuss GARCH models, threshold models, regression with autocorrelated errors, lagged regression or transfer functions, and selected topics in multivariate ARMAX models.

5.2 Long Memory ARMA and Fractional Differencing

The conventional ARMA(p, q) process is often referred to as a short memory process because the coefficients in the representation

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j},$$

obtained by solving

$$\phi(z)\psi(z) = \theta(z),$$

are dominated by exponential decay. As pointed out in §3.3, this result implies the ACF of the short memory process $\rho(h) \to 0$ exponentially fast as $h \to \infty$. When the sample ACF of a time series decays slowly, the advice given in Chapter 3 has been to difference the series until it seems stationary. Following this advice with the glacial varve series first presented in Example 3.31 leads to the first difference of the logarithms of the data being represented as a first-order moving average. In Example 3.37, further analysis of the residuals leads to fitting an ARIMA$(1, 1, 1)$ model,

$$\nabla x_t = \phi \nabla x_{t-1} + w_t + \theta w_{t-1},$$

where we understand x_t is the log-transformed varve series. In particular, the estimates of the parameters (and the standard errors) were $\widehat{\phi} = .23(.05)$, $\widehat{\theta} = -.89(.03)$, and $\widehat{\sigma}_w^2 = .23$. The use of the first difference $\nabla x_t = (1 - B)x_t$ can be too severe a modification in the sense that the nonstationary model might represent an overdifferencing of the original process.

Long memory (or persistent) time series were considered in Hosking (1981) and Granger and Joyeux (1980) as intermediate compromises between the short memory ARMA type models and the fully integrated nonstationary processes in the Box–Jenkins class. The easiest way to generate a long memory series is to think of using the difference operator $(1 - B)^d$ for fractional values of d, say, $0 < d < .5$, so a basic long memory series gets generated as

$$(1 - B)^d x_t = w_t, \tag{5.1}$$

where w_t still denotes white noise with variance σ_w^2. Now, d becomes a parameter to be estimated along with σ_w^2. Differencing the original process, as in the Box–Jenkins approach, may be thought of as simply assigning a value of $d = 1$. This idea has been extended to the class of fractionally integrated ARMA, or ARFIMA models, where we allow $-.5 < d < .5$; when d is negative, the term antipersistent is used. Long memory processes occur in hydrology (see Hurst, 1951, and McLeod and Hipel, 1978) and in environmental series, such as the varve data we have previously analyzed, to mention a few examples. Long memory time series data tend to exhibit sample autocorrelations that are not necessarily large (as in the case of $d = 1$), but persist for a long time. Figure 5.1 shows the sample ACF, to lag 100, of the log-transformed varve series, which exhibits classic long memory behavior.

The fractionally differenced series (5.1), for $|d| < .5$, is often called *fractional noise*. To investigate its properties, we can use the binomial expansion ($d > -.5$) to write

$$w_t = (1 - B)^d x_t = \sum_{j=0}^{\infty} \pi_j B^j x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} \tag{5.2}$$

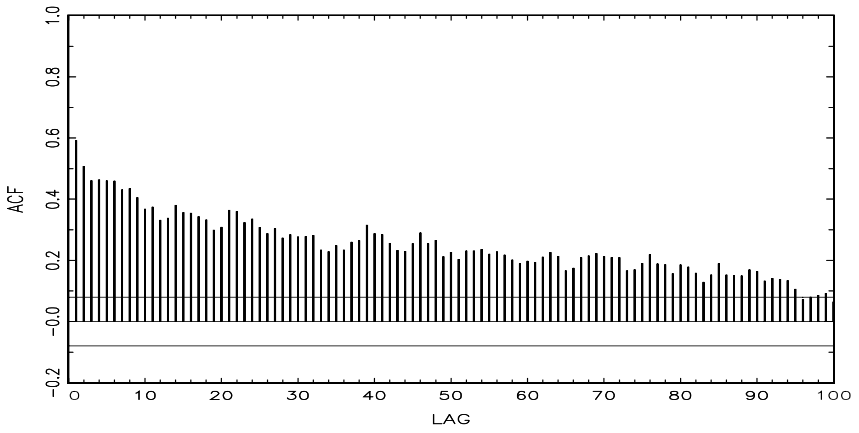


Figure 5.1 Sample ACF of the log transformed varve series.

where

$$\pi_j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} \tag{5.3}$$

with $\Gamma(x+1) = x\Gamma(x)$ being the gamma function. Similarly ($d < .5$), we can write

$$x_t = (1-B)^{-d} w_t = \sum_{j=0}^{\infty} \psi_j B^j w_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} \tag{5.4}$$

where

$$\psi_j = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)}. \tag{5.5}$$

The processes (5.2) and (5.4) are well-defined stationary processes (see Brockwell and Davis, 1991, for details). In the case of fractional differencing, however, the coefficients satisfy $\sum \pi_j^2 < \infty$ and $\sum \psi_j^2 < \infty$ as opposed to the absolute summability of the coefficients in ARMA processes.

Using the representation (5.4)–(5.5), the ACF of x_t is seen to be

$$\rho(h) = \frac{\Gamma(h+d)\Gamma(1-d)}{\Gamma(h-d+1)\Gamma(d)} \sim h^{2d-1} \tag{5.6}$$

for large h. From this we see that for $0 < d < .5$

$$\sum_{h=-\infty}^{\infty} |\rho(h)| = \infty$$

and hence the term *long memory*.

In order to examine a series such as the varve series for a possible long memory pattern, it is convenient to look at ways of estimating d. Using (5.3) it is easy to derive the recursions

$$\pi_{j+1}(d) = \frac{(j-d)\pi_j(d)}{(j+1)}, \tag{5.7}$$

for $j = 0, 1, \ldots$, with $\pi_0(d) = 1$. Maximizing the joint likelihood of the errors under normality, say, $w_t(d)$, will involve minimizing the sum of squared errors

$$Q(d) = \sum w_t^2(d).$$

The usual Gauss–Newton method, described in §3.6, leads to the expansion

$$w_t(d) = w_t(d_0) + w_t'(d_0)(d - d_0),$$

where

$$w_t'(d_0) = \left.\frac{\partial w_t}{\partial d}\right|_{d=d_0}$$

and d_0 is an initial estimate (guess) at to the value of d. Setting up the usual regression leads to

$$d = d_0 - \frac{\sum_t w_t'(d_0) w_t(d_0)}{\sum_t w_t'(d_0)^2}. \tag{5.8}$$

The derivatives are computed recursively by differentiating (5.7) successively with respect to d: $\pi_{j+1}'(d) = [(j-d)\pi_j'(d) - \pi_j(d)]/(j+1)$, where $\pi_0'(d) = 0$. The errors are computed from an approximation to (5.2), namely,

$$w_t(d) = \sum_{j=0}^{t} \pi_j(d) x_{t-j}. \tag{5.9}$$

It is advisable to omit a number of initial terms from the computation and start the sum, (5.8), at some fairly large value of t to have a reasonable approximation.

Example 5.1 Long Memory Fitting of the Glacial Varve Series

We consider analyzing the glacial varve series discussed in Examples 2.5 and 3.31. Figure 2.6 shows the original and log-transformed series (which we denote by x_t). In Example 3.37, we noted that x_t could be modeled as an ARMA$(1, 1, 1)$ process. We fit the fractionally differenced model, (5.1), to the mean-adjusted series, $x_t - \bar{x}$. Applying the Gauss–Newton iterative procedure previously described, starting with $d = .1$ and omitting the first 30 points from the computation, leads to a final value of $d = .384$, which implies the set of coefficients $\pi_j(.384)$, as given in Figure 5.2 with $\pi_0(.384) = 1$. We can compare roughly the performance

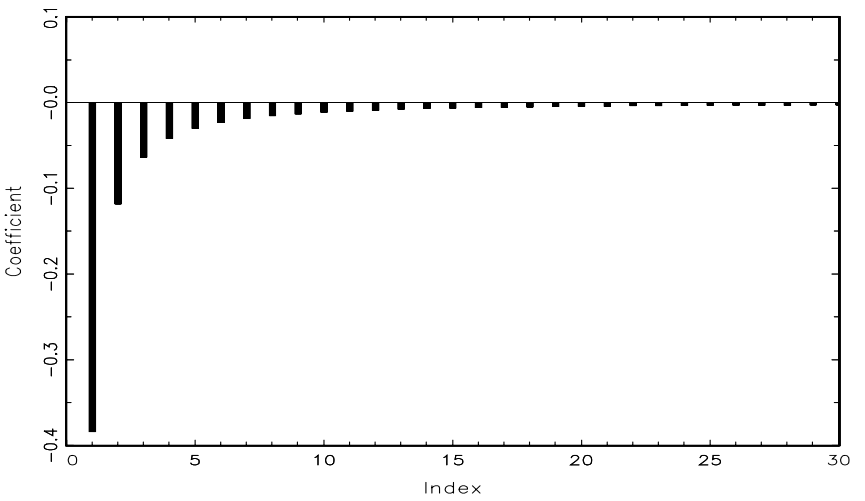


Figure 5.2 Coefficients $\pi_j(.384), j = 1, 2, \ldots, 30$ in the representation (5.7).

of the fractional difference operator with the ARIMA model by examining the autocorrelation functions of the two residual series as shown in Figure 5.3. The ACFs of the two residual series are roughly comparable with the white noise model.

To perform this analysis in R, first download and install the `fracdiff` package from CRAN. Then, load the package and issue the following commands (assuming the data are in `varve`).

```
> lvarve = log(varve)-mean(log(varve))
> varve.fd = fracdiff(lvarve, nar=0, nma=0, M=30)
> varve.fd$d
   [1] 0.3841688
> varve.fd$stderror.dpq
   [1] 4.589514e-06
```

The R package uses a truncated maximum likelihood procedure that was discussed in Haslett and Raftery (1989), which is a little more elaborate than simply zeroing out initial values. The default truncation value in R is $M = 100$. In the default case, the estimate is $\widehat{d} = .37$ with approximately the same standard error. The standard error is obtained from the Hessian as described in Example 3.28. At this time the R package `fracdiff` does not supply the residuals for diagnostics or an estimate of σ_w^2, hence some additional programming would be necessary for a full analysis.

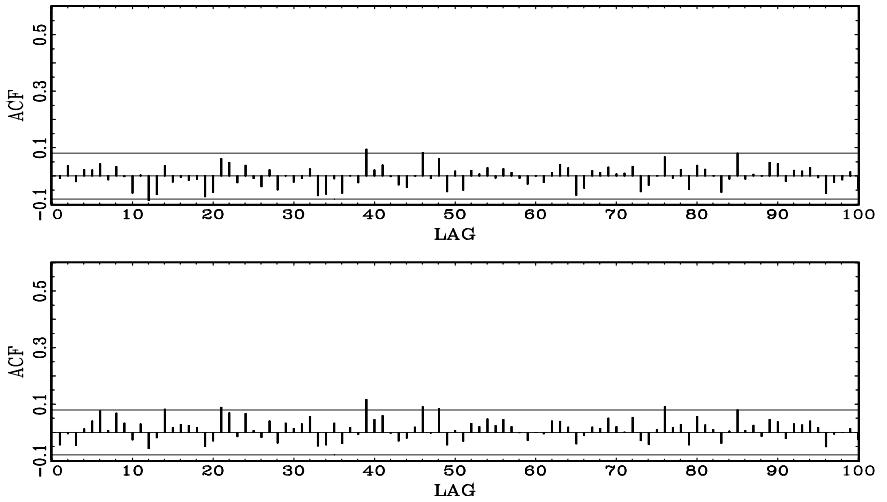


Figure 5.3 ACF of residuals from the ARIMA$(1,1,1)$ fit to the varve series (top) and of the residuals from the long memory model fit, $(1-B)^d x_t = w_t$, with $d = .384$ (bottom).

Forecasting long memory processes is similar to forecasting ARIMA models. That is, (5.2) and (5.7) can be used to obtain the truncated forecasts

$$\widetilde{x}^n_{n+m} = \sum_{j=1}^{n} \pi_j(\widehat{d})\, \widetilde{x}^n_{n+m-j}, \tag{5.10}$$

for $m = 1, 2, \ldots$. Error bounds can be approximated by using

$$P^n_{n+m} = \widehat{\sigma}^2_w \left(\sum_{j=0}^{m-1} \psi^2_j(\widehat{d}) \right) \tag{5.11}$$

where, as in (5.7),

$$\psi_j(\widehat{d}) = \frac{(j+\widehat{d})\psi_j(\widehat{d})}{(j+1)}, \tag{5.12}$$

with $\psi_0(\widehat{d}) = 1$.

No obvious short memory ARMA-type component can be seen in the ACF of the residuals from the fractionally differenced varve series shown in Figure 5.3. It is natural, however, that cases will exist in which substantial short memory-type components will also be present in data that exhibits long memory. Hence, it is natural to define the general ARFIMA(p,d,q), $-.5 < d < .5$ process as

$$\phi(B)\nabla^d(x_t - \mu) = \theta(B)w_t, \tag{5.13}$$

where $\phi(B)$ and $\theta(B)$ are as given in Chapter 3. Writing the model in the form

$$\phi(B)\pi_d(B)(x_t - \mu) = \theta(B)w_t \tag{5.14}$$

makes it clear how we go about estimating the parameters for the more general model. Forecasting for the ARFIMA(p, d, q) series can be easily done, noting that we may equate coefficients in

$$\phi(z)\psi(z) = (1 - z)^{-d}\theta(z) \tag{5.15}$$

and

$$\theta(z)\pi(z) = (1 - z)^{d}\phi(z) \tag{5.16}$$

to obtain the representations

$$x_t = \mu + \sum_{j=0}^{\infty} \psi_j w_{t-j}$$

and

$$w_t = \sum_{j=0}^{\infty} \pi_j (x_{t-j} - \mu).$$

We then can proceed as discussed in (5.10) and (5.11).

A comprehensive treatment of long memory models is given in Beran (1994), and it should be noted that several other techniques for estimating the parameters, especially, the long memory parameter, can be developed in the frequency domain. In this case, we may think of the equations as generated by an infinite order autoregressive series with coefficients π_j given by (5.7) . Using the same approach as before, we obtain

$$f_x(\omega) = \frac{\sigma_w^2}{|\sum_{k=0}^{\infty} \pi_k e^{-2\pi i k\omega}|^2} \tag{5.17}$$

$$= \sigma_w^2 |1 - e^{-2\pi i\omega}|^{-2d} \tag{5.18}$$

$$= [4\sin^2(\pi\omega)]^{-d}\sigma_w^2 \tag{5.19}$$

as equivalent representations of the spectrum of a long memory process. The long memory spectrum approaches infinity as the frequency $\omega \to 0$.

The main reason for defining the Whittle approximation to the log likelihood is to propose its use for estimating the parameter d in the long memory case as an alternative to the time domain method previously mentioned. The time domain approach is useful because of its simplicity and easily computed standard errors. One may also use an exact likelihood approach by developing an innovations form of the likelihood as in Brockwell and Davis (1991).

For the approximate approach using the Whittle likelihood (4.116), we consider using the approach of Fox and Taqqu (1986) who showed that maximizing

the Whittle log likelihood leads to a consistent estimator with the usual asymptotic normal distribution that would be obtained by treating (4.116) as a conventional log likelihood (see also Dahlhaus, 1989; Robinson, 1995; Hurvich et al., 1998). Unfortunately, the periodogram ordinates are not asymptotically independent (Hurvich and Beltrao, 1993), although a quasi-likelihood in the form of the Whittle approximation works well and has good asymptotic properties.

To see how this would work for the purely long memory case, write the long memory spectrum as

$$f_x(\omega_k;\ d, \sigma_w^2) = \sigma_w^2 g_k^{-d}, \tag{5.20}$$

where

$$g_k = 4\sin^2(\pi\omega_k). \tag{5.21}$$

Then, differentiating the log likelihood, say,

$$\ln L(\boldsymbol{x}; d, \sigma_w^2) \approx -m \ln \sigma_w^2 + d\sum_{k=1}^{m} \ln g_k - \frac{1}{\sigma_w^2}\sum_{k=1}^{m} g_k^d\, I(\omega_k) \tag{5.22}$$

at $m = n/2 - 1$ frequencies and solving for σ_w^2 yields

$$\sigma_w^2(d) = \frac{1}{m}\sum_{k=1}^{m} g_k^d\, I(\omega_k) \tag{5.23}$$

as the approximate maximum likelihood estimator for the variance parameter. To estimate d, we use a grid scan of the concentrated log likelihood

$$\ln L(\boldsymbol{x}; d) \approx -m \ln \sigma_w^2(d) - d\sum_{k=1}^{m} \ln g_k - m \tag{5.24}$$

over the interval $(-.5, .5)$, followed by a Newton–Raphson procedure to convergence.

Example 5.2 Long Memory Spectra for the Varve Series

We have previously examined the fit of the long memory model for the glacial varve data that is thought to be a reasonable surrogate for temperature. Fitting the long memory model using the Whittle approximation above gives $\widehat{d} = .394$, with an estimated standard error of .022. The earlier time domain method gave $\widehat{d} = .384$, with a standard error of 4.6×10^{-6}, so the results of the two methods are different. The error variance estimated was $\widehat{\sigma}_w^2 = .2320$. One might also consider fitting an autoregressive model to this data using a procedure similar to that used in Example 4.19. Following this approach gave an autoregressive model with $p = 8$ and $\widehat{\boldsymbol{\phi}} = (.34, .11, .03, .09, .09, .08, .02, .09)'$, with $\widehat{\sigma}_w^2 = .2303$

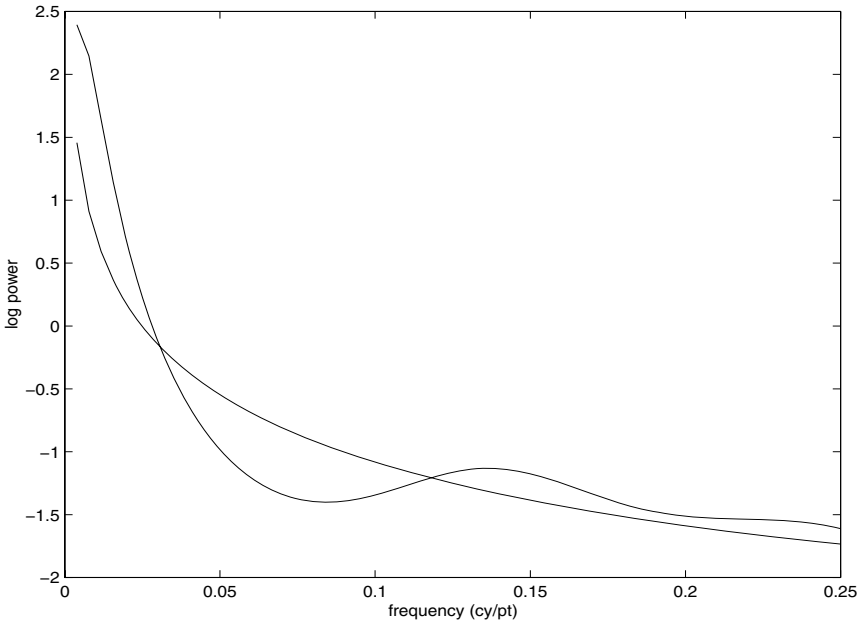


Figure 5.4 Long Memory ($d = .394$) and autoregressive AR(8) spectral estimators for the paleoclimatic glacial varve series.

as the error variance. The two log spectra are plotted in Figure 5.4 for $\omega > 0$, and we note that long memory spectrum is lower for the first frequency estimated ($\omega_1 = 1/512$) but will eventually become infinite, whereas the AR(8) spectrum is higher at that point, but takes a finite value at $\omega = 0$.

It should be noted that there is a strong likelihood that the spectrum will not be purely long memory, as it seemed to be in the example given above. A common situation has the long memory component multiplied by a short memory component, leading to an alternate version of (5.20) of the form

$$f_x(\omega_k;\ d, \theta) = g_k^{-d} f_0(\omega_k; \boldsymbol{\theta}), \tag{5.25}$$

where $f_0(\omega_k; \theta)$ might be the spectrum of an autoregressive moving average process with vector parameter $\boldsymbol{\theta}$, or it might be unspecified. If the spectrum has a parametric form, the Whittle likelihood can be used. However, there is a substantial amount of semiparametric literature that develops the estimators when the underlying spectrum $f_0(\omega; \boldsymbol{\theta})$ is unknown. A class of Gaussian semi-parametric estimators simply uses the same Whittle likelihood (5.24), evaluated over a sub-band of low frequencies, say $m' = \sqrt{n}$. There is some latitude in selecting a band that is relatively free from low frequency interference due to the short memory component in (5.25).

Geweke and Porter–Hudak (1983) developed an approximate method for estimating d based on a regression model, derived from (5.24). Note that we may write a simple equation for the logarithm of the spectrum as

$$\ln f_x(\omega_k; d) = \ln f_0(\omega_k; \boldsymbol{\theta}) - d \ln[4 \sin^2(\pi \omega_k)], \tag{5.26}$$

with the frequencies $\omega_k = k/n$ restricted to a range $k = 1, 2, \ldots, m'$ near the zero frequency with $m' = \sqrt{n}$ as the recommended value. Relationship (5.26) suggests using a simple linear regression model of the form,

$$\ln I(\omega_k) = \beta_0 - d \ln[4 \sin^2(\pi \omega_k)] + e_k \tag{5.27}$$

for the periodogram to estimate the parameters σ_w^2 and d. In this case, one performs least squares using $\ln I(\omega_k)$ as the dependent variable, and $\ln[4 \sin^2(\pi \omega_k)]$ as the independent variable for $k = 1, 2, \ldots, m$. The resulting slope estimate is then used as an estimate of $-d$. For a good discussion of various alternative methods for selecting m, see Hurvich and Deo (1999).

One of the above two procedures works well for estimating the long memory component but there will be cases (such as ARFIMA) where there will be a parameterized short memory component $f_0(\omega_k; \boldsymbol{\theta})$ that needs to be estimated. If the spectrum is highly parameterized, one might estimate using the Whittle log likelihood (5.21) and

$$f_x(\omega_k; \boldsymbol{\theta}) = g_k^{-d} f_0(\omega_k; \boldsymbol{\theta})$$

and jointly estimating the parameters d and $\boldsymbol{\theta}$ using the Newton–Raphson method. If we are interested in a nonparametric estimator, using the conventional smoothed spectral estimator for the periodogram, adjusted for the long memory component, say $g_k^d \, I(\omega_k)$ might be a possible approach.

5.3 GARCH Models

Recent problems in finance have motivated the study of the volatility, or variability, of a time series. Although ARMA models assume a constant variance, models such as the autoregressive conditionally heteroscedastic or ARCH model, first introduced by Engle (1982), were developed to model changes in volatility. These models were later extended to generalized ARCH, or GARCH models by Bollerslev (1986).

In §3.8, we discussed the return or growth rate of a series. For example, if x_t is the value of a stock at time t, then the return or relative gain, y_t, of the stock at time t is

$$y_t = \frac{x_t - x_{t-1}}{x_{t-1}}. \tag{5.28}$$

Definition (5.28) implies that $x_t = (1 + y_t) x_{t-1}$. Thus, based on the discussion in §3.8, if the return represents a small (in magnitude) percentage change then

$$\nabla[\ln(x_t)] \approx y_t. \tag{5.29}$$

Either value, $\nabla[\ln(x_t)]$ or $(x_t - x_{t-1})/x_{t-1}$, will be called the return, and will be denoted by y_t. It is the study of y_t that is the focus of ARCH, GARCH, and other volatility models. Recently there has been interest in stochastic volatility models and we will discuss these models in Chapter 6 because they are state-space models.

Typically, for financial series, the return y_t, does not have a constant variance, and highly volatile periods tend to be clustered together. In other words, there is a strong dependence of sudden bursts of variability in a return on the series own past. For example, Figure 1.4 shows the daily returns of the New York Stock Exchange (NYSE) from February 2, 1984 to December 31, 1991. In this case, as is typical, the return y_t is fairly stable, except for short-term bursts of high volatility.

The simplest ARCH model, the ARCH(1), models the return as

$$\begin{aligned} y_t &= \sigma_t \epsilon_t && (5.30) \\ \sigma_t^2 &= \alpha_0 + \alpha_1 y_{t-1}^2, && (5.31) \end{aligned}$$

where ϵ_t is standard Gaussian white noise; that is, $\epsilon_t \sim$ iid $\mathrm{N}(0,1)$. As with ARMA models, we must impose some constraints on the model parameters to obtain desirable properties. One obvious constraint is that α_1 must not be negative, or else σ_t^2 may be negative.

As we shall see, the ARCH(1) models return as a white noise process with nonconstant conditional variance, and that conditional variance depends on the previous return. First, notice that the conditional distribution of y_t given y_{t-1} is Gaussian:

$$y_t \mid y_{t-1} \sim \mathrm{N}(0, \alpha_0 + \alpha_1 y_{t-1}^2). \tag{5.32}$$

In addition, it is possible to write the ARCH(1) model as a non-Gaussian AR(1) model in the square of the returns y_t^2. To do this, rewrite (5.30)-(5.31) as

$$\begin{aligned} y_t^2 &= \sigma_t^2 \epsilon_t^2 \\ \alpha_0 + \alpha_1 y_{t-1}^2 &= \sigma_t^2, \end{aligned}$$

and subtract the two equations to obtain

$$y_t^2 - (\alpha_0 + \alpha_1 y_{t-1}^2) = \sigma_t^2 \epsilon_t^2 - \sigma_t^2.$$

Now, write this equation as

$$y_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + v_t, \tag{5.33}$$

where $v_t = \sigma_t^2(\epsilon_t^2 - 1)$. Because ϵ_t^2 is the square of a $\mathrm{N}(0,1)$ random variable, $\epsilon_t^2 - 1$ is a shifted (to have mean-zero), χ_1^2 random variable.

To explore the properties of ARCH, we define $Y_s = \{y_s, y_{s-1}, ...\}$. Then, using (5.32), we immediately see that y_t has a zero mean:

$$E(y_t) = EE(y_t \mid Y_{t-1}) = EE(y_t \mid y_{t-1}) = 0. \tag{5.34}$$

Because $E(y_t \mid Y_{t-1}) = 0$, the process y_t is said to be a *martingale difference*.

Because y_t is a martingale difference, it is also an uncorrelated sequence. For example, with $h > 0$,

$$\begin{aligned} \text{cov}(y_{t+h}, y_t) &= E(y_t y_{t+h}) = EE(y_t y_{t+h} \mid Y_{t+h-1}) \\ &= E\{y_t E(y_{t+h} \mid Y_{t+h-1})\} = 0. \end{aligned} \tag{5.35}$$

The last line of (5.35) follows because y_t belongs to the information set Y_{t+h-1} for $h > 0$, and, $E(y_{t+h} \mid Y_{t+h-1}) = 0$, as determined in (5.34).

An argument similar to (5.34) and (5.35) will establish the fact that the error process v_t in (5.33) is also a martingale difference and, consequently, an uncorrelated sequence. If the variance of v_t is finite and constant with respect to time, and $0 \leq \alpha_1 < 1$, then based on Property P3.1, (5.33) specifies a causal AR(1) process for y_t^2. Therefore, $E(y_t^2)$ and $\text{var}(y_t^2)$ must be constant with respect to time t. This, implies that

$$E(y_t^2) = \text{var}(y_t) = \frac{\alpha_0}{1 - \alpha_1} \tag{5.36}$$

and, after some manipulations,

$$E(y_t^4) = \frac{3\alpha_0^2}{(1 - \alpha_1)^2} \frac{1 - \alpha_1^2}{1 - 3\alpha_1^2}, \tag{5.37}$$

provided $3\alpha_1^2 < 1$. These results imply that the kurtosis, κ, of y_t is

$$\kappa = \frac{E(y_t^4)}{[E(y_t^2)]^2} = 3\frac{1 - \alpha_1^2}{1 - 3\alpha_1^2}, \tag{5.38}$$

which is always larger than 3 (unless $\alpha_1 = 0$), the kurtosis of the normal distribution. Thus, the marginal distribution of the returns, y_t, is leptokurtic, or has "fat tails."

In summary, an ARCH(1) process, y_t, as given by (5.30)-(5.31), or equivalently (5.32), is characterized by the following properties.

- If $0 \leq \alpha_1 < 1$, the process y_t itself is white noise and its unconditional distribution is symmetrically distributed around zero; this distribution is leptokurtic.
- If, in addition, $3\alpha_1^2 < 1$, the square of the process, y_t^2, follows a causal AR(1) model with ACF given by $\rho_{y^2}(h) = \alpha_1^h \geq 0$, for all $h > 0$. If $3\alpha_1 \geq 1$, but $\alpha_1 < 1$, then y_t^2 is strictly stationary with infinite variance.

Estimation of the parameters α_0 and α_1 of the ARCH(1) model is typically accomplished by conditional MLE. The conditional likelihood of the data $y_2, ..., y_n$ given y_1, is given by

$$L(\alpha_0, \alpha_1 \mid y_1) = \prod_{t=2}^{n} f_{\alpha_0, \alpha_1}(y_t \mid y_{t-1}), \tag{5.39}$$

where the density $f_{\alpha_0,\alpha_1}(y_t \mid y_{t-1})$ is the normal density specified in (5.32). Hence, the criterion function to be minimized, $l(\alpha_0, \alpha_1) \propto -\ln L(\alpha_0, \alpha_1 \mid y_1)$ is given by

$$l(\alpha_0, \alpha_1) = \frac{1}{2}\sum_{t=2}^{n} \ln(\alpha_0 + \alpha_1 y_{t-1}^2) + \frac{1}{2}\sum_{t=2}^{n}\left(\frac{y_t^2}{\alpha_0 + \alpha_1 y_{t-1}^2}\right). \tag{5.40}$$

Estimation is accomplished by numerical methods, as described in §3.6. In this case, analytic expressions for the gradient vector, $l^{(1)}(\alpha_0, \alpha_1)$, and Hessian matrix, $l^{(2)}(\alpha_0, \alpha_1)$, as described in Example 3.28, can be obtained by straightforward calculations. For example, the 2×1 gradient vector, $l^{(1)}(\alpha_0, \alpha_1)$, is given by

$$\begin{pmatrix} \partial l/\partial\alpha_0 \\ \partial l/\partial\alpha_1 \end{pmatrix} = \sum_{t=2}^{n} \begin{pmatrix} 1 \\ y_{t-1}^2 \end{pmatrix} \times \frac{\alpha_0 + \alpha_1 y_{t-1}^2 - y_t^2}{2\left(\alpha_0 + \alpha_1 y_{t-1}^2\right)^2}. \tag{5.41}$$

The calculation of the Hessian matrix is left as an exercise (Problem 5.7). The likelihood of the ARCH model tends to be flat unless n is very large. A discussion of this problem can be found in Shephard (1996).

It is also possible to combine a regression or an ARMA model for the mean with an ARCH model for the errors. For example, a regression with ARCH(1) errors model would have the observations x_t as linear function of p regressors, $\boldsymbol{z}_t = (z_{t1}, ..., z_{tp})'$, and ARCH(1) noise y_t, say,

$$x_t = \boldsymbol{\beta}'\boldsymbol{z}_t + y_t,$$

where y_t satisfies (5.30)-(5.31), but, in this case, is unobserved. Similarly, for example, an AR(1) model for data x_t exhibiting ARCH(1) errors would be

$$x_t = \phi_0 + \phi_1 x_{t-1} + y_t.$$

These types of models were explored by Weiss (1984).

Example 5.3 Analysis of U.S. GNP

In Example 3.35, we fit an MA(2) model and an AR(1) model to the U.S. GNP series and we concluded that the residuals from both fits appeared to behave like a white noise process. In Example 3.39 we concluded that the AR(1) is probably the better model in this case. It has been suggested that the U.S. GNP series has ARCH errors, and in this example, we will investigate this claim. If the GNP noise term is ARCH, the squares of the residuals from the fit should behave like a non-Gaussian AR(1) process, as pointed out in (5.33). Figure 5.5 shows the ACF and PACF of the squared residuals it appears that there may be some dependence, albeit small, left in the residuals.

We used the S-PLUS GARCH module to fit an AR(1)-ARCH(1) model to the U.S. GNP returns with the following results:

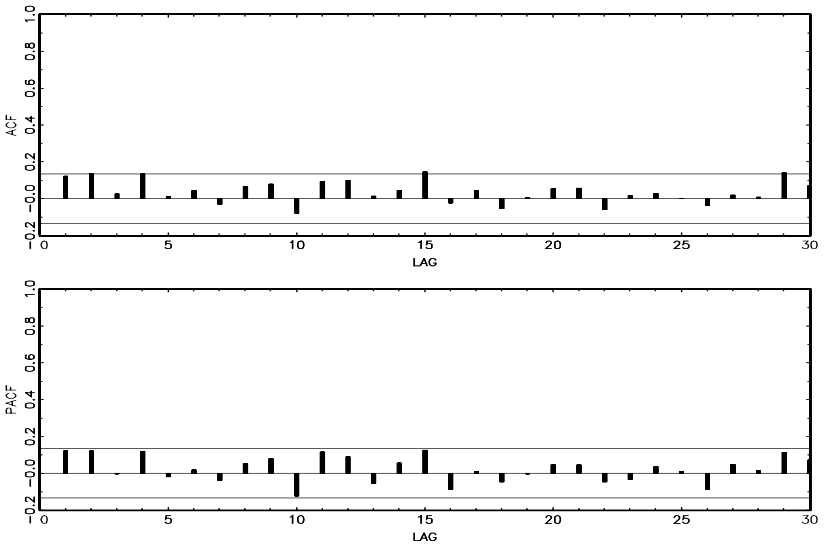


Figure 5.5 ACF and PACF of the squares of the residuals from the AR(1) fit on U.S. GNP.

```
> gnp96 <- matrix(scan("/mydata/gnp96.dat"),ncol=2,byrow=T)
> gnpr <- diff(log(gnp96[,2]))   # gnp returns
> gnpr.mod <- garch(gnpr~ar(1),~garch(1,0))   # model call
> summary(gnpr.mod)

Estimated Coefficients:
          Value  Std.Error t value   Pr(>|t|)
      C 0.00522 8.264e-004   6.326 6.990e-010  # AR cnst
  AR(1) 0.36721 7.888e-002   4.656 2.798e-006  # AR coef
      A 0.00007 6.978e-006  10.349 0.000e+000  # ARCH cnst
ARCH(1) 0.20242 7.031e-002   2.879 2.193e-003  # ARCH coef

Residual Tests:
 Jarque-Bera P-value   # tests normal skewness & kurtosis
       8.643 0.01328
Shapiro-Wilk P-value   # tests normal order statistics
      0.9827  0.4829
Q-Statistic P-value  Chi^2-d.f.
      13.88  0.3087         12
```

In this example, we obtain $\widehat{\phi}_0 = .005$ and $\widehat{\phi}_1 = .367$ for the AR(1) parameter estimates; in Example 3.35 the values were .005 and .347, respectively. The ARCH(1) parameter estimates are $\widehat{\alpha}_0 = 0$ for the constant and $\widehat{\alpha}_1 = .202$, which is highly significant with a p-value of

about .002. The Jarque–Bera statistic tests the residuals of the fit for normality based on the observed skewness and kurtosis, and it appears that the residuals have some non-normal skewness and kurtosis. The Shapiro–Wilk statistic tests the residuals of the fit for normality based on the empirical order statistics. In this case, the residuals appear to be normal. Finally, the Q-statistic is used on the squared residuals, and we conclude that the squared residuals appear to be an uncorrelated sequence.

To repeat the analysis in R without the simultaneous estimation, download the package `tseries` from CRAN and load it. Then, perform the AR estimation first and use those residuals for the ARCH fit as follows (assuming `gnpr` is available as in the S-PLUS example). We note that the results are similar to the simultaneous estimation results from S-PLUS.

```
> gnpr.ar = ar.mle(gnpr, order.max=1)  # recall phi1 =.347
> y = gnpr.ar$resid[2:length(gnpr)]    # first resid is NA
> arch.y = garch(y,order=c(0,1))
> summary.garch(arch.y)              # partial output below

   Coefficient(s):
     Estimate  Std. Error  t value Pr(>|t|)
a0 7.403e-05   7.275e-06   10.175  < 2e-16   # ARCH cnst
a1 1.939e-01   6.781e-02    2.859  0.00425   # ARCH coef

Jarque Bera Test:
   X-squared = 8.4801, df = 2, p-value = 0.01441
Box-Ljung test (squared residuals):
   X-squared = 3e-04, df = 1, p-value = 0.9865
```

The ARCH(1) model can be extended to the general ARCH(m) model in an obvious way. That is, (5.30) is retained,

$$y_t = \sigma_t \epsilon_t, \tag{5.30}$$

but (5.31) is extended to

$$\sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_m y_{t-m}^2. \tag{5.42}$$

Estimation for ARCH(m) also follows in an obvious way from the discussion of estimation for ARCH(1) models. That is, the conditional likelihood of the data $y_{m+1}, ..., y_n$ given $y_1, \ldots, y_m$, is given by

$$L(\boldsymbol{\alpha} \mid y_1, \ldots, y_m) = \prod_{t=m+1}^{n} f_{\boldsymbol{\alpha}}(y_t \mid y_{t-1}, \ldots, y_{t-m}), \tag{5.43}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_m)$ and the conditional densities $f_{\boldsymbol{\alpha}}(\cdot|\cdot)$ in (5.43) are normal densities; that is, for $t > m$,

$$y_t \mid y_{t-1}, \ldots, y_{t-m} \sim \mathrm{N}(0, \alpha_0 + \alpha_1 y_{t-1}^2 + \cdots + \alpha_m y_{t-m}^2).$$

Another extension of ARCH is the generalized ARCH or GARCH model developed by Bollerslev (1986). For example, a GARCH$(1,1)$ model retains (5.30),

$$y_t = \sigma_t \epsilon_t, \tag{5.30}$$

but extends (5.31) as follows:

$$\sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \beta_1 \sigma_{t-1}^2. \tag{5.44}$$

Under the condition that $\alpha_1 + \beta_1 < 1$, using similar manipulations as in (5.33), the GARCH$(1,1)$ model, (5.30) and (5.44), admits a non-Gaussian ARMA$(1,1)$ model for the squared process

$$y_t^2 = \alpha_0 + (\alpha_1 + \beta_1) y_{t-1}^2 + v_t - \beta_1 v_{t-1}, \tag{5.45}$$

where v_t is as defined in (5.33). Representation (5.45) follows by writing (5.30) as

$$\begin{aligned} y_t^2 - \sigma_t^2 &= \sigma_t^2(\epsilon_t^2 - 1) \\ \beta_1(y_{t-1}^2 - \sigma_{t-1}^2) &= \beta_1 \sigma_{t-1}^2(\epsilon_{t-1}^2 - 1), \end{aligned}$$

subtracting the second equation from the first, and using the fact that, from (5.44), $\sigma_t^2 - \beta_1 \sigma_{t-1}^2 = \alpha_0 + \alpha_1 y_{t-1}^2$, on the left-hand side of the result. The GARCH(m,r) model retains (5.30) and extends (5.44) to

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{m} \alpha_j y_{t-j}^2 + \sum_{j=1}^{r} \beta_j \sigma_{t-j}^2. \tag{5.46}$$

Conditional maximum likelihood estimation of the GARCH(m,r) model parameters is similar to the ARCH(m) case, wherein the conditional likelihood, (5.43), is the product of N$(0,\sigma_t^2)$ densities with σ_t^2 given by (5.46) and where the conditioning is on the first $\max(m,r)$ observations, with $\sigma_1^2 = \cdots = \sigma_r^2 = 0$. Once the parameter estimates are obtained, the model can be used to obtain one-step-ahead forecasts of the volatility, say $\widehat{\sigma}_{t+1}^2$, given by

$$\widehat{\sigma}_{t+1}^2 = \widehat{\alpha}_0 + \sum_{j=1}^{m} \widehat{\alpha}_j y_{t+1-j}^2 + \sum_{j=1}^{r} \widehat{\beta}_j \widehat{\sigma}_{t+1-j}^2. \tag{5.47}$$

We explore these concepts in the following example.

Example 5.4 GARCH Analysis of the NYSE Returns

As previously mentioned, the daily returns of the NYSE shown in Figure 1.4 exhibit classic GARCH features. We used the R `tseries` package to fit a GARCH$(1,1)$ model to the series with the following results:

```
> nyse = scan("/mydata/nyse.dat")
> nyse.g = garch(nyse, order=c(1,1))
> summary.garch(nyse.g)

Coefficient(s):
    Estimate  Std. Error  t value Pr(>|t|)
a0 6.552e-06   6.761e-07    9.691   <2e-16 # alpha0
a1 1.118e-01   4.056e-03   27.554   <2e-16 # alpha1
b1 8.086e-01   1.292e-02   62.566   <2e-16 # beta1

Diagnostic Tests:
Jarque Bera Test - data:  Residuals
X-squared = 3983.873, df = 2, p-value < 2.2e-16
Box-Ljung Test - data:  Squared.Residuals
X-squared = 1.5874, df = 1, p-value = 0.2077
```

To explore the GARCH predictions, we calculated and plotted the middle of the data along (which includes the October 19, 1987 crash) with the one-step-ahead predictions of the corresponding volatility, σ_t^2. The results are displayed as $\pm\widehat{\sigma}_t$ as a dashed line surrounding the data in Figure 5.6. These predictions can be obtained easily in R using the `tseries` package.

```
> u = predict.garch(nyse.g)
> plot(800:1000, nyse[800:1000], type="l", xlab="Time",
+   ylab="NYSE Returns")
> lines(u[,1], col="blue", lty="dashed")
> lines(u[,2], col="blue", lty="dashed")
```

Some key points can be gleaned from the examples of this section. First, it is apparent that the conditional distribution of the returns is rarely normal. S-PLUS allows for long tailed distributions to be fit to the data, whereas R does not. In particular, aside from the Gaussian distribution (the default), the S-PLUS Garch module allows for t, double exponential, and generalized double exponential[1] conditional distributions. Also, the predictions shown in Figure 5.6 leave something to be desired. It appears the model is better at telling you what the volatility was rather than what it is going to be; basically, increases or decreases in predicted volatility are a day late. In addition to these points, some other drawbacks of the GARCH model are: (i) the model assumes positive and negative returns have the same effect because volatility depends on squared returns; (ii) the model is restrictive because of the tight constraints on the model parameters (e.g., for an ARCH(1), $0 \le \alpha_1^2 < \frac{1}{3}$); (iii) the likelihood is flat unless n is very large; (iv) the model tends to overpredict volatility because it responds slowly to large isolated returns.

[1] $f(x) = p\alpha \exp(-\alpha x) I_{(0,\infty)}(x) + (1-p)\beta \exp(\beta x) I_{(-\infty,0)}(x)$; $0 < p < 1$.

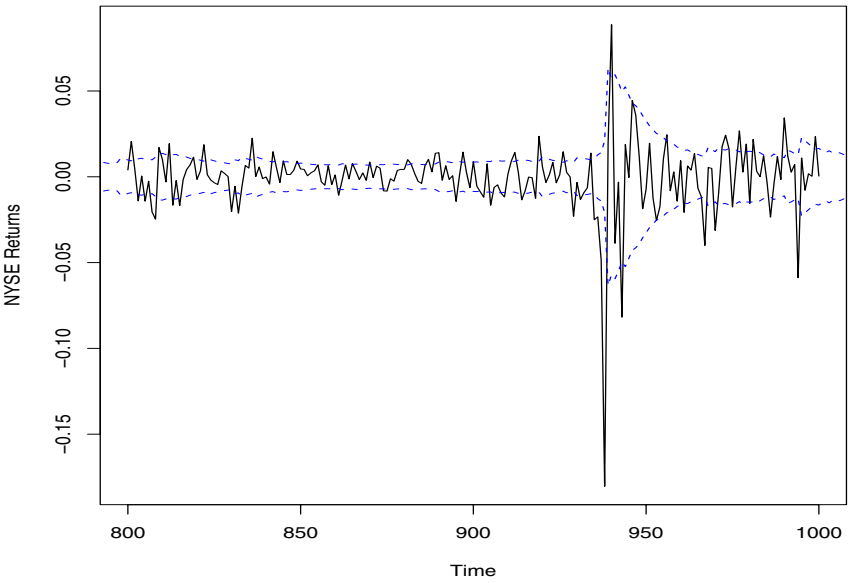


Figure 5.6 GARCH predictions of the NYSE volatility, $\pm\widehat{\sigma}_t$, displayed as dashed lines.

Various extensions to the original model have been proposed to overcome some of the shortcomings we have just mentioned. For example, we have already discussed the fact that the S-PLUS Garch module will fit some non-normal, albeit symmetric, distributions. For asymmetric return dynamics, one can use the EGARCH (exponential GARCH) model, which is a complex model that has different components for positive returns and for negative returns. In the case of persistence in volatility, the integrated GARCH (IGARCH) model may be used. Recall (5.45) where we showed the GARCH$(1,1)$ model can be written as

$$y_t^2 = \alpha_0 + (\alpha_1 + \beta_1)y_{t-1}^2 + v_t - \beta_1 v_{t-1}$$

and y_t^2 is stationary if $\alpha_1 + \beta_1 < 1$. The IGARCH model sets $\alpha_1 + \beta_1 = 1$, in which case the IGARCH$(1,1)$ model is

$$y_t = \sigma_t \epsilon_t \quad \text{and} \quad \sigma_t^2 = \alpha_0 + (1 - \beta_1)y_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

There are many different extensions to the basic ARCH model that were developed to handle the various situations noticed in practice. Interested readers might find the general discussions in Bollerslev et al. (1994) and Shephard (1996) worthwhile reading. Also, Gouriéroux (1997) gives a detailed presentation of ARCH and related models with financial applications and contains an extensive bibliography. Two excellent texts on financial time series analysis are Chan (2002) and Tsay (2001).

Finally, we briefly discuss stochastic volatility models; a detailed treatment of these models is given in Chapter 6. The volatility component, σ_t^2, in the GARCH model is conditionally nonstochastic. In the ARCH(1) model for example, any time the previous return is zero, i.e., $y_{t-1} = 0$, it must be the case that $\sigma_t^2 = \alpha_0$, and so on. This assumption seems a bit unrealistic. The stochastic volatility model adds a stochastic component to the volatility in the following way. In the GARCH model, a return, say y_t, is

$$y_t = \sigma_t \epsilon_t \quad \Rightarrow \quad \log y_t^2 = \log \sigma_t^2 + \log \epsilon_t^2. \tag{5.48}$$

In this way, we see that the observations $\log y_t^2$, are made up of two components, the unobserved volatility $\log \sigma_t^2$, which may be considered a latent variable, and unobserved noise $\log \epsilon_t^2$. While, for example, the GARCH$(1, 1)$ models volatility without error, $\sigma_{t+1}^2 = \alpha_0 + \alpha_1 r_t^2 + \beta_1 \sigma_t^2$, the basic stochastic volatility model assumes the latent variable is an autoregressive process,

$$\log \sigma_{t+1}^2 = \phi_0 + \phi_1 \log \sigma_t^2 + w_t \tag{5.49}$$

where $w_t \sim$ iid N$(0, \sigma_w^2)$. The introduction of the noise term w_t makes the latent volatility process stochastic. Together (5.48) and (5.49) comprise the stochastic volatility model. Given n observations, the goals are to estimate the parameters ϕ_0, ϕ_1 and σ_w^2, and then predict future observations $\log y_{n+m}^2$. Details are provided in §6.10.

5.4 Threshold Models

In §3.5 we discussed the fact that, for a stationary time series, best linear prediction forward in time is the same as best linear prediction backward in time. This result followed from the fact that the variance–covariance matrix of $\boldsymbol{x}_{1:n} = (x_1, x_2, ..., x_n)'$, say, $\Gamma = \{\gamma(i-j)\}_{i,j=1}^n$, is the same as the variance–covariance matrix of $\boldsymbol{x}_{n:1} = (x_n, x_{n-1}, ..., x_1)'$. In addition, if the process is Gaussian, the distributions of $\boldsymbol{x}_{1:n}$ and $\boldsymbol{x}_{n:1}$ are identical. In this case, a time plot of $\boldsymbol{x}_{1:n}$ (that is, the data plotted forward in time) should look similar to a time plot of $\boldsymbol{x}_{n:1}$ (that is, the data plotted backward in time).

There are, however, many series that do not fit into this category. For example, Figure 5.7 shows a plot of monthly pneumonia and influenza deaths per 10,000 in the U.S. for 11 years, 1968 to 1978. Typically, the number of deaths tends to increase slower than it decreases. Thus, if the data were plotted backward in time, the backward series would tend to increase faster than it decreases. Also, if monthly pneumonia and influenza deaths followed a linear Gaussian process, we would not expect to see such large bursts of positive and negative changes that occur periodically in this series. Moreover, although the number of deaths is typically largest during the winter months, the data are

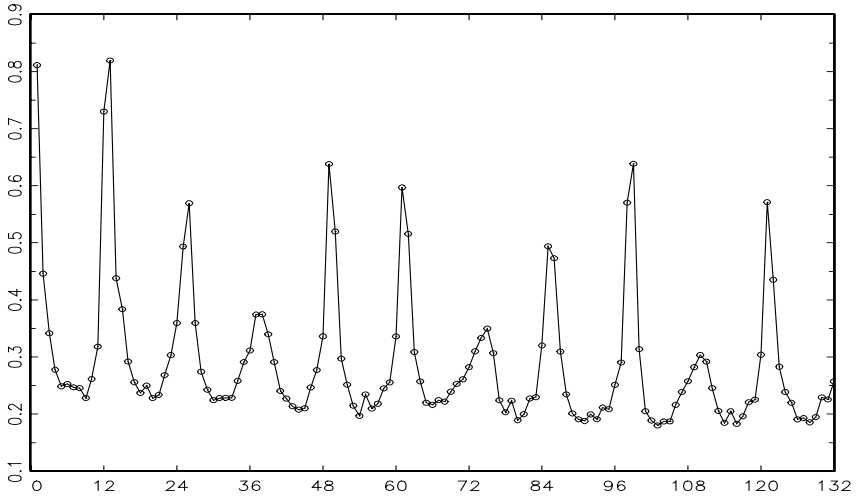


Figure 5.7 U.S. monthly pneumonia and influenza deaths per 10,000 over 11 years from 1968 to 1978.

not perfectly seasonal. That is, although the peak of the series often occurs in January, in other years, the peak occurs in December, February, or March.

If our goal is to predict flu epidemics, then it should be clear that a Gaussian linear model would not be appropriate. Many approaches to modeling nonlinear series exist that could be used (see Priestley, 1988); here, we focus on the class of threshold autoregressive models presented in Tong (1983, 1990). The basic idea of these models is that of fitting local linear AR(p) models, and their appeal is that we can use the intuition from fitting global linear AR(p) models. Suppose we know p, and given the vectors $\boldsymbol{x}_{t-1} = (x_{t-1}, ..., x_{t-p})'$, we can identify r mutually exclusive and exhaustive regions for $\boldsymbol{x}_{t-1}$, say, $R_1, ..., R_r$, where the dynamics of the system changes. The threshold model is then written as r AR(p) models,

$$x_t = \alpha^{(j)} + \phi_1^{(j)} x_{t-1} + \cdots + \phi_p^{(j)} x_{t-p} + w_t^{(j)}, \quad \boldsymbol{x}_{t-1} \in R_j, \tag{5.50}$$

for $j = 1, ..., r$. In (5.50), the $w_t^{(j)}$ are independent white noise series, each with variance σ_j^2, for $j = 1, ..., r$. Model estimation, identification, and diagnostics proceed as in the case in which $r = 1$.

Example 5.5 Threshold Modeling of the Influenza Series

As previously discussed, examination of Figure 5.7 leads us to believe that the monthly pneumonia and influenza deaths time series, say flu_t, is not linear. It is also evident from Figure 5.7 that there is a slight negative trend in the data. We have found that the most convenient way

to fit a threshold model to this data set, while removing the trend, is to work with the first difference of the data. The differenced data,

$$x_t = \text{flu}_t - \text{flu}_{t-1}$$

is exhibited in Figure 5.8 as the dark solid line with circles representing observations. The dashed line with squares in Figure 5.8 are the one-month-ahead predictions, and we will discuss this series later.

The nonlinearity of the data is more pronounced in the plot of the first differences, x_t. Clearly, the change in the numbers of deaths, x_t, slowly rises for some months and, then, sometime in the winter, has a possibility of jumping to a large number once x_t exceeds about .05. If the processes does make a large jump, then a subsequent significant decrease occurs in flu deaths. As an initial analysis, we fit the following threshold model

$$\begin{aligned} x_t &= \alpha^{(1)} + \sum_{j=1}^{p} \phi_j^{(1)} x_{t-j} + w_t^{(1)}, \quad x_{t-1} < .05 \\ x_t &= \alpha^{(2)} + \sum_{j=1}^{p} \phi_j^{(2)} x_{t-j} + w_t^{(2)}, \quad x_{t-1} \geq .05, \end{aligned} \tag{5.51}$$

with $p = 6$, assuming this would be larger than necessary.

Model (5.51) is easy to fit using two linear regression runs. That is, let $\delta_t^{(1)} = 1$ if $x_{t-1} < .05$, and zero otherwise, and let $\delta_t^{(2)} = 1$ if $x_{t-1} \geq .05$, and zero otherwise. Then, using the notation of §2.2, for $t = p+1, ..., n$, either equation in (5.51) can be written as

$$y_t = \boldsymbol{\beta}' \boldsymbol{z}_t + w_t$$

where, for $i = 1, 2$,

$$y_t = \delta_t^{(i)} x_t, \quad \boldsymbol{z}_t', = \delta_t^{(i)}(1, x_{t-1}, ..., x_{t-p}), \quad w_t = \delta_t^{(i)} w_t^{(i)},$$

and

$$\boldsymbol{\beta}' = (\alpha^{(i)}, \phi_1^{(i)}, \phi_2^{(i)}, ..., \phi_p^{(i)}).$$

Parameter estimates can then be obtained using the regression techniques of §2.2 twice, once for $i = 1$ and again for $i = 2$.

For each model, an order $p = 4$ model was finally selected. The final model was

$$\begin{aligned} \widehat{x}_t &= .51_{(.08)} x_{t-1} - .20_{(.06)} x_{t-2} + .12_{(.05)} x_{t-3} \\ &\quad -.11_{(.5)} x_{t-4} + \widehat{w}_t^{(1)}, \qquad \text{when } x_{t-1} < .05 \end{aligned}$$

$$\begin{aligned} \widehat{x}_t &= .40 - .75_{(.17)} x_{t-1} - 1.03_{(.21)} x_{t-2} - 2.05_{(1.05)} x_{t-3} \\ &\quad -6.71_{(1.25)} x_{t-4} + \widehat{w}_t^{(2)}, \qquad \text{when } x_{t-1} \geq .05, \end{aligned}$$

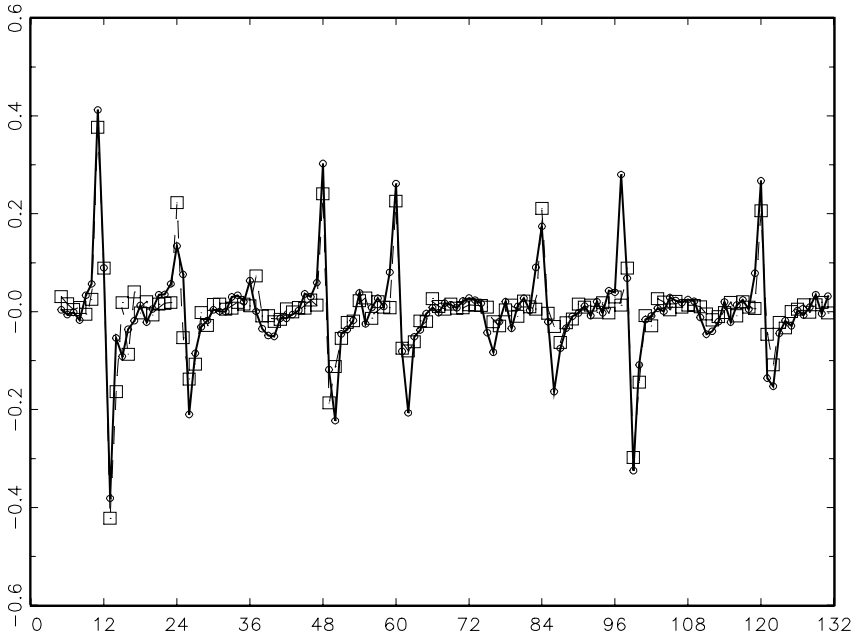


Figure 5.8 First differenced U.S. monthly pneumonia and influenza deaths per 1,000 (solid line - circles); one-month-ahead predictions (dashed line -squares).

where $\widehat{\sigma}_1 = .05$ and $\widehat{\sigma}_2 = .07$. The threshold of .05 was exceeded 17 times. Using the final model, one-month-ahead predictions can be made, and these are shown in Figure 5.8 as a dashed line with squares. The model does extremely well at predicting a flu epidemic; the peak at $t = 96$, however, was missed by this model. When we fit a model with a smaller threshold of .04, flu epidemics were somewhat underestimated, but the flu epidemic in the eighth year was predicted one month early. We chose the model with a threshold of .05 because the residual diagnostics showed no obvious departure from the model assumption (except for one outlier at $t = 96$); the model with a threshold of .04 still had some correlation left in the residuals and there were more than one outliers. Finally, prediction beyond one-month-ahead for this model is very complicated, but some approximate techniques exist (see Tong, 1983).

5.5 Regression with Autocorrelated Errors

In §2.2, we covered the classical regression model with uncorrelated errors w_t. In this section, we discuss the modifications that might be considered when the errors are correlated. That is, consider the regression model

$$y_t = \boldsymbol{\beta}' \boldsymbol{z}_t + x_t, \tag{5.52}$$

$t = 1, \ldots, n$, where x_t is a process with some covariance function $\gamma(s, t)$. Then, we have the matrix form

$$\boldsymbol{y} = Z\boldsymbol{\beta} + \boldsymbol{x}, \tag{5.53}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)'$ is a $n \times 1$ vector with $n \times n$ covariance matrix $\Gamma = \{\gamma(s, t)\}$. Note that $Z = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n]'$ is the $n \times q$ matrix of input variables, as before. If we know the covariance matrix Γ, it is possible to find a transformation matrix A, such that $A\Gamma A' = \sigma^2 I$, where I denotes the $n \times n$ identity matrix. Then, the underlying model can be transformed into

$$\begin{aligned} A\boldsymbol{y} &= AZ\boldsymbol{\beta} + A\boldsymbol{x} \\ &= U\boldsymbol{\beta} + \boldsymbol{w}, \end{aligned}$$

where $U = AZ$ and $\boldsymbol{w}$ is a white noise vector with covariance matrix $\sigma^2 I$ as in §2.2. Then, applying least squares or maximum likelihood to the vector $A\boldsymbol{y}$ gives

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_w &= (U'U)^{-1}U'A\boldsymbol{y} \\ &= (Z'A'AZ)^{-1}Z'A'A\boldsymbol{y} \\ &= (Z'\Gamma^{-1}Z)^{-1}Z'\Gamma^{-1}\boldsymbol{y} \end{aligned} \tag{5.54}$$

because

$$\sigma^2\Gamma^{-1} = A'A.$$

The difficulty in applying (5.54) is, we do not know the form of the matrix Γ.

It may be possible, however, in the time series case, to assume a stationary covariance structure for the error process x_t that corresponds to a linear process and try to find an ARMA representation for x_t. For example, if we have a pure AR(p) error, then

$$\phi(B)x_t = w_t,$$

and $\phi(B)$ is the linear transformation that, when applied to the error process, produces the white noise w_t. Regarding this transformation as the appropriate matrix A of the preceding paragraph produces the transformed regression equation

$$\phi(B)y_t = \boldsymbol{\beta}'\phi(B)\boldsymbol{z}_t + w_t,$$

and we are back to the same model as before. Defining $u_t = \phi(B)y_t$ and $\boldsymbol{v}_t = \phi(B)\boldsymbol{z}_t$ leads to the simple regression problem

$$u_t = \boldsymbol{\beta}'\boldsymbol{v}_t + w_t \tag{5.55}$$

considered before. The preceding discussion suggests an algorithm, due to Cochrane and Orcutt (1949), for fitting a regression model with autocorrelated errors.

(i) First, run an ordinary regression of y_t on z_t (acting as if the errors are uncorrelated). Retain the residuals.

(ii) Fit an ARMA model to the residuals $\widehat{x}_t = y_t - \widehat{\boldsymbol{\beta}}' \boldsymbol{z}_t$, say,

$$\widehat{\phi}(B)\widehat{x}_t = \widehat{\theta}(B)w_t \tag{5.56}$$

(iii) Then, apply the ARMA transformation to both sides (5.52), that is,

$$u_t = \frac{\widehat{\phi}(B)}{\widehat{\theta}(B)} y_t$$

and

$$\boldsymbol{v}_t = \frac{\widehat{\phi}(B)}{\widehat{\theta}(B)} \boldsymbol{z}_t,$$

to obtain the transformed regression model (5.55).

(iv) Run an ordinary least squares regression model assuming uncorrelated errors on the transformed regression model (5.55), obtaining

$$\widehat{\boldsymbol{\beta}}_w = (V'V)^{-1}V'\boldsymbol{u}, \tag{5.57}$$

where $V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n]'$ and $\boldsymbol{u} = (u_1, \ldots, u_n)'$ are the corresponding transformed components.

The above procedure can be repeated until convergence and will approach the maximum likelihood solution under normality of the errors (for details, see Sargan, 1964).

Example 5.6 Pollution, Temperature, Mortality with Correlated Errors

We consider further the best regression obtained in Example 2.2 of Chapter 2, relating adjusted temperature $T_t - T.$, $(T_t - T.)^2$ and particulate levels P_t to cardiovascular mortality M_t. Identifying the vectors

$$\boldsymbol{z}_t = (1, t, (T_t - T.), (T_t - T.)^2, P_t)'$$

leads to a model of the form (5.52). Taking the residuals from the least squares regression, as described in Step (i), the sample ACF and PACF, shown in Figure 5.9, suggest an AR(2) model for the residuals. Note, $\widehat{\sigma}^2 = 40.77$ and $R^2 = .59$ for this model.

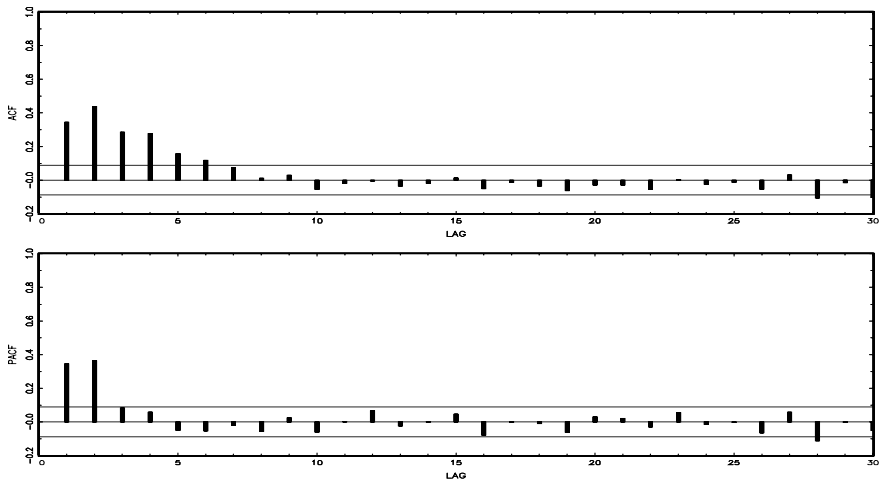


Figure 5.9 Sample ACF and PACF of the mortality residuals indicating an AR(2) process.

For the residuals, we obtain a second-order autoregressive model with operator

$$\phi(B) = 1 - .2207B - .3627B^2$$

which is applied to both sides of the defining equation (5.52) to produce the transformed equation (5.55), as in Step (ii) above. Running the regression, as in Step (iii), yields the model

$$\begin{aligned}\widehat{M_t} &= 83.54 - .028_{(.004)}t - .196_{(.039)}(T_t - 74.6) \\ &\quad + .017_{(.002)}(T_t - 74.6)^2 + .229_{(.023)}P_t\end{aligned}$$

as the model for transformed mortality, where the coefficients and estimated variances have changed slightly because of the transformation. The linear temperature component has decreased in magnitude from $-.473$ to $-.196$, whereas the other components stayed almost the same. The new residuals from the transformed model have sample ACF and PACF in Figure 5.10 that show no prominent peaks and can probably be taken as white noise.

5.6 Lagged Regression: Transfer Function Modeling

In §4.10, we considered lagged regression in a frequency domain approach based on coherency. In this section we focus on a time domain approach to the same

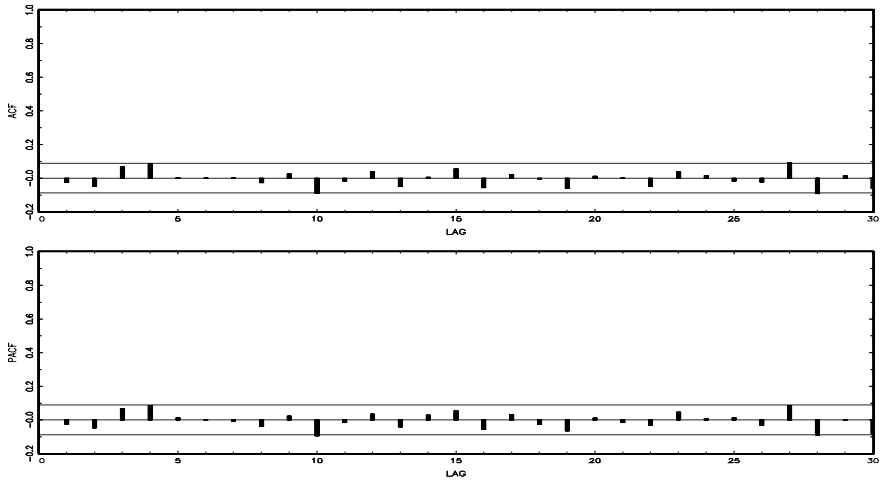


Figure 5.10 Sample ACF and PACF of the mortality residuals after fitting an AR(2) model.

problem. In the previous section, we looked at autocorrelated errors but, still regarded the input series $\boldsymbol{z}_t$ as being fixed unknown functions of time. This consideration made sense for the time argument t, but was less satisfactory for the other inputs, which are probably stochastic processes. For example, consider the SOI and Recruitment series that were presented in Example 1.5. The series are displayed in Figure 1.5. In this case, the interest is in predicting the output Recruitment series, say, y_t, from the input SOI, say x_t. We might consider the lagged regression model

$$y_t = \sum_{j=0}^{\infty} \alpha_j x_{t-j} + \eta_t = \alpha(B)x_t + \eta_t, \tag{5.58}$$

where $\sum_j |\alpha_j| < \infty$. We assume the input process x_t and noise process η_t in (5.58) are both stationary and mutually independent. The coefficients $\alpha_0, \alpha_1, \ldots$ describe the weights assigned to past values of x_t used in predicting y_t and we have used the notation

$$\alpha(B) = \sum_{j=0}^{\infty} \alpha_j B^j. \tag{5.59}$$

In the Box and Jenkins (1970) formulation, we assign ARIMA models, say, ARIMA(p, d, q) and ARIMA(p_η, d_η, q_η), to the series x_t and η_t, respectively. The components of (5.58) in backshift notation, for the case of simple ARMA(p, q) modeling of the input and noise, would have the representation

$$\phi(B)x_t = \theta(B)w_t \tag{5.60}$$

and

$$\phi_\eta(B)\eta_t = \theta_\eta(B)z_t, \tag{5.61}$$

where w_t and z_t are independent white noise processes with variances σ_w^2 and σ_z^2, respectively. Box and Jenkins (1970) proposed that systematic patterns often observed in the coefficients α_j, for $j = 1, 2, ...$, could often be expressed as a ratio of polynomials involving a small number of coefficients, along with a specified delay, d, so

$$\alpha(B) = \frac{\delta(B)B^d}{\omega(B)}, \tag{5.62}$$

where

$$\omega(B) = 1 - \omega_1 B - \omega_2 B^2 - \cdots - \omega_r B^r \tag{5.63}$$

and

$$\delta(B) = \delta_0 + \delta_1 B + \cdots + \delta_s B^s \tag{5.64}$$

are the indicated operators; in this section, we find it convenient to represent the inverse of an operator, say, $[\omega(B)]^{-1}$, as $1/\omega(B)$.

Determining a parsimonious model involving a simple form for $\alpha(B)$ and estimating all of the parameters in the above model are the main tasks in the transfer function methodology. Because of the large number of parameters, it is necessary to develop a sequential methodology. Suppose we focus first on finding the ARIMA model for the input x_t and apply this operator to both sides of (5.58), obtaining the new model

$$\begin{aligned} \tilde{y}_t &= \frac{\phi(B)}{\theta(B)} y_t \\ &= \alpha(B)w_t + \frac{\phi(B)}{\theta(B)}\eta_t \\ &= \alpha(B)w_t + \tilde{\eta}_t, \end{aligned}$$

where w_t and the transformed noise $\tilde{\eta}_t$ are independent.

The series w_t is a prewhitened version of the input series, and its cross-correlation with the transformed output series $\tilde{y}_t$ will be just

$$\begin{aligned} \gamma_{\tilde{y}w}(h) &= E[\tilde{y}_{t+h}w_t] \\ &= E[\sum_{j=0}^{\infty} \alpha_j w_{t+h-j} w_t] \\ &= \sigma_w^2 \alpha_h, \end{aligned} \tag{5.65}$$

because the autocovariance function of white noise will be zero except when $j = h$ in (5.65). Hence, computing the cross-correlation between the prewhitened input series and the transformed output series should yield a rough estimate of the behavior of $\alpha(B)$.

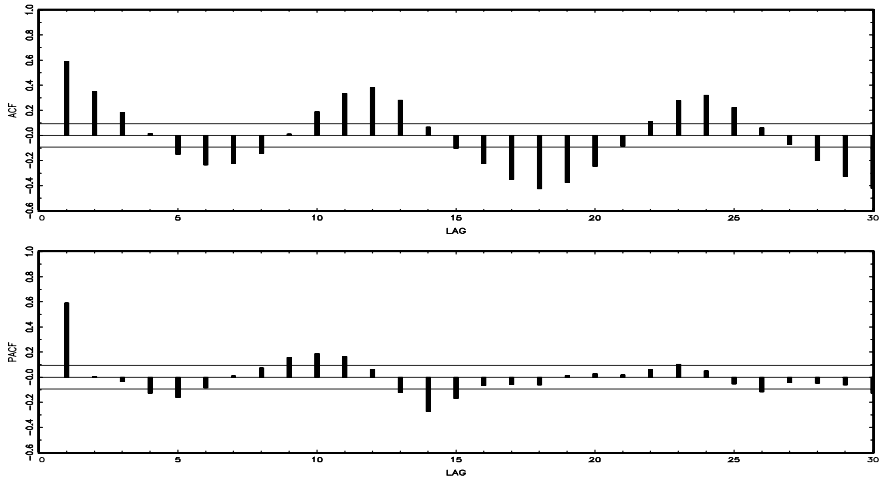


Figure 5.11 Sample ACF and PACF of SOI.

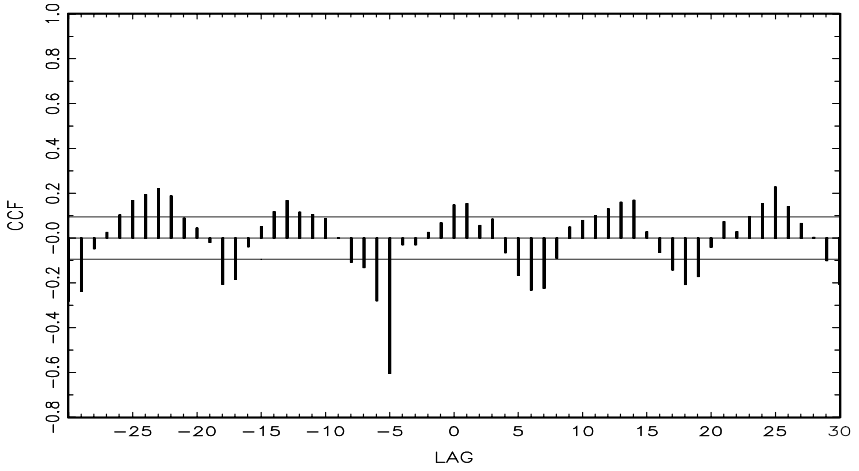


Figure 5.12 Sample CCF of the prewhitened, detrended SOI and the similarly transformed Recruitment series; negative lags indicate that SOI leads Recruitment.

Example 5.7 Relating the Prewhitened SOI to the Transformed Recruitment Series

We give a simple example of the suggested procedure for the SOI and the Recruitment series. Figure 5.11 shows the sample ACF and PACF of the detrended SOI index, and it is clear, from the PACF, that an

autoregressive series with $p = 1$ will do a reasonable job. Fitting the series gave $\widehat{\phi} = .589, \widehat{\sigma}_w^2 = .092$, and we applied the operator $(1 - .589B)$ to both x_t and y_t and computed the cross-correlation function, which is shown in Figure 5.12. Noting the apparent shift of $d = 5$ months and the exponential decrease thereafter, it seems plausible to hypothesize a model of the form

$$\begin{aligned} \alpha(B) &= \delta_0 B^5(1 + \omega_1 B + \omega_1^2 B^2 + \cdots) \\ &= \frac{\delta_0 B^5}{1 - \omega_1 B} \end{aligned}$$

for the transfer function. In this case, we would expect ω_1 to be negative.

In some cases, we may postulate the form of the separate components $\delta(B)$ and $\omega(B)$, so we might write the equation

$$y_t = \frac{\delta(B)B^d}{\omega(B)}x_t + \eta_t$$

as

$$\omega(B)y_t = \delta(B)B^d x_t + \omega(B)\eta_t,$$

or in regression form

$$y_t = \sum_{k=1}^{r} \omega_k y_{t-k} + \sum_{k=0}^{s} \delta_k x_{t-d-k} + u_t, \tag{5.66}$$

where

$$u_t = \omega(B)\eta_t. \tag{5.67}$$

The form of (5.66) suggests doing a regression on the lagged versions of both the input and output series to obtain $\widehat{\boldsymbol{\beta}}$, the estimate of the $(r + s + 1) \times 1$ regression vector

$$\boldsymbol{\beta} = (\omega_1, \ldots, \omega_r, \delta_0, \delta_1, \ldots, \delta_s)'.$$

The residuals from the regression above, say,

$$\widehat{u}_t = y_t - \widehat{\boldsymbol{\beta}}' \boldsymbol{z}_t,$$

where

$$\boldsymbol{z}_t = (y_{t-1}, \ldots, y_{t-r}, x_{t-d}, \ldots, x_{t-d-s})'$$

denotes the usual vector of independent variables, could be used to approximate the best ARMA model for the noise process η_t, because we can compute an estimator for that process from the (5.67), using $\widehat{u}_t$ and $\widehat{\omega}(B)$ and applying the moving average operator to get $\widehat{\eta}_t$. Fitting an ARMA(p_η, q_η) model to the this estimated noise then completes the specification. The preceding suggests the following sequential procedure for fitting the transfer function model to data.

(i) Fit an ARMA model to the input series x_t to estimate the parameters $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma_w^2$ in the specification (5.60). Retain ARMA coefficients for use in Step (ii) and the fitted residuals $\widehat{w}_t$ for use in Step (iii).

(ii) Apply the operator determined in Step (i), that is,

$$\widehat{\phi}(B)y_t = \widehat{\theta}(B)\tilde{y}_t,$$

to determine the transformed output series $\tilde{y}_t$.

(iii) Use the cross-correlation function between $\tilde{y}_t$ and $\widehat{w}_t$ in (i) and (ii) to suggest a form for the components of the polynomial

$$\alpha(B) = \frac{\delta(B)B^d}{\omega(B)}$$

and the estimated time delay d.

(iv) Obtain $\widehat{\boldsymbol{\beta}} = (\widehat{\omega}_1, \ldots, \widehat{\omega}_r, \widehat{\delta}_0, \widehat{\delta}_1, \ldots, \widehat{\delta}_s)$ by fitting a linear regression of the form (5.66). Retain the residuals $\widehat{u}_t$ for use in Step (v).

(v) Apply the moving average transformation (5.67) to the residuals $\widehat{u}_t$ to find the noise series $\widehat{\eta}_t$, and fit an ARMA model to the noise, obtaining the estimated coefficients in $\widehat{\phi}_\eta(B)$ and $\widehat{\theta}_\eta(B)$.

The above procedure is fairly reasonable, but does not have any recognizable overall optimality. Simultaneous least squares estimation, based on the observed x_t and y_t, can be accomplished by noting that the transfer function model can be written as

$$y_t = \frac{\delta(B)B^d}{\omega(B)}x_t + \frac{\theta_\eta(B)}{\phi_\eta(B)}z_t,$$

which can be put in the form

$$\omega(B)\phi_\eta(B)y_t = \phi_\eta(B)\delta(B)B^d x_t + \omega(B)\theta_\eta(B)z_t, \tag{5.68}$$

and it is clear that we may use least squares to minimize $\sum_t z_t^2$, as in earlier sections. We may also express the transfer function in state-space form (see Brockwell and Davis, 1991, Chapter 12). It is often easier to fit a transfer function model in the spectral domain as presented in §4.10.

Example 5.8 Transfer Function Model for the SOI and Recruitment Series

We illustrate the procedure for fitting a transfer function model of the form suggested in Example 5.7 to the detrended SOI series (x_t) and the detrended Recruitment series (y_t). The results reported here can be

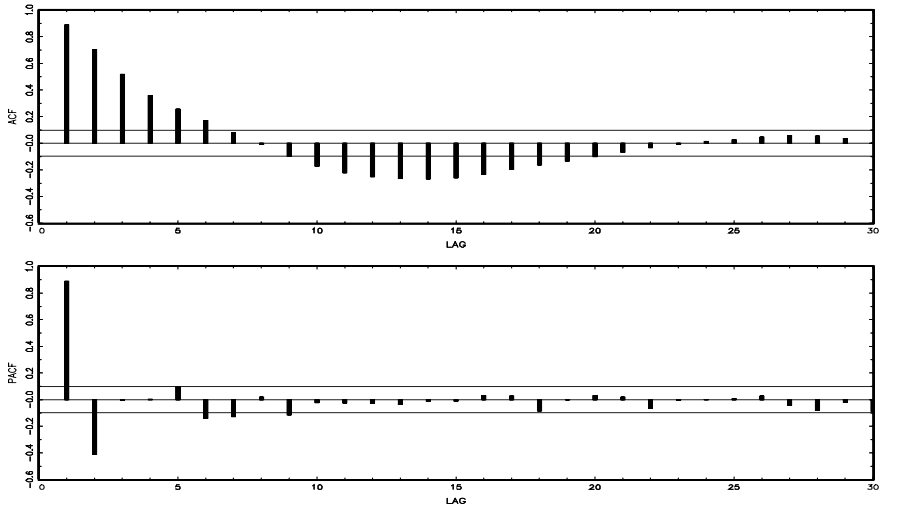


Figure 5.13 ACF and PACF of the estimated noise $\widehat{\eta}_t$ departures from the transfer function model.

compared with the results obtained from the frequency domain approach used in Example 4.23. Note first that Steps (i)-(iii). have already been applied to determine the ARMA model

$$(1 - .589B)x_t = w_t,$$

where $\widehat{\sigma}_w^2 = .092$. Using the model determined in Example 5.7, we run the regression

$$y_t = \omega_1 y_{t-1} + \delta_0 x_{t-5} + u_t,$$

yielding $\widehat{\omega}_1 = .848, \widehat{\delta}_0 = -20.54$, where the residuals satisfy

$$\widehat{u}_t = (1 - .848B)\eta_t.$$

This completes Step (iv). To complete the specification, we apply the moving average operator above to estimate the original noise series η_t and fit a second-order autoregressive model, based on the ACF and PACF shown in Figure 5.13. We obtain

$$(1 - 1.255B + .410B^2)\eta_t = z_t,$$

with $\widehat{\sigma}_z^2 = 52.46$ as the estimated error variance.

5.7 Multivariate ARMAX Models

To understand multivariate time series models and their capabilities, we first present an introduction to multivariate time series regression techniques. A useful extension of the basic univariate regression model presented in §2.2 is the case in which we have more than one output series, that is, multivariate regression analysis. Suppose, instead of a single output variable y_t, a collection of k output variables $y_{t1}, y_{t2}, \ldots, y_{tk}$ exist that are related to the inputs as

$$y_{ti} = \beta_{i1} z_{t1} + \beta_{i2} z_{t2} + \cdots + \beta_{ir} z_{tr} + w_{ti} \tag{5.69}$$

for each of the $i = 1, 2, \ldots, k$ output variables. We assume the w_{ti} variables are correlated over the variable identifier i, but are still independent over time. Formally, we assume $\mathrm{cov}\{w_{si}, w_{tj}\} = \sigma_{ij}$ for $s = t$ and is zero otherwise. Then, writing (5.69) in matrix notation, with $\boldsymbol{y}_t = (y_{t1}, y_{t2}, \ldots, y_{tk})'$ being the vector of outputs, and $\mathcal{B} = \{\beta_{ij}\}, i = 1, \ldots, k,\ j = 1, \ldots, r$ being an $k \times r$ matrix containing the regression coefficients, leads to the simple looking form

$$\boldsymbol{y}_t = \mathcal{B}\boldsymbol{z}_t + \boldsymbol{w}_t. \tag{5.70}$$

Here, the $k \times 1$ vector process $\boldsymbol{w}_t$ is assumed to be a collection of independent vectors with common covariance matrix $E\{\boldsymbol{w}_t \boldsymbol{w}_t'\} = \Sigma_w$, the $k \times k$ matrix containing the covariances σ_{ij}. The maximum likelihood estimator, under the assumption of normality, for the regression matrix in this case is

$$\widehat{\mathcal{B}} = Y'Z(Z'Z)^{-1}, \tag{5.71}$$

where $Z' = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n]$ is as before and $Y' = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n]$. The error covariance matrix Σ_w is estimated by

$$\widehat{\Sigma}_w = \frac{1}{(n-r)} \sum_{t=1}^{n} (\boldsymbol{y}_t - \widehat{\mathcal{B}}\boldsymbol{z}_t)(\boldsymbol{y}_t - \widehat{\mathcal{B}}\boldsymbol{z}_t)'. \tag{5.72}$$

The uncertainty in the estimators can be evaluated from

$$\mathrm{se}(\widehat{\beta}_{ij}) = \sqrt{\widehat{\sigma}_{jj} c_{ii}}, \tag{5.73}$$

for $i = 1, \ldots, r$, $j = 1, \ldots, k$, where se denotes estimated standard error, $\widehat{\sigma}_{jj}$ is the j-th diagonal element of $\widehat{\Sigma}_w$, and c_{ii} is the i-th diagonal element of $(\sum_{t=1}^{n} \boldsymbol{z}_t \boldsymbol{z}_t')^{-1}$.

Also, the information theoretic criterion changes to

$$\mathrm{AIC} = \ln |\widehat{\Sigma}_w| + \frac{2}{n}\left(kr + \frac{k(k+1)}{2}\right). \tag{5.74}$$

and SIC replaces the second term in (5.74) by $K \ln n/n$ where $K = kr + k(k+1)/2$. Bedrick and Tsai (1994) have given a corrected form for AIC in the multivariate case as

$$\mathrm{AICc} = \ln |\widehat{\Sigma}_w| + \frac{k(r+n)}{n-k-r-1}. \tag{5.75}$$

Many data sets involve more than one time series, and we are often interested in the possible dynamics relating all series. In this situation, we are interested in modeling and forecasting $k \times 1$ vector-valued time series $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tk})'$, $t = 0, \pm 1, \pm 2, \ldots$. Unfortunately, extending univariate ARMA models to the multivariate case is not so simple. The multivariate autoregressive model, however, is a straight-forward extension of the univariate AR model.

For the first-order vector autoregressive model, VAR(1), we take

$$\boldsymbol{x}_t = \boldsymbol{\alpha} + \Phi \boldsymbol{x}_{t-1} + \boldsymbol{w}_t, \tag{5.76}$$

where Φ is a $k \times k$ transition matrix that expresses the dependence of $\boldsymbol{x}_t$ on $\boldsymbol{x}_{t-1}$. The vector white noise process $\boldsymbol{w}_t$ is assumed to be multivariate normal with mean-zero and covariance matrix

$$E\left(\boldsymbol{w}_t \boldsymbol{w}_t'\right) = \boldsymbol{\Sigma}_w. \tag{5.77}$$

The vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)'$ appears as the constant in the regression setting. If $E(\boldsymbol{x}_t) = \boldsymbol{\mu}$, then $\boldsymbol{\alpha} = (I - \Phi)\boldsymbol{\mu}$.

Note the similarity between the VAR model and the multivariate linear regression model (5.70). The regression formulas carry over, and we can, on observing $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, set up the model (5.76) with $\boldsymbol{y}_t = \boldsymbol{x}_t$, $\mathcal{B} = (\boldsymbol{\alpha}, \Phi)$ and $\boldsymbol{z}_t = (1, \boldsymbol{x}_{t-1}')'$. Then, write the solution as (5.71) with the conditional maximum likelihood estimator for the covariance matrix given by

$$\widehat{\Sigma}_w = (n-1)^{-1} \sum_{t=2}^{n} (\boldsymbol{x}_t - \widehat{\boldsymbol{\alpha}} - \widehat{\Phi}\boldsymbol{x}_{t-1})(\boldsymbol{x}_t - \widehat{\boldsymbol{\alpha}} - \widehat{\Phi}\boldsymbol{x}_{t-1})'. \tag{5.78}$$

Example 5.9 Pollution, Weather, and Mortality

For example, for the three-dimensional series composed of detrended cardiovascular mortality x_{t1}, temperature x_{t2}, and particulate levels x_{t3}, introduced in Example 2.2, take $\boldsymbol{x}_t = (x_{t1}, x_{t2}, x_{t3})'$ as a vector of dimension $k = 3$. We might envision dynamic relations among the three series defined as the first order relation,

$$x_{t1} = \alpha_1 + \phi_{11} x_{t-1,1} + \phi_{12} x_{t-1,2} + \phi_{13} x_{t-1,3} + w_{t1},$$

which expresses the current value of mortality as a linear combination of its immediate past value and the past values of temperature and particulate levels. Similarly,

$$x_{t2} = \alpha_2 + \phi_{21} x_{t-1,1} + \phi_{22} x_{t-1,2} + \phi_{23} x_{t-1,3} + w_{t2}$$

and

$$x_{t3} = \alpha_3 + \phi_{31} x_{t-1,1} + \phi_{32} x_{t-1,2} + \phi_{33} x_{t-1,3} + w_{t3}$$

express the dependence of temperature and particulate levels on the other series. Of course, methods for the preliminary identification of these models exist, and we will discuss these methods shortly.

For this particular case, we obtain $\widehat{\boldsymbol{\alpha}} = (-4.57, 6.09, 19.78)'$ and

$$\widehat{\Phi} = \begin{pmatrix} .47(.04) & -.36(.03) & .10(.02) \\ -.24(.04) & .49(.04) & -.13(.02) \\ -.13(.08) & -.48(.07) & .58(.04) \end{pmatrix},$$

where the standard errors, computed as in (5.73), are given in parentheses. Hence, for the vector $(x_{t1}, x_{t2}, x_{t3}) = (M_t, T_t, P_t)$, with M_t, T_t and P_t denoting mortality, temperature, and particulate level, respectively, we obtain the prediction equation for mortality,

$$\widehat{M_t} = -4.57 + .47M_{t-1} - .36T_{t-1} + .10P_{t-1}.$$

Comparing observed and predicted mortality with this model leads to an R^2 of about .78, whereas the value in the regression model fitted by the method of Example 2.2 gave an $R^2 = .69$.

It is easy to extend the VAR(1) process to higher orders, VAR(p). To do this, we use the notation of (5.70) and write the vector of regressors as

$$\boldsymbol{z}_t = (1, \boldsymbol{x}'_{t-1}, \boldsymbol{x}'_{t-2}, \ldots \boldsymbol{x}'_{t-p})'$$

and the regression matrix as $\mathcal{B} = (\boldsymbol{\alpha}, \Phi_1, \Phi_2, \ldots, \Phi_p)$. Then, this regression model can be written as

$$\boldsymbol{x}_t = \boldsymbol{\alpha} + \sum_{j=1}^{p} \Phi_j \boldsymbol{x}_{t-j} + \boldsymbol{w}_t \tag{5.79}$$

for $t = p+1, \ldots, n$. The $k \times k$ error sum of products matrix becomes

$$RSP = \sum_{t=p+1}^{n} (\boldsymbol{x}_t - \mathcal{B}\boldsymbol{z}_t)(\boldsymbol{x}_t - \mathcal{B}\boldsymbol{z}_t)', \tag{5.80}$$

so that the conditional maximum likelihood estimator for the error covariance matrix Σ_w is

$$\widehat{\Sigma}_w = RSP/(n-p), \tag{5.81}$$

as in the multivariate regression case, except now only $n-p$ residuals exist in (5.80). For the multivariate case, we have found that the Schwarz criterion

$$\text{SIC} = \log|\widehat{\Sigma}_w| + k^2 p \ln n/n, \tag{5.82}$$

gives more reasonable classifications than either AIC or corrected version AICc. The result is consistent with those reported in simulations by Lütkepohl (1985).

Table 5.1 Summary Statistics for Example 5.10

Order (p)	k^2p	$\lvert\widehat{\Sigma}_w\rvert$	SIC	AICc
1	505	118,520	11.79	14.71
2	503	74,708	11.44	14.26
3	501	70,146	11.49	14.21
4	499	65,268	11.53	14.15
5	497	59,684	11.55	14.08

Example 5.10 Mortality, Pollution and Temperature Data

A trivariate AR(2) model for the data in Example 5.9 yields

$$\widehat{\Phi}_1 = \begin{pmatrix} .30(.04) & -.20(.04) & .04(.02) \\ -.11(.05) & .26(.05) & -.05(.03) \\ .08(.09) & -.39(.09) & .39(.05) \end{pmatrix},$$

$$\widehat{\Phi}_2 = \begin{pmatrix} .28(.04) & -.08(.04) & .07(.03) \\ -.04(.05) & .36(.05) & -.09(.03) \\ -.33(.09) & .05(.09) & .38(.05) \end{pmatrix}.$$

In Table 5.1, fitting successively higher order models beyond $p = 2$ does not improve the value of SIC, and we would tend to settle on the second-order model. Note that the value of AICc continues to decrease as the model order increases.

A $k \times 1$ vector-valued time series $\boldsymbol{x}_t$, for $t = 0, \pm1, \pm2, \ldots$, is said to be VARMA$(p, q)$ if $\boldsymbol{x}_t$ is stationary and

$$\boldsymbol{x}_t = \boldsymbol{\alpha} + \Phi_1\boldsymbol{x}_{t-1} + \cdots + \Phi_p\boldsymbol{x}_{t-p} + \boldsymbol{w}_t + \Theta_1\boldsymbol{w}_{t-1} + \cdots + \Theta_q\boldsymbol{w}_{t-q}, \qquad (5.83)$$

with $\Phi_p \neq 0$, $\Theta_q \neq 0$, and $\Sigma_w > 0$ (that is, Σ_w is positive definite). The coefficient matrices Φ_j; $j = 1, ..., p$ and Θ_j; $j = 1, ..., q$ are, of course, $p \times p$ matrices. If $\boldsymbol{x}_t$ has mean $\boldsymbol{\mu}$ then $\boldsymbol{\alpha} = (I - \Phi_1 - \cdots - \Phi_p)\boldsymbol{\mu}$. As in the univariate case, we will have to place a number of conditions on the multivariate ARMA model to ensure the model is unique and has desirable properties such as causality. These conditions will be discussed shortly.

The special form assumed for the constant component, $\boldsymbol{\alpha}$, of the vector ARMA model in (5.83) can be generalized to include a fixed $r \times 1$ vector of inputs, $\boldsymbol{u}_t$. That is, we could have proposed the vector ARMAX model,

$$\boldsymbol{x}_t = \Gamma\boldsymbol{u}_t + \sum_{j=1}^{p} \Phi_j\boldsymbol{x}_{t-j} + \sum_{k=1}^{q} \Theta_k\boldsymbol{w}_{t-k} + \boldsymbol{w}_t, \qquad (5.84)$$

where Γ is a $p \times r$ parameter matrix. The X in ARMAX refers to the exogenous vector process we have denoted here by $\boldsymbol{u}_t$. The introduction of exogenous

variables through replacing $\boldsymbol{\alpha}$ by $\Gamma \boldsymbol{u}_t$ does not present any special problems in making inferences. For example, the case of the ARX model, that is, $q = 0$ in (5.84), can be estimated using standard regression results. In this case, the model can be written as a multivariate regression model in which the vector of regressors are

$$\boldsymbol{z}_t = (\boldsymbol{u}_t', \boldsymbol{x}_{t-1}', ..., \boldsymbol{x}_{t-p}')' \tag{5.85}$$

and the new regression matrix is

$$\mathcal{B} = [\Gamma, \Phi_1, \Phi_2, ..., \Phi_p]. \tag{5.86}$$

The general VARMA model, (5.83), is a special case of the vector ARMAX model, (5.84), with $r = 1$, $\boldsymbol{u}_t = 1$, and $\Gamma = \boldsymbol{\alpha}$.

As previously indicated, extending univariate AR (or pure MA) models to the vector case is fairly easy, but extending univariate ARMA models to the multivariate case is not a simple matter. Our discussion will be brief, but interested readers can get more details in Lütkepohl (1993), Reinsel (1997), and Tiao and Tsay (1989).

In the multivariate case, the autoregressive operator is

$$\Phi(B) = I - \Phi_1 B - \cdots - \Phi_p B^p, \tag{5.87}$$

and the moving average operator is

$$\Theta(B) = I + \Theta_1 B + \cdots + \Theta_q B^q, \tag{5.88}$$

The zero-mean VARMA(p, q) model is then written in the concise form as

$$\Phi(B)\boldsymbol{x}_t = \Theta(B)\boldsymbol{w}_t. \tag{5.89}$$

The model is said to be causal if the roots of $|\Phi(z)|$ (where $|\cdot|$ denotes determinant) are outside the unit circle, $|z| > 1$; that is, $|\Phi(z)| \neq 0$ for any value z such that $|z| \leq 1$. In this case, we can write

$$\boldsymbol{x}_t = \Psi(B)\boldsymbol{w}_t,$$

where $\Psi(B) = \sum_{j=0}^{\infty} \Psi_j B^j$, $\Psi_0 = I$, and $\sum_{j=0}^{\infty} ||\Psi_j|| < \infty$. The model is said to be invertible if the roots of $|\Theta(z)|$ lie outside the unit circle. Then, we can write

$$\boldsymbol{w}_t = \Pi(B)\boldsymbol{x}_t,$$

where $\Pi(B) = \sum_{j=0}^{\infty} \Pi_j B^j$, $\Pi_0 = I$, and $\sum_{j=0}^{\infty} ||\Pi_j|| < \infty$. Analogous to the univariate case, we can determine the matrices Ψ_j by solving $\Psi(z) = \Phi(z)^{-1}\Theta(z), |z| \leq 1$, and the matrices Π_j by solving $\Pi(z) = \Theta(z)^{-1}\Phi(z), |z| \leq 1$.

For a causal model, we can write $\boldsymbol{x}_t = \Psi(B)\boldsymbol{w}_t$ so the general autocovariance structure of an ARMA(p, q) model is

$$\Gamma(h) = \text{cov}(\boldsymbol{x}_{t+h}, \boldsymbol{x}_t) = E(\boldsymbol{x}_{t+h}\boldsymbol{x}_t') = \sum_{j=0}^{\infty} \Psi_{j+h}\Sigma_w\Psi_j'. \tag{5.90}$$

Note, $\Gamma(-h) = \Gamma'(h)$ so we will only exhibit the autocovariances for $h \geq 0$. For pure MA(q) processes, (5.90) becomes

$$\Gamma(h) = \sum_{j=0}^{q-h} \Theta_{j+h} \Sigma_w \Theta_j', \tag{5.91}$$

where $\Theta_0 = I$. Of course, (5.91) implies $\Gamma(h) = 0$ for $h > q$. For pure AR(p) models, the autocovariance structure leads to the multivariate version of the Yule–Walker equations:

$$\Gamma(h) = \sum_{j=1}^{p} \Phi_j \Gamma(h-j), \quad h = 1, 2, ..., \tag{5.92}$$

$$\Gamma(0) = \sum_{j=1}^{p} \Phi_j \Gamma(-j) + \Sigma_w. \tag{5.93}$$

As in the univariate case, we will need conditions for model uniqueness. These conditions are similar to the condition in the univariate case the the autoregressive and moving average polynomials have no common factors. To explore the uniqueness problems that we encounter with multivariate ARMA models, consider a bivariate AR(1) process, $\boldsymbol{x}_t = (x_{t,1}, x_{t,2})'$, given by

$$\begin{aligned} x_{t,1} &= \phi x_{t-1,2} + w_{t,1}, \\ x_{t,2} &= w_{t,2}, \end{aligned}$$

where $w_{t,1}$ and $w_{t,2}$ are independent white noise processes and $|\phi| < 1$. Both processes, $x_{t,1}$ and $x_{t,2}$ are causal and invertible. Moreover, the processes are jointly stationary because $\text{cov}(x_{t+h,1}, x_{t,2}) = \phi \, \text{cov}(x_{t+h-1,2}, x_{t,2}) \equiv \phi \, \gamma_{2,2}(h-1) = \phi \sigma_{w_2}^2 \delta_1^h$ does not depend on t; note, $\delta_1^h = 1$ when $h = 1$, otherwise, $\delta_1^h = 0$. In matrix notation, we can write this model as

$$\boldsymbol{x}_t = \Phi \boldsymbol{x}_{t-1} + \boldsymbol{w}_t, \tag{5.94}$$

where

$$\Phi = \begin{bmatrix} 0 & \phi \\ 0 & 0 \end{bmatrix}.$$

We can write (5.94) in operator notation as

$$\Phi(B)\boldsymbol{x}_t = \boldsymbol{w}_t$$

where

$$\Phi(z) = \begin{bmatrix} 1 & -\phi z \\ 0 & 1 \end{bmatrix}.$$

In addition, model (5.94) can be written as a bivariate ARMA(1,1) model

$$\boldsymbol{x}_t = \Phi_1 \boldsymbol{x}_{t-1} + \Theta_1 \boldsymbol{w}_{t-1} + \boldsymbol{w}_t, \tag{5.95}$$

where

$$\Phi_1 = \begin{bmatrix} 0 & \phi+\theta \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \Theta_1 = \begin{bmatrix} 0 & -\theta \\ 0 & 0 \end{bmatrix},$$

and θ is arbitrary. To verify this, we write (5.95), as $\Phi_1(B)\boldsymbol{x}_t = \Theta_1(B)\boldsymbol{w}_t$, or

$$\Theta_1(B)^{-1}\Phi_1(B)\boldsymbol{x}_t = \boldsymbol{w}_t,$$

where

$$\Phi_1(z) = \begin{bmatrix} 1 & -(\phi+\theta)z \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \Theta_1(z) = \begin{bmatrix} 1 & -\theta z \\ 0 & 1 \end{bmatrix}.$$

Then,

$$\Theta_1(z)^{-1}\Phi_1(z) = \begin{bmatrix} 1 & \theta z \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & -(\phi+\theta)z \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -\phi z \\ 0 & 1 \end{bmatrix} = \Phi(z),$$

where $\Phi(z)$ is the polynomial associated with the bivariate AR(1) model in (5.94). Because θ is arbitrary, the parameters of the ARMA(1,1) model given in (5.95) are not identifiable. No problem exists, however, in fitting the AR(1) model given in (5.94).

The problem in the previous discussion was caused by the fact that both $\Theta(B)$ and $\Theta(B)^{-1}$ are finite; such a matrix operator is called unimodular. If $U(B)$ is unimodular, $|U(z)|$ is constant. It is also possible for two seemingly different multivariate ARMA(p,q) models, say, $\Phi(B)\boldsymbol{x}_t = \Theta(B)\boldsymbol{w}_t$ and $\Phi_*(B)\boldsymbol{x}_t = \Theta_*(B)\boldsymbol{w}_t$, to be related through a unimodular operator, $U(B)$ as $\Phi_*(B) = U(B)\Phi(B)$ and $\Theta_*(B) = U(B)\Theta(B)$, in such a way that the orders of $\Phi(B)$ and $\Theta(B)$ are the same as the orders of $\Phi_*(B)$ and $\Theta_*(B)$, respectively. For example, consider the bivariate ARMA(1,1) models given by

$$\Phi\boldsymbol{x}_t \equiv \begin{bmatrix} 1 & -\phi B \\ 0 & 1 \end{bmatrix}\boldsymbol{x}_t = \begin{bmatrix} 1 & \theta B \\ 0 & 1 \end{bmatrix}\boldsymbol{w}_t \equiv \Theta w_t$$

and

$$\Phi_*(B)\boldsymbol{x}_t \equiv \begin{bmatrix} 1 & (\alpha-\phi)B \\ 0 & 1 \end{bmatrix}\boldsymbol{x}_t = \begin{bmatrix} 1 & (\alpha+\theta)B \\ 0 & 1 \end{bmatrix}\boldsymbol{w}_t \equiv \Theta_*(B)\boldsymbol{w}_t,$$

where α, ϕ, and θ are arbitrary constants. Note,

$$\Phi_*(B) \equiv \begin{bmatrix} 1 & (\alpha-\phi)B \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & \alpha B \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & -\phi B \\ 0 & 1 \end{bmatrix} \equiv U(B)\Phi(B)$$

and

$$\Theta_*(B) \equiv \begin{bmatrix} 1 & (\alpha+\theta)B \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & \alpha B \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & \theta B \\ 0 & 1 \end{bmatrix} \equiv U(B)\Theta(B).$$

In this case, both models have the same infinite MA representation $\boldsymbol{x}_t = \Psi(B)\boldsymbol{w}_t$, where

$$\Psi(B) = \Phi(B)^{-1}\Theta(B) = \Phi(B)^{-1}U(B)^{-1}U(B)\Theta(B) = \Phi_*(B)^{-1}\Theta_*(B).$$

This result implies the two models have the same autocovariance function $\Gamma(h)$. Two such ARMA(p,q) models are said to be observationally equivalent.

As previously mentioned, in addition to requiring causality and invertiblity, we will need some additional assumptions in the multivariate case to make sure that the model is unique. To ensure the identifiability of the parameters of the multivariate ARMA(p,q) model, we need the following additional two conditions: (i) the matrix operators $\Phi(B)$ and $\Theta(B)$ have no common left factors other than unimodular ones; that is, if $\Phi(B) = U(B)\Phi_*(B)$ and $\Theta(B) = U(B)\Theta_*(B)$, the common factor must be unimodular; and (ii) with q as small as possible and p as small as possible for that q, the matrix $[\Phi_p, \Theta_q]$ must be full rank, k. One suggestion for avoiding most of the aforementioned problems is to fit only vector AR(p) models in multivariate situations. Although this suggestion might be reasonable for many situations, this philosophy is not in accordance with law of parsimony because we might have to fit a large number of parameters to describe the dynamics of a process.

Analogous to the univariate case, we can define a sequence of matrices, Φ_{hh}, for $h = 1, 2, ...$, called the partial autoregression matrices at lag h. These matrices are obtained by solving the Yule–Walker equations of order h, namely,

$$\Gamma(\ell) = \sum_{j=1}^{h} \Phi_{jh}\Gamma(\ell - j), \quad \ell = 1, 2, ..., h. \tag{5.96}$$

The partial autoregression matrices can be viewed as the result of successive AR(h) fits to the data; that is,

$$\boldsymbol{x}_t = \sum_{j=1}^{h} \Phi_{jh}\boldsymbol{x}_{t-j} + \boldsymbol{w}_t, \quad h = 1, 2, \ldots . \tag{5.97}$$

If the process is truly an AR(p), the partial autoregression matrices have the property that $\Phi_{pp} = \Phi_p$ and $\Phi_{hh} = 0$ for $h > p$. Unlike the univariate case, however, the elements of these matrices are not partial correlations, or correlations of any kind. As in the univariate case, the Φ_{hh} can be obtained iteratively using a multivariate extension of the Durbin-Levinson algorithm; details can be found in Reinsel (1997).

The partial canonical correlations can be viewed as the multivariate extension of the PACF in the univariate case. In general, the first canonical correlation, λ_1, between the $k_1 \times 1$ random vector $\boldsymbol{X}_1$ and the $k_2 \times 1$ random vector $\boldsymbol{X}_2$, $k_1 \leq k_2$, with variance–covariance matrices Σ_{11} and Σ_{22}, respectively, is the largest possible correlation between a linear combination of the components of $\boldsymbol{X}_1$, say, $\boldsymbol{\alpha}'\boldsymbol{X}_1$, and a linear combination of the components of $\boldsymbol{X}_2$, say, $\boldsymbol{\beta}'\boldsymbol{X}_2$, where $\boldsymbol{\alpha}$ is $k_1 \times 1$ and $\boldsymbol{\beta}$ is $k_2 \times 1$. That is,

$$\lambda_1 = \max_{\boldsymbol{\alpha}.\boldsymbol{\beta}} \operatorname{corr}\left(\boldsymbol{\alpha}'\boldsymbol{X}_1, \boldsymbol{\beta}'\boldsymbol{X}_2\right),$$

subject to the constraints $\operatorname{var}(\boldsymbol{\alpha}'\boldsymbol{X}_1) = \boldsymbol{\alpha}'\Sigma_{11}\boldsymbol{\alpha} = 1$ and $\operatorname{var}(\boldsymbol{\beta}'\boldsymbol{X}_2) = \boldsymbol{\beta}'\Sigma_{22}\boldsymbol{\beta} = 1$. If we let $\Sigma_{ij} = \operatorname{cov}(\boldsymbol{X}_i, \boldsymbol{X}_j)$, for $i, j = 1, 2$, then λ_1^2 is the largest eigenvalue

of the matrix $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$; see Johnson and Wichern (1992, Chapter 10) for details. We call the solutions $U_1 = \boldsymbol{\alpha}_1'\boldsymbol{X}_1$ and $V_1 = \boldsymbol{\beta}_1'\boldsymbol{X}_2$ the first canonical variates, that is, $\lambda_1 = \text{corr}(U_1, V_1)$, and $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ are the coefficients of the linear combinations that maximize the correlation. In a similar fashion, the second canonical correlation, λ_2, is then the largest possible correlation between $\boldsymbol{\alpha}'\boldsymbol{X}_1$ and $\boldsymbol{\beta}'\boldsymbol{X}_2$ such that $\boldsymbol{\alpha}$ is orthogonal to $\boldsymbol{\alpha}_1$ (that is, $\boldsymbol{\alpha}'\boldsymbol{\alpha}_1 = 0$), and $\boldsymbol{\beta}$ is orthogonal to $\boldsymbol{\beta}_1$ ($\boldsymbol{\beta}'\boldsymbol{\beta}_1 = 0$) . If we call the solutions $U_2 = \boldsymbol{\alpha}_2'\boldsymbol{X}_1$ and $V_2 = \boldsymbol{\beta}_2'\boldsymbol{X}_2$, then $\text{corr}(U_1, U_2) = 0 = \text{corr}(V_1, V_2)$, $\text{corr}(U_i, V_j) = 0$ for $i \neq j$, and by design, $\lambda_1^2 \geq \lambda_2^2$. Also, λ_2^2 is the second largest eigenvalue of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Continuing this way, we obtain the squared canonical correlations $1 \geq \lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_{k_1}^2 \geq 0$ as the ordered eigenvalues of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. The canonical correlations, λ_j, are typically taken to be nonnegative.

We can extend this idea to obtain partial canonical correlations between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ given another random $k_3 \times 1$ vector $\boldsymbol{X}_3$. Let $\Sigma_{ij} = \text{cov}(\boldsymbol{X}_i, \boldsymbol{X}_j)$, for $i, j = 1, 2, 3$. The regression of $\boldsymbol{X}_1$ on $\boldsymbol{X}_3$ is $\Sigma_{13}\Sigma_{33}^{-1}\boldsymbol{X}_3$ so that $\boldsymbol{X}_{1|3} = \boldsymbol{X}_1 - \Sigma_{13}\Sigma_{33}^{-1}\boldsymbol{X}_3$ can be thought of as $\boldsymbol{X}_1$ with the linear effects of $\boldsymbol{X}_3$ removed (*partialled out*). Similarly, $\boldsymbol{X}_{2|3} = \boldsymbol{X}_2 - \Sigma_{23}\Sigma_{33}^{-1}\boldsymbol{X}_3$ can be thought of as $\boldsymbol{X}_2$ with the linear effects of $\boldsymbol{X}_3$ partialled out. The partial variance–covariance matrices are $\Sigma_{ij|3} = \text{cov}(\boldsymbol{X}_{i|3}, \boldsymbol{X}_{j|3}) = \Sigma_{ij} - \Sigma_{i3}\Sigma_{33}^{-1}\Sigma_{3j}$, for $i, j = 1, 2$. The squared partial canonical correlations between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ given $\boldsymbol{X}_3$ are the ordered eigenvalues of $\Sigma_{11|3}^{-1}\Sigma_{12|3}\Sigma_{22|3}^{-1}\Sigma_{21|3}$.

For a stationary vector process $\boldsymbol{x}_t$, the partial canonical correlations at lag h, for $h = 2, 3, ...$, denoted $\lambda_1(h) \geq \lambda_2(h) \geq \cdots \geq \lambda_k(h) \geq 0$, are defined to be the partial canonical correlations between $\boldsymbol{x}_h$ and $\boldsymbol{x}_0$ with the effects of $\boldsymbol{X} = (\boldsymbol{x}_{h-1}', ..., \boldsymbol{x}_1')'$ removed. For ease of notation, we put $r = h - 1$. Let $\Sigma_{00|X} = \Gamma(0) - \Gamma_1^{(r)}\Gamma_{r,r}^{-1}\Gamma_1^{(r)'}$, where $\Gamma_{r,r} = \{\Gamma(i-j)\}_{i,j=1}^{r}$ is a $kr \times kr$ symmetric matrix, and $\Gamma_1^{(r)} = [\Gamma(r)', \Gamma(r-1)', ..., \Gamma(1)']$ is $k \times kr$. Similarly, let $\Sigma_{hh|X} = \Gamma(0) - \Gamma_r^{(1)}\Gamma_{r,r}^{-1}\Gamma_r^{(1)'}$, where $\Gamma_r^{(1)} = [\Gamma(1), \Gamma(2), ..., \Gamma(r)]$ is $k \times kr$. Also needed are $\Sigma_{h0|X} = \Gamma(r) - \Gamma_r^{(1)}\Gamma_{r,r}^{-1}\Gamma_1^{(r)'}$ and $\Sigma_{0h|X} = \Sigma_{h0|X}'$. The squared partial canonical correlations, $\lambda_j^2(h)$, $j = 1, ..., k$ at lag h, $h = 2, 3, ...$, are given by the ordered eigenvalues of $\Sigma_{00|X}^{-1}\Sigma_{0h|X}\Sigma_{hh|X}^{-1}\Sigma_{h0|X}$. The inversion of $\Gamma_{r,r}$, when h is large will, be a problem; see Reinsel (1997) for methods that avoid having to invert $\Gamma_{r,r}$. Finally, we will define the partial canonical correlations between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ to be the lag-one canonical correlations. In this case, $\lambda_j^2(1)$, $j = 1, ..., k$ are the ordered eigenvalues of $\Gamma(0)^{-1}\Gamma(1)\Gamma(0)^{-1}\Gamma(1)'$.

Prediction and estimation for identifiable multivariate ARMA models follow analogously to the univariate case, except in the general case, the estimation of the coefficient parameters and Σ_w must be done simultaneously. Preliminary identification of the model uses the sample autocovariance matrices, the sample partial autoregression matrices, and the sample partial canonical correlations. We illustrate the techniques using the mortality data of Examples 2.2, 5.9, and 5.10.

Example 5.11 Identification, Estimation and Prediction for the Mortality Series

As in Example 5.10, we consider the trivariate series composed of detrended cardiovascular mortality x_{t1}, temperature x_{t2}, and particulate levels x_{t3}, and set $\boldsymbol{x}_t = (x_{t1}, x_{t2}, x_{t3})'$ as the three-dimensional data vector.

Estimation of the autocovariance matrix is similar to the univariate case, that is, with $\bar{\boldsymbol{x}} = n^{-1}\sum_{t=1}^{n}\boldsymbol{x}_t$, as an estimate of $\boldsymbol{\mu} = E\boldsymbol{x}_t$,

$$\widehat{\Gamma}(h) = n^{-1}\sum_{t=1}^{n-h}(\boldsymbol{x}_{t+h} - \bar{\boldsymbol{x}})(\boldsymbol{x}_t - \bar{\boldsymbol{x}})', \quad h = 0, 1, 2, .., n-1, \tag{5.98}$$

and $\widehat{\Gamma}(-h) = \widehat{\Gamma}(h)'$. If $\widehat{\gamma}_{i,j}(h)$ denotes the element in the i-th row and j-th column of $\widehat{\Gamma}(h)$, the cross-correlation functions (CCF), as discussed in (1.35), are estimated by

$$\widehat{\rho}_{i,j}(h) = \frac{\widehat{\gamma}_{i,j}(h)}{\sqrt{\widehat{\gamma}_{i,i}(0)}\sqrt{\widehat{\gamma}_{j,j}(0)}} \quad h = 0, 1, 2, .., n-1. \tag{5.99}$$

When $i = j$ in (5.99), we get the estimated autocorrelation function (ACF) of the individual series. The first six estimated autocovariance matrices, $\widehat{\Gamma}(h)$, $h = 0, 1, ..., 5$, are (we have rounded the entries to integers to ease the display):

$$\widehat{\Gamma}(0) = \begin{bmatrix} 79 & -37 & 62 \\ -37 & 81 & -2 \\ 62 & -2 & 227 \end{bmatrix} \quad \widehat{\Gamma}(1) = \begin{bmatrix} 56 & -46 & 52 \\ -45 & 49 & -45 \\ 44 & -35 & 125 \end{bmatrix}$$

$$\widehat{\Gamma}(2) = \begin{bmatrix} 56 & -42 & 62 \\ -42 & 50 & -48 \\ 35 & -20 & 136 \end{bmatrix} \quad \widehat{\Gamma}(3) = \begin{bmatrix} 47 & -42 & 59 \\ -41 & 44 & -55 \\ 27 & -18 & 123 \end{bmatrix} \tag{5.100}$$

$$\widehat{\Gamma}(4) = \begin{bmatrix} 44 & -34 & 72 \\ -39 & 46 & -53 \\ 16 & -9 & 120 \end{bmatrix} \quad \widehat{\Gamma}(5) = \begin{bmatrix} 38 & -35 & 68 \\ -39 & 39 & -67 \\ 7 & 3 & 104 \end{bmatrix}.$$

Inspecting the autocovariance matrices, we find mortality, x_{t1}, and temperature, x_{t2}, are negatively correlated at about the same strength for both positive and negative lags. The strongest cross-correlation occurs at lag ± 1, where $\widehat{\rho}_{12}(-1) \approx -45/\sqrt{79}\sqrt{81} = -.56$, and $\widehat{\rho}_{12}(1) \approx -46/\sqrt{79}\sqrt{81} = -.58$. Also, mortality x_{t1} and particulates x_{t3} are positively correlated, the strongest correlation being when particulates leads mortality by about one month, $\widehat{\rho}_{13}(4) \approx 72/\sqrt{79}\sqrt{227} = .54$. Finally, we note that particulates and temperature are negatively correlated, the strongest displayed value (which is approximately the strongest overall

correlation between the two series) is when particulates leads temperature by about five weeks, $\widehat{\rho}_{23}(5) \approx -67/\sqrt{81}\sqrt{227} = .49$. The autocovariance matrices do not cut off at any small lag, and hence a pure moving average model is not indicated.

Replacing $\Gamma(h)$ by $\widehat{\Gamma}(h)$ in (5.96), we can obtain estimates of the partial autoregression matrices. The first four estimated matrices are

$$\widehat{\Phi}_{11} = \begin{bmatrix} .47 & -.36 & .10 \\ -.25 & .49 & -.13 \\ -.12 & -.48 & .58 \end{bmatrix} \quad \widehat{\Phi}_{22} = \begin{bmatrix} .27 & -.08 & .07 \\ -.04 & .35 & -.09 \\ -.33 & .05 & .38 \end{bmatrix}$$

$$\widehat{\Phi}_{33} = \begin{bmatrix} -.04 & .02 & -.01 \\ .00 & .11 & -.03 \\ -.21 & .07 & .17 \end{bmatrix} \quad \widehat{\Phi}_{44} = \begin{bmatrix} -.04 & .08 & .06 \\ -.07 & .17 & .01 \\ -.26 & .12 & .13 \end{bmatrix}.$$

As explained above (5.97), we can use (5.96) to estimate successive AR(h) models with parameter estimates $\widehat{\Phi}_j = \widehat{\Phi}_{jh}$, $j = 1, \ldots, h$, and $h = 1, 2, \ldots$. Note, $\widehat{\Phi}_{11}$ is practically the same as $\widehat{\Phi}$ in Example 5.9, and $\widehat{\Phi}_{22}$ is practically the same as $\widehat{\Phi}_2$ in Example 5.10. The only difference in the estimates is that we are using Yule–Walker here, whereas regression was used in the other examples. These matrices contain small components after lag two, indicating the AR(2) relationship, although there is evidence of some relationship between mortality and particulates at lags of three and four weeks.

The estimated autocovariance matrices can also be used to obtain estimates of the partial canonical correlations. For example, to estimate the lag $h = 3$ partial canonical correlations, $\{\widehat{\lambda}_1^2(3), \widehat{\lambda}_2^2(3), \widehat{\lambda}_3^2(3)\}$, we would put

$$\widehat{\Gamma}_{22} = \begin{bmatrix} \widehat{\Gamma}(0) & \widehat{\Gamma}(1) \\ \widehat{\Gamma}(1)' & \widehat{\Gamma}(0) \end{bmatrix}, \tag{5.101}$$

which represents, in this case, a 6×6 matrix of the estimated autocovariances that were displayed in (5.100). In addition, we will need the matrices

$$\widehat{\Gamma}_1^{(2)} = \left[\widehat{\Gamma}(2)', \widehat{\Gamma}(1)'\right] \quad \text{and} \quad \widehat{\Gamma}_2^{(1)} = \left[\widehat{\Gamma}(1), \widehat{\Gamma}(2)\right],$$

which are both, in this example, 3×6 matrices. From these matrices, we construct the 3×3 matrices

$$\widehat{\Sigma}_{00|21} = \widehat{\Gamma}(0) - \widehat{\Gamma}_1^{(2)}\widehat{\Gamma}_{22}^{-1}\widehat{\Gamma}_1^{(2)'},$$

$$\widehat{\Sigma}_{33|21} = \widehat{\Gamma}(0) - \widehat{\Gamma}_2^{(1)}\widehat{\Gamma}_{22}^{-1}\widehat{\Gamma}_2^{(1)'},$$

and

$$\widehat{\Sigma}_{30|21} = \widehat{\Gamma}(2) - \widehat{\Gamma}_2^{(1)}\widehat{\Gamma}_{22}^{-1}\widehat{\Gamma}_1^{(2)'} = \widehat{\Sigma}_{03|21}'.$$

Finally, the squared partial canonical correlations, $\lambda_j^2(3)$, for $j = 1, 2, 3$, are obtained as the ordered eigenvalues of $\widehat{\Sigma}_{00|21}^{-1}\widehat{\Sigma}_{03|21}\widehat{\Sigma}_{33|21}^{-1}\widehat{\Sigma}_{33|21}$.

In this example we obtain

$$\left(\widehat{\lambda}_1^2(h), \widehat{\lambda}_2^2(h), \widehat{\lambda}_3^2(h)\right) = \begin{cases} (.81, .24, .02) & h = 1 \\ (.22, .14, .06) & h = 2 \\ (.05, .01, .00) & h = 3 \\ (.05, .02, .00) & h = 4, \end{cases}$$

which also suggests an AR(2) model for the data.

In addition, successive Yule–Walker estimates, for $h = 1, 2, ...,$ of the error variance–covariance matrix can be obtained from (5.93), that is,

$$\widehat{\Sigma}_w^{(h)} = \widehat{\Gamma}(0) - \sum_{j=1}^{h} \widehat{\Phi}_{jh}\widehat{\Gamma}(-j). \tag{5.102}$$

For this data, we obtained (entries are rounded to integers)

$$\widehat{\Sigma}_w^{(1)} = \begin{bmatrix} 31 & 6 & 17 \\ 6 & 41 & 42 \\ 17 & 42 & 144 \end{bmatrix}, \quad \widehat{\Sigma}_w^{(2)} = \begin{bmatrix} 28 & 7 & 16 \\ 7 & 37 & 40 \\ 16 & 40 & 123 \end{bmatrix},$$

$$\widehat{\Sigma}_w^{(3)} = \begin{bmatrix} 28 & 7 & 16 \\ 7 & 37 & 40 \\ 16 & 40 & 118 \end{bmatrix}, \quad \widehat{\Sigma}_w^{(4)} = \begin{bmatrix} 27 & 6 & 14 \\ 6 & 36 & 38 \\ 14 & 38 & 114 \end{bmatrix}.$$

The estimates stabilize (except for perhaps the variance of the particulate series) after $h = 2$, indicating the AR(3) and AR(4) fits do not improve much over the AR(2) fit. Recall the comparison of the autoregressions of order one to five using the SIC, as reported in Table 5.1 also indicated the AR(2) model.

At this point, we would settle on the AR(2) model estimated in Example 5.10 on the detrended data. We will write the estimated model as

$$\widehat{\boldsymbol{x}}_t = \widehat{\Phi}_1\boldsymbol{x}_{t-1} + \widehat{\Phi}_2 x_{t-2} + \widehat{\boldsymbol{w}}_t, \tag{5.103}$$

where $\widehat{\Phi}_1$ and $\widehat{\Phi}_2$ are given in Example 5.10. The estimate of Σ_w for this model is $\widehat{\Sigma}_w^{(2)}$, which is listed below (5.102). Residual analysis, performed on the residuals $\widehat{\boldsymbol{w}}_t = \widehat{\boldsymbol{x}}_t - \widehat{\Phi}_1\boldsymbol{x}_{t-1} - \widehat{\Phi}_2\boldsymbol{x}_{t-2}$, for t=3,...,508, suggests the model fits well. Individual residual analyses on the $\widehat{\boldsymbol{w}}_{ti}$, for $i = 1, 2, 3$, show, except for the particulate series, w_{t3}, the residuals are Gaussian white noise. For the particulate series, a small, but significant, amount of autocorrelation is still left in that series. In this case, we may wish to fit a higher order (higher than two) model to the particulate series only. In addition, we might be inclined to fit a reduced rank model, and we

will discuss this matter later. Inspection of the pairwise CCF between all residual series shows no obvious departures from independence.

Once the model has been estimated, estimated forecasts can be obtained. Analogous to the univariate case, the m-step-ahead forecast, $m = 1, 2, ...,$ in this example ($n = 508$), is obtained as follows:

$$\widehat{\boldsymbol{x}}_{n+m}^{n} = \widehat{\Phi}_1 \widehat{\boldsymbol{x}}_{n+m-1}^{n} + \widehat{\Phi}_2 \widehat{\boldsymbol{x}}_{n+m-2}^{n}, \tag{5.104}$$

where $\widehat{\boldsymbol{x}}_{t}^{n} = \boldsymbol{x}_t$ when $1 \leq t \leq n$. The mean square prediction error matrices can be calculated in a manner similar to the univariate case, (3.67). In the general case of vector ARMA or ARMAX models, forecasts and their mean square prediction errors can be obtained by using the state-space formulation of the model and the Kalman filter (see §6.6). Analogous to (3.67), the general form of the m-step-ahead mean square prediction error matrix is,

$$\begin{aligned} P_{n+m}^{n} &= E\left(\boldsymbol{x}_{n+m} - \boldsymbol{x}_{n+m}^{n}\right)\left(\boldsymbol{x}_{n+m} - \boldsymbol{x}_{n+m}^{n}\right)' && (5.105) \\ &= \Gamma(0) - \Gamma_{n}^{(m)} \Gamma_{nn}^{-1} \Gamma_{n}^{(m)'}, && (5.106) \end{aligned}$$

where $\Gamma_{n}^{(m)} = [\Gamma(m), \Gamma(m+1), ..., \Gamma(m+n-1)]$, is a $k \times nk$ matrix, and $\Gamma_{nn} = \{\Gamma(i-j)\}_{i,j=1}^{n}$, is an $nk \times nk$ symmetric matrix. Of course, P_{n+m}^{n} can be estimated by substituting $\widehat{\Gamma}(h)$ for $\Gamma(h)$ in (5.106). The analogue of (3.77) for multivariate ARMA models is

$$P_{n+m}^{n} = \sum_{j=0}^{m-1} \Psi_j \Sigma_w \Psi_j'. \tag{5.107}$$

When the model is autoregressive, as in this example, a simplification occurs by noticing a k-dimensional AR(p) model can be written as a kp-dimensional AR(1) model. For example, we can write the vector AR(2) model as

$$\boldsymbol{X}_t = \boldsymbol{\alpha} + A(\boldsymbol{X}_{t-1} - \boldsymbol{\alpha}) + \boldsymbol{\eta}_t, \tag{5.108}$$

where

$$\boldsymbol{X}_t = \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix} \quad \boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix} \quad A = \begin{bmatrix} \Phi_1 & \Phi_2 \\ I & 0 \end{bmatrix} \quad \boldsymbol{\eta}_t = \begin{bmatrix} \boldsymbol{w}_t \\ \boldsymbol{0} \end{bmatrix}.$$

Of course, this technique generalizes to any dimension k and any order p. From (5.108) we immediately obtain the forecasts and mean square prediction errors as

$$\boldsymbol{X}_{n+m}^{n} = \boldsymbol{\alpha} + A^m(\boldsymbol{X}_n - \boldsymbol{\alpha})$$

and

$$\begin{aligned} Q_{n+m}^{n} &= E\left(\boldsymbol{X}_{n+m} - \boldsymbol{X}_{n+m}^{n}\right)\left(\boldsymbol{X}_{n+m} - \boldsymbol{X}_{n+m}^{n}\right)' \\ &= \Gamma_X(0) - A^m \Gamma_X(0) A'^m, \end{aligned}$$

where

$$\Gamma_X(0) = \begin{bmatrix} \Gamma(0) & \Gamma(1) \\ \Gamma(1)' & \Gamma(0) \end{bmatrix}.$$

We can then obtain the desired mean square prediction error matrices P_{n+m}^n as submatrices of Q_{n+m}^n. In addition, Yule–Walker estimation and forecasting can be accomplished by substituting autocovariance matrices by their sample equivalents obtained via (5.98).

For this numerical example,

$$\widehat{A} = \begin{bmatrix} \widehat{\Gamma}(1) & \widehat{\Gamma}(2) \\ \widehat{\Gamma}(0) & \widehat{\Gamma}(1) \end{bmatrix} \begin{bmatrix} \widehat{\Gamma}(0) & \widehat{\Gamma}(1) \\ \widehat{\Gamma}(1)' & \widehat{\Gamma}(0) \end{bmatrix}^{-1} = \begin{bmatrix} \widehat{\Phi}_1 & \widehat{\Phi}_2 \\ I & 0 \end{bmatrix},$$

where $\widehat{\Phi}_1$ and $\widehat{\Phi}_2$ are as given in Example 5.10. In the general case, we obtain the coefficient estimates from the top k rows of $\widehat{A}$. Similarly, estimated forecasts in this example are found as follows:

$$\begin{bmatrix} \widehat{\boldsymbol{x}}_{n+m}^n \\ \widehat{\boldsymbol{x}}_{n+m-1}^n \end{bmatrix} = \widehat{A}^m \begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{x}_{n-1} \end{bmatrix}.$$

Because $\boldsymbol{x}_{507} = (8.62, -1.85, 12.16)'$ and $\boldsymbol{x}_{508} = (4.71, -4.67, 17.20)'$, we can, for example, calculate the one-step-ahead and two-step-ahead forecasts by putting $m = 2$ and using the numerical values given in Example 5.10 to construct $\widehat{A}^2$,

$$\begin{bmatrix} \widehat{\boldsymbol{x}}_{510}^{508} \\ \widehat{\boldsymbol{x}}_{509}^{508} \end{bmatrix} = \widehat{A}^2 \begin{bmatrix} \widehat{\boldsymbol{x}}_{508} \\ \widehat{\boldsymbol{x}}_{507} \end{bmatrix} = \begin{bmatrix} 6.13 \\ -5.94 \\ 11.23 \\ \\ 6.43 \\ -4.77 \\ 10.53 \end{bmatrix}.$$

Substituting autocovariance matrices with their estimates, we may write

$$\begin{aligned} \widehat{Q}_{510}^{508} &= \begin{bmatrix} \widehat{\Gamma}(0) & \widehat{\Gamma}(1) \\ \widehat{\Gamma}(1)' & \widehat{\Gamma}(0) \end{bmatrix} - \widehat{A}^2 \begin{bmatrix} \widehat{\Gamma}(0) & \widehat{\Gamma}(1) \\ \widehat{\Gamma}(1)' & \widehat{\Gamma}(0) \end{bmatrix} \widehat{A}'^2 \\ &= \begin{bmatrix} \widehat{P}_{510}^{508} & \widehat{P}_{510,509}^{508} \\ \\ \widehat{P}_{509,510}^{508} & \widehat{P}_{509}^{508} \end{bmatrix}, \end{aligned}$$

where we have written $\widehat{P}_{s,t}^n$ to be the estimate of $E\{(\boldsymbol{x}_s - \boldsymbol{x}_s^n)(\boldsymbol{x}_t - \boldsymbol{x}_t^n)'\}$.

In this example, we found (entries are rounded)

$$\widehat{Q}_{510}^{508} = \begin{bmatrix} 31 & 5 & 19 & 8 & -4 & 2 \\ 5 & 39 & 38 & -2 & 7 & 2 \\ 19 & 38 & 135 & 6 & 2 & 33 \\ 8 & -2 & 6 & 28 & 7 & 16 \\ -4 & 7 & 2 & 7 & 37 & 40 \\ 2 & 2 & 33 & 16 & 40 & 123 \end{bmatrix}.$$

Note, $\widehat{P}_{509}^{508} = \widehat{\Sigma}_w = \widehat{\Sigma}_w^{(2)}$. The diagonal elements of $\widehat{Q}_{510}^{508}$ give the individual mean-square prediction errors. For example, an approximate 95% prediction interval for $x_{510,1}^{508}$ is $6.13 \pm 2\sqrt{31}$ or $(-5.0, 17.2)$.

Although the estimation in Example 5.11 was performed using Yule–Walker estimation, we could have also used conditional or unconditional maximum likelihood estimation, or conditional (as in Example 5.10) or unconditional least squares estimation. Because, as we have seen, any k-dimensional AR(p) model can be written as a kp-dimensional AR(1) model, any of these estimation techniques are straightforward multivariate extensions to the univariate case presented in equations (2.124)-(2.133). Also, as in the univariate case, the Yule–Walker estimators, the maximum likelihood estimators, and the least squares estimators are asymptotically equivalent. To exhibit the asymptotic distribution of the autoregression parameter estimators, we write

$$\boldsymbol{\phi} = \text{vec}\,(\Phi_1, ..., \Phi_p)\,,$$

where the vec operator stacks the columns of a matrix into a vector. For example, for a bivariate AR(2) model,

$$\boldsymbol{\phi} = \text{vec}\,(\Phi_1, \Phi_2) = (\Phi_{1_{11}}, \Phi_{1_{21}}, \Phi_{1_{12}}, \Phi_{1_{22}} \Phi_{2_{11}}, \Phi_{2_{21}}, \Phi_{2_{12}}, \Phi_{2_{22}})'\,,$$

where $\Phi_{\ell_{ij}}$ is the ij-th element of Φ_ℓ, $\ell = 1, 2$. Because $(\Phi_1, ..., \Phi_p)$ is a $k \times kp$ matrix, $\boldsymbol{\phi}$ is a $k^2p \times 1$ vector. We now state the following property.

Property P5.1: Large Sample Distribution of the Vector Autoregression Estimators
Let $\widehat{\boldsymbol{\phi}}$ denote the vector of parameter estimators (obtained via Yule–Walker, least squares, or maximum likelihood) for a k-dimensional AR(p) model. Then,

$$\sqrt{n}\left(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}\right) \sim \text{AN}(\mathbf{0}, \Sigma_w \otimes \Gamma_{pp}^{-1}), \tag{5.109}$$

where $\Gamma_{pp} = \{\Gamma(i-j)\}_{i,j=1}^{p}$ is a $kp \times kp$ matrix, $\Sigma_w \otimes \Gamma_{pp}^{-1} = \{\sigma_{ij}\Gamma_{pp}^{-1}\}_{i,j=1}^{k}$ is a $k^2p \times k^2p$ matrix, and σ_{ij} is the ij-th element of Σ_w.

The variance–covariance matrix of the estimator $\widehat{\boldsymbol{\phi}}$ is approximated by replacing Σ_w by $\widehat{\Sigma}_w$, and replacing $\Gamma(h)$ by $\widehat{\Gamma}(h)$ in Γ_{pp}. The square root of the

diagonal elements of $\widehat{\Sigma}_w \otimes \widehat{\Gamma}_{pp}^{-1}$ divided by $\sqrt{n}$ gives the individual standard errors. For the mortality data example, the estimated standard errors for the VAR(2) fit are listed in Example 5.10; although those standard errors were taken from a regression run, they could have also been calculated using Property P5.1 along with the numerical values taken from $\widehat{\Sigma}_w^{(2)}$ given below (5.102) and $\widehat{\Gamma}_{22}$ given in (5.101).

Asymptotic inference for the general case of vector ARMA models is more complicated than pure AR models; details can be found in Reinsel (1997) or Lütkepohl (1993), for example. We also note that estimation for VARMA models can be recast into the problem of estimation for state-space models that will be discussed in Chapter 6.

A simple algorithm for fitting multivariate ARMA models from Spliid (1983) is worth mentioning because it repeatedly uses the multivariate regression equations. Consider a general ARMA(p, q) model for a time series with a nonzero mean

$$\boldsymbol{x}_t = \boldsymbol{\alpha} + \Phi_1 \boldsymbol{x}_{t-1} + \cdots + \Phi_p \boldsymbol{x}_{t-p} + \boldsymbol{w}_t + \Theta_1 \boldsymbol{w}_{t-1} + \cdots + \Theta_q \boldsymbol{w}_{t-q}. \tag{5.110}$$

If $\boldsymbol{\mu} = E\boldsymbol{x}_t$, then $\boldsymbol{\alpha} = (I - \Phi_1 - \cdots - \Phi_p)\boldsymbol{\mu}$. If $\boldsymbol{w}_{t-1}, ..., \boldsymbol{w}_{t-q}$ were observed, we could rearrange (5.110) as a multivariate regression model

$$\boldsymbol{x}_t = \mathcal{B}\boldsymbol{z}_t + \boldsymbol{w}_t, \tag{5.111}$$

with

$$\boldsymbol{z}_t = (1, \boldsymbol{x}'_{t-1}, ..., \boldsymbol{x}'_{t-p}, \boldsymbol{w}'_{t-1}, ..., \boldsymbol{w}'_{t-q})' \tag{5.112}$$

and

$$\mathcal{B} = [\boldsymbol{\alpha}, \Phi_1, ..., \Phi_p, \Theta_1, ..., \Theta_q], \tag{5.113}$$

for $t = p + 1, ..., n$. Given an initial estimator $\mathcal{B}_0$, of $\mathcal{B}$, we can reconstruct $\{\boldsymbol{w}_{t-1}, ..., \boldsymbol{w}_{t-q}\}$ by setting

$$\boldsymbol{w}_{t-j} = \boldsymbol{x}_{t-j} - \mathcal{B}_0 \boldsymbol{z}_{t-j}, \quad t = p + 1, ..., n, \quad j = 1, ..., q, \tag{5.114}$$

where, if $q > p$, we put $\boldsymbol{w}_{t-j} = \boldsymbol{0}$ for $t-j \leq 0$. The new values of $\{\boldsymbol{w}_{t-1}, ..., \boldsymbol{w}_{t-q}\}$ are then put into the regressors $\boldsymbol{z}_t$ and a new estimate, say, $\mathcal{B}_1$, is obtained. The initial value, $\mathcal{B}_0$, can be computed by fitting a pure autoregression of order p or higher, and taking $\Theta_1 = \cdots = \Theta_q = \boldsymbol{0}$. The procedure is then iterated until the parameter estimates stabilize. The algorithm usually converges, but not to the maximum likelihood estimators. Experience suggests the estimators are reasonably close to the maximum likelihood estimators.

As previously discussed, the special form assumed for the constant component, $\boldsymbol{\alpha}$, of the general ARMA model in (5.110) can be generalized to include a fixed $r \times 1$ vector of inputs, say, $\boldsymbol{u}_t$. In this case we have a k-dimensional ARMAX model:

$$\boldsymbol{x}_t = \Gamma \boldsymbol{u}_t + \sum_{j=1}^{p} \Phi_j \boldsymbol{x}_{t-j} + \sum_{j=1}^{q} \Theta_j \boldsymbol{w}_{t-j} + \boldsymbol{w}_t, \tag{5.115}$$

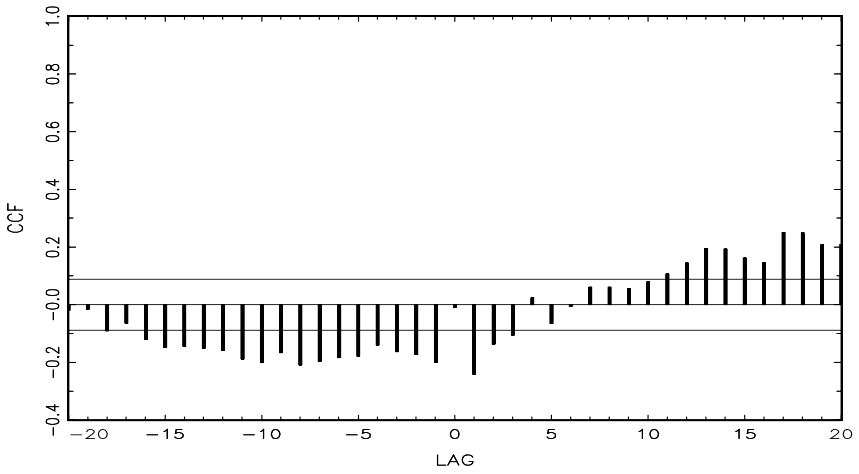


Figure 5.14 CCF between prewhitened mortality and temperature (positive lag means temperature leads mortality).

where Γ is a $k \times r$ parameter matrix. Recall the X in ARMAX refers to the exogenous vector process we have denoted here by $\boldsymbol{u}_t$ and the introduction of exogenous variables through setting $\boldsymbol{\alpha} = \Gamma \boldsymbol{u}_t$ does not present any special problems in making inferences.

Example 5.12 An ARMAX Model for Cardiovascular Mortality

In Example 2.2, we regressed the cardiovascular mortality series, M_t, on time t, temperature T_t, and particulate pollution P_t. There, the interest was an analysis of the effect of temperature and pollution on cardiovascular mortality. In Example 5.10, we fit a multivariate ARMA model to the trivariate vector (M_t, T_t, P_t), as if modeling the behavior of temperature and pollution was equally as important as modeling the behavior of mortality. In this example, we are interested in using temperature and pollution to explain some of the variation in the mortality series.

To examine the CCF between mortality and temperature, and between mortality and pollution, we first prewhitened mortality by fitting an AR(2) to the detrended data. That is, we first fit the model

$$M_t = \beta_0 + \beta_1 t + \phi_1 M_{t-1} + \phi_2 M_{t-2} + \epsilon_t.$$

Using the residuals of the fit, say, $\widehat{\epsilon}_t$, we then calculated the CCF between $\widehat{\epsilon}_t$ and T_t, and between $\widehat{\epsilon}_t$ and P_t. Figure 5.14 shows the cross-correlation of prewhitened mortality and temperature (positive lag means temperature leads mortality) and a significant correlation is seen at lag $h = 1$.

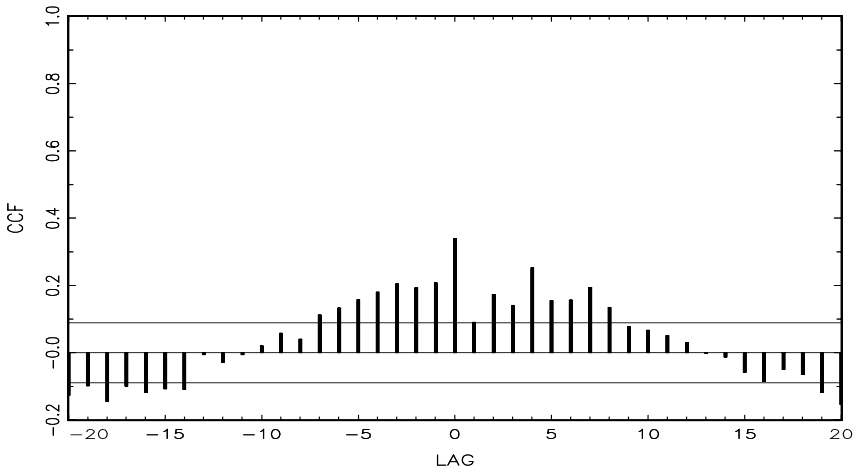


Figure 5.15 CCF between prewhitened mortality and particulate pollution (positive lag means pollution leads mortality).

Figure 5.15 shows a similar plot for the CCF of prewhitened mortality and pollution, and significant correlations are seen at lags $h = 0, 2, 4, 7$. After some preliminary fitting, the final model uses the exogenous variables $\boldsymbol{u}_t = (1, t, T_{t-1}, T_{t-1}^2, P_t, P_{t-4})'$, along with an AR(2) on mortality, M_t; the inclusion of particulate pollution at lags two and seven were not significant when lags zero and four are in the model. In this case, the ARMAX model is

$$M_t = \Gamma \boldsymbol{u}_t + \phi_1 M_{t-1} + \phi_2 M_{t-2} + w_t,$$

where $\Gamma = [\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4]$.

Estimation was accomplished using the regression approach described in (5.85) and (5.86). In this case, the fitted model was (values are rounded)

$$\begin{aligned} \widehat{M}_t &= 42.9 - .01_{(.002)}t - .18_{(.03)}T_{t-1} + .11_{(.02)}P_t + .05_{(.02)}P_{t-4} \\ &+ .31_{(.04)}M_{t-1} + .30_{(.04)}M_{t-2} + \widehat{w}_t, \end{aligned}$$

where $\widehat{\sigma}_w^2 = 25.7$ and $R^2 = 74.3\%$. Each coefficient is significant, as seen from the estimated standard errors listed below each parameter estimate. Finally, an analysis of the residuals, $\widehat{w}_t$, shows, except for a few outliers, the model fits well. The value of the Ljung–Box–Pierce statistic for H=24 was Q=25.7, which when compared to a χ^2_{22}, is not significant. In addition, a Q-Q plot shows no departure from the Gaussian assumption, except for the few outliers. Our general conclusions are that decrease in cardiovascular mortality occurred during the period studied, and an

increase in mortality is associated with lower temperatures the previous week and higher particulate pollution both currently and one month prior.

Problems

Section 5.2

5.1 The data set labeled `fracdiff.dat` is $n = 1000$ simulated observations from a fractionally differenced ARIMA$(1, 1, 0)$ model with $\phi = .75$ and $d = .4$.

(a) Plot of the data and comment.

(b) Plot the ACF and PACF of the data and comment.

(c) Estimate the parameters and test for the significance of the estimates $\widehat{\phi}$ and $\widehat{d}$.

(d) Explain why, using the results of part (a) and (b), it would seem reasonable to difference the data prior to the analysis. That is, if x_t represents the data, explain why we might choose to fit an ARMA model to ∇x_t.

(e) Plot the ACF and PACF of ∇x_t and comment.

(f) Fit an ARMA model to ∇x_t and comment.

5.2 The data in `globtemp2.dat` are annual global temperature deviations from 1880 to 2004 (there are three columns in the data file; work with the annual means and not the 5-year smoothed data). The data are an update to the Hansen–Lebedeff global temperature data displayed in Figure 1.2. The URL of the data source is in the file, you can go there for further explanation of the data. Fit an ARFIMA model to this series.

5.3 Compute the sample ACF of the absolute values of the NYSE returns displayed in Figure 1.4 up to lag 200 and comment on whether the ACF indicates long memory. Fit an ARFIMA model to the absolute values and comment.

Section 5.3

5.4 Investigate whether the monthly returns of a stock dividend yield listed in the file `sdyr.dat` exhibit GARCH behavior. If so, fit an appropriate model to the returns. The data are monthly returns of a stock dividend yield from January 1947 through May 1993 and are taken from Hamilton and Lin (1996).

5.5 Investigate whether the growth rate of the monthly Oil Prices exhibit GARCH behavior. If so, fit an appropriate model to the growth rate.

5.6 The `stats` package of R contains the daily closing prices of four major European stock indices; type `help(EuStockMarkets)` for details. Fit a GARCH model to the returns of these series and discuss your findings. (Note: The data set contains actual values, and not returns. Hence, the data must be transformed prior to the model fitting.)

5.7 The 2×1 gradient vector, $l^{(1)}(\alpha_0, \alpha_1)$, given for an ARCH(1) model was displayed in (5.41). Verify (5.41) and then use the result to calculate the 2×2 Hessian matrix

$$l^{(2)}(\alpha_0, \alpha_1) = \begin{pmatrix} \partial^2 l/\partial\alpha_0^2 & \partial^2 l/\partial\alpha_0\partial\alpha_1 \\ \partial^2 l/\partial\alpha_0\partial\alpha_1 & \partial^2 l/\partial\alpha_1^2 \end{pmatrix}.$$

Section 5.4

5.8 The sunspot data are plotted in Chapter 4, Figure 4.31. From a time plot of the data, discuss why is it reasonable to fit a threshold model to the data, and then fit a threshold model.

Section 5.5

5.9 Let S_t represent the monthly sales data listed in `sales.dat` ($n = 150$), and let L_t be the leading indicator listed in `lead.dat`. Fit the regression model $\nabla S_t = \beta_0 + \beta_1 \nabla L_{t-3} + x_t$, where x_t is an ARMA process.

5.10 Consider the correlated regression model, defined in the text by (5.53), say,

$$\boldsymbol{y} = Z\boldsymbol{\beta} + \boldsymbol{x},$$

where $\boldsymbol{x}$ has mean zero and covariance matrix Γ. In this case, we know that the weighted least squares estimator is (5.54), namely,

$$\widehat{\boldsymbol{\beta}}_w = (Z'\Gamma^{-1}Z)^{-1}Z'\Gamma^{-1}\boldsymbol{y}.$$

Now, a problem of interest in spatial series can be formulated in terms of this basic model. Suppose $y_i = y(\sigma_i), i = 1, 2, \ldots, n$ is a function of the spatial vector coordinates $\sigma_i = (s_{i1}, s_{i2}, \ldots, s_{ir})'$, the error is $x_i = x(\sigma_i)$, and the rows of Z are defined as $\boldsymbol{z}(\sigma_i)', i = 1, 2, \ldots, n$. The Kriging estimator is defined as the best spatial predictor of $y_0 = \boldsymbol{z}_0'\boldsymbol{\beta} + x_0$ using the estimator

$$\widehat{y}_0 = \boldsymbol{a}'\boldsymbol{y},$$

subject to the unbiased condition $E\widehat{y}_0 = Ey_0$, and such that the mean square prediction error

$$\text{MSE} = E[(y_0 - \widehat{y}_0)^2]$$

is minimized.

(a) Prove the estimator is unbiased when $Z'\boldsymbol{a} = \boldsymbol{z}_0$.

(b) Show the MSE is minimized by solving the equations

$$\Gamma\boldsymbol{a} + Z\boldsymbol{\lambda} = \boldsymbol{\gamma}_0$$

and

$$Z'\boldsymbol{a} = \boldsymbol{z}_0,$$

where $\boldsymbol{\gamma}_0 = E[\boldsymbol{x}x_0]$ represents the vector of covariances between the error vector of the observed data and the error of the new point the vector $\boldsymbol{\lambda}$ is a $q \times 1$ vector of LaGrangian multipliers.

(c) Show the predicted value can be expressed as

$$\widehat{y}_0 = \boldsymbol{z}_0'\widehat{\boldsymbol{\beta}}_w + \boldsymbol{\gamma}_0'\Gamma^{-1}(\boldsymbol{y} - Z\widehat{\boldsymbol{\beta}}_w),$$

so the optimal prediction is a linear combination of the usual predictor and the least squares residuals.

Section 5.6

5.11 The file labeled `clim-hyd` has 454 months of measured values for the climatic variables air temperature, dew point, cloud cover, wind speed, precipitation (p_t), and inflow (i_t), at Shasta Lake. We would like to look at possible relations between the weather factors and the inflow to Shasta Lake.

(a) Fit ARIMA$(0,0,0) \times (0,1,1)_{12}$ models to (i) transformed precipitation $P_t = \sqrt{p_t}$ and (ii) transformed inflow $I_t = \log i_t$.

(b) Apply the ARIMA model fitted in part (a) for transformed precipitation to the flow series to generate the prewhitened flow residuals assuming the precipitation model. Compute the cross-correlation between the flow residuals using the precipitation ARIMA model and the precipitation residuals using the precipitation model and interpret. Use the coefficients from the ARIMA model to construct the transformed flow residuals.

5.12 Consider predicting the transformed flows $I_t = \log i_t$ from transformed precipitation values $P_t = \sqrt{p_t}$ using a transfer function model of the form

$$(1 - B^{12})I_t = \alpha(B)(1 - B^{12})P_t + n_t,$$

where we assume that seasonal differencing is a reasonable thing to do. The data are the 454 monthly values of precipitation and inflow from the Shasta Lake reservoir in the file *clim-hyd*. You may think of it as fitting

$$y_t = \alpha(B)x_t + n_t,$$

where y_t and x_t are the seasonally differenced transformed flows and precipitations.

(a) Argue that x_t can be fitted by a first-order seasonal moving average, and use the transformation obtained to prewhiten the series x_t.

(b) Apply the transformation applied in (a) to the series y_t, and compute the cross-correlation function relating the prewhitened series to the transformed series. Argue for a transfer function of the form

$$\alpha(B) = \frac{\delta_0}{1 - \omega_1 B}.$$

(c) Write the overall model obtained in regression form to estimate δ_0 and ω_1. Note that you will be minimizing the sums of squared residuals for the transformed noise series $(1 - \widehat{\omega}_1 B)n_t$. Retain the residuals for further modeling involving the noise n_t. The observed residual is $u_t = (1 - \widehat{\omega}_1 B)n_t$.

(d) Fit the noise residuals obtained in (c) with an ARMA model, and give the final form suggested by your analysis in the previous parts.

(e) Discuss the problem of forecasting y_{t+m} using the infinite past of y_t and the present and infinite past of x_t. Determine the predicted value and the forecast variance.

Section 5.7

5.13 Consider the data set containing quarterly U.S. unemployment, U.S. GNP, consumption, and government and private investment from 1948-III to 1988-II. The seasonal component has been removed from the data. Concentrating on unemployment (U_t), GNP (G_t), and consumption (C_t), fit a vector ARMA model to the data after first logging each series, and then removing the linear trend. That is, fit a vector ARMA model to $\boldsymbol{x}_t = (x_{1t}, x_{2t}, x_{3t})'$, where, for example, $x_{1t} = \log(U_t) - \widehat{\beta}_0 - \widehat{\beta}_1 t$, where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the least squares estimates for the regression of $\log(U_t)$ on time, t. Run a complete set of diagnostics on the residuals.

Chapter 6

State-Space Models

6.1 Introduction

A very general model that seems to subsume a whole class of special cases of interest in much the same way that linear regression does is the state-space model or the dynamic linear model, which was introduced in Kalman (1960) and Kalman and Bucy (1961). Although the model was originally introduced as a method primarily for use in aerospace-related research, it has been applied to modeling data from economics (Harrison and Stevens, 1976; Harvey and Pierse, 1984; Harvey and Todd, 1983; Kitagawa and Gersch 1984, Shumway and Stoffer, 1982), medicine (Jones, 1984) and the soil sciences (Shumway, 1985). An excellent modern treatment of time series analysis based on the state space model is the text by Durbin and Koopman (2001).

Although there are some packages available for R that focus on various aspects of state-space modeling and Kalman filtering, we prefer to write our own code. As a result, the code we have written is long and will most likely be subject to frequent updates. Hence, we have decided to distribute the R code for this chapter on the website for the text.

The state-space model or dynamic linear model (DLM), in its basic form, employs an order one, vector autoregression as the state equation,

$$\boldsymbol{x}_t = \Phi \boldsymbol{x}_{t-1} + \boldsymbol{w}_t, \tag{6.1}$$

where the state equation determines the rule for the generation of the $p \times 1$ state vector $\boldsymbol{x}_t$ from the past $p \times 1$ state $\boldsymbol{x}_{t-1}$, for time points $t = 1, \ldots, n$. We assume the $\boldsymbol{w}_t$ are $p \times 1$ independent and identically distributed, zero-mean normal vectors with covariance matrix Q. In the DLM, we assume the process starts with a normal vector $\boldsymbol{x}_0$ that has mean $\boldsymbol{\mu}_0$ and $p \times p$ covariance matrix Σ_0.

The DLM, however, adds an additional component to the model in assuming we do not observe the state vector $\boldsymbol{x}_t$ directly, but only a linear transformed

version of it with noise added, say

$$\boldsymbol{y}_t = A_t \boldsymbol{x}_t + \boldsymbol{v}_t \tag{6.2}$$

where A_t is a $q \times p$ measurement or observation matrix; equation (6.2) is called the observation equation. The model arose originally in the space tracking setting, where the state equation defines the motion equations for the position or state of a spacecraft with location $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ reflects information that can be observed from a tracking device such as velocity and azimuth. The observed data are in the $q \times 1$ vectors $\boldsymbol{y}_t$, which can be larger than or smaller than p, the dimension of the underlying series of interest. The additive observation noise $\boldsymbol{v}_t$ is assumed to be white and Gaussian with $q \times q$ covariance matrix R. In addition, we initially assume, for simplicity, $\{\boldsymbol{w}_t\}$ and $\{\boldsymbol{v}_t\}$ are uncorrelated; this assumption is not necessary, but it helps in the explanation of first concepts. The case of correlated errors is discussed in §6.6. Of course, we can further modify the basic model, (6.1) and (6.2), to include exogenous variables, and we will also discuss this in §6.6.

As in the ARMAX model of §5.7, exogenous variables, or fixed inputs, may enter into the states or into the observations. In this case, we suppose we have an $r \times 1$ vector of inputs $\boldsymbol{u}_t$, and write the model as

$$\boldsymbol{x}_t = \Phi \boldsymbol{x}_{t-1} + \Upsilon \boldsymbol{u}_t + \boldsymbol{w}_t \tag{6.3}$$

$$\boldsymbol{y}_t = A_t \boldsymbol{x}_t + \Gamma \boldsymbol{u}_t + \boldsymbol{v}_t \tag{6.4}$$

where Υ is $p \times r$ and Γ is $q \times r$.

Example 6.1 A Biomedical Example

Suppose we consider the problem of monitoring the level of several biomedical parameters monitored after a cancer patient undergoes a bone marrow transplant. The data in Figure 6.1, used by Jones (1984), are measurements made for 91 days on the three variables, log(white blood count), log(platelet), and hematocrit (HCT), denoted y_{t1}, y_{t2}, and y_{t3}, respectively. Approximately 40% of the values are missing, with missing values occurring primarily after the 35th day. The main objectives are to model the three variables using the state-space approach, and to estimate the missing values. According to Jones, "Platelet count at about 100 days post transplant has previously been shown to be a good indicator of subsequent long term survival." For this particular situation, we model the three variables in terms of the state equation (6.1); that is,

$$\begin{pmatrix} x_{t1} \\ x_{t2} \\ x_{t3} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{pmatrix} \begin{pmatrix} x_{t-1,1} \\ x_{t-1,2} \\ x_{t-1,3} \end{pmatrix} + \begin{pmatrix} w_{t1} \\ w_{t2} \\ w_{t3} \end{pmatrix}. \tag{6.5}$$

The 3×3 observation matrix, A_t, is either the identity matrix, or the identity matrix with all zeros in a row when that variable is missing. The covariance matrices R and Q are 3×3 matrices with $R = \text{diag}\{r_{11}, r_{22}, r_{33}\}$, a diagonal matrix, required for a simple approach when data are missing.

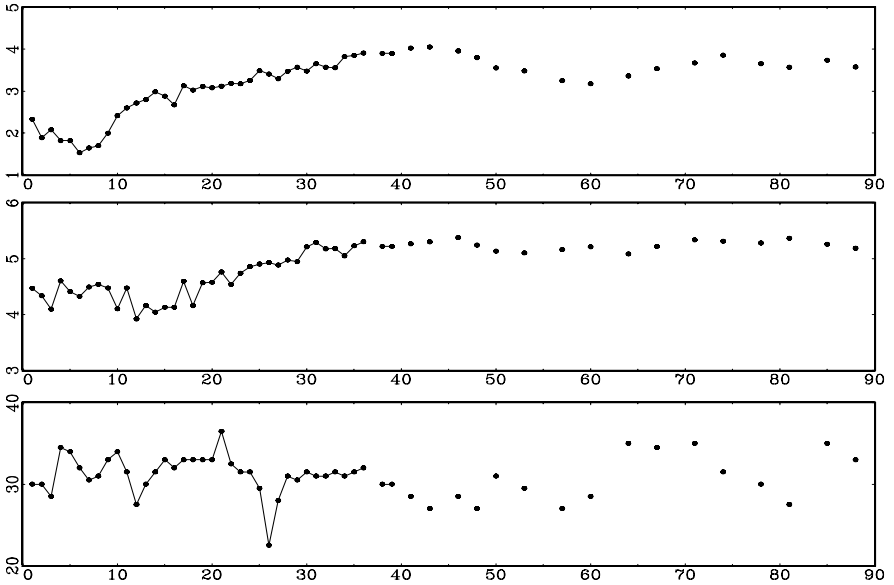


Figure 6.1 Longitudinal series of blood parameter levels monitored, log (white blood count) [top], log (platelet) [middle], and hematocrit (HCT) [bottom], after a bone marrow transplant ($n = 91$ days).

The model given in (6.1) involving only a single lag is not unduly restrictive. A multivariate model with m lags, such as the VAR(m) discussed in §5.7, could be developed by replacing the $p \times 1$ state vector, $\boldsymbol{x}_t$, by the $pm \times 1$ state vector $\boldsymbol{X}_t = (\boldsymbol{x}_t', \boldsymbol{x}_{t-1}', \ldots, \boldsymbol{x}_{t-m+1}')'$ and the transition matrix by

$$\Phi = \begin{pmatrix} \Phi_1 & \Phi_2 & \ldots & \Phi_{m-1} & \Phi_m \\ I & 0 & \ldots & 0 & 0 \\ 0 & I & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & I & 0 \end{pmatrix}. \tag{6.6}$$

Letting $\boldsymbol{W}_t = (\boldsymbol{w}_t', \mathbf{0}', \ldots, \mathbf{0}')'$ be the new $pm \times 1$ state error vector, the new state equation will be

$$\boldsymbol{X}_t = \Phi \boldsymbol{X}_{t-1} + \boldsymbol{W}_t, \tag{6.7}$$

where the new matrix "Q" now has the form of a $pm \times pm$ matrix with Q in the upper right-hand corner and zeros elsewhere. The observation equation can then be written as

$$\boldsymbol{y}_t = \left[A_t \mid 0 \mid \cdots \mid 0\right] \boldsymbol{X}_t + \boldsymbol{v}_t. \tag{6.8}$$

This simple recoding shows one way of handling higher order lags within the

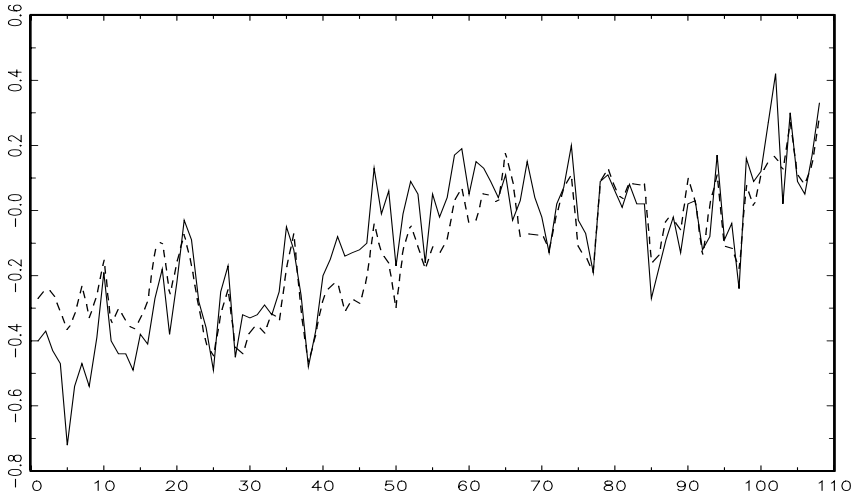


Figure 6.2 Two average global temperature deviations for $n = 108$ years in degrees centigrade (1880-1987). The solid line is the land-based series whereas the dotted line shows the marine-based series.

context of the single lag structure. Further discussion of this notion is given in §6.6.

The real advantages of the state-space formulation, however, do not really come through in the simple example given above. The special forms that can be developed for various versions of the matrix A_t and for the transition scheme defined by the matrix Φ allow fitting more parsimonious structures with fewer parameters needed to describe a multivariate time series. We will give some examples of structural models in §6.5, but the simple example shown below is instructive.

Example 6.2 Global Warming

Figure 6.2 shows two different estimators for the global temperature series from 1880 to 1987, plotted on the same scale. The solid line is considered in the first chapter (see Jones, 1994), which gives average surface air temperature computed from land-based observation stations. The second series (see Parker et al., 1996) gives averages from a number of marine-based stations. Conceptually, both series should be measuring the same underlying climatic signal, and we may consider the problem of extracting this underlying signal. We suppose both series are observing the same signal with different noises; that is,

$$y_{t1} = x_t + v_{t1}$$

and

$$y_{t2} = x_t + v_{t2},$$

where x_t is the unknown common signal. Suppose it can be modeled as a random walk of the form

$$x_t = x_{t-1} + w_t, \tag{6.9}$$

which we can argue for by noting the stability of the first difference as has been noted in Chapter 2. Furthermore, the first difference of the observed series will be a first-order moving average under this model, arguing from the fact that the first difference of the land-based series has a peak at lag 1. In this example, $p = 1, q = 2$, and $\Phi = 1$ is held at a constant value. The observation equation (6.2) can be written in the form

$$\begin{pmatrix} y_{t1} \\ y_{t2} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} x_t + \begin{pmatrix} v_{t1} \\ v_{t2} \end{pmatrix}, \tag{6.10}$$

and we have the covariance matrices given by $Q = q_{11}$ and

$$R = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}.$$

The introduction of the state-space approach as a tool for modeling data in the social and biological sciences requires model identification and parameter estimation because there is rarely a well-defined differential equation describing the state transition. The questions of general interest for the dynamic linear model (6.3) and (6.4) relate to estimating the unknown parameters contained in $\Phi, \Upsilon, Q, \Gamma, A_t$, and R, that define the particular model, and estimating or forecasting values of the underlying unobserved process $\boldsymbol{x}_t$. The advantages of the state-space formulation are in the ease with which we can treat various missing data configurations and in the incredible array of models that can be generated from (6.1) and (6.2). The analogy between the observation matrix A_t and the design matrix in the usual regression and analysis of variance setting is a useful one. We can generate fixed and random effect structures that are either constant or vary over time simply by making appropriate choices for the matrix A_t and the transition structure Φ. We will give only a few examples in this chapter; for further examples, see Durbin and Koopman (2001), Harvey (1993) or Shumway (1988) to mention a few.

Before continuing our investigation of the more complex model, it is instructive to consider a simple univariate state-space model.

Example 6.3 An AR(1) Process with Observational Noise

Consider a univariate state-space model where the observations satisfy

$$y_t = x_t + v_t, \tag{6.11}$$

and the signal (state) is an AR(1) process,

$$x_t = \phi x_{t-1} + w_t, \tag{6.12}$$

for $t = 1, 2, \ldots, n$, where $v_t \sim$ iid $\mathrm{N}(0, \sigma_v^2)$, $w_t \sim$ iid $\mathrm{N}(0, \sigma_w^2)$, and $x_0 \sim \mathrm{N}\big(0, \sigma_w^2/(1-\phi^2)\big)$; $\{v_t\}, \{w_t\}, x_0$ are independent.

In Chapter 3, we investigated the properties of the state, x_t, because it is a stationary AR(1) process (recall Problem 3.2e). For example, we know the autocovariance function of x_t is

$$\gamma_x(h) = \frac{\sigma_w^2}{1-\phi^2}\ \phi^h, \quad h = 0, 1, 2, \ldots\ . \tag{6.13}$$

But, here, we must investigate how the addition of observation noise affects the dynamics. Although it is not a necessary assumption, we have assumed in this example that x_t is stationary. In this case, the observations are also stationary because y_t is the sum of two independent stationary components x_t and v_t. We have

$$\gamma_y(0) = \mathrm{var}(y_t) = \mathrm{var}(x_t + v_t) = \frac{\sigma_w^2}{1-\phi^2} + \sigma_v^2, \tag{6.14}$$

and, when $h \geq 1$,

$$\gamma_y(h) = \mathrm{cov}(y_t, y_{t-h}) = \mathrm{cov}(x_t + v_t, x_{t-h} + v_{t-h}) = \gamma_x(h). \tag{6.15}$$

Consequently, for $h \geq 1$, the ACF of the observations is

$$\rho_y(h) = \frac{\gamma_y(h)}{\gamma_y(0)} = \left(1 + \frac{\sigma_v^2}{\sigma_w^2}(1-\phi^2)\right)^{-1} \phi^h. \tag{6.16}$$

It should be clear from the correlation structure given by (6.16) the observations, y_t, are not AR(1) unless $\sigma_v^2 = 0$. In addition, the autocorrelation structure of y_t is identical to the autocorrelation structure of an ARMA(1,1) process, as presented in Example 3.11. Thus, the observations can also be written in an ARMA(1,1) form,

$$y_t = \phi y_{t-1} + \theta u_{t-1} + u_t,$$

where u_t is Gaussian white noise with variance σ_u^2, and with θ and σ_u^2 suitably chosen (see Example 6.11).

Although an equivalence exists between stationary ARMA models and stationary state-space models (see §6.6), it is sometimes easier to work with one form than another. As previously mentioned, in the case of missing data, complex multivariate systems, mixed effects, and certain types of nonstationarity, it is easier to work in the framework of state-space models; in this chapter, we explore some of these situations.

6.2 Filtering, Smoothing, and Forecasting

From a practical view, the primary aims of any analysis involving the state-space model as defined by (6.1)-(6.2), or by (6.3)-(6.4), would be to produce estimators for the underlying unobserved signal $\boldsymbol{x}_t$, given the data $Y_s = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s\}$, to time s. When $s < t$, the problem is called forecasting or prediction. When $s = t$, the problem is called filtering, and when $s > t$, the problem is called smoothing. In addition to these estimates, we would also want to measure their precision. The solution to these problems is accomplished via the Kalman filter and smoother and is the focus of this section.

Throughout this chapter, we will use the following definitions:

$$\boldsymbol{x}_t^s = E(\boldsymbol{x}_t \mid Y_s) \tag{6.17}$$

and

$$P_{t_1,t_2}^s = E\left\{(\boldsymbol{x}_{t_1} - \boldsymbol{x}_{t_1}^s)(\boldsymbol{x}_{t_2} - \boldsymbol{x}_{t_2}^s)'\right\}. \tag{6.18}$$

When $t_1 = t_2$ ($= t$ say) in (6.18), we will write P_t^s for convenience.

In obtaining the filtering and smoothing equations, we will rely heavily on the Gaussian assumption. Some knowledge of the material covered in Appendix B, §B.1, will be helpful in understanding the details of this section (although these details may be skipped on a casual reading of the material). Even in the non-Gaussian case, the estimators we obtain are the minimum mean-squared error estimators within the class of linear estimators. That is, we can think of E in (6.17) as the projection operator in the sense of §B.1 rather than expectation and P_t^s as the corresponding mean-squared error. When we assume, as in this section, the processes are Gaussian, (6.18) is also the conditional error covariance; that is,

$$P_{t_1,t_2}^s = E\left\{(\boldsymbol{x}_{t_1} - \boldsymbol{x}_{t_1}^s)(\boldsymbol{x}_{t_2} - \boldsymbol{x}_{t_2}^s)' \mid Y_s\right\}.$$

This fact can be seen, for example, by noting the covariance matrix between $(\boldsymbol{x}_t - \boldsymbol{x}_t^s)$ and Y_s, for any t and s, is zero; we could say they are orthogonal in the sense of §B.1. This result implies that $(\boldsymbol{x}_t - \boldsymbol{x}_t^s)$ and Y_s are independent (because of the normality), and hence, the conditional distribution of $(\boldsymbol{x}_t - \boldsymbol{x}_t^s)$ given Y_s is the unconditional distribution of $(\boldsymbol{x}_t - \boldsymbol{x}_t^s)$. Derivations of the filtering and smoothing equations from a Bayesian perspective are given in Meinhold and Singpurwalla (1983); more traditional approaches based on the concept of projection and on multivariate normal distribution theory are given in Jazwinski (1970) and Anderson and Moore (1979).

First, we present the Kalman filter, which gives the filtering and forecasting equations. The name filter comes from the fact that $\boldsymbol{x}_t^t$ is a linear filter of the observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t$; that is, $\boldsymbol{x}_t^t = \sum_{s=1}^t B_s \boldsymbol{y}_s$ for suitably chosen $p \times q$ matrices B_s. The advantage of the Kalman filter is that it specifies how to update the filter from $\boldsymbol{x}_{t-1}^{t-1}$ to $\boldsymbol{x}_t^t$ once a new observation $\boldsymbol{y}_t$ is obtained, without having to reprocess the entire data set $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t$.

Property P6.1: The Kalman Filter
For the state-space model specified in (6.3) and (6.4), with initial conditions $\boldsymbol{x}_0^0 = \boldsymbol{\mu}_0$ *and* $P_0^0 = \Sigma_0$*, for* $t = 1, \ldots, n$*,*

$$\boldsymbol{x}_t^{t-1} = \Phi \boldsymbol{x}_{t-1}^{t-1} + \Upsilon \boldsymbol{u}_t, \tag{6.19}$$

$$P_t^{t-1} = \Phi P_{t-1}^{t-1} \Phi' + Q, \tag{6.20}$$

with

$$\boldsymbol{x}_t^t = \boldsymbol{x}_t^{t-1} + K_t(\boldsymbol{y}_t - A_t \boldsymbol{x}_t^{t-1} - \Gamma \boldsymbol{u}_t), \tag{6.21}$$

$$P_t^t = [I - K_t A_t] P_t^{t-1}, \tag{6.22}$$

where

$$K_t = P_t^{t-1} A_t' [A_t P_t^{t-1} A_t' + R]^{-1} \tag{6.23}$$

is called the Kalman gain. Prediction for $t > n$ *is accomplished via (6.19) and (6.20) with initial conditions* $\boldsymbol{x}_n^n$ *and* P_n^n*.*

Proof. The derivations of (6.19) and (6.20) follow from straight forward calculations, because from (6.3) we have

$$\boldsymbol{x}_t^{t-1} = E(\boldsymbol{x}_t \mid Y_{t-1}) = E(\Phi \boldsymbol{x}_{t-1} + \Upsilon \boldsymbol{u}_t + \boldsymbol{w}_t \mid Y_{t-1}) = \Phi x_{t-1}^{t-1} + \Upsilon \boldsymbol{u}_t,$$

and thus

$$\begin{aligned} P_t^{t-1} &= E\left\{(\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1})(\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1})'\right\} \\ &= E\left\{\left[\Phi(\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-1}^{t-1}) + \boldsymbol{w}_t\right]\left[\Phi(\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-1}^{t-1}) + \boldsymbol{w}_t\right]'\right\} \\ &= \Phi P_{t-1}^{t-1} \Phi' + Q. \end{aligned}$$

To derive (6.21), we first define the innovations as

$$\boldsymbol{\epsilon}_t = \boldsymbol{y}_t - E(\boldsymbol{y}_t \mid Y_{t-1}) = \boldsymbol{y}_t - A_t \boldsymbol{x}_t^{t-1} - \Gamma \boldsymbol{u}_t, \tag{6.24}$$

for $t = 1, \ldots, n$. Note, $E(\boldsymbol{\epsilon}_t) = \mathbf{0}$ and

$$\Sigma_t \stackrel{\text{def}}{=} \operatorname{var}(\boldsymbol{\epsilon}_t) = \operatorname{var}[A_t(\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1}) + \boldsymbol{v}_t] = A_t P_t^{t-1} A_t' + R \tag{6.25}$$

In addition, $E(\boldsymbol{\epsilon}_t \boldsymbol{y}_s') = 0$ for $s < t$, which in view of the fact the innovation sequence is a Gaussian process, implies that the innovations are independent of the past observations. Furthermore, the conditional covariance between $\boldsymbol{x}_t$ and $\boldsymbol{\epsilon}_t$ given Y_{t-1} is

$$\begin{aligned} \operatorname{cov}(\boldsymbol{x}_t,\ \boldsymbol{\epsilon}_t \mid Y_{t-1}) &= \operatorname{cov}(\boldsymbol{x}_t,\ \boldsymbol{y}_t - A_t \boldsymbol{x}_t^{t-1} - \Gamma \boldsymbol{u}_t \mid Y_{t-1}) \\ &= \operatorname{cov}(\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1},\ \boldsymbol{y}_t - A_t \boldsymbol{x}_t^{t-1} - \Gamma \boldsymbol{u}_t \mid Y_{t-1}) \\ &= \operatorname{cov}[\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1},\ A_t(\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1}) + \boldsymbol{v}_t] \\ &= P_t^{t-1} A_t'. \end{aligned} \tag{6.26}$$

Using these results we have that joint conditional distribution of $\boldsymbol{x}_t$ and $\boldsymbol{\epsilon}_t$ given Y_{t-1} is normal

$$\begin{pmatrix} \boldsymbol{x}_t \\ \boldsymbol{\epsilon}_t \end{pmatrix} \Big| \, Y_{t-1} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{x}_t^{t-1} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} P_t^{t-1} & P_t^{t-1} A_t' \\ A_t P_t^{t-1} & \Sigma_t \end{bmatrix} \right). \tag{6.27}$$

Thus, using (B.9) of Appendix B, we can write

$$\boldsymbol{x}_t^t = E(\boldsymbol{x}_t \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}, \boldsymbol{y}_t) = E(\boldsymbol{x}_t \mid Y_{t-1}, \boldsymbol{\epsilon}_t) = \boldsymbol{x}_t^{t-1} + K_t \boldsymbol{\epsilon}_t, \tag{6.28}$$

where

$$K_t = P_t^{t-1} A_t' \Sigma_t^{-1} = P_t^{t-1} A_t' (A_t P_t^{t-1} A_t' + R)^{-1}.$$

The evaluation of P_t^t is easily computed from (6.27) [see (B.10)] as

$$P_t^t = \mathrm{cov}\left(\boldsymbol{x}_t \mid Y_{t-1}, \boldsymbol{\epsilon}_t\right) = P_t^{t-1} - P_t^{t-1} A_t' \Sigma_t^{-1} A_t P_t^{t-1},$$

which simplifies to (6.22). ∎

Next, we explore the model, prediction, and filtering from a density point of view. For the sake of brevity, consider the Gaussian DLM without inputs, as described in (6.1) and (6.2); that is,

$$\boldsymbol{x}_t = \Phi \boldsymbol{x}_{t-1} + \boldsymbol{w}_t \quad \text{and} \quad \boldsymbol{y}_t = A_t \boldsymbol{x}_t + \boldsymbol{v}_t.$$

Recall $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are independent, white Gaussian sequences, and the initial state is normal, say, $\boldsymbol{x}_0 \sim \mathrm{N}(\boldsymbol{\mu}_0, \Sigma_0)$; we will denote the initial p-variate state normal density by $f_0(\boldsymbol{x}_0)$. Now, letting $p_\Theta(\cdot)$ denote a generic density function with parameters represented by Θ, we could describe the state relationship as

$$p_\Theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2}, \ldots, \boldsymbol{x}_0) = p_\Theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = f_w(\boldsymbol{x}_t - \Phi \boldsymbol{x}_{t-1}), \tag{6.29}$$

where $f_w(\cdot)$ denotes the p-variate normal density with mean zero and variance-covariance matrix Q. In (6.29), we are stating the process is Markovian, linear, and Gaussian. The relationship of the observations to the state process is written as

$$p_\Theta(\boldsymbol{y}_t \mid \boldsymbol{x}_t, Y_{t-1}) = p_\Theta(\boldsymbol{y}_t \mid \boldsymbol{x}_t) = f_v(\boldsymbol{y}_t - A_t \boldsymbol{x}_t), \tag{6.30}$$

where $f_v(\cdot)$ denotes the q-variate normal density with mean zero and variance-covariance matrix R. In (6.30), we are stating the observations are conditionally independent given the state, and the observations are linear and Gaussian. Note, (6.29), (6.30), and the initial density, $f_0(\cdot)$, completely specify the model in terms of densities, namely,

$$p_\Theta(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = f_0(\boldsymbol{x}_0) \prod_{t=1}^{n} f_w(\boldsymbol{x}_t - \Phi \boldsymbol{x}_{t-1}) f_v(\boldsymbol{y}_t - A_t \boldsymbol{x}_t), \tag{6.31}$$

where $\Theta = \{\boldsymbol{\mu}_0, \Sigma_0, \Phi, Q, R\}$.

Given the data, $Y_{t-1} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}\}$, and the current filter density, $p_\Theta(\boldsymbol{x}_{t-1} \,|\, Y_{t-1})$, Property P6.1 tells us, via conditional means and variances, how to recursively generate the Gaussian forecast density, $p_\Theta(\boldsymbol{x}_t \,|\, Y_{t-1})$, and how to update the density given the current observation, $\boldsymbol{y}_t$, to obtain the Gaussian filter density, $p_\Theta(\boldsymbol{x}_t \,|\, Y_t)$. In terms of densities, the Kalman filter can be seen as a simple Bayesian updating scheme, where, to determine the forecast and filter densities, we have

$$\begin{aligned} p_\Theta(\boldsymbol{x}_t \mid Y_{t-1}) &= \int_{R^p} p_\Theta(\boldsymbol{x}_t, \boldsymbol{x}_{t-1} \mid Y_{t-1})\, d\boldsymbol{x}_{t-1} \\ &= \int_{R^p} p_\Theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) p_\Theta(\boldsymbol{x}_{t-1} \mid Y_{t-1})\, d\boldsymbol{x}_{t-1} \\ &= \int_{R^p} f_w(\boldsymbol{x}_t - \Phi \boldsymbol{x}_{t-1}) p_\Theta(\boldsymbol{x}_{t-1} \mid Y_{t-1}) d\boldsymbol{x}_{t-1}, \end{aligned} \tag{6.32}$$

which simplifies to the p-variate $\mathrm{N}(\boldsymbol{x}_t^{t-1}, P_t^{t-1})$ density, and

$$\begin{aligned} p_\Theta(\boldsymbol{x}_t \mid Y_t) &= p_\Theta(\boldsymbol{x}_t \mid \boldsymbol{y}_t, Y_{t-1}) \\ &\propto p_\Theta(\boldsymbol{y}_t \mid \boldsymbol{x}_t)\; p_\Theta(\boldsymbol{x}_t \mid Y_{t-1}), \\ &= f_v(\boldsymbol{y}_t - A_t \boldsymbol{x}_t) p_\Theta(\boldsymbol{x}_t \mid Y_{t-1}), \end{aligned} \tag{6.33}$$

from which we can deduce $p_\Theta(\boldsymbol{x}_t \mid Y_t)$ is the p-variate $\mathrm{N}(\boldsymbol{x}_t^t, P_t^t)$ density. These statements are true for $t = 1, \ldots, n$, with initial condition $p_\Theta(\boldsymbol{x}_0 \mid Y_0) = f_0(\boldsymbol{x}_0)$. The prediction and filter recursions of Property P6.1 could also have been calculated directly from the density relationships (6.32) and (6.33) using multivariate normal distribution theory. The following example illustrates the Bayesian updating scheme.

Example 6.4 Bayesian Analysis of a Local Level Model

In this example, we suppose that we observe a univariate series y_t that consists of a trend component, μ_t, and a noise component, v_t, where

$$y_t = \mu_t + v_t \tag{6.34}$$

and $v_t \sim$ iid $\mathrm{N}(0, \sigma_v^2)$. In particular, we assume the trend is a random walk given by

$$\mu_t = \mu_{t-1} + w_t \tag{6.35}$$

where $w_t \sim$ iid $\mathrm{N}(0, \sigma_w^2)$ is independent of $\{v_t\}$. Recall Example 6.2, where we suggested this type of trend model for the global temperature series.

The model is, of course, a state-space model with (6.34) being the observation equation, and (6.35) being the state equation. For forecasting, we seek the posterior density $p(\mu_t \mid Y_{t-1})$. We will use the following notation introduced in Blight (1974) for the multivariate case. Let

$$\{x; \mu, \sigma^2\} = \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \tag{6.36}$$

then simple manipulation shows

$$\{x;\mu,\sigma^2\} = \{\mu;x,\sigma^2\} \tag{6.37}$$

and

$$\begin{aligned}\{x;\mu_1,\sigma_1^2\}\{x;\mu_2,\sigma_2^2\} &= \left\{x;\frac{\mu_1/\sigma_1^2+\mu_2/\sigma_2^2}{1/\sigma_1^2+1/\sigma_2^2},(1/\sigma_1^2+1/\sigma_2^2)^{-1}\right\}\\ &\times \quad \left\{\mu_1;\mu_2,\sigma_1^2+\sigma_2^2\right\}. \end{aligned}\tag{6.38}$$

Thus, using (6.32), (6.37) and (6.38) we have

$$\begin{aligned} p(\mu_t \mid Y_{t-1}) &\propto \int \left\{\mu_t;\mu_{t-1},\sigma_w^2\right\}\left\{\mu_{t-1};\mu_{t-1}^{t-1},P_{t-1}^{t-1}\right\}d\mu_{t-1}\\ &= \int \left\{\mu_{t-1};\mu_t,\sigma_w^2\right\}\left\{\mu_{t-1};\mu_{t-1}^{t-1},P_{t-1}^{t-1}\right\}d\mu_{t-1}\\ &= \left\{\mu_t;\mu_{t-1}^{t-1},P_{t-1}^{t-1}+\sigma_w^2\right\}. \end{aligned}\tag{6.39}$$

From (6.39) we conclude that

$$\mu_t \mid Y_{t-1} \sim \mathrm{N}(\mu_t^{t-1},P_t^{t-1}) \tag{6.40}$$

where

$$\mu_t^{t-1}=\mu_{t-1}^{t-1} \quad \text{and} \quad P_t^{t-1}=P_{t-1}^{t-1}+\sigma_w^2 \tag{6.41}$$

which agrees with the first part of Property P6.1.

To derive the filter density using (6.33) and (6.37) we have

$$\begin{aligned} p(\mu_t \mid Y_t) &\propto \left\{y_t;\mu_t,\sigma_v^2\right\}\left\{\mu_t;\mu_t^{t-1},P_t^{t-1}\right\}\\ &= \left\{\mu_t;y_t,\sigma_v^2\right\}\left\{\mu_t;\mu_t^{t-1},P_t^{t-1}\right\}. \end{aligned}\tag{6.42}$$

An application of (6.38) gives

$$\mu_t \mid Y_t \sim \mathrm{N}(\mu_t^t,P_t^t) \tag{6.43}$$

with

$$\mu_t^t=\frac{\sigma_v^2\mu_t^{t-1}}{P_t^{t-1}+\sigma_v^2}+\frac{P_t^{t-1}y_t}{P_t^{t-1}+\sigma_v^2}=\mu_t^{t-1}+K_t(y_t-\mu_t^{t-1}), \tag{6.44}$$

where we have defined

$$K_t=\frac{P_t^{t-1}}{P_t^{t-1}+\sigma_v^2}, \tag{6.45}$$

and

$$P_t^t=\left(\frac{1}{\sigma_v^2}+\frac{1}{P_t^{t-1}}\right)^{-1}=\frac{\sigma_v^2P_t^{t-1}}{P_t^{t-1}+\sigma_v^2}=(1-K_t)P_t^{t-1}. \tag{6.46}$$

The filter for this specific case, of course, agrees with Property P6.1.

Next, we consider the problem of obtaining estimators for $\boldsymbol{x}_t$ based on the entire data sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, where $t \leq n$, namely, $\boldsymbol{x}_t^n$. These estimators are called smoothers because a time plot of the sequence $\{\boldsymbol{x}_t^n;\ t = 1, \ldots, n\}$ is typically smoother than the forecasts $\{\boldsymbol{x}_t^{t-1};\ t = 1, \ldots, n\}$ or the filters $\{\boldsymbol{x}_t^t;\ t = 1, \ldots, n\}$. As is obvious from the above remarks, smoothing implies that each estimated value is a function of the present, future, and past, whereas the filtered estimator depends on the present and past. The forecast depends only on the past, as usual.

Property P6.2: The Kalman Smoother
For the state-space model specified in (6.3) and (6.4), with initial conditions $\boldsymbol{x}_n^n$ *and* P_n^n *obtained via Property P6.1, for* $t = n, n-1, \ldots, 1$,

$$\boldsymbol{x}_{t-1}^n = \boldsymbol{x}_{t-1}^{t-1} + J_{t-1}\left(\boldsymbol{x}_t^n - \boldsymbol{x}_t^{t-1}\right), \tag{6.47}$$

$$P_{t-1}^n = P_{t-1}^{t-1} + J_{t-1}\left(P_t^n - P_t^{t-1}\right) J_{t-1}', \tag{6.48}$$

where

$$J_{t-1} = P_{t-1}^{t-1}\Phi' \left[P_t^{t-1}\right]^{-1}. \tag{6.49}$$

Proof. The smoother can be derived in many ways. Here we provide a proof that was given in Ansley and Kohn (1982). First, for $1 \leq t \leq n$, define

$$Y_{t-1} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}\} \quad \text{and} \quad \eta_t = \{\boldsymbol{v}_t, \ldots, \boldsymbol{v}_n, \boldsymbol{w}_{t+1}, \ldots, \boldsymbol{w}_n\},$$

with Y_0 being empty, and let

$$\boldsymbol{q}_{t-1} = E\{\boldsymbol{x}_{t-1} \mid Y_{t-1},\, \boldsymbol{x}_t - \boldsymbol{x}_t^{t-1},\, \eta_t\}.$$

Then, because Y_{t-1}, $\{\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1}\}$, and η_t are mutually independent, and $\boldsymbol{x}_{t-1}$ and η_t are independent, using (B.9) we have

$$\boldsymbol{q}_{t-1} = \boldsymbol{x}_{t-1}^{t-1} + J_{t-1}(\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1}), \tag{6.50}$$

where

$$J_{t-1} = \mathrm{cov}(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t - \boldsymbol{x}_t^{t-1})[P_t^{t-1}]^{-1} = P_{t-1}^{t-1}\Phi'[P_t^{t-1}]^{-1}.$$

Finally, because Y_{t-1}, $\boldsymbol{x}_t - \boldsymbol{x}_t^{t-1}$, and η_t generate $Y_n = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$,

$$x_{t-1}^n = E\{x_{t-1} \mid Y_n\} = E\{\boldsymbol{q}_{t-1} \mid Y_n\} = \boldsymbol{x}_{t-1}^{t-1} + J_{t-1}(\boldsymbol{x}_t^n - \boldsymbol{x}_t^{t-1}),$$

which establishes (6.47).

The recursion for the error covariance, P_{t-1}^n, is obtained by straight-forward calculation. Using (6.47) we obtain

$$\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-1}^n = \boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-1}^{t-1} - J_{t-1}\left(\boldsymbol{x}_t^n - \Phi\boldsymbol{x}_{t-1}^{t-1}\right),$$

or

$$\left(\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-1}^{n}\right) + J_{t-1}\boldsymbol{x}_{t}^{n} = \left(\boldsymbol{x}_{t-1} - \boldsymbol{x}_{t-1}^{t-1}\right) + J_{t-1}\Phi\boldsymbol{x}_{t-1}^{t-1}. \tag{6.51}$$

Multiplying each side of (6.51) by the transpose of itself and talking expectation, we have

$$P_{t-1}^{n} + J_{t-1}E(\boldsymbol{x}_{t}^{n}\boldsymbol{x}_{t}^{n'})J_{t-1}' = P_{t-1}^{t-1} + J_{t-1}\Phi E(\boldsymbol{x}_{t-1}^{t-1}\boldsymbol{x}_{t-1}^{t-1'})\Phi' J_{t-1}', \tag{6.52}$$

using the fact the cross-product terms are zero. But,

$$E(\boldsymbol{x}_{t}^{n}\boldsymbol{x}_{t}^{n'}) = E(\boldsymbol{x}_{t}\boldsymbol{x}_{t}') - P_{t}^{n} = \Phi E(\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}')\Phi' + Q - P_{t}^{n},$$

and

$$E(\boldsymbol{x}_{t-1}^{t-1}\boldsymbol{x}_{t-1}^{t-1'}) = E(\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}') - P_{t-1}^{t-1},$$

so (6.52) simplifies to (6.48). ∎

Example 6.5 Prediction, Filtering and Smoothing for the Local Level Model

For this example, we simulated $n = 50$ observations from the local level trend model discussed in Example 6.4. We generated a random walk

$$\mu_t = \mu_{t-1} + w_t \tag{6.53}$$

with $w_t \sim$ iid $\mathrm{N}(0, 1)$ and $\mu_0 \sim \mathrm{N}(0, 1)$. We then supposed that we observe a univariate series y_t consisting of the trend component, μ_t, and a noise component, $v_t \sim$ iid $\mathrm{N}(0, 1)$, where

$$y_t = \mu_t + v_t. \tag{6.54}$$

The sequences $\{w_t\}$, $\{v_t\}$ and μ_0 were generated independently. We then ran the Kalman filter and smoother, Properties P6.1 and P6.2, using the actual parameters. The top panel of Figure 6.3 shows the actual values of μ_t as points, and the predictions μ_t^{t-1} superimposed on the graph as a line. In addition, we display $\mu_t^{t-1} \pm 2\sqrt{P_t^{t-1}}$ as dashed lines on the plot. The middle panel displays the filters, μ_t^t as a line with $\mu_t^t \pm 2\sqrt{P_t^t}$ as dashed lines. The bottom panel of Figure 6.3 shows a similar plot for the smoothers μ_t^n.

Table 6.1 shows the first 10 observations as well as the corresponding state values, the predictions, filters and smoothers. Note that in Table 6.1, one-step-ahead prediction is more uncertain than the corresponding filtered value, which, in turn, is more uncertain than the corresponding smoother value (that is $P_t^{t-1} > P_t^t > P_t^n$). Also, in each case, the error variances stabilize quickly. The R code for this example may be found on the website for the text.

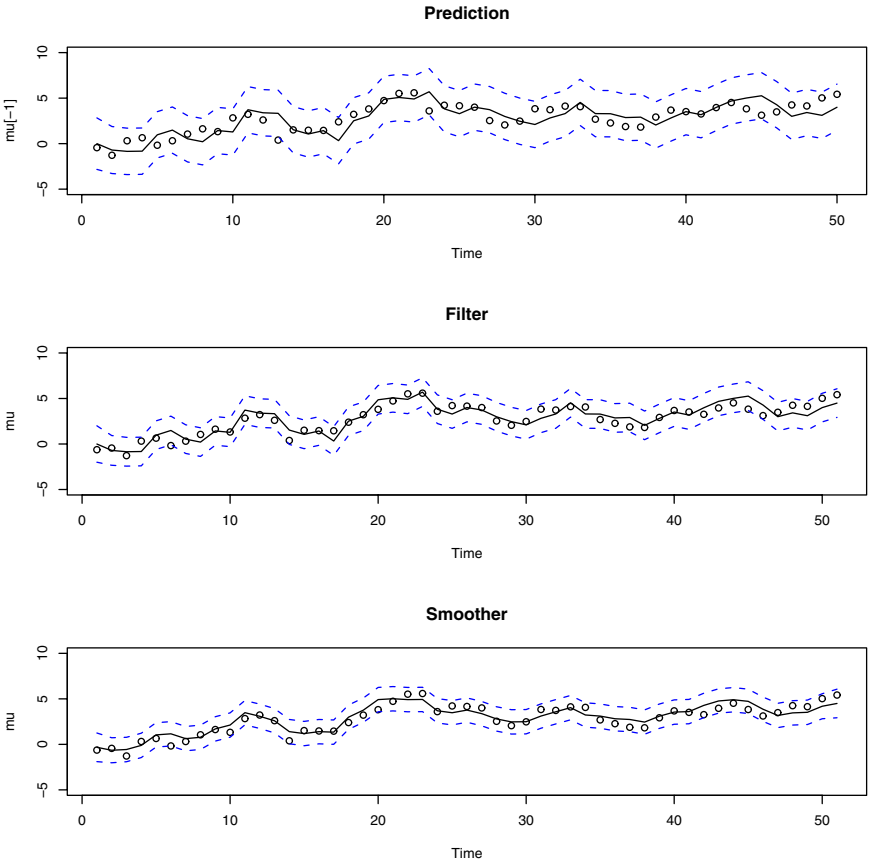


Figure 6.3 Displays for Example 6.5. The simulated values of μ_t, for $t = 1,\ldots,50$, given by (6.53) are shown as points. *Top:* The predictions μ_t^{t-1} obtained via the Kalman filter are shown as a line. Error bounds, $\mu_t^{t-1} \pm 2\sqrt{P_t^{t-1}}$, are shown as dashed lines. *Middle:* The filter μ_t^t obtained via the Kalman filter are shown as a line. Error bounds, $\mu_t^t \pm 2\sqrt{P_t^t}$, are shown as dashed lines. *Bottom:* The smoothers μ_t^n obtained via the Kalman smoother are shown as a line. Error bounds, $\mu_t^n \pm 2\sqrt{P_t^n}$, are shown as dashed lines.

In the next section, we will need a set of recursions for obtaining $P_{t,t-1}^n$, as defined in (6.18). We give the necessary recursion in the following property.

Property P6.3: The Lag-One Covariance Smoother
For the state-space model specified in (6.3) and (6.4), with K_t, J_t $(t = 1,\ldots,n)$, and P_n^n obtained from Properties P6.1 and P6.2, and with initial condition

$$P_{n,n-1}^n = (I - K_n A_n)\Phi P_{n-1}^{n-1}, \tag{6.55}$$

Table 6.1 Forecasts, Filters, and Smoothers for Example 6.5.

t	y_t	μ_t	μ_t^{t-1}	P_t^{t-1}	μ_t^t	P_t^t	μ_t^n	P_t^n
0	—	−.63	—	—	.00	1.00	−.32	.62
1	−1.05	−.44	.00	2.00	−.70	.67	−.65	.47
2	−.94	−1.28	−.70	1.67	−.85	.63	−.57	.45
3	−.81	.32	−.85	1.63	−.83	.62	−.11	.45
4	2.08	.65	−.83	1.62	.97	.62	1.04	.45
5	1.81	−.17	.97	1.62	1.49	.62	1.16	.45
6	−.05	.31	1.49	1.62	.53	.62	.63	.45
7	.01	1.05	.53	1.62	.21	.62	.78	.45
8	2.20	1.63	.21	1.62	1.44	.62	1.70	.45
9	1.19	1.32	1.44	1.62	1.28	.62	2.12	.45
10	5.24	2.83	1.28	1.62	3.73	.62	3.48	.45

for $t = n, n-1, \ldots, 2$,

$$P_{t-1,t-2}^n = P_{t-1}^{t-1} J_{t-2}' + J_{t-1}\left(P_{t,t-1}^n - \Phi P_{t-1}^{t-1}\right) J_{t-2}'. \tag{6.56}$$

Proof. Because we are computing covariances, we may assume $\boldsymbol{u}_t \equiv \mathbf{0}$ without loss of generality. To derive the initial term (6.55), we first define

$$\widetilde{\boldsymbol{x}}_t^s = \boldsymbol{x}_t - \boldsymbol{x}_t^s.$$

Then, using (6.21) and (6.47), we write

$$\begin{aligned} P_{t,t-1}^t &= E\left(\widetilde{\boldsymbol{x}}_t^t\, \widetilde{\boldsymbol{x}}_{t-1}^{t'}\right) \\ &= E\left\{[\widetilde{\boldsymbol{x}}_t^{t-1} - K_t(\boldsymbol{y}_t - A_t\boldsymbol{x}_t^{t-1})][\widetilde{\boldsymbol{x}}_{t-1}^{t-1} - J_{t-1}K_t(\boldsymbol{y}_t - A_t\boldsymbol{x}_t^{t-1})]'\right\} \\ &= E\left\{[\widetilde{\boldsymbol{x}}_t^{t-1} - K_t(A_t\widetilde{\boldsymbol{x}}_t^{t-1} + \boldsymbol{v}_t)][\widetilde{\boldsymbol{x}}_{t-1}^{t-1} - J_{t-1}K_t(A_t\widetilde{\boldsymbol{x}}_t^{t-1} + \boldsymbol{v}_t)]'\right\}. \end{aligned}$$

Expanding terms and taking expectation, we arrive at

$$P_{t,t-1}^t = P_{t,t-1}^{t-1} - P_t^{t-1}A_t'K_t'J_{t-1}' - K_tA_tP_{t,t-1}^{t-1} + K_t(A_tP_t^{t-1}A_t' + R)K_t'J_{t-1}',$$

noting $E(\widetilde{\boldsymbol{x}}_t^{t-1}\boldsymbol{v}_t') = \mathbf{0}$. The final simplification occurs by realizing that $K_t(A_tP_t^{t-1}A_t' + R) = P_t^{t-1}A_t'$, and $P_{t,t-1}^{t-1} = \Phi P_{t-1}^{t-1}$. These relationships hold for any $t = 1, \ldots, n$, and (6.55) is the case $t = n$.

We give the basic steps in the derivation of (6.56). The first step is to use (6.47) to write

$$\widetilde{\boldsymbol{x}}_{t-1}^n + J_{t-1}\boldsymbol{x}_t^n = \widetilde{\boldsymbol{x}}_{t-1}^{t-1} + J_{t-1}\Phi\boldsymbol{x}_{t-1}^{t-1} \tag{6.57}$$

and

$$\widetilde{\boldsymbol{x}}_{t-2}^{n} + J_{t-2}\boldsymbol{x}_{t-1}^{n} = \widetilde{\boldsymbol{x}}_{t-2}^{t-2} + J_{t-2}\Phi\boldsymbol{x}_{t-2}^{t-2}. \tag{6.58}$$

Next, multiply the left-hand side of (6.57) by the transpose of the left-hand side of (6.58), and equate that to the corresponding result of the right-hand sides of (6.57) and (6.58). Then, taking expectation of both sides, the left-hand side result reduces to

$$P_{t-1,t-2}^{n} + J_{t-1}E(\boldsymbol{x}_{t}^{n}\boldsymbol{x}_{t-1}^{n'})J_{t-2}' \tag{6.59}$$

and the right-hand side result reduces to

$$\begin{aligned} P_{t-1,t-2}^{t-2} &- K_{t-1}A_{t-1}P_{t-1,t-2}^{t-2} + J_{t-1}\Phi K_{t-1}A_{t-1}P_{t-1,t-2}^{t-2} \\ &+ J_{t-1}\Phi E(\boldsymbol{x}_{t-1}^{t-1}\boldsymbol{x}_{t-2}^{t-2'})\Phi' J_{t-2}'. \end{aligned} \tag{6.60}$$

In (6.59), write

$$E(\boldsymbol{x}_{t}^{n}\boldsymbol{x}_{t-1}^{n'}) = E(\boldsymbol{x}_{t}\boldsymbol{x}_{t-1}') - P_{t,t-1}^{n} = \Phi E(\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-2}')\Phi' + \Phi Q - P_{t,t-1}^{n},$$

and in (6.60), write

$$E(\boldsymbol{x}_{t-1}^{t-1}\boldsymbol{x}_{t-2}^{t-2'}) = E(\boldsymbol{x}_{t-1}^{t-2}\boldsymbol{x}_{t-2}^{t-2'}) = E(\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-2}') - P_{t-1,t-2}^{t-2}.$$

Equating (6.59) to (6.60) using these relationships and simplifying the result leads to (6.56). ∎

6.3 Maximum Likelihood Estimation

The estimation of the parameters that specify the state-space model, (6.3) and (6.4), is quite involved. We use $\Theta = \{\boldsymbol{\mu}_0, \Sigma_0, \Phi, Q, R, \Upsilon, \Gamma\}$ to represent the vector of parameters containing the elements of the initial mean and covariance $\boldsymbol{\mu}_0$ and Σ_0, the transition matrix Φ, and the state and observation covariance matrices Q and R and the input coefficient matrices, Υ and Γ. We use maximum likelihood under the assumption that the initial state is normal, $\boldsymbol{x}_0 \sim \mathrm{N}(\boldsymbol{\mu}_0, \Sigma_0)$, and the errors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are jointly normal and uncorrelated vector variables. We continue to assume, for simplicity, $\{\boldsymbol{w}_t\}$ and $\{\boldsymbol{v}_t\}$ are uncorrelated.

The likelihood is computed using the innovations $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_n$, defined by (6.24),

$$\boldsymbol{\epsilon}_t = \boldsymbol{y}_t - A_t\boldsymbol{x}_t^{t-1} - \Gamma\boldsymbol{u}_t.$$

The innovations form of the likelihood function, which was first given by Schweppe (1965), is obtained using an argument similar to the one leading to (3.105) and proceeds by noting the innovations are independent Gaussian random vectors with zero means and, as shown in (6.25), covariance matrices

$$\Sigma_t = A_t P_t^{t-1} A_t' + R. \tag{6.61}$$

Hence, ignoring a constant, we may write the likelihood, $L_Y(\Theta)$, as

$$-\ln L_Y(\Theta) = \frac{1}{2}\sum_{t=1}^{n}\log|\Sigma_t(\Theta)| + \frac{1}{2}\sum_{t=1}^{n}\boldsymbol{\epsilon}_t(\Theta)'\Sigma_t(\Theta)^{-1}\boldsymbol{\epsilon}_t(\Theta), \tag{6.62}$$

where we have emphasized the dependence of the innovations on the parameters Θ. Of course, (6.62) is a highly nonlinear and complicated function of the unknown parameters. The usual procedure is to fix $\boldsymbol{x}_0$ and then develop a set of recursions for the log likelihood function and its first two derivatives (for example, Gupta and Mehra, 1974). Then, a Newton–Raphson algorithm (see Example 3.28) can be used successively to update the parameter values until the negative of the log likelihood is minimized. This approach is advocated, for example, by Jones (1980), who developed ARMA estimation by putting the ARMA model in state-space form. For the univariate case, (6.62) is identical, in form, to the likelihood for the ARMA model given in (3.105).

The steps involved in performing a Newton–Raphson estimation procedure are as follows.

1. Select initial values for the parameters, say, $\Theta^{(0)}$.

2. Run the Kalman filter, Property P6.1, using the initial parameter values, $\Theta^{(0)}$, to obtain a set of innovations and error covariances, say, $\{\boldsymbol{\epsilon}_t^{(0)};\ t = 1,\ldots,n\}$ and $\{\Sigma_t^{(0)};\ t = 1,\ldots,n\}$.

3. Run one iteration of a Newton–Raphson procedure with $-\ln L_Y(\Theta)$ as the criterion function (refer to Example 3.28 for details), to obtain a new set of estimates, say $\Theta^{(1)}$.

4. At iteration j, $(j = 1, 2, \ldots)$, repeat step 2 using $\Theta^{(j)}$ in place of $\Theta^{(j-1)}$ to obtain a new set of innovation values $\{\boldsymbol{\epsilon}_t^{(j)};\ t = 1,\ldots,n\}$ and $\{\Sigma_t^{(j)};\ t = 1,\ldots,n\}$. Then repeat step 3 to obtain a new estimate $\Theta^{(j+1)}$. Stop when the estimates or the likelihood stabilize; for example, stop when the values of $\Theta^{(j+1)}$ differ from $\Theta^{(j)}$, or when $L_Y(\Theta^{(j+1)})$ differs from $L_Y(\Theta^{(j)})$, by some predetermined, but small amount.

Example 6.6 Newton–Raphson for Example 6.3

In this example, we generated $n = 100$ observations, $y_1,\ldots,y_{100}$, using the model in Example 6.3, to perform a Newton–Raphson estimation of the parameters ϕ, σ_w^2, and σ_v^2. In the notation of §6.2, we would have $\Phi = \phi$, $Q = \sigma_w^2$ and $R = \sigma_v^2$. The actual values of the parameters are $\phi = .8$, $\sigma_w^2 = \sigma_v^2 = 1$.

In the simple case of an AR(1) with observational noise, initial estimation can be accomplished using the results of Example 6.3. For example, using (6.16), we set

$$\phi^{(0)} = \widehat{\rho}_y(2)/\widehat{\rho}_y(1).$$

Similarly, from (6.15), $\gamma_x(1) = \gamma_y(1) = \phi\sigma_w^2/(1-\phi^2)$, so that, initially, we set

$$\sigma_w^{2^{(0)}} = (1-\phi^{(0)^2})\widehat{\gamma}_y(1)/\phi^{(0)}.$$

Finally, using (6.14) we obtain an initial estimate of σ_v^2, namely,

$$\sigma_v^{2^{(0)}} = \widehat{\gamma}_y(0) - [\sigma_w^{2^{(0)}}/(1-\phi^{(0)^2})].$$

Newton–Raphson estimation was accomplished using the R program `optim`. The code used for this example can be obtained on the website for the text. In that program, we must provide an evaluation of the function to be minimized, namely, $-\ln L_Y(\Theta)$. In this case, the "function call" combines steps 2 and 3, using the current values of the parameters, $\Theta^{(j-1)}$, to obtain first the filtered values, then the innovation values, and then calculating the criterion function, $-\ln L_Y(\Theta^{(j-1)})$, to be minimized. We can also provide analytic forms of the gradient or score vector, $-\partial \ln L_Y(\Theta)/\partial\Theta$, and the Hessian matrix, $-\partial^2 \ln L_Y(\Theta)/\partial\Theta\,\partial\Theta'$, in the optimization routine, or allow the program to calculate these values numerically. In this example, we let the program proceed numerically and we note the need to be cautious when calculating gradients numerically. For better stability, we can also provide an iterative solution for obtaining analytic gradients and Hessians of the log likelihood function; for details, see Problems 6.11 and 6.12; also, see Gupta and Mehra (1974). The final estimates, along with their standard errors (in parentheses), were

$$\widehat{\phi} = .81\ (.08), \quad \widehat{\sigma}_w = .85\ (.17), \quad \widehat{\sigma}_v = .87\ (.14),$$

and the algorithm converged in seven steps. The standard errors are a byproduct of the estimation procedure, and we will discuss their evaluation later in this section, after Property P6.4.

Example 6.7 Newton–Raphson for the Global Temperature Series in Example 6.2

In Example 6.2 we considered two different global temperature series of $n = 108$ observations each, and they are plotted in Figure 6.2. In that example, we argued that both series should be measuring the same underlying climatic signal, x_t, which we model as a random walk,

$$x_t = x_{t-1} + w_t.$$

Recall that the observation equation was written as

$$\begin{pmatrix} y_{t1} \\ y_{t2} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} x_t + \begin{pmatrix} v_{t1} \\ v_{t2} \end{pmatrix},$$

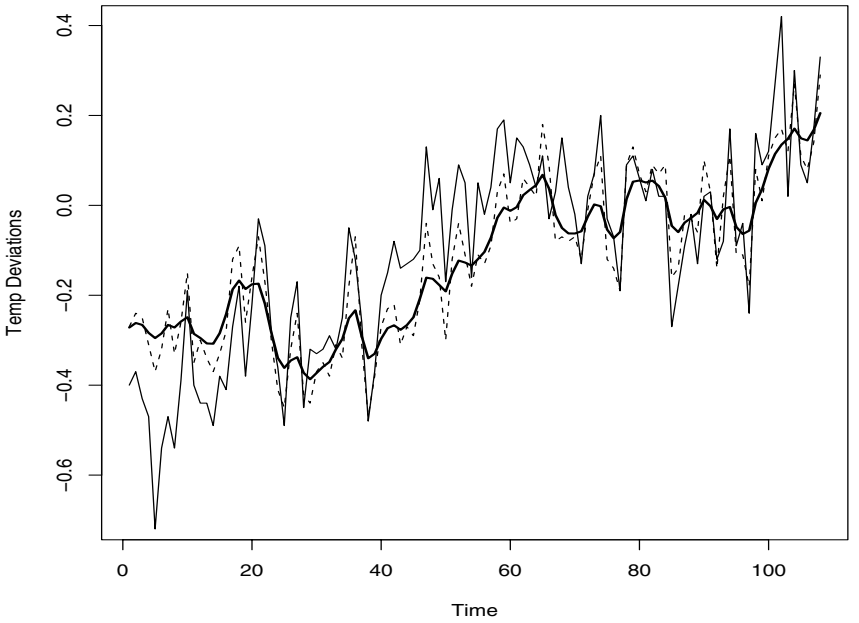


Figure 6.4 Plot for Example 6.7. The thin solid and dashed lines are the two average global temperature deviations shown in Figure 6.2. The thick solid line is the estimated smoother $\widehat{x}_t^n$.

and the model covariance matrices are given by $Q = q_{11}$ and

$$R = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}.$$

Hence, there are four parameters to estimate, namely $q_{11}, r_{11}, r_{12}, r_{22}$, noting that $r_{21} = r_{12}$ We hold the the initial state parameters fixed in this example at $\mu_0 = -.35$ and $\Sigma_0 = .01$ (these are, approximately, the mean and variance of the first observation in each series).

The final estimates are $\widehat{q}_{11} = .05, \widehat{r}_{11} = .019, \widehat{r}_{12} = .006, \widehat{r}_{22} = .005$, with all values being significant. The observations and the smoothed estimate of the signal, $\widehat{x}_t^n$, are displayed in Figure 6.4.

In addition to Newton–Raphson, Shumway and Stoffer (1982) presented a conceptually simpler estimation procedure based on the EM (expectation-maximization) algorithm (Dempster et al., 1977). For the sake of brevity, we ignore the inputs and consider the model in the form of (6.1) and (6.2); the general case is left as an exercise (Problem 6.9). The basic idea is that if we could observe the states, $X_n = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, in addition to the observations $Y_n = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$, then we would consider $\{X_n, Y_n\}$ as the complete data, with

the joint density

$$f_\Theta(X_n, Y_n) = f_{\mu_0,\Sigma_0}(\boldsymbol{x}_0) \prod_{t=1}^{n} f_{\Phi,Q}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \prod_{t=1}^{n} f_R(\boldsymbol{y}_t|\boldsymbol{x}_t). \tag{6.63}$$

Under the Gaussian assumption and ignoring constants, the complete data likelihood, (6.63), can be written as

$$\begin{aligned} -2\ln L_{X,Y}(\Theta) &= \ln|\Sigma_0| + (\boldsymbol{x}_0 - \boldsymbol{\mu}_0)'\Sigma_0^{-1}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0) \\ &+ n\ln|Q| + \sum_{t=1}^{n}(\boldsymbol{x}_t - \Phi\boldsymbol{x}_{t-1})'Q^{-1}(\boldsymbol{x}_t - \Phi\boldsymbol{x}_{t-1}) \\ &+ n\ln|R| + \sum_{t=1}^{n}(\boldsymbol{y}_t - A_t\boldsymbol{x}_t)'R^{-1}(\boldsymbol{y}_t - A_t\boldsymbol{x}_t). \end{aligned} \tag{6.64}$$

Thus, in view of (6.64), if we did have the complete data, we could then use the results from multivariate normal theory to easily obtain the MLEs of Θ. We do not have the complete data; however, the EM algorithm gives us an iterative method for finding the MLEs of Θ based on the incomplete data, Y_n, by successively maximizing the conditional expectation of the complete data likelihood. To implement the EM algorithm, we write, at iteration j, $(j = 1, 2, \ldots)$,

$$Q\left(\Theta \mid \Theta^{(j-1)}\right) = E\left\{-2\ln L_{X,Y}(\Theta) \mid Y_n, \Theta^{(j-1)}\right\}. \tag{6.65}$$

Calculation of (6.65) is the expectation step. Of course, given the current value of the parameters, $\Theta^{(j-1)}$, we can use Property P6.2 to obtain the desired conditional expectations as smoothers. This property yields

$$\begin{aligned} Q\left(\Theta \mid \Theta^{(j-1)}\right) &= \ln|\Sigma_0| + \mathrm{tr}\left\{\Sigma_0^{-1}[P_0^n + (\boldsymbol{x}_0^n - \boldsymbol{\mu}_0)(\boldsymbol{x}_0^n - \boldsymbol{\mu}_0)']\right\} \\ &+ n\ln|Q| + \mathrm{tr}\left\{Q^{-1}[S_{11} - S_{10}\Phi' - \Phi S_{10}' + \Phi S_{00}\Phi']\right\} \\ &+ n\ln|R| \\ &+ \mathrm{tr}\left\{R^{-1}\sum_{t=1}^{n}[(\boldsymbol{y}_t - A_t\boldsymbol{x}_t^n)(\boldsymbol{y}_t - A_t\boldsymbol{x}_t^n)' + A_tP_t^nA_t']\right\}, \end{aligned} \tag{6.66}$$

where

$$S_{11} = \sum_{t=1}^{n}(\boldsymbol{x}_t^n\boldsymbol{x}_t^{n\prime} + P_t^n), \tag{6.67}$$

$$S_{10} = \sum_{t=1}^{n}(\boldsymbol{x}_t^n\boldsymbol{x}_{t-1}^{n\prime} + P_{t,t-1}^n), \tag{6.68}$$

and

$$S_{00} = \sum_{t=1}^{n}(\boldsymbol{x}_{t-1}^n\boldsymbol{x}_{t-1}^{n\prime} + P_{t-1}^n). \tag{6.69}$$

In (6.66)-(6.69), the smoothers are calculated under the current value of the parameters $\Theta^{(j-1)}$; for simplicity, we have not explicitly displayed this fact.

Minimizing (6.66) with respect to the parameters, at iteration j, constitutes the maximization step, and is analogous to the usual multivariate regression approach, which yields the updated estimates

$$\Phi^{(j)} = S_{10}S_{00}^{-1}, \tag{6.70}$$

$$Q^{(j)} = n^{-1}\left(S_{11} - S_{10}S_{00}^{-1}S_{10}'\right), \tag{6.71}$$

and

$$R^{(j)} = n^{-1}\sum_{t=1}^{n}[(\boldsymbol{y}_t - A_t\boldsymbol{x}_t^n)(\boldsymbol{y}_t - A_t\boldsymbol{x}_t^n)' + A_tP_t^nA_t']. \tag{6.72}$$

The updates for the initial mean and variance–covariance matrix are

$$\boldsymbol{\mu}_0^{(j)} = \boldsymbol{x}_0^n \quad \text{and} \quad \Sigma_0^{(j)} = P_0^n \tag{6.73}$$

obtained from minimizing (6.66).

The overall procedure can be regarded as simply alternating between the Kalman filtering and smoothing recursions and the multivariate normal maximum likelihood estimators, as given by (6.70)–(6.73). Convergence results for the EM algorithm under general conditions can be found in Wu (1983). We summarize the iterative procedure as follows.

1. Initialize the procedure by selecting starting values for the parameters $\Theta^{(0)} = \{\boldsymbol{\mu}_0, \Sigma_0, \Phi, Q, R\}$.

On iteration j, $(j = 1, 2, \ldots)$:

2. Compute the incomplete-data likelihood, $-\ln L_Y(\Theta^{(j-1)})$; see equation (6.62).

3. Perform the E-Step. Use Properties 6.1, 6.2, and 6.3 to obtain the smoothed values $\boldsymbol{x}_t^n, P_t^n$ and $P_{t,t-1}^n$, for $t = 1, \ldots, n$, using the parameters $\Theta^{(j-1)}$. Use the smoothed values to calculate S_{11}, S_{10}, S_{00} given in (6.67)–(6.69).

4. Perform the M-Step. Update the estimates, $\boldsymbol{\mu}_0, \Sigma_0, \Phi, Q$, and R using (6.70)–(6.73), to obtain $\Theta^{(j)}$.

5. Repeat Steps 2 – 4 to convergence.

Example 6.8 EM Algorithm for Example 6.3

Using the same data generated in Example 6.6, we performed an EM algorithm estimation of the parameters ϕ, σ_w^2 and σ_v^2 as well as the initial parameters μ_0 and Σ_0.

The convergence rate of the EM algorithm compared with the Newton–Raphson procedure is slow. In this example, with convergence being claimed when the log likelihood does not change by more that .001, convergence was attained after 38 iterations.

The final estimates, along with their standard errors (in parentheses), were

$$\widehat{\phi} = .83\ (.08), \quad \widehat{\sigma}_w = .81\ (.17), \quad \widehat{\sigma}_v = .91\ (.14),$$

with $\widehat{\mu}_0 = -.06$ and $\widehat{\Sigma}_0 = .44$.

Evaluation of the standard errors used a call to `fdHess` in the `nlme` R package to evaluate the Hessian at the final estimates. The `nlme` package must be loaded prior to the call to `fdHess`.

Asymptotic Distribution of the MLEs

The asymptotic distribution of estimators of the model parameters, say, $\widehat{\Theta}_n$, is studied extensively in Caines (1988, Chapters 7 and 8), and in Hannan and Deistler (1988, Chapter 4). In both of these references, the consistency and asymptotic normality of the estimators is established under general conditions. Although we will only state the basic result, some crucial elements are needed to establish large sample properties of the estimators. An essential condition is the stability of the filter. Stability of the filter assures that, for large t, the innovations $\boldsymbol{\epsilon}_t$ are basically copies of each other (that is, independent and identically distributed) with a stable covariance matrix Σ that does not depend on t and that, asymptotically, the innovations contain all of the information about the unknown parameters. Although it is not necessary, for simplicity, we shall assume here that $A_t \equiv A$ for all t. Details on departures from this assumption can be found in Jazwinski (1970, Sections 7.6 and 7.8). We also drop the inputs as use the model in the form of (6.1) and (6.2).

For stability of the filter, we assume the eigenvalues of Φ are less than one in absolute value; this assumption can be weakened (for example, see Harvey, 1991, Section 4.3), but we retain it for simplicity. This assumption is enough to ensure the stability of the filter in that, as $t \to \infty$, the filter error covariance matrix P_t^t converges to P, the steady-state error covariance matrix, the gain matrix K_t converges to K, the steady-state gain matrix, from which it follows that the innovation variance–covariance matrix Σ_t converges to Σ, the steady-state variance–covariance matrix of the stable innovations; details can be found in Jazwinski (1970, Sections 7.6 and 7.8) and Anderson and Moore (1979, Section 4.4). In particular, the steady-state filter error covariance matrix, P,

satisfies the Riccati equation:

$$P = \Phi[P - PA'(APA' + R)^{-1}AP]\Phi' + Q;$$

the steady-state gain matrix satisfies $K = PA'[APA' + R]^{-1}$. In Example 6.5, for all practical purposes, stability was reached by the fourth observation.

When the process is in steady-state, we may consider $\boldsymbol{x}_{t+1}^t$ as the steady-state predictor and interpret it as $\boldsymbol{x}_{t+1}^t = E(\boldsymbol{x}_{t+1} \mid \boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots)$. As can be seen from (6.19) and (6.21), the steady-state predictor can be written as

$$\begin{aligned} \boldsymbol{x}_{t+1}^t &= \Phi[I - KA]\boldsymbol{x}_t^{t-1} + \Phi K y_t \\ &= \Phi \boldsymbol{x}_t^{t-1} + \Phi K \boldsymbol{\epsilon}_t, \end{aligned} \tag{6.74}$$

where $\boldsymbol{\epsilon}_t$ is the steady-state innovation process given by

$$\boldsymbol{\epsilon}_t = y_t - E(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \ldots).$$

In this case, $\boldsymbol{\epsilon}_t \sim$ iid $\mathrm{N}(\mathbf{0}, \Sigma)$, where $\Sigma = APA' + R$. In steady-state, the observations can be written as

$$\boldsymbol{y}_t = A\boldsymbol{x}_t^{t-1} + \boldsymbol{\epsilon}_t. \tag{6.75}$$

Together, (6.74) and (6.75) make up the steady-state innovations form of the dynamic linear model.

Two other conditions worth mentioning are observability and controllability. Observability focuses on the question of how much information can be gained about the p-dimensional state vector $\boldsymbol{x}_t$ from p future observations $\{\boldsymbol{y}_t, \boldsymbol{y}_{t+1}, \ldots, \boldsymbol{y}_{t+p-1}\}$. Consider the state without any noise term,

$$\boldsymbol{x}_{t+p} = \Phi \boldsymbol{x}_{t+p-1} = \cdots = \Phi^p \boldsymbol{x}_t.$$

Then, the data (without observational noise) satisfy

$$\boldsymbol{y}_{t+j} = A\boldsymbol{x}_{t+j} = A\Phi^j \boldsymbol{x}_t, \quad j = 0, \ldots, p-1,$$

or

$$(\boldsymbol{y}_t', \ldots, \boldsymbol{y}_{t+p-1}') = \boldsymbol{x}_t'[A', \Phi' A', \ldots, \Phi^{'p-1} A'].$$

Hence, if the observability matrix $\mathcal{O}' = [A', \Phi' A', \ldots, \Phi^{'p-1} A']$ has full rank p, we may explicitly solve for $\boldsymbol{x}_t$ in terms of $\boldsymbol{y}_{t:p} = (\boldsymbol{y}_t', \ldots, \boldsymbol{y}_{t+p-1}')'$, namely, $\boldsymbol{x}_t = (\mathcal{O}'\mathcal{O})^{-1}\mathcal{O}'\boldsymbol{y}_{t:p}$, and the system is said to be observable.

In a similar manner, to define controllability, write the state noise as $\boldsymbol{w}_t = B\boldsymbol{u}_t$, where B is $p \times r$ and $\boldsymbol{u}_t$ is an r-dimensional, nonsingular, white noise process. Thus, the state equation is $\boldsymbol{x}_t = \Phi \boldsymbol{x}_{t-1} + B\boldsymbol{u}_t$. If the matrix $\mathcal{C} = [B, \Phi B, \Phi^2 B, \ldots, \Phi^{p-1} B]$ has full rank p, the process is said to be controllable. Controllability has to do with the fact that the state equation satisfies

$$\boldsymbol{x}_{t+p} = \sum_{j=0}^{p-1} \Phi^j B \boldsymbol{u}_{t+p-j} + \Phi^p \boldsymbol{x}_t = \mathcal{C}\boldsymbol{U}_t + \Phi^p \boldsymbol{x}_t,$$

where $\boldsymbol{U}_t = (\boldsymbol{u}'_{t+p}, \ldots, \boldsymbol{u}'_{t+1})'$. If we think of the variables $\{\boldsymbol{u}_{t+p}, \ldots, \boldsymbol{u}_{t+1}\}$ as "controlling" the state output $\boldsymbol{x}_t$, and we act as if we are free to choose the $\boldsymbol{u}_t$ at will, the fact that $\mathcal{C}$ is full rank means any desired value of $\boldsymbol{x}_{t+p}$ can be obtained from any initial state $\boldsymbol{x}_t$ by control of $\boldsymbol{U}_t$. In particular, we can put $\boldsymbol{U}_t = \mathcal{C}'(\mathcal{C}\mathcal{C}')^{-1}\boldsymbol{x}_{t+p} - \Phi^p \boldsymbol{x}_t$.

The key point about controllability and observability is that these conditions are necessary and sufficient to ensure the state-space model has the smallest possible dimension; details can be found in Hannan and Diestler (1988, Section 2.3). As a simple example, suppose the state system is bivariate, $\boldsymbol{x}_t = (x_{t1}, x_{t2})'$, where x_{t1} and x_{t2} are independent components with $\Phi = \text{diag}\{\phi_1, \phi_2\}$, and $y_t = [1, 0]\boldsymbol{x}_t + v_t$; that is, $y_t = x_{t1} + v_t$. Clearly we could not hope to reasonably estimate ϕ_2. This system is not observable because $\mathcal{O}$ has rank one. Additional details on this point can be found in Jazwinski (1970, Section 7.5).

In the following property, we assume the Gaussian state-space model (6.1) and (6.2), is time invariant, i.e., $A_t \equiv A$, the eigenvalues of Φ are within the unit circle and the system is observable and controllable. We denote the true parameters by Θ_0, and we assume the dimension of Θ_0 is the dimension of the parameter space. Although it is not necessary to assume $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are Gaussian, certain additional conditions would have to apply and adjustments to the asymptotic covariance matrix would have to be made (see Caines, 1988, Chapter 8).

Property P6.4: Asymptotic Distribution of the Estimators
Under general conditions, let $\widehat{\Theta}_n$ be the estimator of Θ_0 obtained by maximizing the innovations likelihood, $L_Y(\Theta)$, as given in (6.62). Then, as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\Theta}_n - \Theta_0\right) \xrightarrow{d} \text{N}\left[0,\ \mathcal{I}(\Theta_0)^{-1}\right],$$

where $\mathcal{I}(\Theta)$ is the asymptotic information matrix given by

$$\mathcal{I}(\Theta) = \lim_{n\to\infty}\ n^{-1} E\left[-\partial^2 \ln L_Y(\Theta)/\partial\Theta\ \partial\Theta'\right].$$

Precise details and the proof of Property P6.4 are given in Caines (1988, Chapter 7) and in Hannan and Deistler (1988, Chapter 4). For a Newton procedure, the Hessian matrix (as described in Example 6.6) at the time of convergence can be used as an estimate of $n\mathcal{I}(\Theta_0)$ to obtain estimates of the standard errors. In the case of the EM algorithm, no derivatives are calculated, but we may include a numerical evaluation of the Hessian matrix at the time of convergence to obtain estimated standard errors. Also, extensions of the EM algorithm exist, such as the SEM algorithm (Meng and Rubin, 1991), that include a procedure for the estimation of standard errors. In the examples of this section, the estimated standard errors were obtained from the numerical Hessian matrix of $-\ln L_Y(\widehat{\Theta})$, where $\widehat{\Theta}$ is the vector of parameters estimates at the time of convergence.

6.4 Missing Data Modifications

An attractive feature available within the state-space framework is its ability to treat time series that have been observed irregularly over time. For example, Palma and Chan (1997) used the state-space model for estimation and forecasting of long memory (specifically, fractionally integrated ARMA, or ARFIMA, processes) time series with missing observations. Throughout this section we assume the model is of the form (6.1) and (6.2). The EM algorithm allows parts of the observation vector $\boldsymbol{y}_t$ to be missing at a number of observation times. Shumway and Stoffer (1982) described the modifications necessary for the special case in which the subvectors of $\boldsymbol{v}_t$ corresponding to the observed and unobserved parts of $\boldsymbol{y}_t$ happen to be uncorrelated. Here, we will also discuss the general case.

Suppose, at a given time t, we define the partition of the $q \times 1$ observation vector $\boldsymbol{y}_t = (\boldsymbol{y}_t^{(1)'}, \boldsymbol{y}_t^{(2)'})'$, where the first $q_{1t} \times 1$ component is observed and the second $q_{2t} \times 1$ component is unobserved, $q_{1t} + q_{2t} = q$. Then, write the partitioned observation equation

$$\begin{pmatrix} \boldsymbol{y}_t^{(1)} \\ \boldsymbol{y}_t^{(2)} \end{pmatrix} = \begin{bmatrix} A_t^{(1)} \\ A_t^{(2)} \end{bmatrix} \boldsymbol{x}_t + \begin{pmatrix} \boldsymbol{v}_t^{(1)} \\ \boldsymbol{v}_t^{(2)} \end{pmatrix}, \tag{6.76}$$

where $A_t^{(1)}$ and $A_t^{(2)}$ are, respectively, the $q_{1t} \times p$ and $q_{2t} \times p$ partitioned observation matrices, and

$$\operatorname{cov} \begin{pmatrix} \boldsymbol{v}_t^{(1)} \\ \boldsymbol{v}_t^{(2)} \end{pmatrix} = \begin{bmatrix} R_{11t} & R_{12t} \\ R_{21t} & R_{22t} \end{bmatrix} \tag{6.77}$$

denotes the covariance matrix of the measurement errors between the observed and unobserved parts. Stoffer (1982) established the filtering equations, Property P6.1, hold for the missing data case if, at update t, we make the replacements

$$\boldsymbol{y}_{(t)} = \begin{pmatrix} \boldsymbol{y}_t^{(1)} \\ \boldsymbol{0} \end{pmatrix}, \quad A_{(t)} = \begin{bmatrix} A_t^{(1)} \\ 0 \end{bmatrix}, \quad R_{(t)} = \begin{bmatrix} R_{11t} & 0 \\ 0 & R_{22t} \end{bmatrix}, \tag{6.78}$$

for $\boldsymbol{y}_t$, A_t, and R, respectively, in (6.21)–(6.23).

Once the "missing data" filtered values have been obtained, Stoffer (1982) also established the smoother values can be processed using Properties P6.2 and P6.3 with the values obtained from the missing data-filtered values. *The implication of these results is that, if $\boldsymbol{y}_t$ is incomplete, the filtered and smoothed estimators can be calculated from the usual equations by entering zeros in the observation vector when data are missing, by zeroing out the corresponding rows of the design matrix A_t, and by entering zeros in the off-diagonal elements of R that correspond to R_{12t} and R_{21t} at update t in the filter equation (6.23).* In doing this procedure, the state estimators are

$$\boldsymbol{x}_t^{(s)} = E\left(\boldsymbol{x}_t \mid \boldsymbol{y}_1^{(1)}, \ldots, \boldsymbol{y}_s^{(1)}\right), \tag{6.79}$$

with error variance–covariance matrix

$$P_t^{(s)} = E\left\{\left(\boldsymbol{x}_t - \boldsymbol{x}_t^{(s)}\right)\left(\boldsymbol{x}_t - \boldsymbol{x}_t^{(s)}\right)'\right\}. \tag{6.80}$$

The missing data lag-one smoother covariances will be denoted by $P_{t,t-1}^{(n)}$.

The maximum likelihood estimators, as computed in the EM procedure, must also be modified in the missing data case. Now, we consider

$$Y_n^{(1)} = \{\boldsymbol{y}_1^{(1)}, \ldots, \boldsymbol{y}_n^{(1)}\} \tag{6.81}$$

as the incomplete data, and X_n, Y_n, as defined in (6.63), as the complete data. In this case, the complete data likelihood, (6.63), or equivalently (6.64), is the same, but to implement the E-step, at iteration j, we must calculate

$$\begin{aligned}
Q\left(\Theta \mid \Theta^{(j-1)}\right) &= E\left\{-2\ln L_{X,Y}(\Theta) \mid Y_n^{(1)}, \Theta^{(j-1)}\right\} \\
&= E_*\left\{\ln|\Sigma_0| + \text{tr}\ \Sigma_0^{-1}(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)' \mid Y_n^{(1)}\right\} \\
&+ E_*\left\{n\ln|Q| + \sum_{t=1}^{n}\text{tr}\ \left[Q^{-1}(\boldsymbol{x}_t - \Phi\boldsymbol{x}_{t-1})(\boldsymbol{x}_t - \Phi\boldsymbol{x}_{t-1})'\right] \mid Y_n^{(1)}\right\} \\
&+ E_*\left\{n\ln|R| + \sum_{t=1}^{n}\text{tr}\ \left[R^{-1}(\boldsymbol{y}_t - A_t\boldsymbol{x}_t)(\boldsymbol{y}_t - A_t\boldsymbol{x}_t)'\right] \mid Y_n^{(1)}\right\},
\end{aligned} \tag{6.82}$$

where E_* denotes the conditional expectation under $\Theta^{(j-1)}$ and *tr* denotes trace. The first two terms in (6.82) will be like the first two terms of (6.66) with the smoothers $\boldsymbol{x}_t^n$, P_t^n, and $P_{t,t-1}^n$ replaced by their missing data counterparts, $\boldsymbol{x}_t^{(n)}$, $P_t^{(n)}$, and $P_{t,t-1}^{(n)}$. What changes in the missing data case is the third term of (6.82), where we must evaluate $E_*(\boldsymbol{y}_t^{(2)} \mid Y_n^{(1)})$ and $E_*(\boldsymbol{y}_t^{(2)}\boldsymbol{y}_t^{(2)'} \mid Y_n^{(1)})$. In Stoffer (1982), it is shown that

$$\begin{aligned}
&E_*\left\{(\boldsymbol{y}_t - A_t\boldsymbol{x}_t)(\boldsymbol{y}_t - A_t\boldsymbol{x}_t)' \mid Y_n^{(1)}\right\} \\
&= \begin{pmatrix} \boldsymbol{y}_t^{(1)} - A_t^{(1)}\boldsymbol{x}_t^{(n)} \\ R_{*21t}R_{*11t}^{-1}(\boldsymbol{y}_t^{(1)} - A_t^{(1)}\boldsymbol{x}_t^{(n)}) \end{pmatrix}\begin{pmatrix} \boldsymbol{y}_t^{(1)} - A_t^{(1)}\boldsymbol{x}_t^{(n)} \\ R_{*21t}R_{*11t}^{-1}(\boldsymbol{y}_t^{(1)} - A_t^{(1)}\boldsymbol{x}_t^{(n)}) \end{pmatrix}' \\
&+ \begin{pmatrix} A_t^{(1)} \\ R_{*21t}R_{*11t}^{-1}A_t^{(1)} \end{pmatrix} P_t^{(n)} \begin{pmatrix} A_t^{(1)} \\ R_{*21t}R_{*11t}^{-1}A_t^{(1)} \end{pmatrix}' \\
&+ \begin{pmatrix} 0 & 0 \\ 0 & R_{*22t} - R_{*21t}R_{*11t}^{-1}R_{*12t} \end{pmatrix}.
\end{aligned} \tag{6.83}$$

In (6.83), the values of R_{*ikt}, for $i, k = 1, 2$, are the current values specified by $\Theta^{(j-1)}$. In addition, $\boldsymbol{x}_t^{(n)}$ and $P_t^{(n)}$ are the values obtained by running the smoother under the current parameter estimates specified by $\Theta^{(j-1)}$.

In the case in which observed and unobserved components have uncorrelated errors, that is, R_{*12t} is the zero matrix, (6.83) can be simplified to

$$\begin{aligned} & E_* \left\{ (\boldsymbol{y}_t - A_t \boldsymbol{x}_t)(\boldsymbol{y}_t - A_t \boldsymbol{x}_t)' \mid Y_n^{(1)} \right\} \\ = & \left(\boldsymbol{y}_{(t)} - A_{(t)} \boldsymbol{x}_t^{(n)} \right) \left(\boldsymbol{y}_{(t)} - A_{(t)} \boldsymbol{x}_t^{(n)} \right)' + A_{(t)} P_t^{(n)} A_{(t)}' \\ + & \begin{pmatrix} 0 & 0 \\ 0 & R_{*22t} \end{pmatrix}, \end{aligned} \tag{6.84}$$

where $\boldsymbol{y}_{(t)}$ and $A_{(t)}$ are defined in (6.78).

In this simplified case, the "missing data" M-step looks like the M-step given in (6.67)-(6.73). That is, with

$$S_{(11)} = \sum_{t=1}^{n} (\boldsymbol{x}_t^{(n)} \boldsymbol{x}_t^{(n)\prime} + P_t^{(n)}), \tag{6.85}$$

$$S_{(10)} = \sum_{t=1}^{n} (\boldsymbol{x}_t^{(n)} \boldsymbol{x}_{t-1}^{(n)\prime} + P_{t,t-1}^{(n)}), \tag{6.86}$$

and

$$S_{(00)} = \sum_{t=1}^{n} (\boldsymbol{x}_{t-1}^{(n)} \boldsymbol{x}_{t-1}^{(n)\prime} + P_{t-1}^{(n)}), \tag{6.87}$$

where the smoothers are calculated under the present value of the parameters $\Theta^{(j-1)}$ using the missing data modifications, at iteration j, the *maximization step* is

$$\Phi^{(j)} = S_{(10)} S_{(00)}^{-1}, \tag{6.88}$$

$$Q^{(j)} = n^{-1} \left(S_{(11)} - S_{(10)} S_{(00)}^{-1} S_{(10)}' \right), \tag{6.89}$$

and

$$\begin{aligned} R^{(j)} = & n^{-1} \sum_{t=1}^{n} D_t \left\{ \left(\boldsymbol{y}_{(t)} - A_{(t)} \boldsymbol{x}_t^{(n)} \right) \left(\boldsymbol{y}_{(t)} - A_{(t)} \boldsymbol{x}_t^{(n)} \right)' + A_{(t)} P_t^{(n)} A_{(t)}' \right. \\ + & \left. \begin{pmatrix} 0 & 0 \\ 0 & R_{22t}^{(j-1)} \end{pmatrix} \right\} D_t', \end{aligned} \tag{6.90}$$

where D_t is a permutation matrix that reorders the variables at time t in their original order and $\boldsymbol{y}_{(t)}$ and $A_{(t)}$ are defined in (6.78). For example, suppose $q = 3$ and at time t, y_{t2} is missing. Then,

$$\boldsymbol{y}_{(t)} = \begin{pmatrix} y_{t1} \\ y_{t3} \\ 0 \end{pmatrix}, \quad A_{(t)} = \begin{bmatrix} A_{t1} \\ A_{t3} \\ \mathbf{0}' \end{bmatrix}, \quad \text{and} \quad D_t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

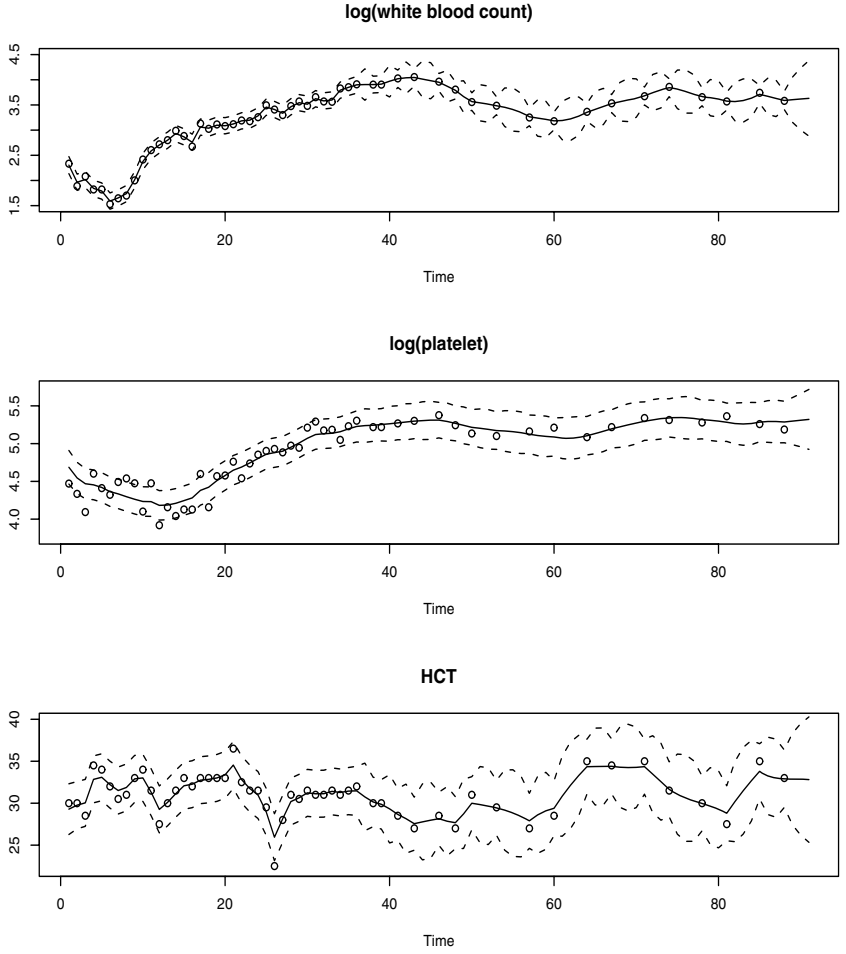


Figure 6.5 Smoothed values for various components in the blood parameter tracking problem. The actual data are shown as points, the smoothed values are shown as solid lines, and ± 3 standard error bounds are shown as dashed lines.

where A_{ti} is the ith row of A_t and $\mathbf{0}'$ is a $1 \times p$ vector of zeros. In (6.90), only R_{11t} gets updated, and R_{22t} at iteration j is simply set to its value from the previous iteration, $j-1$. Of course, if we cannot assume $R_{12t} = 0$, (6.90) must be changed accordingly using (6.83), but (6.88) and (6.89) remain the same. As before, the parameter estimates for the initial state are updated as

$$\boldsymbol{\mu}_0^{(j)} = \boldsymbol{x}_0^{(n)} \quad \text{and} \quad \Sigma_0^{(j)} = P_0^{(n)}. \tag{6.91}$$

Example 6.9 Longitudinal Biomedical Data

We consider the biomedical data in Example 6.1 which has portions of the three-dimensional vector missing after the 40th day. The maximum likelihood procedure yielded the estimators

$$\widehat{\Phi} = \begin{pmatrix} 1.02 & -.09 & .01 \\ .08 & .90 & .01 \\ -.90 & 1.42 & .87 \end{pmatrix}, \quad \widehat{Q} = \begin{pmatrix} .018 & .002 & .000 \\ .002 & .004 & .017 \\ .000 & .017 & 2.27 \end{pmatrix},$$

and $\widehat{R} = \text{diag}\{.004, .022, 1.69\}$ for the transition, state error covariance and observation error covariance matrices, respectively. The coupling between the first and second series is relatively weak, whereas the third series HCT is strongly related to the first two; that is,

$$\widehat{x}_{t3} = -.90x_{t-1,1} + 1.42x_{t-1,2} + .87x_{t-1,3}.$$

Hence, the HCT is negatively correlated with white blood count and positively correlated with platelet count. Byproducts of the procedure are estimated trajectories for all three longitudinal series and their respective prediction intervals. In particular, Figure 6.5 shows the data as points, the estimated smoothed values $\widehat{x}_t^{(n)}$ as solid lines, and error bounds, $\widehat{x}_t^{(n)} \pm 3\sqrt{\widehat{P}_t^{(n)}}$ as dotted lines, for critical post-transplant platelet count.

6.5 Structural Models: Signal Extraction and Forecasting

In order to develop computing techniques for handling a versatile cross section of possible models, it is necessary to restrict the state-space model somewhat, and we consider one possible class of specializations in this section. The components of the model are taken as linear processes that can be adapted to represent fixed and disturbed trends and periodicities as well as classical autoregressions. The observed series is regarded as being a sum of component signal series. To illustrate the possibilities, consider the economic example given below that shows how to fit a sum of trend, seasonal, and irregular components the quarterly earnings data that we have considered before.

Example 6.10 Johnson & Johnson Quarterly Earnings

Consider the quarterly earnings series from the U.S. company Johnson & Johnson as given in Figure 1.1. The series is highly nonstationary, and there is both a trend signal that is gradually increasing over time and a seasonal component that cycles every four quarters or once per year. The

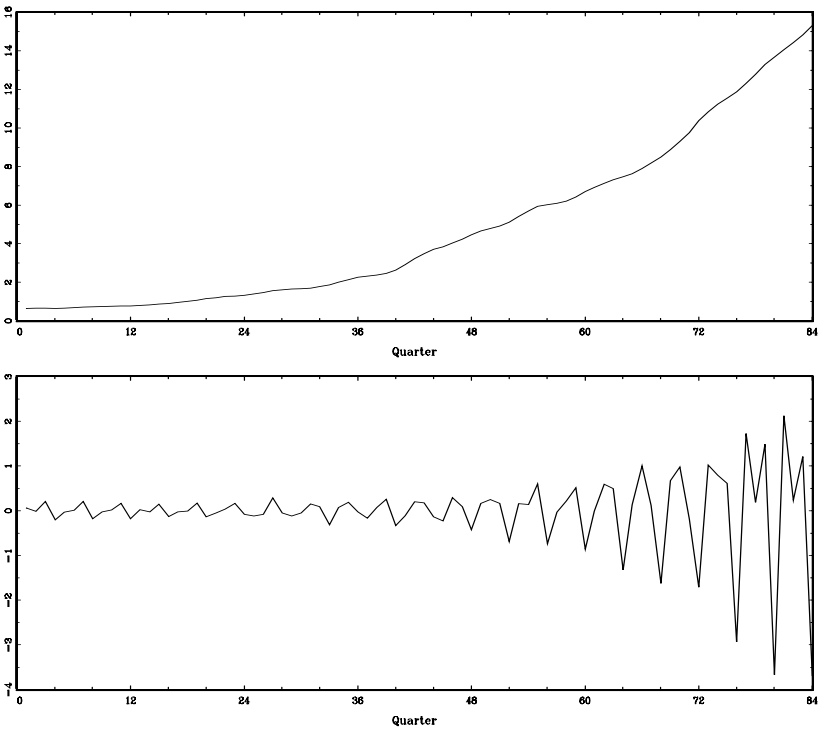


Figure 6.6 Estimated trend component, T_t^n (top), and estimated trend plus seasonal component, S_t^n (bottom), for the Johnson and Johnson quarterly earnings series.

seasonal component is getting larger over time as well. Transforming into logarithms or even taking the nth root does not seem to make the series stationary, as there is a slight bend to the transformed curve. Suppose, however, we consider the series to be the sum of a trend component, a seasonal component, and a white noise. That is, let the observed series be expressed as

$$y_t = T_t + S_t + v_t, \tag{6.92}$$

where T_t is trend and S_t is the seasonal component. Suppose we allow trend to increase exponentially; that is,

$$T_t = \phi T_{t-1} + w_{t1}, \tag{6.93}$$

where the coefficient $\phi > 1$ characterizes the increase. Let the seasonal component be modeled as

$$S_t + S_{t-1} + S_{t-2} + S_{t-3} = w_{t2}, \tag{6.94}$$

which corresponds to assuming the seasonal component is expected to sum to zero over a complete period or four quarters. To express this model in state-space form, let $\boldsymbol{x}_t' = (T_t, S_t, S_{t-1}, S_{t-2})$ be the state vector so the observation equation (6.2) can be written as

$$y_t = \begin{pmatrix} 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} T_t \\ S_t \\ S_{t-1} \\ S_{t-2} \end{pmatrix} + v_t,$$

with the state equation written as

$$\begin{pmatrix} T_t \\ S_t \\ S_{t-1} \\ S_{t-2} \end{pmatrix} = \begin{pmatrix} \phi & 0 & 0 & 0 \\ 0 & -1 & -1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} T_{t-1} \\ S_{t-1} \\ S_{t-2} \\ S_{t-3} \end{pmatrix} + \begin{pmatrix} w_{t1} \\ w_{t2} \\ 0 \\ 0 \end{pmatrix},$$

where $R = r_{11}$ and

$$Q = \begin{pmatrix} q_{11} & 0 & 0 & 0 \\ 0 & q_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The model reduces to state-space form, (6.1) and (6.2), with $p = 4$ and $q = 1$. The parameters to be estimated are r_{11}, the noise variance in the measurement equations, q_{11} and q_{22}, the model variances corresponding to the trend and seasonal components and ϕ, the transition parameter that models the growth rate. Growth is about 3% per year, and we began with $\phi = 1.03$. The initial mean was fixed at $\boldsymbol{\mu}_0 = (.5, .3, .2, .1)'$, with uncertainty modeled by the diagonal covariance matrix with $\Sigma_{0ii} = .01$, for $i = 1, \ldots, 4$. Initial state covariance values were taken as $q_{11} = .01, q_{22} = .10$, corresponding to relatively low uncertainty in the trend model compared with that in the seasonal model. The measurement error covariance was started at $r_{11} = .04$. After 70 iterations of the EM algorithm the transition parameter stabilized at $\widehat{\phi} = 1.035$, corresponding to exponential growth with inflation at about 3.5% per year. The measurement uncertainty was small at $\widehat{r}_{11} = .0086$, compared with the model uncertainties $\widehat{q}_{11} = .0169$ and $\widehat{q}_{22} = .0497$. From initial guesses, the trend uncertainty increased and the seasonal uncertainty decreased. Figure 6.6 shows the smoothed trend estimate and the exponentially increasing seasonal components. We may also consider forecasting the Johnson & Johnson series, and the result of a 12-quarter forecast is shown in Figure 6.7 as basically an extension of the latter part of the observed data.

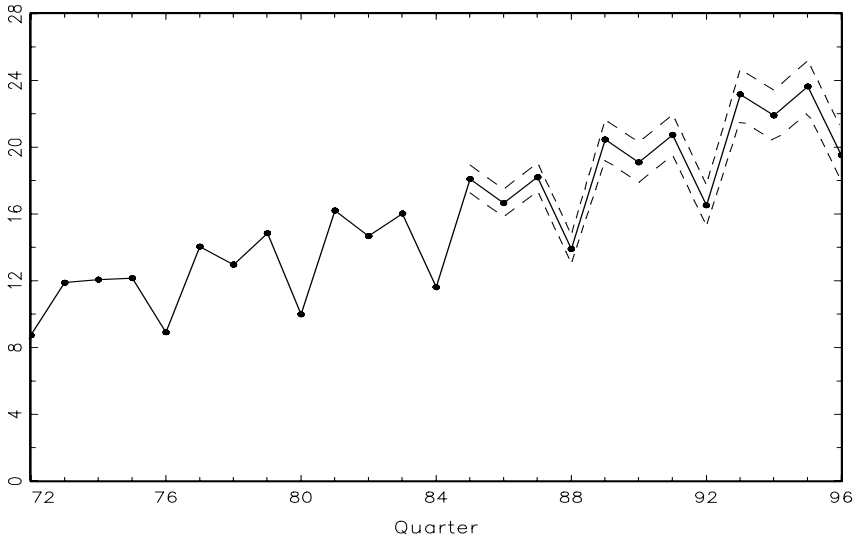


Figure 6.7 A 12-quarter forecast for the Johnson & Johnson quarterly earnings series. The last three years of data (quarters 72-84), are shown as points connected by a solid line. The forecasts are shown as points connected by a solid line (quarters 85-96) and dotted lines are upper and lower 95% prediction intervals.

6.6 ARMAX Models in State-Space Form

Sometimes, it is advantageous to write the state-space model in a slightly different way, as is done by numerous authors; for example, Anderson and Moore (1970) and Hannan and Deistler (1988). Here, we write the state-space model as

$$\boldsymbol{x}_{t+1} = \Phi \boldsymbol{x}_t + \Upsilon \boldsymbol{u}_t + \boldsymbol{w}_t \qquad t = 0, 1, \ldots, n \tag{6.95}$$

$$\boldsymbol{y}_t = A_t \boldsymbol{x}_t + \Gamma \boldsymbol{u}_t + \boldsymbol{v}_t \qquad t = 1, \ldots, n \tag{6.96}$$

where, in the state equation, $\boldsymbol{x}_0 \sim \mathrm{N}(\boldsymbol{\mu}_0, \Sigma_0)$, Φ is $p \times p$, and Υ is $p \times r$. In the observation equation, A_t is $q \times p$ and Γ is $q \times r$. Now, $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are still white noise series (both independent of $\boldsymbol{x}_0$), with $\mathrm{var}(\boldsymbol{w}_t) = Q$, $\mathrm{var}(\boldsymbol{v}_t) = R$, but we also allow the state noise and observation noise to be correlated at time t; that is,

$$\mathrm{cov}(\boldsymbol{w}_t, \boldsymbol{v}_t) = E(\boldsymbol{w}_t \boldsymbol{v}_t') = S \tag{6.97}$$

and zero otherwise; note, S is a $p \times q$ matrix. To obtain the innovations, $\boldsymbol{\epsilon}_t = \boldsymbol{y}_t - A_t \boldsymbol{x}_t^{t-1} - \Gamma \boldsymbol{u}_t$, and the innovation variance $\Sigma_t = A_t P_t^{t-1} A_t' + R$, in this case, we need the one-step-ahead state predictions. Of course, the filtered estimates will also be of interest, and they will be needed for smoothing. Property P6.2 (the smoother) as displayed in §6.2 still holds. The following

property generates the predictor $\boldsymbol{x}_{t+1}^t$ from the past predictor $\boldsymbol{x}_t^{t-1}$ when the noise terms are correlated and exhibits the filter update.

Property P6.5: The Kalman Filter with Correlated State and Measurement Noise

For the state-space model specified in (6.95) and (6.96), with initial conditions $\boldsymbol{x}_1^0$ *and* P_1^0*, for* $t = 1, \ldots, n$*,*

$$\boldsymbol{x}_{t+1}^t = \Phi \boldsymbol{x}_t^{t-1} + \Upsilon \boldsymbol{u}_t + K_t^*(\boldsymbol{y}_t - A_t \boldsymbol{x}_t^{t-1} - \Gamma \boldsymbol{u}_t), \tag{6.98}$$

$$P_{t+1}^t = [\Phi - K_t^* A_t] P_t^{t-1} [\Phi - K_t^* A_t]' + Q + K_t^* R K_t^{*'} - S K_t^{*'} - K_t^* S', \tag{6.99}$$

where the new gain matrix is given by

$$K_t^* = [\Phi P_t^{t-1} A_t' + S][[A_t P_t^{t-1} A_t' + R]^{-1}. \tag{6.100}$$

The filter update, given a new observation $\boldsymbol{y}_{t+1}$ *and input* $\boldsymbol{u}_{t+1}$ *is given by*

$$\boldsymbol{x}_{t+1}^{t+1} = \boldsymbol{x}_{t+1}^t + P_{t+1}^t A_{t+1}' \left[A_{t+1} P_{t+1}^t A_{t+1}' + R\right]^{-1} \boldsymbol{\epsilon}_{t+1}, \tag{6.101}$$

$$P_{t+1}^{t+1} = P_{t+1}^t - P_{t+1}^t A_{t+1}' \left[A_{t+1} P_{t+1}^t A_{t+1}' + R\right]^{-1} A_{t+1} P_{t+1}^t. \tag{6.102}$$

The derivation of Property P6.5 is similar to the derivation of the Kalman filter in Property P6.1 (Problem 6.17). Note, (6.101) and (6.102) are identical to (6.19) and (6.20).

Consider a p-dimensional ARMAX model given by,

$$\boldsymbol{y}_t = \Gamma \boldsymbol{u}_t + \sum_{j=1}^{m} \Phi_j \boldsymbol{y}_{t-j} + \sum_{k=1}^{q} \Theta_k \boldsymbol{v}_{t-k} + \boldsymbol{v}_t. \tag{6.103}$$

The Φs and Θs are $p \times p$ matrices, Γ is $p \times r$, and $\boldsymbol{v}_t$ is a $p \times 1$ white noise process; in fact, (6.103) and (5.84) are identical models, but here, we have written the observations as $\boldsymbol{y}_t$. We now have the following property.

Property P6.6: A State-Space Form of ARMAX

For $m \geq q$*, the state-space model given by*

$$\boldsymbol{x}_{t+1} = \begin{bmatrix} \Phi_1 & I & 0 & \cdots & 0 \\ \Phi_2 & 0 & I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Phi_{m-1} & 0 & 0 & \cdots & I \\ \Phi_m & 0 & 0 & \cdots & 0 \end{bmatrix} \boldsymbol{x}_t + \begin{bmatrix} \Theta_1 + \Phi_1 \\ \vdots \\ \Theta_q + \Phi_q \\ \Phi_{q+1} \\ \vdots \\ \Phi_m \end{bmatrix} \boldsymbol{v}_t, \tag{6.104}$$

$$\boldsymbol{y}_t = [\, I, 0, \cdots, 0 \,] \boldsymbol{x}_t + \Gamma \boldsymbol{u}_t + \boldsymbol{v}_t, \tag{6.105}$$

implies the ARMAX model (6.103). The state process, $\boldsymbol{x}_t$*, is* $pm \times 1$*, and the observations process* $\boldsymbol{y}_t$ *is* $p \times 1$*. If* $m < q$*, set* $\Phi_{m+1} = \cdots = \Phi_q = 0$*, in which case* $m = q$ *and (6.104)–(6.105) still apply.*

This form of the model is somewhat different than the form suggested in §5.1, equations (6.6)-(6.8). For example, in (6.8), by setting A_t equal to the $p \times p$ identity matrix (for all t) and setting $R = 0$ implies the data y_t in (6.8) follow a VAR(m) process. In doing so, however, we do not make use of the ability to allow for correlated state and observation error, so a singularity is introduced into the system in the form of $R = 0$. The method in Property P6.6 avoids that problem, and points out the fact that the same model can take many forms. We do not prove Property P6.6 directly, but the following example should suggest how to establish the general result.

Example 6.11 Univariate ARMA$(1,1)$ in State-Space Form

Consider the univariate ARMA(1, 1) model $y_t = \phi y_{t-1} + \theta v_{t-1} + v_t$. Using Property P6.6, we can write the model as

$$x_{t+1} = \phi x_t + w_t, \qquad \text{(state eqn)}, \tag{6.106}$$

where $w_t = (\theta + \phi)v_t$ and

$$y_t = x_t + v_t, \qquad \text{(obs eqn)}. \tag{6.107}$$

In this case, $\text{cov}(w_t, v_t) = (\theta + \phi)\text{var}(v_t) = (\theta + \phi)R$, and $\text{cov}(w_t, v_s) = 0$ when $s \neq t$, so Property P6.5 would apply. To verify (6.106) and (6.107) specify an ARMA(1,1) model, we have

$$\begin{aligned} y_t &= x_t + v_t && \text{from (6.107)} \\ &= \phi x_{t-1} + (\theta + \phi)v_{t-1} + v_t && \text{from (6.106)} \\ &= \phi(x_{t-1} + v_{t-1}) + \theta v_{t-1} + v_t \\ &= \phi y_{t-1} + \theta v_{t-1} + v_t, && \text{from (6.107).} \end{aligned}$$

Properties P6.5 and P6.6, together, can be used to accomplish maximum likelihood estimation for ARMAX models. In this case, the likelihood would be in the innovations form given in Chapter 2, equation (3.106), or equivalently (6.62), and estimation could be accomplished using Newton–Raphson or the EM algorithm as described §6.3.

6.7 Bootstrapping State-Space Models

Although, in §6.3, we discussed the fact that, under general conditions (which we assume to hold in this section), the MLEs of the parameters of a DLM are consistent and asymptotically normal, time series data are often of short or moderate length. Several researchers have found evidence that samples must be fairly large before asymptotic results are applicable (Dent and Min, 1978; Ansley and Newbold, 1980). Moreover, as we discussed in Example 3.31, problems occur if the parameters are near the boundary of the parameter

space. In this section, we discuss an algorithm for bootstrapping state-space models; this algorithm and its justification, including the non-Gaussian case, along with numerous examples, can be found in Stoffer and Wall (1991) and in Stoffer and Wall (2004). In view of §6.6, anything we do or say here about DLMs applies equally to ARMAX models.

Using the DLM given by (6.95)–(6.97) and Property P6.5, we write the innovations form of the filter as

$$\begin{aligned}
\boldsymbol{\epsilon}_t &= \boldsymbol{y}_t - A_t\boldsymbol{x}_t^{t-1} - \Gamma\boldsymbol{u}_t, && (6.108)\\
\Sigma_t &= A_tP_t^{t-1}A_t' + R, && (6.109)\\
\boldsymbol{x}_{t+1}^t &= \Phi\boldsymbol{x}_t^{t-1} + \Upsilon\boldsymbol{u}_t + K_t\boldsymbol{\epsilon}_t, && (6.110)\\
K_t &= [\Phi P_t^{t-1}A_t' + S]\Sigma_t^{-1}, && (6.111)\\
P_{t+1}^t &= \Phi P_t^{t-1}\Phi' + Q - K_t\Sigma_tK_t'. && (6.112)
\end{aligned}$$

This form of the filter is just a rearrangement of the filter given in Property P6.5; we have dropped the * in the new form of the gain matrix.

In addition, we can rewrite the model to obtain the innovations form of the model,

$$\begin{aligned}
\boldsymbol{x}_{t+1}^t &= \Phi\boldsymbol{x}_t^{t-1} + \Upsilon\boldsymbol{u}_t + K_t\boldsymbol{\epsilon}_t, && (6.113)\\
\boldsymbol{y}_t &= A_t\boldsymbol{x}_t^{t-1} + \Gamma\boldsymbol{u}_t + \boldsymbol{\epsilon}_t. && (6.114)
\end{aligned}$$

This form of the model is a rewriting of (6.108) and (6.110), and it accommodates the bootstrapping algorithm.

As discussed in Example 6.5, although the innovations $\boldsymbol{\epsilon}_t$ are uncorrelated, initially, Σ_t can be different for different time points t. Thus, in a resampling procedure, we can either ignore the first few values of $\boldsymbol{\epsilon}_t$ until Σ_t stabilizes or we can work with the standardized innovations

$$\boldsymbol{e}_t = \Sigma_t^{-1/2}\boldsymbol{\epsilon}_t, \tag{6.115}$$

so we are guaranteed these innovations have, at least, the same first two moments. In (6.115), $\Sigma_t^{1/2}$ denotes the unique square root matrix of Σ_t defined by $\Sigma_t^{1/2}\Sigma_t^{1/2} = \Sigma_t$. In what follows, we base the bootstrap procedure on the standardized innovations, but we stress the fact that, even in this case, ignoring startup values might be necessary, as noted by Stoffer and Wall (1991).

The model coefficients and the correlation structure of the model are uniquely parameterized by a $k\times 1$ parameter vector Θ_0; that is, $\Phi = \Phi(\Theta_0)$, $\Upsilon = \Upsilon(\Theta_0)$, $Q = Q(\Theta_0)$, $A_t = A_t(\Theta_0)$, $\Gamma = \Gamma(\Theta_0)$, and $R = R(\Theta_0)$. Recall the innovations form of the Gaussian likelihood (ignoring a constant) is

$$\begin{aligned}
-2\ln L_Y(\Theta) &= \sum_{t=1}^{n}\left[\ln|\Sigma_t(\Theta)| + \boldsymbol{\epsilon}_t(\Theta)'\Sigma_t(\Theta)^{-1}\boldsymbol{\epsilon}_t(\Theta)\right]\\
&= \sum_{t=1}^{n}\left[\ln|\Sigma_t(\Theta)| + \boldsymbol{e}_t(\Theta)'\boldsymbol{e}_t(\Theta)\right]. && (6.116)
\end{aligned}$$

We stress the fact that it is not necessary for the model to be Gaussian to consider (6.116) as the criterion function to be used for parameter estimation.

Let $\widehat{\Theta}$ denote the MLE of Θ_0, that is, $\widehat{\Theta} = \text{argmax}_\Theta L_Y(\Theta)$, obtained by the methods discussed in §6.3. Let $\boldsymbol{\epsilon}_t(\widehat{\Theta})$ and $\Sigma_t(\widehat{\Theta})$ be the innovation values obtained by running the filter, (6.108)–(6.112), under $\widehat{\Theta}$. Once this has been done, the bootstrap procedure is accomplished by the following steps.

1. Construct the standardized innovations
$$\boldsymbol{e}_t(\widehat{\Theta}) = \Sigma_t^{-1/2}(\widehat{\Theta})\boldsymbol{\epsilon}_t(\widehat{\Theta}).$$

2. Sample, with replacement, n times from the set $\{\boldsymbol{e}_1(\widehat{\Theta}), \ldots, \boldsymbol{e}_n(\widehat{\Theta})\}$ to obtain $\{\boldsymbol{e}_1^*(\widehat{\Theta}), \ldots, \boldsymbol{e}_n^*(\widehat{\Theta})\}$, a bootstrap sample of standardized innovations.

3. Construct a bootstrap data set $\{\boldsymbol{y}_1^*, \ldots, \boldsymbol{y}_n^*\}$ as follows. Define the $(p+q)\times 1$ vector $\boldsymbol{\xi}_t = (\boldsymbol{x}_{t+1}^{t'}, \boldsymbol{y}_t')'$. Stacking (6.113) and (6.114) results in a vector first-order equation for $\boldsymbol{\xi}_t$ given by
$$\boldsymbol{\xi}_t = F_t\boldsymbol{\xi}_{t-1} + G\boldsymbol{u}_t + H_t\boldsymbol{e}_t, \tag{6.117}$$
where
$$F_t = \begin{bmatrix} \Phi & 0 \\ A_t & 0 \end{bmatrix}, \quad G = \begin{bmatrix} \Upsilon \\ \Gamma \end{bmatrix}, \quad H_t = \begin{bmatrix} K_t\Sigma_t^{-1/2} \\ \Sigma_t^{-1/2} \end{bmatrix}.$$
Thus, to construct the bootstrap data set, solve (6.117) using $\boldsymbol{e}_t^*(\widehat{\Theta})$ in place of $\boldsymbol{e}_t$. The exogenous variables $\boldsymbol{u}_t$ and the initial conditions of the Kalman filter remain fixed at their given values, and the parameter vector is held fixed at $\widehat{\Theta}$.

4. Using the bootstrap data set $\{\boldsymbol{y}_t^*; t = 1, \ldots, n\}$, construct a likelihood, $L_{Y^*}(\Theta)$, and obtain the MLE of Θ, say, $\widehat{\Theta}^*$.

5. Repeat steps 2 through 4, a large number, B, of times, obtaining a bootstrapped set of parameter estimates $\{\widehat{\Theta}_b^*;\ b = 1, \ldots, B\}$. The finite sample distribution of $\widehat{\Theta} - \Theta_0$ may be approximated by the distribution of $\widehat{\Theta}_b^* - \widehat{\Theta}$, $b = 1, \ldots, B$.

In the next example, we discuss the case of a linear regression model, but where the regression coefficients are stochastic and allowed to vary with time. The state-space model provides a convenient setting for the analysis of such models.

Example 6.12 Stochastic Regression

Figure 6.8 shows the interest rate recorded for three-month treasury bills (line–squares), y_t, and the quarterly inflation rate (dotted line–circles) in

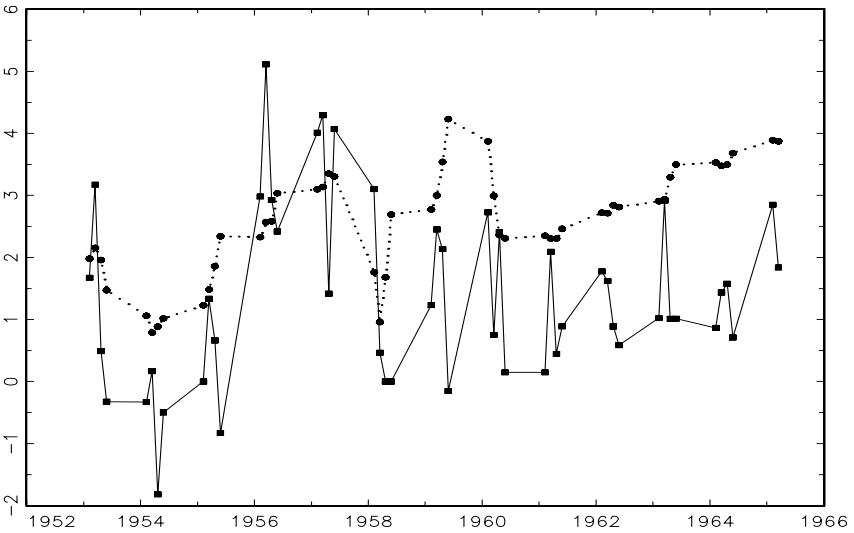


Figure 6.8 Interest rate for three-month treasury bills (line–squares) and quarterly inflation rate (dotted line–circles) in the Consumer Price Index, 1953:1 to 1965:2.

the Consumer Price Index, z_t, from the first quarter of 1953 through the second quarter of 1965, $n = 50$ observations. These data were analyzed by Newbold and Bos (1985, pp. 61-73).

In this analysis, the treasury bill interest rate is modeled as being linearly related to quarterly inflation as

$$y_t = \alpha + \beta_t z_t + v_t,$$

where α is a fixed constant, β_t is a stochastic regression coefficient, and v_t is white noise with variance σ_v^2. The stochastic regression term, which comprises the state variable, is specified by a first-order autoregression,

$$(\beta_t - b) = \phi(\beta_{t-1} - b) + w_t,$$

where b is a constant, and w_t is white noise with variance σ_w^2. The noise processes, v_t and w_t, are assumed to be uncorrelated.

Using the notation of the state-space model (6.95) and (6.96), we have in the state equation, $\boldsymbol{x}_t = \beta_t$, $\Phi = \phi$, $\boldsymbol{u}_t \equiv 1$, $\Upsilon = (1 - \phi)b$, $Q = \sigma_w^2$, and in the observation equation, $A_t = z_t$, $\Gamma = \alpha$, $R = \sigma_v^2$, and $S = 0$. The parameter vector is $\Theta = (\phi, \alpha, b, \sigma_w, \sigma_v)'$. The results of the Newton–Raphson estimation procedure are listed in Table 6.2. Also shown in the Table 6.2 are the corresponding standard errors obtained from $B = 500$ runs of the bootstrap. These standard errors are simply

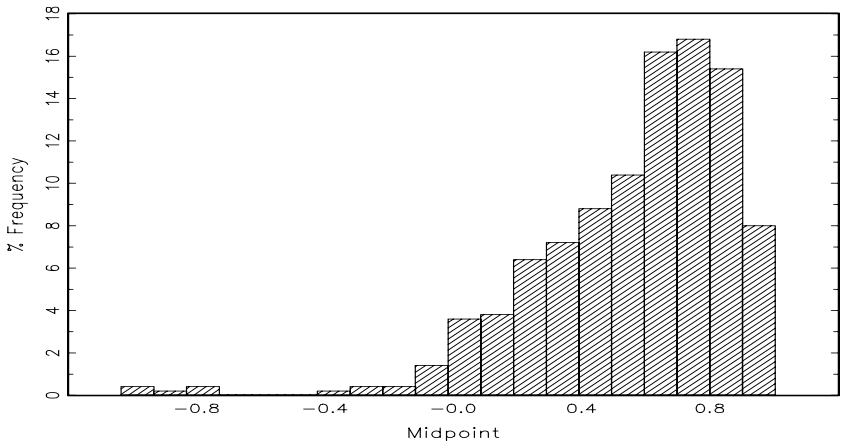


Figure 6.9 Bootstrap distribution, $B = 500$, of the estimator of ϕ.

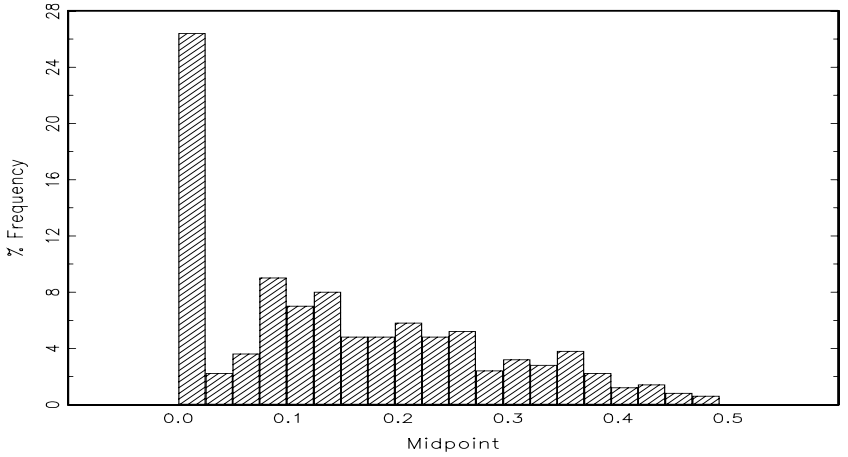


Figure 6.10 Bootstrap distribution, $B = 500$, of the estimator of σ_w.

the standard deviations of the bootstrapped estimates, that is, the square root of $\sum_{b=1}^{B}(\Theta_{ib}^{*} - \bar{\Theta}_{i}^{*})^2/(B-1)$, where Θ_i, represents the ith parameter, $i = 1, \ldots, 5$, and $\bar{\Theta}_{i}^{*} = \sum_{b=1}^{B} \Theta_{ib}^{*}/B$.

The asymptotic standard errors listed in Table 6.2 are typically smaller than those obtained from the bootstrap. This result is the most pronounced in the estimates of ϕ, σ_w, and σ_v, where the bootstrapped standard errors are about 50% larger than the corresponding asymptotic value. Also, asymptotic theory prescribes the use of normal theory when dealing with the parameter estimates. The bootstrap, however, allows us to investigate the small sample distribution of the estimators and, hence, provides more insight into the data analysis.

Table 6.2 Comparison of Asymptotic Standard Errors and Bootstrapped Standard Errors ($B = 500$)

Parameter	MLE	Asymptotic Standard Error	Bootstrap Standard Error
ϕ	.841	.200	.304
α	−.771	.645	.645
b	.855	.278	.277
σ_w	.127	.092	.182
σ_v	1.131	.142	.217

For example, Figure 6.9 shows the bootstrap distribution of the estimator of ϕ. This distribution is highly skewed with values concentrated around .8, but with a long tail to the left. Some quantiles of the bootstrapped distribution of ϕ are -.09 (2.5%), .03 (5%), .16 (10%), .87 (90%), .92 (95%), .94 (97.5%), and they can be used to obtain confidence intervals. For example, a 90% confidence interval for ϕ would be approximated by (.03, .92). This interval is rather wide, and we will interpret this after we discuss the results of the estimation of σ_w.

Figure 6.10 shows the bootstrap distribution of $\widehat{\sigma}_w$. The distribution is concentrated at two locations, one at approximately $\widehat{\sigma}_w = .15$ and the other at $\widehat{\sigma}_w = 0$. The cases in which $\widehat{\sigma}_w \approx 0$ correspond to deterministic state dynamics. When $\sigma_w = 0$ and $|\phi| < 1$, then $\beta_t \approx b$ for large t, so the approximately 25% of the cases in which $\widehat{\sigma}_w \approx 0$ suggest a fixed state, or constant coefficient model. The cases in which $\widehat{\sigma}_w$ is away from zero would suggest a truly stochastic regression parameter. To investigate this matter further, Figure 6.11 shows the joint bootstrapped estimates, $(\widehat{\phi}, \widehat{\sigma}_w)$, for positive values of $\widehat{\phi}$. The joint distribution suggests $\widehat{\sigma}_w > 0$ corresponds to $\widehat{\phi} \approx 0$. When $\phi = 0$, the state dynamics are given by $\beta_t = b + w_t$. If, in addition, σ_w is small relative to b, the system is nearly deterministic; that is, $\beta_t \approx b$. Considering these results, the bootstrap analysis leads us to conclude the dynamics of the data are best described in terms of a fixed regression effect.

6.8 Dynamic Linear Models with Switching

The problem of modeling changes in regimes for vector-valued time series has been of interest in many different fields. In §5.4, we explored the idea that the dynamics of the system of interest might change over the course of time. In Example 5.5, we saw that pneumonia and influenza mortality rates behave differently when a flu epidemic occurs than when no epidemic occurs. As another example, some authors (for example, Hamilton, 1989, or McCulloch and Tsay,

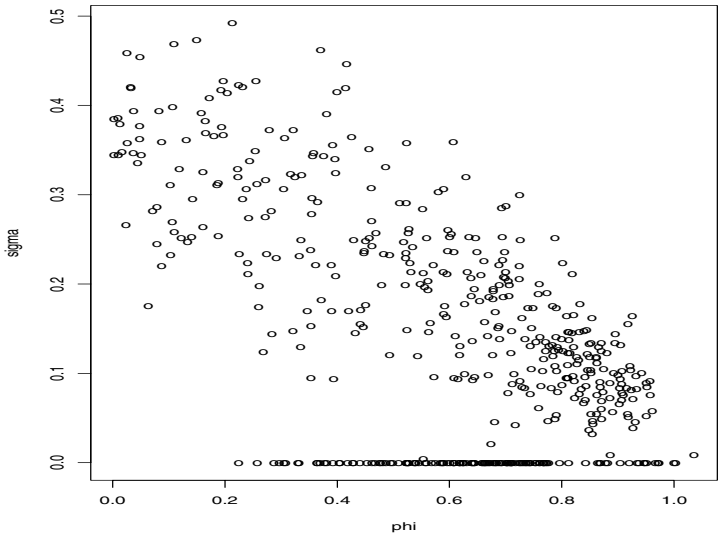


Figure 6.11 Joint bootstrap distribution, $B = 500$, of the estimators of ϕ and σ_w. Only the values corresponding to $\widehat{\phi}^* \geq 0$ are shown.

1993) have explored the possibility the dynamics of the quarterly U.S. GNP series (say, y_t) analyzed in Example 3.33 might be different during expansion ($\nabla \log y_t > 0$) than during contraction ($\nabla \log y_t < 0$). In this section, we will concentrate on the method presented in Shumway and Stoffer (1991). One way of modeling change in an evolving time series is by assuming the dynamics of some underlying model changes discontinuously at certain undetermined points in time. Our starting point is the DLM given by (6.1) and (6.2), namely,

$$\boldsymbol{x}_t = \Phi \boldsymbol{x}_{t-1} + \boldsymbol{w}_t, \tag{6.118}$$

to describe the $p \times 1$ state dynamics, and

$$\boldsymbol{y}_t = A_t \boldsymbol{x}_t + \boldsymbol{v}_t \tag{6.119}$$

to describe the $q \times 1$ observation dynamics. Recall $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are Gaussian white noise sequences with $\text{var}(\boldsymbol{w}_t) = Q$, $\text{var}(\boldsymbol{v}_t) = R$, and $\text{cov}(\boldsymbol{w}_t, \boldsymbol{v}_s) = 0$ for all s and t.

Generalizations of (6.118) and (6.119) to include the possibility of changes occurring over time have been approached by allowing changes in the error covariances (Harrison and Stevens, 1976, Gordon and Smith, 1988, 1990) or by assigning mixture distributions to the observation errors $\boldsymbol{v}_t$ (Peña and Guttman, 1988). Approximations to filtering were derived in all of the aforementioned articles. An application to monitoring renal transplants was described in Smith and West (1983) and in Gordon and Smith (1990). Changes can also be modeled in the classical regression case by allowing switches in the design matrix, as in Quandt (1972).

Switching via a stationary Markov chain with independent observations has been developed by Lindgren (1978) and Goldfeld and Quandt (1973). In the Markov chain approach, we declare the dynamics of the system at time t is generated by one of m possible regimes evolving according to a Markov chain over time. As a simple example, suppose the dynamics of a univariate time series, y_t, is generated by either the model (1) $y_t = \beta_1 y_{t-1} + w_t$ or the model (2) $y_t = \beta_2 y_{t-1} + w_t$. We will write the model as $y_t = \phi_t y_{t-1} + w_t$ such that $\Pr(\phi_t = \beta_j) = \pi_j$, $j = 1, 2$, $\pi_1 + \pi_2 = 1$, and with the Markov property

$$\Pr(\phi_t = \beta_j \mid \phi_{t-1} = \beta_i, \phi_{t-2} = \beta_{i_2}, \ldots) = \Pr(\phi_t = \beta_j \mid \phi_{t-1} = \beta_i) = \pi_{ij},$$

for $i, j = 1, 2$ (and $i_2, \ldots = 1, 2$). As previously mentioned, Markov switching for dependent data has been applied by Hamilton (1989) to detect changes between positive and negative growth periods in the economy. Applications to speech recognition have been considered by Juang and Rabiner (1985). The case in which the particular regime is unknown to the observer comes under the heading of hidden Markov models, and the techniques related to analyzing these models are summarized in Rabiner and Juang (1986). An application of the idea of switching to the tracking of multiple targets has been considered in Bar-Shalom (1978), who obtained approximations to Kalman filtering in terms of weighted averages of the innovations.

Example 6.13 Tracking Multiple Targets

The approach of Shumway and Stoffer (1991) was motivated primarily by the problem of tracking a large number of moving targets using a vector $\boldsymbol{y}_t$ of sensors. In this problem, we do not know at any given point in time which target any given sensor has detected. Hence, it is the structure of the measurement matrix A_t in (6.119) that is changing, and not the dynamics of the signal $\boldsymbol{x}_t$ or the noises, $\boldsymbol{w}_t$ or $\boldsymbol{v}_t$. As an example, consider a 3×1 vector of satellite measurements $\boldsymbol{y}_t = (y_{t1}, y_{t2}, y_{t3})'$ that are observations on some combination of a 3×1 vector of targets or signals, $\boldsymbol{x}_t = (x_{t1}, x_{t2}, x_{t3})'$. For the measurement matrix

$$A_t = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

in the model (6.119), all sensors are observing the first target, x_{t1}, whereas for the measurement matrix

$$A_t = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

the first sensor, y_{t1}, observes the second target, x_{t2}; the second sensor, y_{t2}, observes the first target, x_{t1}; and the third sensor, y_{t3}, observes the

third target, x_{t3}. All possible detection configurations will define a set of possible values for A_t, say, $\{M_1, M_2, \ldots, M_m\}$, as a collection of plausible measurement matrices.

Example 6.14 Modeling Economic Change

As another example of the switching model presented in this section, consider the case in which the dynamics of the linear model changes suddenly over the history of a given realization. For example, Lam (1990) has given the following generalization of Hamilton's (1989) model for detecting positive and negative growth periods in the economy. Suppose the data are generated by

$$y_t = z_t + n_t, \tag{6.120}$$

where z_t is an autoregressive series and n_t is a random walk with a drift that switches between two values α_0 and $\alpha_0 + \alpha_1$. That is,

$$n_t = n_{t-1} + \alpha_0 + \alpha_1 S_t, \tag{6.121}$$

with $S_t = 0$ or 1, depending on whether the system is in state 1 or state 2. For the purpose of illustration, suppose

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + w_t \tag{6.122}$$

is an AR(2) series with $\text{var}(w_t) = \sigma_w^2$. Lam (1990) wrote (6.120) in a differenced form

$$\nabla y_t = z_t - z_{t-1} + \alpha_0 + \alpha_1 S_t, \tag{6.123}$$

which we may take as the observation equation (6.119) with state vector

$$\boldsymbol{x}_t = (z_t, z_{t-1}, \alpha_0, \alpha_1)' \tag{6.124}$$

and

$$M_1 = [1, -1, 1, 0] \quad \text{and} \quad M_2 = [1, -1, 1, 1] \tag{6.125}$$

determining the two possible economic conditions. The state equation, (6.118), is of the form

$$\begin{pmatrix} z_t \\ z_{t-1} \\ \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} z_{t-1} \\ z_{t-2} \\ \alpha_0 \\ \alpha_1 \end{pmatrix} + \begin{pmatrix} w_t \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{6.126}$$

The observation equation, (6.119), in this case is

$$\nabla y_t = A_t \boldsymbol{x}_t + v_t, \tag{6.127}$$

where $\Pr(A_t = M_1) = 1 - \Pr(A_t = M_2)$, with M_1 and M_2 given in (6.125).

To incorporate a reasonable switching structure for the measurement matrix into the DLM that is compatible with both practical situations previously described, we assume that the m possible configurations are states in a nonstationary, independent process defined by the time-varying probabilities

$$\pi_j(t) = \Pr(A_t = M_j), \tag{6.128}$$

for $j = 1, \ldots, m$ and $t = 1, 2, \ldots, n$. Important information about the current state of the measurement process is given by the filtered probabilities of being in state j, defined as the conditional probabilities

$$\pi_j(t|t) = \Pr(A_t = M_j|Y_t), \tag{6.129}$$

which also vary as a function of time. In (6.129), we have used the notation $Y_s = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s\}$. The filtered probabilities (6.129) give the time-varying estimates of the probability of being in state j given the data to time t.

It will be important for us to obtain estimators of the configuration probabilities, $\pi_j(t|t)$, the predicted and filtered state estimators, $\boldsymbol{x}_t^{t-1}$ and $\boldsymbol{x}_t^t$, and the corresponding error covariance matrices P_t^{t-1} and P_t^t. Of course, the predictor and filter estimators will depend on the parameters, Θ, of the DLM. In many situations, the parameters will be unknown and we will have to estimate them. Our focus will be on maximum likelihood estimation, but other authors have taken a Bayesian approach that assigns priors to the parameters, and then seeks posterior distributions of the model parameters; see, for example, Gordon and Smith (1990), Peña and Guttman (1988), or McCulloch and Tsay (1993).

We now establish the recursions for the filters associated with the state $\boldsymbol{x}_t$ and the switching process, A_t. As discussed in §6.3, the filters are also an essential part of the maximum likelihood procedure. The predictors, $\boldsymbol{x}_t^{t-1} = E(\boldsymbol{x}_t|Y_{t-1})$, and filters, $\boldsymbol{x}_t^t = E(\boldsymbol{x}_t|Y_t)$, and their associated error variance–covariance matrices, P_t^{t-1} and P_t^t, are given by

$$\boldsymbol{x}_t^{t-1} = \Phi \boldsymbol{x}_{t-1}^{t-1}, \tag{6.130}$$

$$P_t^{t-1} = \Phi P_{t-1}^{t-1} \Phi' + Q, \tag{6.131}$$

$$\boldsymbol{x}_t^t = \boldsymbol{x}_t^{t-1} + \sum_{j=1}^{m} \pi_j(t|t) K_{tj} \boldsymbol{\epsilon}_{tj}, \tag{6.132}$$

$$P_t^t = \sum_{j=1}^{m} \pi_j(t|t)(I - K_{tj} M_j) P_t^{t-1}, \tag{6.133}$$

$$K_{tj} = P_t^{t-1} M_j' \Sigma_{tj}^{-1}, \tag{6.134}$$

where the innovation values in (6.132) and (6.134) are

$$\boldsymbol{\epsilon}_{tj} = \boldsymbol{y}_t - M_j \boldsymbol{x}_t^{t-1}, \tag{6.135}$$

$$\Sigma_{tj} = M_j P_t^{t-1} M_j' + R, \tag{6.136}$$

for $j = 1, \ldots, m$.

Equations (6.130)-(6.134) exhibit the filter values as weighted linear combinations of the m innovation values, (6.135)-(6.136), corresponding to each of the possible measurement matrices. The equations are similar to the approximations introduced by Bar-Shalom and Tse (1975), by Gordon and Smith (1990), and Peña and Guttman (1988).

To verify (6.132), let the indicator $I(A_t = M_j) = 1$ when $A_t = M_j$, and zero otherwise. Then, using (6.21),

$$\begin{aligned}
\boldsymbol{x}_t^t &= E(\boldsymbol{x}_t|Y_t) = E[E(\boldsymbol{x}_t|Y_t, A_t) \mid Y_t] \\
&= E\left\{\sum_{j=1}^{m} E(\boldsymbol{x}_t|Y_t, A_t = M_j) I(A_t = M_j) \mid Y_t\right\} \\
&= E\left\{\sum_{j=1}^{m} [\boldsymbol{x}_t^{t-1} + K_{tj}(\boldsymbol{y}_t - M_j \boldsymbol{x}_t^{t-1})] I(A_t = M_j) \mid Y_t\right\} \\
&= \sum_{j=1}^{m} \pi_j(t|t)[\boldsymbol{x}_t^{t-1} + K_{tj}(\boldsymbol{y}_t - M_j \boldsymbol{x}_t^{t-1})],
\end{aligned}$$

where K_{tj} is given by (6.134). Equation (6.133) is derived in a similar fashion; the other relationships, (6.130), (6.131), and (6.134), follow from straightforward applications of the Kalman filter results given in Property P6.1.

Next, we derive the filters $\pi_j(t|t)$. Let $f_j(t|t-1)$ denote the conditional density of $\boldsymbol{y}_t$ given the past $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}$, and $A_t = M_j$, for $j = 1, \ldots, m$. Then,

$$\pi_j(t|t) = \frac{\pi_j(t) f_j(t|t-1)}{\sum_{k=1}^{m} \pi_k(t) f_k(t|t-1)}, \tag{6.137}$$

where we assume the distribution $\pi_j(t)$, for $j = 1, \ldots, m$ has been specified before observing $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t$ (details follow as in Example 6.15 below). If the investigator has no reason to prefer one state over another at time t, the choice of uniform priors, $\pi_j(t) = m^{-1}$, for $j = 1, \ldots, m$, will suffice. Smoothness can be introduced by letting

$$\pi_j(t) = \sum_{i=1}^{m} \pi_i(t-1|t-1)\pi_{ij}, \tag{6.138}$$

where the non-negative weights π_{ij} are chosen so $\sum_{i=1}^{m} \pi_{ij} = 1$. If the A_t process was Markov with transition probabilities π_{ij}, then (6.138) would be the update for the filter probability, as shown in the next example.

Example 6.15 Hidden Markov Chain Model

If $\{A_t\}$ is a hidden Markov chain with stationary transition probabilities $\pi_{ij} = \Pr(A_t = M_j | A_{t-1} = M_i)$, for $i, j = 1, \ldots, m$, letting $p(\cdot)$ denote a generic probability function, we have

$$\begin{aligned}
\pi_j(t|t) &= \frac{p(A_t = M_j, \boldsymbol{y}_t, Y_{t-1})}{p(\boldsymbol{y}_t, Y_{t-1})} \\
&= \frac{p(Y_{t-1})p(A_t = M_j \mid Y_{t-1})p(\boldsymbol{y}_t \mid A_t = M_j, Y_{t-1})}{p(Y_{t-1})p(\boldsymbol{y}_t \mid Y_{t-1})} \\
&= \frac{\pi_j(t|t-1)f_j(t|t-1)}{\sum_{k=1}^{m} \pi_k(t|t-1)f_k(t|t-1)}. \qquad (6.139)
\end{aligned}$$

In the Markov case, the conditional probabilities

$$\pi_j(t|t-1) = \Pr(A_t = M_j \big| Y_{t-1})$$

in (6.139) replace the unconditional probabilities, $\pi_j(t) = \Pr(A_t = M_j)$, in (6.137).

To evaluate (6.139), we must be able to calculate $\pi_j(t|t-1)$ and $f_j(t|t-1)$. We will discuss the calculation of $f_j(t|t-1)$ after this example. To derive $\pi_j(t|t-1)$, note,

$$\begin{aligned}
\pi_j(t|t-1) &= \Pr(A_t = M_j \big| Y_{t-1}) \\
&= \sum_{i=1}^{m} \Pr(A_t = M_j, A_{t-1} = M_i \big| Y_{t-1}) \\
&= \sum_{i=1}^{m} \Pr(A_t = M_j \big| A_{t-1} = M_i) \Pr(A_{t-1} = M_i \big| Y_{t-1}) \\
&= \sum_{i=1}^{m} \pi_{ij}\pi_i(t-1|t-1). \qquad (6.140)
\end{aligned}$$

Expression (6.138) comes from equation (6.140), where, as previously noted, we replace $\pi_j(t|t-1)$ by $\pi_j(t)$.

The difficulty in extending the approach here to the Markov case is the dependence among the $\boldsymbol{y}_t$, which makes it necessary to enumerate over all possible histories to derive the filtering equations. This problem will be evident when we derive the conditional density $f_j(t|t-1)$. Equation (6.138) has $\pi_j(t)$ as a function of the past observations, Y_{t-1}, which is inconsistent with our model assumption. Nevertheless, this seems to be a reasonable compromise that allows the data to modify the probabilities $\pi_j(t)$, without having to develop a highly computer-intensive technique.

As previously suggested, the computation of $f_j(t|t-1)$, without some approximations, is highly computer-intensive. To evaluate $f_j(t|t-1)$, consider the event

$$A_1 = M_{j_1}, \ \ldots \ , A_{t-1} = M_{j_{t-1}}, \qquad (6.141)$$

for $j_i = 1, \ldots, m$, and $i = 1, \ldots, t-1$, which specifies a specific set of measurement matrices through the past; we will write this event as $A_{(t-1)} = M_{(\ell)}$. Because m^{t-1} possible outcomes exist for $A_1, \ldots, A_{t-1}$, the index ℓ runs through $\ell = 1, \ldots, m^{t-1}$. Using this notation, we may write

$$\begin{aligned} & f_j(t|t-1) \\ &= \sum_{\ell=1}^{m^{t-1}} \Pr\{A_{(t-1)} = M_{(\ell)} \mid Y_{t-1}\} f(\boldsymbol{y}_t \mid Y_{t-1}, A_t = M_j, A_{(t-1)} = M_{(\ell)}) \\ &\equiv \sum_{\ell=1}^{m^{t-1}} \alpha(\ell)\ \mathrm{N}\left(\boldsymbol{y}_t \;\middle|\; \boldsymbol{\mu}_{tj}(\ell), \Sigma_{tj}(\ell)\right), \quad j = 1, \ldots, m, \end{aligned} \tag{6.142}$$

where the notation $\mathrm{N}(\cdot \mid \boldsymbol{b}, B)$ represents the normal density with mean vector $\boldsymbol{b}$ and variance–covariance matrix B. That is, $f_j(t|t-1)$ is a mixture of normals with non-negative weights $\alpha(\ell) = \Pr\{A_{(t-1)} = M_{(\ell)} \mid Y_{t-1}\}$ such that $\sum_\ell \alpha(\ell) = 1$, and with each normal distribution having mean vector

$$\boldsymbol{\mu}_{tj}(\ell) = M_j \boldsymbol{x}_t^{t-1}(\ell) = M_j E[\boldsymbol{x}_t \mid Y_{t-1}, A_{(t-1)} = M_{(\ell)}] \tag{6.143}$$

and covariance matrix

$$\Sigma_{tj}(\ell) = M_j P_t^{t-1}(\ell) M_j' + R. \tag{6.144}$$

This result follows because the conditional distribution of $\boldsymbol{y}_t$ in (6.142) is identical to the fixed measurement matrix case presented in Section 4.2. The values in (6.143) and (6.144), and hence the densities, $f_j(t|t-1)$, for $j = 1, \ldots, m$, can be obtained directly from the Kalman filter, Property P6.1, with the measurement matrices $A_{(t-1)}$ fixed at $M_{(\ell)}$.

Although $f_j(t|t-1)$ is given explicitly in (6.142), its evaluation is highly computer intensive. For example, with $m = 2$ states and $n = 20$ observations, we have to filter over $2 + 2^2 + \cdots + 2^{20}$ possible sample paths; note, $2^{20} = 1,048,576$. One remedy is to trim (remove), at each t, highly improbable sample paths; that is, remove events in (6.141) with extremely small probability of occurring, and then evaluate $f_j(t|t-1)$ as if the trimmed sample paths could not have occurred. Another alternative, as suggested by Gordon and Smith (1990) and Shumway and Stoffer (1991), is to approximate $f_j(t|t-1)$ using the closest (in the sense of Kulback–Leibler distance) normal distribution. In this case, the approximation leads to choosing normal distribution with the same mean and variance associated with $f_j(t|t-1)$; that is, we approximate $f_j(t|t-1)$ by a normal with mean $M_j \boldsymbol{x}_t^{t-1}$ and variance Σ_{tj} given in (6.136).

To develop a procedure for maximum likelihood estimation, the joint density of the data is

$$\begin{aligned} f(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) &= \prod_{t=1}^{n} f(\boldsymbol{y}_t | Y_{t-1}) \\ &= \prod_{t=1}^{n} \sum_{j=1}^{m} \Pr(A_t = M_j | Y_{t-1}) f(\boldsymbol{y}_t | A_t = M_j, Y_{t-1}), \end{aligned}$$

and hence, the likelihood can be written as

$$\ln L_Y(\Theta) = \sum_{t=1}^{n} \ln \left(\sum_{j=1}^{m} \pi_j(t) f_j(t|t-1) \right). \tag{6.145}$$

For the hidden Markov model, $\pi_j(t)$ would be replaced by $\pi_j(t|t-1)$. In (6.145), we will use the normal approximation to $f_j(t|t-1)$. That is, henceforth, we will consider $f_j(t|t-1)$ as the normal, $\mathrm{N}(M_j \boldsymbol{x}_t^{t-1}, \Sigma_{tj})$, density, where $\boldsymbol{x}_t^{t-1}$ is given in (6.130) and Σ_{tj} is given in (6.136). We may consider maximizing (6.145) directly as a function of the parameters $\Theta = \{\boldsymbol{\mu}_0, \Phi, Q, R\}$ using a Newton method, or we may consider applying the EM algorithm to the complete data likelihood.

To apply the EM algorithm as in §6.3, we call $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, $A_1, \ldots, A_n$, and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, the complete data, with likelihood given by

$$\begin{aligned} -2 \ln L_{X,A,Y}(\Theta) &= \ln |\Sigma_0| + (\boldsymbol{x}_0 - \boldsymbol{\mu}_0)' \Sigma_0^{-1} (\boldsymbol{x}_0 - \boldsymbol{\mu}_0) \\ &\quad + n \ln |Q| + \sum_{t=1}^{n} (\boldsymbol{x}_t - \Phi \boldsymbol{x}_{t-1})' Q^{-1} (\boldsymbol{x}_t - \Phi \boldsymbol{x}_{t-1}) \\ &\quad - 2 \sum_{t=1}^{n} \sum_{j=1}^{m} I(A_t = M_j) \ln \pi_j(t) + n \ln |R| \\ &\quad + \sum_{t=1}^{n} \sum_{j=1}^{m} I(A_t = M_j) (\boldsymbol{y}_t - A_t \boldsymbol{x}_t)' R^{-1} (\boldsymbol{y}_t - A_t \boldsymbol{x}_t). \end{aligned} \tag{6.146}$$

As discussed in §6.3, we require the minimization of the conditional expectation

$$Q\left(\Theta \mid \Theta^{(k-1)}\right) = E\left\{ -2 \ln L_{X,A,Y}(\Theta) \mid Y_n, \Theta^{(k-1)} \right\}, \tag{6.147}$$

with respect to Θ at each iteration, $k = 1, 2, \ldots$. The calculation and maximization of (6.147) is similar to the case of (6.65). In particular, with

$$\pi_j(t|n) = E[I(A_t = M_j) \mid Y_n], \tag{6.148}$$

we obtain on iteration k,

$$\pi_j^{(k)}(t) = \pi_j(t|n), \tag{6.149}$$

$$\boldsymbol{\mu}_0^{(k)} = \boldsymbol{x}_0^n, \tag{6.150}$$

$$\Phi^{(k)} = S_{10} S_{00}^{-1}, \tag{6.151}$$

$$Q^{(k)} = n^{-1} \left(S_{11} - S_{10} S_{00}^{-1} S_{10}' \right), \tag{6.152}$$

and

$$R^{(k)} = n^{-1} \sum_{t=1}^{n} \sum_{j=1}^{m} \pi_j(t|n) \left[(\boldsymbol{y}_t - M_j \boldsymbol{x}_t^n)(\boldsymbol{y}_t - M_j \boldsymbol{x}_t^n)' + M_j P_t^n M_j' \right]. \tag{6.153}$$

where S_{11}, S_{10}, S_{00} are given in (6.67)-(6.69). As before, at iteration k, the filters and the smoothers are calculated using the current values of the parameters, $\Theta^{(k-1)}$, and Σ_0 is held fixed. Filtering is accomplished by using (6.130)-(6.134). Smoothing is derived in a similar manner to the derivation of the filter, and one is led to the smoother given in Property P6.2 and P6.3, with one exception, the initial smoother covariance, (6.55), is now

$$P^n_{n,n-1} = \sum_{j=1}^{m} \pi_j(n|n)(I - K_{tj}M_j)\Phi P^{n-1}_{n-1}. \tag{6.154}$$

Unfortunately, the computation of $\pi_j(t|n)$ is excessively complicated, and requires integrating over mixtures of normal distributions. Shumway and Stoffer (1991) suggest approximating the smoother $\pi_j(t|n)$ by the filter $\pi_j(t|t)$, and find the approximation works well.

Example 6.16 Analysis of Influenza Data

We use the results of this section to analyze the U.S. monthly pneumonia and influenza mortality data presented in §5.4, Figure 5.7. Letting y_t denote the mortality caused by pneumonia and influenza at month t, we model y_t in terms of a structural component model coupled with a hidden Markov process that determines whether a flu epidemic exists.

The model consists of three structural components. The first component, x_{t1}, is an AR(2) process chosen to represent the periodic (seasonal) component of the data,

$$x_{t1} = \alpha_1 x_{t-1,1} + \alpha_2 x_{t-2,1} + w_{t1}, \tag{6.155}$$

where w_{t1} is white noise, with $\text{var}(w_{t1}) = \sigma_1^2$. The second component, x_{t2}, is an AR(1) process with a nonzero constant term, which is chosen to represent the sharp rise in the data during an epidemic,

$$x_{t2} = \beta_0 + \beta_1 x_{t-1,2} + w_{t2}, \tag{6.156}$$

where w_{t2} is white noise, with $\text{var}(w_{t2}) = \sigma_2^2$. The third component, x_{t3}, is a fixed trend component given by,

$$x_{t3} = x_{t-1,3} + w_{t3}, \tag{6.157}$$

where $\text{var}(w_{t3}) = 0$. The case in which $\text{var}(w_{t3}) > 0$, which corresponds to a stochastic trend (random walk), was tried here, but the estimation became unstable, and lead to us fitting a fixed, rather than stochastic, trend. Thus, in the final model, the trend component satisfies $\nabla x_{t3} = 0$; recall in Example 2.42 the data were also differenced once before fitting the model.

Table 6.3 Estimation Results for Influenza Data

Parameter	Initial Estimate	Final Estimate
α_1	1.401 (.079)	1.379 (.073)
α_2	−.618 (.091)	−.575 (.075)
β_0	.162 (.042)	.201 (.028)
β_1	.156 (.142)	—
σ_1	.023 (.001)	.023 (.001)
σ_2	.105 (.015)	.108 (.016)
σ_v	.000 (.032)	—

Estimated standard errors are shown in parentheses.

Throughout the years, periods of normal influenza mortality (state 1) are modeled as

$$y_t = x_{t1} + x_{t3} + v_t, \tag{6.158}$$

where the measurement error, v_t, is white noise with $\text{var}(v_t) = \sigma_v^2$. When an epidemic occurs (state 2), mortality is modeled as

$$y_t = x_{t1} + x_{t2} + x_{t3} + v_t. \tag{6.159}$$

The model specified in (6.155)–(6.159) can be written in the general state-space form. The state equation is

$$\begin{pmatrix} x_{t1} \\ x_{t-1,1} \\ x_{t2} \\ x_{t3} \end{pmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & \beta_1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} x_{t-1,1} \\ x_{t-2,1} \\ x_{t-1,2} \\ x_{t-1,3} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \beta_0 \\ 0 \end{pmatrix} + \begin{pmatrix} w_{t1} \\ 0 \\ w_{t2} \\ 0 \end{pmatrix}. \tag{6.160}$$

Of course, (6.160) can be written in the standard state-equation form as

$$\boldsymbol{x}_t = \Phi \boldsymbol{x}_{t-1} + \Gamma u_t + \boldsymbol{w}_t, \tag{6.161}$$

where $\boldsymbol{x}_t = (x_{t1}, x_{t-1,1}, x_{t2}, x_{t3})'$, $\Gamma = (0, 0, \beta_0, 0)'$, $u_t \equiv 1$, and Q is a 4×4 matrix with σ_1^2 as the (1,1)-element, σ_2^2 as the (3,3)-element, and the remaining elements set equal to zero. The observation equation is

$$y_t = A_t \boldsymbol{x}_t + v_t, \tag{6.162}$$

where A_t is 1×4, and v_t is white noise with $\text{var}(v_t) = R = \sigma_v^2$. We assume all components of variance w_{t1}, w_{t2}, and v_t are uncorrelated.

As discussed in (6.158) and (6.159), A_t can take one of two possible forms

$$\begin{aligned} A_t &= M_1 = [1, 0, 0, 1] \quad \text{no epidemic,} \\ A_t &= M_2 = [1, 0, 1, 1] \quad \text{epidemic,} \end{aligned}$$

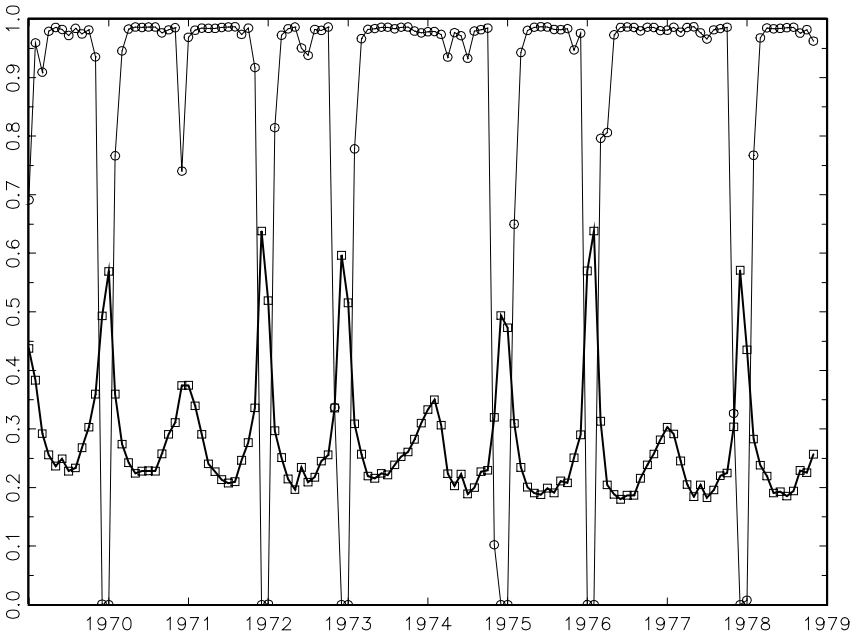


Figure 6.12 Influenza data, y_t, (dark line–squares) and the predicted probability that no epidemic occurs in month t given the past, $\widehat{\pi}_1(t|t-1)$ (line–circles) for the ten-year period 1969-1978; 1968 is not shown.

corresponding to the two possible states of (1) no flu epidemic and (2) flu epidemic, such that $\Pr(A_t = M_1) = 1 - \Pr(A_t = M_2)$. In this example, we will assume A_t is a hidden Markov chain, and hence we use the updating equations given in Example 6.15, (6.139) and (6.140), with transition probabilities $\pi_{11} = \pi_{22} = .75$ (and, thus, $\pi_{12} = \pi_{21} = .25$).

Parameter estimation was accomplished using a quasi-Newton–Raphson procedure to maximize the approximate log likelihood given in (6.145), with initial values of $\pi_1(1|0) = \pi_2(1|0) = .5$. Table 6.3 shows the results of the estimation procedure. On the initial fit, two estimates are not significant, namely, $\widehat{\beta}_1$ and $\widehat{\sigma}_v$. When $\sigma_v^2 = 0$, there is no measurement error, and the variability in data is explained solely by the variance components of the state system, namely, σ_1^2 and σ_2^2. The case in which $\beta_1 = 0$ corresponds to a simple level shift during a flu epidemic. In the final model, with β_1 and σ_v^2 removed, the estimated level shift ($\widehat{\beta}_0$) corresponds to an increase in mortality by about .2 per 1000 during a flu epidemic. The estimates for the final model are also listed in Table 6.3.

Figure 6.12 shows a plot of the data, y_t, for the ten-year period of 1969-1978 as well as the estimated approximate conditional probabili-

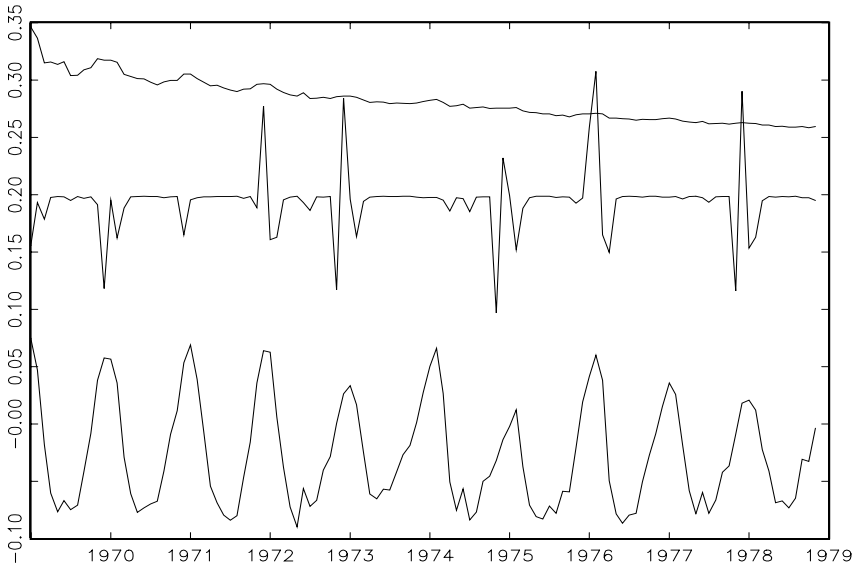


Figure 6.13 The three filtered structural components of influenza mortality: $\widehat{x}_{t1}^t$ (*cyclic trace*), $\widehat{x}_{t2}^t$ (*spiked trace*), and $\widehat{x}_{t3}^t$ (*negative linear trace*) for the ten-year period 1969-1978.

ties $\widehat{\pi}_1(t|t-1)$, that is, the predicted probability no epidemic occurs in month t given the past, $y_1, \ldots, y_{t-1}$. The results for the first year of the data, 1968, are not included in the figure because of initial instabilities of the filter. Of course, the estimated predicted probability a flu epidemic will occur next month is $\widehat{\pi}_2(t|t-1) = 1 - \widehat{\pi}_1(t|t-1)$. Thus, a good estimator would have small values of $\widehat{\pi}_1(t|t-1)$ corresponding to peaks in y_t. Except for initial values where instability exists, the estimated prediction probabilities are right on the mark. That is, the predicted probability of a flu epidemic exceeds the probability of no epidemic when indeed a flu epidemic occurred the next month.

Figure 6.13 shows the estimated filtered values (that is, filtering is done using the parameter estimates) of the three components of the model, x_{t1}^t, x_{t2}^t, and x_{t3}^t. Except for initial instability (which is not shown), $\widehat{x}_{t1}^t$ represents the seasonal (cyclic) aspect of the data, $\widehat{x}_{t2}^t$ represents the spikes during a flu epidemic, and $\widehat{x}_{t3}^t$ represents the slow decline in flu mortality over the ten-year period of 1969-1978.

One-month-ahead prediction, say, $\widehat{y}_t^{t-1}$, is obtained as follows,

$$\widehat{y}_t^{t-1} = M_1 \widehat{\boldsymbol{x}}_t^{t-1} \quad \text{if} \quad \widehat{\pi}_1(t|t-1) > \widehat{\pi}_2(t|t-1),$$

$$\widehat{y}_t^{t-1} = M_2 \widehat{\boldsymbol{x}}_t^{t-1} \quad \text{if} \quad \widehat{\pi}_1(t|t-1) \leq \widehat{\pi}_2(t|t-1).$$

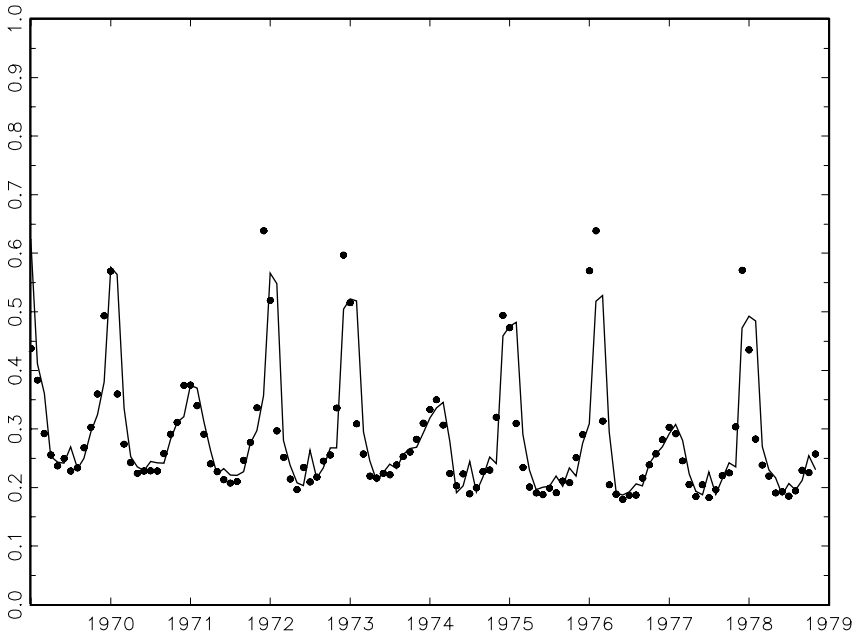


Figure 6.14 One-month-ahead prediction, $\widehat{y}_t^{t-1}$ (line), of the number of deaths caused by pneumonia and influenza, y_t (points) for 1969-1978. The standard error of the prediction is .02 when a flu epidemic is not predicted, and .11 when a flu epidemic is predicted.

Of course, $\widehat{\boldsymbol{x}}_t^{t-1}$ is the estimated state prediction, obtained via the filter presented in (6.130)-(6.134) (with the addition of the constant term in the model) using the estimated parameters. The results are shown in Figure 6.14. The precision of the forecasts can be measured by the innovation variances, Σ_{t1} when no epidemic is predicted, and Σ_{t2} when an epidemic is predicted. These values become stable quickly, and when no epidemic is predicted, the estimated standard error of the prediction is approximately .02 (this is the square root of Σ_{t1} for t large); when a flu epidemic is predicted, the estimated standard error of the prediction is approximately .11.

The results of this analysis are impressive given the small number of parameters and the degree of approximation that was made to obtain a computationally simple method for fitting a complex model. In particular, as seen in Figure 6.12, the model is never fooled as to when a flu epidemic will occur. This result is particularly impressive, given that, for example, in the third year, around $t = 36$, it appeared as though an epidemic was about to begin, but it never was realized, and the model

predicted no flu epidemic that year. As seen in Figure 6.14, the predicted mortality tends to be underestimated during the peaks, but the true values are typically within one standard error of the predicted value. Further evidence of the strength of this technique can be found in the example given in Shumway and Stoffer (1991).

6.9 Nonlinear and Non-normal State-Space Models Using Monte Carlo Methods

Most of this chapter has focused on linear dynamic models assumed to be Gaussian processes. Historically, these models were convenient because analyzing the data was a relatively simple matter. These assumptions cannot cover every situation, and it is advantageous to explore departures from these assumptions. As seen in §6.8, the solution to the nonlinear and non-Gaussian case will require computer-intensive techniques currently in vogue because of the availability of cheap and fast computers. In this section, we take a Bayesian approach to forecasting as our main objective; see West and Harrison (1997) for a detailed account of Bayesian forecasting with dynamic models. Prior to the mid-1980s, a number of approximation methods were developed to filter non-normal or nonlinear processes in an attempt to circumvent the computational complexity of the analysis of such models. For example, the extended Kalman filter and the Gaussian sum filter (Alspach and Sorensen, 1972) are two such methods described in detail in Anderson and Moore (1979). As in the previous section, these techniques typically rely on approximating the non-normal distribution by one or several Gaussian distributions or by some other parametric function.

With the advent of cheap and fast computing, a number of authors developed computer-intensive methods based on numerical integration. For example, Kitagawa (1987) proposed a numerical method based on piecewise linear approximations to the density functions for prediction, filtering, and smoothing for non-Gaussian and nonstationary state-space models. Pole and West (1988) used Gaussian quadrature techniques in a Bayesian analysis of nonlinear dynamic models; West and Harrison (1997, Chapter 13) provide a detailed explanation of these and similar methods. Markov chain Monte Carlo (MCMC) methods refer to Monte Carlo integration methods that use a Markovian updating scheme. We will describe the method in more detail later. The most common MCMC method is the Gibbs sampler, which is essentially a modification of the Metropolis algorithm (Metropolis et al., 1953) developed by Hastings (1970) in the statistical setting and by Geman and Geman (1984) in the context of image restoration. Later, Tanner and Wong (1987) used the ideas in their substitution sampling approach, and Gelfand and Smith (1990) developed the Gibbs sampler for a wide class of parametric models. This technique

was first used by Carlin et al. (1992) in the context of general nonlinear and non-Gaussian state-space models. Frühwirth-Schnatter (1994) and Carter and Kohn (1994) built on these ideas to develop efficient Gibbs sampling schemes for more restrictive models.

If the model is linear, that is, (6.1) and (6.2) hold, but the distributions are not Gaussian, a non-Gaussian likelihood can be defined by (6.31) in §6.2, but where $f_0(\cdot)$, $f_w(\cdot)$ and $f_v(\cdot)$ are not normal densities. In this case, prediction and filtering can be accomplished using numerical integration techniques (e.g., Kitagawa, 1987; Pole and West, 1988) or Monte Carlo techniques (e.g. Frühwirth-Schnatter, 1994; Carter and Kohn, 1994) to evaluate (6.32) and (6.33). Of course, the prediction and filter densities $p_\Theta(\boldsymbol{x}_t \mid Y_{t-1})$ and $p_\Theta(\boldsymbol{x}_t \mid Y_t)$ will no longer be Gaussian and will not generally be of the location-scale form as in the Gaussian case. A rich class of non-normal densities is given in (6.173).

In general, the state-space model can be given by the following equations:

$$\boldsymbol{x}_t = F_t(\boldsymbol{x}_{t-1}, \boldsymbol{w}_t) \quad \text{and} \quad \boldsymbol{y}_t = H_t(\boldsymbol{x}_t, \boldsymbol{v}_t), \tag{6.163}$$

where F_t and H_t are known functions that may depend on parameters Θ and $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are white noise processes. The main component of the model retained by (6.163) is that the states are Markov, and the observations are conditionally independent, but we do not necessarily assume F_t and H_t are linear, or $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are Gaussian. Of course, if $F_t(\boldsymbol{x}_{t-1}, \boldsymbol{w}_t) = \Phi_t \boldsymbol{x}_{t-1} + \boldsymbol{w}_t$ and $H_t(\boldsymbol{x}_t, \boldsymbol{v}_t) = A_t \boldsymbol{x}_t + \boldsymbol{v}_t$ and $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are Gaussian, we have the standard DLM (exogenous variables can be added to the model in the usual way). In the general model, (6.163), the likelihood is given by

$$L_{X,Y}(\Theta) = p_\Theta(\boldsymbol{x}_0) \prod_{t=1}^{n} p_\Theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) p_\Theta(\boldsymbol{y}_t \mid \boldsymbol{x}_t), \tag{6.164}$$

and the prediction and filter densities, as given by (6.32) and (6.33) in Section 4.2, still hold.

Because our focus is on simulation using MCMC methods, we first describe the technique in a general context.

Example 6.17 MCMC Techniques and the Gibbs Sampler

The goal of a Monte Carlo technique, of course, is to simulate a pseudo-random sample of vectors from a desired density function $p_\Theta(\boldsymbol{z})$. In Markov chain Monte Carlo, we simulate an ordered sequence of pseudo-random vectors, $\boldsymbol{z}_0 \mapsto \boldsymbol{z}_1 \mapsto \boldsymbol{z}_2 \mapsto \cdots$ by specifying a starting value, $\boldsymbol{z}_0$ and then sampling successive values from a transition density $\pi(\boldsymbol{z}_t | \boldsymbol{z}_{t-1})$, for $t = 1, 2, \ldots$. In this way, conditional on $\boldsymbol{z}_{t-1}$, the t-th pseudo-random vector, $\boldsymbol{z}_t$, is simulated independent of its predecessors. This technique alone does not yield a pseudo-random sample because contiguous draws are dependent on each other (that is, we obtain a first-order dependent

sequence of pseudo-random vectors). If done appropriately, the dependence between the pseudo-variates $\boldsymbol{z}_t$ and $\boldsymbol{z}_{t+m}$ decays exponentially in m, and we may regard the collection $\{\boldsymbol{z}_{t+\ell m}; \ell = 1, 2, \ldots\}$ for t and m suitably large, as a pseudo-random sample. Alternately, one may repeat the process in parallel, retaining the m-th value, on run $g = 1, 2, \ldots$, say, $\boldsymbol{z}_m^{(g)}$, for large m. Under general conditions, the Markov chain converges in the sense that, eventually, the sequence of pseudo-variates appear stationary and the individual $\boldsymbol{z}_t$ are marginally distributed according to the stationary "target" density $p_\Theta(\boldsymbol{z})$. Technical details may be found in Tierney (1994).

For Gibbs sampling, suppose we have a collection $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k\}$ of random vectors with complete conditional densities denoted generically by

$$p_\Theta(\boldsymbol{z}_j \mid \boldsymbol{z}_i, i \neq j) \equiv p_\Theta(\boldsymbol{z}_j \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{j-1}, \boldsymbol{z}_{j+1}, \ldots, \boldsymbol{z}_k),$$

for $j = 1, \ldots, k$, available for sampling. Here, available means pseudo-samples may be generated by some method given the values of the appropriate conditioning random vectors. Under mild conditions, these complete conditionals uniquely determine the full joint density $p_\Theta(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)$ and, consequently, all marginals, $p_\Theta(\boldsymbol{z}_j)$ for $j = 1, \ldots, k$; details may be found in Besag (1974). The Gibbs sampler generates pseudo-samples from the joint distribution as follows. Start with an arbitrary set of starting values, say, $\{\boldsymbol{z}_{1[0]}, \ldots, \boldsymbol{z}_{k[0]}\}$. Draw $\boldsymbol{z}_{1[1]}$ from $p_\Theta(\boldsymbol{z}_1 | \boldsymbol{z}_{2[0]}, \ldots, \boldsymbol{z}_{k[0]})$, then draw $\boldsymbol{z}_{2[1]}$ from $p_\Theta(\boldsymbol{z}_2 | \boldsymbol{z}_{1[1]}, \boldsymbol{z}_{3[0]}, \ldots, \boldsymbol{z}_{k[0]})$, and so on up to $\boldsymbol{z}_{k[1]}$ from $p_\Theta(\boldsymbol{z}_k | \boldsymbol{z}_{1[1]}, \ldots, \boldsymbol{z}_{k-1[1]})$, to complete one iteration. After ℓ such iterations, we have the collection $\{\boldsymbol{z}_{1[\ell]}, \ldots, \boldsymbol{z}_{k[\ell]}\}$. Geman and Geman (1984) showed that under mild conditions, $\{\boldsymbol{z}_{1[\ell]}, \ldots, \boldsymbol{z}_{k[\ell]}\}$ converges ($\ell \to \infty$) in distribution to a random observation from $p_\Theta(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k)$. For this reason, we typically drop the subscript $[\ell]$ from the notation, assuming ℓ is sufficiently large for the generated sample to be thought of as a realization from the joint density; hence, we denote this first realization as $\{\boldsymbol{z}_{1[\ell]}^{(1)}, \ldots, \boldsymbol{z}_{k[\ell]}^{(1)}\} \equiv \{\boldsymbol{z}_1^{(1)}, \ldots, \boldsymbol{z}_k^{(1)}\}$. This entire process is replicated in parallel, a large number, G, of times providing pseudo-random iid collections $\{\boldsymbol{z}_1^{(g)}, \ldots, \boldsymbol{z}_k^{(g)}\}$, for $g = 1, \ldots, G$ from the joint distribution. These simulated values can the be used to estimate the marginal densities. In particular, if $p_\Theta(\boldsymbol{z}_j | \boldsymbol{z}_i, i \neq j)$ is available in closed form, then

$$\widehat{p}_\Theta(\boldsymbol{z}_j) = G^{-1} \sum_{g=1}^{G} p_\Theta(\boldsymbol{z}_j \mid \boldsymbol{z}_i^{(g)}, i \neq j). \tag{6.165}$$

Approximation (6.165) is based on the fact that, for random variables x and y with joint density $p(x, y)$, the marginal density of x is obtained as follows: $p(x) = \int p(x, y)dy = \int p(x|y)p(y)dy$. Because of the relatively recent appearance of Gibbs sampling methodology, several important theoretical and practical issues are under investigation. These issues

include the diagnosis of convergence, modification of the sampling order, efficient estimation, and sequential sampling schemes (as opposed to the parallel processing described above) to mention a few. At this time, the best advice can be obtained from the texts by Gelman et al. (1995) and Gilks et al. (1996), and we are certain that many more will follow.

Finally, it may be necessary to nest rejection sampling within the Gibbs sampling procedure. The need for rejection sampling arises when we want to sample from a density, say, $f(\boldsymbol{z})$, but $f(\boldsymbol{z})$ is known only up to a proportionality constant, say, $p(\boldsymbol{z}) \propto f(\boldsymbol{z})$. If a density $g(\boldsymbol{z})$ is available, and there is a constant c for which $p(\boldsymbol{z}) \leq cg(\boldsymbol{z})$ for all $\boldsymbol{z}$, the rejection algorithm generates pseudo-variates from $f(\boldsymbol{z})$ by generating a value, $\boldsymbol{z}^*$ from $g(\boldsymbol{z})$ and accepting it as a value from $f(\boldsymbol{z})$ with probability $\pi(\boldsymbol{z}^*) = p(\boldsymbol{z}^*)/[cg(\boldsymbol{z}^*)]$. This algorithm can be quite inefficient if $\pi(\cdot)$ is close to zero; in such cases, more sophisticated envelope functions may be needed. Further discussion of these matters in the case of nonlinear state-space models can be found in Carlin et al. (1992, Examples 1.2 and 3.2).

In Example 6.17, the generic random vectors $\boldsymbol{z}_j$ can represent parameter values, such as components of Θ, state values $\boldsymbol{x}_t$, or future observations $\boldsymbol{y}_{n+m}$, for $m \geq 1$. This will become evident in the following examples. Before discussing the general case of nonlinear and non-normal state-space models, we briefly introduce MCMC methods for the Gaussian DLM, as presented in Frühwirth-Schnatter (1994) and Carter and Kohn (1994).

Example 6.18 Assessing Model Parameters for the Gaussian DLM

Consider the Gaussian DLM given by

$$\boldsymbol{x}_t = \Phi_t \boldsymbol{x}_{t-1} + \boldsymbol{w}_t \quad \text{and} \quad y_t = \boldsymbol{a}_t' \boldsymbol{x}_t + v_t. \tag{6.166}$$

The observations are univariate, and the state process is p-dimensional; this DLM includes the structural models presented in §6.5. The prior on the initial state is $\boldsymbol{x}_0 \sim \mathrm{N}(\boldsymbol{\mu}_0, \Sigma_0)$, and we assume that $\boldsymbol{w}_t \sim$ iid $\mathrm{N}(\boldsymbol{0}, Q_t)$, independent of $v_t \sim$ iid $\mathrm{N}(0, r_t)$. The collection of unknown model parameters will be denoted by Θ.

To explore how we would assess the values of Θ using an MCMC technique, we focus on the problem obtaining the posterior distribution, $p(\Theta \mid Y_n)$, of the parameters given the data, $Y_n = \{y_1, \ldots, y_n\}$ and a prior $\pi(\Theta)$. Of course, these distributions depend on "hyperparameters" that are assumed to be known. (Some authors consider the states $\boldsymbol{x}_t$ as the first level of parameters because they are unobserved. In this case, the values in Θ are regarded as the hyperparameters, and the parameters of their distributions are regarded as hyper-hyperparameters.) Denoting

the entire set of state vectors as $X_n = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, the posterior can be written as

$$p(\Theta \mid Y_n) = \int p(\Theta \mid X_n, Y_n)\; p(X_n, \Theta^* \mid Y_n)\; dX_n\; d\Theta^*. \tag{6.167}$$

Although the posterior, $p(\Theta \mid Y_n)$, may be intractable, conditioning on the states can make the problem manageable in that

$$p(\Theta \mid X_n, Y_n) \propto \pi(\Theta)\; p(x_0 \mid \Theta) \prod_{t=1}^{n}\; p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \Theta)\; p(y_t \mid \boldsymbol{x}_t, \Theta) \tag{6.168}$$

can be easier to work with (either as members of conjugate families or using some rejection scheme); we will discuss this in more detail when we present the nonlinear, non-Gaussian case, but we will assume for the present $p(\Theta \mid X_n, Y_n)$ is in closed form.

Suppose we can obtain G pseudo-random draws, $X_n^{(g)} \equiv (X_n, \Theta^*)^{(g)}$, for $g = 1, \ldots, G$, from the joint posterior density $p(X_n, \Theta^* \mid Y_n)$. Then (6.167) can be approximated by

$$\widehat{p}(\Theta \mid Y_n) = G^{-1} \sum_{g=1}^{G} p(\Theta \mid X_n^{(g)}, Y_n).$$

A sample from $p(X_n, \Theta^* \mid Y_n)$ is obtained using two different MCMC methods. First, the Gibbs sampler is used, for each g, as follows: sample $X_{n[\ell]}$ given $\Theta^*_{[\ell-1]}$ from $p(X_n \mid \Theta^*_{[\ell-1]}, Y_n)$, and then a sample $\Theta^*_{[\ell]}$ from $p(\Theta \mid X_{n[\ell]}, Y_n)$ as given by (6.168), for $\ell = 1, 2, \ldots$. Stop when ℓ is sufficiently large, and retain the final values as $X_n^{(g)}$. This process is repeated G times.

The first step of this method requires simultaneous generation of the state vectors. Because we are dealing with a Gaussian linear model, we can rely on the existing theory of the Kalman filter to accomplish this step. This step is conditional on Θ, and we assume at this point that Θ is fixed and known. In other words, our goal is to sample the entire set of state vectors, $X_n = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, from the multivariate normal posterior density $p_\Theta(X_n \mid Y_n)$, where $Y_n = \{y_1, \ldots, y_n\}$ represents the observations. Because of the Markov structure, we can write,

$$p_\Theta(X_n \mid Y_n) = p_\Theta(\boldsymbol{x}_n \mid Y_n) p_\Theta(\boldsymbol{x}_{n-1} \mid \boldsymbol{x}_n, Y_{n-1}) \cdots p_\Theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1). \tag{6.169}$$

In view of (6.169), it is possible to sample the entire set of state vectors, X_n, by sequentially simulating the individual states backward. This process yields a simulation method that Frühwirth–Schnatter (1994) called the forward-filtering, backward-sampling algorithm. In particular,

because the processes are Gaussian, we need only obtain the conditional means and variances, say, $\boldsymbol{m}_t = E_\Theta(\boldsymbol{x}_t \mid Y_t, \boldsymbol{x}_{t+1})$, and $V_t = \mathrm{var}_\Theta(\boldsymbol{x}_t \mid Y_t, \boldsymbol{x}_{t+1})$. This conditioning argument is akin to having $\boldsymbol{x}_{t+1}$ as an additional observation on state $\boldsymbol{x}_t$. In particular, using standard multivariate normal distribution theory,

$$\begin{aligned} \boldsymbol{m}_t &= \boldsymbol{x}_t^t + J_t(\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t+1}^t), \\ V_t &= P_t^t - J_t P_{t+1}^t J_t', \end{aligned} \tag{6.170}$$

for $t = n-1, n-2, \ldots, 0$, where J_t is defined in (6.49). To verify (6.170), the essential part of the Gaussian density (that is, the exponent) of $\boldsymbol{x}_t \mid Y_t, \boldsymbol{x}_{t+1}$ is

$$(\boldsymbol{x}_{t+1} - \Phi_{t+1}\boldsymbol{x}_t)'[Q_{t+1}]^{-1}(\boldsymbol{x}_{t+1} - \Phi_{t+1}\boldsymbol{x}_t) + (\boldsymbol{x}_t - \boldsymbol{x}_t^t)'[P_t^t]^{-1}(\boldsymbol{x}_t - \boldsymbol{x}_t^t),$$

and we simply complete the square; see Frühwirth–Schnatter (1994) or West and Harrison (1997, Section 4.7). Hence, the algorithm is to first sample $\boldsymbol{x}_n$ from a $\mathrm{N}(\boldsymbol{x}_n^n, P_n^n)$, where $\boldsymbol{x}_n^n$ and P_n^n are obtained from the Kalman filter, Property P6.1, and then sample $\boldsymbol{x}_t$ from a $\mathrm{N}(\boldsymbol{m}_t, V_t)$, for $t = n-1, n-2, \ldots, 0$, where the conditioning value of $\boldsymbol{x}_{t+1}$ is the value previously sampled; $\boldsymbol{m}_t$ and V_t are given in (6.170).

Next, we address an MCMC approach to nonlinear and non-Gaussian state-space modeling that was first presented in Carlin et al. (1992). We consider the general model given in (6.163), but with additive errors:

$$\boldsymbol{x}_t = F_t(\boldsymbol{x}_{t-1}) + \boldsymbol{w}_t \quad \text{and} \quad \boldsymbol{y}_t = H_t(\boldsymbol{x}_t) + \boldsymbol{v}_t, \tag{6.171}$$

where F_t and H_t are given, but may also depend on unknown parameters, say, Φ_t and A_t, respectively, the collection of which will be denoted by Θ. The errors are independent white noise sequences with $\mathrm{var}(\boldsymbol{w}_t) = Q_t$ and $\mathrm{var}(\boldsymbol{v}_t) = R_t$. Although time-varying variance–covariance matrices are easily incorporated in this framework, to ease the discussion we focus on the case $Q_t \equiv Q$ and $R_t \equiv R$. Also, although it is not necessary, we assume the initial state condition $\boldsymbol{x}_0$ is fixed and known; this is merely for notational convenience, so we do not have to carry along the additional terms involving $\boldsymbol{x}_0$ throughout the discussion.

In general, the likelihood specification for the model is given by

$$L_{X,Y}(\Theta, Q, R) = \prod_{t=1}^{n} f_1(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \Theta, Q)\ f_2(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \Theta, R), \tag{6.172}$$

where it is assumed the densities $f_1(\cdot)$ and $f_2(\cdot)$ are scale mixtures of normals. Specifically, for $t = 1, \ldots, n$,

$$\begin{aligned} f_1(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \Theta, Q) &= \int f(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \Theta, Q, \lambda_t)\ p_1(\lambda_t)\ d\lambda_t, \\ f_2(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \Theta, R) &= \int f(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \Theta, R, \omega_t)\ p_2(\omega_t)\ d\omega_t, \end{aligned} \tag{6.173}$$

where conditional on the independent sequences of nuisance parameters $\boldsymbol{\lambda} = (\lambda_t; t = 1, \ldots, n)$ and $\boldsymbol{\omega} = (\omega_t; t = 1, \ldots, n)$,

$$\begin{aligned} \boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \Theta, Q, \lambda_t &\sim \mathrm{N}\Big(F_t(\boldsymbol{x}_{t-1}; \Theta),\ \lambda_t Q\Big), \\ \boldsymbol{y}_t \mid \boldsymbol{x}_t, \Theta, R, \omega_t &\sim \mathrm{N}\Big(H_t(\boldsymbol{x}_t; \Theta),\ \omega_t R\Big). \end{aligned} \tag{6.174}$$

By varying $p_1(\lambda_t)$ and $p_2(\omega_t)$, we can have a wide variety of non-Gaussian error densities. These densities include, for example, double exponential, logistic, and t distributions in the univariate case and a rich class of multivariate distributions; this is discussed further in Carlin et al. (1992). The key to the approach is the introduction of the nuisance parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$ and the structure (6.174), which lends itself naturally to the Gibbs sampler and allows for the analysis of this general nonlinear and non-Gaussian problem.

According to Example 6.17, to implement the Gibbs sampler, we must be able to sample from the following complete conditional distributions:

(i) $\boldsymbol{x}_t \mid \boldsymbol{x}_{s\neq t}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \Theta, Q, R, Y_n \quad t = 1, \ldots, n,$

(ii) $\lambda_t \mid \lambda_{s\neq t}, \boldsymbol{\omega}, \Theta, Q, R, Y_n, X_n \sim \lambda_t \mid \Theta, Q, \boldsymbol{x}_t, \boldsymbol{x}_{t-1} \quad t = 1, \ldots, n,$

(iii) $\omega_t \mid \omega_{s\neq t}, \boldsymbol{\lambda}, \Theta, Q, R, Y_n, X_n \sim \omega_t \mid \Theta, R, \boldsymbol{y}_t, \boldsymbol{x}_t \quad t = 1, \ldots, n,$

(iv) $Q \mid \boldsymbol{\lambda}, \boldsymbol{\omega}, \Theta, R, Y_n, X_n \sim Q \mid \boldsymbol{\lambda}, Y_n, X_n,$

(v) $R \mid \boldsymbol{\lambda}, \boldsymbol{\omega}, \Theta, Q, Y_n, X_n \sim R \mid \boldsymbol{\omega}, Y_n, X_n,$

(vi) $\Theta \mid \boldsymbol{\lambda}, \boldsymbol{\omega}, Q, R, Y_n, X_n \sim \Theta \mid Y_n, X_n,$

where $X_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $Y_n = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$. The main difference between this method and the linear Gaussian case is that, because of the generality, we sample the states one-at-a-time rather than simultaneously generating all of them. As discussed in Carter and Kohn (1994), if possible, it is more efficient to generate the states simultaneously as in Example 6.18.

We will discuss items (i) and (ii) above. The third item follows in a similar manner to the second, and items (iv)-(vi) will follow from standard multivariate normal distribution theory and from Wishart distribution theory because of the conditioning on $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$. We will discuss this matter further in the next example. First, consider the linear model, $F_t(\boldsymbol{x}_{t-1}) = \Phi_t \boldsymbol{x}_{t-1}$, and $H_t(\boldsymbol{x}_t) = A_t \boldsymbol{x}_t$ in (6.171). In this case, for $t = 1, \ldots, n$, $\boldsymbol{x}_t \mid \boldsymbol{x}_{s\neq t}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \Theta, Q, R, Y_n$ has a p-dimensional $\mathrm{N}_p(B_t \boldsymbol{b}_t, B_t)$ distribution, with

$$\begin{aligned} B_t^{-1} &= \frac{Q^{-1}}{\lambda_t} + \frac{A_t' R^{-1} A_t}{\omega_t} + \frac{\Phi_{t+1}' Q^{-1} \Phi_{t+1}}{\lambda_{t+1}}, \\ \boldsymbol{b}_t &= \frac{\boldsymbol{x}_{t-1} \Phi_t' Q^{-1}}{\lambda_t} + \frac{\boldsymbol{y}_t R^{-1} A_t}{\omega_t} + \frac{\boldsymbol{x}_{t+1} Q^{-1} \Phi_{t+1}}{\lambda_{t+1}}, \end{aligned} \tag{6.175}$$

where, when $t = n$ in (6.175), terms in the sum with elements having a subscript of $n+1$ are dropped (this is assumed to be the case in what follows, although we do not explicitly state it). This result follows by noting the essential part of the multivariate normal distribution (that is, the exponent) of $\boldsymbol{x}_t \mid \boldsymbol{x}_{s\neq t}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \Theta, Q, R, Y_n$ is

$$(\boldsymbol{x}_t - \Phi_t \boldsymbol{x}_{t-1})'(\lambda_t Q)^{-1}(\boldsymbol{x}_t - \Phi_t \boldsymbol{x}_{t-1}) + (\boldsymbol{y}_t - A_t \boldsymbol{x}_t)'(\omega_t R)^{-1}(\boldsymbol{y}_t - A_t \boldsymbol{x}_t)$$
$$+(\boldsymbol{x}_{t+1} - \Phi_{t+1} \boldsymbol{x}_t)'(\lambda_{t+1} Q)^{-1}(\boldsymbol{x}_{t+1} - \Phi_{t+1} \boldsymbol{x}_t), \tag{6.176}$$

which upon manipulation yields (6.175).

Example 6.19 Nonlinear Models

In the case of nonlinear models, we can use (6.175) with slight modifications. For example, consider the case in which F_t is nonlinear, but H_t is linear, so the observations are $\boldsymbol{y}_t = A_t \boldsymbol{x}_t + \boldsymbol{v}_t$. Then,

$$\boldsymbol{x}_t \mid \boldsymbol{x}_{s\neq t}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \Theta, Q, R, Y_n \propto \eta_1(\boldsymbol{x}_t) \mathrm{N}_p(B_{1t}\boldsymbol{b}_{1t}, B_{1t}), \tag{6.177}$$

where

$$\begin{aligned} B_{1t}^{-1} &= \frac{Q^{-1}}{\lambda_t} + \frac{A_t' R^{-1} A_t}{\omega_t}, \\ \boldsymbol{b}_{1t} &= \frac{F_t'(\boldsymbol{x}_{t-1}) Q^{-1}}{\lambda_t} + \frac{\boldsymbol{y}_t R^{-1} A_t}{\omega_t}, \end{aligned}$$

and

$$\eta_1(\boldsymbol{x}_t) = \exp\left\{ -\frac{1}{2\lambda_{t+1}} \Big(\boldsymbol{x}_{t+1} - F_{t+1}(\boldsymbol{x}_t)\Big)' Q^{-1} \Big(\boldsymbol{x}_{t+1} - F_{t+1}(\boldsymbol{x}_t)\Big) \right\}.$$

Because $0 \leq \eta_1(\boldsymbol{x}_t) \leq 1$, for all $\boldsymbol{x}_t$, the distribution we want to sample from is dominated by the $\mathrm{N}_p(B_{1t}\boldsymbol{b}_{1t}, B_{1t})$ density. Hence, we may use rejection sampling as discussed in Example 6.17 to obtain an observation from the required density. That is, we generate a pseudo-variate from the $\mathrm{N}_p(B_{1t}\boldsymbol{b}_{1t}, B_{1t})$ density and accept it with probability $\eta_1(\boldsymbol{x}_t)$.

We proceed analogously in the case in which $F_t(\boldsymbol{x}_{t-1}) = \Phi_t \boldsymbol{x}_{t-1}$ is linear and $H_t(\boldsymbol{x}_t)$ is nonlinear. In this case,

$$\boldsymbol{x}_t \mid \boldsymbol{x}_{s\neq t}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \Theta, Q, R, Y_n \propto \eta_2(\boldsymbol{x}_t) \mathrm{N}_p(B_{2t}\boldsymbol{b}_{2t}, B_{2t}), \tag{6.178}$$

where

$$\begin{aligned} B_{2t}^{-1} &= \frac{Q^{-1}}{\lambda_t} + \frac{\Phi_{t+1}' Q^{-1} \Phi_{t+1}}{\lambda_{t+1}}, \\ \boldsymbol{b}_{2t} &= \frac{\boldsymbol{x}_{t-1} \Phi_t' Q^{-1}}{\lambda_t} + \frac{\boldsymbol{x}_{t+1} Q^{-1} \Phi_{t+1}}{\lambda_{t+1}}, \end{aligned}$$

and

$$\eta_2(\boldsymbol{x}_t) = \exp\left\{-\frac{1}{2\omega_t}\Big(\boldsymbol{y}_t - H_t(\boldsymbol{x}_t)\Big)' R^{-1}\Big(\boldsymbol{y}_t - H_t(\boldsymbol{x}_t)\Big)\right\}.$$

Here, we generate a pseudo-variate from the $\mathrm{N}_p(B_{2t}\boldsymbol{b}_{2t}, B_{2t})$ density and accept it with probability $\eta_2(\boldsymbol{x}_t)$.

Finally, in the case in which both F_t and H_t are nonlinear, we have

$$\boldsymbol{x}_t \mid \boldsymbol{x}_{s\neq t}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \Theta, Q, R, Y_n \propto \eta_1(\boldsymbol{x}_t)\eta_2(\boldsymbol{x}_t)\mathrm{N}_p(F_t(\boldsymbol{x}_{t-1}), \lambda_t Q), \quad (6.179)$$

so we sample from a $\mathrm{N}_p(F_t(\boldsymbol{x}_{t-1}), \lambda_t Q)$ density and accept it with probability $\eta_1(\boldsymbol{x}_t)\eta_2(\boldsymbol{x}_t)$.

Determination of (ii), $\lambda_t \mid \Theta, Q, \boldsymbol{x}_t, \boldsymbol{x}_{t-1}$ follows directly from Bayes theorem; that is, $p(\lambda_t \mid \Theta, Q, \boldsymbol{x}_t, \boldsymbol{x}_{t-1}) \propto p_1(\lambda_t)p(\boldsymbol{x}_t \mid \lambda_t, \boldsymbol{x}_{t-1}, \Theta, Q)$. By (6.173), however, we know the normalization constant is given by $f_1(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \Theta, Q)$, and thus the complete conditional density for λ_t is of a known functional form.

Many examples of these techniques are given in Carlin et al. (1992), including the problem of model choice. In the next example, we consider a univariate nonlinear model in which the state noise process has a t-distribution. As noted in Meinhold and Singpurwalla (1989), using t-distributions for the error processes is a way of robustifying the Kalman filter against outliers. In this example we present a brief discussion of a detailed analysis presented in Carlin et al. (1992, Example 4.2); readers interested in more detail may find it in that article.

Example 6.20 Analysis of a Nonlinear, Non-Gaussian State-Space Model

Kitagawa (1987) considered the analysis of data generated from the following univariate nonlinear model:

$$x_t = F_t(x_{t-1}) + w_t \quad \text{and} \quad y_t = H_t(x_t) + v_t \quad t = 1, \ldots, 100, \quad (6.180)$$

with

$$\begin{aligned} F_t(x_{t-1}) &= \alpha x_{t-1} + \beta x_{t-1}/(1 + x_{t-1}^2) + \gamma\cos[1.2(t-1)], \\ H_t(x_t) &= x_t^2/20, \end{aligned} \quad (6.181)$$

where $x_0 = 0$, w_t are independent random variables having a central t-distribution with $\nu = 10$ degrees and scaled so $\mathrm{var}(w_t) = \sigma_w^2 = 10$ [we denote this generically by $t(0, \sigma, \nu)$], and v_t is white standard Gaussian noise, $\mathrm{var}(v_t) = \sigma_v^2 = 1$. The state noise and observation noise are mutually independent. Kitagawa (1987) discussed the analysis of data generated from this model with $\alpha = .5$, $\beta = 25$, and $\gamma = 8$ assumed

known. We will use these values of the parameters in this example, but we will assume they are unknown. Figure 6.15 shows a typical data sequence y_t and the corresponding state process x_t.

Our goal here will be to obtain an estimate of the prediction density $p(x_{101} \mid Y_{100})$. To accomplish this, we use $n = 101$ and consider y_{101} as a latent variable (we will discuss this in more detail shortly). The priors on the variance components are chosen from a conjugate family, that is, $\sigma_w^2 \sim \text{IG}(a_0, b_0)$ independent of $\sigma_v^2 \sim \text{IG}(c_0, d_0)$, where IG denotes the inverse (reciprocal) gamma distribution [z has an inverse gamma distribution if $1/z$ has a gamma distribution; general properties can be found, for example, in Box and Tiao (1973, Section 8.5)]. Then,

$$\begin{aligned}
\sigma_w^2 \mid \boldsymbol{\lambda}, Y_n, X_n \sim \\
&\text{IG}\left(a_0 + \frac{n}{2},\ \left\{\frac{1}{b_0} + \frac{1}{2}\sum_{t=1}^{n}[x_t - F(x_{t-1})]^2/\lambda_t\right\}^{-1}\right), \\
\sigma_v^2 \mid \boldsymbol{\omega}, Y_n, X_n \sim \\
&\text{IG}\left(c_0 + \frac{n}{2},\ \left\{\frac{1}{d_0} + \frac{1}{2}\sum_{t=1}^{n}[y_t - H(x_t)]^2/\omega_t\right\}^{-1}\right). \quad (6.182)
\end{aligned}$$

Next, letting $\nu/\lambda_t \sim \chi_\nu^2$, we get that, marginally, $w_t \mid \sigma_w \sim t(0, \sigma_w, \nu)$, as required, leading to the complete conditional $\lambda_t \mid \sigma_w, \alpha, \beta, \gamma, Y_n, X_n$, for $t = 1, \ldots, n$, being distributed as

$$\text{IG}\left(\frac{\nu+1}{2},\ 2\left\{\frac{[x_t - F(x_{t-1})]^2}{\sigma_w^2} + \nu\right\}^{-1}\right). \quad (6.183)$$

We take $\omega_t \equiv 1$ for $t = 1, \ldots, n$, because the observation noise is Gaussian.

For the states, x_t, we take a normal prior on the initial state, $x_0 \sim \text{N}(\mu_0, \sigma_0^2)$, and then we use rejection sampling to conditionally generate a state value x_t, for $t = 1, \ldots, n$, as described in Example 6.19, equation (6.179). In this case, $\eta_1(x_t)$ and $\eta_2(x_t)$ are given in (6.177) and (6.178), respectively, with F_t and H_t given by (6.181), $\Theta = (\alpha, \beta, \gamma)'$, $Q = \sigma_w^2$ and $R = \sigma_v^2$. Endpoints take some special consideration; we generate x_0 from a $\text{N}(\mu_0, \sigma_0^2)$ and accept it with probability $\eta_1(x_0)$, and we generate x_{101} as usual and accept it with probability $\eta_2(x_{101})$. The last complete conditional depends on y_{101}, a latent data value not observed but instead generated according to its complete conditional, which is $\text{N}(x_{101}^2/20, \sigma_v^2)$, because $\omega_{101} = 1$.

The prior on $\Theta = (\alpha, \beta, \gamma)'$ is taken to be trivariate normal with mean $(\mu_\alpha, \mu_\beta, \mu_\gamma)'$ and diagonal variance–covariance matrix $\text{diag}\{\sigma_\alpha^2, \sigma_\beta^2, \sigma_\gamma^2\}$. The necessary conditionals can be found using standard normal theory,

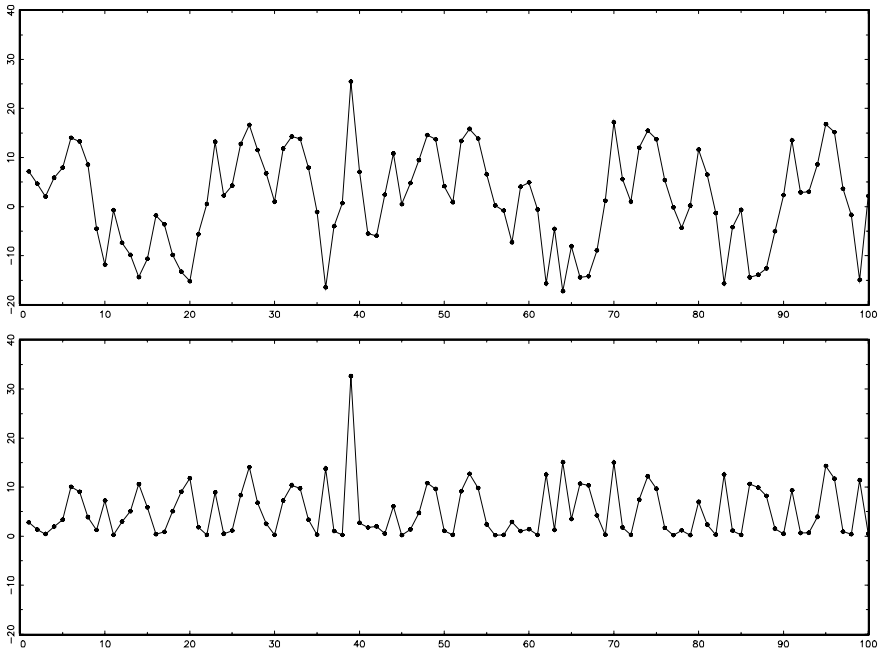


Figure 6.15 The state process, x_t (top), and the observations, y_t (bottom), for $t = 1, \ldots, 100$ generated from the model (6.180).

as done in (6.175). For example, the complete conditional distribution of α is of the form $\mathrm{N}(Bb, B)$, where

$$B^{-1} = \frac{1}{\sigma_\alpha^2} + \frac{1}{\sigma_w^2} \sum_{t=1}^{n} \frac{x_{t-1}^2}{\lambda_t}$$

and

$$b = \frac{\mu_\alpha}{\sigma_\alpha^2} + \frac{1}{\sigma_w^2} \sum_{t=1}^{n} \frac{x_{t-1}}{\lambda_t} \left(x_t - \beta \frac{x_{t-1}}{1 + x_{t-1}^2} - \gamma \cos[1.2(t-1)] \right).$$

The complete conditional for β has the same form, with

$$B^{-1} = \frac{1}{\sigma_\beta^2} + \frac{1}{\sigma_w^2} \sum_{t=1}^{n} \frac{x_{t-1}^2}{\lambda_t (1 + x_{t-1}^2)^2}$$

and

$$b = \frac{\mu_\beta}{\sigma_\beta^2} + \frac{1}{\sigma_w^2} \sum_{t=1}^{n} \frac{x_{t-1}}{\lambda_t (1 + x_{t-1}^2)} \left(x_t - \alpha x_{t-1} - \gamma \cos[1.2(t-1)] \right),$$

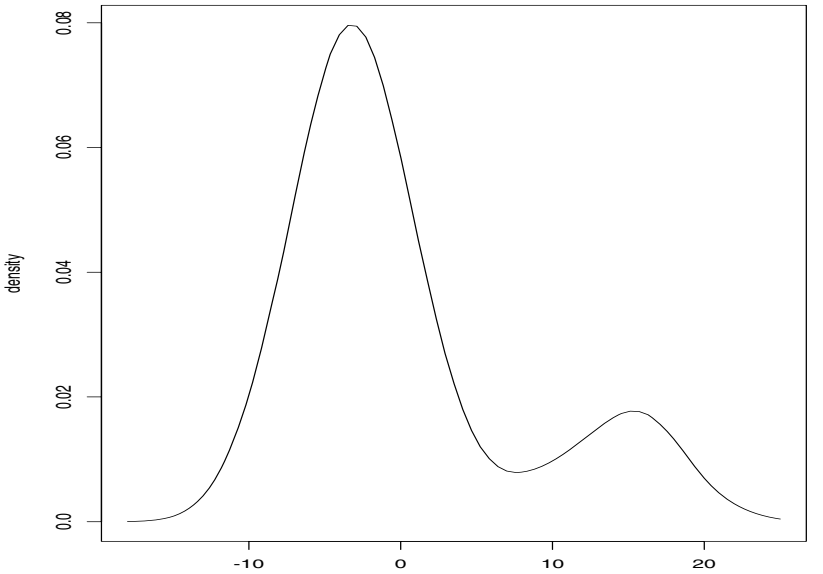


Figure 6.16 Estimated one-step-ahead prediction posterior density $\widehat{p}(x_{101}|Y_{100})$ of the state process for the nonlinear and non-normal model given by (6.180) using Gibbs sampling, $G = 500$.

and for γ the values are

$$B^{-1} = \frac{1}{\sigma_\gamma^2} + \frac{1}{\sigma_w^2}\sum_{t=1}^{n}\frac{\cos^2[1.2(t-1)]}{\lambda_t}$$

and

$$b = \frac{\mu_\gamma}{\sigma_\gamma^2} + \frac{1}{\sigma_w^2}\sum_{t=1}^{n}\frac{\cos[1.2(t-1)]}{\lambda_t}\left(x_t - \alpha x_{t-1} - \beta\frac{x_{t-1}}{1+x_{t-1}^2}\right).$$

In this example, we put $\mu_0 = 0$, $\sigma_0^2 = 10$, and $a_0 = 3$, $b_0 = .05$ (so the prior on σ_w^2 has mean and standard deviation equal to 10), and $c_0 = 3$, $d_0 = .5$ (so the prior on σ_v^2 has mean and standard deviation equal to one). The normal prior on $\Theta = (\alpha, \beta, \gamma)'$ had corresponding mean vector equal to $(\mu_\alpha = .5, \mu_\beta = 25, \mu_\gamma = 8)'$ and diagonal variance matrix equal to $\text{diag}\{\sigma_\alpha^2 = .25, \sigma_\beta^2 = 10, \sigma_\gamma^2 = 4\}$. The Gibbs sampler ran for $\ell = 50$ iterations for $G = 500$ parallel replications per iteration. We estimate the marginal posterior density of x_{101} as

$$\widehat{p}(x_{101} \mid Y_{100}) = G^{-1}\sum_{g=1}^{G}\text{N}\left(x_{101} \;\middle|\; [F_t(x_{t-1})]^{(g)}, \lambda_{101}^{(g)}\sigma_w^{2(g)}\right), \quad (6.184)$$

where $\mathrm{N}(\cdot|a, b)$ denotes the normal density with mean a and variance b, and

$$[F_t(x_{t-1})]^{(g)} = \alpha^{(g)} x_{t-1}^{(g)} + \beta^{(g)} x_{t-1}^{(g)} / (1 + x_{t-1}^{2(g)}) + \gamma^{(g)} \cos[1.2(t-1)].$$

The estimate, (6.184), with $G = 500$, is shown in Figure 6.16. Other aspects of the analysis, for example, the marginal posteriors of the elements of Θ, can be found in Carlin et al. (1992).

6.10 Stochastic Volatility

Recently, there has been considerable interest in stochastic volatility models. These models are similar to the ARCH models presented in Chapter 5, but they add a stochastic noise term to the equation for σ_t. Recall from §5.2 that a GARCH(1, 1) model for a return, which we denote here by r_t, is given by

$$r_t = \sigma_t \epsilon_t \tag{6.185}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \tag{6.186}$$

where ϵ_t is Gaussian white noise. If we define

$$h_t = \log \sigma_t^2 \quad \text{and} \quad y_t = \log r_t^2,$$

then (6.185) can be written as

$$y_t = h_t + \log \epsilon_t^2. \tag{6.187}$$

Equation (6.187) is considered the observation equation, and the stochastic variance h_t is considered to be an unobserved state process. Similar to (6.186), the volatility process follows, in its basic form, an autoregression,

$$h_t = \phi_0 + \phi_1 h_{t-1} + w_t, \tag{6.188}$$

where w_t is white Gaussian noise with variance σ_w^2.

Together, (6.187) and (6.188) make up the stochastic volatility model due to Harvey, Ruiz and Shephard (1994). If ϵ_t^2 had a log-normal distribution, (6.187)-(6.188) would form a Gaussian state-space model, and we could then use standard DLM results to fit the model to data. Unfortunately, $y_t = \log r_t^2$ is rarely normal, so we typically keep the ARCH normality assumption on ϵ_t; in which case, $\log \epsilon_t^2$ is distributed as the log of a chi-squared random variable with one degree of freedom. This density is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(e^x - x)\right\} \quad -\infty < x < \infty, \tag{6.189}$$

and its mean and variance are -1.27 and $\pi^2/2$, respectively; the density (6.189) is highly skewed with a long tail on the left (see Figure 6.18).

Various approaches to the fitting of stochastic volatility models have been examined; these methods include a wide range of assumptions on the observational noise process. A good summary of the proposed techniques, both Bayesian (via MCMC) and non-Bayesian approaches (such as quasi-maximum likelihood estimation and the EM algorithm), can be found in Jacquier et al. (1994), and Shephard (1996). Simulation methods for classical inference applied to stochastic volatility models are discussed in Danielson (1994) and Sandmann and Koopman (1998).

Kim, Shephard and Chib (1998) proposed modeling the log of a chi-squared random variable by a mixture of seven normals to approximate the first four moments of the observational error distribution; the mixture is fixed and no additional model parameters are added by using this technique. In an effort to keep matters simple, and perhaps somewhat more general (in that we allow the observational error dynamics to depend on parameters that will be fitted), our method of fitting stochastic volatility models is to retain the Gaussian state equation (6.188), but to write the observation equation, with $y_t = \log r_t^2$, as

$$y_t = \alpha + h_t + \eta_t, \tag{6.190}$$

where η_t is white noise, whose distribution is a mixture of two normals, one centered at zero. In particular, we write

$$\eta_t = u_t z_{t0} + (1 - u_t) z_{t1}, \tag{6.191}$$

where u_t is an iid Bernoulli process, $\Pr\{u_t = 0\} = \pi_0$, $\Pr\{u_t = 1\} = \pi_1$ $(\pi_0 + \pi_1 = 1)$, $z_{t0} \sim$ iid $\mathrm{N}(0, \sigma_0^2)$, and $z_{t1} \sim$ iid $\mathrm{N}(\mu_1, \sigma_1^2)$.

The advantage to this model is that it is easy to fit because it uses normality. In fact, the model equations (6.188) and (6.190)-(6.191) are similar to those presented in Peña and Guttman (1988), who used the idea to obtain a robust Kalman filter, and, as previously mentioned, in Kim, Shephard and Chib (1998). The material presented in §6.8 applies here, and in particular, the filtering equations for this model are

$$h_{t+1}^t = \phi_0 + \phi_1 h_t^{t-1} + \sum_{j=0}^{1} \pi_{tj} K_{tj} \epsilon_{tj}, \tag{6.192}$$

$$P_{t+1}^t = \phi_1^2 P_t^{t-1} + \sigma_w^2 - \sum_{j=0}^{1} \pi_{tj} K_{tj}^2 \Sigma_{tj}, \tag{6.193}$$

$$\epsilon_{t0} = y_t - \alpha - h_t^{t-1}, \tag{6.194}$$

$$\epsilon_{t1} = y_t - \alpha - h_t^{t-1} - \mu_1, \tag{6.195}$$

$$\Sigma_{t0} = P_t^{t-1} + \sigma_0^2, \tag{6.196}$$

$$\Sigma_{t1} = P_t^{t-1} + \sigma_1^2, \tag{6.197}$$

$$K_{t0} = \phi_1 P_t^{t-1} \,/\, \Sigma_{t0}, \tag{6.198}$$

$$K_{t1} = \phi_1 P_t^{t-1} \,/\, \Sigma_{t1}. \tag{6.199}$$

To complete the filtering, we must be able to assess the probabilities $\pi_{t1} = \Pr(u_t = 1 \mid y_1, \ldots, y_t)$, for $t = 1, \ldots, n$; of course, $\pi_{t0} = 1 - \pi_{t1}$. Let $f_j(t \mid t-1)$ denote the conditional density of y_t given the past $y_1, \ldots, y_{t-1}$, and $u_t = j$ $(j = 0, 1)$. Then,

$$\pi_{t1} = \frac{\pi_1 f_1(t \mid t-1)}{\pi_0 f_0(t \mid t-1) + \pi_1 f_1(t \mid t-1)}, \tag{6.200}$$

where we assume the distribution π_j, for $j = 0, 1$ has been specified *a priori*. If the investigator has no reason to prefer one state over another the choice of uniform priors, $\pi_1 = 1/2$, will suffice. Unfortunately, it is computationally difficult to obtain the exact values of $f_j(t \mid t-1)$; although we can give an explicit expression of $f_j(t \mid t-1)$, the actual computation of the conditional density is prohibitive. A viable approximation, however, is to choose $f_j(t \mid t-1)$ to be the normal density, $\mathrm{N}(h_t^{t-1} + \mu_j, \Sigma_{tj})$, for $j = 0, 1$ and $\mu_0 = 0$; see §6.8 for details.

The innovations filter given in (6.192)–(6.137) can be derived from the Kalman filter by a simple conditioning argument. For example, to derive (6.192), we write

$$\begin{aligned}
\mathrm{E}\left(h_{t+1} \middle| y_1, \ldots, y_t\right) &= \sum_{j=0}^{1} \mathrm{E}\left(h_{t+1} \middle| y_1, \ldots, y_t, u_t = j\right) \Pr(u_t = j \,|\, y_1, \ldots, y_t) \\
&= \sum_{j=0}^{1} \left(\phi_0 + \phi_1 h_t^{t-1} + K_{tj}\epsilon_{tj}\right) \pi_{tj} \\
&= \phi_0 + \phi_1 h_t^{t-1} + \sum_{j=0}^{1} \pi_{tj} K_{tj} \epsilon_{tj}.
\end{aligned}$$

Estimation of the parameters, $\Theta = (\phi_0, \phi_1, \sigma_0^2, \mu_1, \sigma_1^2, \sigma_w^2)'$, is accomplished via MLE based on the likelihood given by

$$\ln L_Y(\Theta) = \sum_{t=1}^{n} \ln \left(\sum_{j=0}^{1} \pi_j \; f_j(t \mid t-1) \right), \tag{6.201}$$

where the density $f_j(t \mid t-1)$ is approximated by the normal density, $\mathrm{N}(h_t^{t-1} + \mu_j, \sigma_j^2)$, previously mentioned. We may consider maximizing (6.201) directly as a function of the parameters Θ using a Newton method, or we may consider applying the EM algorithm to the complete data likelihood.

Example 6.21 Analysis of the New York Stock Exchange Returns

Figure 6.17 shows the log of the squares of returns, $y_t = \log r_t^2$, of 2000 daily observations of the NYSE previously displayed in Figure 1.4.

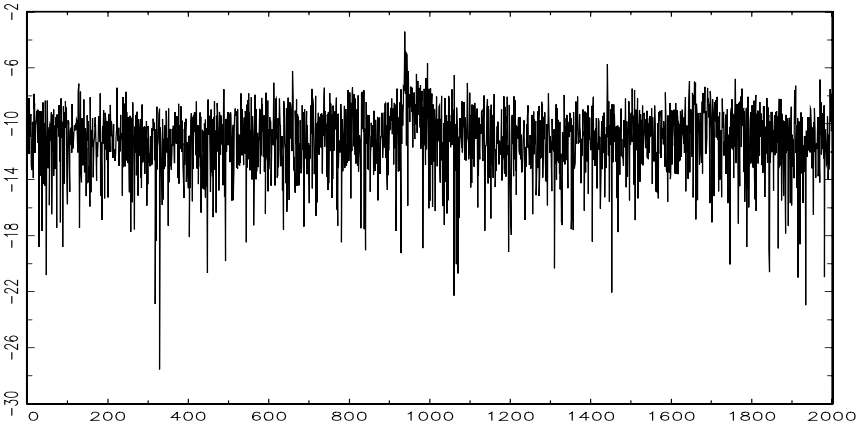


Figure 6.17 Graph of $y_t = \log r_t^2$, where r_t is the daily return of the NYSE, 2000 observations.

Table 6.4 Estimation Results for the NYSE Fit

Parameter	Estimate	Estimated Standard Error
ϕ_0	$-.006$	.016
ϕ_1	.988	.007
σ_w	.091	.027
α	-9.607	1.266
σ_0	1.220	.065
μ_1	-2.292	.204
σ_1	2.683	.105

Model (6.188) and (6.190)-(6.191), with and π_1 fixed at .5, was fit to the data using a quasi-Newton–Raphson method to maximize (6.201). The results are given in Table 6.4. Figure 6.18 compares the density of the log of a χ_1^2 with the fitted normal mixture; we note the data indicate a substantial amount of probability in the upper tail that the log-χ_1^2 distribution misses.

Finally, Figure 6.19 shows y_t for $800 \le t \le 1000$, which includes the crash of October 19, 1987, with $y_t^{t-1} = \widehat{\alpha} + h_t^{t-1}$ superimposed on the graph; compare with Figure 5.6. Also displayed are error bounds.

It is possible to use the bootstrap procedure described in §6.7 for the stochastic volatility model, with some minor changes. The following procedure was described in Stoffer and Wall (2004). We develop a vector first-order equation, as was done in (6.117). First, using (6.194)–(6.195), and noting that

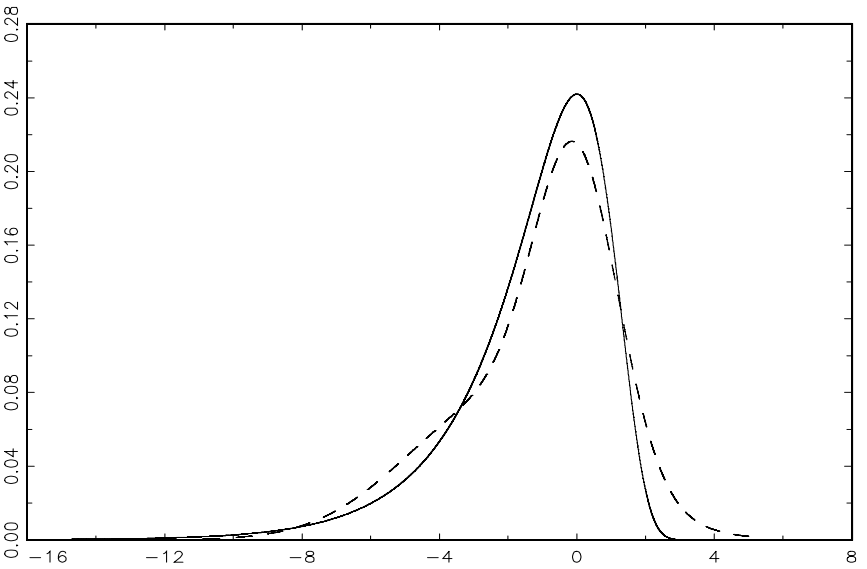


Figure 6.18 Density of the log of a χ^2_1 as given by (6.189) (solid line) and the fitted normal mixture (dashed line) form the NYSE example.

$y_t = \pi_{t0} y_t + \pi_{t1} y_t$, we may write

$$y_t = \alpha + h_t^{t-1} + \pi_{t0}\epsilon_{t0} + \pi_{t1}(\epsilon_{t1} + \mu_1). \tag{6.202}$$

Consider the standardized innovations

$$e_{tj} = \Sigma_{tj}^{-1/2}\epsilon_{tj}, \qquad j = 0, 1, \tag{6.203}$$

and define the 2×1 vector

$$\boldsymbol{e}_t = \begin{bmatrix} e_{t0} \\ e_{t1} \end{bmatrix}.$$

Also, define the 2×1 vector

$$\boldsymbol{\xi}_t = \begin{bmatrix} h_{t+1}^t \\ y_t \end{bmatrix}.$$

Combining (6.192) and (6.202) results in a vector first-order equation for $\boldsymbol{\xi}_t$ given by

$$\boldsymbol{\xi}_t = F\boldsymbol{\xi}_{t-1} + G_t + H_t\boldsymbol{e}_t, \tag{6.204}$$

where

$$F = \begin{bmatrix} \phi_1 & 0 \\ 1 & 0 \end{bmatrix}, \quad G_t = \begin{bmatrix} \phi_0 \\ \alpha + \pi_{t1}\mu_1 \end{bmatrix}, \quad H_t = \begin{bmatrix} \pi_{t0}K_{t0}\Sigma_{t0}^{1/2} & \pi_{t1}K_{t1}\Sigma_{t1}^{1/2} \\ \pi_{t0}\Sigma_{t0}^{1/2} & \pi_{t1}\Sigma_{t1}^{1/2} \end{bmatrix}.$$

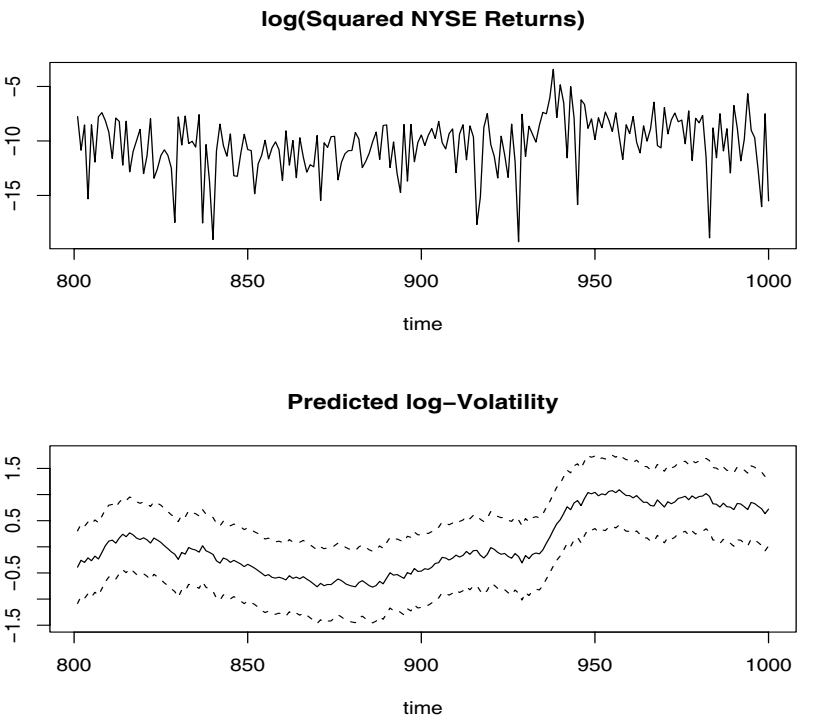


Figure 6.19 Two hundred observations of $y_t = \log r_t^2$, for $801 \le t \le 1000$, where r_t is the daily return of the NYSE (top). Corresponding one-step-ahead predicted log volatility, $\log \sigma_t^2$, with ± 2 standard prediction errors (bottom).

Hence, the steps in bootstrapping for this case are the same as steps 1 through 5 described in §5.7, but with (6.117) replaced by the following first-order equation:

$$\boldsymbol{\xi}_t^* = F(\widehat{\Theta})\boldsymbol{\xi}_{t-1}^* + G_t(\widehat{\Theta}; \widehat{\pi}_{t1}) + H_t(\widehat{\Theta}; \widehat{\pi}_{t1})\boldsymbol{e}_t^*, \tag{6.205}$$

where $\widehat{\Theta} = (\widehat{\phi}_0, \widehat{\phi}_1, \widehat{\sigma}_0^2, \widehat{\alpha}, \widehat{\mu}_1, \widehat{\sigma}_1^2, \widehat{\sigma}_w^2)'$ is the MLE of Θ, and $\widehat{\pi}_{t1}$ is estimated via (6.200), replacing $f_1(t \mid t-1)$ and $f_0(t \mid t-1)$ by their respective estimated normal densities ($\widehat{\pi}_{t0} = 1 - \widehat{\pi}_{t1}$).

Example 6.22 Analysis of the U.S. GNP Growth Rate

In Example 5.3, we fit an ARCH model to the U.S. GNP growth rate. In this example, we will fit a stochastic volatility model to the residuals from the MA(2) fit on the growth rate (see Example 3.35).

Figure 6.20 shows the log of the squared residuals, say y_t, from the MA(2) fit on the U.S. GNP series. The stochastic volatility model (6.187)–(6.191) was then fit to y_t. Table 6.5 shows the MLEs of the model

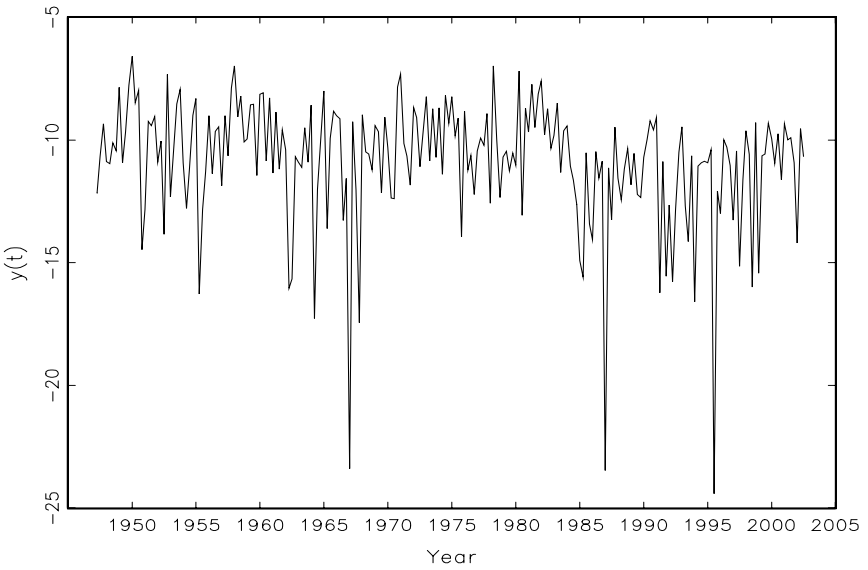


Figure 6.20 Log of the squared residuals from an MA(2) fit on GNP growth rate.

parameters along with their asymptotic SEs assuming the model is correct. Also displayed in Table 6.5 are the means and SEs of $B = 500$ bootstrapped samples. There is some amount of agreement between the asymptotic values and the bootstrapped values. The interest here, however, is not so much in the SEs, but in the actual sampling distribution of the estimates. For example, Figure 6.21 compares the bootstrap histogram and asymptotic normal distribution of $\widehat{\phi}_1$. In this case, the bootstrap distribution exhibits positive kurtosis and skewness which is missed by the assumption of asymptotic normality.

6.11 State-Space and ARMAX Models for Longitudinal Data Analysis

In some studies, we may observe several independent k-dimensional time series, say, $\boldsymbol{y}_{t\ell}$, for $\ell = 1, \ldots, N$. For example, a new treatment may be given to N patients with high blood pressure, and the systolic and diastolic blood pressures (SBP and DBP) are recorded at equal time intervals, for some time, using an ambulatory device. We may think of $\boldsymbol{y}_{t\ell}$ as being the bivariate, $k = 2$, recordings of SBP and DBP at time t for person ℓ. It is also reasonable to assume, in this example, exogenous variables may have been collected on each

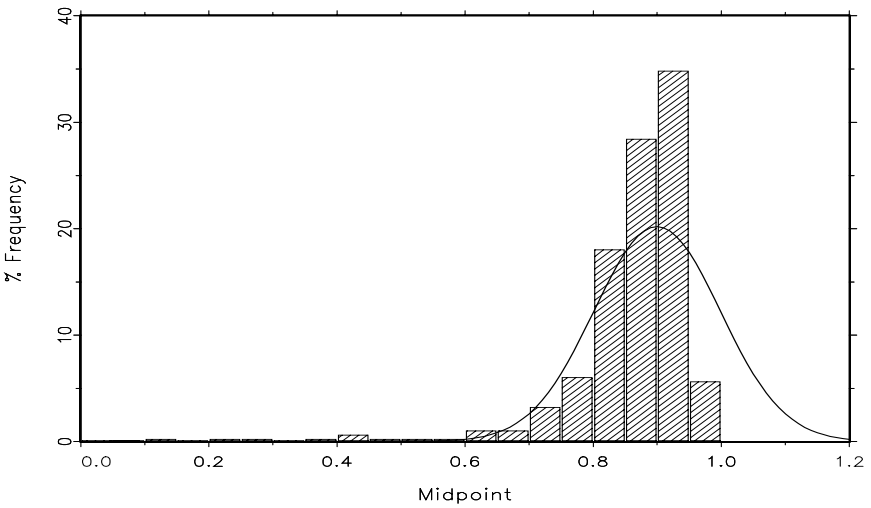


Figure 6.21 Bootstrap histogram and asymptotic distribution of $\widehat{\phi}_1$ for the U.S. GNP example.

Table 6.5 Estimates and Their Asymptotic and Bootstrap Standard Errors for U.S. GNP Example.

Parameter	MLE	Asymptotic SE	Bootstrap Mean†	Bootstrap SE†
ϕ_0	.068	.274	−.010	.353
ϕ_1	.900	.099	.864	.102
σ_w	.378	.208	.696	.375
α	−10.524	2.321	−10.792	.748
μ_1	−2.164	.567	−1.941	.416
σ_1	3.007	.377	2.891	.422
σ_0	.935	.198	.692	.362

† Based on 500 bootstrapped samples.

subject to help explain the variation in blood pressure (for example, gender, race, age, activity, and so on). We might expect to encounter missing data or irregularly spaced observations in this type of experiment; these problems are easier to handle from a state-space perspective.

An extension of the ARMAX model given in (6.103) that might handle the case of cross-sectional data, $\boldsymbol{y}_{t\ell}$, is

$$\boldsymbol{y}_{t\ell} = \Gamma \boldsymbol{u}_{t\ell} + \sum_{j=1}^{p} \Phi_j \boldsymbol{y}_{t-j,\ell} + \sum_{j=1}^{q} \Theta_j \boldsymbol{w}_{t-j,\ell} + \boldsymbol{w}_{t\ell}, \tag{6.206}$$

where, for $\ell = 1, \ldots, N$, $\mathrm{var}(\boldsymbol{w}_{t\ell}) = \Sigma_w$ and $\boldsymbol{u}_{t\ell}$ represents the $r \times 1$ vector of exogenous variables. As in §6.6, Property P6.6, we can write (6.206) in terms

of a state-space model. That is, for $\ell = 1, \ldots, N$,

$$\begin{aligned} \boldsymbol{x}_{t+1,\ell} &= F\boldsymbol{x}_{t,\ell} + G\boldsymbol{w}_{t,\ell}, & (6.207) \\ \boldsymbol{y}_{t,\ell} &= [I, 0, \cdots, 0]\boldsymbol{x}_{t\ell} + \Gamma\boldsymbol{u}_{t\ell} + \boldsymbol{w}_{t\ell}, & (6.208) \end{aligned}$$

where matrices F and G are as in (6.104), $\boldsymbol{x}_{t,\ell}$ represents the unobserved state, and $\boldsymbol{y}_{t,\ell}$ is the observation at time t, replication ℓ. Maximum likelihood estimation for state space models with cross-sectional data, such as the example given here, was investigated by Goodrich and Caines (1979), and can be carried out with minor modifications to the methods described in §6.3. In particular, given data $\boldsymbol{y}_{t,\ell}$, $t = 1, \ldots, n$, $\ell = 1, \ldots, N$, we can use Newton–Raphson to minimize the criterion function, which is, up to a constant term, proportional to the negative of the log likelihood function,

$$l(\Theta) = N^{-1} \sum_{\ell=1}^{N} \left(\sum_{t=1}^{n} \log\left|\Sigma_{t,\ell}(\Theta)\right| + \sum_{t=1}^{n} \boldsymbol{\epsilon}_{t,\ell}(\Theta)' \Sigma_{t,\ell}(\Theta)^{-1} \boldsymbol{\epsilon}_{t,\ell}(\Theta) \right), \quad (6.209)$$

where $\boldsymbol{\epsilon}_{t,\ell}(\Theta)$ and $\Sigma_{t,\ell}(\Theta)$ are the innovations and their variance–covariance matrices, respectively. For details, see Goodrich and Caines (1979).

Anderson (1978) did an extensive study of replicated ARX models, that is, the case in which $q = 0$ in (6.206). We can write this model using regression notation as

$$\boldsymbol{y}_{t\ell} = \mathcal{B}\boldsymbol{z}_{t\ell} + \boldsymbol{w}_{t\ell}, \quad (6.210)$$

for $\ell = 1, \ldots, N$ and $t = p+1, \ldots, n$, where

$$\boldsymbol{z}_{t\ell} = (\boldsymbol{u}'_{t\ell}, \boldsymbol{y}'_{t-1,\ell}, \ldots, \boldsymbol{y}'_{t-p,\ell})' \quad (6.211)$$

and the matrix of regression coefficients is

$$\mathcal{B} = [\Gamma, \Phi_1, \Phi_2, \ldots, \Phi_p]. \quad (6.212)$$

The estimate of the regression matrix $\mathcal{B}$ in this case is

$$\widehat{\mathcal{B}} = \left(\sum_{\ell=1}^{N} \sum_{t=p+1}^{n} \boldsymbol{y}_{t\ell}\boldsymbol{z}'_{t\ell} \right) \left(\sum_{\ell=1}^{N} \sum_{t=p+1}^{n} \boldsymbol{z}_{t\ell}\boldsymbol{z}'_{t\ell} \right)^{-1}, \quad (6.213)$$

and an estimate of Σ_w is

$$\widehat{\Sigma}_w = \frac{1}{N(n-p)} \sum_{\ell=1}^{N} \sum_{t=p+1}^{n} (\boldsymbol{y}_{t\ell} - \widehat{\mathcal{B}}\boldsymbol{z}_{t\ell})(\boldsymbol{y}_{t\ell} - \widehat{\mathcal{B}}\boldsymbol{z}_{t\ell})'. \quad (6.214)$$

Inference for $\widehat{\mathcal{B}}$ follows as in multivariate regression. That is, the large sample standard error of the ij-th element of $\mathcal{B}$ is $\sqrt{\widehat{\sigma}_{jj} c_{ii}}$, where $\widehat{\sigma}_{jj}$ is the j-th diagonal element of $\widehat{\Sigma}_w$ and c_{ii} is the i-th diagonal element of

$$\left(\sum_{\ell=1}^{N} \sum_{t=p+1}^{n} \boldsymbol{z}_{t\ell}\boldsymbol{z}'_{t\ell} \right)^{-1}.$$

Model (6.206) may be somewhat restrictive in its assumption that the parameters do not change over time. Because replications exist, extending the model to the case of time-varying parameters is easy. The case of time-varying parameters in (6.210) was also presented in Anderson (1978). In particular, the model is written as

$$\boldsymbol{y}_{t\ell} = \Gamma_t \boldsymbol{u}_{t\ell} + \sum_{j=1}^{p_t} \Phi_{tj} \boldsymbol{y}_{t-j,\ell} + \boldsymbol{w}_{t\ell}, \tag{6.215}$$

and $\text{var}(\boldsymbol{w}_{t\ell}) = \Sigma_t$, for $\ell = 1, \ldots, N$. The order of the model, p_t, is also allowed to vary with time, and the equal spacing of time is not required. Of course, we can still use regression for estimation because the time-varying model can be written as n regressions, one for each point in time,

$$\boldsymbol{y}_{t\ell} = \mathcal{B}_t \boldsymbol{z}_{t\ell} + \boldsymbol{w}_{t\ell}, \tag{6.216}$$

for $\ell = 1, \ldots, N$, where $\boldsymbol{z}_{t\ell}$ is as in (6.211), but with p replaced by p_t, and where now,

$$\mathcal{B}_t = [\Gamma_t, \Phi_{t1}, \Phi_{t2}, \ldots, \Phi_{tp_t}], \tag{6.217}$$

assuming $t > p_t$. The estimate of $\mathcal{B}_t$, for any time t, is now given by

$$\widehat{\mathcal{B}}_t = \left(\sum_{\ell=1}^{N} \boldsymbol{y}_{t\ell} \boldsymbol{z}_{t\ell}' \right) \left(\sum_{\ell=1}^{N} \boldsymbol{z}_{t\ell} \boldsymbol{z}_{t\ell}' \right)^{-1}, \tag{6.218}$$

and an estimate of Σ_t is

$$\widehat{\Sigma}_t = \frac{1}{N - p_t - 1} \sum_{\ell=1}^{N} (\boldsymbol{y}_{t\ell} - \widehat{\mathcal{B}}_t \boldsymbol{z}_{t\ell})(\boldsymbol{y}_{t\ell} - \widehat{\mathcal{B}}_t \boldsymbol{z}_{t\ell})'. \tag{6.219}$$

Example 6.23 The Effect of Prenatal Smoking on Growth

In this example, we use data taken from an epidemiologic study at the University of Pittsburgh that focused on the effects of substance use during pregnancy. In particular, we focus on the growth of $N = 318$ children followed from birth to six years of age. In this longitudinal study, the children were examined at birth ($t = 0$), and at eight months ($t = 1$), 18 months ($t = 2$), 36 months ($t = 3$), and 72 months ($t = 4$) of age. At times $t = 1, 2, 3, 4$, a growth index, say, $y_{t\ell}$, was calculated for each child $\ell = 1, \ldots, 318$. The growth index is essentially a standardized score for a child's weight adjusting for that child's age, gender, and height, against the national averages. At birth, $y_{0\ell}$ represents the standardized birthweight of child ℓ.

We might consider that children not prenatally exposed to teratogens would follow a certain growth curve, whereas exposed children would

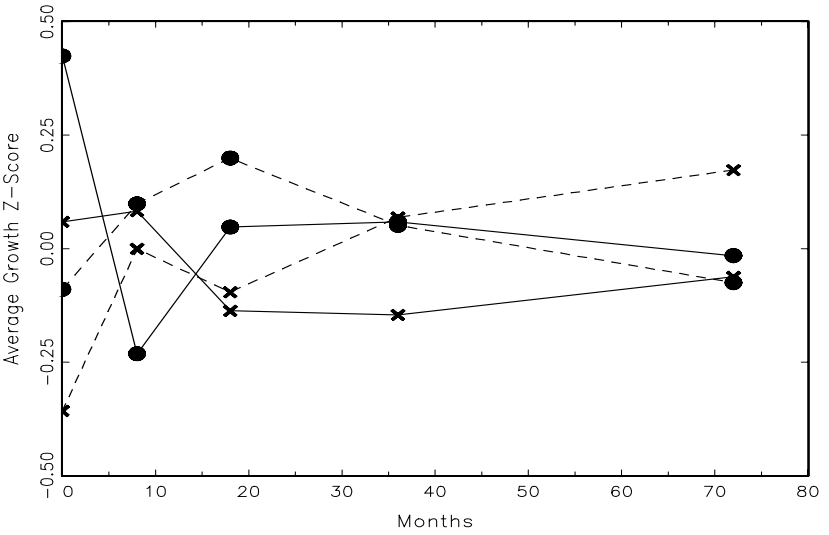


Figure 6.22 Average growth scores across time for four groups of children. A solid line represents children not prenatally exposed to cigarette smoke; a dashed line represents children prenatally exposed to cigarette smoke. A circle represents white children, and a cross represents black children.

follow another. To investigate this hypothesis, we propose the following time-varying ARX model for growth:

$$\begin{aligned} y_{t\ell} &= \gamma_{0t} + \gamma_{1t} S_\ell + \gamma_{2t} R_\ell + \gamma_{3t} S_\ell R_\ell \\ &+ \sum_{j=1}^{t} \phi_{tj} \left(y_{t-j,\ell} - \widehat{y}_{t-j,\ell} \right) + w_{t\ell}, \end{aligned} \tag{6.220}$$

for $t = 0, 1, 2, 3, 4$, where $\text{var}(w_{t\ell}) = \sigma_t^2$, for $\ell = 1, \ldots, 318$. The exogenous variables in the model are, S_ℓ, the average number of cigarettes per day the mother smoked during the second trimester of pregnancy, and R_ℓ, which indicates race (0 = black, 1 = white). The model is written in terms of the innovation sequences, $(y_{t-j,\ell} - \widehat{y}_{t-j,\ell})$, where $\widehat{y}_{t,\ell}$ is the prediction of $y_{t,\ell}$ from the previous model. We did this to remove any effect of smoking or race on previous growth. Figure 6.22 shows the average growth scores over time for four groups: 68 black children not exposed to smoke prenatally (solid line-cross), 92 white children not exposed to smoke (solid line-circle), 83 black children exposed to smoke (dashed line-cross), and 75 white children exposed to smoke (dashed line-circle). For display purposes in Figure 6.22, smoking has been dichotomized to no exposure versus any exposure, but in the analysis, the smoking variable is in average cigarettes per day.

For example, the model for birthweight, $t = 0$, is

$$y_{0\ell} = \gamma_{00} + \gamma_{10} S_\ell + \gamma_{20} R_\ell + \gamma_{30} S_\ell R_\ell + w_{0\ell}.$$

Once the model has been estimated, the predicted values are calculated

$$\widehat{y}_{0\ell} = \widehat{\gamma}_{00} + \widehat{\gamma}_{10} S_\ell + \widehat{\gamma}_{20} R_\ell + \widehat{\gamma}_{30} S_\ell R_\ell.$$

Then, the model for growth at eight months, $t = 1$, is

$$y_{1\ell} = \gamma_{01} + \gamma_{11} S_\ell + \gamma_{21} R_\ell + \gamma_{31} S_\ell R_\ell + \phi_{11}(y_{0,\ell} - \widehat{y}_{0,\ell}) + w_{1\ell},$$

where $(y_{0,\ell} - \widehat{y}_{0,\ell})$ represents birthweight with the effect of smoking and race removed. In this way, only S_ℓ represents smoking and R_ℓ represents race, because their effect on birthweight has been removed. The other cases, for $t = 2, 3, 4$ continue in the same way.

The following estimates are the results of the fit; we only report the final models. At birth,

$$\widehat{y}_{0\ell} = 3.295 - .011_{(.002)} S_\ell + .215_{(.056)} R_\ell,$$

with $\widehat{\sigma}_0 = .472$; estimated standard errors are shown in parenthesis. We conclude that prenatal smoking significantly reduces birthweight, white babies are born slightly bigger, and no interaction exists between smoking and race. At eight months,

$$\begin{aligned} \widehat{y}_{1\ell} &= -.015_{(.011)} S_\ell - .335_{(.147)} R_\ell \\ &+ .029_{(.012)} S_\ell R_\ell + .214_{(.127)} (y_{0,\ell} - \widehat{y}_{0,\ell}), \end{aligned}$$

with $\widehat{\sigma}_1 = 1.066$. The interaction term is significant, indicating that white, unexposed babies are slightly smaller than the others.

The estimated model for 18 months is,

$$\begin{aligned} \widehat{y}_{2\ell} &= .340 + .278_{(.125)} R_\ell \\ &+ .661_{(.056)} (y_{1,\ell} - \widehat{y}_{1,\ell}) + .357_{(.126)} (y_{0,\ell} - \widehat{y}_{0,\ell}), \end{aligned}$$

with $\widehat{\sigma}_2 = 1.059$. Now, the effect of prenatal smoking is gone at 18 months, and, at this age, the white kids tend to be larger. The result at 36 months ($t = 3$) is that prenatal smoking becomes significant again, but exposed children are slightly bigger at this age, and race is no longer significant (this result is not as unusual as it might seem; in fact, it has been hypothesized that children exposed prenatally to cigarette smoke tend to become obese as they grow older):

$$\begin{aligned} \widehat{y}_{3\ell} &= .334 + .008_{(.004)} S_\ell + .310_{(.044)} (y_{2,\ell} - \widehat{y}_{2,\ell}) \\ &+ .450_{(.043)} (y_{1,\ell} - \widehat{y}_{1,\ell}) + .465_{(.098)} (y_{0,\ell} - \widehat{y}_{0,\ell}), \end{aligned}$$

with $\widehat{\sigma}_2 = .817$. Finally, the result for 72 months is

$$\begin{aligned}\widehat{y}_{4\ell} &= .330 + .933_{(.082)}(y_{3,\ell} - \widehat{y}_{3,\ell}) \\ &+ .462_{(.063)}(y_{2,\ell} - \widehat{y}_{2,\ell}) + .484_{(.062)}(y_{1,\ell} - \widehat{y}_{1,\ell}),\end{aligned}$$

with $\widehat{\sigma}_2 = 1.176$. At this age, the effect of prenatal smoking and the effect of race are gone. Also growth at eight months ($t = 1$) is still a predictor of growth at 72 months, but the effect of birthweight ($t = 0$) is gone.

Mixed Linear Models in State-Space Form

A widely used general mixed model for longitudinal data was introduced by Laird and Ware (1982). In this case, responses $\boldsymbol{y}_\ell = \{y_{t,\ell}, t = 1, \ldots, n_\ell\}$ are obtained on N subjects, $\ell = 1, \ldots, N$. Each response vector is modeled as

$$\boldsymbol{y}_\ell = X_\ell \boldsymbol{\beta} + Z_\ell \boldsymbol{\gamma}_\ell + \boldsymbol{\epsilon}_\ell, \tag{6.221}$$

where X_ℓ is an $n_\ell \times b$ design matrix, $\boldsymbol{\beta}$ is a $b \times 1$ vector of fixed parameters, and Z_ℓ is an $n_\ell \times g$ design matrix corresponding to the random $g \times 1$ vector of random effects, $\boldsymbol{\gamma}_\ell$, which is assumed to be independent across subject, and distributed as $\boldsymbol{\gamma}_\ell \sim \mathrm{N}(\mathbf{0}, D)$, where $D > 0$ is an arbitrary variance–covariance matrix. The within-subject errors, $\boldsymbol{\epsilon}_\ell$, are independently distributed as $\boldsymbol{\epsilon}_\ell \sim \mathrm{N}(\mathbf{0}, \Sigma_\ell)$; often, Σ_ℓ is of the form $\sigma^2 I$. A good introduction to these models can be found in many texts; for example, Diggle et al. (1994), Jones (1993), and Fahrmeir and Tutz (1994). Jones (1993) focuses on the state-space approach, and so will we.

The model, (6.221), can be written as

$$\boldsymbol{y}_\ell \sim \mathrm{N}\left(X_\ell \boldsymbol{\beta}, V_\ell\right), \tag{6.222}$$

independently, for $\ell = 1, \ldots, N$, where

$$V_\ell = Z_\ell D Z_\ell' + \Sigma_\ell. \tag{6.223}$$

An example of a typical covariance structure for V_ℓ is compound symmetry, wherein $g = 1$, Z_ℓ is a vector of ones, $D = \sigma_\gamma^2$ is a scalar, and $\Sigma_\ell = \sigma^2 I$. In this way, V_ℓ is an $n_\ell \times n_\ell$ matrix given by

$$V_\ell = \begin{pmatrix} \sigma^2 + \sigma_\gamma^2 & \sigma_\gamma^2 & \ldots & \sigma_\gamma^2 \\ \sigma_\gamma^2 & \sigma^2 + \sigma_\gamma^2 & \ldots & \sigma_\gamma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\gamma^2 & \sigma_\gamma^2 & \ldots & \sigma^2 + \sigma_\gamma^2 \end{pmatrix}. \tag{6.224}$$

Another useful covariance structure is the autoregressive structure, where $g = 0$ (that is, no random effects exist) and

$$V_\ell = \Sigma_\ell = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n_\ell - 1} \\ \rho & 1 & \rho & \ldots & \rho^{n_\ell - 2} \\ \vdots & \vdots & \ddots & & \vdots \\ \rho^{n_\ell - 1} & \rho^{n_\ell - 2} & & \ldots & 1 \end{pmatrix}, \tag{6.225}$$

with $|\rho| < 1$.

For a particular subject, ℓ, the vector $\boldsymbol{y}_\ell$ consists of observations, $y_{t\ell}$, taken over time $t = 1, 2, \ldots, n_\ell$. For subject ℓ, model (6.221) states

$$y_{t\ell} = \boldsymbol{x}'_{t\ell}\boldsymbol{\beta} + \boldsymbol{z}'_{t\ell}\boldsymbol{\gamma}_\ell + \epsilon_{t\ell}, \tag{6.226}$$

where $\boldsymbol{x}'_{t\ell}$ is the t-th row of X_ℓ and $\boldsymbol{z}'_{t\ell}$ is the t-th row of Z_ℓ. Using the form of the model given by (6.226), $y_{t\ell}$ is normal with

$$\begin{aligned} E(y_{t\ell}) &= \boldsymbol{x}'_{t\ell}\boldsymbol{\beta}, \\ \operatorname{cov}(y_{t\ell}, y_{s\ell}) &= \boldsymbol{z}'_{t\ell} D \boldsymbol{z}_{s\ell} + \sigma_{\ell,ts}, \\ \operatorname{cov}(y_{t\ell}, y_{sk}) &= 0 \qquad \ell \neq k, \end{aligned}$$

where $\sigma_{\ell,ts}$ is the ts-th element of Σ_ℓ. For the example given in (6.224), we would have

$$\operatorname{var}(y_{t\ell}) = \sigma^2 + \sigma^2_\gamma \quad \text{and} \quad \operatorname{cov}(y_{t\ell}, y_{s\ell}) = \sigma^2_\gamma,$$

for any $t \neq s$, so the correlation between two observations on the same subject is given by $\rho = \sigma^2_\gamma/(\sigma^2+\sigma^2_\gamma)$. In the autoregressive case, (6.225), the correlation between two observations $y_{t\ell}$ and $y_{s\ell}$ on the same subject $t-s$ time units apart is, of course, $\rho^{|t-s|}$.

The Laird–Ware model has a state space formulation; Jones (1993) provides a detailed presentation of these and related topics. If random effects exist, that is $g \geq 1$, and $\Sigma_\ell = \sigma^2 I$, let $\boldsymbol{s}_{t,\ell}$ denote a $g \times 1$ state vector with initial condition $\boldsymbol{s}_{0,\ell} \sim \mathrm{N}(\boldsymbol{0}, D)$. Then, for each $\ell = 1, \ldots, N$, (6.226) can be written as

$$\boldsymbol{s}_{t,\ell} = \boldsymbol{s}_{t-1,\ell} + \boldsymbol{w}_{t,\ell}, \tag{6.227}$$

$$y_{t\ell} = \boldsymbol{x}'_{t\ell}\boldsymbol{\beta} + \boldsymbol{z}'_{t\ell}\boldsymbol{s}_{t,\ell} + \epsilon_{t\ell}, \tag{6.228}$$

for $t = 1, \ldots, n_\ell$, where $\boldsymbol{w}_{t,\ell} \equiv \boldsymbol{0}$, or, equivalently, $\boldsymbol{w}_{t,\ell} \sim \mathrm{N}(\boldsymbol{0}, Q)$, where $Q = 0$ is the zero matrix. All other values are as defined in (6.226). The data $y_{t\ell}$ as written in (6.227)-(6.228) have the same properties as the data written in (6.226).

If $g = 0$, that is, no random effects exist, and the variance–covariance structure is autoregressive, as in (6.225), the state-space model can be written as

$$s_{t\ell} = \rho s_{t-1,\ell} + w_{t,\ell}, \tag{6.229}$$

$$y_{t\ell} = \boldsymbol{x}'_{t\ell}\boldsymbol{\beta} + s_{t\ell}, \tag{6.230}$$

where, now, the autoregressive structure is entered into the data via the (scalar, in this example) state, and there is no measurement error. In this case, $R = 0$, which does not present a problem in running the Kalman filter, provided $P^0_0 > 0$. To obtain a matrix of the form given in (6.225), $w_{t\ell}$ is white Gaussian noise, with $Q = \sigma^2$, and the initial state satisfies $s_{0,\ell} \sim \mathrm{N}(0, \sigma^2/(1-\rho^2))$. In this case, recall the states, $s_{t\ell}$, for a given subject ℓ, form a stationary AR(1) process with variance $\sigma^2/(1-\rho^2)$ and ACF given by $\rho(h) = \rho^{|h|}$.

In the more general case in which both random effects, $g > 0$, and an autoregressive error structure exist, we can combine the ideas used to get (6.227)-(6.228) and (6.229)-(6.230). In this case, the state equation would be a $(g+1)\times 1$ process made by stacking (6.227) and (6.229), and the observation equation would be

$$y_{t\ell} = \boldsymbol{x}'_{t\ell}\boldsymbol{\beta} + A_t s_{t\ell},$$

where $A_t = [z'_{t\ell}, 1]$.

We immediately see from (6.227)-(6.228), or from (6.229)-(6.230), that the likelihood of the data is the same as the one given in (6.209), but with n set to n_ℓ. Consequently, the methods presented in §5.3 can be used to estimate the parameters of the Laird–Ware model, namely, $\boldsymbol{\beta}$, and variance components in V_ℓ, for $\ell = 1, \ldots, n_\ell$. For simplicity, let Θ represent the vector of all of the parameters associated with the model.

In the notation of the algorithm presented in §6.3, Step 1 is to find initial estimates, $\Theta^{(0)}$, of the parameters Θ. If the V_ℓ were known, using a weighted least squares argument (see §4.4), the least squares estimate of $\boldsymbol{\beta}$ in the model (6.222)-(6.223) is given by

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{\ell=1}^{N} X'_\ell V_\ell^{-1} X_\ell\right)^{-1} \left(\sum_{\ell=1}^{N} X'_\ell V_\ell^{-1} \boldsymbol{y}_\ell\right). \tag{6.231}$$

Initial guesses for V_ℓ should reflect the variance–covariance structure of the model. We can use (6.231) with the initial values chosen for V_ℓ to obtain the initial regression coefficients, $\boldsymbol{\beta}^{(0)}$.

To accomplish Step 2 of the algorithm, for each $\ell = 1, \ldots, N$, run the Kalman filter (Property P6.1 with the states denoted by $\boldsymbol{s}_t$) for $t = 1, \ldots, n_\ell$ to obtain the initial innovations and their covariance matrices. For example, if the model is of the form given in (6.227)-(6.228), run the Kalman filter with $\Phi = I$, $Q = 0$, $A_t = \boldsymbol{z}'_{t\ell}$, $R = [\sigma^{(0)}]^2$, and initial conditions $\boldsymbol{s}_0^0 = \mathbf{0}$, $P_0^0 = D^{(0)} > 0$. In addition, $y_{t\ell}$ replaced by $y_{t\ell} - \boldsymbol{x}'_{t\ell}\boldsymbol{\beta}^{(0)}$; this is also equivalent to running Property P6.6 with uncorrelated noises, wherein the rows of the fixed effects design matrix, X_ℓ, are the exogenous variables. The Newton–Raphson procedure (steps 3 and 4 of the algorithm in §6.3) is performed on the criterion function given in (6.209). The following example may help in understanding the technique.

Example 6.24 Response to Medication

As a simple example of how we can use the state-space formulation of the Laird–Ware model, we analyze the S+ data set *drug.mult*. The data are taken from an experiment in which six subjects are given a dose of medication and then observed immediately and at weekly intervals for three weeks. The data are given in Table 6.6.

Table 6.6 Weekly Response to Medication

		Week 0	Week 1	Week 2	Week 3
ℓ	Gender	y_1	y_2	y_3	y_4
1	F	75.9	74.3	80.0	78.9
2	F	78.3	75.5	79.6	79.2
3	F	80.3	78.2	80.4	76.2
4	M	80.7	77.2	82.0	83.8
5	M	80.3	78.6	81.4	81.5
6	M	80.1	81.1	81.9	86.4

We fit model (6.229)-(6.230) to this data using gender as a grouping variable. In particular, if $\boldsymbol{y}_{t\ell}$ is the 4×1 vector of observations over time for a female ($\ell = 1, 2, 3$), the model is

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix},$$

and for a male ($\ell = 4, 5, 6$), the model is

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix},$$

where the ϵ_t, in general, form an AR(1) process given by

$$\begin{aligned} \epsilon_0 &= w_0/\sqrt{1-\rho^2}, \\ \epsilon_t &= \rho\epsilon_{t-1} + w_t \quad t = 1, 2, 3, 4, \end{aligned}$$

where w_t is white Gaussian noise, with $\text{var}(w_t) = \sigma_w^2$. Recall $\text{var}(\epsilon_t) = \sigma_\epsilon^2 = \sigma_w^2/(1-\rho^2)$ and $\rho_\epsilon(h) = \rho^{|h|}$. A different value of ρ was selected for each gender group, say, ρ_1 for female subjects and ρ_2 for male subjects.

We initialized the estimation procedure with $\rho_1^{(0)} = \rho_2^{(0)} = 0$, $\sigma_w^{(0)} = 2$, which, upon using (6.231), yields $\boldsymbol{\beta}^{(0)} = (78.07, 3.18)'$. The final estimates (and their estimated standard errors) were

$$\widehat{\beta}_1 = 78.20(.56), \quad \widehat{\beta}_2 = 3.24(.89),$$

$$\widehat{\rho}_1 = -.47(.36), \quad \widehat{\rho}_2 = .07(.53), \quad \widehat{\sigma}_w = 2.17(.36).$$

Because $\widehat{\rho}_1$ and $\widehat{\rho}_2$ are not significantly different from zero, this would suggest either a simple linear regression is sufficient to describe the results, or the model is not correct.

Next, we fit the compound symmetry model using (6.227)-(6.228) with $g = 1$. In this case, the model for a female subject is

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \gamma_1 + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix},$$

and for a male subject, the model is

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \gamma_2 + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix},$$

where $\gamma_1 \sim \mathrm{N}(0, \sigma^2_{\gamma_1})$, $\gamma_2 \sim \mathrm{N}(0, \sigma^2_{\gamma_2})$, and the ϵ_t, for $t = 1, 2, 3, 4$ are uncorrelated with variance σ^2_ϵ.

In this case, the state variable is a scalar process with $D = \sigma^2_{\gamma_1}$ for female subjects ($\ell = 1, 2, 3$) and $D = \sigma^2_{\gamma_2}$ for male subjects ($\ell = 4, 5, 6$). Starting the estimation process off with $\sigma^{(0)}_{\gamma_1} = \sigma^{(0)}_{\gamma_2} = 1$, $\sigma^{(0)}_\epsilon = 2$, and $\boldsymbol{\beta}^{(0)} = (78, 3)'$, the final estimates were

$$\widehat{\beta}_1 = 78.03\,(.67), \quad \widehat{\beta}_2 = 3.51\,(1.05),$$

$$\widehat{\sigma}_{\gamma_1} = 2.05\,(.45), \quad \widehat{\sigma}_{\gamma_2} = 2.51\,(.59), \quad \widehat{\sigma}_\epsilon = 2.00\,(.13).$$

This model fits the data well.

Problems

Section 6.1

6.1 Consider a system process given by

$$x_t = -.9x_{t-2} + w_t \quad t = 1, \ldots, n$$

where $x_0 \sim \mathrm{N}(0, \sigma^2_0)$, $x_{-1} \sim N(0, \sigma^2_1)$, and w_t is Gaussian white noise with variance σ^2_w. The system process is observed with noise, say,

$$y_t = x_t + v_t,$$

where v_t is Gaussian white noise with variance σ^2_v. Further, suppose x_0, x_{-1}, $\{w_t\}$ and $\{v_t\}$ are independent.

(a) Write the system and observation equations in the form of a state space model.

(b) Find the values of σ_0^2 and σ_1^2 that make the observations, y_t, stationary.

(c) Generate $n = 100$ observations with $\sigma_w = 1$, $\sigma_v = 1$ and using the values of σ_0^2 and σ_1^2 found in (b). Do a time plot of x_t and of y_t and compare the two processes. Also, compare the sample ACF and PACF of x_t and of y_t.

(d) Repeat (c), but with $\sigma_v = 10$.

6.2 Consider the state-space model presented in Example 6.3. Let $x_t^{t-1} = E(x_t|y_{t-1}, \ldots, y_1)$ and let $P_t^{t-1} = E(x_t - x_t^{t-1})^2$. The innovation sequence or residuals are $\epsilon_t = y_t - y_t^{t-1}$, where $y_t^{t-1} = E(y_t|y_{t-1}, \ldots, y_1)$. Find $\mathrm{cov}(\epsilon_s, \epsilon_t)$ in terms of x_t^{t-1} and P_t^{t-1} for (i) $s \neq t$ and (ii) $s = t$.

Section 6.2

6.3 Simulate $n = 100$ observations from the following state-space model:

$$x_t = .8x_{t-1} + w_t \quad \text{and} \quad y_t = x_t + v_t$$

where $x_0 \sim \mathrm{N}(0, 2.78)$, $w_t \sim$ iid $\mathrm{N}(0, 1)$, and $v_t \sim$ iid $\mathrm{N}(0, 1)$ are all mutually independent. Compute and plot the data, y_t, the one-step-ahead predictors, y_t^{t-1} along with the root mean square prediction errors, $E^{1/2}(y_t - y_t^{t-1})^2$ using Figure 6.3 as a guide.

6.4 Suppose the vector $\boldsymbol{z} = (\boldsymbol{x}', \boldsymbol{y}')'$, where $\boldsymbol{x}$ $(p \times 1)$ and $\boldsymbol{y}$ $(q \times 1)$ are jointly distributed with mean vectors $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ and with covariance matrix

$$\mathrm{cov}(\boldsymbol{z}) = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}.$$

Consider projecting $\boldsymbol{x}$ on $\mathcal{M} = \overline{\mathrm{sp}}\{\mathbf{1}, \boldsymbol{y}\}$, say, $\widehat{\boldsymbol{x}} = \boldsymbol{b} + B\boldsymbol{y}$.

(a) Show the orthogonality conditions can be written as

$$E(\boldsymbol{x} - \boldsymbol{b} - B\boldsymbol{y}) = 0,$$

$$E[(\boldsymbol{x} - \boldsymbol{b} - B\boldsymbol{y})\boldsymbol{y}'] = 0,$$

leading to the solutions

$$\boldsymbol{b} = \boldsymbol{\mu}_x - B\boldsymbol{\mu}_y \quad \text{and} \quad B = \Sigma_{xy}\Sigma_{yy}^{-1}.$$

(b) Prove the mean square error matrix is

$$MSE = E[(\boldsymbol{x} - \boldsymbol{b} - B\boldsymbol{y})\boldsymbol{x}'] = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}.$$

(c) How can these results be used to justify the claim that, in the absence of normality, Property P6.1 yields the best linear estimate of the state $\boldsymbol{x}_t$ given the data Y_t, namely, $\boldsymbol{x}_t^t$, and its corresponding MSE, namely, P_t^t?

6.5 *Derivation of Property P6.2 Based on the Projection Theorem.* Throughout this problem, we use the notation of Property P6.2 and of the Projection Theorem given in Appendix B, where $\mathcal{H}$ is L^2. If $\mathcal{L}_{k+1} = \overline{\text{sp}}\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{k+1}\}$, and $\mathcal{V}_{k+1} = \overline{\text{sp}}\{\boldsymbol{y}_{k+1} - \boldsymbol{y}_{k+1}^k\}$, for $k = 0, 1, \ldots, n-1$, where $\boldsymbol{y}_{k+1}^k$ is the projection of $\boldsymbol{y}_{k+1}$ on $\mathcal{L}_k$, then, $\mathcal{L}_{k+1} = \mathcal{L}_k \oplus \mathcal{V}_{k+1}$. We assume $P_0^0 > 0$ and $R > 0$.

(a) Show the projection of $\boldsymbol{x}_k$ on $\mathcal{L}_{k+1}$, that is, $\boldsymbol{x}_k^{k+1}$, is given by

$$\boldsymbol{x}_k^{k+1} = \boldsymbol{x}_k^k + H_{k+1}(\boldsymbol{y}_{k+1} - \boldsymbol{y}_{k+1}^k),$$

where H_{k+1} can be determined by the orthogonality property

$$E\left\{\left(\boldsymbol{x}_k - H_{k+1}(\boldsymbol{y}_{k+1} - \boldsymbol{y}_{k+1}^k)\right)\left(\boldsymbol{y}_{k+1} - \boldsymbol{y}_{k+1}^k\right)'\right\} = 0.$$

Show

$$H_{k+1} = P_k^k \Phi' A_{k+1}' \left[A_{k+1} P_{k+1}^k A_{k+1}' + R\right]^{-1}.$$

(b) Define $J_k = P_k^k \Phi' [P_{k+1}^k]^{-1}$, and show

$$\boldsymbol{x}_k^{k+1} = \boldsymbol{x}_k^k + J_k(\boldsymbol{x}_{k+1}^{k+1} - \boldsymbol{x}_{k+1}^k).$$

(c) Repeating the process, show

$$\boldsymbol{x}_k^{k+2} = \boldsymbol{x}_k^k + J_k(\boldsymbol{x}_{k+1}^{k+1} - \boldsymbol{x}_{k+1}^k) + H_{k+2}(\boldsymbol{y}_{k+2} - \boldsymbol{y}_{k+2}^{k+1}),$$

solving for H_{k+2}. Simplify and show

$$\boldsymbol{x}_k^{k+2} = \boldsymbol{x}_k^k + J_k(\boldsymbol{x}_{k+1}^{k+2} - \boldsymbol{x}_{k+1}^k).$$

(d) Using induction, conclude

$$\boldsymbol{x}_k^n = \boldsymbol{x}_k^k + J_k(\boldsymbol{x}_{k+1}^n - \boldsymbol{x}_{k+1}^k),$$

which yields the smoother with $k = t - 1$.

Section 6.3

6.6 (a) Consider the univariate state-space model given by state conditions $x_0 = w_0$, $x_t = x_{t-1} + w_t$ and observations $y_t = x_t + v_t$, $t = 1, 2, \ldots$, where w_t and v_t are independent, Gaussian, white noise processes with $\text{var}(w_t) = \sigma_w^2$ and $\text{var}(v_t) = \sigma_v^2$. Show the data follow an IMA(1,1) model, that is, ∇y_t follows an MA(1) model.

(b) Fit the model specified in part (a) to the logarithm of the glacial varve series and compare the results to those presented in Example 3.31.

6.7 Let y_t represent the land-based global temperature series shown in Figure 6.2. The data file for this problem is `HL.dat` on the website.

(a) Using regression, fit a third-degree polynomial in time to y_t, that is, fit the model

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon_t,$$

where ϵ_t is white noise. Do a time plot of the data fit, $\widehat{y}_t$, superimposed on the data, y_t, for t=1,...,108.

(b) Write the model $y_t = x_t + v_t$ with $\nabla^2 x_t = w_t$, where w_t and v_t are independent white noise processes, in state-space form. Hint: The state will be a 2×1 vector, say, $\boldsymbol{x}_t = (x_t, x_{t-1})'$. Fit the state-space model to the data, and do a time plot of the estimated filter, $\widehat{x}_t^{t-1}$, and the estimated smoother, $\widehat{x}_t^n$, superimposed on the data, y_t, for t=1,...,108. Compare these results with the results of part (a).

6.8 Consider the model

$$y_t = x_t + v_t,$$

where v_t is Gaussian white noise with variance σ_v^2, x_t are independent Gaussian random variables with mean zero and $\mathrm{var}(x_t) = r_t \sigma_x^2$ with x_t independent of v_t, and $r_1, \ldots, r_n$ are known constants. Show that applying the EM algorithm to the problem of estimating σ_x^2 and σ_v^2 leads to updates (represented by hats)

$$\widehat{\sigma}_x^2 = \frac{1}{n}\sum_{t=1}^{n} \frac{\sigma_t^2 + \mu_t^2}{r_t} \quad \text{and} \quad \widehat{\sigma}_v^2 = \frac{1}{n}\sum_{t=1}^{n}[(y_t - \mu_t)^2 + \sigma_t^2],$$

where, based on the current estimates (represented by tildes),

$$\mu_t = \frac{r_t \widetilde{\sigma}_x^2}{r_t \widetilde{\sigma}_x^2 + \widetilde{\sigma}_v^2}\, y_t \quad \text{and} \quad \sigma_t^2 = \frac{r_t \widetilde{\sigma}_x^2 \widetilde{\sigma}_v^2}{r_t \widetilde{\sigma}_x^2 + \widetilde{\sigma}_v^2}.$$

6.9 Develop the EM algorithm for the model with inputs, (6.3) and (6.4).

6.10 To explore the stability of the filter, consider a univariate state-space model. That is, for $t = 1, 2, \ldots$, the observations are $y_t = x_t + v_t$ and the state equation is $x_t = \phi x_{t-1} + w_t$, where $\sigma_w = \sigma_v = 1$ and $|\phi| < 1$. The initial state, x_0, has zero mean and variance one.

(a) Exhibit the recursion for P_t^{t-1} in Property P6.1 in terms of P_{t-1}^{t-2}.

(b) Use the result of (a) to verify P_t^{t-1} approaches a limit ($t \to \infty$) P that is the positive solution of $P^2 - \phi^2 P - 1 = 0$.

(c) With $K = \lim_{t\to\infty} K_t$ as given in Property P6.1, show $|1 - K| < 1$.

(d) Show, in steady-state, the one-step-ahead predictor, $y_{n+1}^n = E(y_{n+1} \mid y_n, y_{n-1}, \ldots)$, of a future observation satisfies

$$y_{n+1}^n = \sum_{j=0}^{\infty} \phi^j K(1-K)^{j-1} y_{n+1-j}.$$

6.11 In §6.3, we discussed that it is possible to obtain a recursion for the gradient vector, $-\partial \ln L_Y(\Theta)/\partial\Theta$. Assume the model is given by (6.1) and (6.2) and A_t is a known design matrix that does not depend on Θ, in which case Property P6.1 applies. For the gradient vector, show

$$\begin{aligned} \partial \ln L_Y(\Theta)/\partial\Theta_i &= \sum_{t=1}^{n} \left\{ \boldsymbol{\epsilon}_t' \Sigma_t^{-1} \frac{\partial \boldsymbol{\epsilon}_t}{\partial \Theta_i} - \frac{1}{2} \boldsymbol{\epsilon}_t' \Sigma_t^{-1} \frac{\partial \Sigma_t}{\partial \Theta_i} \Sigma_t^{-1} \boldsymbol{\epsilon}_t \right. \\ &\qquad \left. + \frac{1}{2} \mathrm{tr} \left(\Sigma_t^{-1} \frac{\partial \Sigma_t}{\partial \Theta_i} \right) \right\}, \end{aligned}$$

where the dependence of the innovation values on Θ is understood. In addition, with the general definition $\partial_i g = \partial g(\Theta)/\partial\Theta_i$, show the following recursions, for $t = 2, \ldots, n$ apply:

- $\partial_i \boldsymbol{\epsilon}_t = -A_t \; \partial_i \boldsymbol{x}_t^{t-1}$,
- $\partial_i \boldsymbol{x}_t^{t-1} = \; \partial_i \Phi \; \boldsymbol{x}_{t-1}^{t-2} + \Phi \; \partial_i \boldsymbol{x}_{t-1}^{t-2} + \; \partial_i K_{t-1} \; \boldsymbol{\epsilon}_{t-1} + K_{t-1} \; \partial_i \boldsymbol{\epsilon}_{t-1}$,
- $\partial_i \Sigma_t = A_t \; \partial_i P_t^{t-1} A_t' + \; \partial_i R$,
- $\partial_i K_t = \left[\; \partial_i \Phi \; P_t^{t-1} A_t' + \Phi \; \partial_i P_t^{t-1} \; A_t' - K_t \; \partial_i \Sigma_t \right] \; \Sigma_t^{-1}$,
- $\partial_i P_t^{t-1} = \; \partial_i \Phi \; P_{t-1}^{t-2} \Phi' + \Phi \; \partial_i P_{t-1}^{t-2} \; \Phi' + \Phi \; P_{t-1}^{t-2} \; \partial_i \Phi' + \; \partial_i Q,$
  $\qquad - \; \partial_i K_{t-1} \; \Sigma_t K_{t-1}' - K_{t-1} \; \partial_i \Sigma_t \; K_{t-1}' - K_{t-1} \Sigma_t \; \partial_i K_{t-1}',$

using the fact that $P_t^{t-1} = \Phi P_{t-1}^{t-2} \Phi' + Q - K_{t-1} \Sigma_t K_{t-1}'$.

6.12 Continuing with the previous problem, consider the evaluation of the Hessian matrix and the numerical evaluation of the asymptotic variance–covariance matrix of the parameter estimates. The information matrix satisfies

$$E\left\{ -\frac{\partial^2 \ln L_Y(\Theta)}{\partial\Theta \; \partial\Theta'} \right\} = E\left\{ \left(\frac{\partial \ln L_Y(\Theta)}{\partial\Theta} \right) \left(\frac{\partial \ln L_Y(\Theta)}{\partial\Theta} \right)' \right\};$$

see Anderson (1984, Section 4.4), for example. Show the (i, j)-th element of the information matrix, say, $\mathcal{I}_{ij}(\Theta) = E\left\{ -\partial^2 \ln L_Y(\Theta)/\partial\Theta_i \; \partial\Theta_j \right\}$, is

$$\begin{aligned} \mathcal{I}_{ij}(\Theta) &= \sum_{t=1}^{n} E\Big\{ \; \partial_i \boldsymbol{\epsilon}_t' \; \Sigma_t^{-1} \; \; \partial_j \boldsymbol{\epsilon}_t + \frac{1}{2} \mathrm{tr}\big(\Sigma_t^{-1} \; \partial_i \Sigma_t \; \Sigma_t^{-1} \; \partial_j \Sigma_t \big) \\ &\qquad + \frac{1}{4} \mathrm{tr}\left(\Sigma_t^{-1} \; \partial_i \Sigma_t \right) \mathrm{tr}\big(\Sigma_t^{-1} \; \partial_j \Sigma_t \big) \Big\}. \end{aligned}$$

Consequently, an approximate Hessian matrix can be obtained from the sample by dropping the expectation, E, in the above result and using only the recursions needed to calculate the gradient vector.

Section 6.4

6.13 As an example of the way the state-space model handles the missing data problem, suppose the first-order autoregressive process

$$x_t = \phi x_{t-1} + w_t$$

has an observation missing at $t = m$, leading to the observations $y_t = A_t x_t$, where $A_t = 1$ for all t, except $t = m$ wherein $A_t = 0$. Assume $x_0 = 0$ with variance $\sigma_w^2/(1-\phi^2)$, where the variance of w_t is σ_w^2. Show the Kalman smoother estimators in this case are

$$x_t^n = \begin{cases} \phi y_1, & t = 0, \\ \frac{\phi}{1+\phi^2}(y_{m-1} + y_{m+1}), & t = m, \\ y_t, & t \neq 0, m, \end{cases}$$

with mean square covariances determined by

$$P_t^n = \begin{cases} \sigma_w^2, & t = 0, \\ \frac{\sigma_w^2}{1+\phi^2}, & t = m, \\ 0 & t \neq 0, m. \end{cases}$$

6.14 The data set labeled `ar1miss.dat` is $n = 100$ observations generated from an AR(1) process, $x_t = \phi x_{t-1} + w_t$, with $\phi = .9$ and $\sigma_w = 1$, where 10% of the data has been zeroed out at random. Considering the zeroed out data to be missing data, use the results of Problem 6.13 to estimate the parameters of the model, ϕ and σ_w, using the EM algorithm, and then estimate the missing values.

Section 6.5

6.15 Using Example 6.10 as a guide, fit a structural model to the Federal Reserve Board Production Index data and compare it with the model fit in Example 3.43.

Section 6.6

6.16 Use Property P6.6 to complete the following exercises.

(a) Write a univariate AR(1) model, $y_t = \phi y_{t-1} + v_t$, in state-space form. Verify your answer is indeed an AR(1).

(b) Repeat (a) for an MA(1) model, $y_t = v_t + \theta v_{t-1}$.

(c) Write an IMA(1,1) model, $y_t = y_{t-1} + v_t + \theta v_{t-1}$, in state-space form.

6.17 Verify Property P6.5.

6.18 Verify Property P6.6.

Section 6.7

6.19 Repeat the bootstrap analysis of Example 6.12 on the entire three-month treasury bills and rate of inflation data set of 110 observations. Do the conclusions of Example 6.12—that the dynamics of the data is best described in terms of a fixed, rather than stochastic, regression—still hold?

Section 6.8

6.20 Argue that a switching model is reasonable in explaining the behavior of the number of sunspots (see Figure 4.31) and then fit a switching model to the sunspot data.

Section 6.9

6.21 Use the material presented in Example 6.18 to perform a Bayesian analysis of the model for the Johnson & Johnson data presented in Example 6.10.

6.22 Verify (6.169) and (6.170).

6.23 Verify (6.175) and (6.182).

Section 6.10

6.24 Fit a stochastic volatility model to the returns of one (or more) of the four financial time series available in the R datasets package as `EuStockMarkets`.

Section 6.11

6.25 In a small pilot study, a psychiatrist wanted to examine the effects of the drug lithium on bulimics (bulimics have continuous abnormal hunger and frequently go on eating binges). Although evidence of the effectiveness of lithium on bulimics has been shown, he was not sure if depressed subjects would respond differently than those without depression. He treated eight teenage female patients with lithium for 12 weeks; four of the subjects were diagnosed with depression, and half of the subjects received behavioral therapy. At the end of each four-week period, he recorded the number of binges each subject had during that week. The following are the results:

Weekly Number of Binges

Subject	Depression	Week 0	Week 4	Week 8	Week 12
1	No	13	3	0	0
2	No	15	4	3	1
3	No	16	4	3	2
4	No	14	2	1	2
5	Yes	10	7	4	3
6	Yes	18	7	2	4
7	Yes	16	6	5	4
8	Yes	19	8	5	7

Fit a longitudinal model that addresses the concerns of the psychiatrist. Because the data are counts (number of occurrences), consider a square root transformation prior to the analysis.

Chapter 7

Statistical Methods in the Frequency Domain

7.1 Introduction

In previous chapters, we saw many applied time series problems that involved relating series to each other or to evaluating the effects of treatments or design parameters that arise when time-varying phenomena are subjected to periodic stimuli. In many cases, the nature of the physical or biological phenomena under study are best described by their Fourier components rather than by the difference equations involved in ARIMA or state-space models. The fundamental tools we use in studying periodic phenomena are the discrete Fourier transforms (DFTs) of the processes and their statistical properties. Hence, in §7.2, we review the properties of the DFT of a multivariate time series and discuss various approximations to the likelihood function based on the large-sample properties and the properties of the complex multivariate normal distribution. This enables extension of the classical techniques discussed in the following paragraphs to the multivariate time series case.

An extremely important class of problems in classical statistics develops when we are interested in relating a collection of input series to some output series. For example, in Chapter 2, we have previously considered relating temperature and various pollutant levels to daily mortality, but have not investigated the frequencies that appear to be driving the relation and have not looked at the possibility of leading or lagging effects. In Chapter 4, we isolated a definite lag structure that could be used to relate sea surface temperature to the number of new recruits. In Problem 5.11 of Chapter 5, the possible driving processes that could be used to explain inflow to Shasta Lake were hypothesized in terms of the possible inputs precipitation, cloud cover, temperature, and other variables. Identifying the combination of input factors in Figure 4.33

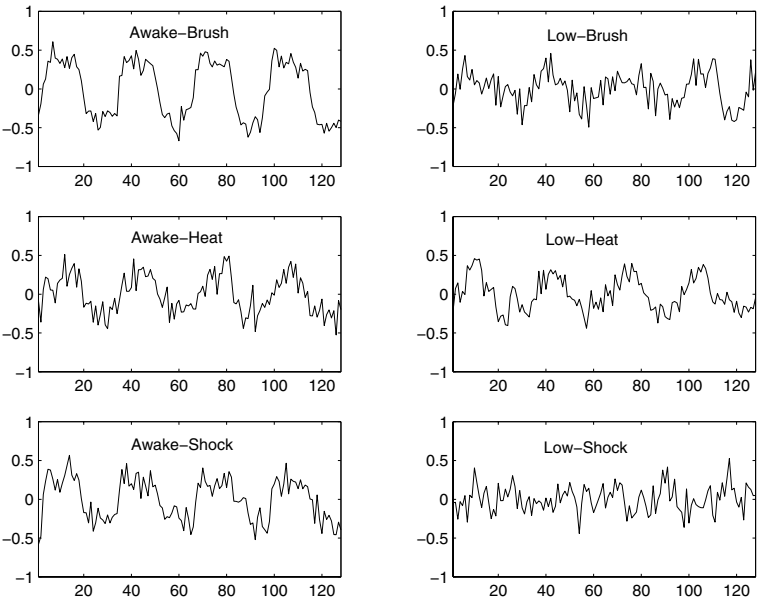


Figure 7.1 Mean response of subjects to various combinations of periodic stimulae measured at the cortex (primary somatosensory, contralateral).

that produce the best prediction for inflow is an example of multiple regression in the frequency domain, with the models treated theoretically by considering the regression, conditional on the random input processes.

A situation somewhat different from that above would be one in which the input series are regarded as fixed and known. In this case, we have a model analogous to that occurring in analysis of variance, in which the analysis now can be performed on a frequency by frequency basis. This analysis works especially well when the inputs are dummy variables, depending on some configuration of treatment and other design effects and when effects are largely dependent on periodic stimuli. As an example, we will look at a designed experiment measuring the fMRI brain responses of a number of awake and mildly anesthetized subjects to several levels of periodic brushing, heat, and shock effects. Some limited data from this experiment have been discussed previously in Example 1.6 of Chapter 1. Figure 7.1 shows mean responses to various levels of periodic heat, brushing, and shock stimuli for subjects awake and subjects under mild anesthesia. The stimuli were periodic in nature, applied alternately for 32 seconds (16 points) and then stopped for 32 seconds. The periodic input signal comes through under all three design conditions when the subjects are awake, but is somewhat attenuated under anesthesia. The mean shock level response hardly shows on the input signal; shock levels were designed to simulate surgical incision without inflicting tissue damage. The

means in Figure 7.1 are from a single location. Actually, for each individual, some nine series were recorded at various locations in the brain. It is natural to consider testing the effects of brushing, heat, and shock under the two levels of consciousness, using a time series generalization of analysis of variance.

A generalization to random coefficient regression is also considered, paralleling the univariate approach to signal extraction and detection presented in §4.9. This method enables a treatment of multivariate ridge-type regressions and inversion problems. Also, the usual random effects analysis of variance in the frequency domain becomes a special case of the random coefficient model.

The extension of frequency domain methodology to more classical approaches to multivariate discrimination and clustering is of interest in the frequency dependent case. Many time series differ in their means and in their autocovariance functions, making the use of both the mean function and the spectral density matrices relevant. As an example of such data, consider the bivariate series consisting of the P and S components derived from several earthquakes and explosions, such as those shown in Figure 7.2, where the P and S components, representing different arrivals have been separated from the first and second halves, respectively, of wave forms like those shown originally in Figure 1.7 of Chapter 1.

Two earthquakes and two explosions from a set of eight earthquakes and explosions are shown in Figure 7.2 and some essential differences exist that might be used to characterize the two classes of events. Also, the frequency content of the two components of the earthquakes appears to be lower than those of the explosions, and relative amplitudes of the two classes appear to differ. For example, the ratio of the S to P amplitudes in the earthquake group is much higher for this restricted subset. Spectral differences were also noticed in Chapter 4, where the explosion processes had a stronger high-frequency component relative to the low-frequency contributions. Examples like these are typical of applications in which the essential differences between multivariate time series can be expressed by the behavior of either the frequency-dependent mean value functions or the spectral matrix. In discriminant analysis, these types of differences are exploited to develop combinations of linear and quadratic classification criteria. Such functions can then be used to classify events of unknown origin, such as the Novaya Zemlya event shown in Figure 7.2, which tends to bear a visual resemblance to the explosion group.

Finally, for multivariate processes, the structure of the spectral matrix is also of great interest. We might reduce the dimension of the underlying process to a smaller set of input processes that explain most of the variability in the cross-spectral matrix as a function of frequency. Principal component analysis can be used to decompose the spectral matrix into a smaller subset of component factors that explain decreasing amounts of power. For example, the hydrological data might be explained in terms of a component process that weights heavily on precipitation and inflow and one that weights heavily on temperature and cloud cover. Perhaps these two components could explain most of the power in the spectral matrix at a given frequency. The ideas

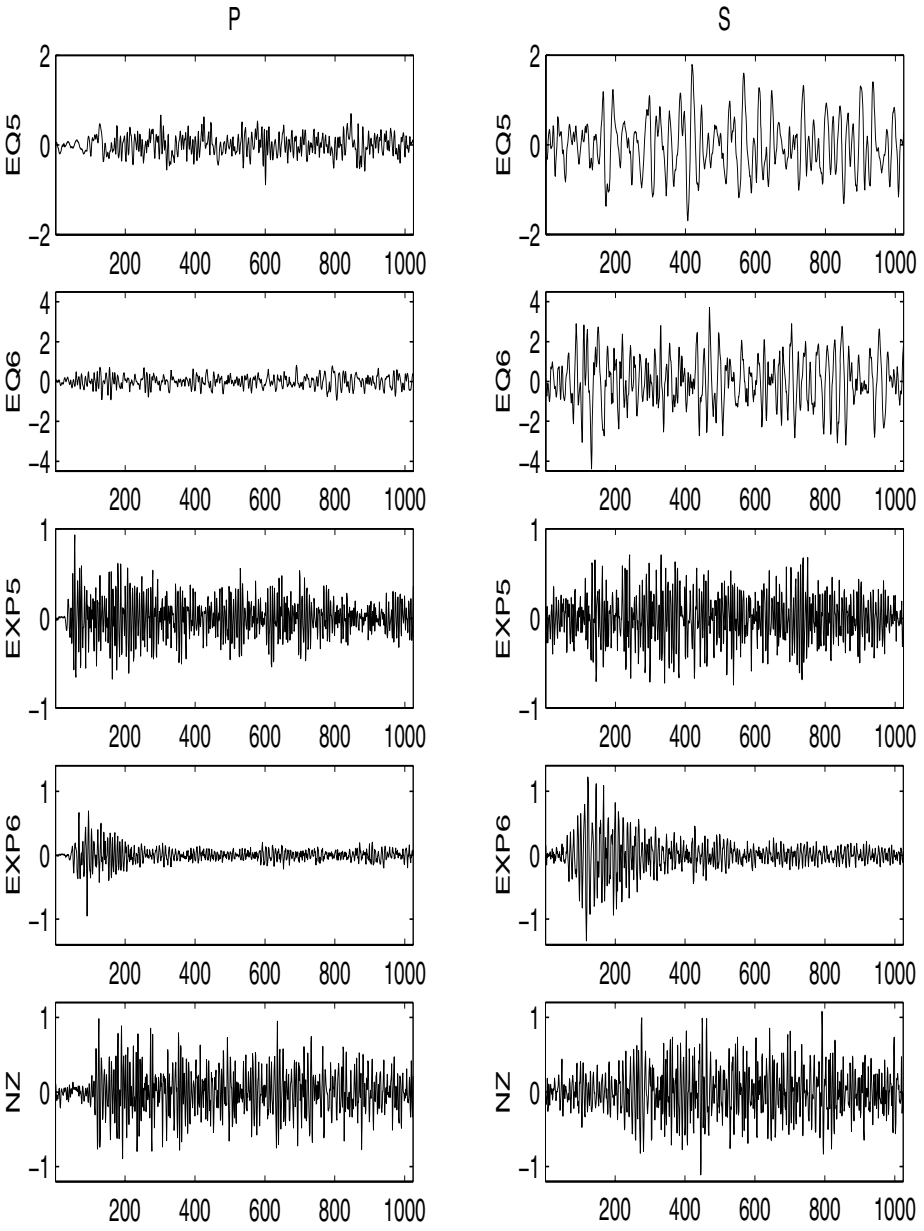


Figure 7.2 Bivariate earthquakes and explosions (40 pts/sec) compared with an event NZ (Novaya Zemlya) of unknown origin.

behind principal component analysis can also be generalized to include an optimal scaling methodology for categorical data called the spectral envelope (see Stoffer et al., 1993). In succeeding sections, we also give an introduction to dynamic Fourier analysis and to wavelet analysis.

7.2 Spectral Matrices and Likelihood Functions

We have previously argued for an approximation to the log likelihood based on the joint distribution of the DFTs in (4.116), where we used approximation as an aid in estimating parameters for certain parameterized spectra. In this chapter, we make heavy use of the fact that the sine and cosine transforms of the $p \times 1$ vector process $\boldsymbol{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tp})'$ with mean $E\boldsymbol{x}_t = \boldsymbol{\mu}_t$, say, with DFT[1]

$$\begin{aligned} \boldsymbol{X}(\omega_k) &= n^{-1/2} \sum_{t=1}^{n} \boldsymbol{x}_t \, \mathrm{e}^{-2\pi i \omega_k t} \\ &= \boldsymbol{X}_c(\omega_k) - i\boldsymbol{X}_s(\omega_k) \end{aligned} \tag{7.1}$$

and mean

$$\begin{aligned} \boldsymbol{M}(\omega_k) &= n^{-1/2} \sum_{t=1}^{n} \boldsymbol{\mu}_t \, \mathrm{e}^{-2\pi i \omega_k t} \\ &= \boldsymbol{M}_c(\omega_k) - i\boldsymbol{M}_s(\omega_k) \end{aligned} \tag{7.2}$$

will be approximately uncorrelated, where we evaluate at the usual Fourier frequencies $\{\omega_k = k/n,\ 0 < |\omega_k| < 1/2\}$. By Theorem C.6, the approximate $2p \times 2p$ covariance matrix of the cosine and sine transforms, say, $\boldsymbol{X}(\omega_k) = (\boldsymbol{X}_c(\omega_k)', \boldsymbol{X}_s(\omega_k)')'$, is

$$\Sigma(\omega_k) = \frac{1}{2} \begin{pmatrix} C(\omega_k) & -Q(\omega_k) \\ Q(\omega_k) & C(\omega_k) \end{pmatrix}, \tag{7.3}$$

and the real and imaginary parts are jointly normal. This result implies, by the results stated in Appendix C, the density function of the vector DFT, say, $\boldsymbol{X}(\omega_k)$, can be approximated as

$$p(\omega_k) \approx |f(\omega_k)|^{-1} \exp\big\{-\big(\boldsymbol{X}(\omega_k) - \boldsymbol{M}(\omega_k)\big)^* f^{-1}(\omega_k)\big(\boldsymbol{X}(\omega_k) - \boldsymbol{M}(\omega_k)\big)\big\},$$

where the spectral matrix is the usual

$$f(\omega_k) = C(\omega_k) - iQ(\omega_k). \tag{7.4}$$

[1]In previous chapters, the DFT of a process x_t was denoted by $d_x(\omega_k)$. In this chapter, we will consider the Fourier transforms of many different processes, and so to avoid the overuse of subscripts and hence, to ease the notation, we use a capital letter, e.g., $X(\omega_k)$, to denote Fourier transform of x_t.

Certain computations that we do in the section on discriminant analysis will involve approximating the joint likelihood by the product of densities like the one given above over subsets of the frequency band $0 < \omega_k < 1/2$.

To use the likelihood function for estimating the spectral matrix, for example, we appeal to the limiting result implied by Theorem C.7 and again choose L frequencies in the neighborhood of some target frequency ω, say, $\boldsymbol{X}(\omega_k \pm k/n)$, for $k = 1, \ldots, m$ and $L = 2m + 1$. Then, let $\boldsymbol{X}_\ell$, for $\ell = 1, \ldots, L$ denote the indexed values, and note the DFTs of the mean adjusted vector process are approximately jointly normal with mean zero and complex covariance matrix $f = f(\omega)$. Then, write the log likelihood over the L sub-frequencies as

$$\ln L(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_L; f) \approx -L \ln |f| - \sum_{\ell=1}^{L} (\boldsymbol{X}_\ell - \boldsymbol{M}_\ell)^* f^{-1} (\boldsymbol{X}_\ell - \boldsymbol{M}_\ell), \qquad (7.5)$$

where we have suppressed the argument of $f = f(\omega)$ for ease of notation. The use of spectral approximations to the likelihood has been fairly standard, beginning with the work of Whittle (1961) and continuing in Brillinger (1981) and Hannan (1970). In this case, assuming the mean adjusted series are available, i.e., that $\boldsymbol{M}_\ell$ is known, so that we may assume that $\boldsymbol{X}_\ell$ is the mean-adjusted series. We may obtain the maximum likelihood estimator for f by writing the joint log likelihood of the real and imaginary parts in terms of $\boldsymbol{Z}_\ell = (\boldsymbol{X}'_{c\ell}, \boldsymbol{X}'_{s\ell})'$ and obtaining the maximum likelihood estimators for C and Q, the real and imaginary parts of f. Problem 7.2 shows we will obtain

$$\widehat{f} = L^{-1} \sum_{\ell=1}^{L} (\boldsymbol{X}_\ell - \boldsymbol{M}_\ell)(\boldsymbol{X}_\ell - \boldsymbol{M}_\ell)^*, \qquad (7.6)$$

which is just the usual mean-adjusted estimator for the spectral matrix.

7.3 Regression for Jointly Stationary Series

In §4.8, we considered a model of the form

$$y_t = \sum_{r=-\infty}^{\infty} \beta_{1r} x_{t-r,1} + v_t, \qquad (7.7)$$

where x_{t1} is a single observed input series and y_t is the observed output series, and we are interested in estimating the filter coefficients β_{1r} relating the adjacent lagged values of x_{t1} to the output series y_t. In the case of the SOI and Recruitment series, we identified the El Niño driving series as x_{t1}, the input and y_t, the Recruitment series, as the output. In general, more than a single plausible input series may exist. For example, the hydrological data shown in Figure 4.33 suggests there may be at least five possible series driving the inflow. Hence, we may envision a $q \times 1$ input vector of driving series,

say, $\boldsymbol{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tq})'$, and a set of $q \times 1$ vector of regression functions $\boldsymbol{\beta}_r = (\beta_{1r}, \beta_{2r,}, \ldots, \beta_{qr})'$, which are related as

$$y_t = \sum_{r=-\infty}^{\infty} \boldsymbol{\beta}_r' \boldsymbol{x}_{t-r} + v_t. \tag{7.8}$$

Writing the matrix form out as

$$y_t = \sum_{j=1}^{q} \sum_{r=-\infty}^{\infty} \beta_{jr} x_{t-r,j} + v_t \tag{7.9}$$

shows the output is basically a sum of linearly filtered versions of the input processes and a stationary noise process v_t, assumed to be uncorrelated with $\boldsymbol{x}_t$. Each filtered component in the sum over j gives the contribution of lagged values of the j-th input series to the output series. We assume the regression functions β_{jr} are fixed and unknown.

The model given by (7.8) is useful under several different scenarios, corresponding to a number of different assumptions that can be made about the components. Assuming the input and output processes are jointly stationary with zero means leads to the conventional regression analysis given in this section. The analysis depends on theory that assumes we observe the output process y_t conditional on fixed values of the input vector $\boldsymbol{x}_t$; this is the same as the assumptions made in conventional regression analysis. Assumptions considered later involve letting the coefficient vector $\boldsymbol{\beta}_t$ be a random unknown signal vector that can be estimated by Bayesian arguments, using the conditional expectation given the data. The answers to this approach, given in §7.5, allow signal extraction and deconvolution problems to be handled. Assuming the inputs are fixed allows various experimental designs and analysis of variance to be done for both fixed and random effects models. Estimation of the frequency-dependent random effects variance components in the analysis of variance model is also considered in §7.5.

For the approach in this section, assume the inputs and outputs have zero means and are jointly stationary with the $(q+1) \times 1$ vector process $(\boldsymbol{x}_t', y_t)'$ of inputs $\boldsymbol{x}_t$ and outputs y_t assumed to have a spectral matrix of the form

$$f(\omega) = \begin{pmatrix} f_{xx}(\omega) & f_{xy}(\omega) \\ f_{yx}(\omega) & f_{yy}(\omega) \end{pmatrix}, \tag{7.10}$$

where $f_{yx}(\omega) = (f_{yx_1}(\omega), f_{yx_2}(\omega), \ldots, f_{yx_q}(\omega))$ is the $1 \times q$ vector of cross-spectra relating the q inputs to the output and $f_{xx}(\omega)$ is the $q \times q$ spectral matrix of the inputs. Generally, we observe the inputs and search for the vector of regression functions $\boldsymbol{\beta}_t$ relating the inputs to the outputs. We assume all autocovariance functions satisfy the absolute summability conditions of the form

$$\sum_{h=-\infty}^{\infty} |h||\gamma_{jk}(h)| \ < \infty. \tag{7.11}$$

$(j, k = 1, \ldots, q+1)$, where $\gamma_{jk}(h)$ is the autocovariance corresponding to the cross-spectrum $f_{jk}(\omega)$ in (7.10). We also need to assume a linear process of the form (C.35) as a condition for using Theorem C.7 on the joint distribution of the discrete Fourier transforms in the neighborhood of some fixed frequency.

Estimation of the Regression Function

In order to estimate the regression function $\boldsymbol{\beta}_r$, the Projection Theorem (Appendix B) applied to minimizing

$$MSE = E\big[(y_t - \sum_{r=-\infty}^{\infty} \boldsymbol{\beta}_r' \boldsymbol{x}_{t-r})^2\big] \tag{7.12}$$

leads to the orthogonality conditions

$$E\big[(y_t - \sum_{r=-\infty}^{\infty} \boldsymbol{\beta}_r' \boldsymbol{x}_{t-r})\, \boldsymbol{x}_{t-s}'\big] = \mathbf{0}' \tag{7.13}$$

for all $s = 0, \pm 1, \pm 2, \ldots$, where $\mathbf{0}'$ denotes the $1 \times q$ zero vector. Taking the expectations inside and substituting for the definitions of the autocovariance functions appearing and leads to the normal equations

$$\sum_{r=-\infty}^{\infty} \boldsymbol{\beta}_r'\, \Gamma_{xx}(s-r) = \boldsymbol{\gamma}_{yx}'(s), \tag{7.14}$$

for $s = 0, \pm 1, \pm 2, \ldots$, where $\Gamma_{xx}(s)$ denotes the $q \times q$ autocovariance matrix of the vector series $\boldsymbol{x}_t$ at lag s and $\boldsymbol{\gamma}_{yx}(s) = (\gamma_{yx_1}(s), \ldots, \gamma_{yx_q}(s))$ is a $1 \times q$ vector containing the lagged covariances between y_t and $\boldsymbol{x}_t$. Again, a frequency domain approximate solution is easier in this case because the computations can be done frequency by frequency using cross-spectra that can be estimated from sample data using the DFT. In order to develop the frequency domain solution, substitute the representation into the normal equations, using the same approach as used in the simple case derived in §4.8. This approach yields

$$\int_{-1/2}^{1/2} \sum_{r=-\infty}^{\infty} \boldsymbol{\beta}_r'\, \mathrm{e}^{2\pi i \omega (s-r)}\, f_{xx}(\omega)\, d\omega = \boldsymbol{\gamma}_{yx}'(s).$$

Now, because $\boldsymbol{\gamma}_{yx}'(s)$ is the Fourier transform of the cross-spectral vector $f_{yx}(\omega) = f_{xy}^*(\omega)$, we might write the system of equations in the frequency domain, using the uniqueness of the Fourier transform, as

$$\boldsymbol{B}'(\omega) f_{xx}(\omega) = f_{xy}^*(\omega), \tag{7.15}$$

where $f_x(\omega)$ is the $q \times q$ spectral matrix of the inputs and $\boldsymbol{B}(\omega)$ is the $q \times 1$ vector Fourier transform of $\boldsymbol{\beta}_t$. Multiplying (7.15) on the right by $f_{xx}^{-1}(\omega)$, assuming $f_{xx}(\omega)$ is nonsingular at ω, leads to the frequency domain estimator

$$\boldsymbol{B}'(\omega) = f_{xy}^*(\omega) f_{xx}^{-1}(\omega). \tag{7.16}$$

Note, (7.16) implies the regression function would take the form

$$\boldsymbol{\beta}_t = \int_{-1/2}^{1/2} \boldsymbol{B}(\omega)\ \mathrm{e}^{2\pi i\omega t}\ d\omega. \tag{7.17}$$

As before, it is conventional to introduce the DFT as the approximate estimator for the integral (7.17) and write

$$\boldsymbol{\beta}_t^M = M^{-1} \sum_{k=0}^{M-1} \boldsymbol{B}(\omega_k)\ \mathrm{e}^{2\pi i\omega_k t}, \tag{7.18}$$

where $\omega_k = k/M, M << n$. The approximation was shown in Problem 4.35 to hold exactly as long as $\boldsymbol{\beta}_t = \mathbf{0}$ for $|t| \geq M$ and to have a mean squared error bounded by a function of the zero-lag autocovariance and the absolute sum of the neglected coefficients.

The mean squared error (7.12) can be written using the orthogonality principle, giving

$$MSE = \int_{-1/2}^{1/2} f_{y\cdot x}(\omega)\ d\omega, \tag{7.19}$$

where

$$f_{y\cdot x}(\omega) = f_{yy}(\omega) - f_{xy}^*(\omega) f_{xx}^{-1}(\omega) f_{xy}(\omega) \tag{7.20}$$

denotes the residual or error spectrum The resemblance of (7.20) to the usual equations in regression analysis is striking. It is useful to pursue the multiple regression analogy further by noting a squared multiple coherence can be defined as

$$\rho_{y\cdot x}^2(\omega) = \frac{f_{xy}^*(\omega) f_{xx}^{-1}(\omega) f_{xy}(\omega)}{f_{yy}(\omega)}. \tag{7.21}$$

This expression leads to the mean squared error in the form

$$MSE = \int_{-1/2}^{1/2} f_{yy}(\omega)[1 - \rho_{y\cdot x}^2(\omega)]\ d\omega, \tag{7.22}$$

and we have an interpretation of $\rho_{y\cdot x}^2(\omega)$ as the proportion of power accounted for by the lagged regression on $\boldsymbol{x}_t$ at frequency ω. If $\rho_{y\cdot x}^2(\omega) = 0$ for all ω, we have

$$MSE = \int_{-1/2}^{1/2} f_{yy}(\omega)\ d\omega = E[y_t^2],$$

which is the mean squared error when no predictive power exists. As long as $f_{xx}(\omega)$ is positive definite at all frequencies, $MSE \geq 0$, and we will have

$$0 \leq \rho_{y\cdot x}^2(\omega) \leq 1 \tag{7.23}$$

for all ω. If the multiple coherence is unity for all frequencies, the mean squared error in (7.22) is zero and the output series is perfectly predicted by a linearly

filtered combination of the inputs. Problem 7.3 shows the ordinary squared coherence between the series y_t and the linearly filtered combinations of the inputs appearing in (7.12) is exactly (7.21).

Estimation Using Sampled Data

Clearly, the matrices of spectra and cross-spectra will not ordinarily be known, so the regression computations need to be based on sampled data. We assume, therefore, the inputs $x_{t1}, x_{t2}, \ldots, x_{tq}$ and output y_t series are available at the time points $t = 1, 2, \ldots, n$, as in Chapter 4. In order to develop reasonable estimates for the spectral quantities, some replication must be assumed. Often, only one replication of each of the inputs and the output will exist, so it is necessary to assume a band exists over which the spectra and cross-spectra are approximately equal to $f_{xx}(\omega)$ and $f_{xy}(\omega)$, respectively. Then, let $Y(\omega_k + \ell/n)$ and $\boldsymbol{X}(\omega_k + \ell/n)$ be the DFTs of y_t and $\boldsymbol{x}_t$ over the band, say, at frequencies of the form

$$\omega_k \pm \ell/n, \quad \ell = 1, \ldots, m,$$

where $L = 2m + 1$ as before. Then, simply substitute the sample spectral matrix

$$\widehat{f}_{xx}(\omega) = L^{-1} \sum_{\ell=-m}^{m} \boldsymbol{X}(\omega_k + \ell/n)\boldsymbol{X}^*(\omega_k + \ell/n) \tag{7.24}$$

and the vector of sample cross-spectra

$$\widehat{f}_{xy}(\omega) = L^{-1} \sum_{\ell=-m}^{m} \boldsymbol{X}(\omega_k + \ell/n)\overline{Y(\omega_k + \ell/n)} \tag{7.25}$$

for the respective terms in (7.16) to get the regression estimator $\widehat{\boldsymbol{B}}(\omega)$. For the regression estimator (7.18), we may use

$$\widehat{\beta}_t^M = \frac{1}{M} \sum_{k=0}^{M-1} \widehat{f}_{xy}^*(\omega_k)\widehat{f}_{xx}^{-1}(\omega_k)\, \mathrm{e}^{2\pi i \omega_k t} \tag{7.26}$$

for $t = 0, \pm 1, \pm 2, \ldots, \pm(M/2 - 1)$, as the estimated regression function.

Tests of Hypotheses

The estimated squared multiple coherence, corresponding to the theoretical coherence (7.21), becomes

$$\widehat{\rho}_{y\cdot x}^2(\omega) = \frac{\widehat{f}_{xy}^*(\omega)\widehat{f}_{xx}^{-1}(\omega)\widehat{f}_{xy}(\omega)}{\widehat{f}_{yy}(\omega)}. \tag{7.27}$$

We may obtain a distributional result for the multiple coherence function analogous to that obtained in the univariate case by writing the multiple regression model in the frequency domain, as was done in §4.6. We obtain the statistic

$$F_{2q,2(L-q)} = \frac{(L-q)}{q} \frac{\widehat{\rho}^2_{y\cdot x}(\omega)}{[1-\widehat{\rho}^2_{y\cdot x}(\omega)]}, \tag{7.28}$$

which has an F-distribution with $2q$ and $2(L-q)$ degrees of freedom under the null hypothesis that $\rho^2_{y\cdot x}(\omega) = 0$, or equivalently, that $\boldsymbol{B}(\omega) = 0$, in the model

$$Y(\omega_k + \ell/n) = \boldsymbol{B}'(\omega)X(\omega_k + \ell/n) + V(\omega_k + \ell/n), \tag{7.29}$$

where the spectral density of the error $V(\omega_k + \ell/n)$ is $f_{y\cdot x}(\omega)$. Problem 7.5 sketches a derivation of this result.

A second kind of hypothesis of interest is one that might be used to test whether a full model with q inputs is significantly better than some submodel with $q_1 < q$ components. In the time domain, this hypothesis implies, for a partition of the vector of inputs into q_1 and q_2 components ($q_1 + q_2 = q$), say, $\boldsymbol{x}_t = (\boldsymbol{x}'_{t1}, \boldsymbol{x}'_{t2})'$, and the similarly partitioned vector of regression functions $\boldsymbol{\beta}_t = (\boldsymbol{\beta}'_{1t}, \boldsymbol{\beta}'_{2t})'$, we would be interested in testing whether $\boldsymbol{\beta}_{2t} = \mathbf{0}$ in the partitioned regression model

$$y_t = \sum_{r=-\infty}^{\infty} \boldsymbol{\beta}'_{1r}\boldsymbol{x}_{t-r,1} + \sum_{r=-\infty}^{\infty} \boldsymbol{\beta}'_{2r}\boldsymbol{x}_{t-r,2} + v_t. \tag{7.30}$$

Rewriting the regression model (7.30) in the frequency domain in a form that is similar to (7.29) establishes that, under the partitions of the spectral matrix into its $q_i \times q_j$ $(i, j = 1, 2)$ submatrices, say,

$$\widehat{f}_{xx}(\omega) = \begin{pmatrix} \widehat{f}_{11}(\omega) & \widehat{f}_{12}(\omega) \\ \widehat{f}_{21}(\omega) & \widehat{f}_{22}(\omega) \end{pmatrix}, \tag{7.31}$$

and the cross-spectral vector into its $q_i \times 1$ $(i = 1, 2)$ subvectors,

$$\widehat{f}_{xy}(\omega) = \begin{pmatrix} \widehat{f}_{1y}(\omega) \\ \widehat{f}_{2y}(\omega) \end{pmatrix}, \tag{7.32}$$

we may test the hypothesis $\boldsymbol{\beta}_{2t} = \mathbf{0}$ at frequency ω by comparing the estimated residual power

$$\widehat{f}_{y\cdot x}(\omega) = \widehat{f}_{yy}(\omega) - \widehat{f}^*_{xy}(\omega)\widehat{f}^{-1}_{xx}(\omega)\widehat{f}_{xy}(\omega) \tag{7.33}$$

under the full model with that under the reduced model, given by

$$\widehat{f}_{y\cdot 1}(\omega) = \widehat{f}_{yy}(\omega) - \widehat{f}^*_{1y}(\omega)\widehat{f}^{-1}_{11}(\omega)\widehat{f}_{1y}(\omega). \tag{7.34}$$

The power due to regression can be written as

$$\mathrm{SSR}(\omega) = L[\widehat{f}_{y\cdot 1}(\omega) - \widehat{f}_{y\cdot x}(\omega)], \tag{7.35}$$

Table 7.1 Analysis of Power (ANOPOW) for Testing No Contribution from the Subset $\boldsymbol{x}_{t2}$ in the Partitioned Regression Model

Source	Power	Degrees of Freedom
$x_{t,q_1+1}, \ldots, x_{t,q_1+q_2}$	SSR(ω) (7.35)	$2q_2$
Error	SSE(ω) (7.36)	$2(L-q_1-q_2)$
Total	$L\widehat{f}_{y\cdot 1}(\omega)$	$2(L-q_1)$

with the usual error power given by

$$\mathrm{SSE}(\omega) = L\widehat{f}_{y\cdot x}(\omega). \tag{7.36}$$

The test of no regression proceeds using the F-statistic

$$F_{2q_2,2(L-q)} = \frac{(L-q)}{q_2}\frac{\mathrm{SSR}(\omega)}{\mathrm{SSE}(\omega)}. \tag{7.37}$$

The distribution of this F-statistic with $2q_2$ numerator degrees of freedom and $2(L-q)$ denominator degrees of freedom follows from an argument paralleling that given in Chapter 4 for the case of a single input. The test results can be summarized in an Analysis of Power (ANOPOW) table that parallels the usual analysis of variance (ANOVA) table. Table 7.1 shows the components of power for testing $\boldsymbol{\beta}_{2t} = \mathbf{0}$ at a particular frequency ω. The ratio of the two components divided by their respective degrees of freedom just yields the F-statistic (7.37) used for testing whether the q_2 add significantly to the predictive power of the regression on the q_1 series.

Example 7.1 Predicting Shasta Lake Inflow

We illustrate some of the preceding ideas by considering the problem of predicting the transformed inflow series shown in Figure 4.33 from some combination of the inputs. First, look for the best single input predictor using the squared coherence function (7.27). The results, exhibited in Figure 7.3, show transformed precipitation produces the most consistently high squared coherence values at all frequencies ($L = 41$), with the seasonal frequencies .08, .17, .25, and .33 cycles per month corresponding to 12-month, six-month, four-month, and three-month periods contributing most significantly at the $\alpha = .001$ level.

Other inputs, with the exception of wind speed, also appear to be plausible contributors. In order to evaluate the other contributors, we consider partitioned tests with models including each of the other variables and precipitation tested against models including precipitation alone. Figure 7.4 shows a plot of the F-statistic (7.28) as a function of frequency for testing each of the inputs as a possible additional component. We see here some isolated significance points, particularly in the temperature series at some of the higher seasonal components, although the strong

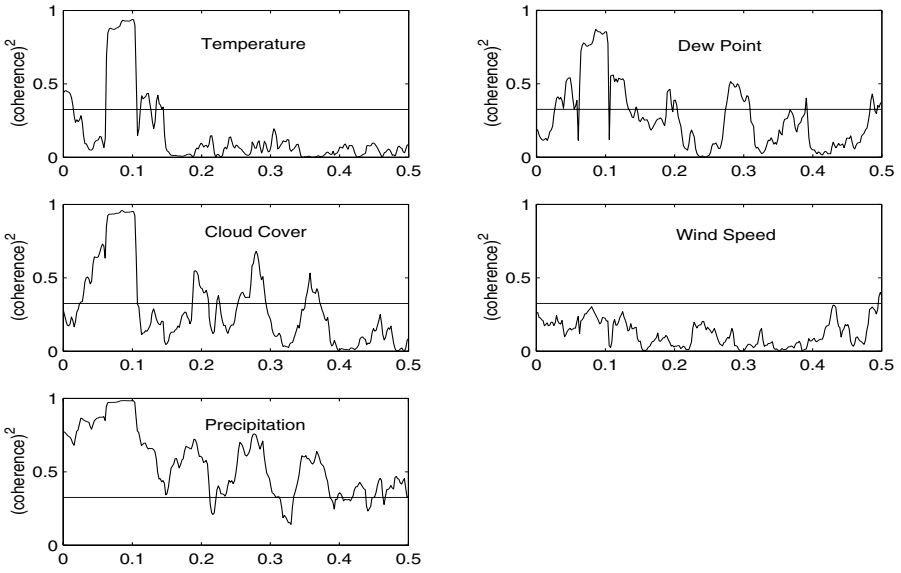


Figure 7.3 Univariate coherence functions relating Shasta Lake inflow to various inputs (frequency scale is cycles per month).

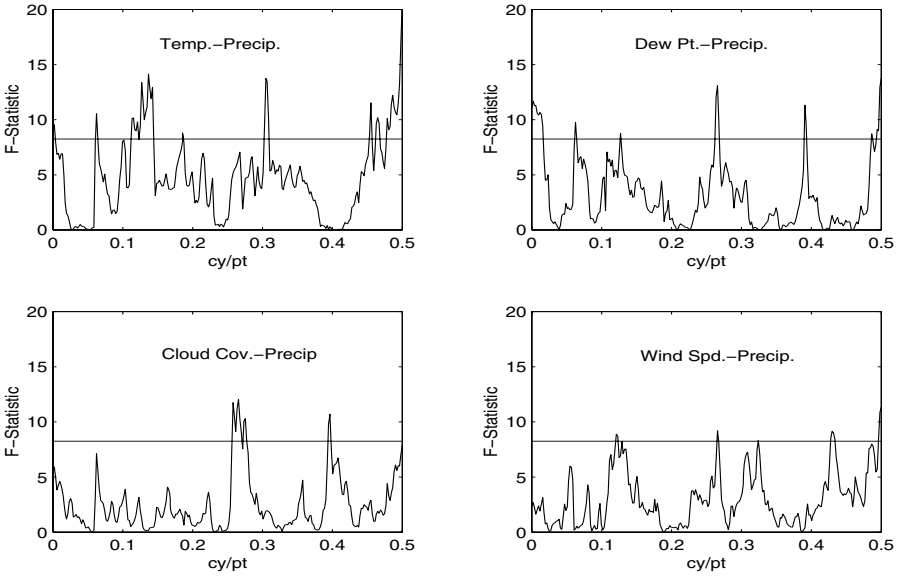


Figure 7.4 *F*-statistics for testing whether various inputs combined with precipitation add to the ability to predict Shasta Lake inflow.

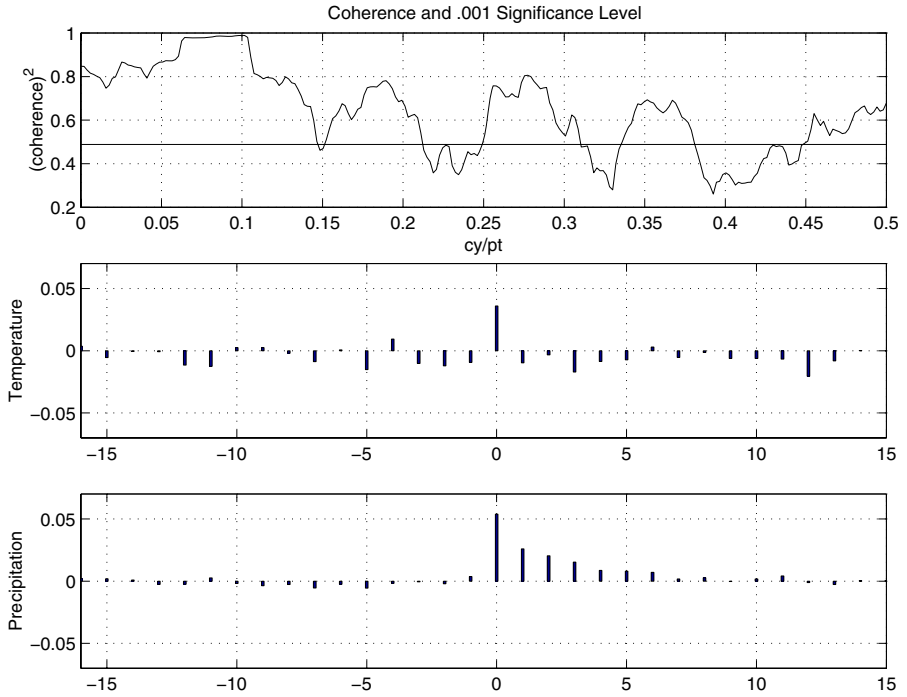


Figure 7.5 Multiple coherence between inflow and combined precipitation and temperature along with multiple impulse response functions for the regression relations.

coherence at the 12-month frequency seems to have been essentially eliminated by the incorporation of precipitation.

The additional contribution of temperature to the model seems somewhat marginal because the multiple coherence (7.27), shown in the top panel of Figure 7.5, seems only slightly better than the univariate coherence with precipitation shown in Figure 7.3. It is, however, instructive to produce the multiple regression functions, using (7.26) to see if a simple model for inflow exists that would involve some regression combination of inputs temperature and precipitation that would be useful for predicting inflow to Shasta Lake. With this in mind, denoting the possible inputs P_t for transformed precipitation and T_t for transformed temperature, the regression function has been plotted in the lower two panels of Figure 7.5. The time axes run over both positive and negative values and are centered at time $t = 0$. Hence, the relation with temperature seems to be instantaneous and positive and an exponentially decaying relation to precipitation exists that has been noticed previously in the analysis in Problem 4.37 of Chapter 4. The plots suggest a transfer function model

of the general form fitted to the Recruitment and SOI series in Example 5.7 of Chapter 5. We might propose fitting the inflow output, say, I_t, using the model

$$I_t = \alpha_0 + \frac{\delta_0}{(1-\omega_1 B)} P_t + \alpha_2 T_t + \eta_t,$$

which is the transfer function model, without the temperature component, considered in that section.

7.4 Regression with Deterministic Inputs

The previous section considered the case in which the input and output series were jointly stationary, but there are many circumstances where in which we might want to assume that the input functions are fixed and have a known functional form. This case happens in the analysis of data from designed experiments. For example, we may want to take a collection of earthquakes and explosions such as are shown in Figure 7.2 and test whether the mean functions are the same for either the P or S components or, perhaps, for them jointly. In certain other signal detection problems using arrays, the inputs are used as dummy variables to express lags corresponding to the arrival times of the signal at various elements, under a model corresponding to that of a plane wave from a fixed source propagating across the array. In Figure 7.1, we plotted the mean responses of the cortex as a function of various underlying design configurations corresponding to various stimuli applied to awake and mildly anesthetized subjects.

It is necessary to introduce a replicated version of the underlying model to handle even the univariate situation, and we replace (7.8) by

$$y_{jt} = \sum_{r=-\infty}^{\infty} \boldsymbol{\beta}_r' \boldsymbol{z}_{j,t-r} + v_{jt} \tag{7.38}$$

for $j = 1, 2, \ldots, N$ series, where we assume the vector of known deterministic inputs, $\boldsymbol{z}_{jt} = (z_{jt1}, \ldots, \boldsymbol{z}_{jtq})'$, satisfies

$$\sum_{t=-\infty}^{\infty} |t||z_{jtk}| < \infty$$

for $j = 1, \ldots, N$ replicates of an underlying process involving $k = 1, \ldots, q$ regression functions. The model can also be treated under the assumption that the deterministic function satisfy Grenanders' conditions, as in Hannan (1970), but we do not need those conditions here and simply follow the approach in Shumway (1983, 1988).

It will sometimes be convenient in what follows to represent the model in matrix notation, writing (7.38) as

$$\boldsymbol{y}_t = \sum_{r=-\infty}^{\infty} z_{t-r}\, \boldsymbol{\beta}_r + \boldsymbol{v}_t, \tag{7.39}$$

where $z_t = (\boldsymbol{z}_{1t}, \ldots, \boldsymbol{z}_{Nt})'$ are the $N \times q$ matrices of independent inputs and $\boldsymbol{y}_t$ and $\boldsymbol{v}_t$ are the $N \times 1$ output and error vectors. The error vector $\boldsymbol{v}_t = (v_{1t}, \ldots, v_{Nt})'$ is assumed to be a multivariate, zero-mean, stationary, normal process with spectral matrix $f_v(\omega)I_N$ that is proportional to the $N \times N$ identity matrix. That is, we assume the error series v_{jt} are independently and identically distributed with spectral densities $f_v(\omega)$.

Example 7.2 An Infrasonic Signal from a Nuclear Explosion

Often, we will observe a common signal, say, β_t on an array of sensors, with the response at the j-th sensor denoted by $y_{jt}, j = 1, \ldots, N$ For example, Figure 7.6 shows an infrasonic or low-frequency acoustic signal from a nuclear explosion, as observed on a small triangular array of $N = 3$ acoustic sensors. These signals appear at slightly different times. Because of the way signals propagate, a plane wave signal of this kind, from a given source, traveling at a given velocity, will arrive at elements in the array at predictable time delays. In the case of the infrasonic signal in Figure 7.6, the delays were approximated by computing the cross-correlation between elements and simply reading off the time delay corresponding to the maximum. For a detailed discussion of the statistical analysis of array signals, see Shumway et al. (1999).

A simple additive signal plus noise model of the form

$$y_{jt} = \beta_{t-\tau_j} + v_{jt} \tag{7.40}$$

can be assumed, where $\tau_j, j = 1, 2, \ldots, N$ are the time delays that determine the start point of the signal at each element of the array. The model (7.40) is written in the form (7.38) by letting $z_{jt} = \delta_{t-\tau_j}$, where $\delta_t = 1$ when $t = 0$ and is zero otherwise. In this case, we are interested in both the problem of detecting the presence of the signal and in estimating its waveform β_t. In this case, a plausible estimator of the waveform would be the unbiased beam, say,

$$\widehat{\beta}_t = \frac{\sum_{j=1}^{N} y_{j,t+\tau_j}}{N}, \tag{7.41}$$

where time delays in this case were measured as $\tau_1 = -17, \tau_2 = 0$, and $\tau_3 = 22$ from the cross-correlation function. The bottom panel of Figure 7.6 shows the computed beam in this case, and the noise in the individual channels has been reduced and the essential characteristics of the common signal are retained in the average.

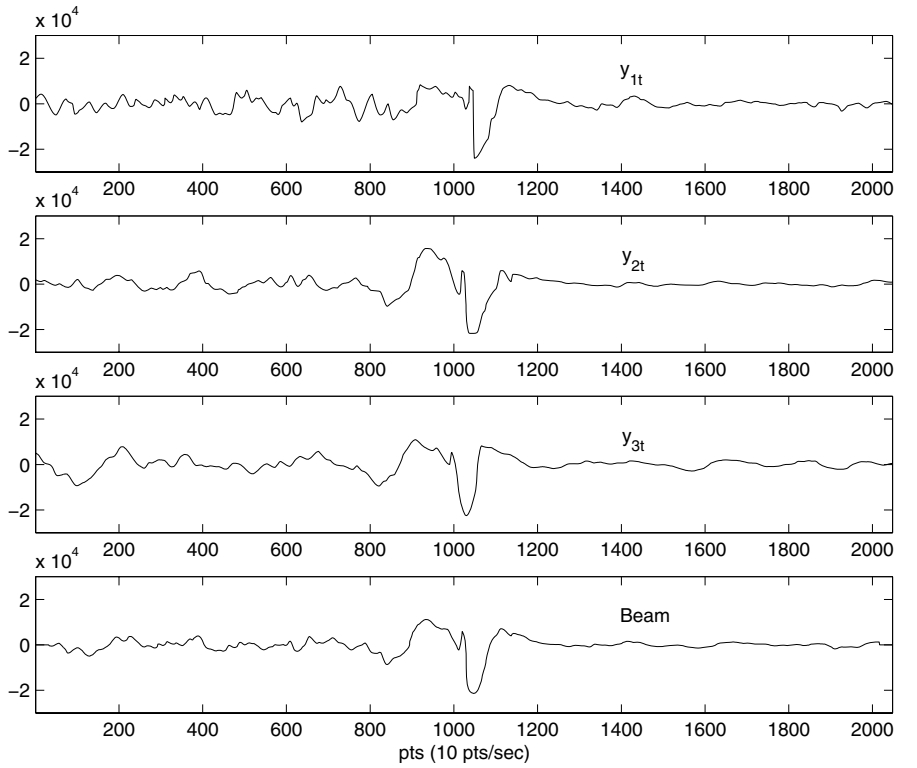


Figure 7.6 Three series for a nuclear explosion detonated 25 km south of Christmas Island and the delayed average or beam.

The above discussion and example serve to motivate a more detailed look at the estimation and detection problems in the case in which the input series $\boldsymbol{z}_{jt}$ are fixed and known. We consider the modifications needed for this case in the following sections.

Estimation of the Regression Relation

Because the regression model (7.38) involves fixed functions, we may parallel the usual approach using the Gauss–Markov theorem to search for linear-filtered estimators of the form

$$\widehat{\boldsymbol{\beta}}_t = \sum_{j=1}^{N} \sum_{r=-\infty}^{\infty} \boldsymbol{h}_{jr} y_{j,t-r}, \tag{7.42}$$

where $\boldsymbol{h}_{jt} = (h_{jt1} \ldots, h_{jtq})'$ is a vector of filter coefficients, determined so the estimators are unbiased and have minimum variance. The equivalent matrix

form is

$$\widehat{\boldsymbol{\beta}}_t = \sum_{r=-\infty}^{\infty} h_r \, \boldsymbol{y}_{t-r}, \tag{7.43}$$

where $h_t = (\boldsymbol{h}_{1t}, \ldots, \boldsymbol{h}_{Nt})$ is a $q \times N$ matrix of filter functions. The matrix form resembles the usual classical regression case and is more convenient for extending the the Gauss–Markov theorem to lagged regression. The unbiased condition is considered in Problem 7.7. It can be shown (see Shumway and Dean, 1968) that $\boldsymbol{h}_{js}$ can be taken as the Fourier transform of

$$\boldsymbol{H}_j(\omega) = S_z^{-1}(\omega)\overline{\boldsymbol{Z}_j(\omega)}, \tag{7.44}$$

where

$$\boldsymbol{Z}_j(\omega) = \sum_{t=-\infty}^{\infty} \boldsymbol{z}_{jt} \mathrm{e}^{-2\pi i\omega t} \tag{7.45}$$

is the infinite Fourier transform of $\boldsymbol{z}_{jt}$. The matrix

$$S_z(\omega) = \sum_{j=1}^{N} \overline{\boldsymbol{Z}_j(\omega)}\boldsymbol{Z}_j'(\omega) \tag{7.46}$$

can be written in the form

$$S_z(\omega) = Z^*(\omega)Z(\omega), \tag{7.47}$$

where the $N \times q$ matrix $Z(\omega)$ is defined by $Z(\omega) = (\boldsymbol{Z}_1(\omega), \ldots, \boldsymbol{Z}_N(\omega))'$. In matrix notation, the Fourier transform of the optimal filter becomes

$$H(\omega) = S_z^{-1}(\omega)Z^*(\omega), \tag{7.48}$$

where $H(\omega) = (\boldsymbol{H}_1(\omega), \ldots, \boldsymbol{H}_N(\omega))$ is the $q \times N$ matrix of frequency response functions. The optimal filter then becomes the Fourier transform

$$h_t = \int_{-1/2}^{1/2} H(\omega)\mathrm{e}^{2\pi i\omega t} \, d\omega. \tag{7.49}$$

If the transform is not tractable to compute, an approximation analogous to (7.26) may be used.

Example 7.3 Estimation of the Infrasonic Signal in Example 7.2

We consider the problem of producing a best linearly filtered unbiased estimator for the infrasonic signal in Example 7.2. In this case, $q = 1$ and (7.45) becomes

$$Z_j(\omega) = \sum_{t=-\infty}^{\infty} \delta_{t-\tau_j} \mathrm{e}^{-2\pi i\omega t} = \mathrm{e}^{-2\pi i\omega\tau_j}$$

and $S_z(\omega) = N$. Hence, we have

$$H_j(\omega) = \frac{1}{N}\, \mathrm{e}^{2\pi i \omega \tau_j}.$$

Using (7.49), we obtain $h_{jt} = \frac{1}{N}\delta(t + \tau_j)$. Substituting in (7.42), we obtain the best linear unbiased estimator as the beam, computed as in (7.41).

Tests of Hypotheses

We consider first testing the hypothesis that the complete vector $\boldsymbol{\beta}_t$ is zero, i.e., that the vector signal is absent. We develop a test at each frequency ω by taking single adjacent frequencies of the form $\omega_k = k/n$, as in the initial section. We may approximate the DFT of the observed vector in the model (7.38) using a representation of the form

$$Y_j(\omega_k) = \boldsymbol{B}'(\omega_k)\boldsymbol{Z}_j(\omega_k) + V_j(\omega_k) \tag{7.50}$$

for $j = 1, \ldots, N$, where the error terms will be uncorrelated with common variance $f(\omega_k)$, the spectral density of the error term. The independent variables $\boldsymbol{Z}_j(\omega_k)$ can either be the infinite Fourier transform, or they can be approximated by the DFT. Hence, we can obtain the matrix version of a complex regression model, written in the form

$$\boldsymbol{Y}(\omega_k) = Z(\omega_k)\boldsymbol{B}(\omega_k) + \boldsymbol{V}(\omega_k), \tag{7.51}$$

where the $N \times q$ matrix $Z(\omega_k)$ has been defined previously below (7.47) and $\boldsymbol{Y}(\omega_k)$ and $\boldsymbol{V}(\omega_k)$ are $N \times 1$ vectors with the error vector $\boldsymbol{V}(\omega_k)$ having mean zero, with covariance matrix $f(\omega_k)I_N$. The usual regression arguments show that the maximum likelihood estimator for the regression coefficient will be

$$\widehat{\boldsymbol{B}}(\omega_k) = S_z^{-1}(\omega_k)\boldsymbol{s}_{zy}(\omega_k), \tag{7.52}$$

where $S_z(\omega_k)$ is given by (7.47) and

$$\begin{aligned} \boldsymbol{s}_{zy}(\omega_k) &= Z^*(\omega_k)\boldsymbol{Y}(\omega_k) \\ &= \sum_{j=1}^{N} \overline{\boldsymbol{Z}_j(\omega_k)} Y_j(\omega_k). \end{aligned} \tag{7.53}$$

Also, the maximum likelihood estimator for the error spectral matrix is proportional to

$$\begin{aligned} s^2_{y\cdot z}(\omega_k) &= \sum_{j=1}^{N} |Y_j(\omega_k) - \widehat{\boldsymbol{B}}(\omega k)'\boldsymbol{Z}_j(\omega_k)|^2 \\ &= \boldsymbol{Y}^*(\omega_k)\boldsymbol{Y}(\omega_k) - \boldsymbol{Y}^*(\omega_k)Z(\omega_k)[Z^*(\omega_k)Z(\omega_k)]^{-1}Z^*(\omega_k)\boldsymbol{Y}(\omega_k) \\ &= s^2_y(\omega_k) - \boldsymbol{s}^*_{zy}(\omega_k)S_z^{-1}(\omega_k)\boldsymbol{s}_{zy}(\omega_k), \end{aligned} \tag{7.54}$$

Table 7.2 Analysis of Power (ANOPOW) for Testing No Contribution from the Independent Series at Frequency ω in the Fixed Input Case

Source	Power	Degrees of Freedom
Regression	SSR(ω)(7.56)	$2Lq$
Error	SSE(ω) (7.57)	$2L(N-q)$
Total	SST(ω)	$2LN$

where

$$s_y^2(\omega_k) = \sum_{j=1}^{N} |Y_j(\omega_k)|^2. \tag{7.55}$$

Under the null hypothesis that the regression coefficient $\boldsymbol{B}(\omega_k) = \mathbf{0}$, the estimator for the error power is just $s_y^2(\omega_k)$ If smoothing is needed, we may replace the (7.54) and (7.55) by smoothed components over the frequencies $\omega_k + \ell/n$, for $\ell = -m, \ldots, m$ and $L = 2m + 1$, close to ω. In that case, we obtain the regression and error spectral components as

$$\mathrm{SSR}(\omega) = \sum_{\ell=-m}^{m} \boldsymbol{s}_{zy}^*(\omega_k + \ell/n) S_z^{-1}(\omega_k + \ell/n) \boldsymbol{s}_{zy}(\omega_k + \ell/n) \tag{7.56}$$

and

$$\mathrm{SSE}(\omega) = \sum_{\ell=-m}^{m} s_{y\cdot z}^2(\omega_k + \ell/n). \tag{7.57}$$

The F-statistic for testing no regression relation is

$$F_{2Lq,2L(N-q)} = \frac{N-q}{q} \frac{\mathrm{SSR}(\omega)}{\mathrm{SSE}(\omega)}. \tag{7.58}$$

The analysis of power pertaining to this situation appears in Table 7.2.

In the fixed regression case, the partitioned hypothesis that is the analog of $\boldsymbol{\beta}_{2t} = 0$ in (7.28) with $\boldsymbol{x}_{t1}, \boldsymbol{x}_{t2}$ replaced by $\boldsymbol{z}_{t1}, \boldsymbol{z}_{t2}$. Here, we partition $S_z(\omega)$ into $q_i \times q_j$ $(i, j = 1, 2)$ submatrices, say,

$$S_z(\omega_k) = \begin{pmatrix} S_{11}(\omega_k) & S_{12}(\omega_k) \\ S_{21}(\omega_k) & S_{22}(\omega_k) \end{pmatrix}, \tag{7.59}$$

and the cross-spectral vector into its $q_i \times 1$, for $i = 1, 2$, subvectors

$$\boldsymbol{s}_{zy}(\omega_k) = \begin{pmatrix} \boldsymbol{s}_{1y}(\omega_k) \\ \boldsymbol{s}_{2y}(\omega_k) \end{pmatrix}. \tag{7.60}$$

Here, we test the hypothesis $\boldsymbol{\beta}_{2t} = \mathbf{0}$ at frequency ω by comparing the residual power (7.54) under the full model with the residual power under the reduced model, given by

$$s_{y\cdot 1}^2(\omega_k) = s_y^2(\omega_k) - \boldsymbol{s}_{1y}^*(\omega_k) S_{11}^{-1}(\omega_k) \boldsymbol{s}_{1y}(\omega_k). \tag{7.61}$$

Table 7.3 Analysis of Power (ANOPOW) for Testing No Contribution from the Last q_2 Inputs in the Fixed Input Case

Source	Power	Degrees of Freedom
Regression	SSR(ω)(7.62)	$2Lq_2$
Error	SSE(ω) (7.63)	$2L(N-q)$
Total	SST(ω)	$2L(N-q_1)$

Again, it is desirable to add over adjacent frequencies with roughly comparable spectra so the regression and error power components can be taken as

$$\mathrm{SSR}(\omega) = \sum_{\ell=-m}^{m} \left[s^2_{y\cdot 1}(\omega_k + \ell/n) - s^2_{y\cdot z}(\omega_k + \ell/n) \right] \tag{7.62}$$

and

$$\mathrm{SSE}(\omega) = \sum_{\ell=-m}^{m} s^2_{y\cdot z}(\omega_k + \ell/n). \tag{7.63}$$

The information can again be summarized as in Table 7.3, where the ratio of mean power regression and error components leads to the F-statistic

$$F_{2Lq_2,2L(N-q)} = \frac{(N-q)}{q_2} \frac{\mathrm{SSR}(\omega)}{\mathrm{SSE}(\omega)}. \tag{7.64}$$

We illustrate the analysis of power procedure using the infrasonic signal detection procedure of Example 7.2.

Example 7.4 Detecting the Infrasonic Signal Using ANOPOW

We consider the problem of detecting the common signal for the three infrasonic series observing the common signal, as shown in Figure 7.3. The presence of the signal is obvious in the waveforms shown, so the test here mainly confirms the statistical significance and isolates the frequencies containing the strongest signal components. Each series contained $n = 1024$ points, sampled at 10 points per second. We use the model in (7.40) so $Z_j(\omega) = \mathrm{e}^{-2\pi i \omega \tau_j}$ and $S_z(\omega) = N$ as in Example 7.3, with $s_{zy}(\omega_k)$ given as

$$s_{zy}(\omega_k) = \sum_{j=1}^{N} \mathrm{e}^{2\pi i \omega \tau_j} Y_j(\omega_k),$$

using (7.46) and (7.53). The above expression can be interpreted as being proportional to the weighted mean or beam, computed in frequency, and we introduce the notation

$$B_w(\omega_k) = \frac{1}{N} \sum_{j=1}^{N} \mathrm{e}^{2\pi i \omega \tau_j} Y_j(\omega_k) \tag{7.65}$$

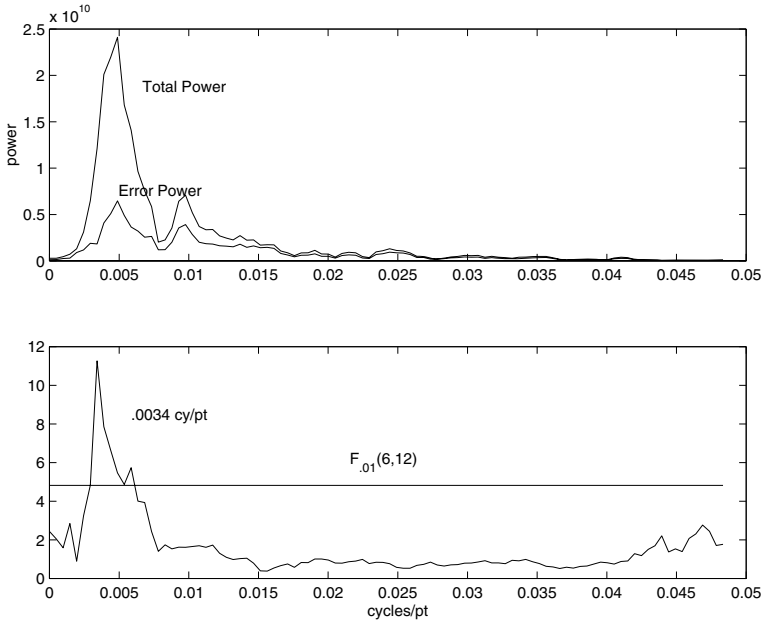


Figure 7.7 Analysis of power for infrasound array (top panel) and F-statistic (bottom panel) showing detection at .033 cy/sec (10 pts/sec).

for that term. Substituting for the power components in Table 7.3 yields

$$\boldsymbol{s}_{zy}^*(\omega_k) S_z^{-1}(\omega_k) \boldsymbol{s}_{zy}(\omega_k) = N|B_w(\omega_k)|^2$$

and

$$\begin{aligned} s_{y\cdot z}^2(\omega_k) &= \sum_{j=1}^{N} |Y_j(\omega_k) - B_w(\omega_k)|^2 \\ &= \sum_{j=1}^{N} |Y_j(\omega_k)|^2 - N|B_w(\omega_k)|^2 \end{aligned}$$

for the regression signal and error components, respectively. Because only three elements in the array and a reasonable number of points in time exist, it seems advisable to employ some smoothing over frequency to obtain additional degrees of freedom. In this case, $L = 3$, yielding $2(3) = 6$ and $2(3)(3-1) = 12$ degrees of freedom for the numerator and denominator of the F-statistic (7.58). Figure 7.7 shows the analysis of power components due to error and the total power. The power is maximum at about .0044 cycles per point or about .044 cycles per second. The F-statistic is compared with the 1% significance level $F_{.01}(6, 12) =$

4.82 in the bottom panel and has the strongest detection at about .034 cycles per second, a result mainly because the error power is decreasing more quickly than the regression or signal power in that band. Little power of consequence appears to exist in the higher range (.3-.5 cycles per second).

Although there are examples of detecting multiple regression functions of the general type considered above (see, for example, Shumway, 1983), we do not consider additional examples of partitioning in the fixed input case here. The reason is that several examples exist in the section on designed experiments that illustrate the partitioned approach.

7.5 Random Coefficient Regression

The lagged regression models considered so far have assumed the input process is either stochastic or fixed and the components of the vector of regression function $\boldsymbol{\beta}_t$ are fixed and unknown parameters to be estimated. There are many cases in time series analysis in which it is more natural to regard the regression vector as an unknown stochastic signal. For example, we have studied the state-space model in Chapter 6, where the state equation can be considered as involving a random parameter vector that is essentially a multivariate autoregressive process. In §4.10, we considered estimating the univariate regression function β_t as a signal extraction problem.

In this section, we consider a random coefficient regression model of (7.39) in the equivalent form

$$\boldsymbol{y}_t = \sum_{r=-\infty}^{\infty} z_{t-r}\, \boldsymbol{\beta}_r + \boldsymbol{v}_t, \tag{7.66}$$

where $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{Nt})'$ is the $N \times 1$ response vector and $z_t = (\boldsymbol{z}_{1t}, \ldots, \boldsymbol{z}_{Nt})'$ are the $N \times q$ matrices containing the fixed input processes. Here, the components of the $q \times 1$ regression vector $\boldsymbol{\beta}_t$ are zero-mean, uncorrelated, stationary series with common spectral matrix $f_\beta(\omega) I_q$ and the error series $\boldsymbol{v}_t$ have zero-means and spectral matrix $f_v(\omega) I_N$, where I_N is the $N \times N$ identity matrix. Then, defining the $N \times q$ matrix $Z(\omega) = (\boldsymbol{Z}_1(\omega), \boldsymbol{Z}_2(\omega), \ldots, \boldsymbol{Z}_N(\omega))'$ of Fourier transforms of z_t, as in (7.45), it is easy to show the spectral matrix of the response vector $\boldsymbol{y}_t$ is given by

$$f_y(\omega) = f_\beta(\omega) Z(\omega) Z^*(\omega) + f_v(\omega) I_N. \tag{7.67}$$

The regression model with a stochastic stationary signal component is a general version of the simple additive noise model

$$y_t = \beta_t + v_t,$$

considered by Wiener (1949) and Kolmogorov (1941), who derived the minimum mean squared error estimators for β_t, as in §4.10. The more general multivariate version (7.66) represents the series as a convolution of the signal vector $\boldsymbol{\beta}_t$ and a known set of vector input series contained in the matrix $\boldsymbol{z}_t$. Restricting the the covariance matrices of signal and noise to diagonal form is consistent with what is done in statistics using random effects models, which we consider here in a later section. The problem of estimating the regression function $\boldsymbol{\beta}_t$ is often called *deconvolution* in the engineering and geophysical literature.

Estimation of the Regression Relation

The regression function $\boldsymbol{\beta}_t$ can be estimated by a general filter of the form (7.43), where we write that estimator in matrix form

$$\widehat{\boldsymbol{\beta}}_t = \sum_{r=-\infty}^{\infty} h_t \, \boldsymbol{y}_{t-r}, \tag{7.68}$$

where $h_t = (\boldsymbol{h}_{1t}, \ldots, \boldsymbol{h}_{Nt})$, and apply the orthogonality principle, as in §3.9. A generalization of the argument in that section (see Problem 7.8) leads to the estimator

$$H(\omega) = [S_z(\omega) + \theta(\omega) I_q]^{-1} Z^*(\omega) \tag{7.69}$$

for the Fourier transform of the minimum mean-squared error filter, where the parameter

$$\theta(\omega) = \frac{f_v(\omega)}{f_\beta(\omega)} \tag{7.70}$$

is the inverse of the signal-to-noise ratio. It is clear from the frequency domain version of the linear model (7.51), the comparable version of the estimator (7.52) can be written as

$$\widehat{\boldsymbol{B}}(\omega) = [S_z(\omega) + \theta(\omega) I_q]^{-1} \boldsymbol{s}_{zy}(\omega). \tag{7.71}$$

This version exhibits the estimator in the stochastic regressor case as the usual estimator, with a ridge correction, $\theta(\omega)$, that is proportional to the inverse of the signal-to-noise ratio.

The mean-squared covariance of the estimator is shown to be

$$E[(\widehat{\boldsymbol{B}} - \boldsymbol{B})(\widehat{\boldsymbol{B}} - \boldsymbol{B})^*] = f_v(\omega)[S_z(\omega) + \theta(\omega) I_q]^{-1}, \tag{7.72}$$

which again exhibits the close connection between this case and the variance of the estimator (7.52), which can be shown to be $f_v(\omega) S_z^{-1}(\omega)$.

Example 7.5 Estimating the Random Infrasonic Signal

In Example 7.4, we have already determined the components needed in (7.69) and (7.70) to obtain the estimators for the random signal. The Fourier transform of the optimum filter at series j has the form

$$H_j(\omega) = \frac{e^{2\pi i\omega\tau_j}}{N + \theta(\omega)} \tag{7.73}$$

with the mean-squared error given by $f_v(\omega)/[N+\theta(\omega)]$ from (7.72). The net effect of applying the filters will be the same as filtering the beam with the frequency response function

$$\begin{aligned} H_0(\omega) &= \frac{N}{N + \theta(\omega)} \\ &= \frac{N f_\beta(\omega)}{f_v(\omega) + N f_\beta(\omega)}, \end{aligned} \tag{7.74}$$

where the last form is more convenient in cases in which portions of the signal spectrum are essentially zero.

The optimal filters h_t have frequency response functions that depend on the signal spectrum $f_\beta(\omega)$ and noise spectrum $f_v(\omega)$, so we will need estimators for these parameters to apply the optimal filters. Sometimes, there will be values, suggested from experience, for the signal-to-noise ratio $1/\theta(\omega)$ as a function of frequency. The analogy between the model here and the usual variance components model in statistics, however, suggests we try an approach along those lines as in the next section.

Detection and Parameter Estimation

The analogy to the usual variance components situation suggests looking at the regression and error components of Table 7.2 under the stochastic signal assumptions. We consider the components of (7.56) and (7.57) at a single frequency ω_k. In order to estimate the spectral components $f_\beta(\omega)$ and $f_v(\omega)$, we reconsider the linear model (7.51) under the assumption that $\boldsymbol{B}(\omega_k)$ is a random process with spectral matrix $f_\beta(\omega_k)I_q$. Then, the spectral matrix of the observed process is (7.67), evaluated at frequency ω_k.

Consider first the component of the regression power, defined as

$$\begin{aligned} SSR(\omega_k) &= \boldsymbol{s}^*_{zy}(\omega_k) S_z^{-1}(\omega_k) \boldsymbol{s}_{zy}(\omega_k) \\ &= \boldsymbol{Y}^*(\omega_k) Z(\omega_k) S_z^{-1}(\omega_k) Z^*(\omega_k) \boldsymbol{Y}(\omega_k). \end{aligned}$$

A computation shows

$$E[SSR(\omega_k)] = f_\beta(\omega_k)\ \mathrm{tr}\{S_z(\omega_k)\} + q f_v(\omega_k),$$

where tr denotes the trace of a matix. If we can find a set of frequencies of the form $\omega_k + \ell/n$, where the spectra and the Fourier transforms $S_z(\omega_k + \ell/n) \approx S_z(\omega)$ are relatively constant, the expectation of the averaged values in (7.56) yields

$$E[SSR(\omega)] = Lf_\beta(\omega)\text{tr } [S_z(\omega)] + Lqf_v(\omega). \tag{7.75}$$

A similar computation establishes

$$E[SSE(\omega)] = L(N-q)f_v(\omega). \tag{7.76}$$

We may obtain an approximately unbiased estimator for the spectra $f_v(\omega)$ and $f_\beta(\omega)$ by replacing the expected power components by their values and solving (7.75) and (7.76).

Example 7.6 Estimating the Power Components and the Random Infrasonic Signal

In order to provide an optimum estimator for the infrasonic signal, we need to have estimators for the signal and noise spectra $f_\beta(\omega)$ and $f_v(\omega)$ for the case considered in Example 7.5. The form of the filter is $H_0(\omega)$, given in (7.74), and with $q = 1$ and the matrix $S_z(\omega) = N$ at all frequencies in this example simplifies the computations considerably. We may estimate

$$\widehat{f}_v(\omega) = \frac{SSE(\omega)}{L(N-1)} \tag{7.77}$$

and

$$\widehat{f}_\beta(\omega) = (LN)^{-1}\left(SSR(\omega) - \frac{SSE(\omega)}{(N-1)}\right), \tag{7.78}$$

using (7.75) and (7.76) for this special case. Cases will exist in which (7.78) is negative and the estimated signal spectrum can be set to zero for those frequencies. The estimators can be substituted into the optimal filters to apply to the beam, say, $H_0(\omega)$ in (7.74), or to use in the filter applied to each level (7.73).

The analysis of variance estimators can be computed using the analysis of power given in Figure 7.7, and the results of that computation and applying (7.77) and (7.78) are shown in the top panel of Figure 7.8 for a bandwidth of $B = 7/2048 =$ cycles per point or about .03 cycles per second (Hz). Neither spectrum contains any significant power for frequencies greater than .04 cycles per point or about .4 Hz. As expected, the signal spectral estimator is substantial over a narrow band, and this leads to an estimated filter, with estimated frequency response function $\widehat{H}_0(\omega)$, shown on the left-hand side of the second panel. The estimated optimal filter essentially deletes frequencies above .014 Hz and, subject to slight modification, differs little from a standard low-pass filter with

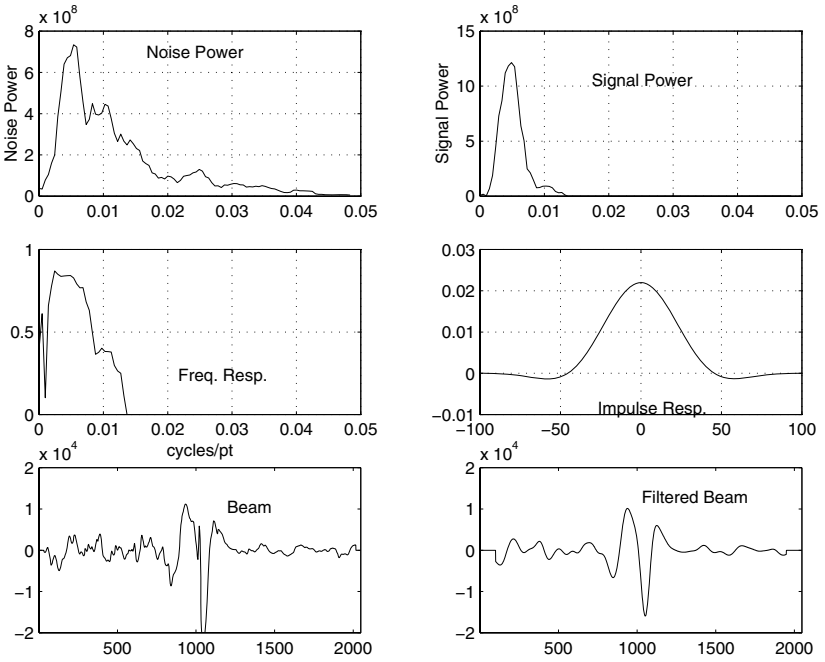


Figure 7.8 Estimated signal and noise spectra, filter responses, and beams.

that cutoff. Computing the time version with a cutoff at $M = 201$ points and using a taper leads to the estimated impulse response function $\widehat{h}_0(t)$, as shown on the right-hand side of the middle panel. Finally, we apply the optimal filter to the beam and get the filtered beam $\widehat{\beta}_t$ shown in the bottom right-hand panel. It is smoother than the left-hand bottom panel, where we have reproduced the beam shown earlier in Figure 7.3. The analysis shows the primary signal as basically a low-frequency signal with primary power at about .05 Hz or, essentially, a wave with a 20-second period.

7.6 Analysis of Designed Experiments

An important special case (see Brillinger, 1973, 1980) of the regression model (7.50) occurs when the regression (7.39) is of the form

$$\boldsymbol{y}_t = z\boldsymbol{\beta}_t + \boldsymbol{v}_t, \tag{7.79}$$

where $z = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N)'$ is a matrix that determines what is observed by the j-th series; i.e.,

$$y_{jt} = \boldsymbol{z}_j'\boldsymbol{\beta}_t + v_{jt}. \tag{7.80}$$

In this case, the the matrix $\boldsymbol{z}$ of independent variables is constant and we will have the frequency domain model.

$$\boldsymbol{Y}(\omega_k) = Z\boldsymbol{B}(\omega_k) + \boldsymbol{V}(\omega_k) \tag{7.81}$$

corresponding to (7.51), where the matrix $Z(\omega_k)$ was a function of frequency ω_k. The matrix is purely real, in this case, but the equations (7.52)-(7.58) can be applied with $Z(\omega_k)$ replaced by the constant matrix Z.

Equality of Means

A typical general problem that we encounter in analyzing real data is a simple equality of means test in which there might be a collection of time series y_{ijt}, $i = 1, \ldots, I$, $j = 1, \ldots, N_i$, belonging to I possible groups, with N_i series in group i. To test equality of means, we may write the regression model in the form

$$y_{ijt} = \mu_t + \alpha_{it} + v_{ijt}, \tag{7.82}$$

where μ_t denotes the overall mean and α_{it} denotes the effect of the i-th group at time t and we require that $\sum_i \alpha_{it} = 0$ for all t. In this case, the full model can be written in the general regression notation as

$$y_{ijt} = \boldsymbol{z}'_{ij}\boldsymbol{\beta}_t + v_{ijt}$$

where

$$\boldsymbol{\beta}_t = (\mu_t, \alpha_{1t}, \alpha_{2t}, \ldots, \alpha_{I-1,t})'$$

denotes the regression vector, subject to the constraint. The reduced model becomes

$$y_{ijt} = \mu_t + v_{ijt} \tag{7.83}$$

under the assumption that the group means are equal. In the full model, there are I possible values for the $I \times 1$ design vectors $\boldsymbol{z}_{ij}$; the first component is always one for the mean, and the rest have a one in the i-th position for $i = 1, \ldots, I-1$ and zeros elsewhere. The vectors for the last group take the value -1 for $i = 2, 3, \ldots, I-1$. Under the reduced model, each $\boldsymbol{z}_{ij}$ is a single column of ones. The rest of the analysis follows the approach summarized in (7.52)-(7.58). In this particular case, the power components in Table 7.3 (before smoothing) simplify to

$$SSR(\omega_k) = \sum_{i=1}^{I}\sum_{j=1}^{N_i} |Y_{i\cdot}(\omega_k) - Y_{\cdot\cdot}(\omega_k)|^2 \tag{7.84}$$

and

$$SSE(\omega_k) = \sum_{i=1}^{I}\sum_{j=1}^{N_i} |Y_{ij}(\omega_k) - Y_{i\cdot}(\omega_k)|^2, \tag{7.85}$$

which are analogous to the usual sums of squares in analysis of variance. Note that a dot $(\cdot)$ stands for a mean, taken over the appropriate subscript, so the

regression power component $SSR(\omega_k)$ is basically the power in the residuals of the group means from the overall mean and the error power component $SSE(\omega_k)$ reflects the departures of the group means from the original data values. Smoothing each component over L frequencies leads to the usual F-statistic (7.64) with $2L(I-1)$ and $2L(\sum_i N_i - I)$ degrees of freedom at each frequency ω of interest.

Example 7.7 Means Test for the Magnetic Resonance Imaging Data

Figure 7.1 showed the mean responses of subjects to various levels of periodic stimulation while awake and while under anesthesia, as collected in a pain perception experiment of Antognini et al. (1997). Three types of periodic stimuli were presented to awake and anesthetized subjects, namely, brushing, heat, and shock. The periodicity was introduced by applying the stimuli, brushing, heat, and shocks in on-off sequences lasting 32 seconds each and the sampling rate was one point every two seconds. The blood oxygenation level (BOLD) signal intensity (Ogawa et al., 1990) was measured at nine locations in the brain. Areas of activation were determined using a technique first described by Bandettini et al. (1993). The specfic locations of the brain where the signal was measured were Cortex 1: Primary Somatosensory, Contralateral, Cortex 2: Primary Somatosensory, Ipsilateral, Cortex 3: Secondary Somatosensory, Contralateral, Cortex 4: Secondary Somatosensory, Ipsilateral, Caudate, Thalamus 1: Contralateral, Thalamus 2: Ipsilateral, Cerebellum 1: Contralateral and Cerebellum 2: Ipsilateral. Figure 7.1 shows the mean response of subjects at Cortex 1 for each of the six treatment combinations, 1: Awake-Brush (5 subjects), 2: Awake-Heat (4 subjects), 3: Awake-Shock (5 subjects), 4: Low-Brush (3 subjects), 5: Low-Heat (5 subjects), and 6: Low-Shock(4 subjects). The objective of this first analysis is to test equality of these six group means, paying special attention to the 64-second period band (1/64 cycles per second) expected from the periodic driving stimuli. Because a test of equality is needed at each of the nine brain locations, we took $\alpha = .001$ to control for the overall error rate. Figure 7.9 shows F-statistics, computed from (7.64), with $L = 3$, and we see substantial signals for the four cortex locations and for the second cerebellum trace, but the effects are nonsignificant in the caudate and thalamus regions. Hence, we will retain the four cortex locations and the second cerebellum location for further analysis.

An Analysis of Variance Model

The arrangement of treatments for the fMRI data in Figure 7.1 suggests more information might be available than was obtained from the simple equality of means test. Separate effects caused by state of consciousness as well as the

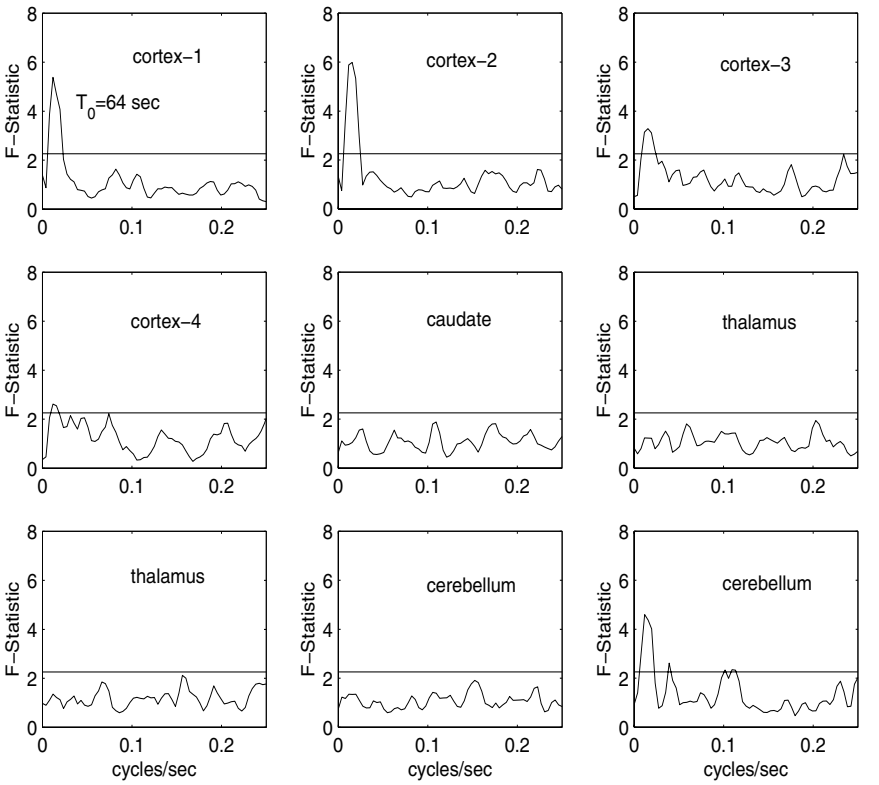


Figure 7.9 Frequency-dependent equality of means tests for fMRI data at 9 brain locations. $L = 3$ and critical value $F_{.001}(30, 120) = 2.26$.

separate treatments brush, heat, and shock might exist. The reduced signal present in the low shock mean suggests a possible interaction between the treatments and level of consciousness. The arrangement in the classical two-way table suggests looking at the analog of the two factor analysis of variance as a function of frequency. In this case, we would obtain a different version of the regression model (7.82) of the form

$$y_{ijkt} = \mu_t + \alpha_{it} + \beta_{jt} + \gamma_{ijt} + v_{ijkt} \tag{7.86}$$

for k-th individual undergoing the i-th level of some factor A and the j-th level of some other factor B, $i = 1, \ldots I, j = 1 \ldots, J, k = 1, \ldots n_{ij}$. The number of individuals in each cell can be different, as for the fMRI data in the next example. In the above model, we assume the response can be modeled as the sum of a mean, μ_t, a row effect (type of stimulus), α_{it}, a column effect (level

Table 7.4 Rows of the Design Matrix $\boldsymbol{z}_j'$ for fMRI Data. Number of Observations per Cell in Parentheses

	Awake	Low Anesthesia
Brush	1 1 0 1 1 0 (5)	1 1 0 -1 -1 0 (3)
Heat	1 0 1 1 0 1 (4)	1 0 1 -1 0 -1 (5)
Shock	1 -1 -1 1 -1 -1 (5)	1 -1 -1 -1 1 1 (4)

of consciousness), β_{jt} and an interaction, γ_{ijt}, with the usual restrictions

$$\sum_i \alpha_{it} = \sum_j \beta_{jt} = \sum_i \gamma_{ijt} = \sum_j \gamma_{ijt} = 0$$

required for a full rank design matrix Z in the overall regression model (7.81). If the number of observations in each cell were the same, the usual simple analogous version of the power components (7.84) and (7.85) would exist for testing various hypotheses. In the case of (7.86), we are interested in testing hypotheses obtained by dropping one set of terms at a time out of (7.86), so an A factor (testing $\alpha_{it} = 0$), a B factor ($\beta_{jt} = 0$), and an interaction term ($\gamma_{ijt} = 0$) will appear as components in the analysis of power. Because of the unequal numbers of observations in each cell, we often put the model in the form of the regression model (7.79)-(7.81).

Example 7.8 Analysis of Power Tests for the Magnetic Resonance Imaging Data

For the fMRI data given as the means in Figure 7.1, a model of the form (7.86) is plausible and will yield more detailed information than the simple equality of means test described earlier. The results of that test, shown in Figure 7.9, were that the means were different for the four cortex locations and for the second cerebellum location. We may examine these differences further by testing whether the mean differences are because of the nature of the stimulus or the consciousness level, or perhaps due to an interaction between the two factors. Unequal numbers of observations exist in the cells that contributed the means in Figure 7.1. For the regression vector,

$$(\mu_t, \alpha_{1t}, \beta_{1t}, \beta_{2t}, \gamma_{11t}, \gamma_{21t})',$$

the rows of the design matrix are as specified in Table 7.4. Note the restrictions given above for the parameters.

The results of testing the three hypotheses are shown in Figure 7.10 for the four cortex locations and the cerebellum, the components that showed some significant differences in the means in Figure 7.9. Again, the regression power components were smoothed over $L = 3$ frequencies.

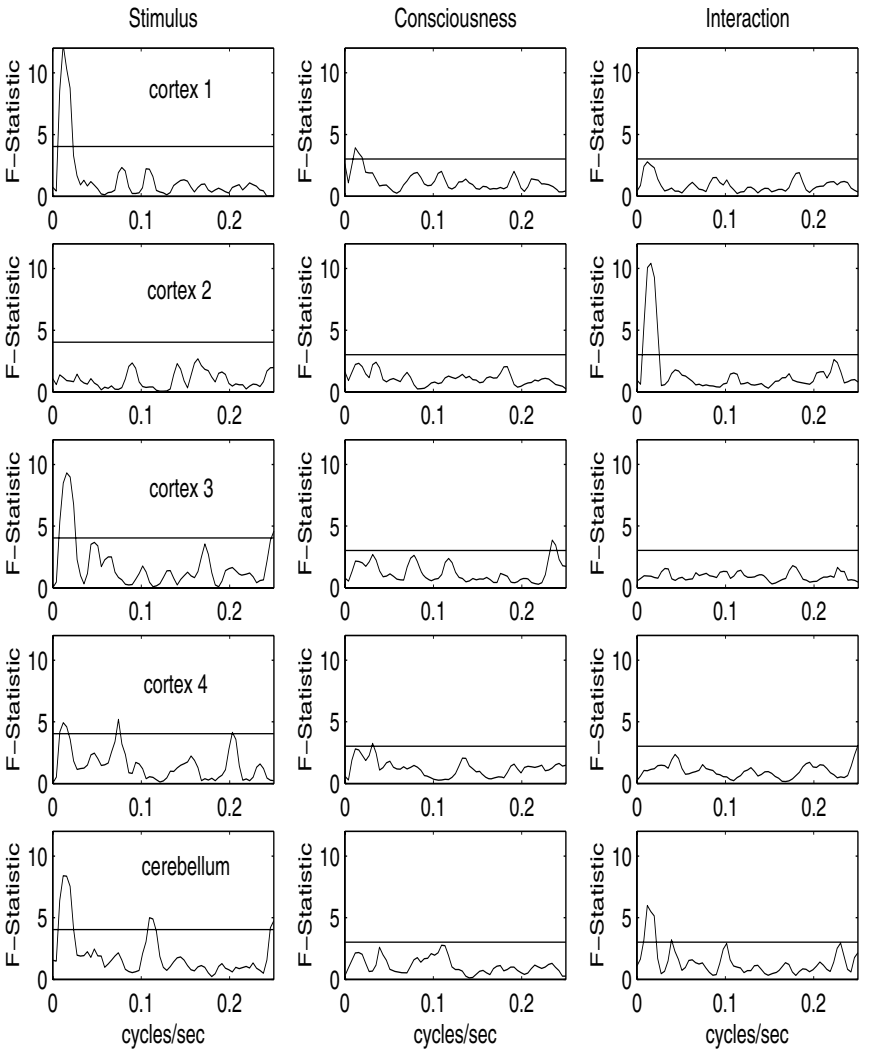


Figure 7.10 Analysis of power for fMRI data at five locations, $L = 3$ and critical values $F_{.001}(6, 120) = 4.04$ for stimulus and $F_{.001}(12, 120) = 3.02$ for consciousness and interaction.

Appealing to the ANOPOW results summarized in Table 7.3 for each of the subhypotheses, $q_2 = 1$ when the stimulus effect is dropped, and $q_2 = 2$ when either the conciousness effect or the interaction terms are dropped. Hence, $2Lq_2 = 6, 12$ for the two cases, with $N = \sum_{ij} n_{ij} = 26$ total observations. Here, the form of the stimulus has the major effect, with the brushing, heat, and shock means substantially different at the

probe frequency in four out of five cases. The level of consciousness was less significant and did not show the strong component at the signal frequency. A significant interaction occurred, however, at the ipsilateral component of the primary somatosensory cortex location. The more detailed model does separate the stimuli as having the major effect, but does not isolate which of the three might be more substantial than the other two.

Simultaneous Inference

In the previous examples involving the fMRI data, it would be helpful to focus on the components that contributed most to the rejection of the equal means hypothesis. One way to accomplish this is to develop a test for the significance of an arbitrary linear compound of the form

$$\Psi(\omega_k) = \boldsymbol{A}^*(\omega_k)\boldsymbol{B}(\omega_k), \tag{7.87}$$

where the components of the vector $\boldsymbol{A}(\omega_k) = (A_1(\omega_k), A_2(\omega_k), \ldots, A_q(\omega_k))'$ are chosen in such a way as to isolate particular linear functions of parameters in the regression vector $\boldsymbol{B}(\omega_k)$ in the regression model (7.81). This argument suggests developing a test of the hypothesis $\Psi(\omega_k) = 0$ for all possible values of the linear coefficients in the compound (7.87) as is done in the conventional analysis of variance approach (see, for example, Scheffé, 1959).

Recalling the material involving the regression models of the form (7.51), the linear compound (7.87) can be estimated by

$$\widehat{\Psi}(\omega_k) = \boldsymbol{A}^*(\omega_k)\widehat{\boldsymbol{B}}(\omega_k), \tag{7.88}$$

where $\widehat{\boldsymbol{B}}(\omega_k)$ is the estimated vector of regression coefficients given by (7.52) and independent of the error spectrum $s^2_{y\cdot z}(\omega_k)$ in (7.54). It is possible to show the maximum of the ratio

$$F(\boldsymbol{A}) = \frac{N-q}{q}\,\frac{|\widehat{\Psi}(\omega_k) - \Psi(\omega_k)|^2}{s^2_{y\cdot z}(\omega_k)Q(\boldsymbol{A})}, \tag{7.89}$$

where

$$Q(\boldsymbol{A}) = \boldsymbol{A}^*(\omega_k)S_z^{-1}(\omega_k)\boldsymbol{A}(\omega_k) \tag{7.90}$$

is bounded by a statistic that has an F-distribution with $2q$ and $2(N-q)$ degrees of freedom. Testing the hypothesis that the compound has a particular value, usually $\Psi(\omega_k) = 0$, then proceeds naturally, by comparing the statistic (7.89) evaluated at the hypothesized value with the α level point on an $F_{2q,2(N-q)}$ distribution. We can choose an infinite number of compounds of the form (7.87) and the test will still be valid at level α. As before, arguing the error spectrum is relatively constant over a band enables us to smooth the numerator and denominator of (7.89) separately over L frequencies so distribution involving the smooth components is $F_{2Lq,2L(N-q)}$.

Example 7.9 Simultaneous Inference for Magnetic Resonance Imaging Data

As an example, consider the previous tests for significance of the fMRI factors, in which we have indicated the primary effects are among the stimuli but have not investigated which of the stimuli, heat, brushing, or shock, had the most effect. To analyze this further, consider the means model (7.82) and a 6×1 contrast vector of the form

$$\widehat{\Psi} = \boldsymbol{A}^*(\omega_k)\widehat{\boldsymbol{B}}(\omega_k) = \sum_{i=1}^{6} A_i^*(\omega_k)\boldsymbol{Y}_{i\cdot}(\omega_k), \tag{7.91}$$

where the means are easily shown to be the regression coefficients in this particular case. In this case, the means are ordered by columns; the first three means are the the three levels of stimuli for the awake state, and the last three means are the levels for the anesthetized state. In this special case, the denominator terms are

$$Q = \sum_{i=1}^{6} \frac{|A_i(\omega_k)|^2}{N_i}, \tag{7.92}$$

with $SSE(\omega_k)$ available in (7.85). In order to evaluate the effect of a particular stimulus, like brushing over the two levels of consciousness, we may take $A_1(\omega_k) = A_4(\omega_k) = 1$ for the two brush levels and $A(\omega_k) = 0$ zero otherwise. From Figure 7.11, we see that, at the first and third cortex locations, brush and heat are both significant, whereas the fourth cortex shows only brush and the second cerebellum shows only heat. Shock appears to be transmitted relatively weakly, when averaged over the awake and mildly anesthetized states.

Multivariate Tests

Although it is possible to develop multivariate regression along lines analogous to the usual real valued case, we will only look at tests involving equality of group means and spectral matrices, because these tests appear to be used most often in applications. For these results, consider the p-variate time series $\boldsymbol{y}_{ijt} = (y_{ijt1}, \ldots, y_{ijtp})'$ to have arisen from observations on $j = 1, \ldots, N_i$ individuals in group i, all having mean $\boldsymbol{\mu}_{it}$ and stationary autocovariance matrix $\Gamma_i(h)$. Denote the DFTs of the group mean vectors as $\boldsymbol{Y}_{i\cdot}(\omega_k)$ and the $p \times p$ spectral matrices as $\widehat{f}_i(\omega_k)$ for the $i = 1, 2, \ldots, I$ groups. Assume the same general properties as for the vector series considered in §7.3.

In the multivariate case, we obtain the analogous versions of (7.84) and (7.85) as the between cross-power and within cross-power matrices

$$SPR(\omega_k) = \sum_{i=1}^{I}\sum_{j=1}^{N_i} \big(\boldsymbol{Y}_{i\cdot}(\omega_k) - \boldsymbol{Y}_{\cdot\cdot}(\omega_k)\big)\big(\boldsymbol{Y}_{i\cdot}(\omega_k) - \boldsymbol{Y}_{\cdot\cdot}(\omega_k)\big)^* \tag{7.93}$$

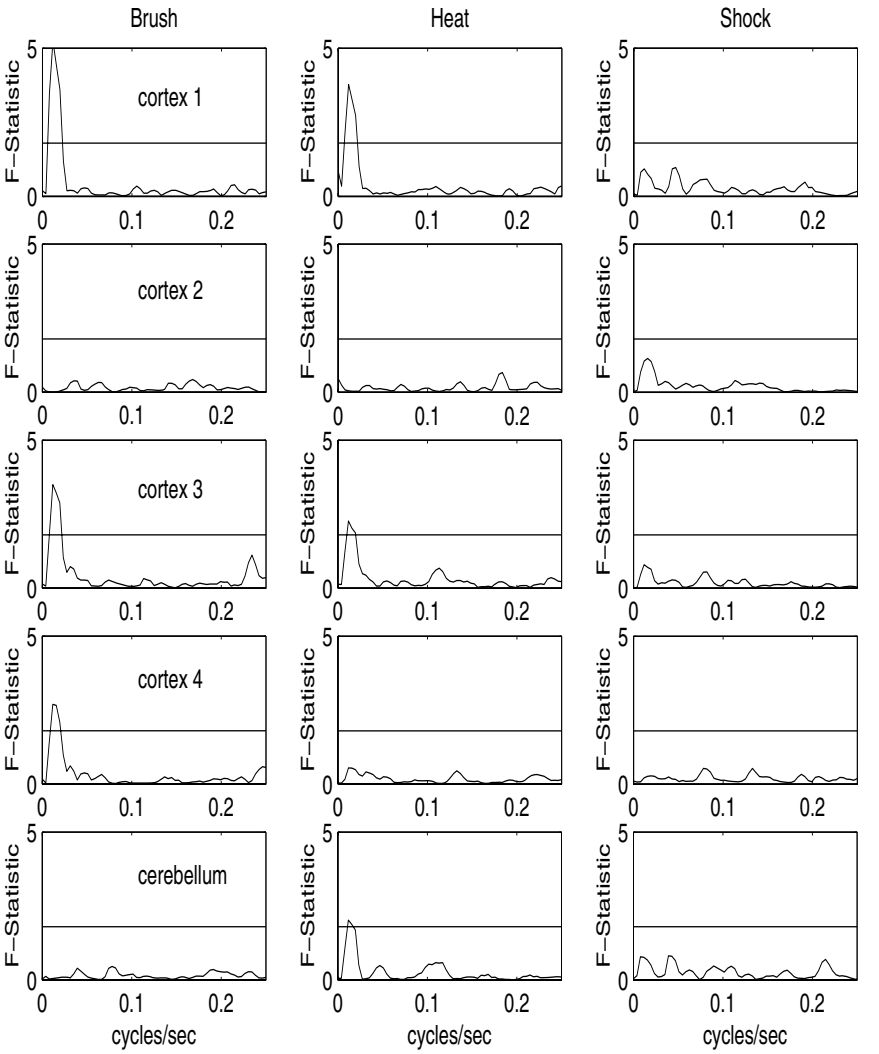


Figure 7.11 Power in simultaneous linear compounds at five locations, enhancing brush, heat, and shock effects, $L = 3, F_{.001}(36, 120) = 1.80$.

and

$$SPE(\omega_k) = \sum_{i=1}^{I} \sum_{j=1}^{N_i} \big(\boldsymbol{Y}_{ij}(\omega_k) - \boldsymbol{Y}_{i\cdot}(\omega_k)\big)\big(\boldsymbol{Y}_{ij}(\omega_k) - \boldsymbol{Y}_{i\cdot}(\omega_k)\big)^*. \tag{7.94}$$

The equality of means test is rejected using the fact that the likelihood ratio

test yields a monotone function of

$$\Lambda(\omega_k) = \frac{|SPE(\omega_k)|}{|SPE(\omega_k) + SPR(\omega_k)|}. \tag{7.95}$$

Khatri (1965) and Hannan (1970) give the approximate distribution of the statistic

$$\chi^2_{2(I-1)p} = -2\left(\sum N_i - I - p - 1\right) \log \Lambda(\omega_k) \tag{7.96}$$

as chi-squared with $2(I-1)p$ degrees of freedom when the group means are equal.

The case of $I = 2$ groups reduces to Hotelling's T^2, as has been shown by Giri (1965), where

$$T^2 = \frac{N_1 N_2}{(N_1 + N_2)} \big[\boldsymbol{Y}_{1\cdot}(\omega_k) - \boldsymbol{Y}_{2\cdot}(\omega_k)\big]^* \widehat{f}_v^{-1}(\omega_k) \big[\boldsymbol{Y}_{1\cdot}(\omega_k) - \boldsymbol{Y}_{2\cdot}(\omega_k)\big], \tag{7.97}$$

where

$$\widehat{f}_v(\omega_k) = \frac{SPE(\omega_k)}{\sum_i N_i - I} \tag{7.98}$$

is the pooled error spectrum given in (7.94),with $I = 2$. The test statistic, in this case, is

$$F_{2p,2(N_1+N_2-p-1)} = \frac{(N_1 + N_2 - 2)p}{(N_1 + N_2 - p - 1)} T^2, \tag{7.99}$$

which was shown by Giri (1965) to have the indicated limiting F-distribution with $2p$ and $2(N_1 + N_2 - p - 1)$ degrees of freedom when the means are the same. The classical t-test for inequality of two univariate means will be just (7.98) and (7.99) with $p = 1$.

Testing equality of the spectral matrices is also of interest, not only for discrimination and pattern recognition, as considered in the next section, but also as a test indicating whether the equality of means test, which assumes equal spectral matrices, is valid. The test evolves from the likelihood ration criterion, which compares the single group spectral matrices

$$\widehat{f}_i(\omega_k) = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} \big(\boldsymbol{Y}_{ij}(\omega_k) - \boldsymbol{Y}_{i\cdot}(\omega_k)\big) \big(\boldsymbol{Y}_{ij}(\omega_k) - \boldsymbol{Y}_{i\cdot}(\omega_k)\big)^* \tag{7.100}$$

with the pooled spectral matrix (7.98). A modification of the likelihood ratio test, which incorporates the degrees of freedom $M_i = N_i - 1$ and $M = \sum M_i$ rather than the sample sizes into the likelihood ratio statistic, uses

$$L'(\omega_k) = \frac{M^{Mp}}{\prod_{i=1}^{I} M_i^{M_i p}} \frac{\prod |M_i \widehat{f}_i(\omega_k)|^{M_i}}{|M \widehat{f}_v(\omega_k)|^M}. \tag{7.101}$$

Krishnaiah et al. (1976) have given the moments of $L'(\omega_k)$ and calculated 95% critical points for $p = 3, 4$ using a Pearson Type I approximation. For

reasonably large samples involving smoothed spectral estimators, the approximation involving the first term of the usual chi-squared series will suffice and Shumway (1982) has given

$$\chi^2_{(I-1)p^2} = -2r \log L'(\omega_k), \tag{7.102}$$

where

$$1 - r = \frac{(p+1)(p-1)}{6p(I-1)} \left(\sum_i M_i^{-1} - M^{-1} \right), \tag{7.103}$$

with an approximate chi-squared distribution with $(I-1)p^2$ degrees of freedom when the spectral matrices are equal. Introduction of smoothing over L frequencies leads to replacing M_j and M by LM_j and LM in the equations above.

Of course, it is often of great interest to use the above result for testing equality of two univariate spectra, and it is obvious from the material in Chapter 4

$$F_{2LM_1, 2LM_2} = \frac{\widehat{f}_1(\omega)}{\widehat{f}_2(\omega)} \tag{7.104}$$

will have the requisite F-distribution with $2LM_1$ and $2LM_2$ degrees of freedom when spectra are smoothed over L frequencies.

Example 7.10 Equality of Means and Spectral Matrices for Earthquakes and Explosions

An interesting problem arises when attempting to develop a methodology for discriminating between waveforms originating from explosions and those that came from the more commonly occurring earthquakes. Figure 7.2 shows a small subset of a larger population of bivariate series consisting of two phases from each of eight earthquakes and eight explosions. If the large–sample approximations to normality hold for the DFTs of these series, it is of interest to known whether the differences between the two classes are better represented by the mean functions or by the spectral matrices. The tests described above can be applied to look at these two questions. The upper left panel of Figure 7.12 shows the test statistic (7.99) with the straight line denoting the critical level for $\alpha = .001$, i.e., $F_{.001}(4, 26) = 7.36$, for equal means using $L = 1$, and the test statistic remains well below its critical value at all frequencies, implying that the means of the two classes of series are not significantly different. Checking Figure 7.2 shows little reason exists to suspect that either the earthquakes or explosions have a nonzero mean signal. Checking the equality of the spectra and the spectral matrices, however, leads to a different conclusion. Some smoothing ($L = 21$) is useful here, and univariate tests on both the P and S components using (7.104) and $N_1 = N_2 = 8$ lead to strong rejections of the equal spectra

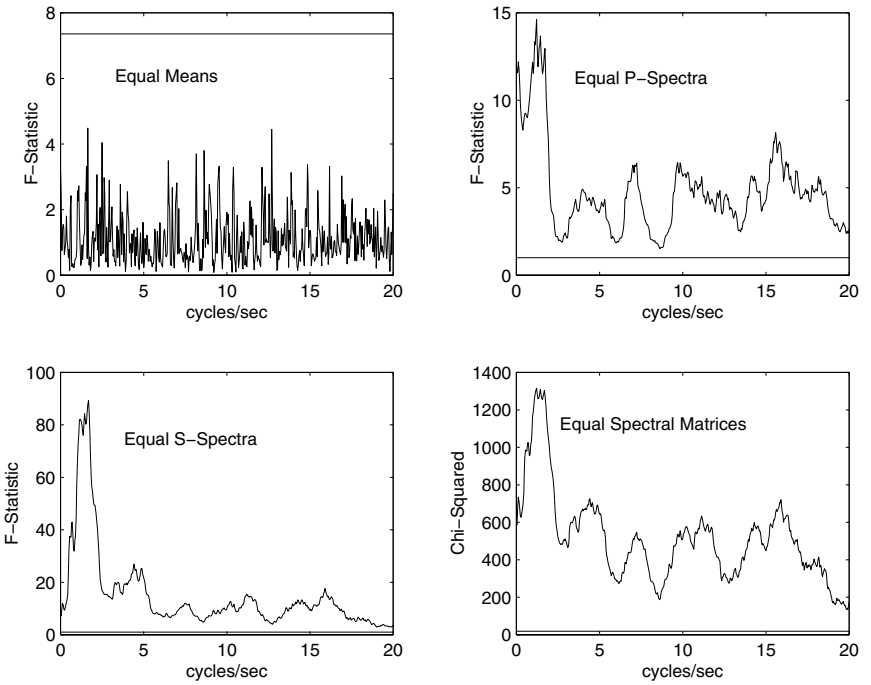


Figure 7.12 Tests for equality of means, spectra, and spectral matrices for the earthquake and explosion data $p = 2, L = 21, n = 1024$ points at 40 points per second.

hypotheses, with $F_{.001}(\infty, \infty) = 1.00$ exceeded at almost all frequencies. The rejection seems stronger for the S component and we might tentatively identify that component as being dominant. Testing equality of the spectral matrices using (7.102) and $\chi^2_{.001}(4) = 18.47$ shows a similar strong rejection of the equality of spectral matrices. We use these results to suggest optimal discriminant functions based on spectral differences in the next section.

7.7 Discrimination and Cluster Analysis

The extension of classical pattern-recognition techniques to experimental time series is a problem of great practical interest. A series of observations indexed in time often produces a pattern that may form a basis for discriminating between different classes of events. As an example, consider Figure 7.2, which shows regional (100-2000 km) recordings of several typical Scandinavian earthquakes and mining explosions measured by stations in Scandinavia. A listing

of the events is given in Kakizawa et al. (1998). The problem of discriminating between mining explosions and earthquakes is a reasonable proxy for the problem of discriminating between nuclear explosions and earthquakes. This latter problem is one of critical importance for monitoring a comprehensive test-ban treaty. Time series classification problems are not restricted to geophysical applications, but occur under many and varied circumstances in other fields. Traditionally, the detecting of a signal embedded in a noise series has been analyzed in the engineering literature by statistical pattern recognition techniques (see Problems 7.13 and 7.14).

The historical approaches to the problem of discriminating among different classes of time series can be divided into two distinct categories. The optimality approach, as found in the engineering and statistics literature, makes specific Gaussian assumptions about the probability density functions of the separate groups and then develops solutions that satisfy well-defined minimum error criteria. Typically, in the time series case, we might assume the difference between classes is expressed through differences in the theoretical mean and covariance functions and use likelihood methods to develop an optimal classification function. A second class of techniques, which might be described as a feature extraction approach, proceeds more heuristically by looking at quantities that tend to be good visual discriminators for well-separated populations and have some basis in physical theory or intuition. Less attention is paid to finding functions that are approximations to some well-defined optimality criterion.

As in the case of regression, both time domain and frequency domain approaches to discrimination will exist. For relatively short univariate series, a time domain approach that follows conventional multivariate discriminant analysis as described in conventional multivariate texts, such as Anderson (1984) or Johnson and Wichern (1992) may be preferable. We might even characterize differences by the autocovariance functions generated by different ARMA or state-space models. For longer multivariate time series that can be regarded as stationary after the common mean has been subtracted, the frequency domain approach will be easier computationally because the np dimensional vector in the time domain, represented here as $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_t', \ldots, \boldsymbol{x}_n')'$, with $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})'$, will reduced to separate computations made on the p-dimensional DFTs. This happens because of the approximate independence of the DFTs, $\boldsymbol{X}(\omega_k), 0 \leq \omega_k \leq 1$, a property that we have often used in preceding chapters.

Finally, the grouping properties of measures like the discrimination information and likelihood-based statistics can be used to develop measures of disparity for clustering multivariate time series. In this section, we define a measure of disparity between two multivariate times series by the spectral matrices of the two processes and then apply hierarchical clustering and partitioning techniques to identify natural groupings within the bivariate earthquake and explosion populations.

The General Discrimination Problem

The general problem of classifying a vector time series $\boldsymbol{x}$ occurs in the following way. We observe a time series $\boldsymbol{x}$ known to belong to one of g populations, denoted by $\Pi_1, \Pi_2, \ldots, \Pi_g$. The general problem is to assign or classify this observation into one of the g groups in some optimal fashion. An example might be the $g = 2$ populations of earthquakes and explosions shown in Figure 7.2. We would like to classify the unknown event, shown as NZ in the bottom two panels, as belonging to either the earthquake (Π_1) or explosion (Π_2) populations. To solve this problem, we need an optimality criterion that leads to a statistic $T(\boldsymbol{x})$ that can be used to assign the NZ event to either the earthquake or explosion populations. To measure the success of the classification, we need to evaluate errors that can be expected in the future relating to the number of earthquakes classified as explosions (false alarms) and the number of explosions classified as earthquakes (missed signals).

The problem can be formulated by assuming the observed series $\boldsymbol{x}$ has a probability density $p_i(\boldsymbol{x})$ when the observed series is from population Π_i for $i = 1, \ldots, g$. Then, partition the space spanned by the np-dimensional process $\boldsymbol{x}$ into g mutually exclusive regions $R_1, R_2, \ldots, R_g$ such that, if $\boldsymbol{x}$ falls in R_i, we assign $\boldsymbol{x}$ to population Π_i. The misclassification probability is defined as the probability of classifying the observation into population Π_j when it belongs to Π_i, for $j \neq i$ and would be given by the expression

$$P(j|i) = \int_{R_j} p_i(\boldsymbol{x}) \; d\boldsymbol{x}. \tag{7.105}$$

The overall total error probability depends also on the prior probabilities, say, $\pi_1, \pi_2, \ldots, \pi_g$, of belonging to one of the g groups. For example, the probability that an observation $\boldsymbol{x}$ originates from Π_i and is then classified into Π_j is obviously $\pi_i P(j|i)$, and the total error probability becomes

$$P_e = \sum_{i=1}^{g} \pi_i \sum_{j \neq i} P(j|i). \tag{7.106}$$

Although costs have not been incorporated into (7.106), it is easy to do so by multiplying $P(j|i)$ by $C(j|i)$, the cost of assigning a series from population Π_i to Π_j.

The overall error P_e is minimized by classifying $\boldsymbol{x}$ into Π_i if

$$\frac{p_i(\boldsymbol{x})}{p_j(\boldsymbol{x})} > \frac{\pi_j}{\pi_i} \tag{7.107}$$

for all $j \neq i$ (see, for example, Anderson, 1984). A quantity of interest, from the Bayesian perspective, is the posterior probability an observation belongs to population Π_i, conditional on observing $\boldsymbol{x}$, say,

$$P(\Pi_i|\boldsymbol{x}) = \frac{\pi_i p_i(\boldsymbol{x})}{\sum_j \pi_j(\boldsymbol{x}) p_j(\boldsymbol{x})}. \tag{7.108}$$

The procedure that classifies $\boldsymbol{x}$ into the population Π_i for which the posterior probability is largest is equivalent to that implied by using the criterion (7.107). The posterior probabilities give an intuitive idea of the relative odds of belonging to each of the plausible populations.

Many situations occur, such as in the classification of earthquakes and explosions, in which there are only $g = 2$ populations of interest. For two populations, the Neyman–Pearson lemma implies, in the absence of prior probabilities, classifying an observation into Π_1 when

$$\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} > K \tag{7.109}$$

minimizes each of the error probabilities for a fixed value of the other. The rule is identical to the Bayes rule (7.107) when $K = \pi_2/\pi_1$.

The theory given above takes a simple form when the vector $\boldsymbol{x}$ has a p-variate normal distribution with mean vectors $\boldsymbol{\mu}_j$ and covariance matrices Σ_j under Π_j for $j = 1, 2, \ldots, g$. In this case, simply use

$$p_j(\boldsymbol{x}) = (2\pi)^{-p/2}|\Sigma_j|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_j)'\Sigma_j^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_j)\right\}. \tag{7.110}$$

The classification functions are conveniently expressed by quantities that are proportional to the logarithms of the densities, say,

$$g_j(\boldsymbol{x}) = -\frac{1}{2}\ \ln|\Sigma_j| - \frac{1}{2}\ \boldsymbol{x}'\Sigma_j^{-1}\boldsymbol{x} + \boldsymbol{\mu}_j'\Sigma_j^{-1}\boldsymbol{x} - \frac{1}{2}\ \boldsymbol{\mu}_j'\Sigma_j^{-1}\boldsymbol{\mu}_j + \ln\pi_j. \tag{7.111}$$

In expressions involving the log likelihood, we will generally ignore terms involving the constant $-\ln 2\pi$. For this case, we may assign an observation $\boldsymbol{x}$ to population Π_i whenever

$$g_i(\boldsymbol{x}) > g_j(\boldsymbol{x}) \tag{7.112}$$

for $j \neq i, j = 1, \ldots, g$ and the posterior probability (7.108) has the form

$$P(\Pi_i|\boldsymbol{x}) = \frac{\exp\{g_i(\boldsymbol{x})\}}{\sum_j \exp\{g_j(\boldsymbol{x})\}}.$$

A common situation occurring in applications involves classification for $g = 2$ groups under the assumption of multivariate normality and equal covariance matrices; i.e., $\Sigma_1 = \Sigma_2 = \Sigma$. Then, the criterion (7.112) can be expressed in terms of the linear discriminant function

$$\begin{aligned} d_l(\boldsymbol{x}) &= g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \ln\frac{\pi_1}{\pi_2}, \end{aligned} \tag{7.113}$$

where we classify into Π_1 or Π_2 according to whether $d_l(\boldsymbol{x}) \geq 0$ or $d_l(\boldsymbol{x}) < 0$. The linear discriminant function is clearly a combination of normal variables

and, for the case $\pi_1 = \pi_2 = .5$, will have mean $D^2/2$ under Π_1 and mean $-D^2/2$ under Π_2, with variances given by D^2 under both hypotheses, where

$$D^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{7.114}$$

is the Mahalanobis distance between the mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. In this case, the two misclassification probabilities (7.1) are

$$\begin{aligned} P(1|2) &= P(2|1) \\ &= \Phi\left(-\frac{D}{2}\right), \end{aligned} \tag{7.115}$$

and the performance is directly related to the Mahalanobis distance (7.114).

For the case in which the covariance matrices cannot be assumed to be the the same, the discriminant function takes a different form, with the difference $g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$ taking the form

$$\begin{aligned} d_q(\boldsymbol{x}) &= -\frac{1}{2}\ \ln\frac{|\Sigma_1|}{|\Sigma_2|} - \frac{1}{2}\ \boldsymbol{x}'(\Sigma_1^{-1} - \Sigma_2^{-1})\boldsymbol{x} \\ &\quad +(\boldsymbol{\mu}_1'\Sigma_1^{-1} - \boldsymbol{\mu}_2'\Sigma_2^{-1})\boldsymbol{x} + \ln\frac{\pi_1}{\pi_2} \end{aligned} \tag{7.116}$$

for $g = 2$ groups. This discriminant function differs from the equal covariance case in the linear term and in a nonlinear quadratic term involving the differing covariance matrices. The distribution theory is not tractable for the quadratic case so no convenient expression like (7.115) is available for the error probabilities for the quadratic discriminant function.

A difficulty in applying the above theory to real data is that the group mean vectors $\boldsymbol{\mu}_j$ and covariance matrices Σ_j are seldom known. Some engineering problems, such as the detection of a signal in white noise, assume the means and covariance parameters are known exactly, and this can lead to an optimal solution (see Problems 7.14 and 7.15). In the classical multivariate situation, it is possible to collect a sample of N_i training vectors from group Π_i, say, $\boldsymbol{x}_{ij}$, for $j = 1, \ldots, N_i$, and use them to estimate the mean vectors and covariance matrices for each of the groups $i = 1, 2, \ldots, g$; i.e., simply choose $\boldsymbol{x}_{i\cdot}$ and

$$S_i = (N_i - 1)^{-1}\sum_{j=1}^{N_i}(\boldsymbol{x}_{ij} - \boldsymbol{x}_{i\cdot})(\boldsymbol{x}_{ij} - \boldsymbol{x}_{i\cdot})' \tag{7.117}$$

as the estimators for $\boldsymbol{\mu}_i$ and Σ_i, respectively. In the case in which the covariance matrices are assumed to be equal, simply use the pooled estimator

$$S = \left(\sum_i N_i - g\right)^{-1}\sum_i (N_i - 1)S_i. \tag{7.118}$$

For the case of a linear discriminant function, we may use

$$\widehat{g_i(\boldsymbol{x})} = \boldsymbol{x}_{i\cdot}'S^{-1}\boldsymbol{x} - \frac{1}{2}\ \boldsymbol{x}_{i\cdot}'S^{-1}\boldsymbol{x}_{i\cdot} + \log\pi_i \tag{7.119}$$

as a simple estimator for $g_i(\boldsymbol{x})$. For large samples, $\boldsymbol{x}_{i\cdot}$ and S converge to $\boldsymbol{\mu}_i$ and Σ in probability so $\widehat{g_i(\boldsymbol{x})}$ converges in distribution to $g_i(\boldsymbol{x})$ in that case. The procedure works reasonably well for the case in which $N_i, i = 1, \ldots g$ are large, relative to the length of the series n, a case that is relatively rare in time series analysis. For this reason, we will resort to using spectral approximations for the case in which data are given as long time series.

The performance of sample discriminant functions can be evaluated in several different ways. If the population parameters are known, (7.114) and (7.115) can be evaluated directly. If the parameters are estimated, the estimated Mahalanobis distance $\widehat{D^2}$ can be substituted for the theoretical value in very large samples. Another approach is to calculate the apparent error rates using the result of applying the classification procedure to the training samples. If n_{ij} denotes the number of observations from population Π_j classified into Π_i, the sample error rates can be estimated by the ratio

$$\widehat{P(i|j)} = \frac{n_{ij}}{\sum_i n_{ij}} \tag{7.120}$$

for $i \neq j$. If the training samples are not large, this procedure may be biased and a resampling option like cross validation or the bootstrap can be employed. A simple version of cross validation is the jacknife procedure proposed by Lachenbruch and Mickey (1968), which holds out the observation to be classified, deriving the classification function from the remaining observations. Repeating this procedure for each of the members of the training sample and computing (7.120) for the holdout samples leads to better estimators of the error rates.

Example 7.11 Discriminant Analysis Using Amplitudes from Earthquakes and Explosions

We can give a simple example of applying the above procedures to the logarithms of the amplitudes of the separate P and S components of the original earthquake and explosion traces. The logarithms (base 10) of the maximum peak-to-peak amplitudes of the P and S components, denoted by $\log_{10} P$ and $\log_{10} S$, can be considered as two-dimensional feature vectors, say, $\boldsymbol{x} = (x_1, x_2)' = (\log_{10} P, \log_{10} S)'$, from a bivariate normal population with differering means and covariances. The original data, from Kakizawa et al. (1998), are shown in Table 7.5 and in the left-hand panel of Figure 7.13. The table includes the Novaya Zemlya (NZ) event of unknown origin. The tendency of the earthquakes to have higher values for $\log_{10} S$, relative to $\log_{10} P$ has been noted by many and the use of the logarithm of the ratio, i.e., $\log_{10} P - \log_{10} S$ in some references (see Lay, 1997, pp. 40-41) is a tacit indicator that a linear function of the two parameters will be a useful discriminant.

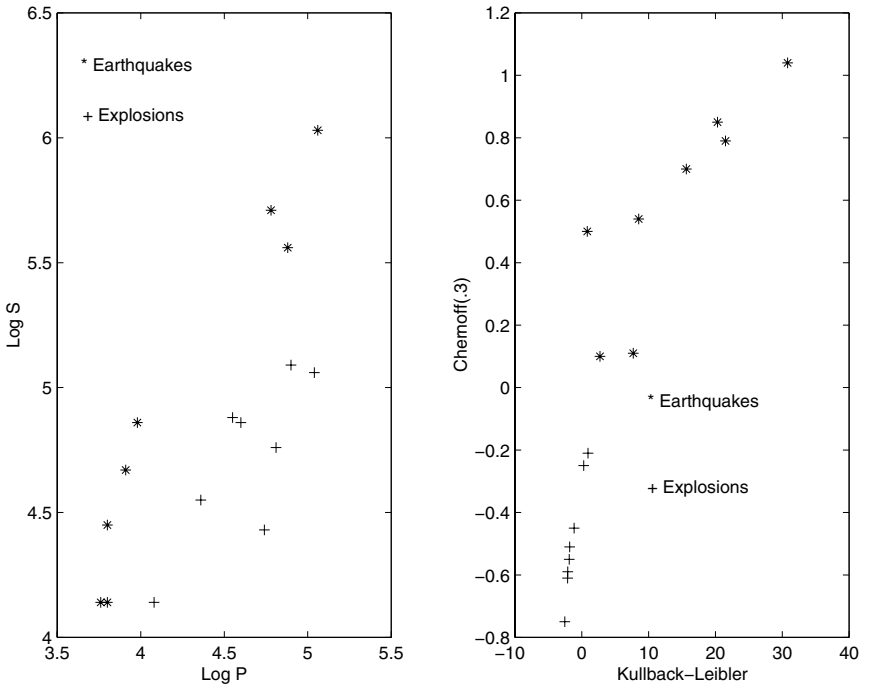


Figure 7.13 Classification of earthquakes and explosions using the magnitude features (left panel) and the K-L and Chernoff disparity measures (right panel).

The sample means $\boldsymbol{x}_{1\cdot} = (4.25, 4.95)'$ and $\boldsymbol{x}_{2\cdot} = (4.64, 4.73)'$, and covariance matrices

$$S_1 = \begin{pmatrix} .3096 & .3954 \\ .3954 & .5378 \end{pmatrix}$$

and

$$S_2 = \begin{pmatrix} .0954 & .0804 \\ .0804 & .1070 \end{pmatrix}$$

are immediate from (7.117), with the pooled covariance matrix given by

$$S = \begin{pmatrix} .2025 & .2379 \\ .2379 & .3238 \end{pmatrix}$$

from (7.118). Although the covariance matrices are not equal, we try the linear discriminant function anyway, which yields (with equal prior probabilities $\pi_1 = \pi_2 = .5$) the sample discriminant functions

$$\widehat{g_1(\boldsymbol{x})} = 22.12x_1 - .98x_2 - 45.23$$

and

$$\widehat{g_2(\boldsymbol{x})} = 42.61x_2 - 16.8x_2 - 59.80$$

Table 7.5 Logarithms of Maximum Peak-to-Peak Amplitudes from P and S Components for Eight Earthquakes and Eight Explosions

EQ	$\log_{10} P$	$\log_{10} S$	EXP	$\log_{10} P$	$\log_{10} S$
1	3.91	4.67	1	4.55	4.88
2	4.78	5.71	2	4.74	4.43
3	3.98	4.86	3	4.90	5.09
4	3.76	4.14	3	4.60	4.86
5	3.80	4.14	5	4.81	4.76
6	4.88	5.56	6	4.36	4.55
7	5.06	6.03	6	5.04	5.06
8	3.80	4.45	8	4.08	4.14
NZ	3.18	3.27			

from (7.119), with the estimated linear discriminant function (7.113) as

$$\widehat{d_l(\boldsymbol{x})} = -20.49x_1 + 15.82x_2 + 14.57,$$

indicating $\log_{10} S - \log_{10} P = x_2 - x_1$ is not far from the optimal linear discriminant function. The jacknifed posterior probabilities of being an earthquake for the earthquake group ranged from .791 to 1.000, whereas the explosion probabilities for the explosion group ranged from .814 to .998, except for the first explosion, which was classified as an earthquake with a posterior probability of .949. Hence, $n_{12} = 1$ for this particular example. The unknown event, NZ, was classified as an earthquake, with posterior probability .753. Components of the vector for the unknown event NZ were well outside the range of the values spanned by the training set, so the classification here is somewhat suspect. The quadratic discriminant might be more appropriate here, given the observed differences in the two covariance matrices. Applying the sample version of (7.116) leads to essentially the same results, namely, the misclassification of the first earthquake as an explosion with a posterior probability of .807 and the classification of the unknown NZ event into the earthquake group.

Frequency Domain Discrimination

The feature extraction approach often works well for discriminating between classes of univariate or multivariate series when there is a simple low-dimensional vector that seems to capture the essence of the differences between the classes. It still seems sensible, however, to develop optimal methods for classification that exploit the differences between the multivariate means and covariance matrices in the time series case. Such methods can be based on the Whittle approximation to the log likelihood given in §7.2. In this case, the vector DFTs, say, $\boldsymbol{X}(\omega_k)$, are assumed to be approximately normal, with means $\boldsymbol{M}_j(\omega_k)$ and spectral matrices $f_j(\omega_k)$ for population Π_j at frequencies

$\omega_k = k/n$, for $k = 0, 1, \ldots [n/2]$, and are approximately uncorrelated at different frequencies, say, ω_k and ω_ℓ for $k \neq \ell$. Then, writing the complex normal densities as in §7.2 leads to a criterion similar to (7.111); namely,

$$\begin{aligned} g_j(\boldsymbol{X}) &= \ln \pi_j - \sum_{0<\omega_k<1/2} \Big[\ln |f_j(\omega_k)| + \boldsymbol{X}^*(\omega_k) f_j^{-1}(\omega_k) \boldsymbol{X}(\omega_k) \\ &\quad - 2\boldsymbol{M}_j^*(\omega_k) f_j^{-1}(\omega_k) \boldsymbol{X}(\omega_k) + \boldsymbol{M}_j^*(k) f_j^{-1}(\omega_k) \boldsymbol{M}_j(\omega_k) \Big], \quad (7.121) \end{aligned}$$

where the sum goes over frequencies for which $|f_j(\omega_k)| \neq 0$. The periodicity of the spectral density matrix and DFT allows adding over $0 < k < 1/2$. The classification rule is as in (7.112).

In the time series case, it is more likely the discriminant analysis involves assuming the covariance matrices are different and the means are equal. For example, the tests, shown in Figure 7.12, imply, for the earthquakes and explosions, the primary differences are in the bivariate spectral matrices and the means are essentially the same. For this case, it will be convenient to write the Whittle approximation to the log likelihood in the form

$$\ln p_j(\boldsymbol{X}) = \sum_{0<\omega_k<1/2} \Big[- \ln |f_j(\omega_k)| - \boldsymbol{X}^*(\omega_k) f_j^{-1}(\omega_k) \boldsymbol{X}(\omega_k) \Big], \quad (7.122)$$

where we have omitted the prior probabilities from the equation. The quadratic detector in this case can be written in the form

$$\ln p_j(\boldsymbol{X}) = \sum_{0<\omega_k<1/2} \Big[- \ln |f_j(\omega_k)| - \mathrm{tr}\big\{ I(\omega_k) f_j^{-1}(\omega_k) \big\} \Big], \quad (7.123)$$

where

$$I(\omega_k) = \boldsymbol{X}(\omega_k) \boldsymbol{X}^*(\omega_k) \quad (7.124)$$

denotes the periodogram matrix. For equal prior probabilities, we may assign an observation $\boldsymbol{x}$ into population Π_i whenever

$$\ln p_i(\boldsymbol{X}) > \ln p_j(\boldsymbol{X}) \quad (7.125)$$

for $j \neq i, j = 1, 2, \ldots, g$.

Numerous authors have considered various versions of discriminant analysis in the frequency domain. Shumway and Unger (1974) considered (7.121) for $p = 1$ and equal covariance matrices, so the criterion reduces to a simple linear one. They apply the criterion to discriminating between earthquakes and explosions using teleseismic P wave data in which the means over the two groups might be considered as fixed. Alagón (1989) and Dargahi-Noubary and Laycock (1981) considered discriminant functions of the form (7.121) in the univariate case when the means are zero and the spectra for the two groups are different. Taniguchi et al. (1994) adopted (7.122) as a criterion and discussed

its non-Gaussian robustness. Shumway (1982) reviews general discriminant functions in both the univariate and multivariate time series cases.

Measures of Disparity

Before proceeding to examples of discriminant and cluster analysis, it is useful to consider the relation to the Kullback–Leibler (K-L) discrimination information, as defined in Problem 2.4 of Chapter 2. Using the spectral approximation and noting the periodogram matrix has the approximate expectation

$$E_j I(\omega_k) = f_j(\omega_k)$$

under the assumption that the data come from population Π_j, and approximating the ratio of the densities by

$$\ln \frac{p_1(\boldsymbol{X})}{p_2(\boldsymbol{X})} = \sum_{0<\omega_k<1/2} \left[-\ln \frac{|f_1(\omega_k)|}{|f_2(\omega_k)|} - \mathrm{tr}\left\{ \left(f_2^{-1}(\omega_k) - f_1^{-1}(\omega_k)\right) I(\omega_k) \right\} \right],$$

we may write the approximate discrimination information as

$$\begin{aligned} I(f_1; f_2) &= \frac{1}{n} E_1 \ln \frac{p_1(\boldsymbol{X})}{p_2(\boldsymbol{X})} \\ &= \frac{1}{n} \sum_{0<\omega_k<1/2} \left[\mathrm{tr}\left\{ f_1(\omega_k) f_2^{-1}(\omega_k) \right\} - \ln \frac{|f_1(\omega_k)|}{|f_2(\omega_k)|} - p \right]. \quad (7.126) \end{aligned}$$

The approximation may be carefully justified by noting the multivatiate normal time series $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2' \ldots, \boldsymbol{x}_n')$ with zero means and $np \times np$ stationary covariance matrices Γ_1 and Γ_2 will have p, $n \times n$ blocks, with elements of the form $\gamma_{ij}^{(l)}(s-t), s, t = 1, \ldots, n, i, j = 1, \ldots, p$ for population $\Pi_\ell, \ell = 1, 2$. The discrimination information, under these conditions, becomes

$$\begin{aligned} I(1; 2 : \boldsymbol{x}) &= \frac{1}{n} E_1 \ln \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} \\ &= \frac{1}{2n} \left[\mathrm{tr}\left\{ \Gamma_1 \Gamma_2^{-1} \right\} - \ln \frac{|\Gamma_1|}{|\Gamma_2|} - np \right]. \quad (7.127) \end{aligned}$$

The limiting result

$$\lim_{n\to\infty} I(1; 2 : \boldsymbol{x}) = \frac{1}{2} \int_{-1/2}^{1/2} \left[\mathrm{tr}\{ f_1(\omega) f_2^{-1}(\omega) \} - \ln \frac{|f_1(\omega)|}{|f_2(\omega)|} - p \right] d\omega$$

has been shown, in various forms, by Pinsker (1964), Hannan (1970), and Kazakos and Papantoni-Kazakos (1980). The discrete version of (7.126) is just the approximation to the integral of the limiting form. The K-L measure of disparity is not a true distance, but it can be shown that $I(1; 2) \geq 0$, with equality if and only if $f_1(\omega) = f_2(\omega)$ almost everywhere. This result makes it potentially suitable as a measure of disparity between the two densities.

A connection exists, of course, between the discrimination information number, which is just the expectation of the likelihood criterion and the likelihood itself. For example, we may measure the disparity between the sample and the process defined by the theoretical spectrum $f_j(\omega_k)$ corresponding to population Π_j in the sense of Kullback (1978), as $I(\widehat{f}; f_j)$, where

$$\widehat{f}(\omega_k) = L^{-1} \sum_{\ell=-m}^{m} I(\omega_k + \ell/n) \tag{7.128}$$

denotes the smoothed spectral matrix. The likelihood ratio criterion can be thought of as measuring the disparity between the periodogram and the theoretical spectrum for each of the populations. To make the discrimination information finite, we replace the periodogram implied by the log likelihood by the sample spectrum. In this case, the classification procedure can be regarded as finding the population closest, in the sense of minimizing disparity between the sample and theoretical spectral matrices. The classification in this case proceeds by simply choosing the population Π_j that minimizes $I(\widehat{f}; f_j)$, i.e., assigning $\boldsymbol{x}$ to population Π_i whenever

$$I(\widehat{f}; f_i) < I(\widehat{f}; f_j) \tag{7.129}$$

for $j \neq i, j = 1, 2, \ldots, g$.

Kakizawa et al. (1998) proposed using the Chernoff (CH) information measure (Chernoff, 1952, Renyi, 1961), defined as

$$B_\alpha(1;2) = -\ln E_2\left\{\left(\frac{p_2(\boldsymbol{x})}{p_1(\boldsymbol{x})}\right)^\alpha\right\}, \tag{7.130}$$

where the measure is indexed by a regularizing parameter α, for $0 < \alpha < 1$. When $\alpha = .5$, the Chernoff measure is the symmetric divergence proposed by Bhattacharya (1943). For the multivariate normal case,

$$B_\alpha(1;2:\boldsymbol{x}) = \frac{1}{n}\left[\ln \frac{|\alpha\Gamma_1 + (1-\alpha)\Gamma_2|}{|\Gamma_2|} - \alpha \ln \frac{|\Gamma_1|}{|\Gamma_2|}\right]. \tag{7.131}$$

The large sample spectral approximation to the Chernoff information measure is analogous to that for the discrimination information, namely,

$$\begin{aligned} B_\alpha(f_1; f_2) &= \frac{1}{2n} \sum_{0<\omega_k<1/2} \left[\ln \frac{|\alpha f_1(\omega_k) + (1-\alpha) f_2(\omega_k)|}{|f_2(\omega_k)|}\right. \\ &\quad \left. -\alpha \ln \frac{|f_1(\omega_k)|}{|f_2(\omega_k)|}\right]. \end{aligned} \tag{7.132}$$

The Chernoff measure, when divided by $\alpha(1-\alpha)$, behaves like the discrimination information in the limit in the sense that it converges to $I(1;2:\boldsymbol{x})$ for $\alpha \to 0$ and to $I(2;1:\boldsymbol{x})$ for $\alpha \to 1$. Hence, near the boundaries of the parameter α, it tends to behave like discrimination information and for other values

represents a compromise between the two information measures. The classification rule for the Chernoff measure reduces to assigning $\boldsymbol{x}$ to population Π_i whenever

$$B_\alpha(\widehat{f}; f_i) < B_\alpha(\widehat{f}; f_j) \tag{7.133}$$

for $j \neq i, j = 1, 2, \ldots, g$.

Although the classification rules above are well defined if the group spectral matrices are known, this will not be the case in general. If there are g training samples, $\boldsymbol{x}_{ij}, j = 1, \ldots, N_i, i = 1 \ldots, g$, with N_i vector observations available in each group, the natural estimator for the spectral matrix of the group i is just the single-group spectral matrix (7.100), namely, with $\boldsymbol{X}_{ij}(\omega_k)$ denoting the vector DFTs,

$$\widehat{f}_i(\omega_k) = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} \big(\boldsymbol{X}_{ij}(\omega_k) - \boldsymbol{X}_{i\cdot}(\omega_k)\big)^* \big(\boldsymbol{X}_{ij}(\omega_k) - \boldsymbol{X}_{i\cdot}(\omega_k)\big), \tag{7.134}$$

A second consideration is the choice of the regularization parameter α for the Chernoff criterion, (7.132). For the case of $g = 2$ groups, it should be chosen to maximize the disparity between the two group spectra, as defined in (7.132). Kakizawa et al. (1998) simply plot (7.132) as a function of α, using the estimated group spectra in (7.134), choosing the value that gives the maximum disparity between the two groups.

Example 7.12 Discriminant Analysis for Earthquakes and Explosions

The simplest approaches to discriminating between the earthquake and explosion groups have been based on either the relative amplitudes of the P and S phases, as in Figure 7.4 or on relative power components in various frequency bands. Considerable effort has been expended on using various spectral ratios involving the bivariate P and S phases as discrimination features. Kakizawa et al. (1998) mention a number of measures that have be used in the seismological literature as features. These features include ratios of power for the two phases and ratios of power components in high- and low-frequency bands. The use of such features of the spectrum suggests an optimal procedure based on discriminating between the spectral matrices of two stationary processes would be reasonable. The fact that the hypothesis that the spectral matrices were equal, tested in Example 7.10, was also soundly rejected suggests the use of a discriminant function based on spectral differences. Recall the sampling rate is 40 points per second, leading to a folding frequency of 20 Hz. To avoid numerical problems, we used a broad band (2Hz, $L = 51$) and the criteria (7.126) and (7.132), summed over the interval from 0 to 8 Hz, where the spectra were both positive. Narrowing the bandwidth and summing over a broader interval did not substantially change the results.

Table 7.6 Discriminant Scores $I = I(\widehat{f}; f_1) - I(\widehat{f}; f_2)$ and $B = B_{.3}(\widehat{f}; f_1) - B_{.3}(\widehat{f}; f_2)$ for Earthquakes and Explosions

EQ	I	B	EXP	I	B
1	8.51	.54	1	.29	−.25
2	.81	.50	2	−2.55	−.75
3	30.80	1.04	4	−1.82	−.61
4	2.73	.10	4	−1.89	−.44
5	7.69	.11	5	−1.16	−.45
6	21.50	.79	6	−2.12	−.61
7	20.31	.85	7	−2.10	−.59
8	15.54	.70	8	.93	−.21

The maximum value of the estimated Chernoff disparity $B_\alpha(\widehat{f}_1; \widehat{f}_2)$ occurs for $\alpha = .3$, and we use that value in the discriminant criterion (7.132). Discriminant scores using the holdout classification functions are shown in Table 7.6 for both criteria. We note the generally good performance of the Chernoff measure, which separates the two populations well and makes no errors; the discrimination information misclassified explosions one and eight as earthquakes. The values for the two sets of scores are plotted in the right-hand panel of Figure 7.13, and the earthquake variances of the discrimination information have larger variances than do those for the explosions (the standard deviations were 9.34 and 1.25, respectively). The Chernoff discriminant scores are distributed on either side of the decision point 0, with means .58 and -.48 for the earthquake and explosion groups, respectively; the standard deviations of the two samples were .34 and .20. The NZ event was also classified using the average spectral matrices of the eight earthquakes and explosions, giving the value -.49 for the discimination information and -.31 for the Chernoff measure, putting the event in the explosion population by this criterion. Previously, in Example 7.11, the extracted log amplitudes classified this event in the earthquake group. The Russians have asserted no mine blasting or nuclear testing occurred in the area in question, so the event remains as somewhat of a mystery. The fact that it was relatively removed geographically from the test set may also have introduced some uncertainties into the procedure.

Cluster Analysis

For the purpose of clustering, it may be more useful to consider a symmetric disparity measures and we introduce the J-Divergence measure

$$J(f_1; f_2) = I(f_1; f_2) + I(f_2; f_1) \tag{7.135}$$

and the symmetric Chernoff number

$$JB_\alpha(f_1; f_2) = B_\alpha(f_1; f_2) + B_\alpha(f_2; f_1) \tag{7.136}$$

for that purpose. In this case, we define the disparity between the sample spectral matrix of a single vector, $\boldsymbol{x}$, and the population Π_j as

$$J(\widehat{f}; f_j) = I(\widehat{f}; f_j) + I(f_j; \widehat{f}) \tag{7.137}$$

and

$$JB_\alpha(\widehat{f}; f_j) = B_\alpha(\widehat{f}; f_j) + B_\alpha(f_j; \widehat{f}), \tag{7.138}$$

respectively and use these as quasi-distances between the vector and population Π_j.

The measures of disparity can be used to cluster multivariate time series. The symmetric measures of disparity, as defined above ensure that the disparity between f_i and f_j is the same as the disparity between f_j and f_i. Hence, we will consider the symmetric forms (7.137) and (7.138) as quasi-distances for the purpose of defining a distance matrix for input into one of the standard clustering procedures (see Johnson and Wichern, 1992). In general, we may consider either hierarchical or partitioned clustering methods using the quasi-distance matrix as an input.

For purposes of illustration, we may use the symmetric divergence (7.137), which implies the quasi-distance between sample series with estimated spectral matrices $\widehat{f}_i$ and $\widehat{f}_j$ would be (7.137); i.e.,

$$J(\widehat{f}_i; \widehat{f}_j) = \frac{1}{n} \sum_{0<\omega_k<1/2} \Big[\mathrm{tr}\big\{ \widehat{f}_i(\omega_k) \widehat{f}_j^{-1}(\omega_k) \big\} + \mathrm{tr}\big\{ \widehat{f}_j(\omega_k) \widehat{f}_i^{-1}(\omega_k) \big\} - 2p \Big], \tag{7.139}$$

for $i \neq j$. We can also use the comparable form for the Chernoff divergence, but we may not want to make an assumption for the regularization parameter α.

For hierarchical clustering, we begin by clustering the two members of the population that minimize the disparity measure (7.139). Then, these two items form a cluster, and we can compute distances between unclustered items as before. The distance between unnclustered items and a current cluster is defined here as the average of the distances to elements in the cluster. Again, we combine objects that are closest together. We may also compute the distance between the unclustered items and clustered items as the closest distance, rather than the average. Once a series is in a cluster, it stays there. At each stage, we have a fixed number of clusters, depending on the merging stage.

Alternatively, we may think of clustering as a partitioning of the sample into a prespecified number of groups. MacQueen (1967) has proposed this using k-means clustering, using the Mahalonobis distance between an observation and the group mean vectors. At each stage, a reassignment of an observation into its closest affinity group is possible. To see how this procedure applies

Table 7.7 Clustering Results for Earthquakes and Explosions

Beginning	Cluster 1	Cluster 2	Cluster 3
Two Groups:			
Random	EQ 123678	EX 12345678 EQ 45 NZ	
Hierarchical	EQ 12345678	EX 12345678 NZ	
Three Groups:			
Random	EQ 123678	EX 1234567 EQ 4 NZ	EQ 5 EX 8
Hierarchical	EQ 123678	EX 1234567 NZ	EQ 45 EX 8

in the current context, consider a preliminary partition into a fixed number of groups and define the disparity between the spectral matrix of the observation, say, $\widehat{f}$, and the average spectral matrix of the group, say, $\widehat{f}_i$, as $J(\widehat{f}; \widehat{f}_i)$, where the group spectral matrix can be estimated by (7.134). At any pass, a single series is reassigned to the group for which its disparity is minimized. The reassignment procedure is repeated until all observations stay in their current groups. Of course, the number of groups must be specified for each repetition of the partitioning algorithm and a starting partition must be chosen. This assignment can either be random or chosen from a preliminary hierarchical clustering, as described above. kip

Example 7.13 Cluster Analysis for Earthquakes and Explosions

It is instructive to try the clustering procedure on the population of known earthquakes and explosions. Table 7.7 shows the results of applying partitioned clustering under the assumption that either two or three groups are appropriate. Two groups would be simple assuming the vectors classified naturally into the earthquake and explosion classes, whereas three groups would imply possible outliers from the two primary groups. The starting partitions were defined by either randomly assigning observations to groups or using the result of the hierarchichal clustering procedure. The two-group partition with the hierarchical start configuration tends to produce a final partition that agrees closely with the known configuration, assuming the NZ event is an explosion. The random starting partition puts two of the earthquakes into the explosion group. For the three-group partitions, one or two earthquakes and the last explosion join the third cluster that we have designated as the outlying group.

7.8 Principal Components and Factor Analysis

In this section, we introduce the related topics of spectral domain *principal components analysis* and *factor analysis* for time series. The topics of principal components and canonical analysis in the frequency domain are rigorously presented in Brillinger (1981, Chapters 9 and 10) and many of the details concerning these concepts can be found there.

The techniques presented here are related to each other in that they focus on extracting pertinent information from spectral matrices. This information is important because dealing directly with a high-dimensional spectral matrix $f(\omega)$ itself is somewhat cumbersome because it is a function into the set of complex, nonnegative-definite, Hermitian matrices. We can view these techniques as easily understood, parsimonious tools for exploring the behavior of vector-valued time series in the frequency domain with minimal loss of information. Because our focus is on spectral matrices, we assume for convenience that the time series of interest have zero means; the techniques are easily adjusted in the case of nonzero means.

In this and subsequent sections, it will be convenient to work occasionally with complex-valued time series. A $p \times 1$ complex-valued time series can be represented as $\boldsymbol{x}_t = \boldsymbol{x}_{1t} - i\boldsymbol{x}_{2t}$, where $\boldsymbol{x}_{1t}$ is the real part and $\boldsymbol{x}_{2t}$ is the imaginary part of $\boldsymbol{x}_t$. The process is said to be stationary if $E(\boldsymbol{x}_t)$ and $E(\boldsymbol{x}_{t+h}\boldsymbol{x}_t^*)$ exist and are independent of time t. The $p \times p$ autocovariance function,

$$\Gamma_{xx}(h) = E(\boldsymbol{x}_{t+h}\boldsymbol{x}_t^*) - E(\boldsymbol{x}_{t+h})E(\boldsymbol{x}_t^*),$$

of $\boldsymbol{x}_t$ satisfies conditions similar to those of the real-valued case. Writing $\Gamma_{xx}(h) = \{\gamma_{ij}(h)\}$, for $i, j = 1, \ldots, p$, we have (i) $\gamma_{ii}(0) \geq 0$ is real, (ii) $|\gamma_{ij}(h)|^2 \leq \gamma_{ii}(0)\gamma_{jj}(0)$ for all integers h, and (iii) $\Gamma_{xx}(h)$ is Hermitian, that is, $\Gamma_{xx}(h) = \Gamma_{xx}(h)^*$. The spectral theory of complex-valued vector time series is analogous to the real-valued case. For example, $\Gamma_x(h)$ is a nonnegative-definite function on the integers, and if $\sum_h ||\Gamma_{xx}(h)|| < \infty$, the spectral density matrix of the complex series $\boldsymbol{x}_t$ is given by

$$f_{xx}(\omega) = \sum_{h=-\infty}^{\infty} \Gamma_{xx}(h) \exp(-2\pi i h\omega).$$

Principal Components

Classical principal component analysis (PCA) is concerned with explaining the variance–covariance structure among p variables, $\boldsymbol{x} = (x_1, \ldots, x_p)'$, through a few linear combinations of the components of $\boldsymbol{x}$. Suppose we wish to find a linear combination

$$y = \boldsymbol{c}'\boldsymbol{x} = c_1 x_1 + \cdots + c_p x_p \tag{7.140}$$

of the components of $\boldsymbol{x}$ such that $\mathrm{var}(y)$ is as large as possible. Because $\mathrm{var}(y)$ can be increased by simply multiplying $\boldsymbol{c}$ by a constant, it is common to restrict

$\boldsymbol{c}$ to be of unit length; that is, $\boldsymbol{c}'\boldsymbol{c} = 1$. Noting that $\text{var}(y) = \boldsymbol{c}'\Sigma_{xx}\boldsymbol{c}$, where Σ_{xx} is the $p \times p$ variance–covariance matrix of $\boldsymbol{x}$, another way of stating the problem is to find $\boldsymbol{c}$ such that

$$\max_{\boldsymbol{c}\neq\boldsymbol{0}} \frac{\boldsymbol{c}'\Sigma_{xx}\boldsymbol{c}}{\boldsymbol{c}'\boldsymbol{c}}. \tag{7.141}$$

Denote the eigenvalue–eigenvector pairs of Σ_{xx} by $\{(\lambda_1, \boldsymbol{e}_1), \ldots, (\lambda_p, \boldsymbol{e}_p)\}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, and the eigenvectors are of unit length. The solution to (7.141) is to choose $\boldsymbol{c} = \boldsymbol{e}_1$, in which case the linear combination $y_1 = \boldsymbol{e}_1'\boldsymbol{x}$ has maximum variance, $\text{var}(y_1) = \lambda_1$. In other words,

$$\max_{\boldsymbol{c}\neq\boldsymbol{0}} \frac{\boldsymbol{c}'\Sigma_{xx}\boldsymbol{c}}{\boldsymbol{c}'\boldsymbol{c}} = \frac{\boldsymbol{e}_1'\Sigma_{xx}\boldsymbol{e}_1}{\boldsymbol{e}_1'\boldsymbol{e}_1} = \lambda_1. \tag{7.142}$$

The linear combination, $y_1 = \boldsymbol{e}_1'\boldsymbol{x}$, is called the first principal component. Because the eigenvalues of Σ_{xx} are not necessarily unique, the first principal component is not necessarily unique.

The second principal component is defined to be the linear combination $y_2 = \boldsymbol{c}'\boldsymbol{x}$ that maximizes $\text{var}(y_2)$ subject to $\boldsymbol{c}'\boldsymbol{c} = 1$ and such that $\text{cov}(y_1, y_2) = 0$. The solution is to choose $\boldsymbol{c} = \boldsymbol{e}_2$, in which case, $\text{var}(y_2) = \lambda_2$. In general, the k-th principal component, for $k = 1, 2, \ldots, p$, is the the linear combination $y_k = \boldsymbol{c}'\boldsymbol{x}$ that maximizes $\text{var}(y_k)$ subject to $\boldsymbol{c}'c = 1$ and such that $\text{cov}(y_k, y_j) = 0$, for $j = 1, 2, \ldots, k-1$. The solution is to choose $\boldsymbol{c} = \boldsymbol{e}_k$, in which case $\text{var}(y_k) = \lambda_k$.

One measure of the importance of a principal component is to assess the proportion of the total variance attributed to that principal component. The total variance of $\boldsymbol{x}$ is defined to be the sum of the variances of the individual components; that is, $\text{var}(x_1) + \cdots + \text{var}(x_p) = \sigma_{11} + \cdots + \sigma_{pp}$, where σ_{jj} is the j-th diagonal element of Σ_{xx}. This sum is also denoted as $\text{tr}(\Sigma_{xx})$, or the trace of Σ_{xx}. Because $\text{tr}(\Sigma_{xx}) = \lambda_1 + \cdots + \lambda_p$, the proportion of the total variance attributed to the k-th principal component is given simply by $\text{var}(y_k) \,/\, \text{tr}(\Sigma_{xx}) = \lambda_k \,/\, \sum_{j=1}^{p} \lambda_j$.

Given a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the sample principal components are defined as above, but with Σ_{xx} replaced by the sample variance–covariance matrix, $S_{xx} = (n-1)^{-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'$. Further details can be found in the introduction to classical principal component analysis in Johnson and Wichern (1992, Chapter 9).

For the case of time series, suppose we have a zero mean, $p \times 1$, stationary vector process $\boldsymbol{x}_t$ that has a $p \times p$ spectral density matrix given by $f_{xx}(\omega)$. Recall $f_{xx}(\omega)$ is a complex-valued, nonnegative-definite, Hermitian matrix. Using the analogy of classical principal components, and in particular (7.140) and (7.141), suppose, for a fixed value of ω, we want to find a complex-valued univariate process $y_t(\omega) = \boldsymbol{c}(\omega)^* x_t$, where $\boldsymbol{c}(\omega)$ is complex, such that the spectral density of $y_t(\omega)$ is maximized at frequency ω, and $\boldsymbol{c}(\omega)$ is of unit length, $\boldsymbol{c}(\omega)^*\boldsymbol{c}(\omega) = 1$. Because, at frequency ω, the spectral density of $y_t(\omega)$

is $f_y(\omega) = \boldsymbol{c}(\omega)^* f_{xx}(\omega)\boldsymbol{c}(\omega)$, the problem can be restated as: Find complex vector $\boldsymbol{c}(\omega)$ such that

$$\max_{\boldsymbol{c}(\omega)\neq\boldsymbol{0}} \frac{\boldsymbol{c}(\omega)^* f_{xx}(\omega)\boldsymbol{c}(\omega)}{\boldsymbol{c}(\omega)^*\boldsymbol{c}(\omega)}. \tag{7.143}$$

Let $\{(\lambda_1(\omega), \boldsymbol{e}_1(\omega)), \ldots, (\lambda_p(\omega), \boldsymbol{e}_p(\omega))\}$ denote the eigenvalue–eigenvector pairs of $f_{xx}(\omega)$, where $\lambda_1(\omega) \geq \lambda_2(\omega) \geq \cdots \geq \lambda_p(\omega) \geq 0$, and the eigenvectors are of unit length. We note that the eigenvalues of a Hermitian matrix are real. The solution to (7.143) is to choose $\boldsymbol{c}(\omega) = \boldsymbol{e}_1(\omega)$; in which case the desired linear combination is $y_t(\omega) = \boldsymbol{e}_1(\omega)^* \boldsymbol{x}_t$. For this choice,

$$\max_{\boldsymbol{c}(\omega)\neq\boldsymbol{0}} \frac{\boldsymbol{c}(\omega)^* f_{xx}(\omega)\boldsymbol{c}(\omega)}{\boldsymbol{c}(\omega)^*\boldsymbol{c}(\omega)} = \frac{\boldsymbol{e}_1(\omega)^* f_x(\omega)\boldsymbol{e}_1(\omega)}{\boldsymbol{e}_1(\omega)^*\boldsymbol{e}_1(\omega)} = \lambda_1(\omega). \tag{7.144}$$

This process may be repeated for any frequency ω, and the complex-valued process, $y_{t1}(\omega) = \boldsymbol{e}_1(\omega)^*\boldsymbol{x}_t$, is called the first principal component at frequency ω. The k-th principal component at frequency ω, for $k = 1, 2, \ldots, p$, is the complex-valued time series $y_{tk}(\omega) = \boldsymbol{e}_k(\omega)^*\boldsymbol{x}_t$, in analogy to the classical case. In this case, the spectral density of $y_{tk}(\omega)$ at frequency ω is $f_{y_k}(\omega) = \boldsymbol{e}_k(\omega)^* f_{xx}(\omega)\boldsymbol{e}_k(\omega) = \lambda_k(\omega)$.

The previous development of spectral domain principal components is related to the spectral envelope methodology first discussed in Stoffer et al. (1993). We will present the spectral envelope in the next section, where we motivate the use of principal components as it is presented above. Another way to motivate the use of principal components in the frequency domain was given in Brillinger (1981, Chapter 9). Although this technique leads to the same analysis, the motivation may be more satisfactory to the reader at this point. In this case, we suppose we have a stationary, p-dimensional, vector-valued process $\boldsymbol{x}_t$ and we are only able to keep a univariate process y_t such that, when needed, we may reconstruct the vector-valued process, $\boldsymbol{x}_t$, according to an optimality criterion.

Specifically, we suppose we want to approximate a mean-zero, stationary, vector-valued time series, $\boldsymbol{x}_t$, with spectral matrix $f_{xx}(\omega)$, by a univariate process y_t defined by

$$y_t = \sum_{j=-\infty}^{\infty} \boldsymbol{c}_{t-j}^* \boldsymbol{x}_j, \tag{7.145}$$

where $\{\boldsymbol{c}_j\}$ is a $p \times 1$ vector-valued filter, such that $\{\boldsymbol{c}_j\}$ is absolutely summable; that is, $\sum_{j=-\infty}^{\infty} |\boldsymbol{c}_j| < \infty$. The approximation is accomplished so the reconstruction of $\boldsymbol{x}_t$ from y_t, say,

$$\widehat{\boldsymbol{x}}_t = \sum_{j=-\infty}^{\infty} \boldsymbol{b}_{t-j} y_j, \tag{7.146}$$

where $\{\boldsymbol{b}_j\}$ is an absolutely summable $p \times 1$ filter, is such that the mean square approximation error

$$E\{(\boldsymbol{x}_t - \widehat{\boldsymbol{x}}_t)^*(\boldsymbol{x}_t - \widehat{\boldsymbol{x}}_t)\} \tag{7.147}$$

is minimized.

Let $\boldsymbol{b}(\omega)$ and $\boldsymbol{c}(\omega)$ be the transforms of $\{\boldsymbol{b}_j\}$ and $\{\boldsymbol{c}_j\}$, respectively. For example,

$$\boldsymbol{c}(\omega) = \sum_{j=-\infty}^{\infty} \boldsymbol{c}_j \exp(-2\pi i j \omega), \tag{7.148}$$

and, consequently,

$$\boldsymbol{c}_j = \int_{-1/2}^{1/2} \boldsymbol{c}(\omega) \exp(2\pi i j \omega) d\omega. \tag{7.149}$$

Brillinger (1981, Theorem 9.3.1) shows the solution to the problem is to choose $\boldsymbol{c}(\omega)$ to satisfy (7.143) and to set $\boldsymbol{b}(\omega) = \overline{\boldsymbol{c}(\omega)}$. This is precisely the previous problem, with the solution given by (7.144). That is, we choose $\boldsymbol{c}(\omega) = \boldsymbol{e}_1(\omega)$ and $\boldsymbol{b}(\omega) = \overline{\boldsymbol{e}_1(\omega)}$; the filter values can be obtained via the inversion formula given by (7.149). Using these results, in view of (7.145), we may form the first principal component series, say y_{t1}.

This technique may be extended by requesting another series, say, y_{t2}, for approximating $\boldsymbol{x}_t$ with respect to minimum mean square error, but where the coherency between y_{t2} and y_{t1} is zero. In this case, we choose $\boldsymbol{c}(\omega) = \boldsymbol{e}_2(\omega)$. Continuing this way, we can obtain the first $q \leq p$ principal components series, say, $\boldsymbol{y}_t = (y_{t1}, \ldots, y_{tq})'$, having spectral density $f_q(\omega) = \text{diag}\{\lambda_1(\omega), \ldots, \lambda_q(\omega)\}$. The series y_{tk} is the k-th principal component series.

As in the classical case, given observations, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, from the process $\boldsymbol{x}_t$, we can form an estimate $\widehat{f}_{xx}(\omega)$ of $f_{xx}(\omega)$ and define the sample principal component series by replacing $f_{xx}(\omega)$ with $\widehat{f}_{xx}(\omega)$ in the previous discussion. Precise details pertaining to the asymptotic ($n \to \infty$) behavior of the principal component series and their spectra can be found in Brillinger (1981, Chapter 9). To give a basic idea of what we can expect, we focus on the first principal component series and on the spectral estimator obtained by smoothing the periodogram matrix, $I_n(\omega_j)$; that is

$$\widehat{f}_{xx}(\omega_j) = \sum_{\ell=-m}^{m} h_\ell I_n(\omega_j + \ell/n), \tag{7.150}$$

where $L = 2m + 1$ is odd and the weights are chosen so $h_\ell = h_{-\ell}$ are positive and $\sum_\ell h_\ell = 1$. Under the conditions for which $\widehat{f}_{xx}(\omega_j)$ is a well-behaved estimator of $f_{xx}(\omega_j)$, and for which the largest eigenvalue of $f_{xx}(\omega_j)$ is unique,

$$\left\{ \eta_n \left[\widehat{\lambda}_1(\omega_j) - \lambda_1(\omega_j) \right] \Big/ \lambda_1(\omega_j);\ \eta_n \left[\widehat{\boldsymbol{e}}_1(\omega_j) - \boldsymbol{e}_1(\omega_j) \right];\ j = 1, \ldots, J \right\} \tag{7.151}$$

converges ($n \to \infty$) jointly in distribution to independent, zero-mean normal distributions, the first of which is standard normal. In (7.151), $\eta_n^{-2} = \sum_{\ell=-m}^{m} h_\ell^2$, noting we must have $L \to \infty$ and $\eta_n \to \infty$, but $L/n \to 0$ as $n \to \infty$. The asymptotic variance–covariance matrix of $\widehat{\boldsymbol{e}}_1(\omega)$, say, $\Sigma_{e_1}(\omega)$, is

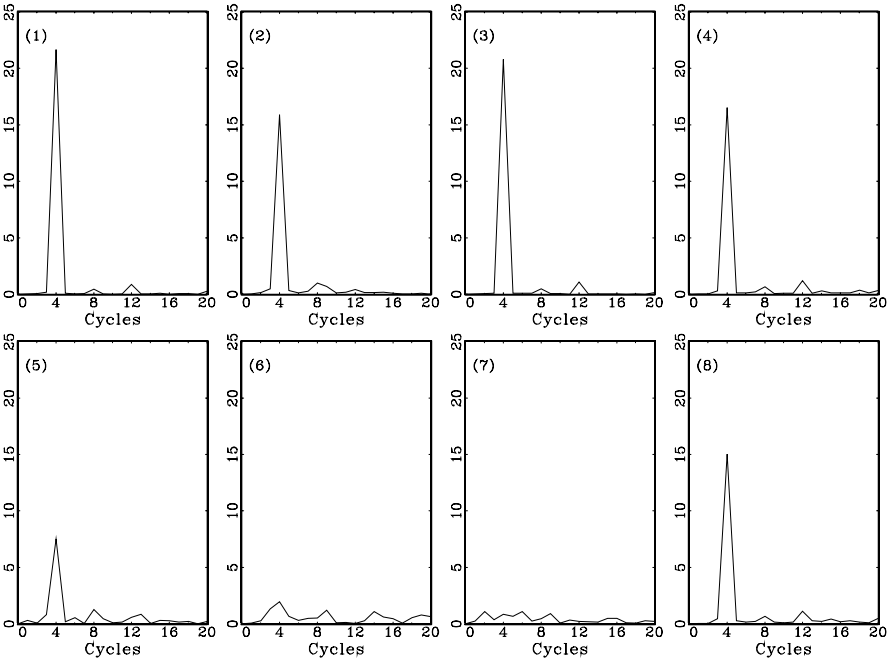


Figure 7.14 The individual periodograms of x_{tk}, for $k = 1, \ldots, 8$, in Example 7.14.

given by

$$\Sigma_{e_1}(\omega) = \eta_n^{-2}\lambda_1(\omega) \sum_{\ell=2}^{p} \lambda_\ell(\omega) \{\lambda_1(\omega) - \lambda_\ell(\omega)\}^{-2} \boldsymbol{e}_\ell(\omega)\boldsymbol{e}_\ell^*(\omega). \tag{7.152}$$

The distribution of $\widehat{\boldsymbol{e}}_1(\omega)$ depends on the other latent roots and vectors of $f_x(\omega)$. Writing $\widehat{\boldsymbol{e}}_1(\omega) = (\widehat{e}_{11}(\omega), \widehat{e}_{12}(\omega), \ldots, \widehat{e}_{1p}(\omega))'$, we may use this result to form confidence regions for the components of $\widehat{\boldsymbol{e}}_1$ by approximating the distribution of

$$\frac{2\,|\widehat{\boldsymbol{e}}_{1,j}(\omega) - \boldsymbol{e}_{1,j}(\omega)|^2}{s_j^2(\omega)}, \tag{7.153}$$

for $j = 1, \ldots, p$, by a χ^2 distribution with two degrees of freedom. In (7.153), $s_j^2(\omega)$ is the j-th diagonal element of $\widehat{\Sigma}_{e_1}(\omega)$, the estimate of $\Sigma_{e_1}(\omega)$. We can use (7.153) to check whether the value of zero is in the confidence region by comparing $2|\widehat{\boldsymbol{e}}_{1,j}(\omega)|^2/s_j^2(\omega)$ with $\chi_2^2(1-\alpha)$, the $1-\alpha$ upper tail cutoff of the χ_2^2 distribution.

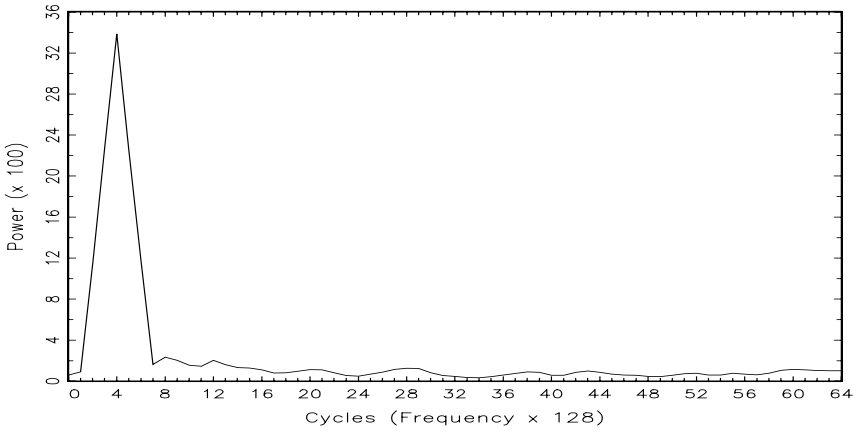


Figure 7.15 The estimated spectral density, $\widehat{\lambda}_1(j/128)$, of the first principal component series in Example 7.14.

Example 7.14 Principal Component Analysis of the fMRI Data

Recall Example 1.6 of Chapter 1, where the vector time series $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{t8})'$, $t = 1, \ldots, 128$, represents consecutive measures of average blood oxygenation level dependent (BOLD) signal intensity, which measures areas of activation in the brain. Recall subjects were given a non-painful brush on the hand and the stimulus was applied for 32 seconds and then stopped for 32 seconds; thus, the signal period is 64 seconds (the sampling rate was one observation every two seconds for 256 seconds). The series x_{tk} for $k = 1, 2, 3, 4$ represent locations in cortex, series x_{t5} and x_{t6} represent locations in the thalamus, and x_{t7} and x_{t8} represent locations in the cerebellum.

As is evident from Figure 1.6 in Chapter 1, different areas of the brain are responding differently, and a principal component analysis may help in indicating which locations are responding with the most spectral power, and which locations do not contribute to the spectral power at the stimulus signal period. In this analysis, we will focus primarily on the signal period of 64 seconds, which translates to four cycles in 256 seconds or $\omega = 4/128$ cycles per time point. In addition, all calculations were performed using the standardized series; that is, we used x_{tk}/s_k, for $k = 1, \ldots, 8$, where s_k is the sample standard deviation of the the k-th series, in the computations.

Figure 7.14 shows individual periodograms of the series x_{tk} for $k = 1, \ldots, 8$. As was evident from Figure 1.6, a strong response to the brush stimulus occurred in areas of the cortex. To estimate the spectral density of $\boldsymbol{x}_t$, we used (7.150) with $L = 5$ and $\{h_0 = 3/9, h_{\pm1} = 2/9, h_{\pm2} = 1/9\}$. Calling the estimated spectrum $\widehat{f}_{xx}(j/128)$, for $j = 0, 1, \ldots, 64$, we can

Table 7.8 Magnitudes of the PC Vector at the Stimulus Frequency in Example 7.14

Location	1	2	3	4	5	6	7	8
$\left\lvert \widehat{\boldsymbol{e}}_1(\frac{4}{128}) \right\rvert$	0.46	0.40	0.45	0.40	0.28	0.15	0.09*	0.39

*The value of zero is in an approximate 99% confidence region for this component.

obtain the estimated spectrum of the first principal component series y_{t1} by calculating the largest eigenvalue, $\widehat{\lambda}_1(j/128)$, of $\widehat{f}_{xx}(j/128)$ for each $j = 0, 1, \ldots, 64$. The result, $\widehat{\lambda}_1(j/128)$, is shown in Figure 7.15. As expected, there is a large peak at the stimulus frequency 4/128, wherein $\widehat{\lambda}_1(4/128) = 0.339$. The total power at the stimulus frequency is $\mathrm{tr}\Big(\widehat{f}_{xx}(4/128)\Big) = 0.353$, so the proportion of the power at frequency 4/128 attributed to the first principal component series is 0.339/0.353 = 96%. Because the first principal component explains nearly all of the total power at the stimulus frequency, there is no need to explore the other principal component series at this frequency.

The estimated first principal component series at frequency 4/128 is given by $\widehat{y}_{t1}(4/128) = \widehat{\boldsymbol{e}}_1^*(4/128)\boldsymbol{x}_t$, and the components of $\widehat{\boldsymbol{e}}_1(4/128)$ can give insight as to which locations of the brain are responding to the brush stimulus. Table 7.8 shows the magnitudes of $\widehat{\boldsymbol{e}}_1(4/128)$. In addition, an approximate 99% confidence interval was obtained for each component using (7.153). As expected, the analysis indicates that location 7 is not contributing to the power at this frequency, but surprisingly, the analysis suggests location 6 is responding to the stimulus.

Factor Analysis

Classical factor analysis is similar to classical principal component analysis. Suppose $\boldsymbol{x}$ is a mean-zero, $p\times 1$, random vector with variance–covariance matrix Σ_{xx}. The factor model proposes that $\boldsymbol{x}$ is dependent on a few unobserved common factors, $z_1, \ldots, z_q$, plus error. In this model, one hopes that q will be much smaller than p. The factor model is given by

$$\boldsymbol{x} = \mathcal{B}\boldsymbol{z} + \boldsymbol{\epsilon}, \tag{7.154}$$

where $\mathcal{B}$ is a $p \times q$ matrix of factor loadings, $\boldsymbol{z} = (z_1, \ldots, z_q)'$ is a random $q \times 1$ vector of factors such that $E(\boldsymbol{z}) = \mathbf{0}$ and $E(\boldsymbol{z}\boldsymbol{z}') = I_q$, the $q \times q$ identity matrix. The $p \times 1$ unobserved error vector $\boldsymbol{\epsilon}$ is assumed to be independent of the factors, with zero mean and diagonal variance-covariance matrix $D =$

$\text{diag}\{\delta_1^2, \ldots, \delta_p^2\}$. Note, (7.154) differs from the multivariate regression model in §5.7 because the factors, $\boldsymbol{z}$, are unobserved. Equivalently, the factor model, (7.154), can be written in terms of the covariance structure of $\boldsymbol{x}$,

$$\Sigma_{xx} = \mathcal{B}\mathcal{B}' + D; \tag{7.155}$$

i.e., the variance-covariance matrix of $\boldsymbol{x}$ is the sum of a symmetric, nonnegative-definite rank $q \leq p$ matrix and a nonnegative-definite diagonal matrix. If $q = p$, then Σ_{xx} can be reproduced exactly as $\mathcal{B}\mathcal{B}'$, using the fact that $\Sigma_{xx} = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1' + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p'$, where $(\lambda_i, \boldsymbol{e}_i)$ are the eigenvalue–eigenvector pairs of Σ_{xx}. As previously indicated, however, we hope q will be much smaller than p. Unfortunately, most covariance matrices cannot be factored as (7.155) when q is much smaller than p.

To motivate factor analysis, suppose the components of $\boldsymbol{x}$ can be grouped into meaningful groups. Within each group, the components are highly correlated, but the correlation between variables that are not in the same group is small. A group is supposedly formed by a single construct, represented as an unobservable factor, responsible for the high correlations within a group. For example, a person competing in a decathlon performs $p = 10$ athletic events, and we may represent the outcome of the decathlon as a 10×1 vector of scores. The events in a decathlon involve running, jumping, or throwing, and it is conceivable the 10×1 vector of scores might be able to be factored into $q = 4$ factors, (1) arm strength, (2) leg strength, (3) running speed, and (4) running endurance. The model (7.154) specifies that $\text{cov}(\boldsymbol{x}, \boldsymbol{z}) = \mathcal{B}$, or $\text{cov}(x_i, z_j) = b_{ij}$ where b_{ij} is the ij-th component of the factor loading matrix $\mathcal{B}$, for $i = 1, \ldots, p$ and $j = 1, \ldots, q$. Thus, the elements of $\mathcal{B}$ are used to identify which hypothetical factors the components of $\boldsymbol{x}$ belong to, or load on.

At this point, some ambiguity is still associated with the factor model. Let Q be a $q \times q$ orthogonal matrix; that is $Q'Q = QQ' = I_q$. Let $\mathcal{B}_* = \mathcal{B}Q$ and $\boldsymbol{z}_* = Q'\boldsymbol{z}$ so (7.154) can be written as

$$\boldsymbol{x} = \mathcal{B}\boldsymbol{z} + \boldsymbol{\epsilon} = \mathcal{B}QQ'\boldsymbol{z} + \boldsymbol{\epsilon} = \mathcal{B}_*\boldsymbol{z}_* + \boldsymbol{\epsilon}. \tag{7.156}$$

The model in terms of $\mathcal{B}_*$ and $\boldsymbol{z}_*$ fulfills all of the factor model requirements, for example, $\text{cov}(\boldsymbol{z}_*) = Q'\text{cov}(\boldsymbol{z})Q = QQ' = I_q$, so

$$\Sigma_{xx} = \mathcal{B}_*\text{cov}(\boldsymbol{z}_*)\mathcal{B}_*' + D = \mathcal{B}QQ'\mathcal{B}' + D = \mathcal{B}\mathcal{B}' + D. \tag{7.157}$$

Hence, on the basis of observations on $\boldsymbol{x}$, we cannot distinguish between the loadings $\mathcal{B}$ and the rotated loadings $\mathcal{B}_* = \mathcal{B}Q$. Typically, Q is chosen so the matrix $\mathcal{B}$ is easy to interpret, and this is the basis of what is called factor rotation.

Given a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, a number of methods are used to estimate the parameters of the factor model, and we discuss two of them here. The first method is the principal component method. Let S_{xx} denote the sample variance–covariance matrix, and let $(\widehat{\lambda}_i, \widehat{\boldsymbol{e}}_i)$ be the eigenvalue–eigenvector pairs

of S_{xx}. The $p \times q$ matrix of estimated factor loadings is found by setting

$$\widehat{\mathcal{B}} = \left[\widehat{\lambda}_1^{1/2}\,\widehat{\boldsymbol{e}}_1 \;\middle|\; \widehat{\lambda}_2^{1/2}\,\widehat{\boldsymbol{e}}_2 \;\middle|\; \cdots \;\middle|\; \widehat{\lambda}_q^{1/2}\,\widehat{\boldsymbol{e}}_q\right]. \tag{7.158}$$

The argument here is that if q factors exist, then

$$S_{xx} \approx \widehat{\lambda}_1\widehat{\boldsymbol{e}}_1\widehat{\boldsymbol{e}}_1' + \cdots + \widehat{\lambda}_q\widehat{\boldsymbol{e}}_q\widehat{\boldsymbol{e}}_q' = \widehat{\mathcal{B}}\widehat{\mathcal{B}}', \tag{7.159}$$

because the remaining eigenvalues, $\widehat{\lambda}_{q+1}, \ldots, \widehat{\lambda}_p$, will be negligible. The estimated diagonal matrix of error variances is then obtained by setting $\widehat{D} = \mathrm{diag}\{\widehat{\delta}_1^2, \ldots, \widehat{\delta}_p^2\}$, where $\widehat{\delta}_j^2$ is the j-th diagonal element of $S_{xx} - \widehat{\mathcal{B}}\widehat{\mathcal{B}}'$.

The second method, which can give answers that are considerably different from the principal component method is maximum likelihood. Upon further assumption that in (7.154), $\boldsymbol{z}$ and ϵ are multivariate normal, the log likelihood of $\mathcal{B}$ and D ignoring a constant is

$$-2\ln L(\mathcal{B}, D) = n\ln|\Sigma_{xx}| + \sum_{j=1}^{n} \boldsymbol{x}_j'\Sigma_{xx}^{-1}\boldsymbol{x}_j. \tag{7.160}$$

The likelihood depends on $\mathcal{B}$ and D through (7.155), $\Sigma_{xx} = \mathcal{B}\mathcal{B}' + D$. As discussed in (7.156)-(7.157), the likelihood is not well defined because $\mathcal{B}$ can be rotated. Typically, restricting $\mathcal{B}D^{-1}\mathcal{B}'$ to be a diagonal matrix is a computationally convenient uniqueness condition. The actual maximization of the likelihood is accomplished using numerical methods.

One obvious method of performing maximum likelihood for the Gaussian factor model is the EM algorithm. For example, suppose the factor vector $\boldsymbol{z}$ is known. Then, the factor model is simply the multivariate regression model given in §5.7, that is, write $X' = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ and $Z' = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n]$, and note that X is $n \times p$ and Z is $n \times q$. Then, the MLE of $\mathcal{B}$ is

$$\widehat{\mathcal{B}} = X'Z(Z'Z)^{-1} = \left(n^{-1}\sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{z}_j'\right)\left(n^{-1}\sum_{j=1}^{n}\boldsymbol{z}_j\boldsymbol{z}_j'\right)^{-1} \stackrel{\text{def}}{=} C_{xz}C_{zz}^{-1}, \tag{7.161}$$

and the MLE of D is

$$\widehat{D} = \mathrm{Diag}\left\{n^{-1}\sum_{j=1}^{n}\left(\boldsymbol{x}_j - \widehat{\mathcal{B}}\boldsymbol{z}_j\right)\left(\boldsymbol{x}_j - \widehat{\mathcal{B}}\boldsymbol{z}_j\right)'\right\}; \tag{7.162}$$

that is, only the diagonal elements of the right-hand side of (7.162) are used. The bracketed quantity in (7.162) reduces to

$$C_{xx} - C_{xz}C_{zz}^{-1}C_{xz}', \tag{7.163}$$

where $C_{xx} = n^{-1}\sum_{j=1}^{n}\boldsymbol{x}_j\boldsymbol{x}_j'$.

Based on the derivation of the EM algorithm for the state-space model, (4.66)-(4.75), we conclude that, to employ the EM algorithm here, given the current parameter estimates, in C_{xz}, we replace $\boldsymbol{x}_j\boldsymbol{z}_j'$ by $\boldsymbol{x}_j\widetilde{\boldsymbol{z}}_j'$, where $\widetilde{\boldsymbol{z}}_j = E(\boldsymbol{z}_j \mid \boldsymbol{x}_j)$, and in C_{zz}, we replace $\boldsymbol{z}_j\boldsymbol{z}_j'$ by $P_z + \widetilde{\boldsymbol{z}}_j\widetilde{\boldsymbol{z}}_j'$, where $P_z = \text{var}(\boldsymbol{z}_j \mid \boldsymbol{x}_j)$. Using the fact that the $(p+q)\times 1$ vector $(\boldsymbol{x}_j', \boldsymbol{z}_j')'$ is multivariate normal with mean-zero, and variance–covariance matrix given by

$$\begin{pmatrix} \mathcal{B}\mathcal{B}' + D & \mathcal{B} \\ \mathcal{B}' & I_q \end{pmatrix}, \tag{7.164}$$

we have

$$\widetilde{\boldsymbol{z}}_j \equiv E(\boldsymbol{z}_j \mid \boldsymbol{x}_j) = \mathcal{B}'(\mathcal{B}'\mathcal{B} + D)^{-1}\boldsymbol{x}_j \tag{7.165}$$

and

$$P_z \equiv \text{var}(\boldsymbol{z}_j \mid \boldsymbol{x}_j) = I_q - \mathcal{B}'(\mathcal{B}'\mathcal{B} + D)^{-1}\mathcal{B}. \tag{7.166}$$

For time series, suppose $\boldsymbol{x}_t$ is a stationary $p\times 1$ process with $p\times p$ spectral matrix $f_{xx}(\omega)$. Analogous to the classical model displayed in (7.155), we may postulate that at a given frequency of interest, ω, the spectral matrix of $\boldsymbol{x}_t$ satisfies

$$f_{xx}(\omega) = \mathcal{B}(\omega)\mathcal{B}(\omega)^* + D(\omega), \tag{7.167}$$

where $\mathcal{B}(\omega)$ is a complex-valued $p\times q$ matrix with $\text{rank}\big(\mathcal{B}(\omega)\big) = q \le p$ and $D(\omega)$ is a real, nonnegative-definite, diagonal matrix. Typically, we expect q will be much smaller than p.

As an example of a model that gives rise to (7.167), let $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})'$, and suppose

$$x_{tj} = c_j s_{t-\tau_j} + \epsilon_{tj}, \quad j = 1, \ldots, p, \tag{7.168}$$

where $c_j \ge 0$ are individual amplitudes and s_t is a common unobserved signal (factor) with spectral density $f_{ss}(\omega)$. The values τ_j are the individual phase shifts. Assume s_t is independent of $\boldsymbol{\epsilon}_t = (\epsilon_{t1}, \ldots, \epsilon_{tp})'$ and the spectral matrix of $\boldsymbol{\epsilon}_t$, $D_{\epsilon\epsilon}(\omega)$, is diagonal. The DFT of x_{tj} is given by

$$X_j(\omega) = n^{-1/2}\sum_{t=1}^{n} x_{tj}\exp(-2\pi i t\omega)$$

and, in terms of the model (7.168),

$$X_j(\omega) = a_j(\omega)X_s(\omega) + X_{\epsilon_j}(\omega), \tag{7.169}$$

where $a_j(\omega) = c_j\exp(-2\pi i\tau_j\omega)$, and $X_s(\omega)$ and $X_{\epsilon_j}(\omega)$ are the respective DFTs of the signal s_t and the noise ϵ_{tj}. Stacking the individual elements of (7.169), we obtain a complex version of the classical factor model with one factor,

$$\begin{pmatrix} X_1(\omega) \\ \vdots \\ X_p(\omega) \end{pmatrix} = \begin{pmatrix} a_1(\omega) \\ \vdots \\ a_p(\omega) \end{pmatrix} X_s(\omega) + \begin{pmatrix} X_{\epsilon_1}(\omega) \\ \vdots \\ X_{\epsilon_p}(\omega) \end{pmatrix},$$

or more succinctly,

$$\boldsymbol{X}(\omega) = \boldsymbol{a}(\omega)X_s(\omega) + \boldsymbol{X}_\epsilon(\omega). \tag{7.170}$$

From (7.170), we can identify the spectral components of the model; that is,

$$f_{xx}(\omega) = \boldsymbol{b}(\omega)\boldsymbol{b}(\omega)^* + D_{\epsilon\epsilon}(\omega), \tag{7.171}$$

where $\boldsymbol{b}(\omega)$ is a $p \times 1$ complex-valued vector, $\boldsymbol{b}(\omega)\boldsymbol{b}(\omega)^* = \boldsymbol{a}(\omega)f_{ss}(\omega)\boldsymbol{a}(\omega)^*$. Model (7.171) could be considered the one-factor model for time series. This model can be extended to more than one factor by adding other independent signals into the original model (7.168). More details regarding this and related models can be found in Stoffer (1999).

Example 7.15 Single Factor Analysis of the fMRI Data

The fMRI data analyzed in Example 7.14 is well suited for a single factor analysis using the model (7.168), or, equivalently, the complex-valued, single factor model (7.170). In terms of (7.168), we can think of the signal s_t as representing the brush stimulus signal. As before, the frequency of interest is $\omega = 4/128$, which corresponds to a period of 32 time points, or 64 seconds.

A simple way to estimate the components $\boldsymbol{b}(\omega)$ and $D_{\epsilon\epsilon}(\omega)$, as specified in (7.171), is to use the principal components method. Let $\widehat{f}_{xx}(\omega)$ denote the estimate of the spectral density of $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{t8})'$ obtained in Example 7.14. Then, analogous to (7.158) and (7.159), we set

$$\widehat{\boldsymbol{b}}(\omega) = \sqrt{\widehat{\lambda}_1(\omega)}\, \widehat{\boldsymbol{e}}_1(\omega),$$

where $\left(\widehat{\lambda}_1(\omega), \widehat{\boldsymbol{e}}_1(\omega)\right)$ is the first eigenvalue–eigenvector pair of $\widehat{f}_{xx}(\omega)$. The diagonal elements of $\widehat{D}_{\epsilon\epsilon}(\omega)$ are obtained from the diagonal elements of $\widehat{f}_{xx}(\omega) - \widehat{\boldsymbol{b}}(\omega)\widehat{\boldsymbol{b}}(\omega)^*$. The appropriateness of the model can be assessed by checking the elements of the residual matrix, $\widehat{f}_{xx}(\omega) - [\widehat{\boldsymbol{b}}(\omega)\widehat{\boldsymbol{b}}(\omega)^* + \widehat{D}_{\epsilon\epsilon}(\omega)]$, are negligible in magnitude.

Concentrating on the stimulus frequency, recall $\widehat{\lambda}_1(4/128) = 0.339$. The magnitudes of $\widehat{\boldsymbol{e}}_1(4/128)$ are displayed in Table 7.8, indicating all locations load on the stimulus factor except for location 7, and location 6 could be considered borderline. The diagonal elements of $\widehat{f}_{xx}(\omega) - \widehat{\boldsymbol{b}}(\omega)\widehat{\boldsymbol{b}}(\omega)^*$ yield

$$\widehat{D}_{\epsilon\epsilon}(4/128) = 0.001 \times \mathrm{diag}\{0.27, 1.06, 0.45, 1.26, 1.64, 4.22, 4.38, 1.08\}.$$

The magnitudes of the elements of residual matrix at $\omega = 4/128$ are

$$0.001 \times \begin{pmatrix} 0.00 & 0.19 & 0.14 & 0.19 & 0.49 & 0.49 & 0.65 & 0.46 \\ 0.19 & 0.00 & 0.49 & 0.86 & 0.71 & 1.11 & 1.80 & 0.58 \\ 0.14 & 0.49 & 0.00 & 0.62 & 0.67 & 0.65 & 0.39 & 0.22 \\ 0.19 & 0.86 & 0.62 & 0.00 & 1.02 & 1.33 & 1.16 & 0.14 \\ 0.49 & 0.71 & 0.67 & 1.02 & 0.00 & 0.85 & 1.11 & 0.57 \\ 0.49 & 1.11 & 0.65 & 1.33 & 0.85 & 0.00 & 1.81 & 1.36 \\ 0.65 & 1.80 & 0.39 & 1.16 & 1.11 & 1.81 & 0.00 & 1.57 \\ 0.46 & 0.58 & 0.22 & 0.14 & 0.57 & 1.36 & 1.57 & 0.00 \end{pmatrix},$$

indicating the model fit is good.

A number of authors have considered factor analysis in the spectral domain, for example Priestley et al. (1974); Priestley and Subba Rao (1975); Geweke (1977), and Geweke and Singleton (1981), to mention a few. An obvious extension of simple model (7.168) is the factor model

$$\boldsymbol{x}_t = \sum_{j=-\infty}^{\infty} \Lambda_j \boldsymbol{s}_{t-j} + \boldsymbol{\epsilon}_t, \tag{7.172}$$

where $\{\Lambda_j\}$ is a real-valued $p \times q$ filter, $\boldsymbol{s}_t$ is a $q \times 1$ stationary, unobserved signal, with independent components, and $\boldsymbol{\epsilon}_t$ is white noise. We assume the signal and noise process are independent, $\boldsymbol{s}_t$ has $q \times q$ real, diagonal spectral matrix $f_{ss}(\omega) = \text{diag}\{f_{s1}(\omega), \ldots, f_{sq}(\omega)\}$, and $\boldsymbol{\epsilon}_t$ has a real, diagonal, $p \times p$ spectral matrix given by $D_{\epsilon\epsilon}(\omega) = \text{diag}\{f_{\epsilon 1}(\omega), \ldots, f_{\epsilon p}(\omega)\}$. If, in addition, $\sum ||\Lambda_j|| < \infty$, the spectral matrix of $\boldsymbol{x}_t$ can be written as

$$f_{xx}(\omega) = \Lambda(\omega) f_{ss}(\omega) \Lambda(\omega)^* + D_{\epsilon\epsilon}(\omega) = \mathcal{B}(\omega)\mathcal{B}(\omega)^* + D_{\epsilon\epsilon}(\omega), \tag{7.173}$$

where

$$\Lambda(\omega) = \sum_{t=-\infty}^{\infty} \Lambda_t \exp(-2\pi i t \omega) \tag{7.174}$$

and $\mathcal{B}(\omega) = \Lambda(\omega) f_{ss}^{1/2}(\omega)$. Thus, by (7.173), the model (7.172) is seen to satisfy the basic requirement of the spectral domain factor analysis model; that is, the $p \times p$ spectral density matrix of the process of interest, $f_{xx}(\omega)$, is the sum of a rank $q \le p$ matrix, $\mathcal{B}(\omega)\mathcal{B}(\omega)^*$, and a real, diagonal matrix, $D_{\epsilon\epsilon}(\omega)$. For the purpose of identifiability we set $f_{ss}(\omega) = I_q$ for all ω; in which case, $\mathcal{B}(\omega) = \Lambda(\omega)$. As in the classical case [see (7.157)], the model is specified only up to rotations; for details, see Bloomfield and Davis (1994).

Parameter estimation for the model (7.172), or equivalently (7.173), can be accomplished using the principal component method. Let $\widehat{f}_{xx}(\omega)$ be an estimate of $f_{xx}(\omega)$, and let $\big(\widehat{\lambda}_j(\omega), \widehat{\boldsymbol{e}}_j(\omega)\big)$, for $j = 1, \ldots, p$, be the eigenvalue–eigenvector pairs, in the usual order, of $\widehat{f}_{xx}(\omega)$. Then, as in the classical case, the $p \times q$ matrix $\mathcal{B}$ is estimated by

$$\widehat{\mathcal{B}}(\omega) = \Big[\widehat{\lambda}_1(\omega)^{1/2}\, \widehat{\boldsymbol{e}}_1(\omega) \,\Big|\, \widehat{\lambda}_2(\omega)^{1/2}\, \widehat{\boldsymbol{e}}_2(\omega) \,\Big|\, \cdots \,\Big|\, \widehat{\lambda}_q(\omega)^{1/2}\, \widehat{\boldsymbol{e}}_q(\omega)\Big]. \tag{7.175}$$

The estimated diagonal spectral density matrix of errors is then obtained by setting $\widehat{D}_{\epsilon\epsilon}(\omega) = \text{diag}\{\widehat{f}_{\epsilon 1}(\omega), \ldots, \widehat{f}_{\epsilon p}(\omega)\}$, where $\widehat{f}_{\epsilon j}(\omega)$ is the j-th diagonal element of $\widehat{f}_{xx}(\omega) - \widehat{\mathcal{B}}(\omega)\widehat{\mathcal{B}}(\omega)^*$.

Alternatively, we can estimate the parameters by approximate likelihood methods. As in (7.170), let $\boldsymbol{X}(\omega_j)$ denote the DFT of the data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ at frequency $\omega_j = j/n$. Similarly, let $\boldsymbol{X}_s(\omega_j)$ and $\boldsymbol{X}_\epsilon(\omega_j)$ be the DFTs of the signal and of the noise processes, respectively. Then, under certain conditions (see Pawitan and Shumway, 1989), for $\ell = 0, \pm 1, \ldots, \pm m$,

$$\boldsymbol{X}(\omega_j + \ell/n) = \Lambda(\omega_j)\boldsymbol{X}_s(\omega_j + \ell/n) + \boldsymbol{X}_\epsilon(\omega_j + \ell/n) + \text{o}_{as}(n^{-\alpha}), \qquad (7.176)$$

where $\Lambda(\omega_j)$ is given by (7.174) and $\text{o}_{as}(n^{-\alpha}) \to 0$ almost surely for some $0 \le \alpha < 1/2$ as $n \to \infty$. In (7.176), the $\boldsymbol{X}(\omega_j + \ell/n)$ are the DFTs of the data at the L odd frequencies $\{\omega_j + \ell/n;\ \ell = 0, \pm 1, \ldots, \pm m\}$ surrounding the central frequency of interest $\omega_j = j/n$.

Under appropriate conditions $\{\boldsymbol{X}(\omega_j + \ell/n);\ \ell = 0, \pm 1, \ldots, \pm m\}$ in (7.176) are approximately ($n \to \infty$) independent, complex Gaussian random vectors with variance–covariance matrix $f_{xx}(\omega_j)$. The approximate likelihood is given by

$$\begin{aligned} &-2 \ln L\big(\mathcal{B}(\omega_j), D_{\epsilon\epsilon}(\omega_j)\big) \\ &\quad = n \ln\big|f_{xx}(\omega_j)\big| + \sum_{\ell=-m}^{m} \boldsymbol{X}^*(\omega_j + \ell/n) f_{xx}^{-1}(\omega_j) \boldsymbol{X}(\omega_j + \ell/n), \end{aligned} \qquad (7.177)$$

with the constraint $f_{xx}(\omega_j) = \mathcal{B}(\omega_j)\mathcal{B}(\omega_j)^* + D_{\epsilon\epsilon}(\omega_j)$. As in the classical case, we can use various numerical methods to maximize $L\big(\mathcal{B}(\omega_j), D_{\epsilon\epsilon}(\omega_j)\big)$ at every frequency, ω_j, of interest. For example, the EM algorithm discussed for the classical case, (7.161)-(7.166), can easily be extended to this case.

Assuming $f_{ss}(\omega) = I_q$, the estimate of $\mathcal{B}(\omega_j)$ is also the estimate of $\Lambda(\omega_j)$. Calling this estimate $\widehat{\Lambda}(\omega_j)$, the time domain filter can be estimated by

$$\widehat{\Lambda}_t^M = M^{-1} \sum_{j=0}^{M-1} \widehat{\Lambda}(\omega_j) \exp(2\pi i j t/n), \qquad (7.178)$$

for some $0 < M \le n$, which is the discrete and finite version of the inversion formula given by

$$\Lambda_t = \int_{-1/2}^{1/2} \Lambda(\omega) \exp(2\pi i \omega t) d\omega. \qquad (7.179)$$

Note that we have used this approximation earlier in Chapter 4, (4.135), for estimating the time response of a frequency response function defined over a finite number of frequencies.

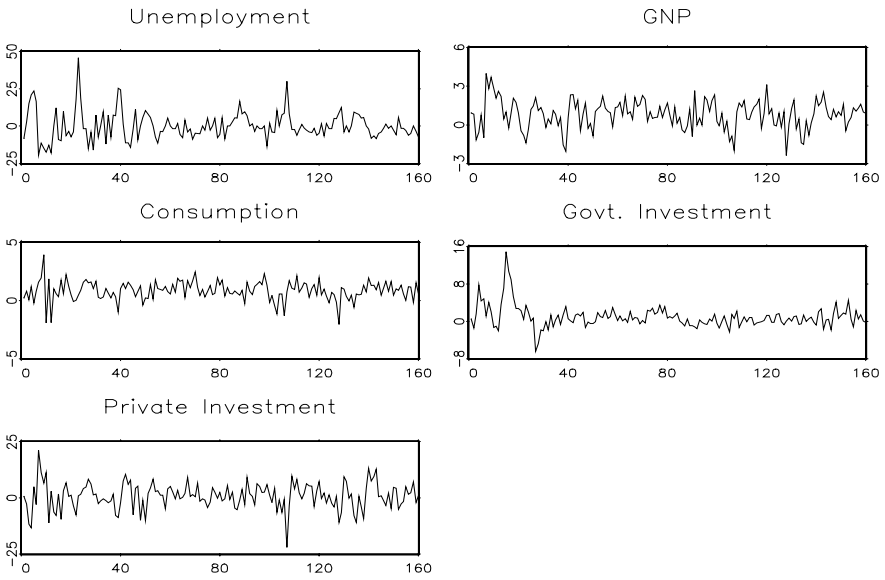


Figure 7.16 The seasonally adjusted, quarterly growth rate (as percentages) of five macroeconomic series, Unemployment, GNP, Consumption, Government Investment, and Private Investment in the U.S. between 1948 and 1988, $n = 160$ values.

Example 7.16 Government Spending, Private Investment, and Unemployment in the U.S.

Figure 7.16 shows the seasonally adjusted, quarterly growth rate (as percentages) of five macroeconomic series, Unemployment, GNP, Consumption, Government Investment, and Private Investment in the U.S. between 1948 and 1988, $n = 160$ values. These data are analyzed in the time domain by Young and Pedregal (1998), who were investigating how government spending and private capital investment influenced the rate of unemployment.

Spectral estimation was performed on the detrended, standardized, and tapered (using a cosine bell) growth rate values. Then, as described in (7.150), a set of $L = 11$ triangular weights, $\{h_0 = 6/36, h_{\pm 1} = 5/36, h_{\pm 2} = 4/36, h_{\pm 3} = 3/36, h_{\pm 4} = 2/36, h_{\pm 5} = 1/36\}$, were used to smooth in 5×5 periodogram matrices. Figure 7.17 shows the individual estimated spectra (scaled by 1000) of each series in terms of the number of cycles. We focus on three interesting frequencies. First, we note the lack of spectral power near 40 cycles ($\omega = 40/160 = 1/4$; one cycle every four quarters, or one year), indicating the data have been seasonally adjusted. In addition, because of the seasonal adjustment, some

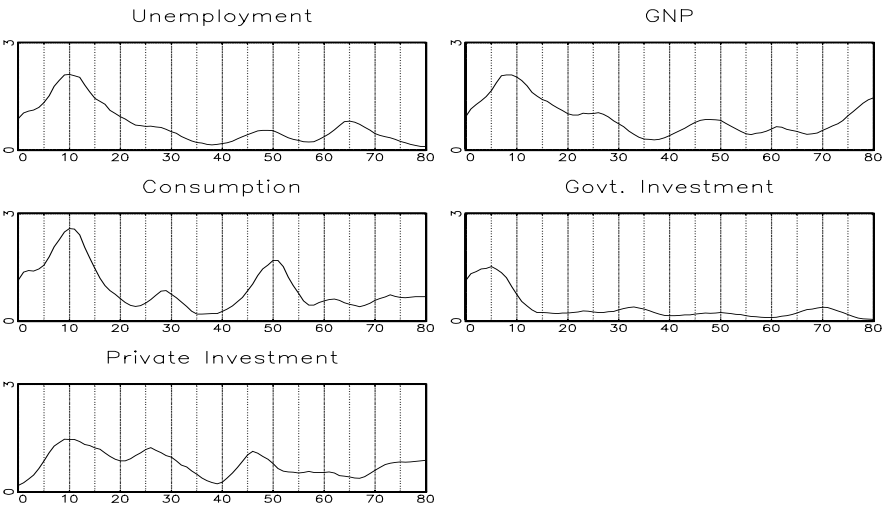


Figure 7.17 The individual estimated spectra (scaled by 1000) of each series show in Figure 7.16 in terms of the number of cycles in 160 quarters.

spectral power appears near the seasonal frequency; this is a distortion apparently caused by the method of seasonally adjusting the data. Next, we note spectral power appears near 10 cycles ($\omega = 10/160 = 1/16$; one cycle every four years) in Unemployment, GNP, Consumption, and, to lesser degree, in Private Investment. Finally, spectral power appears near five cycles ($\omega = 5/160 = 1/32$; one cycle every 8 years) in Government Investment, and perhaps to lesser degrees in Unemployment, GNP, and Consumption.

Figure 7.18 shows the coherences among various series. At the frequencies of interest ($\omega = 5/160$ and $10/160$), pairwise, GNP, Unemployment, Consumption, and Private Investment (except for Unemployment and Private Investment) are coherent. Government Investment is either not coherent or minimally coherent with the other series.

Figure 7.19 shows $\widehat{\lambda}_1(\omega)$ and $\widehat{\lambda}_2(\omega)$, the first and second eigenvalues of the estimated spectral matrix $\widehat{f}_{xx}(\omega)$. These eigenvalues suggest the first factor is identified by the frequency of one cycle every four years, whereas the second factor is identified by the frequency of one cycle every eight years. The modulus of the corresponding eigenvectors at the frequencies of interest, $\widehat{\boldsymbol{e}}_1(10/160)$ and $\widehat{\boldsymbol{e}}_2(5/160)$, are shown in Table 7.9. These values confirm Unemployment, GNP, Consumption, and Private Investment load on the first factor, and Government Investment loads on the second factor. The remainder of the details involving the factor analysis of these data is left as an exercise.

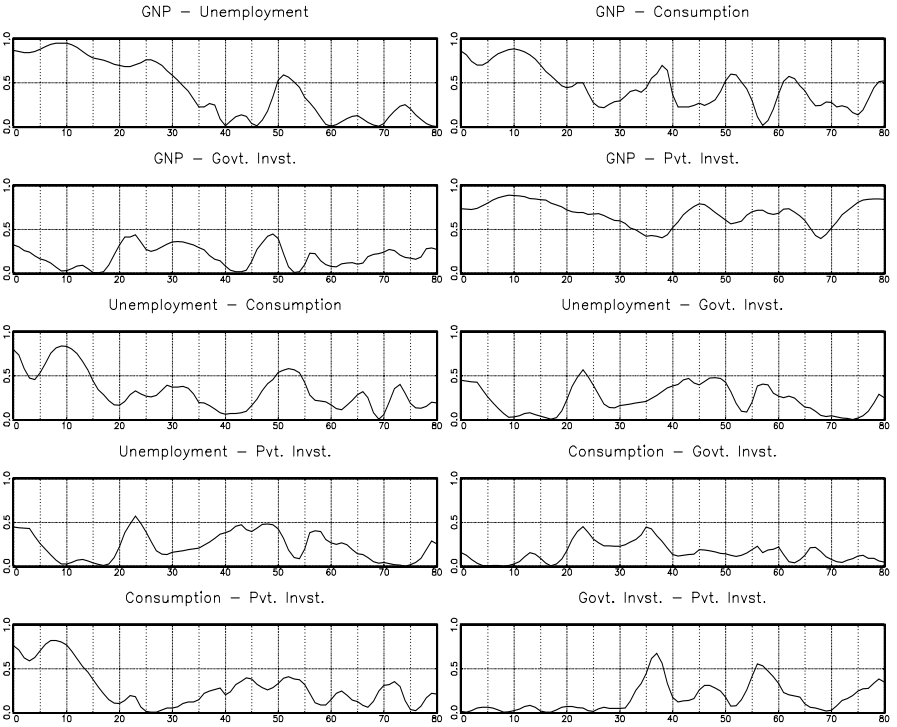


Figure 7.18 The squared coherencies between the various series displayed in Figure 7.16 in terms of the number of cycles in 160 quarters.

Table 7.9 Magnitudes of the Eigenvectors in Example 7.16

Series	Unemp	GNP	Cons	G. Inv.	P. Inv.
$\lvert \widehat{\boldsymbol{e}}_1(\frac{10}{160}) \rvert$	0.51	0.51	0.57	0.05	0.41
$\lvert \widehat{\boldsymbol{e}}_2(\frac{5}{160}) \rvert$	0.17	0.03	0.39	0.87	0.27

7.9 The Spectral Envelope

The concept of spectral envelope for the spectral analysis and scaling of categorical time series was first introduced in Stoffer et al. (1993). Since then, the idea has been extended in various directions (not only restricted to categorical time series), and we will explore these problems as well. First, we give a brief

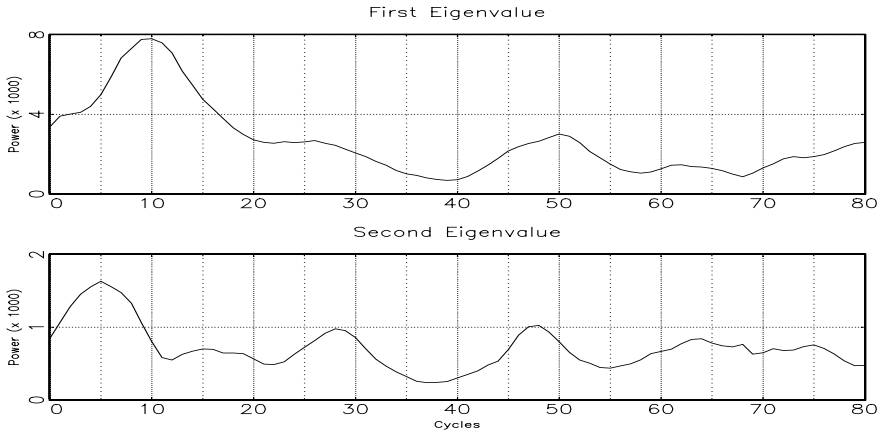


Figure 7.19 The first, $\widehat{\lambda}_1(\omega)$, and second, $\widehat{\lambda}_2(\omega)$, eigenvalues (scaled by 1000) of the estimated spectral matrix $\widehat{f}_{xx}(\omega)$ in terms of the number of cycles in 160 quarters.

introduction to the concept of scaling time series.

The spectral envelope was motivated by collaborations with researchers who collected categorical-valued time series with an interest in the cyclic behavior of the data. For example, Table 7.10 shows the per-minute sleep state of an infant taken from a study on the effects of prenatal exposure to alcohol. Details can be found in Stoffer et al. (1988), but, briefly, an electroencephalographic (EEG) sleep recording of approximately two hours is obtained on a full-term infant 24 to 36 hours after birth, and the recording is scored by a pediatric neurologist for sleep state. Sleep state is categorized, per minute, into one of six possible states: `qt`: quiet sleep - trace alternant, `qh`: quiet sleep - high voltage, `tr`: transitional sleep, `al`: active sleep - low voltage, `ah`: active sleep - high voltage, and `aw`: awake. This particular infant was never awake during the study.

It is not difficult to notice a pattern in the data if we concentrate on active vs. quiet sleep (that is, focus on the first letter). But, it would be difficult to try to assess patterns in a longer sequence, or if more categories were present, without some graphical aid. One simple method would be to *scale* the data, that is, *assign numerical values to the categories* and then draw a time plot of the scaled series. Because the states have an order, one obvious scaling is

$$1 = \texttt{qt} \quad 2 = \texttt{qh} \quad 3 = \texttt{tr} \quad 4 = \texttt{al} \quad 5 = \texttt{ah} \quad 6 = \texttt{aw}, \tag{7.180}$$

and Figure 7.20 shows the time plot using this scaling. Another interesting scaling might be to combine the quiet states and the active states:

$$1 = \texttt{qt} \quad 1 = \texttt{qh} \quad 2 = \texttt{tr} \quad 3 = \texttt{al} \quad 3 = \texttt{ah} \quad 4 = \texttt{aw}. \tag{7.181}$$

The time plot using (7.181) would be similar to Figure 7.20 as far as the

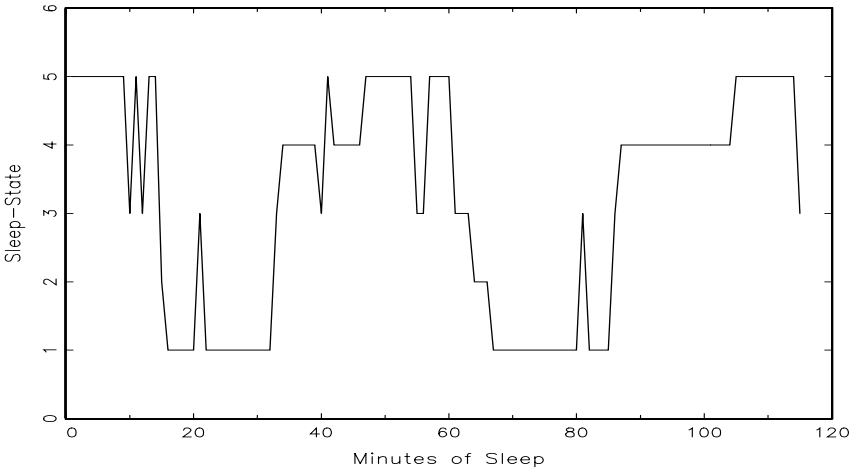


Figure 7.20 Time plot of the EEG sleep state data in Table 7.10 using the scaling in (7.180).

cyclic (in and out of quiet sleep) behavior of this infant's sleep pattern. Figure 7.21 shows the periodogram of the sleep data using the scaling in (7.180). A large peak exists at the frequency corresponding to one cycle every 60 minutes. As we might imagine, the general appearance of the periodogram using the scaling (7.181) (not shown) is similar to Figure 7.20. Most of us would feel comfortable with this analysis even though we made an arbitrary and ad hoc choice about the particular scaling. It is evident from the data (without any

Table 7.10 Infant EEG Sleep-States (per minute)
(read down and across)

ah	qt	qt	al	tr	qt	al	ah
ah	qt	qt	ah	tr	qt	al	ah
ah	qt	tr	ah	tr	qt	al	ah
ah	qt	al	ah	qh	qt	al	ah
ah	qt	al	ah	qh	qt	al	ah
ah	tr	al	ah	qt	qt	al	ah
ah	qt	al	ah	qt	qt	al	ah
ah	qt	al	ah	qt	qt	al	ah
tr	qt	tr	tr	qt	qt	al	tr
ah	qt	ah	tr	qt	tr	al	
tr	qt	al	ah	qt	al	al	
ah	qt	al	ah	qt	al	al	
ah	qt	al	ah	qt	al	al	
qh	qt	al	ah	qt	al	ah	

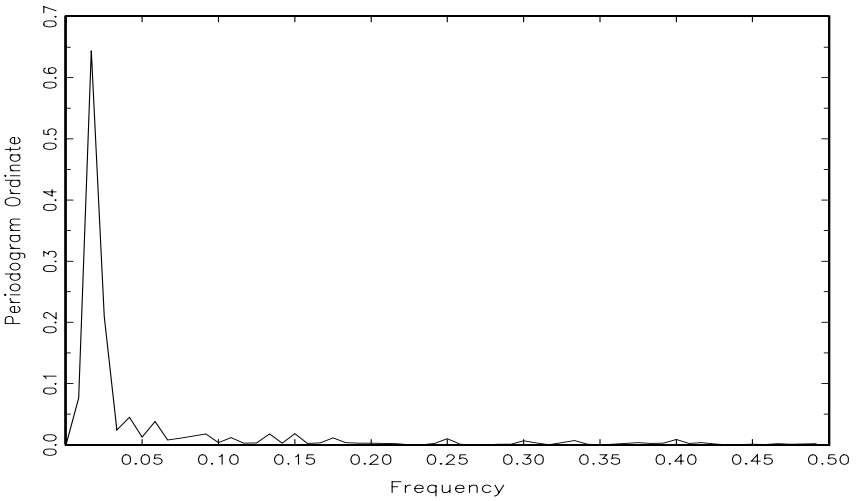


Figure 7.21 Periodogram of the EEG sleep state data in Figure 7.20 based on the scaling in (7.180). The peak corresponds to a frequency of approximately one cycle every 60 minutes.

scaling) that if the interest is in infant sleep cycling, this particular sleep study indicates an infant cycles between active and quiet sleep at a rate of about one cycle per hour.

The intuition used in the previous example is lost when we consider a long DNA sequence. Briefly, a DNA strand can be viewed as a long string of linked nucleotides. Each nucleotide is composed of a nitrogenous base, a five carbon sugar, and a phosphate group. There are four different bases, and they can be grouped by size; the pyrimidines, thymine (`T`) and cytosine (`C`), and the purines, adenine (`A`) and guanine (`G`). The nucleotides are linked together by a backbone of alternating sugar and phosphate groups with the $5'$ carbon of one sugar linked to the $3'$ carbon of the next, giving the string direction. DNA molecules occur naturally as a double helix composed of polynucleotide strands with the bases facing inwards. The two strands are complementary, so it is sufficient to represent a DNA molecule by a sequence of bases on a single strand. Thus, a strand of DNA can be represented as a sequence of letters, termed base pairs (bp), from the finite alphabet $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. The order of the nucleotides contains the genetic information specific to the organism. Expression of information stored in these molecules is a complex multistage process. One important task is to translate the information stored in the protein-coding sequences (CDS) of the DNA. A common problem in analyzing long DNA sequence data is in identifying CDS dispersed throughout the sequence and separated by regions of noncoding (which makes up most of the DNA). Table 7.11 shows part of the Epstein–Barr virus (EBV) DNA sequence. The entire EBV DNA sequence

Table 7.11 Part of the Epstein–Barr Virus DNA Sequence (read across and down)

AGAATTCGTC	TTGCTCTATT	CACCCTTACT	TTTCTTCTTG	CCCGTTCTCT	TTCTTAGTAT
GAATCCAGTA	TGCCTGCCTG	TAATTGTTGC	GCCCTACCTC	TTTTGGCTGG	CGGCTATTGC
CGCCTCGTGT	TTCACGGCCT	CAGTTAGTAC	CGTTGTGACC	GCCACCGGCT	TGGCCCTCTC
ACTTCTACTC	TTGGCAGCAG	TGGCCAGCTC	ATATGCCGCT	GCACAAAGGA	AACTGCTGAC
ACCGGTGACA	GTGCTTACTG	CGGTTGTCAC	TTGTGAGTAC	ACACGCACCA	TTTACAATGC
ATGATGTTCG	TGAGATTGAT	CTGTCTCTAA	CAGTTCACTT	CCTCTGCTTT	TCTCCTCAGT
CTTTGCAATT	TGCCTAACAT	GGAGGATTGA	GGACCCACCT	TTTAATTCTC	TTCTGTTTGC
ATTGCTGGCC	GCAGCTGGCG	GACTACAAGG	CATTTACGGT	TAGTGTGCCT	CTGTTATGAA
ATGCAGGTTT	GACTTCATAT	GTATGCCTTG	GCATGACGTC	AACTTTACTT	TTATTTCAGT
TCTGGTGATG	CTTGTGCTCC	TGATACTAGC	GTACAGAAGG	AGATGGCGCC	GTTTGACTGT
TTGTGGCGGC	ATCATGTTTT	TGGCATGTGT	ACTTGTCCTC	ATCGTCGACG	CTGTTTTGCA
GCTGAGTCCC	CTCCTTGGAG	CTGTAACTGT	GGTTTCCATG	ACGCTGCTGC	TACTGGCTTT
CGTCCTCTGG	CTCTCTTCGC	CAGGGGGCCT	AGGTACTCTT	GGTGCAGCCC	TTTTAACATT
GGCAGCAGGT	AAGCCACACG	TGTGACATTG	CTTGCCTTTT	TGCCACATGT	TTTCTGGACA
CAGGACTAAC	CATGCCATCT	CTGATTATAG	CTCTGGCACT	GCTAGCGTCA	CTGATTTTGG
GCACACTTAA	CTTGACTACA	ATGTTCCTTC	TCATGCTCCT	ATGGACACTT	GGTAAGTTTT
CCCTTCCTTT	AACTCATTAC	TTGTTCTTTT	GTAATCGCAG	CTCTAACTTG	GCATCTCTTT
TACAGTGGTT	CTCCTGATTT	GCTCTTCGTG	CTCTTCATGT	CCACTGAGCA	AGATCCTTCT

consists of approximately 172,000 bp.

We could try scaling according to the purine–pyrimidine alphabet, that is $\mathtt{A} = \mathtt{G} = 0$ and $\mathtt{C} = \mathtt{T} = 1$, but this is not necessarily of interest for every CDS of EBV. Numerous possible alphabets of interest exist. For example, we might focus on the strong–weak hydrogen-bonding alphabet $\mathtt{C} = \mathtt{G} = 0$ and $\mathtt{A} = \mathtt{T} = 1$. Although model calculations as well as experimental data strongly agree that some kind of periodic signal exists in certain DNA sequences, a large disagreement about the exact type of periodicity exists. In addition, a disagreement exists about which nucleotide alphabets are involved in the signals.

If we consider the naive approach of arbitrarily assigning numerical values (scales) to the categories and then proceeding with a spectral analysis, the result will depend on the particular assignment of numerical values. For example, consider the artificial sequence ACGTACGTACGT... . Then, setting $\mathtt{A} = \mathtt{G} = 0$ and $\mathtt{C} = \mathtt{T} = 1$ yields the numerical sequence 010101010101..., or one cycle every two base pairs. Another interesting scaling is $\mathtt{A} = 1, \mathtt{C} = 2, \mathtt{G} = 3$, and $\mathtt{T} = 4$, which results in the sequence 123412341234..., or one cycle every four bp. In this example, both scalings (that is, {A, C, G, T} = {0, 1, 0, 1} and {A, C, G, T} = {1, 2, 3, 4}) of the nucleotides are interesting and bring out different properties of the sequence. Hence, we do not want to focus on only one scaling. Instead, the focus should be on finding all possible scalings that bring out all of the interesting features in the data. Rather than choose values arbitrarily, the spectral envelope approach selects scales that help emphasize any periodic feature that exists in a categorical time series of virtually any length in a quick and automated fashion. In addition, the technique can help in determining whether a sequence is merely a random assignment of categories.

The Spectral Envelope for Categorical Time Series

As a general description, the spectral envelope is a frequency-based, principal components technique applied to a multivariate time series. First, we will focus on the basic concept and its use in the analysis of categorical time series. Technical details can be found in Stoffer et al. (1993).

Briefly, in establishing the spectral envelope for categorical time series, the basic question of how to efficiently discover periodic components in categorical time series was addressed. This, was accomplished via nonparametric spectral analysis as follows. Let x_t, $t = 0, \pm 1, \pm 2, \ldots$, be a categorical-valued time series with finite state-space $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$. Assume x_t is stationary and $p_j = \Pr\{x_t = c_j\} > 0$ for $j = 1, 2, \ldots, k$. For $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)' \in \mathbf{R}^k$, denote by $x_t(\boldsymbol{\beta})$ the real-valued stationary time series corresponding to the scaling that assigns the category c_j the numerical value β_j, $j = 1, 2, \ldots, k$. The spectral density of $x_t(\boldsymbol{\beta})$ will be denoted by $f_{xx}(\omega; \boldsymbol{\beta})$. The goal is to find scalings $\boldsymbol{\beta}$, so the spectral density is in some sense interesting, and to summarize the spectral information by what is called the spectral envelope.

In particular, $\boldsymbol{\beta}$ is chosen to maximize the power at each frequency, ω, of interest, relative to the total power $\sigma^2(\boldsymbol{\beta}) = \mathrm{var}\{x_t(\boldsymbol{\beta})\}$. That is, we chose $\boldsymbol{\beta}(\omega)$, at each ω of interest, so

$$\lambda(\omega) = \max_{\beta} \left\{ \frac{f_{xx}(\omega; \boldsymbol{\beta})}{\sigma^2(\boldsymbol{\beta})} \right\}, \tag{7.182}$$

over all $\boldsymbol{\beta}$ not proportional to $\mathbf{1}_k$, the $k \times 1$ vector of ones. Note, $\lambda(\omega)$ is not defined if $\boldsymbol{\beta} = a\mathbf{1}_k$ for $a \in \mathbf{R}$ because such a scaling corresponds to assigning each category the same value a; in this case, $f_{xx}(\omega\,; \boldsymbol{\beta}) \equiv 0$ and $\sigma^2(\boldsymbol{\beta}) = 0$. The optimality criterion $\lambda(\omega)$ possesses the desirable property of being invariant under location and scale changes of $\boldsymbol{\beta}$.

As in most scaling problems for categorical data, it is useful to represent the categories in terms of the unit vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_k$, where $\boldsymbol{u}_j$ represents the $k \times 1$ vector with a one in the j-th row, and zeros elsewhere. We then define a k-dimensional stationary time series $\boldsymbol{y}_t$ by $\boldsymbol{y}_t = \boldsymbol{u}_j$ when $x_t = c_j$. The time series $x_t(\boldsymbol{\beta})$ can be obtained from the $\boldsymbol{y}_t$ time series by the relationship $x_t(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{y}_t$. Assume the vector process $\boldsymbol{y}_t$ has a continuous spectral density denoted by $f_{yy}(\omega)$. For each ω, $f_{yy}(\omega)$ is, of course, a $k \times k$ complex-valued Hermitian matrix. The relationship $x_t(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{y}_t$ implies $f_{xx}(\omega;\, \boldsymbol{\beta}) = \boldsymbol{\beta}' f_{yy}(\omega)\boldsymbol{\beta} = \boldsymbol{\beta}' f^{re}_{yy}(\omega)\boldsymbol{\beta}$, where $f^{re}_{yy}(\omega)$ denotes the real part[2] of $f_{yy}(\omega)$. The imaginary part disappears from the expression because it is skew-symmetric, that is, $f^{im}_{yy}(\omega)' = -f^{im}_{yy}(\omega)$. The optimality criterion can thus be expressed as

$$\lambda(\omega) = \max_{\beta} \left\{ \frac{\boldsymbol{\beta}' f^{re}_{yy}(\omega)\boldsymbol{\beta}}{\boldsymbol{\beta}' V \boldsymbol{\beta}} \right\}, \tag{7.183}$$

[2]In this section, it is more convenient to write complex values in the form $z = z^{re} + iz^{im}$, which represents a change from the notation used previously.

where V is the variance–covariance matrix of $\boldsymbol{y}_t$. The resulting scaling $\boldsymbol{\beta}(\omega)$ is called the optimal scaling.

The $\boldsymbol{y}_t$ process is a multivariate point process, and any particular component of $\boldsymbol{y}_t$ is the individual point process for the corresponding state (for example, the first component of $\boldsymbol{y}_t$ indicates whether the process is in state c_1 at time t). For any fixed t, $\boldsymbol{y}_t$ represents a single observation from a simple multinomial sampling scheme. It readily follows that $V = D - p\,p'$, where $p = (p_1, \ldots, p_k)'$, and D is the $k \times k$ diagonal matrix $D = \text{diag}\{p_1, \ldots, p_k\}$. Because, by assumption, $p_j > 0$ for $j = 1, 2, \ldots, k$, it follows that $\text{rank}(V) = k-1$ with the null space of V being spanned by $\mathbf{1}_k$. For any $k \times (k-1)$ full rank matrix Q whose columns are linearly independent of $\mathbf{1}_k$, $Q'VQ$ is a $(k-1) \times (k-1)$ positive-definite symmetric matrix.

With the matrix Q as previously defined, define $\lambda(\omega)$ to be the largest eigenvalue of the determinantal equation

$$|Q' f_{yy}^{re}(\omega) Q - \lambda(\omega) Q' V Q| = 0,$$

and let $\boldsymbol{b}(\omega) \in \mathbf{R}^{k-1}$ be any corresponding eigenvector, that is,

$$Q' f_{yy}^{re}(\omega) Q \boldsymbol{b}(\omega) = \lambda(\omega) Q' V Q \boldsymbol{b}(\omega).$$

The eigenvalue $\lambda(\omega) \geq 0$ does not depend on the choice of Q. Although the eigenvector $\boldsymbol{b}(\omega)$ depends on the particular choice of Q, the equivalence class of scalings associated with $\boldsymbol{\beta}(\omega) = Q\boldsymbol{b}(\omega)$ does not depend on Q. A convenient choice of Q is $Q = [I_{k-1} \mid \mathbf{0}\,]'$, where I_{k-1} is the $(k-1) \times (k-1)$ identity matrix and $\mathbf{0}$ is the $(k-1) \times 1$ vector of zeros . For this choice, $Q' f_{yy}^{re}(\omega) Q$ and $Q'VQ$ are the upper $(k-1) \times (k-1)$ blocks of $f_{yy}^{re}(\omega)$ and V, respectively. This choice corresponds to setting the last component of $\boldsymbol{\beta}(\omega)$ to zero.

The value $\lambda(\omega)$ itself has a useful interpretation; specifically, $\lambda(\omega)d\omega$ represents the largest proportion of the total power that can be attributed to the frequencies $(\omega, \omega + d\omega)$ for any particular scaled process $x_t(\boldsymbol{\beta})$, with the maximum being achieved by the scaling $\boldsymbol{\beta}(\omega)$. Because of its central role, $\lambda(\omega)$ is defined to be the spectral envelope of a stationary categorical time series.

The name spectral envelope is appropriate since $\lambda(\omega)$ envelopes the standardized spectrum of any scaled process. That is, given any $\boldsymbol{\beta}$ normalized so that $x_t(\boldsymbol{\beta})$ has total power one, $f_{xx}(\omega\,;\,\boldsymbol{\beta}) \leq \lambda(\omega)$ with equality if and only if $\boldsymbol{\beta}$ is proportional to $\boldsymbol{\beta}(\omega)$.

Given observations x_t, for $t = 1, \ldots, n$, on a categorical time series, we form the multinomial point process $\boldsymbol{y}_t$, for $t = 1, \ldots, n$. Then, the theory for estimating the spectral density of a multivariate, real-valued time series can be applied to estimating $f_{yy}(\omega)$, the $k \times k$ spectral density of $\boldsymbol{y}_t$. Given an estimate $\widehat{f}_{yy}(\omega)$ of $f_{yy}(\omega)$, estimates $\widehat{\lambda}(\omega)$ and $\widehat{\boldsymbol{\beta}}(\omega)$ of the spectral envelope, $\lambda(\omega)$, and the corresponding scalings, $\boldsymbol{\beta}(\omega)$, can then be obtained. Details on estimation and inference for the sample spectral envelope and the optimal scalings can be found in Stoffer et al. (1993), but the main result of that paper is as follows:

If $\widehat{f}_{yy}(\omega)$ is a consistent spectral estimator and if for each $j = 1, \ldots, J$, the largest root of $f^{re}_{yy}(\omega_j)$ is distinct, then

$$\left\{\eta_n[\widehat{\lambda}(\omega_j) - \lambda(\omega_j)]/\lambda(\omega_j),\ \eta_n[\widehat{\boldsymbol{\beta}}(\omega_j) - \boldsymbol{\beta}(\omega_j)];\ j = 1, \ldots, J\right\} \tag{7.184}$$

converges ($n \to \infty$) jointly in distribution to independent zero-mean, normal, distributions, the first of which is standard normal; the asymptotic covariance structure of $\widehat{\boldsymbol{\beta}}(\omega_j)$ is discussed in Stoffer et al. (1993). Result (7.184) is similar to (7.151), but in this case, $\boldsymbol{\beta}(\omega)$ and $\widehat{\boldsymbol{\beta}}(\omega)$ are real. The term η_n is the same as in (7.184), and its value depends on the type of estimator being used. Based on these results, asymptotic normal confidence intervals and tests for $\lambda(\omega)$ can be readily constructed. Similarly, for $\boldsymbol{\beta}(\omega)$, asymptotic confidence ellipsoids and chi-square tests can be constructed; details can be found in Stoffer et al. (1993, Theorems 3.1 – 3.3).

Peak searching for the smoothed spectral envelope estimate can be aided using the following approximations. Using a first-order Taylor expansion, we have

$$\log \widehat{\lambda}(\omega) \approx \log \lambda(\omega) + \frac{\widehat{\lambda}(\omega) - \lambda(\omega)}{\lambda(\omega)}, \tag{7.185}$$

so $\eta_n[\log \widehat{\lambda}(\omega) - \log \lambda(\omega)]$ is approximately standard normal. It follows that $E[\log \widehat{\lambda}(\omega)] \approx \log \lambda(\omega)$ and $\mathrm{var}[\log \widehat{\lambda}(\omega)] \approx \eta_n^{-2}$. If no signal is present in a sequence of length n, we expect $\lambda(j/n) \approx 2/n$ for $1 < j < n/2$, and hence approximately $(1-\alpha)\times 100\%$ of the time, $\log \widehat{\lambda}(\omega)$ will be less than $\log(2/n) + (z_\alpha/\eta_n)$, where z_α is the $(1-\alpha)$ upper tail cutoff of the standard normal distribution. Exponentiating, the α critical value for $\widehat{\lambda}(\omega)$ becomes $(2/n)\exp(z_\alpha/\eta_n)$. Useful values of z_α are $z_{.001} = 3.09$, $z_{.0001} = 3.71$, and $z_{.00001} = 4.26$, and from our experience, thresholding at these levels works well.

Example 7.17 Spectral Analysis of DNA Sequences

We give explicit instructions for the calculations involved in estimating the spectral envelope of a DNA sequence, x_t, for $t = 1, \ldots, n$, using the nucleotide alphabet.

- In this example, we hold the scale for `T` fixed at zero. In this case, we form the 3×1 data vectors $\boldsymbol{y}_t$:

$$\begin{array}{llll} \boldsymbol{y}_t = (1,0,0)' & \text{if} & x_t = \texttt{A}; & \qquad \boldsymbol{y}_t = (0,1,0)' \ \text{ if } \ x_t = \texttt{C}; \\ \boldsymbol{y}_t = (0,0,1)' & \text{if} & x_t = \texttt{G}; & \qquad \boldsymbol{y}_t = (0,0,0)' \ \text{ if } \ x_t = \texttt{T}. \end{array}$$

  The scaling vector is $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$, and the scaled process is $x_t(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{y}_t$.

- Calculate the DFT of the data

$$\boldsymbol{Y}(j/n) = n^{-1/2} \sum_{t=1}^{n} \boldsymbol{y}_t \exp(-2\pi i t j/n).$$

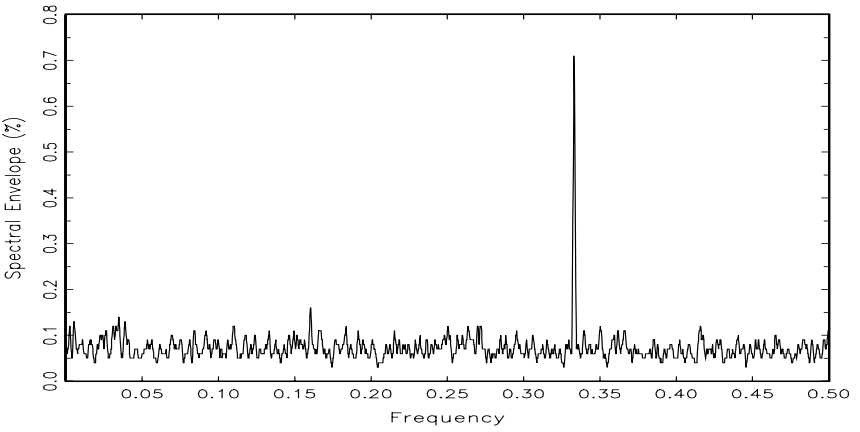


Figure 7.22 Smoothed sample spectral envelope of the BNRF1 gene from the Epstein–Barr virus.

Note $\boldsymbol{Y}(j/n)$ is a 3×1 complex-valued vector. Calculate the periodogram, $I(j/n) = \boldsymbol{Y}(j/n)\boldsymbol{Y}^*(j/n)$, for $j = 1, \ldots, [n/2]$, and retain only the real part, say, $I^{re}(j/n)$.

- Smooth the $I^{re}(j/n)$ to obtain an estimate of $f^{re}_{yy}(j/n)$. For example, using (7.150) with $L = 3$ and triangular weighting, we would calculate

$$\widehat{f}^{re}_{yy}(j/n) = \frac{1}{4}I^{re}\left(\frac{j-1}{n}\right) + \frac{1}{2}I^{re}\left(\frac{j}{n}\right) + \frac{1}{4}I^{re}\left(\frac{j+1}{n}\right).$$

- Calculate the 3×3 sample variance–covariance matrix,

$$S_{yy} = n^{-1}\sum_{t=1}^{n}(\boldsymbol{y}_t - \overline{\boldsymbol{y}})(\boldsymbol{y}_t - \overline{\boldsymbol{y}})',$$

where $\overline{\boldsymbol{y}} = n^{-1}\sum_{t=1}^{n}\boldsymbol{y}_t$ is the sample mean of the data.

- For each $\omega_j = j/n$, $j = 0, 1, \ldots, [n/2]$, determine the largest eigenvalue and the corresponding eigenvector of the matrix $2n^{-1}S_{yy}^{-1/2}\widehat{f}^{re}_{yy}(\omega_j)S_{yy}^{-1/2}$. Note, $S_{yy}^{1/2}$ is the unique square root matrix of S_{yy}.

- The sample spectral envelope $\widehat{\lambda}(\omega_j)$ is the eigenvalue obtained in the previous step. If $\boldsymbol{b}(\omega_j)$ denotes the eigenvector obtained in the previous step, the optimal sample scaling is $\widehat{\boldsymbol{\beta}}(\omega_j) = S_{yy}^{-1/2}\boldsymbol{b}(\omega_j)$; this will result in three values, the value corresponding to the fourth category, `T` being held fixed at zero.

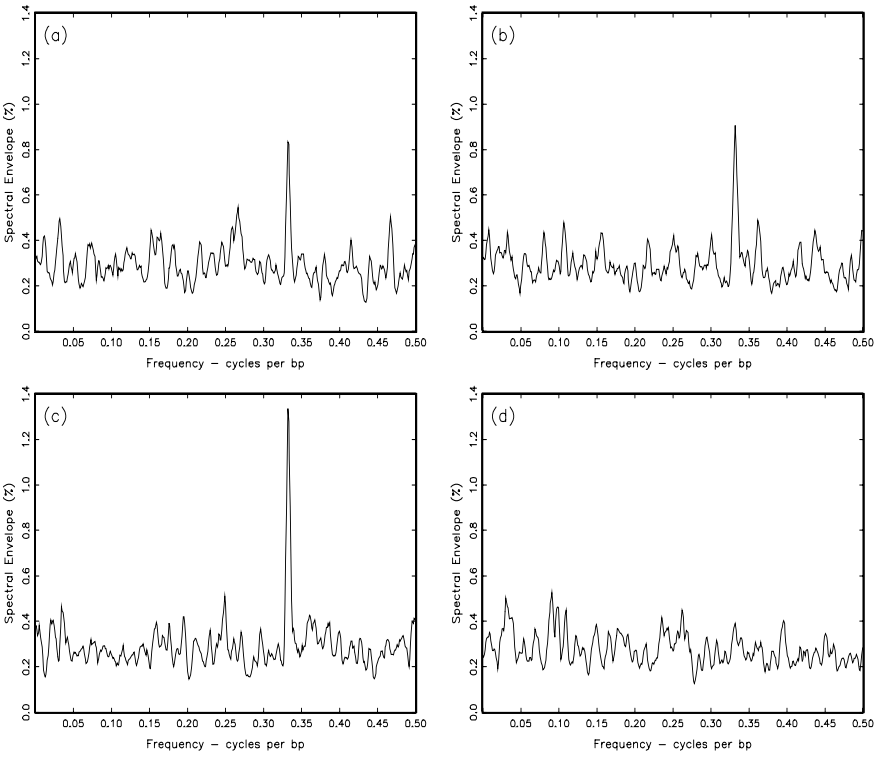


Figure 7.23 Smoothed sample spectral envelope of the BNRF1 gene from the Epstein–Barr virus: (a) first 1000 bp, (b) second 1000 bp, (c) third 1000 bp, and (d) last 954 bp.

Example 7.18 Dynamic Analysis of the Gene Labeled BNRF1 of the Epstein–Barr Virus

In this example, we focus on a dynamic (or sliding-window) analysis of the gene labeled BNRF1 (bp 1736-5689) of Epstein–Barr. Figure 7.22 shows the spectral envelope, using (7.150) with $L = 11$ and $h_0 = 6/36, h_1 = 5/36, \ldots, h_5 = 1/36$, of the entire coding sequence (3954 bp long). The figure also shows a strong signal at frequency 1/3; the corresponding optimal scaling was $\mathtt{A} = 0.04, \mathtt{C} = 0.71, \mathtt{G} = 0.70, \mathtt{T} = 0$, which indicates the signal is in the strong–weak bonding alphabet, $S = \{\mathtt{C}, \mathtt{G}\}$ and $W = \{\mathtt{A}, \mathtt{T}\}$.

Figure 7.23 shows the result of computing the spectral envelope over three nonoverlapping 1000-bp windows and one window of 954 bp, across the CDS, namely, the first, second, third, and fourth quarters of BNRF1.

An approximate 0.0001 significance threshold is .69%. The first three quarters contain the signal at the frequency 1/3 (Figure 7.23a-c); the corresponding sample optimal scalings for the first three windows were (a) $\mathtt{A} = 0.06, \mathtt{C} = 0.69, \mathtt{G} = 0.72, \mathtt{T} = 0$; (b) $\mathtt{A} = 0.09, \mathtt{C} = 0.70, \mathtt{G} = 0.71, \mathtt{T} = 0$; (c) $\mathtt{A} = 0.18, \mathtt{C} = 0.59, \mathtt{G} = 0.77, \mathtt{T} = 0$. The first two windows are consistent with the overall analysis. The third section, however, shows some minor departure from the strong-weak bonding alphabet. The most interesting outcome is that the fourth window shows that no signal is present. This leads to the conjecture that the fourth quarter of BNRF1 of Epstein–Barr is actually noncoding.

The Spectral Envelope for Real-Valued Time Series

The concept of the spectral envelope for categorical time series was extended to real-valued time series, $\{x_t; t = 0, \pm 1, \pm 2, \ldots, \}$, in McDougall et al. (1997). The process x_t can be vector-valued, but here we will concentrate on the univariate case. Further details can be found in McDougall et al. (1997). The concept is similar to projection pursuit (Friedman and Stuetzle, 1981). Let $\mathcal{G}$ denote a k-dimensional vector space of continuous real-valued transformations with $\{g_1, \ldots, g_k\}$ being a set of basis functions satisfying $E[g_i(x_t)^2] < \infty$, $i = 1, \ldots, k$. Analogous to the categorical time series case, define the scaled time series with respect to the set $\mathcal{G}$ to be the real-valued process

$$x_t(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{y}_t = \beta_1 g_1(x_t) + \cdots + \beta_k g_k(x_t)$$

obtained from the vector process

$$\boldsymbol{y}_t = \Big(g_1(X_t), \ldots, g_k(X_t)\Big)',$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)' \in \mathbf{R}^k$. If the vector process, $\boldsymbol{y}_t$, is assumed to have a continuous spectral density, say, $f_{yy}(\omega)$, then $x_t(\boldsymbol{\beta})$ will have a continuous spectral density $f_{xx}(\omega; \boldsymbol{\beta})$ for all $\boldsymbol{\beta} \neq \mathbf{0}$. Noting, $f_{xx}(\omega; \boldsymbol{\beta}) = \boldsymbol{\beta}' f_{yy}(\omega)\boldsymbol{\beta} = \boldsymbol{\beta}' f^{re}_{yy}(\omega)\boldsymbol{\beta}$, and $\sigma^2(\boldsymbol{\beta}) = \mathrm{var}[x_t(\boldsymbol{\beta})] = \boldsymbol{\beta}' V \boldsymbol{\beta}$, where $V = \mathrm{var}(\boldsymbol{y}_t)$ is assumed to be positive definite, the optimality criterion

$$\lambda(\omega) = \sup_{\boldsymbol{\beta} \neq \mathbf{0}} \left\{ \frac{\boldsymbol{\beta}' f^{re}_{yy}(\omega)\boldsymbol{\beta}}{\boldsymbol{\beta}' V \boldsymbol{\beta}} \right\}, \tag{7.186}$$

is well defined and represents the largest proportion of the total power that can be attributed to the frequency ω for any particular scaled process $x_t(\boldsymbol{\beta})$. This interpretation of $\lambda(\omega)$ is consistent with the notion of the spectral envelope introduced in the previous section and provides the following working definition: *The spectral envelope of a time series with respect to the space $\mathcal{G}$ is defined to be $\lambda(\omega)$.*

The solution to this problem, as in the categorical case, is attained by finding the largest scalar $\lambda(\omega)$ such that

$$f^{re}_{yy}(\omega)\boldsymbol{\beta}(\omega) = \lambda(\omega) V \boldsymbol{\beta}(\omega) \tag{7.187}$$

for $\boldsymbol{\beta}(\omega) \neq \mathbf{0}$. That is, $\lambda(\omega)$ is the largest eigenvalue of $f^{re}_{yy}(\omega)$ in the metric of V, and the optimal scaling, $\boldsymbol{\beta}(\omega)$, is the corresponding eigenvector.

If x_t is a categorical time series taking values in the finite state-space $\mathcal{S} = \{c_1, c_2, \ldots, c_k\}$, where c_j represents a particular category, an appropriate choice for $\mathcal{G}$ is the set of indicator functions $g_j(x_t) = I(x_t = c_j)$. Hence, this is a natural generalization of the categorical case. In the categorical case, $\mathcal{G}$ does not consist of linearly independent g's, but it was easy to overcome this problem by reducing the dimension by one. In the vector-valued case, $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})'$, we consider $\mathcal{G}$ to be the class of transformations from $\mathbf{R}^p$ into $\mathbf{R}$ such that the spectral density of $g(\boldsymbol{x}_t)$ exists. One class of transformations of interest are linear combinations of $\boldsymbol{x}_t$. In Tiao et al. (1993), for example, linear transformations of this type are used in a time domain approach to investigate contemporaneous relationships among the components of multivariate time series. Estimation and inference for the real-valued case are analogous to the methods described in the previous section for the categorical case. We focus on two examples here; numerous other examples can be found in McDougall et al. (1997).

Example 7.19 Residual Analysis

A relevant situation may be when x_t is the residual process obtained from some modeling procedure. If the fitted model is appropriate, the residuals should exhibit properties similar to an iid sequence. Departures of the data from the fitted model may suggest model misspecification, non-Gaussian data, or the existence of a nonlinear structure, and the spectral envelope would provide a simple diagnostic tool to aid in a residual analysis.

The series considered here is the quarterly U.S. real GNP which was analyzed in Chapter 3, Examples (3.35) and (3.36). Recall an MA(2) model was fit to the growth rate, and the residuals from this fit are plotted in Figure 3.16 . As discussed in Example (3.36), the residuals from the model fit appear to be uncorrelated; there appears to be one or two outliers, but their magnitudes are not that extreme. In addition, the standard residual analyses showed no obvious structure among the residuals.

Although the MA(2) model appears to be appropriate, Tiao and Tsay (1994) investigated the possibility of nonlinearities in GNP growth rate. Their overall conclusion was that there is subtle nonlinear behavior in the data because the economy behaves differently during expansion periods than during recession periods.

The spectral envelope, used as a diagnostic tool on the residuals, clearly indicates the MA(2) model is not adequate, and that further analysis is warranted. Here, the generating set $\mathcal{G} = \{x, |x|, x^2\}$—which seems natural for a residual analysis—was used to estimate the spectral envelope

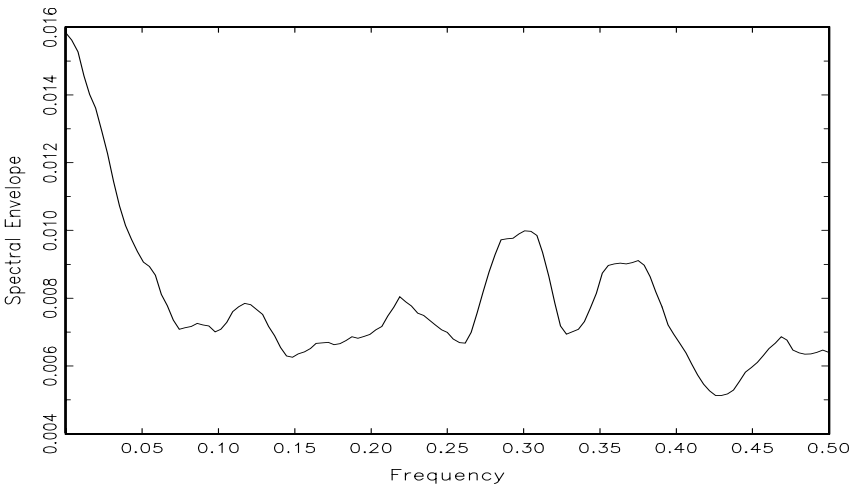


Figure 7.24 Spectral envelope with respect to $\mathcal{G} = \{x, |x|, x^2\}$ of the residuals from an MA(2) fit to the U.S. GNP growth rate data.

for the residuals from the MA(2) fit, and the result is plotted in Figure 7.24. A smoothed periodogram estimate was obtained using $L = 21$ and triangular weighting, $h_0 = 11/121, h_{\pm 1} = 10/121, \ldots, h_{\pm 10} = 1/121$ in (7.150). Clearly, the residuals are not iid, and considerable power is present at the low frequencies. The presence of spectral power at very low frequencies in detrended economic series has been frequently reported and is typically associated with long-range dependence. In fact, our choice of $\mathcal{G}$ was partly influenced by the work of Ding et al. (1993) who applied transformations of the form $|x_t|^d$, for $d \in (0, 3]$, to the S&P 500 stock market series. The estimated optimal transformation at the first nonzero frequency, $\omega = 0.006$, was $\widehat{\boldsymbol{\beta}}(0.006) = (1, 20, -2916)'$, which leads to the transformation

$$y = x + 20|x| - 2916x^2. \tag{7.188}$$

This transformation is plotted in Figure 7.25. The transformation given in (7.188) is basically the absolute value (with some slight curvature and asymmetry) for most of the residual values, but the effect of extreme-valued residuals (outliers) is dampened.

Example 7.20 Optimal Transformations

In this example, we consider a contrived data set, in which we know the optimal transformation, say, g_0, and we determine whether the technology can find the transformation when g_0 is not in $\mathcal{G}$. The data, x_t, are

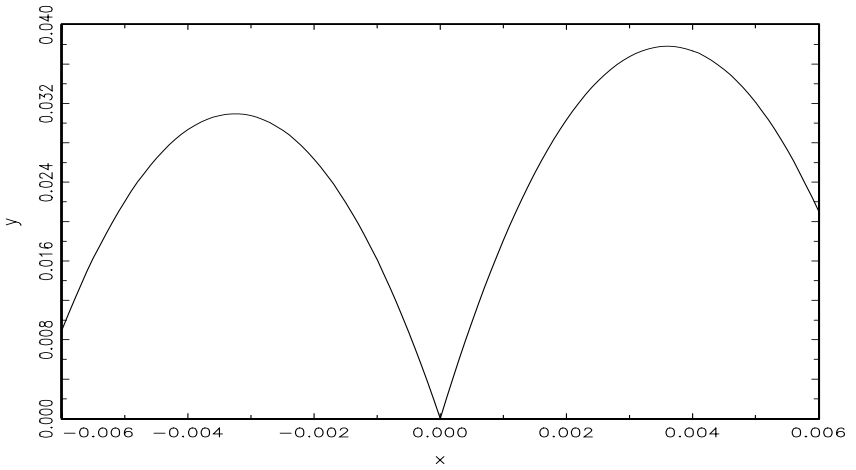


Figure 7.25 Estimated optimal transformation, (7.188), for the GNP residuals at $\omega = 0.006$.

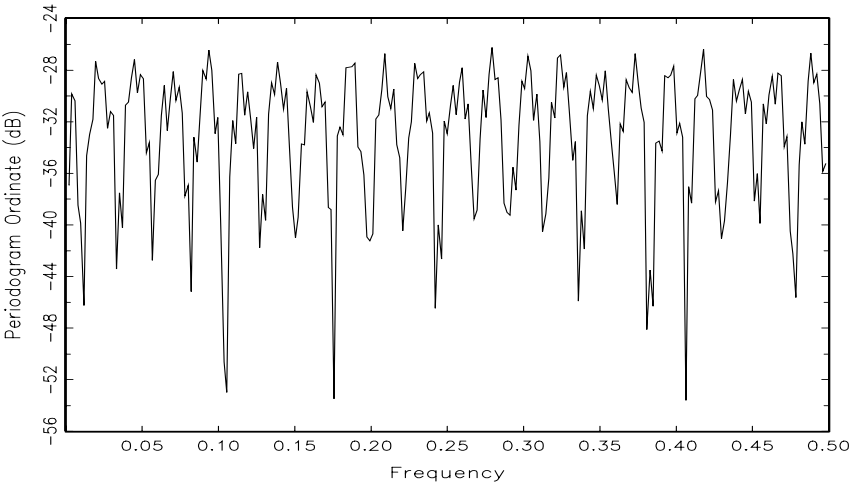


Figure 7.26 Periodogram, in decibels, of the data generated from (7.189) after tapering by a cosine bell.

generated by the nonlinear model

$$x_t = \exp\{3\sin(2\pi t\omega_0) + \epsilon_t\}, \quad t = 1, \ldots, 512, \tag{7.189}$$

where $\omega_0 = 51/512$ and ϵ_t is white Gaussian noise with a variance of 16. This example is adapted from Breiman and Friedman (1985), where the ACE algorithm is introduced. The optimal transformation in this case is $g_0(x_t) = \ln(x_t)$, wherein the data are generated from a sinusoid plus

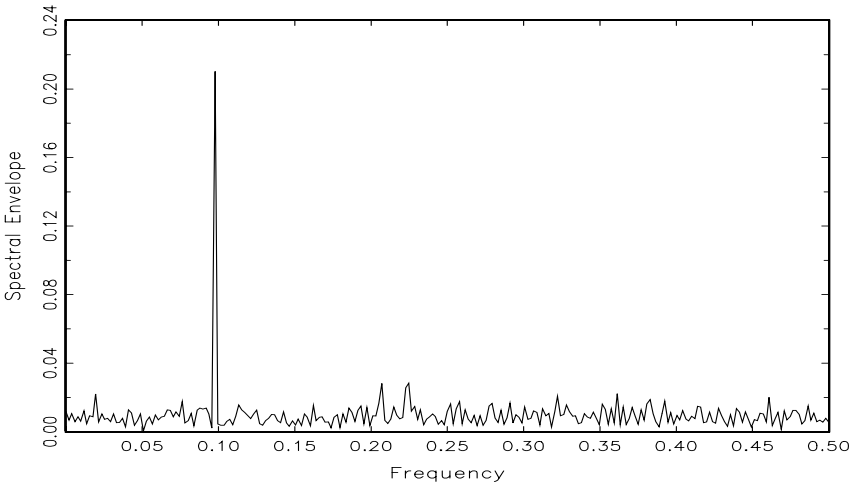


Figure 7.27 Spectral envelope with respect to $\mathcal{G} = \{x, \sqrt{x}, \sqrt[3]{x}\}$ of data generated from (7.189).

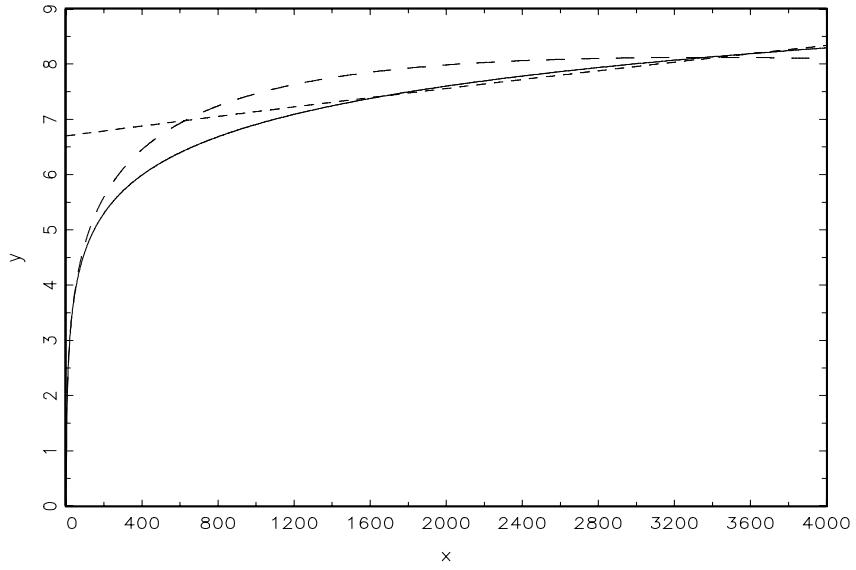


Figure 7.28 Log transformation, $y = \ln(x)$ (solid line), the estimated optimal transformation at ω_0 as given in (7.190) (dashed line), and the estimated optimal transformation at ω_0 using the inappropriate basis $\{x, x^2, x^3\}$ (short-dashed line).

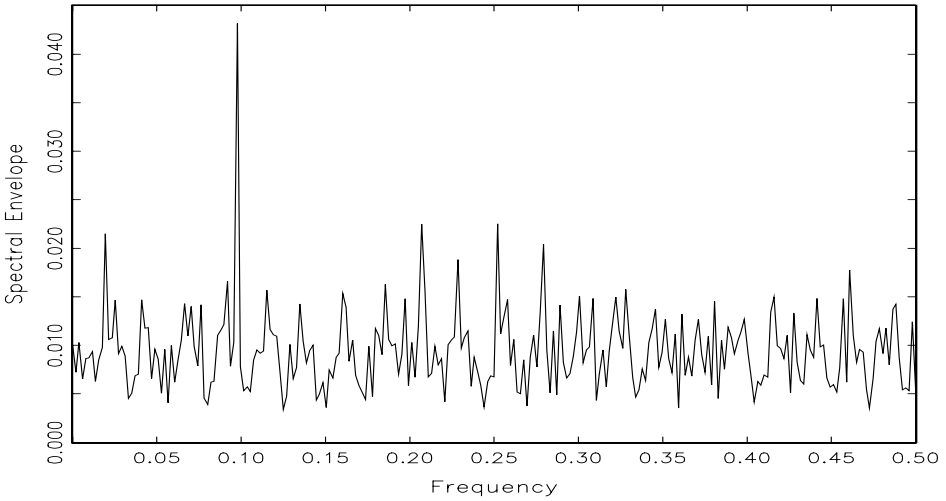


Figure 7.29 Spectral envelope with respect to $\mathcal{G} = \{x, x^2, x^3\}$.

noise. Of the 512 generated data, about 98% were less than 4000. Occasionally, the data values were extremely large (the data exceeded 100,000 about four times). The periodogram, in decibels $[10 \log_{10} X(\omega_j)]$, of the standardized and tapered (by a cosine bell) data is shown in Figure 7.26 and provides no evidence of any dominant frequency, including ω_0.

In contrast, the sample spectral envelope (Figure 7.27) computed with respect to $\mathcal{G} = \{x, \sqrt{x}, \sqrt[3]{x}\}$ has no difficulty in isolating ω_0. No smoothing was used here; so, based on Stoffer et al. (1993, Theorem 3.2), an approximate 0.0001 null significance threshold for the spectral envelope is 4.84% (the null hypothesis being that x_t is iid).

Figure 7.28 compares the estimated optimal transformation with respect to $\mathcal{G}$ with the log transformation for values less than 4000. The estimated transformation at ω_0 is given by

$$y = -.6 + 0.0003x - 0.3638\sqrt{x} + 1.9304\sqrt[3]{x}; \tag{7.190}$$

that is, $\widehat{\boldsymbol{\beta}}(\omega_0) = (0.0003, -0.3638, 1.9304)'$ after rescaling so (7.190) can be compared directly with $y = \ln(x)$.

Finally, it is worth mentioning the result obtained when the rather inappropriate basis, $\{x, x^2, x^3\}$, was used. Surprisingly, the spectral envelope in this case (Figure 7.29) looks similar to that of Figure 7.27. Also, the resulting estimated optimal transformation at ω_0, is close to the log transformation. In fact, as seen in Figure 7.28, it looks like what we would imagine as a linear approximation to $y = \ln(x)$ within the range of most of the data.

Problems

Section 7.2

7.1 Consider the complex Gaussian distribution for the random variable $\boldsymbol{X} = \boldsymbol{X}_c - i\boldsymbol{X}_s$, as defined in (7.1)-(7.3), where the argument ω_k has been suppressed. Now, the $2p \times 1$ real random variable $\boldsymbol{Z} = (\boldsymbol{X}_c', \boldsymbol{X}_s')'$ has a multivariate normal distribution with density

$$p(\boldsymbol{Z}) = (2\pi)^{-p}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{Z}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{Z}-\boldsymbol{\mu})\right\},$$

where $\boldsymbol{\mu} = (\boldsymbol{M}_c', \boldsymbol{M}_s')'$ is the mean vector. Prove

$$|\Sigma| = \left(\frac{1}{2}\right)^{2p}|C - iQ|^2,$$

using the result that the eigenvectors and eigenvalues of Σ occur in pairs, i.e., $(\boldsymbol{v}_c', \boldsymbol{v}_s')'$ and $(\boldsymbol{v}_s', -\boldsymbol{v}_c')'$, where $\boldsymbol{v}_c - i\boldsymbol{v}_s$ denotes the eigenvector of f_{xx}. Show that

$$\frac{1}{2}(\boldsymbol{Z}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{Z}-\boldsymbol{\mu})) = (\boldsymbol{X}-\boldsymbol{M})^* f^{-1}(\boldsymbol{X}-\boldsymbol{M})$$

so $p(\boldsymbol{X}) = p(\boldsymbol{Z})$ and we can identify the density of the complex multivariate normal variable $\boldsymbol{X}$ with that of the real multivariate normal $\boldsymbol{Z}$.

7.2 Prove $\widehat{f}$ in (7.6) maximizes the log likelihood (7.5) by minimizing the negative of the log likelihood

$$L\ln|f| + L\ \text{tr}\ \{\widehat{f}f^{-1}\}$$

in the form

$$L\sum_i\big(\lambda_i - \ln\lambda_i - 1\big) + Lp + L\ln|\widehat{f}|,$$

where the λ_i values correspond to the eigenvalues in a simultaneous diagonalization of the matrices f and $\hat{f}$; i.e., there exists a matrix P such that $P^*fP = I$ and $P^*\widehat{f}P = \text{diag}\ (\lambda_1, \ldots, \lambda_p) = \Lambda$. Note, $\lambda_i - \ln\lambda_i - 1 \geq 0$ with equality if and only if $\lambda_i = 1$, implying $\Lambda = I$ maximizes the log likelihood and $f = \widehat{f}$ is the maximimizing value.

Section 7.3

7.3 Verify (7.19) and (7.20) for the mean-squared prediction error MSE in (7.12). Use the orthogonality principle, which implies

$$MSE = E\left[(y_t - \sum_{r=-\infty}^{\infty}\boldsymbol{\beta}_r'\boldsymbol{x}_{t-r})y_t\right]$$

and gives a set of equations involving the autocovariance functions. Then, use the spectral representations and Fourier transform results to get the final result.

7.4 Consider the predicted series

$$\widehat{y}_t = \sum_{r=-\infty}^{\infty} \boldsymbol{\beta}_r' \boldsymbol{x}_{t-r},$$

where $\boldsymbol{\beta}_r$ satisfies (7.14). Show the ordinary coherence between y_t and $\widehat{y}_t$ is exactly the multiple coherence (7.21).

7.5 Consider the complex regression model (7.29) in the form

$$\boldsymbol{Y} = X\mathrm{B} + \boldsymbol{V},$$

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots Y_L)'$ denotes the observed DFTs after they have been re-indexed and $X = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L)'$ is a matrix containing the reindexed input vectors. The model is a complex regression model with $\boldsymbol{Y} = \boldsymbol{Y}_c - i\boldsymbol{Y}_s, X = X_c - iX_s, \boldsymbol{B} = \boldsymbol{B}_c - i\boldsymbol{B}_s$, and $\boldsymbol{V} = \boldsymbol{V}_c - i\boldsymbol{V}_s$ denoting the representation in terms of the usual cosine and sine transforms. Show the partitioned real regression model involving the $2L \times 1$ vector of cosine and sine transforms, say,

$$\begin{pmatrix} \boldsymbol{Y}_c \\ \boldsymbol{Y}_s \end{pmatrix} = \begin{pmatrix} X_c & -X_s \\ X_s & X_c \end{pmatrix} \begin{pmatrix} \boldsymbol{B}_c \\ \boldsymbol{B}_s \end{pmatrix} + \begin{pmatrix} \boldsymbol{V}_c \\ \boldsymbol{V}_s \end{pmatrix},$$

is isomorphic to the complex regression regression model in the sense that the real and imaginary parts of the complex model appear as components of the vectors in the real regression model. Use the usual regression theory to verify (7.28) holds. For example, writing the real regression model as

$$\boldsymbol{y} = x\boldsymbol{b} + \boldsymbol{v},$$

the isomorphism would imply

$$\begin{aligned} L(\widehat{f}_{yy} - \widehat{f}_{xy}^* \widehat{f}_{xx}^{-1} \widehat{f}_{xy}) &= \boldsymbol{Y}^*\boldsymbol{Y} - \boldsymbol{Y}^* X (X^*X)^{-1} X^* \boldsymbol{Y} \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}' x (x'x)^{-1} x' \boldsymbol{y}. \end{aligned}$$

Section 7.4

7.6 Consider estimating the function

$$\psi_t = \sum_{r=-\infty}^{\infty} \boldsymbol{a}_r' \boldsymbol{\beta}_{t-r}$$

by a linear filter estimator of the form

$$\widehat{\psi}_t = \sum_{r=-\infty}^{\infty} \boldsymbol{a}_r' \widehat{\boldsymbol{\beta}}_{t-r},$$

where $\widehat{\boldsymbol{\beta}}_t$ is defined by (7.43). Show a sufficient condition for $\widehat{\psi}_t$ to be an unbiased estimator; i.e., $E\ \widehat{\psi}_t = \psi_t$, is

$$H(\omega)Z(\omega) = I$$

for all ω. Similarly, show any other unbiased estimator satisfying the above condition has minimum variance (see Shumway and Dean, 1968), so the estimator given is a best linear unbiased (BLUE) estimator.

7.7 Consider a linear model with mean value function μ_t and a signal α_t delayed by an amount τ_j on each sensor, i.e.,

$$y_{jt} = \mu_t + \alpha_{t-\tau_j} + v_{jt}$$

Show the estimators (7.43) for the mean and the signal are the Fourier transforms of

$$\widehat{M}(\omega) = \frac{Y.(\omega) - \overline{\phi(\omega)}B_w(\omega)}{1 - |\phi(\omega)|^2}$$

and

$$\widehat{A}(\omega) = \frac{B_w(\omega) - \phi(\omega)Y.(\omega)}{1 - |\phi(\omega)|^2},$$

where

$$\phi(\omega) = \frac{1}{N} \sum_{j=1}^{N} \boldsymbol{e}^{2\pi i \omega \tau_j}$$

and $B_w(\omega)$ is defined in (7.65).

Section 7.5

7.8 Consider the estimator (7.68) as applied in the context of the random coefficient model (7.66). Prove the filter coefficients for the minimum mean square estimator can be determined from (7.69) and the mean square covariance is given by (7.72).

7.9 For the random coefficient model, verify the expected mean square of the regression power component is

$$\begin{aligned} E[SSR(\omega_k)] &= E[Y^*(\omega_k)Z(\omega_k)S_z^{-1}(\omega_k)Z^*(\omega_k)Y(\omega_k)] \\ &= Lf_\beta(\omega_k)\text{tr}\ \{S_z(\omega_k)\} + Lqf_v(\omega_k). \end{aligned}$$

Recall, the underlying frequency domain model is

$$\boldsymbol{Y}(\omega_k) = Z(\omega_k)\boldsymbol{B}(\omega_k) + \boldsymbol{V}(\omega_k),$$

where $\boldsymbol{B}(\omega_k)$ has spectrum $f_\beta(\omega_k)I_q$ and $\boldsymbol{V}(\omega_k)$ has spectrum $f_v(\omega_k)I_N$ and the two processes are uncorrelated.

Section 7.6

7.10 Suppose we have $I = 2$ groups and the models

$$y_{1jt} = \mu_t + \alpha_{1t} + v_{1jt}$$

for the $j = 1, \ldots, N$ observations in group 1 and

$$y_{2jt} = \mu_t + \alpha_{2t} + v_{2jt}$$

for the $j = 1, \ldots, N$ observations in group 2, with $\alpha_{1t} + \alpha_{2t} = 0$. Suppose we want to test equality of the two group means; i.e.,

$$y_{ijt} = \mu_t + v_{ijt}, \quad i = 1, 2.$$

Derive the residual and error power components corresponding to (7.84) and (7.85) for this particular case.

7.11 Verify the forms of the linear compounds involving the mean given in (7.91) and (7.92), using (7.89) and (7.90).

7.12 Show the ratio of the two smoothed spectra in (7.104) has the indicated F-distribution when $f_1(\omega) = f_2(\omega)$. When the spectra are not equal, show the variable is proportional to an F-distribution, where the proportionality constant depends on the ratio of the spectra.

Section 7.7

7.13 The problem of detecting a signal in noise can be considered using the model

$$x_t = s_t + w_t, \quad t = 1, \ldots, n,$$

for $p_1(\boldsymbol{x})$ when a signal is present and the model

$$x_t = w_t, \quad t = 1, \ldots, n,$$

for $p_2(\boldsymbol{x})$ when no signal is present. Under multivariate normality, we might specialize even further by assuming the vector $\boldsymbol{w} = (w_1, \ldots, w_n)'$ has a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance matrix $\Sigma = \sigma_w^2 I_n$, corresponding to white noise. Assuming the signal vector

$\boldsymbol{s} = (s_1, \ldots, s_n)'$ is fixed and known, show the discriminant function (7.113) becomes the matched filter

$$\frac{1}{\sigma_w^2} \sum_{t=1}^{n} s_t x_t - \frac{1}{2}\left(\frac{S}{N}\right) + \ln \frac{\pi_1}{\pi_2},$$

where

$$\left(\frac{S}{N}\right) = \frac{\sum_{t=1}^{n} s_t^2}{\sigma_w^2}$$

denotes the signal-to-noise ratio. Give the decision criterion if the prior probabilities are assumed to be the same. Express the false alarm and missed signal probabilities in terms of the normal cdf and the signal-to-noise ratio.

7.14 Assume the same additive signal plus noise representations as in the previous problem, except, the signal is now a random process with a zero mean and covariance matrix $\sigma_s^2 I$. Derive the comparable version of (7.116) as a quadratic detector, and characterize its performance under both hypotheses in terms of constant multiples of the chi-squared distribution.

Section 7.8

7.15 The data set `ch5fmri.dat` contains data from other stimulus conditions in the fMRI experiment, as discussed in Example 7.14 (one location—Caudate—was left out of the analysis for brevity). Perform principal component analyses on the stimulus conditions (i) awake-heat and (ii) awake-shock, and compare your results to the results of Example 7.14.

7.16 For this problem, consider the first three earthquake series listed in `eq+exp.dat`.

(a) Estimate and compare the spectral density of the P component and then of the S component for each individual earthquake.

(b) Estimate and compare the squared coherency between the P and S components of each individual earthquake. Comment on the strength of the coherence.

(c) Let x_{ti} be the P component of earthquake $i = 1, 2, 3$, and let $\boldsymbol{x}_t = (x_{t1}, x_{t2}, x_{t3})'$ be the 3×1 vector of P components. Estimate the spectral density, $\lambda_1(\omega)$, of the first principal component series of $\boldsymbol{x}_t$. Compare this to the corresponding spectra calculated in (a).

(d) Analogous to part (c), let $\boldsymbol{y}_t$ denote the 3×1 vector series of S components of the first three earthquakes. Repeat the analysis of part (c) on $\boldsymbol{y}_t$.

7.17 In the factor analysis model (7.155), let $p = 3$, $q = 1$, and

$$\Sigma_{xx} = \begin{bmatrix} 1 & .4 & .9 \\ .4 & 1 & .7 \\ .9 & .7 & 1 \end{bmatrix}.$$

Show there is a unique choice for $\mathcal{B}$ and D, but $\delta_3^2 < 0$, so the choice is not valid.

7.18 Extend the EM algorithm for classical factor analysis, (7.161)-(7.166), to the time series case of maximizing $\ln L\big(\mathcal{B}(\omega_j), D_{\epsilon\epsilon}(\omega_j)\big)$ in (7.177). Then, for the data used in Example 7.16, find the approximate maximum likelihood estimates of $\mathcal{B}(\omega_j)$ and $D_{\epsilon\epsilon}(\omega_j)$, and, consequently, Λ_t.

Section 7.9

7.19 Verify, as stated in (7.182), the imaginary part of a $k \times k$ spectral matrix, $f^{im}(\omega)$, is skew symmetric, and then show $\boldsymbol{\beta}' f_{yy}^{im}(\omega)\boldsymbol{\beta} = 0$ for a real $k \times 1$ vector, $\boldsymbol{\beta}$.

7.20 Repeat the analysis of Example 7.18 on BNRF1 of herpesvirus saimiri (the data file is `bnrf1hvs.dat`), and compare the results with the results obtained for Epstein–Barr.

7.21 For the NYSE returns, say, r_t, analyzed in Chapter 5, Example 5.4:

(a) Estimate the spectrum of the r_t. Does the spectral estimate appear to support the hypothesis that the returns are white?

(b) Examine the possibility of spectral power near the zero frequency for a transformation of the returns, say, $g(r_t)$, using the spectral envelope with Example 7.19 as your guide. Compare the optimal transformation near or at the zero frequency with the usual transformation $y_t = r_t^2$.

Appendix A
Large Sample Theory

A.1 Convergence Modes

The study of the optimality properties of various estimators (such as the sample autocorrelation function) depends, in part, on being able to assess the large-sample behavior of these estimators. We summarize briefly here the kinds of convergence useful in this setting, namely, mean square convergence, convergence in probability, and convergence in distribution.

We consider first a particular class of random variables that plays an important role in the study of second-order time series, namely, the class of random variables belonging to the space L^2, satisfying $E|x|^2 < \infty$. In proving certain properties of the class L^2 we will often use, for random variables $x, y \in L^2$, the Cauchy–Schwarz inequality,

$$|E(xy)|^2 \leq E(|x|^2)E(|y|^2), \tag{A.1}$$

and the Tchebycheff inequality,

$$P\{|x| \geq a\} \leq \frac{E(|x|^2)}{a^2}, \tag{A.2}$$

for $a > 0$.

Next, we investigate the properties of mean square convergence of random variables in L^2.

Definition A.1 *A sequence of L^2 random variables $\{x_n\}$, is said to converge in* **mean square** *to a random variable $x \in L^2$, denoted by*

$$x_n \overset{ms}{\to} x, \tag{A.3}$$

if and only if

$$E|x_n - x|^2 \to 0 \tag{A.4}$$

as $n \to \infty$.

Example A.1 Mean Square Convergence of the Sample Mean

Consider the white noise sequence w_t and the *signal plus noise* series

$$x_t = \mu + w_t.$$

Then, because

$$E|\bar{x}_n - \mu|^2 = \frac{\sigma_w^2}{n} \to 0$$

as $n \to \infty$, where $\bar{x}_n = n^{-1} \sum_{t=1}^{n} x_t$ is the sample mean, we have $\bar{x}_n \overset{ms}{\to} \mu$.

We summarize some of the properties of mean square convergence as follows. If $x_n \overset{ms}{\to} x$, and $y_n \overset{ms}{\to} y$, then, as $n \to \infty$,

$$(i) \qquad E(x_n) \to E(x); \tag{A.5}$$

$$(ii) \qquad E(|x_n|^2) \to E(|x|^2); \tag{A.6}$$

$$(iii) \qquad E(x_n y_n) \to E(xy). \tag{A.7}$$

We also note the L^2 completeness theorem known as the *Riesz–Fisher Theorem.*

Theorem A.1 *Let $\{x_n\}$ be a sequence in L^2. Then, there exists an x in L^2 such that $x_n \overset{ms}{\to} x$ if and only if*

$$E|x_n - x_m|^2 \to 0 \tag{A.8}$$

for $m, n \to \infty$.

Often the condition of Theorem A.1 is easier to verify to establish that a mean square limit x exists without knowing what it is. Sequences that satisfy (A.8) are said to be Cauchy sequences in L^2 and (A.8) is also known as the Cauchy criterion for L^2.

Example A.2 Time Invariant Linear Filter

As an important example of the use of the Riesz–Fisher Theorem A.1 and the properties (i), (ii), and (iii) of mean square convergent series given in (A.5)–(A.7), a time-invariant linear filter is defined as a convolution of the form

$$y_t = \sum_{j=-\infty}^{\infty} a_j x_{t-j} \tag{A.9}$$

for each $t = 0, \pm 1, \pm 2, \ldots$, where x_t is a weakly stationary input series with mean μ_x and autocovariance function $\gamma_x(h)$, and a_j, for $j = 0, \pm 1, \pm 2, \ldots$ are constants satisfying

$$\sum_{j=-\infty}^{\infty} |a_j| < \infty. \tag{A.10}$$

The output series y_t defines a filtering or smoothing of the input series that changes the character of the time series in a predictable way. We need to know the conditions under which the outputs y_t in (A.9) and the linear process (1.31) exist.

Considering the sequence

$$y_t^n = \sum_{j=-n}^{n} a_j x_{t-j}, \tag{A.11}$$

$n = 1, 2, \ldots$, we need to show first that y_t^n has a mean square limit. By Theorem A.1, it is enough to show that

$$E\,|y_t^n - y_t^m|^2 \to 0$$

as $m, n \to \infty$. For $n > m > 0$,

$$\begin{aligned} E\,|y_t^n - y_t^m|^2 &= E\left|\sum_{m<|j|\le n} a_j x_{t-j}\right|^2 \\ &= \sum_{m<|j|\le n}\ \sum_{m\le|k|\le n} a_j a_k E(x_{t-j}x_{t-k}) \\ &\le \sum_{m<|j|\le n}\ \sum_{m\le|k|\le n} |a_j||a_k||E(x_{t-j}x_{t-k})| \\ &\le \sum_{m<|j|\le n}\ \sum_{m\le|k|\le n} |a_j||a_k|(E|x_{t-j}|^2)^{1/2}(E|x_{t-k}|^2)^{1/2} \\ &= \gamma_x(0)\left(\sum_{m\le|j|\le n} |a_j|\right)^2 \to 0 \end{aligned}$$

as $m, n \to \infty$, because $\gamma_x(0)$ is a constant and $\{a_j\}$ is absolutely summable (the second inequality follows from the Cauchy–Schwarz inequality).

Although we know that the sequence $\{y_t^n\}$ given by (A.11) converges in mean square, we have not established its mean square limit. It should be obvious, however, that $y_t^n \overset{ms}{\to} y_t$ as $n \to \infty$, where y_t is given by (A.9).[1]

Finally, we may use (A.5) and (A.7) to establish the mean, μ_y and autocovariance function, $\gamma_y(h)$ of y_t. In particular we have,

$$\mu_y = \mu_x \sum_{j=-\infty}^{\infty} a_j, \tag{A.12}$$

[1]If S denotes the mean square limit of y_t^n, then using Fatou's Lemma, $E|S - y_t|^2 = E\liminf_{n\to\infty}|S - y_t^n|^2 \le \liminf_{n\to\infty} E|S - y_t^n|^2 = 0$, which establishes that y_t is the mean square limit of y_t^n.

and

$$\begin{aligned}\gamma_y(h) &= E\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} a_j(x_{t+h-j}-\mu_x)a_j(x_{t-k}-\mu_x) \\ &= \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} a_j\gamma_x(h-j+k)a_k \end{aligned} \tag{A.13}$$

A second important kind of convergence is convergence in probability.

Definition A.2 *The sequence* $\{x_n\}$*, for* $n = 1, 2, \ldots$*,* **converges in probability** *to a random variable* x*, denoted by*

$$x_n \xrightarrow{p} x, \tag{A.14}$$

if and only if

$$P\{|x_n - x| > \epsilon\} \to 0 \tag{A.15}$$

for all $\epsilon > 0$*, as* $n \to \infty$*.*

An immediate consequence of the Tchebycheff inequality, (A.2), is that

$$P\{|x_n - x| \geq \epsilon\} \leq \frac{E(|x_n - x|^2)}{\epsilon^2},$$

so convergence in mean square implies convergence in probability, i.e.,

$$x_n \xrightarrow{ms} x \;\Rightarrow\; x_n \xrightarrow{p} x. \tag{A.16}$$

This result implies, for example, that the filter (A.9) exists as a limit in probability because it converges in mean square [it is also easily established that (A.9) exists with probability one]. We mention, at this point, the useful Weak Law of Large Numbers which states that, for an independent identically distributed sequence x_n of random variables with mean μ, we have

$$\bar{x}_n \xrightarrow{p} \mu \tag{A.17}$$

as $n \to \infty$, where $\bar{x}_n = n^{-1}\sum_{t=1}^{n} x_t$ is the usual sample mean.

We also will make use of the following concepts.

Definition A.3 *For* **order in probability** *we write*

$$x_n = o_p(a_n) \tag{A.18}$$

if and only if

$$\frac{x_n}{a_n} \xrightarrow{p} 0. \tag{A.19}$$

The term **boundedness in probability**, *written* $x_n = O_p(a_n)$*, means that for every* $\epsilon > 0$*, there exists a* $\delta(\epsilon) > 0$ *such that*

$$P\left\{\left|\frac{x_n}{a_n}\right| > \delta(\epsilon)\right\} \leq \epsilon \tag{A.20}$$

for all n*.*

Under this convention, e.g., the notation for $x_n \xrightarrow{p} x$ becomes $x_n - x = o_p(1)$. The definitions can be compared with their nonrandom counterparts, namely, for a fixed sequence $x_n = o(1)$ if $x_n \to 0$ and $x_n = O(1)$ if x_n, for $n = 1, 2, \ldots$ is bounded. Some handy properties of $o_p(\cdot)$ and $O_p(\cdot)$ are as follows.

(i) If $x_n = o_p(a_n)$ and $y_n = o_p(b_n)$, then $x_n y_n = o_p(a_n b_n)$ and $x_n + y_n = o_p(\max(a_n, b_n))$.

(ii) If $x_n = o_p(a_n)$ and $y_n = O_p(b_n)$, then $x_n y_n = o_p(a_n b_n)$.

(iii) Statement (i) is true if $O_p(\cdot)$ replaces $o_p(\cdot)$.

Example A.3 Convergence and Order in Probability for the Sample Mean

For the sample mean, $\bar{x}_n$, of iid random variables with mean μ and variance σ^2, by the Tchebycheff inequality,

$$\begin{aligned} P\{|\bar{x}_n - \mu| > \epsilon\} &\leq \frac{E[(\bar{x}_n - \mu)^2]}{\epsilon^2} \\ &= \frac{\sigma^2}{n\epsilon^2} \to 0, \end{aligned}$$

as $n \to \infty$. It follows that $\bar{x}_n \xrightarrow{p} \mu$, or $\bar{x}_n - \mu = o_p(1)$. To find the rate, it follows that, for $\delta(\epsilon) > 0$,

$$P\left\{\sqrt{n}\ |\bar{x}_n - \mu| > \delta(\epsilon)\right\} \leq \frac{\sigma^2/n}{\delta^2(\epsilon)/n} = \frac{\sigma^2}{\delta^2(\epsilon)}$$

by Tchebycheff's inequality, so taking $\epsilon = \sigma^2/\delta^2(\epsilon)$ shows that $\delta(\epsilon) = \sigma/\sqrt{\epsilon}$ does the job and

$$\bar{x}_n - \mu = O_p(n^{-1/2}).$$

For $k \times 1$ random vectors $\boldsymbol{x}_n$, convergence in probability, written $\boldsymbol{x}_n \xrightarrow{p} \boldsymbol{x}$ or $\boldsymbol{x}_n - \boldsymbol{x} = o_p(1)$ is defined as element-by-element convergence in probability, or equivalently, as convergence in terms of the Euclidean distance

$$\|\boldsymbol{x}_n - \boldsymbol{x}\| \xrightarrow{p} 0, \tag{A.21}$$

where $\|\boldsymbol{a}\| = \sum_j a_j^2$ for any vector $\boldsymbol{a}$. In this context, we note the result that if $\boldsymbol{x}_n \xrightarrow{p} \boldsymbol{x}$ and $g(\boldsymbol{x}_n)$ is a continuous mapping,

$$g(\boldsymbol{x}_n) \xrightarrow{p} g(\boldsymbol{x}). \tag{A.22}$$

Furthermore, if $\boldsymbol{x}_n - \boldsymbol{a} = O_p(\delta_n)$ with $\delta_n \to 0$ and $g(\cdot)$ is a function with continuous first derivatives continuous in a neighborhood of $\boldsymbol{a} = (a_1, a_2, \ldots, a_k)'$, we have the Taylor series expansion in probability

$$g(\boldsymbol{x}_n) = g(\boldsymbol{a}) + \frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}}\bigg|'_{\boldsymbol{x}=\boldsymbol{a}} (\boldsymbol{x}_n - \boldsymbol{a}) + O_p(\delta_n), \tag{A.23}$$

where

$$\frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}}\bigg|_{\boldsymbol{x}=\boldsymbol{a}} = \left(\frac{\partial g(\boldsymbol{x})}{\partial x_1}\bigg|_{\boldsymbol{x}=\boldsymbol{a}}, \ldots, \frac{\partial g(\boldsymbol{x})}{\partial x_k}\bigg|_{\boldsymbol{x}=\boldsymbol{a}} \right)'$$

denotes the vector of partial derivatives with respect to $x_1, x_2, \ldots, x_k$, evaluated at $\boldsymbol{a}$. This result remains true if $O_p(\delta_n)$ is replaced everywhere by $o_p(\delta_n)$.

Example A.4 Expansion for the Logarithm of the Sample Mean

With the same conditions as Example A.3, consider $g(\bar{x}_n) = \log \bar{x}_n$, which has a derivative at μ, for $\mu > 0$. Then, because $\bar{x}_n - \mu = O_p(n^{-1/2})$ from Example A.3, the conditions for the Taylor expansion in probability, (A.23), are satisfied and we have

$$\log \bar{x}_n = \log \mu + \mu^{-1}(\bar{x}_n - \mu) + O_p(n^{-1/2}).$$

The large sample distributions of sample mean and sample autocorrelation functions defined earlier can be developed using the notion of convergence in distribution.

Definition A.4 *A sequence of $k \times 1$ random vectors $\{\boldsymbol{x}_n\}$ is said to* **converge in distribution**, *written*

$$\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{x} \tag{A.24}$$

if and only if

$$F_n(\boldsymbol{x}) \to F(\boldsymbol{x}) \tag{A.25}$$

at the continuity points of distribution function $F(\cdot)$.

Example A.5 Convergence in Distribution

Consider a sequence $\{x_n\}$ of iid normal random variables with mean zero and variance $1/n$. Now, using the normal cdf (1.10), we have $F_n(x) = \Phi(\sqrt{n}x)$, so

$$F_n(x) \to \begin{cases} 0 & x < 0, \\ 1/2, & x = 0 \\ 1, & x > 0 \end{cases}$$

and we may take

$$F(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0, \end{cases}$$

because the point where the two functions differ is not a continuity point of $F(x)$.

The distribution function relates uniquely to the characteristic function through the Fourier transform, defined as a function with vector argument $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_k)'$, say

$$\begin{aligned} \phi(\boldsymbol{\lambda}) &= E(\exp\{i\boldsymbol{\lambda}'\boldsymbol{x}\}) \\ &= \int \exp\{i\boldsymbol{\lambda}'\boldsymbol{x}\}\, dF(\boldsymbol{x}). \end{aligned} \tag{A.26}$$

Hence, for a sequence $\{\boldsymbol{x}_n\}$ we may characterize convergence in distribution of $F_n(\cdot)$ in terms of convergence of the sequence of characteristic functions $\phi_n(\cdot)$, i.e.,

$$\phi_n(\boldsymbol{\lambda}) \to \phi(\boldsymbol{\lambda}) \;\Leftrightarrow\; F_n(\boldsymbol{x}) \xrightarrow{d} F(\boldsymbol{x}), \tag{A.27}$$

where $\Leftrightarrow$ means that the implication goes both directions. In this connection, the Cramér–Wold device says that for every $\boldsymbol{c} = (c_1, c_2, \ldots, c_k)'$

$$\boldsymbol{c}'\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{c}'\boldsymbol{x} \;\Leftrightarrow \boldsymbol{x}_n \xrightarrow{d} \boldsymbol{x}. \tag{A.28}$$

Also, convergence in probability implies convergence in distribution, namely,

$$\boldsymbol{x}_n \xrightarrow{p} \boldsymbol{x} \;\Rightarrow\; \boldsymbol{x}_n \xrightarrow{d} \boldsymbol{x}, \tag{A.29}$$

but the converse is only true when $\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{c}$, where $\boldsymbol{c}$ is a constant vector. If $\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{x}$ and $\boldsymbol{y}_n \xrightarrow{d} \boldsymbol{c}$ are two sequences of random vectors and $\boldsymbol{c}$ is a constant vector,

$$\boldsymbol{x}_n + \boldsymbol{y}_n \xrightarrow{d} \boldsymbol{x} + \boldsymbol{c} \tag{A.30}$$

and

$$\boldsymbol{y}_n'\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{c}'\boldsymbol{x}. \tag{A.31}$$

For a continuous mapping $h(\boldsymbol{x})$,

$$\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{x} \;\Rightarrow\; h(\boldsymbol{x}_n) \xrightarrow{d} h(\boldsymbol{x}). \tag{A.32}$$

A number of results in time series depend on making a series of approximations to prove convergence in distribution. For example, we have that if $\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{x}$ can be *approximated* by the sequence $\boldsymbol{y}_n$ in the sense that

$$\boldsymbol{y}_n - \boldsymbol{x}_n = o_p(1), \tag{A.33}$$

then we have that $\boldsymbol{y}_n \xrightarrow{d} \boldsymbol{x}$, so the approximating sequence $\boldsymbol{y}_n$ has the same limiting distribution as $\boldsymbol{x}$. We present the following Basic Approximation Theorem (BAT) that will be used later to derive asymptotic distributions for the sample mean and ACF.

Theorem A.2 *Let $\boldsymbol{x}_n$ for $n = 1, 2, \ldots,$ and $\boldsymbol{y}_{mn}$ for $m = 1, 2, \ldots,$ be random $k \times 1$ vectors such that*

(i) $\boldsymbol{y}_{mn} \xrightarrow{d} \boldsymbol{y}_m$ *as* $n \to \infty$ *for each* m*;*

(ii) $\boldsymbol{y}_m \xrightarrow{d} \boldsymbol{y}$ *as* $m \to \infty$*;*

(iii) $\lim_{m\to\infty} \limsup_{n\to\infty} P\{|\boldsymbol{x}_n - \boldsymbol{y}_{mn}| > \epsilon\} = 0$ *for every* $\epsilon > 0$.

Then, $\boldsymbol{x}_n \xrightarrow{d} \boldsymbol{y}$.

As a practical matter, condition (*iii*) is implied by the Tchebycheff inequality if

$$(iii') \qquad E\{|\boldsymbol{x}_n - \boldsymbol{y}_{mn}|^2\} \to 0 \tag{A.34}$$

as $m, n \to \infty$, and (*iii'*) is often much easier to establish than (*iii*).

The theorem allows approximation of the underlying sequence in two steps, through the intermediary sequence $\boldsymbol{y}_{mn}$, depending on two arguments. In the time series case, n is generally the sample length and m is generally the number of terms in an approximation to the linear process of the form (A.11).

Proof. The proof of the theorem is a simple exercise in using the characteristic functions and appealing to (A.27). We need to show

$$|\phi_{\boldsymbol{x}_n} - \phi_{\boldsymbol{y}}| \to 0,$$

where we use the shorthand notation $\phi \equiv \phi(\boldsymbol{\lambda})$ for ease. First,

$$|\phi_{\boldsymbol{x}_n} - \phi_{\boldsymbol{y}}| \le |\phi_{\boldsymbol{x}_n} - \phi_{\boldsymbol{y}_{mn}}| + |\phi_{\boldsymbol{y}_{mn}} - \phi_{\boldsymbol{y}_m}| + |\phi_{\boldsymbol{y}_m} - \phi_{\boldsymbol{y}}|. \tag{A.35}$$

By the condition (*ii*) and (A.27), the last term converges to zero, and by condition (*i*) and (A.27), the second term converges to zero and we only need consider the first term in (A.35). Now, write

$$\begin{aligned}
\left|\phi_{\boldsymbol{x}_n} - \phi_{\boldsymbol{y}_{mn}}\right| &= \left|E(e^{i\boldsymbol{\lambda}'\boldsymbol{x}_n} - e^{i\boldsymbol{\lambda}'\boldsymbol{y}_{mn}})\right| \\
&\le E\left|e^{i\boldsymbol{\lambda}'\boldsymbol{x}_n}\left(1 - e^{i\boldsymbol{\lambda}'(\boldsymbol{y}_{mn}-\boldsymbol{x}_n)}\right)\right| \\
&= E\left|1 - e^{i\boldsymbol{\lambda}'(\boldsymbol{y}_{mn}-\boldsymbol{x}_n)}\right| \\
&= E\left\{\left|1 - e^{i\boldsymbol{\lambda}'(\boldsymbol{y}_{mn}-\boldsymbol{x}_n)}\right| I\{|\boldsymbol{y}_{mn} - \boldsymbol{x}_n| < \delta\}\right\} \\
&\quad + E\left\{\left|1 - e^{i\boldsymbol{\lambda}'(\boldsymbol{y}_{mn}-\boldsymbol{x}_n)}\right| I\{|\boldsymbol{y}_{mn} - \boldsymbol{x}_n| \ge \delta\}\right\},
\end{aligned}$$

where $\delta > 0$ and $I\{A\}$ denotes the indicator function of the set A. Then, given $\boldsymbol{\lambda}$ and $\epsilon > 0$, choose $\delta(\epsilon) > 0$ such that

$$\left|1 - e^{i\boldsymbol{\lambda}'(\boldsymbol{y}_{mn}-\boldsymbol{x}_n)}\right| < \epsilon$$

if $|\boldsymbol{y}_{mn} - \boldsymbol{x}_n| < \delta$, and the first term is less than ϵ, an arbitrarily small constant. For the second term, note that

$$\left|1 - e^{i\boldsymbol{\lambda}'(\boldsymbol{y}_{mn} - \boldsymbol{x}_n)}\right| \leq 2$$

and we have

$$E\left\{\left|1 - e^{i\boldsymbol{\lambda}'(\boldsymbol{y}_{mn} - \boldsymbol{x}_n)}\right| I\{|\boldsymbol{y}_{mn} - \boldsymbol{x}_n| \geq \delta\}\right\} \leq 2P\{|\boldsymbol{y}_{mn} - \boldsymbol{x}_n| \geq \delta\},$$

which converges to zero as $n \to \infty$ by property (iii). ∎

A.2 Central Limit Theorems

We will generally be concerned with the large-sample properties of estimators that turn out to be normally distributed as $n \to \infty$.

Definition A.5 *A sequence of random variables $\{x_n\}$ is said to be* **asymptotically normal** *with mean μ_n and variance σ_n^2 if, as $n \to \infty$,*

$$\sigma_n^{-1}(x_n - \mu_n) \xrightarrow{d} z,$$

where z has the standard normal distribution. We shall abbreviate this as

$$x_n \sim AN(\mu_n, \sigma_n^2), \tag{A.36}$$

where $\sim$ will denote is distributed as.

We state the important Central Limit Theorem, as follows.

Theorem A.3 *Let $x_1, \ldots, x_n$ be independent and identically distributed with mean μ and variance σ^2. If $\bar{x}_n = (x_1 + \cdots + x_n)/n$ denotes the sample mean, then*

$$\bar{x}_n \sim AN(\mu, \sigma^2/n). \tag{A.37}$$

Often, we will be concerned with a sequence of $k \times 1$ vectors $\{\boldsymbol{x}_n\}$. The following definition is motivated by the Cramér–Wold device considered earlier.

Definition A.6 *We define* **asymptotic normality** *for the vector case as*

$$\boldsymbol{x}_n \sim AN(\boldsymbol{\mu}_n, \Sigma_n) \tag{A.38}$$

if and only if

$$\boldsymbol{c}'\boldsymbol{x}_n \sim AN(\boldsymbol{c}'\boldsymbol{\mu}_n, \boldsymbol{c}'\Sigma_n\boldsymbol{c}) \tag{A.39}$$

for all $\boldsymbol{c}$ and Σ_n is positive definite.

In order to begin to consider what happens for dependent data in the limiting case, it is necessary to define, first of all, a particular kind of dependence known as M-dependence. We say that a time series x_t is M-dependent if the set of values $x_s, s \leq t$ is independent of the set of values $x_s, s \geq t + M + 1$, so time points separated by more than M units are independent. A central limit theorem for such dependent processes, used in conjunction with the Basic Approximation Theorem, will allow us to develop large-sample distributional results for the sample mean $\bar{x}$ and the sample ACF $\widehat{\rho}_x(h)$ in the stationary case.

In the arguments that follow, we often make use of the formula for the variance of $\bar{x}_n$ in the stationary case, namely,

$$\text{var } \bar{x}_n = n^{-1} \sum_{u=-(n-1)}^{(n-1)} \left(1 - \frac{|u|}{n}\right)\gamma(u). \tag{A.40}$$

To prove the above formula, letting $u = s - t$ and $v = t$ in

$$\begin{aligned} n^2 E[(\bar{x}_n - \mu)^2] &= \sum_{s=1}^{n}\sum_{t=1}^{n} E[(x_s - \mu)(x_t - \mu)] \\ &= \sum_{s=1}^{n}\sum_{t=1}^{n} \gamma(s-t) \\ &= \sum_{u=-(n-1)}^{-1} \sum_{v=-(u-1)}^{n} \gamma(u) + \sum_{u=0}^{n-1}\sum_{v=1}^{n-u} \gamma(u) \\ &= \sum_{u=-(n-1)}^{-1} (n+u)\gamma(u) + \sum_{u=0}^{n-1}(n-u)\gamma(u) \\ &= \sum_{u=-(n-1)}^{(n-1)} (n - |u|)\gamma(u) \end{aligned}$$

gives the required result. We shall also use the fact that, for

$$\sum_{u=-\infty}^{\infty} |\gamma(u)| \ < \infty,$$

we would have, by dominated convergence,[2]

$$n \text{ var } \bar{x}_n \to \sum_{u=-\infty}^{\infty} \gamma(u), \tag{A.41}$$

because $|(1 - |u|/n)\gamma(u)| \leq |\gamma(u)|$ and $(1 - |u|/n)\gamma(u) \to \gamma(u)$. We may now state the M-Dependent Central Limit Theorem as follows.

Theorem A.4 *If x_t is a strictly stationary M-dependent sequence of random variables with mean zero and autocovariance function $\gamma(\cdot)$ and if*

$$V_M = \sum_{u=-M}^{M} \gamma(u), \tag{A.42}$$

where $V_M \neq 0$,

$$\bar{x}_n \sim AN(0, V_M/n). \tag{A.43}$$

Proof. To prove the theorem, using Theorem A.2, the Basic Approximation Theorem, we may construct a sequence of variables y_{mn} approximating

$$n^{1/2}\bar{x}_n = n^{-1/2} \sum_{t=1}^{n} x_t$$

in the dependent case and then simply verify conditions (i), (ii), and (iii) of Theorem A.2. For $m > 2M$, we may first consider the approximation

$$\begin{aligned} y_{mn} &= n^{-1/2}[(x_1 + \cdots + x_{m-M}) + (x_{m+1} + \cdots + x_{2m-M}) \\ &\quad + (x_{2m+1} + \cdots + x_{3m-M}) + \cdots + (x_{(r-1)m+1} + \cdots + x_{rm-M})] \\ &= n^{-1/2}(z_1 + z_2 + \cdots + z_r), \end{aligned}$$

where $r = [n/m]$, with $[n/m]$ denoting the greatest integer less than or equal to n/m. This approximation contains only part of $n^{1/2}\bar{x}_n$, but the random variables $z_1, z_2, \ldots, z_r$ are independent because they are separated by more than M time points, e.g., $m + 1 - (m - M) = M + 1$ points separate z_1 and z_2. Because of strict stationarity, $z_1, z_2, \ldots, z_r$ are identically distributed with zero means and variances

$$S_{m-M} = \sum_{|u| \leq M} (m - M - |u|)\gamma(u)$$

by a computation similar to that producing (A.40). We now verify the conditions of the Basic Approximation Theorem hold.

[2]Dominated convergence technically relates to convergent sequences (with respect to a sigma-additive measure μ) of measurable functions $f_n \to f$ bounded by an integrable function g, $\int g \, d\mu < \infty$. For such a sequence,

$$\int f_n \, d\mu \to \int f \, d\mu.$$

For the case in point, take $f_n(u) = (1 - |u|/n)\gamma(u)$ for $|u| < n$ and as zero for $|u| \geq n$. Take $\mu(u) = 1, u = \pm 1, \pm 2, \ldots$ to be counting measure.

(*i*): Applying the Central Limit Theorem to the sum y_{mn} gives

$$y_{mn} = n^{-1/2} \sum_{i=1}^{r} z_i = (n/r)^{-1/2} r^{-1/2} \sum_{i=1}^{r} z_i.$$

Because $(n/r)^{-1/2} \to m^{1/2}$ and

$$r^{-1/2} \sum_{i=1}^{r} z_i \xrightarrow{d} N(0, S_{m-M}),$$

it follows from (A.31) that

$$y_{mn} \xrightarrow{d} y_m \sim N(0, S_{m-M}/m).$$

as $n \to \infty$, for a fixed m.

(*ii*): Note that as $m \to \infty$, $S_{m-M}/m \to V_M$ using dominated convergence, where V_M is defined in (A.42). Hence, the characteristic function of y_m, say,

$$\phi_m(\lambda) = \exp\left\{ -\frac{1}{2}\lambda^2 \ \frac{S_{m-M}}{m} \right\} \to \exp\left\{ -\frac{1}{2}\lambda^2 \ V_M \right\},$$

as $m \to \infty$, which is the characteristic function of a random variable $y \sim N(0, V_M)$ and the result follows because of (A.27).

(*iii*): To verify the last condition of the BAT theorem,

$$\begin{aligned} n^{1/2}\bar{x}_n - y_{mn} &= n^{-1/2}[(x_{m-M+1} + \cdots + x_m) \\ &\quad + (x_{2m-M+1} + \cdots + x_{2m}) \\ &\quad + (x_{(r-1)m-M+1} + \cdots + x_{(r-1)m}) \\ &\quad \vdots \\ &\quad + (x_{rm-M+1} + \cdots + x_n)] \\ &= n^{-1/2}(w_1 + w_2 + \cdots + w_r), \end{aligned}$$

so the error is expressed as a scaled sum of iid variables with variance S_M for the first $r-1$ variables and

$$\begin{aligned} \text{var}(w_r) &= \textstyle\sum_{|u| \le m-M} \Big(n - [n/m]m + M - |u| \Big) \gamma(u) \\ &\le \textstyle\sum_{|u| \le m-M} (m + M - |u|)\gamma(u). \end{aligned}$$

Hence,

$$\text{var}\ [n^{1/2}\bar{x} - y_{mn}] = n^{-1}[(r-1)S_M + \text{var}\ w_r],$$

which converges to $m^{-1}S_M$ as $n \to \infty$. Because $m^{-1}S_M \to 0$ as $m \to \infty$, the condition of (iii) holds by the Tchebycheff inequality. ∎

A.3 The Mean and Autocorrelation Functions

The background material in the previous two sections can be used to develop the asymptotic properties of the sample mean and ACF used to evaluate statistical significance. In particular, we are interested in verifying Property P1.1.

We begin with the distribution of the sample mean $\bar{x}_n$, noting that (A.41) suggests a form for the limiting variance. In all of the asymptotics, we will use the assumption that x_t is a linear process, as defined in Definition 1.12, but with the added condition that $\{w_t\}$ is iid. That is, throughout this section we assume

$$x_t = \mu_x + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j} \tag{A.44}$$

where $w_t \sim \text{iid}(0, \sigma_w^2)$, and the coefficients satisfy

$$\sum_{j=-\infty}^{\infty} |\psi_j| < \infty. \tag{A.45}$$

Before proceeding further, we should note that the exact sampling distribution of $\bar{x}_n$ is available if the distribution of the underlying vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$ is multivariate normal. Then, $\bar{x}_n$ is just a linear combination of jointly normal variables that will have the normal distribution

$$\bar{x}_n \sim N\left(\mu_x,\ n^{-1} \sum_{|u|<n} \left(1 - \frac{|u|}{n}\right) \gamma_x(u)\right), \tag{A.46}$$

by (A.40). In the case where x_t are not jointly normally distributed, we have the following theorem.

Theorem A.5 *If x_t is a linear process of the form (A.44) and $\sum_j \psi_j \neq 0$, then*

$$\bar{x}_n \sim AN(\mu_x, n^{-1}V), \tag{A.47}$$

where

$$V = \sum_{h=-\infty}^{\infty} \gamma_x(h) = \sigma_w^2 \left(\sum_{j=-\infty}^{\infty} \psi_j\right)^2 \tag{A.48}$$

and $\gamma_x(\cdot)$ is the autocovariance function of x_t.

Proof. To prove the above, we can again use the Basic Approximation Theorem A.2 by first defining the strictly stationary $2m$-dependent linear process with finite limits

$$x_t^m = \sum_{j=-m}^{m} \psi_j w_{t-j}$$

as an approximation to x_t to use in the approximating mean

$$\bar{x}_{n,m} = n^{-1} \sum_{t=1}^{n} x_t^m.$$

Then, take

$$y_{mn} = n^{1/2}(\bar{x}_{n,m} - \mu_x)$$

as an approximation to $n^{1/2}(\bar{x}_n - \mu_x)$.

(*i*): Applying Theorem A.4, we have

$$y_{mn} \xrightarrow{d} y_m \sim N(0, V_m),$$

as $n \to \infty$, where

$$V_m = \sum_{h=-2m}^{2m} \gamma_x(h) = \sigma_w^2 \Bigg(\sum_{j=-m}^{m} \psi_j \Bigg)^2.$$

To verify the above, we note that for the general linear process with infinite limits, (1.33) implies that

$$\sum_{h=-\infty}^{\infty} \gamma_x(h) = \sigma_w^2 \sum_{h=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \psi_{j+h}\psi_j = \sigma_w^2 \Bigg(\sum_{j=-\infty}^{\infty} \psi_j \Bigg)^2,$$

so taking the special case $\psi_j = 0$, for $|j| > m$, we obtain V_m.

(*ii*): Because $V_m \to V$ in (A.48) as $m \to \infty$, we may use the same characteristic function argument as under (ii) in the proof of Theorem A.4 to note that

$$y_m \xrightarrow{d} y \sim N(0, V),$$

where V is given by (A.48).

(*iii*): Finally,

$$\begin{aligned} \operatorname{var}\Big[n^{1/2}(\bar{x}_n - \mu_x) - y_{mn}\Big] &= n \operatorname{var}\left[n^{-1} \sum_{t=1}^{n} \sum_{|j|>m} \psi_j w_{t-j} \right] \\ &= \sigma_w^2 \left(\sum_{|j|>m} \psi_j \right)^2 \to 0 \end{aligned}$$

as $m \to \infty$. ∎

In order to develop the sampling distribution of the sample autocovariance function, $\widehat{\gamma}_x(h)$, and the sample autocorrelation function, $\widehat{\rho}_x(h)$, we need to develop some idea as to the mean and variance of $\widehat{\gamma}_x(h)$ under some reasonable assumptions. These computations for $\widehat{\gamma}_x(h)$ are messy, and we consider a comparable quantity

$$\widetilde{\gamma}_x(h) = n^{-1}\sum_{t=1}^{n}(x_{t+h} - \mu_x)(x_t - \mu_x) \tag{A.49}$$

as an approximation. By Problem 1.29,

$$n^{1/2}[\widetilde{\gamma}_x(h) - \widehat{\gamma}_x(h)] = o_p(1),$$

so that limiting distributional results proved for $n^{1/2}\widetilde{\gamma}_x(h)$ will hold for $n^{1/2}\widehat{\gamma}_x(h)$ by (A.33).

We begin by proving formulas for the variance and for the limiting variance of $\widetilde{\gamma}_x(h)$ under the assumptions that x_t is a linear process of the form (A.44), satisfying (A.45) with the white noise variates w_t having variance σ_w^2 as before, but also required to have fourth moments satisfying

$$E(w_t^4) = \eta\sigma_w^4 < \infty, \tag{A.50}$$

where η is some constant. We seek results comparable with (A.40) and (A.41) for $\widetilde{\gamma}_x(h)$. To ease the notation, we will henceforth drop the subscript x from the notation.

Using (A.49), $E[\widetilde{\gamma}(h)] = \gamma(h)$. Under the above assumptions, we show now that, for $p, q = 0, 1, 2, \ldots$,

$$\text{cov }[\widetilde{\gamma}(p), \widetilde{\gamma}(q)] = n^{-1}\sum_{u=-(n-1)}^{(n-1)}\left(1 - \frac{|u|}{n}\right)V_u, \tag{A.51}$$

where

$$\begin{aligned} V_u &= \gamma(u)\gamma(u+p-q) + \gamma(u+p)\gamma(u-q) \\ &\quad + (\eta-3)\sigma_w^4\sum_i \psi_{i+u+q}\psi_{i+u}\psi_{i+p}\psi_i. \end{aligned} \tag{A.52}$$

The absolute summability of the ψ_j can then be shown to imply the absolute summability of the V_u.[3] Thus, the dominated convergence theorem implies

$$\begin{aligned} n \text{ cov }[\widetilde{\gamma}(p), \widetilde{\gamma}(q)] &\to \sum_{u=-\infty}^{\infty} V_u \\ &= (\eta-3)\gamma(p)\gamma(q) \\ &\quad + \sum_{u=-\infty}^{\infty}\Big[\gamma(u)\gamma(u+p-q) + \gamma(u+p)\gamma(u-q)\Big]. \end{aligned} \tag{A.53}$$

[3]Note: $\sum_{j=-\infty}^{\infty}|a_j| < \infty$ and $\sum_{\mathrm{j}=-\infty}^{\infty}|\mathrm{b_j}| < \infty$ implies $\sum_{j=-\infty}^{\infty}|a_j b_j| < \infty$.

To verify (A.51) is somewhat tedious, so we only go partially through the calculations, leaving the repetitive details to the reader. First, rewrite (A.44) as

$$x_t = \mu + \sum_{i=-\infty}^{\infty} \psi_{t-i} w_i,$$

so that

$$E[\widetilde{\gamma}(p)\widetilde{\gamma}(q)] = n^{-2} \sum_{s,t} \sum_{i,j,k,\ell} \psi_{s+p-i}\psi_{s-j}\psi_{t+q-k}\ \psi_{t-\ell} E(w_i w_j w_k w_\ell).$$

Then, evaluate, using the easily verified properties of the w_t series

$$E(w_i w_j w_k w_\ell) = \begin{cases} \eta\sigma_w^4 & \text{if } i = j = k = \ell \\ \sigma_w^4 & \text{if } i = j \neq k = \ell \\ 0 & \text{if } i \neq j, i \neq k \text{ and } i \neq \ell. \end{cases}$$

To apply the rules, we break the sum over the subscripts i, j, k, ℓ into four terms, namely,

$$\begin{aligned} \sum_{i,j,k,\ell} &= \sum_{i=j=k=\ell} + \sum_{i=j\neq k=\ell} + \sum_{i=k\neq j=\ell} + \sum_{i=\ell\neq j=k} \\ &= S_1 + S_2 + S_3 + S_4. \end{aligned}$$

Now,

$$\begin{aligned} S_1 &= \eta\sigma_w^4 \sum_i \psi_{s+p-i}\psi_{s-i}\psi_{t+q-i}\psi_{t-i} \\ &= \eta\sigma_w^4 \sum_i \psi_{i+s-t+p}\psi_{i+s-t}\psi_{i+q}\psi_i, \end{aligned}$$

where we have let $i' = t - i$ to get the final form. For the second term,

$$\begin{aligned} S_2 &= \sum_{i=j\neq k=\ell} \psi_{s+p-i}\psi_{s-j}\psi_{t+q-k}\psi_{t-\ell} E(w_i w_j w_k w_\ell) \\ &= \sum_{i\neq k} \psi_{s+p-i}\psi_{s-i}\psi_{t+q-k}\psi_{t-k} E(w_i^2) E(w_k^2). \end{aligned}$$

Then, using the fact that

$$\sum_{i\neq k} = \sum_{i,k} - \sum_{i=k},$$

we have

$$\begin{aligned} S_2 &= \sigma_w^4 \sum_{i,k} \psi_{s+p-i}\psi_{s-i}\psi_{t+q-k}\psi_{t-k} - \sigma_w^4 \sum_i \psi_{s+p-i}\psi_{s-i}\psi_{t+q-i}\psi_{t-i} \\ &= \gamma(p)\gamma(q) - \sigma_w^4 \sum_i \psi_{i+s-t+p}\psi_{i+s-t}\psi_{i+q}\psi_i, \end{aligned}$$

letting $i' = s - i, k' = t - k$ in the first term and $i' = s - i$ in the second term. Repeating the argument for S_3 and S_4 and substituting into the covariance expression yields

$$\begin{aligned} E[\widetilde{\gamma}(p)\widetilde{\gamma}(q)] &= n^{-2}\sum_{s,t}\Big[\gamma(p)\gamma(q) + \gamma(s-t)\gamma(s-t+p-q) \\ &\quad + \gamma(s-t+p)\gamma(s-t-q) \\ &\quad + (\eta-3)\sigma_w^4\sum_i \psi_{i+s-t+p}\psi_{i+s-t}\psi_{i+q}\psi_i\Big]. \end{aligned}$$

Then, letting $u = s - t$ and subtracting $E[\widetilde{\gamma}(p)]E[\widetilde{\gamma}(q)] = \gamma(p)\gamma(q)$ from the summation leads to the result (A.52). Summing (A.52) over u and applying dominated convergence leads to (A.53).

The above results for the variances and covariances of the approximating statistics $\widetilde{\gamma}(\cdot)$ enable proving the following central limit theorem for the autocovariance functions $\widehat{\gamma}(\cdot)$.

Theorem A.6 *If x_t is a stationary linear process of the form (A.44) satisfying the fourth moment condition (A.50), then, for fixed K,*

$$\begin{pmatrix} \widehat{\gamma}(0) \\ \widehat{\gamma}(1) \\ \vdots \\ \widehat{\gamma}(K) \end{pmatrix} \sim AN\left[\begin{pmatrix} \gamma(0) \\ \gamma(1) \\ \vdots \\ \gamma(K) \end{pmatrix}, n^{-1}V\right],$$

where V is the matrix with elements given by

$$\begin{aligned} v_{pq} &= (\eta-3)\gamma(p)\gamma(q) \\ &\quad + \sum_{u=-\infty}^{\infty}\Big[\gamma(u)\gamma(u-p+q) + \gamma(u+q)\gamma(u-p)\Big]. \end{aligned} \tag{A.54}$$

Proof. It suffices to show the result for the approximate autocovariance (A.49) for $\widetilde{\gamma}(\cdot)$ by the remark given below it (see also Problem 1.29). First, define the strictly stationary $(2m + K)$-dependent $(K + 1) \times 1$ vector

$$\boldsymbol{y}_t^m = \begin{pmatrix} (x_t^m - \mu)^2 \\ (x_{t+1}^m - \mu)(x_t^m - \mu) \\ \vdots \\ (x_{t+K}^m - \mu)(x_t^m - \mu) \end{pmatrix},$$

where

$$x_t^m = \mu + \sum_{j=-m}^{m}\psi_j w_{t-j}$$

is the usual approximation. The sample mean of the above vector is

$$\bar{\boldsymbol{y}}_{mn} = n^{-1}\sum_{t=1}^{n}\boldsymbol{y}_t^m = \begin{pmatrix} \widetilde{\gamma}^{mn}(0) \\ \widetilde{\gamma}^{mn}(1) \\ \vdots \\ \widetilde{\gamma}^{mn}(K) \end{pmatrix},$$

where

$$\widetilde{\gamma}^{mn}(h) = n^{-1}\sum_{t=1}^{n}(x_{t+h}^m - \mu)(x_t^m - \mu)$$

denotes the sample autocovariance of the approximating series. Also,

$$E\boldsymbol{y}_t^m = \begin{pmatrix} \gamma^m(0) \\ \gamma^m(1) \\ \vdots \\ \gamma^m(K) \end{pmatrix},$$

where $\gamma^m(h)$ is the theoretical covariance function of the series x_t^m. Then, consider the vector

$$\boldsymbol{y}_{mn} = n^{1/2}[\bar{\boldsymbol{y}}_{mn} - E(\bar{\boldsymbol{y}}_{mn})]$$

as an approximation to

$$\boldsymbol{y}_n = n^{1/2}\left[\begin{pmatrix} \tilde{\gamma}(0) \\ \tilde{\gamma}(1) \\ \vdots \\ \tilde{\gamma}(K) \end{pmatrix} - \begin{pmatrix} \gamma(0) \\ \gamma(1) \\ \vdots \\ \gamma(K) \end{pmatrix}\right],$$

where $E(\bar{\boldsymbol{y}}_{mn})$ is the same as $E(\boldsymbol{y}_t^m)$ given above. The elements of the vector approximation $\boldsymbol{y}_{mn}$ are clearly $n^{1/2}(\tilde{\gamma}^{mn}(h) - \tilde{\gamma}^m(h))$. Note that the elements of $\boldsymbol{y}_n$ are based on the linear process x_t, whereas the elements of $\boldsymbol{y}_{mn}$ are based on the m-dependent linear process x_t^m. To obtain a limiting distribution for $\boldsymbol{y}_n$, we apply the Basic Approximation Theorem A.2 using $\boldsymbol{y}_{mn}$ as our approximation. We now verify (i), (ii), and (iii) of Theorem A.2.

(i): First, let $\boldsymbol{c}$ be a $(K+1)\times 1$ vector of constants, and apply the central limit theorem to the $(2m+K)-$dependent series $\boldsymbol{c}'\boldsymbol{y}_{mn}$ using the Cramér–Wold device (A.28). We obtain

$$\boldsymbol{c}'\boldsymbol{y}_{mn} = n^{1/2}\boldsymbol{c}'[\bar{\boldsymbol{y}}_{mn} - E(\bar{\boldsymbol{y}}_{mn})] \xrightarrow{d} \boldsymbol{c}'\boldsymbol{y}_m \sim N(0, \boldsymbol{c}'V_m\boldsymbol{c}),$$

as $n \to \infty$, where V_m is a matrix containing the finite analogs of the elements v_{pq} defined in (A.54).

(ii): Note that, since $V_m \to V$ as $m \to \infty$, it follows that

$$\boldsymbol{c}'\boldsymbol{y}_m \xrightarrow{d} \boldsymbol{c}'\boldsymbol{y} \sim N(0, \boldsymbol{c}'V\boldsymbol{c}),$$

so, by the Cramér–Wold device, the limiting $(K+1)\times 1$ multivariate normal variable is $N(\boldsymbol{0}, V)$.

(*iii*): To show condition (*iii*) of the Basic Approximation Theorem, we can focus on the element-by-element components of

$$P\{|\boldsymbol{y}_n - \boldsymbol{y}_{mn}| > \epsilon\}.$$

For example, using the Tchebycheff inequality, the h-th element of the probability statement can be bounded by

$$n\epsilon^{-2}\text{var}\,(\tilde{\gamma}(h) - \tilde{\gamma}^m(h))$$

$$= \epsilon^{-2}\left\{n\,\text{var}\,\tilde{\gamma}(h) + n\,\text{var}\,\tilde{\gamma}^m(h) - 2n\,\text{cov}[\tilde{\gamma}(h), \tilde{\gamma}^m(h)]\right\}.$$

Using the results that led to (A.53), we see that the preceding expression approaches

$$(v_{hh} + v_{hh} - 2v_{hh})/\epsilon^2 = 0,$$

as $m, n \to \infty$. ∎

To obtain a result comparable to Theorem A.6 for the autocorrelation function ACF, we note the following theorem.

Theorem A.7 *If x_t is a stationary linear process of the form (1.31) satisfying the fourth moment condition (A.50), then for fixed K,*

$$\begin{pmatrix} \widehat{\rho}(1) \\ \vdots \\ \widehat{\rho}(K) \end{pmatrix} \sim AN\left[\begin{pmatrix} \rho(1) \\ \vdots \\ \rho(K) \end{pmatrix}, n^{-1}W\right],$$

where W is the matrix with elements given by

$$\begin{aligned} w_{pq} &= \sum_{u=-\infty}^{\infty}\Big[\rho(u+p)\rho(u+q) + \rho(u-p)\rho(u+q) + 2\rho(p)\rho(q)\rho^2(u) \\ &\qquad - 2\rho(p)\rho(u)\rho(u+q) - 2\rho(q)\rho(u)\rho(u+p)\Big] \\ &= \sum_{u=1}^{\infty}[\rho(u+p) + \rho(u-p) - 2\rho(p)\rho(u)] \\ &\qquad \times [\rho(u+q) + \rho(u-q) - 2\rho(q)\rho(u)], \end{aligned} \tag{A.55}$$

where the last form is more convenient.

Proof. To prove the theorem, we use the delta method[4] for the limiting distribution of a function of the form

$$\boldsymbol{g}(x_0, x_1, \ldots, x_K) = (x_1/x_0, \ldots, x_K/x_0)',$$

[4]The delta method states that if a k-dimensional vector sequence $\boldsymbol{x}_n \sim AN(\boldsymbol{\mu}, a_n^2\Sigma)$, with $a_n \to 0$, and $\boldsymbol{g}(\boldsymbol{x})$ is an $r \times 1$ continuously differentiable vector function of $\boldsymbol{x}$, then $\boldsymbol{g}(\boldsymbol{x}_n) \sim AN(\boldsymbol{g}(\boldsymbol{\mu}), a_n^2 D\Sigma D')$ where D is the $r \times k$ matrix with elements $d_{ij} = \frac{\partial g_i(\boldsymbol{x})}{\partial x_j}\Big|_{\boldsymbol{\mu}}$.

where $x_h = \widehat{\gamma}(h)$, for $h = 0, 1, \ldots, K$. Hence, using the delta method and Theorem A.6,

$$\boldsymbol{g}\,(\widehat{\gamma}(0), \widehat{\gamma}(1), \ldots, \widehat{\gamma}(K)) = (\widehat{\rho}(1), \ldots, \widehat{\rho}(K))'$$

is asymptotically normal with mean vector $(\rho(1), \ldots, \rho(K))'$ and covariance matrix

$$n^{-1}W = n^{-1}DVD',$$

where V is defined by (A.54) and D is the $(K+1) \times K$ matrix of partial derivatives

$$D = \frac{1}{x_0^2}\begin{pmatrix} -x_1 & x_0 & 0 & \ldots & 0 \\ -x_2 & 0 & x_0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -x_K & 0 & 0 & \ldots & x_0, \end{pmatrix}$$

Substituting $\gamma(h)$ for x_h, we note that D can be written as the patterned matrix

$$D = \frac{1}{\gamma(0)}\,(\,-\boldsymbol{\rho} \quad I_K\,),$$

where $\boldsymbol{\rho} = (\rho(1), \rho(2), \ldots, \rho(K))'$ is the $K \times 1$ matrix of autocorrelations and I_K is the $K \times K$ identity matrix. Then, it follows from writing the matrix V in the partitioned form

$$V = \begin{pmatrix} v_{00} & \boldsymbol{v}_1' \\ \boldsymbol{v}_1 & V_{22} \end{pmatrix}$$

that

$$W = \gamma^{-2}(0)\big[v_{00}\boldsymbol{\rho}\boldsymbol{\rho}' - \boldsymbol{\rho}\boldsymbol{v}_1' - \boldsymbol{v}_1\boldsymbol{\rho}' + V_{22}\big],$$

where $\boldsymbol{v}_1 = (v_{10}, v_{20}, \ldots, v_{K0})'$ and $V_{22} = \{v_{pq};\, p, q = 1, \ldots, K\}$. Hence,

$$\begin{aligned} w_{pq} &= \gamma^{-2}(0)\big[v_{pq} - \rho(p)v_{0q} - \rho(q)v_{p0} + \rho(p)\rho(q)v_{00}\big] \\ &= \sum_{u=-\infty}^{\infty}\Big[\rho(u)\rho(u-p+q) + \rho(u-p)\rho(u+q) + 2\rho(p)\rho(q)\rho^2(u) \\ &\qquad - 2\rho(p)\rho(u)\rho(u+q) - 2\rho(q)\rho(u)\rho(u-p)\Big]. \end{aligned}$$

Interchanging the summations, we get the w_{pq} specified in the statement of the theorem, finishing the proof. ∎

Specializing the theorem to the case of interest in this chapter, we note that if $\{x_t\}$ is iid with finite fourth moment, then $w_{pq} = 1$ for $p = q$ and is zero otherwise. In this case, for $h = 1, \ldots, K$, the $\widehat{\rho}(h)$ are asymptotically independent and jointly normal with

$$\widehat{\rho}(h) \sim AN(0, n^{-1}). \tag{A.56}$$

This justifies the use of (1.38) and the discussion below it as a method for testing whether a series is white noise.

For the cross-correlation, it has been noted that the same kind of approximation holds and we quote the following theorem for the bivariate case, which can be proved using similar arguments (see Brockwell and Davis, 1991, p. 410).

Theorem A.8 *If*

$$x_t = \sum_{j=-\infty}^{\infty} \alpha_j w_{t-j,1}$$

and

$$y_t = \sum_{j=-\infty}^{\infty} \beta_j w_{t-j,2}$$

are two linear processes of the form with absolutely summable coefficients and the two white noise sequences are iid and independent of each other with variances σ_1^2 *and* σ_2^2*, then for* $h \geq 0$*,*

$$\widehat{\rho}_{xy}(h) \sim AN\left(\rho_{xy}(h), n^{-1}\sum_j \rho_x(j)\rho_y(j)\right) \tag{A.57}$$

and the joint distribution of $(\widehat{\rho}_{xy}(h), \widehat{\rho}_{xy}(k))'$ *is asymptotically normal with mean vector zero and*

$$\text{cov}\ \ (\widehat{\rho}_{xy}(h), \widehat{\rho}_{xy}(k)) = n^{-1}\sum_j \rho_x(j)\rho_y(j+k-h). \tag{A.58}$$

Again, specializing to the case of interest in this chapter, as long as at least one of the two series is white (iid) noise, we obtain

$$\widehat{\rho}_{xy}(h) \sim AN\big(0, n^{-1}\big), \tag{A.59}$$

which justifies Property P1.2.

Appendix B
Time Domain Theory

B.1 Hilbert Spaces and the Projection Theorem

Most of the material on mean square estimation and regression can be embedded in a more general setting involving an inner product space that is also complete (that is, satisfies the Cauchy condition). Two examples of inner products are $E(xy^*)$, where the elements are random variables, and $\sum x_i y_i^*$, where the elements are sequences. These examples include the possibility of complex elements, in which case, * denotes the conjugation. We denote an inner product, in general, by the notation $\langle x, y\rangle$. Now, define an inner product space by its properties, namely,

(i) $\langle x, y\rangle = \langle y, x\rangle^*$
(ii) $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$
(iii) $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$
(iv) $\langle x, x\rangle = \|x\|^2 \geq 0$
(v) $\langle x, x\rangle = 0$ iff $x = 0$.

We introduced the notation $\|\cdot\|$ for the norm or distance in property (iv). The norm satisfies the triangle inequality

$$\|x + y\| \leq \|x\| + \|y\| \tag{B.1}$$

and the Cauchy–Schwarz inequality

$$|\langle x, y\rangle|^2 \leq \|x\|^2\|y\|^2, \tag{B.2}$$

which we have seen before for random variables in (A.36). Now, a Hilbert space, $\mathcal{H}$, is defined as an inner product space with the Cauchy property. In other words, $\mathcal{H}$ is a complete inner product space. This means that every Cauchy sequence converges in norm; that is, $x_n \to x \in \mathcal{H}$ if an only if $\|x_n - x_m\| \to 0$

as $m, n \to \infty$. This is just the L^2 completeness Theorem A.1 for random variables.

For a broad overview of Hilbert space techniques that are useful in statistical inference and in probability, see Small and McLeish (1994). Also, Brockwell and Davis (1991, Chapter 2) is a nice summary of Hilbert space techniques that are useful in time series analysis. In our discussions, we mainly use the *projection theorem* (Theorem B.1) and the associated orthogonality principle as a means for solving various kinds of linear estimation problems.

Theorem B.1 *Let $\mathcal{M}$ be a closed subspace of the Hilbert space $\mathcal{H}$ and let y be an element in $\mathcal{H}$. Then, y can be uniquely represented as*

$$y = \widehat{y} + z, \tag{B.3}$$

where $\widehat{y}$ belongs to $\mathcal{M}$ and z is orthogonal to $\mathcal{M}$; that is, $\langle z, w\rangle = 0$ for all w in $\mathcal{M}$. Furthermore, the point $\widehat{y}$ is closest to y in the sense that, for any w in $\mathcal{M}$, $\|y - w\| \geq \|y - \widehat{y}\|$, where equality holds if and only if $w = \widehat{y}$.

We note that (B.3) and the statement following it yield the *orthogonality property*

$$\langle y - \widehat{y}, w\rangle = 0 \tag{B.4}$$

for any w belonging to $\mathcal{M}$, which can sometimes be used easily to find an expression for the projection. The norm of the error can be written as

$$\begin{aligned} \|y - \widehat{y}\|^2 &= \langle y - \widehat{y}, y - \widehat{y}\rangle \\ &= \langle y - \widehat{y}, y\rangle - \langle y - \widehat{y}, \widehat{y}\rangle \\ &= \langle y - \widehat{y}, y\rangle \end{aligned} \tag{B.5}$$

because of orthogonality.

Using the notation of Theorem B.1, we call the mapping $P_{\mathcal{M}} y = \widehat{y}$, for $y \in \mathcal{H}$, the projection mapping of $\mathcal{H}$ onto $\mathcal{M}$. In addition, the closed span of a finite set $\{x_1, \ldots, x_n\}$ of elements in a Hilbert space, $\mathcal{H}$, is defined to be the set of all linear combinations $w = a_1 x_1 + \cdots + a_n x_n$, where $a_1, \ldots, a_n$ are scalars. This subspace of $\mathcal{H}$ is denoted by $\mathcal{M} = \overline{\text{sp}}\{x_1, \ldots, x_n\}$. By the projection theorem, the projection of $y \in \mathcal{H}$ onto $\mathcal{M} = \overline{\text{sp}}\{x_1, \ldots, x_n\}$ is unique and given by

$$P_{\mathcal{M}} y = a_1 x_1 + \cdots + a_n x_n,$$

where $\{a_1, \ldots, a_n\}$ are found using the orthogonality principle

$$\langle y - P_{\mathcal{M}} y, x_j\rangle = 0 \quad j = 1, \ldots, n.$$

Evidently, $\{a_1, \ldots, a_n\}$ can be obtained by solving

$$\sum_{i=1}^{n} a_i \langle x_i, x_j\rangle = \langle y, x_j\rangle \quad j = 1, \ldots, n. \tag{B.6}$$

When the elements of $\mathcal{H}$ are vectors, this problem is the linear regression problem.

Example B.1 Linear Regression Analysis

For the regression model introduced in §2.2, we want to find the regression coefficients β_i that minimize the residual sum of squares. Consider the vectors $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{z}_i = (z_{1i}, \ldots, z_{ni})'$, for $i = 1, \ldots, q$ and the inner product

$$\langle \boldsymbol{z}_i, \boldsymbol{y} \rangle = \sum_{t=1}^{n} z_{ti} y_t = \boldsymbol{z}_i' \, \boldsymbol{y}.$$

We solve the problem of finding a projection of the observed $\boldsymbol{y}$ on the linear space spanned by $\beta_1 \boldsymbol{z}_1 + \cdots + \beta_q \boldsymbol{z}_q$, that is, linear combinations of the $\boldsymbol{z}_i$. The orthogonality principle gives

$$\langle \boldsymbol{y} - \sum_{i=1}^{q} \beta_i \boldsymbol{z}_i, \, \boldsymbol{z}_j \rangle = 0$$

for $j = 1, \ldots, q$. Writing the orthogonality condition, as in (B.6), in vector form gives

$$\boldsymbol{y}' \boldsymbol{z}_j = \sum_{i=1}^{q} \beta_i \boldsymbol{z}_i' \boldsymbol{z}_j \quad j = 1, \ldots, q, \tag{B.7}$$

which can be written in the usual matrix form by letting $Z = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q)$, which is assumed to be full rank. That is, (B.7) can be written as

$$\boldsymbol{y}' Z = \boldsymbol{\beta}' (Z' Z), \tag{B.8}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$. Transposing both sides of (B.8) provides the solution for the coefficients,

$$\widehat{\boldsymbol{\beta}} = (Z' Z)^{-1} Z' \boldsymbol{y}.$$

The mean square error in this case would be

$$\begin{aligned} \| \boldsymbol{y} - \sum_{i=1}^{q} \widehat{\beta}_i \boldsymbol{z}_i \|^2 &= \langle \boldsymbol{y} - \sum_{i=1}^{q} \widehat{\beta}_i \boldsymbol{z}_i \, , \, \boldsymbol{y} \rangle \\ &= \langle \boldsymbol{y}, \boldsymbol{y} \rangle - \sum_{i=1}^{q} \widehat{\beta}_i \langle \boldsymbol{z}_i \, , \, \boldsymbol{y} \rangle \\ &= \boldsymbol{y}' \boldsymbol{y} - \widehat{\boldsymbol{\beta}}' Z' \boldsymbol{y}, \end{aligned}$$

which is in agreement with §2.2.

The extra generality in the above approach hardly seems necessary in the finite dimensional case, where differentiation works perfectly well. It is convenient, however, in many cases to regard the elements of $\mathcal{H}$ as infinite dimensional, so that the orthogonality principle becomes of use. For example, the

projection of the process $\{x_t;\, t = 0 \pm 1, \pm 2, \ldots\}$ on the linear manifold spanned by all filtered convolutions of the form

$$\widehat{x}_t = \sum_{k=-\infty}^{\infty} a_k x_{t-k}$$

would be in this form.

There are some useful results, that we state without proof, pertaining to projection mappings.

Theorem B.2 *Under the established notation and conditions:*

(i) $P_{\mathcal{M}}(ax + by) = aP_{\mathcal{M}}x + bP_{\mathcal{M}}y$, *for* $x, y \in \mathcal{H}$, *where* a *and* b *are scalars.*

(ii) *If* $||y_n - y|| \to 0$, *then* $P_{\mathcal{M}}y_n \to P_{\mathcal{M}}y$, *as* $n \to \infty$.

(iii) $w \in \mathcal{M}$ *if and only if* $P_{\mathcal{M}}w = w$. *Consequently, a projection mapping can be characterized by the property that* $P_{\mathcal{M}}^2 = P_{\mathcal{M}}$, *in the sense that, for any* $y \in \mathcal{H}$, $P_{\mathcal{M}}(P_{\mathcal{M}}y) = P_{\mathcal{M}}y$.

(iv) *Let* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *be closed subspaces of* $\mathcal{H}$. *Then,* $\mathcal{M}_1 \subseteq \mathcal{M}_2$ *if and only if* $P_{\mathcal{M}_1}(P_{\mathcal{M}_2}y) = P_{\mathcal{M}_1}y$ *for all* $y \in \mathcal{H}$.

(v) *Let* $\mathcal{M}$ *be a closed subspace of* $\mathcal{H}$ *and let* $\mathcal{M}_\perp$ *denote the orthogonal complement of* $\mathcal{M}$. *Then,* $\mathcal{M}_\perp$ *is also a closed subspace of* $\mathcal{H}$, *and for any* $y \in \mathcal{H}$, $y = P_{\mathcal{M}}y + P_{\mathcal{M}_\perp}y$.

Part (iii) of Theorem B.2 leads to the well-known result, often used in linear models, that a square matrix M is a projection matrix if and only if it is symmetric and idempotent (that is, $M^2 = M$). For example, using notation of Example B.1 for linear regression, the projection of $\boldsymbol{y}$ onto $\overline{\text{sp}}\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q\}$, the space generated by the columns of Z, is $P_Z(\boldsymbol{y}) = Z\widehat{\boldsymbol{\beta}} = Z(Z'Z)^{-1}Z'\boldsymbol{y}$. The matrix $M = Z(Z'Z)^{-1}Z'$ is an $n \times n$, symmetric and idempotent matrix of rank q (which is the dimension of the space that M projects $\boldsymbol{y}$ onto). Parts (iv) and (v) of Theorem B.2 are useful for establishing recursive solutions for estimation and prediction.

By imposing extra structure, conditional expectation can be defined as a projection mapping for random variables in L^2 with the equivalence relation that, for $x, y \in L^2$, $x = y$ if $\Pr(x = y) = 1$. In particular, for $y \in L^2$, if $\mathcal{M}$ is a closed subspace of L^2 containing 1, the conditional expectation of y given $\mathcal{M}$ is defined to be the projection of y onto $\mathcal{M}$, namely, $E_{\mathcal{M}}y = P_{\mathcal{M}}y$. This means that conditional expectation, $E_{\mathcal{M}}$, must satisfy the orthogonality principle of the Projection Theorem and that the results of Theorem B.2 remain valid (the most widely used tool in this case is item (iv) of the theorem). If we let $\mathcal{M}(x)$ denote the closed subspace of all random variables in L^2 that can be written as a (measurable) function of x, then we may define, for $x, y \in L^2$, the conditional expectation of y given x as $E(y|x) = E_{\mathcal{M}(x)}y$. This idea may

be generalized in an obvious way to define the conditional expectation of y given $\boldsymbol{x} = (x_1, \ldots, x_n)$; that is $E(y|\boldsymbol{x}) = E_{\mathcal{M}(\boldsymbol{x})}y$. Of particular interest to us is the following result which states that, in the Gaussian case, conditional expectation and linear prediction are equivalent.

Theorem B.3 *Under the established notation and conditions, if* $(y, x_1, \ldots, x_n)$ *is multivariate normal, then*

$$E(y \mid x_1, \ldots, x_n) = P_{\overline{\text{sp}}\{1, x_1, \ldots, x_n\}}y.$$

Proof. First, by the projection theorem, the conditional expectation of y given $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is the unique element $E_{\mathcal{M}(\boldsymbol{x})}y$ that satisfies the orthogonality principle,

$$E\left\{\left(y - E_{\mathcal{M}(\boldsymbol{x})}y\right)w\right\} = 0 \quad \text{for all } w \in \mathcal{M}(\boldsymbol{x}).$$

We will show that $\widehat{y} = P_{\overline{\text{sp}}\{1, x_1, \ldots, x_n\}}y$ is that element. In fact, by the projection theorem, $\widehat{y}$ satisfies

$$\langle y - \widehat{y}, x_i \rangle = 0 \quad \text{for } i = 0, 1, \ldots, n,$$

where we have set $x_0 = 1$. But $\langle y - \widehat{y}, x_i \rangle = \text{cov}(y - \widehat{y}, x_i) = 0$, implying that $y - \widehat{y}$ and $(x_1, \ldots, x_n)$ are independent because the vector $(y - \widehat{y}, x_1, \ldots, x_n)'$ is multivariate normal. Thus, if $w \in \mathcal{M}(\boldsymbol{x})$, then w and $y - \widehat{y}$ are independent and, hence, $\langle y - \widehat{y}, w \rangle = E\{(y - \widehat{y})w\} = E(y - \widehat{y})E(w) = 0$, recalling that $0 = \langle y - \widehat{y}, 1 \rangle = E(y - \widehat{y})$. ∎

In the Gaussian case, conditional expectation has an explicit form. Let $\boldsymbol{y} = (y_1, \ldots, y_m)'$, $\boldsymbol{x} = (x_1, \ldots, x_n)'$, and suppose the $(m + n) \times 1$ vector $(\boldsymbol{y}', \boldsymbol{x}')'$ is multivariate normal. Then

$$E(\boldsymbol{y} \mid \boldsymbol{x}) = \boldsymbol{\mu}_y + \Sigma_{yx}\Sigma_{xx}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_x) \tag{B.9}$$

$$\text{var}(\boldsymbol{y} \mid \boldsymbol{x}) = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}, \tag{B.10}$$

where $\boldsymbol{\mu}_y = E(\boldsymbol{y})$ is $m \times 1$, $\Sigma_{yy} = \text{var}(\boldsymbol{y})$ is $m \times m$, $\boldsymbol{\mu}_x = E(\boldsymbol{x})$ is an $n \times 1$ vector, $\Sigma_{yx} = \Sigma_{xy}' = \text{cov}(\boldsymbol{y}, \boldsymbol{x})$ is $m \times n$, and $\Sigma_{xx} = \text{var}(\boldsymbol{x})$ is an $n \times n$ matrix, assumed to be nonsingular.

B.2 Causal Conditions for ARMA Models

In this section, we prove Property P3.1 of §3.2 pertaining to the causality of ARMA models. The proof of Property P3.2, which pertains to invertibility of ARMA models, is similar.

Proof of Property P3.1. Suppose first that the roots of $\phi(z)$, say, $z_1, \ldots, z_p$, lie outside the unit circle. We write the roots in the following order, $1 < |z_1| \leq |z_2| \leq \cdots \leq |z_p|$, noting that $z_1, \ldots, z_p$ are not necessarily unique, and put

$|z_1| = 1 + \epsilon$, for some $\epsilon > 0$. Thus, $\phi(z) \neq 0$ as long as $|z| < |z_1| = 1 + \epsilon$ and, hence, $\phi^{-1}(z)$ exists and has a power series expansion,

$$\frac{1}{\phi(z)} = \sum_{j=0}^{\infty} a_j z^j, \quad |z| < 1 + \epsilon.$$

Now, choose a value δ such that $0 < \delta < \epsilon$, and set $z = 1 + \delta$, which is inside the radius of convergence. It then follows that

$$\phi^{-1}(1 + \delta) = \sum_{j=0}^{\infty} a_j (1 + \delta)^j < \infty. \tag{B.11}$$

Thus, we can bound each of the terms in the sum in (B.11) by constant, say, $|a_j(1 + \delta)^j| < c$, for $c > 0$. In turn, $|a_j| < c(1 + \delta)^{-j}$, from which it follows that

$$\sum_{j=0}^{\infty} |a_j| < \infty. \tag{B.12}$$

Hence, $\phi^{-1}(B)$ exists and we may apply it to both sides of the ARMA model, $\phi(B)x_t = \theta(B)w_t$, to obtain

$$x_t = \phi^{-1}(B)\phi(B)x_t = \phi^{-1}(B)\theta(B)w_t.$$

Thus, putting $\psi(B) = \phi^{-1}(B)\theta(B)$, we have

$$x_t = \psi(B)w_t = \sum_{j=0}^{\infty} \psi_j w_{t-j},$$

where the ψ-weights, which are absolutely summable, can be evaluated by $\psi(z) = \phi^{-1}(z)\theta(z)$, for $|z| \leq 1$.

Now, suppose x_t is a causal process; that is, it has the representation

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}, \qquad \sum_{j=0}^{\infty} |\psi_j| < \infty.$$

In this case, we write

$$x_t = \psi(B)w_t,$$

and premultiplying by $\phi(B)$ yields

$$\phi(B)x_t = \phi(B)\psi(B)w_t. \tag{B.13}$$

In addition to (B.13), the model is ARMA, and can be written as

$$\phi(B)x_t = \theta(B)w_t. \tag{B.14}$$

From (B.13) and (B.14), we see that

$$\phi(B)\psi(B)w_t = \theta(B)w_t. \tag{B.15}$$

Now, let

$$a(z) = \phi(z)\psi(z) = \sum_{j=0}^{\infty} a_j z^j \quad |z| \le 1$$

and, hence, we can write (B.15) as

$$\sum_{j=0}^{\infty} a_j w_{t-j} = \sum_{j=0}^{q} \theta_j w_{t-j}. \tag{B.16}$$

Next, multiply both sides of (B.16) by w_{t-h}, for $h = 0, 1, 2, \ldots$, and take expectation. In doing this, we obtain

$$\begin{aligned} a_h &= \theta_h, \quad h = 0, 1, \ldots, q \\ a_h &= 0, \quad h > q. \end{aligned} \tag{B.17}$$

From (B.17), we conclude that

$$\phi(z)\psi(z) = a(z) = \theta(z), \quad |z| \le 1. \tag{B.18}$$

If there is a complex number in the unit circle, say z_0, for which $\phi(z_0) = 0$, then by (B.18), $\theta(z_0) = 0$. But, if there is such a z_0, then $\phi(z)$ and $\theta(z)$ have a common factor which is not allowed. Thus, we may write $\psi(z) = \theta(z)/\phi(z)$. In addition, by hypothesis, we have that $|\psi(z)| < \infty$ for $|z| \le 1$, and hence

$$|\psi(z)| = \left|\frac{\theta(z)}{\phi(z)}\right| < \infty, \quad \text{for } |z| \le 1. \tag{B.19}$$

Finally, (B.19) implies $\phi(z) \neq 0$ for $|z| \le 1$; that is, the roots of $\phi(z)$ lie outside the unit circle. ∎

B.3 Large Sample Distribution of the AR(p) Conditional Least Squares Estimators

In §3.6 we discussed the conditional least squares procedure for estimating the parameters $\phi_1, \phi_2, \ldots, \phi_p$ and σ_w^2 in the AR(p) model

$$x_t = \sum_{k=1}^{p} \phi_k x_{t-k} + w_t,$$

where we assume $\mu = 0$, for convenience. Write the model as

$$x_t = \boldsymbol{\phi}' \boldsymbol{x}_{t-1} + w_t, \tag{B.20}$$

where $\boldsymbol{x}_{t-1} = (x_{t-1}, x_{t-2}, \ldots, x_{t-p})'$ is a $p \times 1$ vector of lagged values, and $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_p)'$ is the $p \times 1$ vector of regression coefficients. Assuming observations are available at $x_1, \ldots, x_n$, the conditional least squares procedure is to minimize

$$S_c(\boldsymbol{\phi}) = \sum_{t=p+1}^{n} \left(x_t - \boldsymbol{\phi}'\boldsymbol{x}_{t-1}\right)^2$$

with respect to $\boldsymbol{\phi}$. The solution is

$$\widehat{\boldsymbol{\phi}} = \left(\sum_{t=p+1}^{n} \boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}'\right)^{-1} \sum_{t=p+1}^{n} \boldsymbol{x}_{t-1}x_t \tag{B.21}$$

for the regression vector $\boldsymbol{\phi}$; the conditional least squares estimate of σ_w^2 is

$$\widehat{\sigma}_w^2 = \frac{1}{n-p} \sum_{t=p+1}^{n} \left(x_t - \widehat{\boldsymbol{\phi}}'\boldsymbol{x}_{t-1}\right)^2. \tag{B.22}$$

As pointed out following (3.104), Yule–Walker estimators and least squares estimators are approximately the same in that the estimators differ only by inclusion or exclusion of terms involving the endpoints of the data. Hence, it is easy to show the asymptotic equivalence of the two estimators; this is why, for AR(p) models, (3.93) and (3.118), are equivalent. Details on the asymptotic equivalence can be found in Brockwell and Davis (1991, Chapter 8).

Here, we use the same approach as in Appendix A, replacing the lower limits of the sums in (B.21) and (B.22) by one and noting the asymptotic equivalence of the estimators

$$\widetilde{\boldsymbol{\phi}} = \left(\sum_{t=1}^{n} \boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}'\right)^{-1} \sum_{t=1}^{n} \boldsymbol{x}_{t-1}x_t \tag{B.23}$$

and

$$\widetilde{\sigma}_w^2 = \frac{1}{n} \sum_{t=1}^{n} \left(x_t - \widetilde{\boldsymbol{\phi}}'\boldsymbol{x}_{t-1}\right)^2 \tag{B.24}$$

to those two estimators. In (B.23) and (B.24), we are acting as if we are able to observe $x_{1-p}, \ldots, x_0$ in addition to $x_1, \ldots, x_n$. The asymptotic equivalence is then seen by arguing that for n sufficiently large, it makes no difference whether or not we observe $x_{1-p}, \ldots, x_0$. In the case of (B.23) and (B.24), we obtain the following theorem.

Theorem B.4 *Let x_t be a causal AR(p) series with white (iid) noise w_t satisfying $E(w_t^4) = \eta\sigma_w^4$. Then,*

$$\widetilde{\boldsymbol{\phi}} \sim \mathrm{AN}\left(\boldsymbol{\phi},\ n^{-1}\sigma_w^2\Gamma_p^{-1}\right), \tag{B.25}$$

where $\Gamma_p = \{\gamma(i-j)\}_{i,j=1}^p$ *is the* $p \times p$ *autocovariance matrix of the vector* $\boldsymbol{x}_{t-1}$. *We also have, as* $n \to \infty$,

$$n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}' \xrightarrow{p} \Gamma_p \tag{B.26}$$

and

$$\widetilde{\sigma}_w^2 \xrightarrow{p} \sigma_w^2. \tag{B.27}$$

Proof. First, (B.26) follows from the fact that $E(\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}') = \Gamma_p$, recalling that from Theorem 1.6, second-order sample moments converge in probability to their population moments for linear processes in which w_t has a finite fourth moment. To show (B.25), we can write

$$\begin{aligned} \widetilde{\boldsymbol{\phi}} &= \left(\sum_{t=1}^{n} \boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}'\right)^{-1} \sum_{t=1}^{n} \boldsymbol{x}_{t-1}(\boldsymbol{x}_{t-1}'\boldsymbol{\phi} + w_t) \\ &= \boldsymbol{\phi} + \left(\sum_{t=1}^{n} \boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}'\right)^{-1} \sum_{t=1}^{n} \boldsymbol{x}_{t-1} w_t, \end{aligned}$$

so that

$$\begin{aligned} n^{1/2}(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}) &= \left(n^{-1}\sum_{t=1}^{n} \boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}'\right)^{-1} n^{-1/2} \sum_{t=1}^{n} \boldsymbol{x}_{t-1} w_t \\ &= \left(n^{-1}\sum_{t=1}^{n} \boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}'\right)^{-1} n^{-1/2} \sum_{t=1}^{n} \boldsymbol{u}_t, \end{aligned}$$

where $\boldsymbol{u}_t = \boldsymbol{x}_{t-1} w_t$. We use the fact that w_t and $\boldsymbol{x}_{t-1}$ are independent to write $E\boldsymbol{u}_t = E(\boldsymbol{x}_{t-1})E(w_t) = \boldsymbol{0}$, because the errors have zero means. Also,

$$\begin{aligned} E\boldsymbol{u}_t\boldsymbol{u}_t' &= E\boldsymbol{x}_{t-1} w_t w_t \boldsymbol{x}_{t-1}' \\ &= E\boldsymbol{x}_{t-1}\boldsymbol{x}_{t-1}' E w_t^2 \\ &= \sigma_w^2 \Gamma_p. \end{aligned}$$

In addition, we have, for $h > 0$,

$$\begin{aligned} E\boldsymbol{u}_{t+h}\boldsymbol{u}_t' &= E\boldsymbol{x}_{t+h-1} w_{t+h} w_t \boldsymbol{x}_{t-1}' \\ &= E\boldsymbol{x}_{t+h-1} w_t \boldsymbol{x}_{t-1}' E w_{t+h} \\ &= 0. \end{aligned}$$

A similar computation works for $h < 0$.

Next, consider the mean square convergent approximation

$$x_t^m = \sum_{j=0}^{m} \psi_j w_{t-j}$$

for x_t, and define the $(m+p)$-dependent process $\boldsymbol{u}_t^m = w_t(x_{t-1}^m, x_{t-2}^m, \ldots, x_{t-p}^m)'$. Note that we need only look at a central limit theorem for the sum

$$y_{nm} = n^{-1/2} \sum_{t=1}^{n} \boldsymbol{\lambda}' \boldsymbol{u}_t^m,$$

for arbitrary vectors $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$, where y_{nm} is used as an approximation to

$$S_n = n^{-1/2} \sum_{t=1}^{n} \boldsymbol{\lambda}' \boldsymbol{u}_t.$$

First, apply the m-dependent central limit theorem to y_{nm} as $n \to \infty$ for fixed m to establish (i) of Theorem A.4. This result shows $y_{nm} \xrightarrow{d} y_m$, where y_m is asymptotically normal with covariance $\boldsymbol{\lambda}' \Gamma_p^{(m)} \boldsymbol{\lambda}$, where $\Gamma_p^{(m)}$ is the covariance matrix of $\boldsymbol{u}_t^m$. Then, we have $\Gamma_p^{(m)} \to \Gamma_p$, so that y_m converges in distribution to a normal random variable with mean zero and variance $\boldsymbol{\lambda}' \Gamma_p \boldsymbol{\lambda}$ and we have verified part (ii) of Theorem A.4. We verify part (iii) of Theorem A.4 by noting that

$$E[(S_n - y_{nm})^2] = n^{-1} \sum_{t=1}^{n} \boldsymbol{\lambda}' E[(\boldsymbol{u}_t - \boldsymbol{u}_t^m)(\boldsymbol{u}_t - \boldsymbol{u}_t^m)'] \boldsymbol{\lambda}$$

clearly converges to zero as $n, m \to \infty$ because

$$x_t - x_t^m = \sum_{j=m+1}^{\infty} \psi_j w_{t-j}$$

form the components of $\boldsymbol{u}_t - \boldsymbol{u}_t^m$.

Now, the form for $\sqrt{n}(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi})$ contains the premultiplying matrix

$$\left(n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_{t-1} \boldsymbol{x}_{t-1}' \right)^{-1} \xrightarrow{p} \Gamma_p^{-1},$$

because (A.22) can be applied to the function that defines the inverse of the matrix. Then, applying (A.31), shows that

$$n^{1/2} \left(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi} \right) \xrightarrow{d} \mathrm{N}\left(0, \sigma_w^2 \Gamma_p^{-1} \Gamma_p \Gamma_p^{-1}\right),$$

so we may regard it as being multivariate normal with mean zero and covariance matrix $\sigma_w^2 \Gamma_p^{-1}$.

To investigate $\widetilde{\sigma}_w^2$, note

$$\begin{aligned} \widetilde{\sigma}_w^2 &= n^{-1} \sum_{t=1}^{n} \left(x_t - \widetilde{\boldsymbol{\phi}}' \boldsymbol{x}_{t-1} \right)^2 \\ &= n^{-1} \sum_{t=1}^{n} x_t^2 - n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_{t-1}' x_t \left(n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_{t-1} \boldsymbol{x}_{t-1}' \right)^{-1} n^{-1} \sum_{t=1}^{n} \boldsymbol{x}_{t-1} x_t \end{aligned}$$

$$\begin{aligned} &\xrightarrow{p} \quad \gamma(0) - \boldsymbol{\gamma}_p' \Gamma_p^{-1} \boldsymbol{\gamma}_p \\ &= \quad \sigma_w^2, \end{aligned}$$

and we have that the sample estimator converges in probability to σ_w^2, which is written in the form of (3.59). ∎

The arguments above imply that, for sufficiently large n, we may consider the estimator $\widehat{\boldsymbol{\phi}}$ in (B.21) as being approximately multivariate normal with mean $\boldsymbol{\phi}$ and variance–covariance matrix $\sigma_w^2 \Gamma_p^{-1}/n$. Inferences about the parameter $\boldsymbol{\phi}$ are obtained by replacing the σ_w^2 and Γ_p by their estimates given by (B.22) and

$$\widehat{\Gamma}_p = n^{-1} \sum_{t=p+1}^{n} \boldsymbol{x}_{t-1} \boldsymbol{x}_{t-1}',$$

respectively. In the case of a nonzero mean, the data x_t are replaced by $x_t - \bar{x}$ in the estimates and the results of Theorem B.4 remain valid.

B.4 The Wold Decomposition

The ARMA approach to modeling time series is generally implied by the assumption that the dependence between adjacent values in time is best explained in terms of a regression of the current values on the past values. This assumption is partially justified, in theory, by the Wold decomposition.

In this section we assume that $\{x_t;\ t = 0, \pm 1, \pm 2, \ldots\}$ is a stationary, mean-zero process. Using the notation of §B.1, we define

$$\mathcal{M}_n = \overline{\text{sp}}\{x_t,\ -\infty < t \leq n\}, \quad \text{with} \quad \mathcal{M}_{-\infty} = \bigcap_{n=-\infty}^{\infty} \mathcal{M}_n,$$

and

$$\sigma^2 = E\left(x_{n+1} - P_{\mathcal{M}_n} x_{n+1}\right)^2.$$

Next, we say that $\{v_t;\ t = 0, \pm 1, \pm 2, \ldots\}$ is a deterministic process if and only if $\sigma^2 = 0$. That is, a deterministic process is one in which its future is perfectly predictable from its past; a simple example is $v_t = \cos(.2\pi t)$. We are now ready to present the decomposition.

Theorem B.5 (The Wold Decomposition) *Under the conditions and notation of this section, if $\sigma^2 > 0$, then x_t can be expressed as*

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} + v_t$$

where

- $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ *(*$\psi_0 = 1$*)*
- $\{w_t\}$ *is white noise with variance* σ^2
- $w_t \in \mathcal{M}_t$
- $E(w_t v_t) = 0$ *for all* $s, t = 0, \pm 1, \pm 2, \ldots$.
- $v_t \in \mathcal{M}_{-\infty}$
- $\{v_t\}$ *is deterministic.*

The proof of the decomposition follows from the theory of §B.1 by defining the unique sequences:

(*i*) $w_t = x_t - P_{\mathcal{M}_{t-1}} x_t$,

(*ii*) $\psi_j = \sigma^{-2} < x_t, w_{t-j} >= \sigma^{-2} E(x_t w_{t-j})$, and

(*iii*) $v_t = x_t - \sum_{j=0}^{\infty} \psi_j w_{t-j}$.

Although every stationary process can be represented by the Wold decomposition, it does not mean that the decomposition is the best way to describe the process. In addition, there may be some dependence structure among the $\{w_t\}$; we are only guaranteed that the sequence is an uncorrelated sequence. The theorem, in its generality, falls short of our needs because we would prefer the noise process, $\{w_t\}$, to be white independent noise. But, the decomposition does give us the confidence that we will not be completely off the mark by fitting ARMA models to time series data.

Appendix C
Spectral Domain Theory

C.1 Spectral Representation Theorem

In this section, we present a spectral representation for the process x_t itself, which allows us to think of a stationary process as a random sum of sines and cosines as described in (4.4). In addition, we present results that justify representing the autocovariance function $\gamma_x(h)$ of the weakly stationary process x_t in terms of a non-negative spectral density function. The spectral density function essentially measures the variance or power in a particular kind of periodic oscillation in the function. We denote this spectral density of variance function by $f(\omega)$, where the variance is measured as a function of the frequency of oscillation ω, measured in cycles per unit time.

First, we consider developing a representation for the autocovariance function of a stationary, possibly complex, series x_t with zero mean and autocovariance function $\gamma_x(h) = E(x_{t+h}x_t^*)$. We prove the representation for arbitrary non-negative definite functions $\gamma(h)$ and then simply note the autocovariance function is Hermitian non-negative definite, because, for any set of complex constants, $a_t, t = 0 \pm 1, \pm 2, \ldots$, we may write, for any finite subset,

$$E\left|\sum_{s=1}^{n} a_s^* x_s\right|^2 = \sum_{s=1}^{n}\sum_{t=1}^{n} a_s^*\gamma(s-t)a_t \geq 0.$$

The representation is stated in terms of non-negative definite functions and a spectral distribution function $F(\omega)$ that is monotone nondecreasing, and continuous from the right, taking the values $F(-1/2) = 0$ and $F(1/2) = \sigma^2 = \gamma_x(0)$ at $\omega = -1/2$ and $1/2$, respectively.

Theorem C.1 *A function $\gamma(h)$, for $h = 0 \pm 1, \pm 2, \ldots$ is Hermitian non-negative definite if and only if it can be expressed as*

$$\gamma(h) = \int_{-1/2}^{1/2} \exp\{2\pi i\omega h\} dF(\omega) \tag{C.1}$$

where $F(\cdot)$ is monotone non-decreasing. The function $F(\cdot)$ is right continuous, bounded in $[-1/2, 1/2]$, and uniquely determined by the conditions $F(-1/2) = 0, F(1/2) = \gamma(0)$.

Proof. To prove the result, note first if $\gamma(h)$ has the representation above,

$$\begin{aligned}\sum_{s=1}^{n}\sum_{t=1}^{n} a_s^*\gamma(s-t)a_t &= \int_{-1/2}^{1/2} a_s^*\gamma(s-t)a_t e^{2\pi i\omega(s-t)}dF(\omega) \\ &= \int_{-1/2}^{1/2}\left|\sum_{s=1}^{n} a_s e^{-2\pi i\omega s}\right|^2 dF(\omega) \\ &\geq 0\end{aligned}$$

and $\gamma(h)$ is non-negative definite. Conversely, suppose $\gamma(h)$ is a non-negative definite function, and define the non-negative function

$$\begin{aligned}f_n(\omega) &= n^{-1}\sum_{s=1}^{n}\sum_{t=1}^{n} e^{-2\pi i\omega s}\gamma(s-t)e^{2\pi i\omega t} \\ &= n^{-1}\sum_{u=-(n-1)}^{(n-1)}(n-|u|)e^{-2\pi i\omega u}\gamma(u) \qquad \text{(C.2)} \\ &\geq 0.\end{aligned}$$

Now, let $F_n(\omega)$ be the distribution function corresponding to $f_n(\omega)I_{(-1/2,1/2]}$, where $I_{(\cdot)}$ denotes the indicator function of the interval in the subscript. Note that $F_n(\omega) = 0, \omega \leq -1/2$ and $F_n(\omega) = F_n(1/2)$ for $\omega \geq 1/2$. Then,

$$\begin{aligned}\int_{-1/2}^{1/2} e^{2\pi i\omega u}dF_n(\omega) &= \int_{-1/2}^{1/2} e^{2\pi i\omega u}f_n(\omega)\ d\omega \\ &= \begin{cases}(1-|u|/n)\gamma(u), & |u|<n \\ 0, & \text{elsewhere.}\end{cases}\end{aligned}$$

We also have

$$\begin{aligned}F_n(1/2) &= \int_{-1/2}^{1/2} f_n(\omega)\ d\omega \\ &= \int_{-1/2}^{1/2}\sum_{|u|<n}(1-|u|/n)\gamma(u)e^{-2\pi i\omega u}d\omega \\ &= \gamma(0).\end{aligned}$$

Now, by Helly's first convergence theorem (Bhat, 1985, p. 157), there exists a subsequence F_{n_k} converging to F, and by the Helly-Bray Lemma (see Bhat,

p. 157), this implies

$$\int_{-1/2}^{1/2} e^{2\pi i\omega u} dF_{n_k}(\omega) \to \int_{-1/2}^{1/2} e^{2\pi i\omega u} dF(\omega)$$

and, from the right-hand side of the earlier equation,

$$(1 - |u|/n_k)\gamma(u) \to \gamma(u)$$

as $n_k \to \infty$, and the required result follows. ∎

Next, present the version of the Spectral Representation Theorem in terms of a mean-zero, stationary process, x_t. We refer the reader to Hannan (1970, §2.3) for details. This version allows us to think of a stationary process as being generated (approximately) by a random sum of sines and cosines such as described in (4.4).

Theorem C.2 *If x_t is a mean-zero stationary process, with spectral distribution $F(\omega)$ as given in Theorem C.1, then there exists a complex-valued stochastic process $z(\omega)$, on the interval $\omega \in [-1/2, 1/2]$, having stationary uncorrelated increments, such that x_t can be written as the stochastic integral*

$$x_t = \int_{-1/2}^{1/2} \exp(-2\pi i t\omega) dz(\omega)$$

where, for $-1/2 \le \omega_1 \le \omega_2 \le 1/2$,

$$\operatorname{var}\{z(\omega_2) - z(\omega_1)\} = F(\omega_2) - F(\omega_1).$$

An uncorrelated increment process such as $z(\omega)$ is a mean-zero, finite variance, continuous-time stochastic process for which events occurring in non-overlapping intervals are uncorrelated. The integral in this representation is a stochastic integral. To understand its meaning, let $\omega_0, \omega_1, \ldots, \omega_n$ be a partition of the interval $[-1/2, 1/2]$. Define

$$I_n = \sum_{j=1}^{n} \exp(-2\pi i t\omega_j)[z(\omega_j) - z(\omega_{j-1})].$$

Then, assuming it exists, $I = \int_{-1/2}^{1/2} \exp(-2\pi i t\omega_j) dz(\omega)$ is defined to be the mean square limit of I_n as $n \to \infty$. Theorem C.2 allows us to think of a stationary process, approximately, as the random superposition of sines and cosines.

In general, the spectral distribution function can be a mixture of discrete and continuous distributions. The special case of greatest interest is the absolutely continuous case, namely, when $dF(\omega) = f(\omega)d\omega$, and the resulting

function is the spectral density considered in §4.3. What made the proof of Theorem C.1 difficult was that, after we defined

$$f_n(\omega) = \sum_{h=-(n-1)}^{(n-1)} \left(1 - \frac{|h|}{n}\right) \gamma(h) e^{-2\pi i \omega h}$$

in (C.2), we could not simply allow $n \to \infty$ because $\gamma(h)$ may not be absolutely summable. If, however, $\gamma(h)$ is absolutely summable we may define $f(\omega) = \lim_{n\to\infty} f_n(\omega)$, and we have the following result.

Theorem C.3 *If $\gamma(h)$ is the autocovariance function of a stationary process, x_t, with*

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty, \tag{C.3}$$

then the spectral density of x_t is given by

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h}. \tag{C.4}$$

We may extend the representation to the vector case $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})'$ by considering linear combinations of the form

$$y_t = \sum_{j=1}^{p} a_j^* x_{tj},$$

which will be stationary with autocovariance functions of the form

$$\gamma_y(h) = \sum_{j=1}^{p} \sum_{k=1}^{p} a_j^* \gamma_{jk}(h) a_k,$$

where $\gamma_{jk}(h)$ is the usual cross-covariance function between x_{tj} and x_{tk}. To develop the spectral representation of $\gamma_{jk}(h)$ from the representations of the univariate series, consider the linear combinations

$$y_{t1} = x_{tj} + x_{tk}$$

and

$$y_{t2} = x_{tj} + i x_{tk},$$

which are both stationary series with respective representations

$$\begin{aligned} \gamma_{y1}(h) &= \gamma_j(h) + \gamma_{jk}(h) + \gamma_{kj}(h) + \gamma_k(h) \\ &= \int_{-1/2}^{1/2} e^{2\pi i \omega h} dG_1(\omega) \end{aligned}$$

and

$$\begin{aligned}\gamma_{y2}(h) &= \gamma_j(h) + i\gamma_{kj}(h) - i\gamma_{jk}(h) + \gamma_k(h) \\ &= \int_{-1/2}^{1/2} e^{2\pi i\omega h} dG_2(\omega).\end{aligned}$$

Introducing the spectral representations for $\gamma_j(h)$ and $\gamma_k(h)$ yields

$$\gamma_{jk}(h) = \int_{-1/2}^{1/2} e^{2\pi i\omega h} dF_{jk}(\omega),$$

with

$$F_{jk}(\omega) = \frac{1}{2}[G_1(\omega) + iG_2(\omega) - (1+i)(F_j(\omega) + F_k(\omega)].$$

Now, under the summability condition

$$\sum_{h=-\infty}^{\infty} |\gamma_{jk}(h)| < \infty,$$

we have the representation

$$\gamma_{jk}(h) = \int_{-1/2}^{1/2} e^{2\pi i\omega h} f_{jk}(\omega) d\omega,$$

where the cross-spectral density function has the inverse Fourier representation

$$f_{jk}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{jk}(h) e^{-2\pi i\omega h}.$$

The cross-covariance function satisfies $\gamma_{jk}(h) = \gamma_{kj}(-h)$, which implies $f_{jk}(\omega) = f_{kj}(-\omega)$ using the above representation.

Then, defining the autocovariance function of the general vector process $\boldsymbol{x}_t$ as the $p \times p$ matrix

$$\Gamma(h) = E[(\boldsymbol{x}_{t+h} - \boldsymbol{\mu}_x)(\boldsymbol{x}_t - \boldsymbol{\mu}_x)'],$$

and the $p \times p$ spectral matrix as $f(\omega) = \{f_{jk}(\omega), j, k = 1, \ldots, p\}$, we have the representation in matrix form, written as

$$\Gamma(h) = \int_{-1/2}^{1/2} e^{2\pi i\omega h} f(\omega)\ d\omega, \tag{C.5}$$

and the inverse result

$$f(\omega) = \sum_{h=-\infty}^{\infty} \Gamma(h) e^{-2\pi i\omega h}. \tag{C.6}$$

which appears as Property P4.3 in §4.6. Theorem C.2 can also be extended to the multivariate case.

C.2 Large Sample Distribution of the DFT and Smoothed Periodogram

We have previously introduced the DFT, for the stationary zero-mean process x_t, observed at $t = 1, \ldots, n$ as

$$d(\omega) = n^{-1/2} \sum_{t=1}^{n} x_t \, e^{-2\pi i \omega t}, \tag{C.7}$$

as the result of matching sines and cosines of frequency ω against the series x_t. We will suppose now that x_t has an absolutely continuous spectrum $f(\omega)$ corresponding to the absolutely summable autocovariance function $\gamma(h)$. Our purpose in this section is to examine the statistical properties of the complex random variables $d(\omega_k)$, for $\omega_k = k/n$, $k = 0, 1, \ldots, n-1$ in providing a basis for the estimation of $f(\omega)$. To develop the statistical properties, we examine the behavior of

$$\begin{aligned} S_n(\omega, \omega) &= E\left\{|d(\omega)|^2\right\} \\ &= n^{-1} E\left[\sum_{s=1}^{n} x_s \; e^{-2\pi i \omega s} \sum_{t=1}^{n} x_t e^{2\pi i \omega t}\right] \\ &= n^{-1} \sum_{s=1}^{n} \sum_{t=1}^{n} e^{-2\pi i \omega s} e^{2\pi i \omega t} \gamma(s-t) \\ &= \sum_{u=-(n-1)}^{(n-1)} (1 - |u|/n) \gamma(u) e^{-2\pi i \omega u}, \end{aligned} \tag{C.8}$$

where we have let $u = s - t$. Using dominated convergence,

$$S_n(\omega, \omega) \to \sum_{u=-\infty}^{\infty} \gamma(u) e^{-2\pi i \omega u} = f(\omega),$$

as $n \to \infty$, making the large sample variance of the Fourier transform equal to the spectrum evaluated at ω. We have already seen this result in Theorem C.3. For exact bounds it is also convenient to add an absolute summability assumption for the autocovariance function, namely,

$$\theta = \sum_{h=-\infty}^{\infty} |h||\gamma(h)| < \infty. \tag{C.9}$$

Example C.1 Condition (C.9) Verified for an AR(1)

We may write the condition for an AR(1) series. $x_t = \phi x_{t-1} + w_t$, as

$$\theta = \frac{\sigma_w^2}{1-\phi^2} \sum_{h=-\infty}^{\infty} |h|\phi^{|h|}$$

being finite. Note the condition is equivalent to summability of

$$\begin{aligned} \sum_{h=1}^{\infty} h\phi^h &= \phi \sum_{h=1}^{\infty} h\phi^{h-1} \\ &= \phi \frac{\partial}{\partial \phi} \sum_{h=1}^{\infty} \phi^h \\ &= \frac{\phi}{(1-\phi)^2}, \end{aligned}$$

and hence,

$$\theta = \frac{2\sigma_w^2 \phi}{(1-\phi)^3(1+\phi)} < \infty.$$

To elaborate further, we derive two approximation lemmas.

Lemma C.1 *For $S_n(\omega, \omega)$ as defined in (C.8) and θ in (C.9) finite, we have*

$$|S_n(\omega, \omega) - f(\omega)| \leq \frac{\theta}{n} \tag{C.10}$$

or

$$S_n(\omega, \omega) = f(\omega) + O(n^{-1}). \tag{C.11}$$

Proof. To prove the lemma, write

$$\begin{aligned} n|S_n(\omega, \omega) - f_x(\omega)| &= \left| \sum_{|u|<n} (n-|u|)\gamma(u)e^{-2\pi i\omega u} - n\sum_{u=-\infty}^{\infty} \gamma(u)e^{-2\pi i\omega u} \right| \\ &= \left| -n\sum_{|u|\geq n} \gamma(u)e^{-2\pi i\omega u} - \sum_{|u|<n} |u|\gamma(u)e^{-2\pi i\omega u} \right| \\ &\leq \sum_{|u|\geq n} |u||\gamma(u)| + \sum_{|u|<n} |u||\gamma(u)| \\ &= \theta, \end{aligned}$$

which establishes the lemma. ∎

Lemma C.2 *For $\omega_k = k/n$, $\omega_\ell = \ell/n$, $\omega_k - \omega_\ell \neq 0, \pm 1, \pm 2, \pm 3, \ldots$, and θ in (C.9), we have*

$$|S_n(\omega_k, \omega_\ell)| \leq \frac{\theta}{n} = O(n^{-1}), \tag{C.12}$$

where

$$S_n(\omega_k, \omega_\ell) = E\{d(\omega_k)\overline{d(\omega_\ell)}\}. \tag{C.13}$$

Proof. Write

$$\begin{aligned} n|S_n(\omega_k, \omega_\ell)| &= \sum_{u=-(n-1)}^{-1} \gamma(u) \sum_{v=-(u-1)}^{n} e^{-2\pi i(\omega_k - \omega_\ell)v} e^{-2\pi i \omega_k u} \\ &\quad + \sum_{u=0}^{n-1} \gamma(u) \sum_{v=1}^{n-u} e^{-2\pi i(\omega_k - \omega_\ell)v} e^{-2\pi i \omega_k u}. \end{aligned}$$

Now, for the first term, with $u < 0$,

$$\begin{aligned} \sum_{v=-(u-1)}^{n} e^{-2\pi i(\omega_k - \omega_\ell)v} &= \left(\sum_{v=1}^{n} - \sum_{v=1}^{-u}\right) e^{-2\pi i(\omega_k - \omega_\ell)v} \\ &= 0 - \sum_{v=1}^{-u} e^{-2\pi i(\omega_k - \omega_\ell)v}. \end{aligned}$$

For the second term with $u \geq 0$,

$$\begin{aligned} \sum_{v=1}^{n-u} e^{-2\pi i(\omega_k - \omega_\ell)v} &= \left(\sum_{v=1}^{n} - \sum_{v=n-u+1}^{n}\right) e^{-2\pi i(\omega_k - \omega_\ell)v} \\ &= 0 - \sum_{v=n-u+1}^{n} e^{-2\pi i(\omega_k - \omega_\ell)v}. \end{aligned}$$

Consequently,

$$\begin{aligned} n|S_n(\omega_k, \omega_\ell)| &= \Bigg| - \sum_{u=-(n-1)}^{-1} \gamma(u) \sum_{v=1}^{-u} e^{-2\pi i(\omega_k - \omega_\ell)v} e^{-2\pi i \omega_k u} \\ &\qquad - \sum_{u=1}^{n-1} \gamma(u) \sum_{v=n-u+1}^{n} e^{-2\pi i(\omega_k - \omega_\ell)v} e^{-2\pi i \omega_k u} \Bigg| \\ &\leq \sum_{u=-(n-1)}^{0} (-u)|\gamma(u)| + \sum_{u=1}^{n-1} u|\gamma(u)| \\ &= \sum_{u=-(n-1)}^{(n-1)} |u|\ |\gamma(u)|. \end{aligned}$$

Hence, we have

$$S_n(\omega_k, \omega_\ell) \le \frac{\theta}{n},$$

and the asserted relations of the lemma follow. ∎

Because the DFTs are approximately uncorrelated, say, of order $1/n$, when the frequencies are of the form $\omega_k = k/n$, we shall compute at those frequencies. The behavior of $f_(\omega)$ at neighboring frequencies, however, will often be of interest and we shall use Lemma C.3 below to handle such cases

Lemma C.3 *For $|\omega_k - \omega| \le L/2n$ and θ in (C.9), we have*

$$|f(\omega_k) - f(\omega)| \le \frac{\pi\theta L}{n} \tag{C.14}$$

or

$$f(\omega_k) - f(\omega) = O(L/n). \tag{C.15}$$

Proof. To prove Lemma C.3, we write the difference

$$\begin{aligned}
|f(\omega_k) - f(\omega)| &= \left| \sum_{h=-\infty}^{\infty} \gamma(h)\big(e^{-2\pi i\omega_k h} - e^{-2\pi i\omega h}\big) \right| \\
&\le \sum_{h=-\infty}^{\infty} |\gamma(h)||e^{-\pi i(\omega_k-\omega)h} - e^{\pi i(\omega_k-\omega)h}| \\
&= 2\sum_{h=-\infty}^{\infty} |\gamma(h)||\sin[\pi(\omega_k-\omega)h]| \\
&\le 2\pi|\omega_k-\omega| \sum_{h=-\infty}^{\infty} |h||\gamma(h)| \\
&\le \frac{\pi\theta L}{n}
\end{aligned}$$

because $|\sin x| \le |x|$. ∎

The main use of the properties described by Lemmas C.1 and C.2 is in identifying the covariance structure of the DFT, say,

$$\begin{aligned}
d(\omega_k) &= n^{-1/2}\sum_{t=1}^{n} x_t\, e^{-2\pi i\omega_k t} \\
&= d_c(\omega_k) - i d_s(\omega_k),
\end{aligned}$$

where

$$d_c(\omega_k) = n^{-1/2}\sum_{t=1}^{n} x_t \cos(2\pi\omega_k t)$$

and

$$d_s(\omega_k) = n^{-1/2}\sum_{t=1}^{n} x_t \sin(2\pi\omega_k t)$$

are the cosine and sine transforms, respectively, of the observed series, defined previously in (4.24) and (4.25). For example, assuming zero means for convenience, we will have

$$\begin{aligned} E[d_c(\omega_k)d_c(\omega_\ell)] &= \frac{1}{4}n^{-1}\sum_{s=1}^{n}\sum_{t=1}^{n}\gamma(s-t)\big(e^{2\pi i\omega_k s}+e^{-2\pi i\omega_k s}\big) \\ &\quad \times\big(e^{2\pi i\omega_\ell t}+e^{-2\pi i\omega_\ell t}\big) \\ &= \frac{1}{4}\big[S_n(-\omega_k,\omega_l)+S_n(\omega_k,\omega_\ell)+S_n(\omega_\ell,\omega_k) \\ &\quad +S_n(\omega_k,-\omega_\ell)\big]. \end{aligned}$$

Lemmas C.1 and C.2 imply, for $k=\ell$,

$$\begin{aligned} E[d_c(\omega_k)d_c(\omega_\ell)] &= \frac{1}{4}\big[O(n^{-1})+f(\omega_k)+O(n^{-1}) \\ &\quad +f(\omega_k)+O(n^{-1})+O(n^{-1})\big] \\ &= \frac{1}{2}f(\omega_k)+O(n^{-1}). \end{aligned} \tag{C.16}$$

For $k\neq\ell$, all terms are $O(n^{-1})$. Hence, we have

$$E[d_c(\omega_k)d_c(\omega_\ell)] = \begin{cases} \frac{1}{2}f(\omega_k)+O(n^{-1}), & k=\ell \\ O(n^{-1}), & k\neq\ell. \end{cases} \tag{C.17}$$

A similar argument gives

$$E[d_s(\omega_k)d_s(\omega_\ell)] = \begin{cases} \frac{1}{2}f(\omega_k)+O(n^{-1}), & k=\ell, \\ O(n^{-1}), & k\neq\ell \end{cases} \tag{C.18}$$

and we also have $E[d_s(\omega_k)d_c(\omega_\ell)]=O(n^{-1})$ for all k,ℓ. We may summarize the results of Lemmas C.1–C.3 as follows.

Theorem C.4 *For a stationary mean zero process with autocovariance function satisfying (C.9) and frequencies $\omega_{k:n}$, such that $|\omega_{k:n}-\omega|<1/n$, are close to some target frequency ω, the cosine and sine transforms (4.24) and (4.25) are approximately uncorrelated with variances equal to $(1/2)f(\omega)$, and the error in the approximation can be uniformly bounded by $\pi\theta L/n$.*

Now, consider estimating the spectrum in a neighborhood of some target frequency ω, using the periodogram estimator

$$I(\omega_{k:n}) = |d(\omega_{k:n})|^2 = d_c^2(\omega_{k:n})+d_s^2(\omega_{k:n}),$$

where we take $|\omega_{k:n}-\omega|\leq n^{-1}$ for each n. In case the series x_t is Gaussian with zero mean,

$$\begin{pmatrix} d_c(\omega_{k:n}) \\ d_s(\omega_{k:n}) \end{pmatrix} \xrightarrow{d} N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{2}\begin{pmatrix} f(\omega) & 0 \\ 0 & f(\omega) \end{pmatrix}\right\},$$

and we have that

$$\frac{2\ I(\omega_{k:n})}{f(\omega)} \xrightarrow{d} \chi^2_2,$$

where χ^2_ν denotes a chi-squared random variable with ν degrees of freedom, as usual. Unfortunately, the distribution does not become more concentrated as $n \to \infty$, because the variance of the periodogram estimator does not go to zero.

We develop a fix for the deficiencies mentioned above by considering the average of the periodogram over a set of frequencies in the neighborhood of ω. For example, we can always find a set of $L = 2m + 1$ frequencies of the form $\{\omega_{j:n} + k/n;\ k = 0, \pm 1, \pm 2, \ldots, m\}$, for which

$$f(\omega_{j:n} + k/n) = f(\omega) + O(Ln^{-1})$$

by Lemma C.3. As n increases, the values of the separate frequencies change.

Now, we can consider the smoothed periodogram estimator, $\widehat{f}(\omega)$, given in (4.56); this case includes the averaged periodogram, $\bar{f}(\omega)$. First, we note that (C.9), $\theta = \sum_{h=-\infty}^{\infty} |h||\gamma(h)| < \infty$, is a crucial condition in the estimation of spectra. In investigating local averages of the periodogram, we will require a condition on the rate of (C.9), namely

$$\sum_{h=-n}^{n} |h||\gamma(h)| = O(n^{-1/2}). \tag{C.19}$$

One can show that a sufficient condition for (C.19) is that the time series is the linear process given by,

$$x_t = \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}, \qquad \sum_{j=0}^{\infty} \sqrt{j}\,|\psi_j| < \infty \tag{C.20}$$

where $w_t \sim \text{iid}(0, \sigma_w^2)$ and w_t has finite fourth moment,

$$E(w_t^4) = \eta\sigma_w^4 < \infty.$$

We leave it to the reader (Problem 4.40) to show (C.20) implies (C.19) under the condition that $w_t \sim wn(0, \sigma_w^2)$.

We now state the following lemma.

Lemma C.4 *Suppose x_t is the linear process given by (C.20), and let $I(\omega_j)$ be the periodogram of the data $\{x_1, \ldots, x_n\}$. Then*

$$\text{cov}\left(I(\omega_j), I(\omega_k)\right) = \begin{cases} 2f^2(\omega_j) + o(1) & \omega_j = \omega_k = 0,\ 1/2 \\ f^2(\omega_j) + o(1) & \omega_j = \omega_k \neq 0,\ 1/2 \\ O(n^{-1}) & \omega_j \neq \omega_k. \end{cases}$$

The proof of Lemma C.4 is straight forward but tedious, and details may be found in Fuller (1976, Theorem 7.2.1) or in Brockwell and Davis (1991, Theorem 10.3.2). For demonstration purposes, we present the proof of the lemma for the pure white noise case; i.e., $x_t = w_t$, in which case $f(\omega) \equiv \sigma_w^2$. By definition, the periodogram in this case is

$$I(\omega_j) = n^{-1} \sum_{s=1}^{n} \sum_{t=1}^{n} w_s w_t e^{2\pi i \omega_j (t-s)},$$

where $\omega_j = j/n$, and hence

$$E\{I(\omega_j) I(\omega_k)\} = n^{-2} \sum_{s=1}^{n} \sum_{t=1}^{n} \sum_{u=1}^{n} \sum_{v=1}^{n} w_s w_t w_u w_v e^{2\pi i \omega_j (t-s)} e^{2\pi i \omega_k (u-v)}.$$

Now when all the subscripts match, $E(w_s w_t w_u w_v) = \eta \sigma_w^4$, when two of the subscripts match, $E(w_s w_t w_u w_v) = \sigma_w^4$, otherwise, $E(w_s w_t w_u w_v) = 0$. Thus,

$$E\{I(\omega_j) I(\omega_k)\} = n^{-1}(\eta - 3)\sigma_w^4 + \sigma_w^4 \left(1 + n^{-2}[A(\omega_j + \omega_k) + A(\omega_k - \omega_j)]\right),$$

where

$$A(u) = \left| \sum_{t=1}^{n} e^{2\pi i u t} \right|^2 .$$

Noting that $EI(\omega_j) = n^{-1} \sum_{t=1}^{n} E(w_t^2) = \sigma_w^2$, we have

$$\begin{aligned} \mathrm{cov}\{I(\omega_j), I(\omega_k)\} &= E\{I(\omega_j) I(\omega_k)\} - \sigma_w^4. \\ &= n^{-1}(\eta - 3)\sigma_w^4 + n^{-2}\sigma_w^4 [A(\omega_j + \omega_k) + A(\omega_k - \omega_j)]. \end{aligned}$$

Thus we conclude that

$$\mathrm{var}\{I(\omega_j)\} = n^{-1}(\eta - 3)\sigma_w^4 + \sigma_w^4 \quad \text{for} \quad \omega_j \neq 0,\, 1/2$$

$$\mathrm{var}\{I(\omega_j)\} = n^{-1}(\eta - 3)\sigma_w^4 + 2\sigma_w^4 \quad \text{for} \quad \omega_j = 0,\, 1/2$$

$$\mathrm{cov}\{I(\omega_j), I(\omega_k)\} = n^{-1}(\eta - 3)\sigma_w^4 \quad \text{for} \quad \omega_j \neq \omega_k,$$

which establishes the result in this case. We also note that if w_t is Gaussian, then $\eta = 3$ and the periodogram ordinates are independent. Using Lemma C.4, we may establish the following fundamental result.

Theorem C.5 *Suppose x_t is the linear process given by (C.20). Then, with $\widehat{f}(\omega)$ defined in (4.56) and corresponding conditions on the weights h_k, we have, as $n \to \infty$,*

(i) $E\left(\widehat{f}(\omega)\right) \to f(\omega)$

(ii) $\left(\sum_{k=-m}^{m} h_k^2\right)^{-1} \mathrm{cov}\left(\widehat{f}(\omega), \widehat{f}(\lambda)\right) \to f^2(\omega) \quad \text{for } \omega = \lambda \neq 0, 1/2.$

In (ii), replace $f^2(\omega)$ *by 0 if* $\omega \neq \lambda$ *and by* $2f^2(\omega)$ *if* $\omega = \lambda = 0$ *or* $1/2$.

Proof. (i): First, recall (4.29)

$$E\left[I(\omega_{j:n})\right] = \sum_{h=-(n-1)}^{n-1} \left(\frac{n-|h|}{n}\right) \gamma(h) e^{-2\pi i \omega_{j:n} h} \stackrel{\text{def}}{=} f_n(\omega_{j:n}).$$

But since $f_n(\omega_{j:n}) \to f(\omega)$ uniformly, and $|f(\omega_{j:n}) - f(\omega_{j:n} + k/n)| \to 0$ by the continuity of f, we have

$$\begin{aligned} E\widehat{f}(\omega) &= \sum_{k=-m}^{m} h_k EI(\omega_{j:n} + k/n) = \sum_{k=-m}^{m} h_k f_n(\omega_{j:n} + k/n) \\ &= \sum_{k=-m}^{m} h_k \left[f(\omega) + o(1)\right] \to f(\omega_j), \end{aligned}$$

because $\sum_{k=-m}^{m} h_k = 1$.

(ii): First, suppose we have $\omega_{j:n} \to \omega_1$ and $\omega_{\ell:n} \to \omega_2$, and $\omega_1 \neq \omega_2$. Then, for n large enough to separate the bands, using Lemma C.4, we have

$$\begin{aligned} \left|\mathrm{cov}\left(\widehat{f}(\omega_1), \widehat{f}(\omega_2)\right)\right| &= \left|\sum_{|k|\le m} \sum_{|r|\le m} h_k\, h_r \mathrm{cov}\left[I(\omega_{j:n}+k/n), I(\omega_{\ell:n}+r/n)\right]\right| \\ &= \left|\sum_{|k|\le m} \sum_{|r|\le m} h_k\, h_r\, O(n^{-1})\right| \\ &\le \frac{c}{n}\left(\sum_{|k|\le m} h_k\right)^2 \qquad \text{(where c is a constant)} \\ &\le \frac{cL}{n}\left(\sum_{|k|\le m} h_k^2\right), \end{aligned}$$

which establishes (ii) for the case of different frequencies. The case of the same frequencies, i.e., $\mathrm{var}[\widehat{f}(\omega_1)]$ is established in a similar manner to the above arguments. ∎

Theorem C.5 justifies the distributional properties used throughout §4.5 and in Chapter 7. We may extend the results of this section to vector series of the form $\boldsymbol{x}_t = (x_{t1}, \ldots, x_{tp})'$, when the cross-spectrum is given by

$$\begin{aligned} f_{ij}(\omega) &= \sum_{h=-\infty}^{\infty} \gamma_{ij}(h) e^{-2\pi i \omega h} \\ &= c_{ij}(\omega) - i q_{ij}(\omega), \end{aligned} \tag{C.21}$$

where

$$c_{ij}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{ij}(h)\cos(2\pi\omega h) \tag{C.22}$$

and

$$q_{ij}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{ij}(h)\sin(2\pi\omega h) \tag{C.23}$$

denote the cospectrum and quadspectrum, respectively. We denote the DFT of the series x_{tj} by

$$\begin{aligned} d_j(\omega_k) &= n^{-1/2}\sum_{t=1}^{n} x_{tj}\ e^{-2\pi i\omega_k t} \\ &= d_{cj}(\omega_k) - id_{sj}(\omega_k), \end{aligned}$$

where d_{cj} and d_{sj} are the cosine and sine transforms of x_{tj}, for $j = 1, 2, \ldots, p$. We bound the covariance structure as before and summarize the results as follows.

Theorem C.6 *The covariance structure of the multivariate cosine and sine transforms, subject to*

$$\theta_{ij} = \sum_{h=-\infty}^{\infty} |h||\gamma_{ij}(h)| < \infty, \tag{C.24}$$

is given by

$$E[d_{ci}(\omega_k)d_{cj}(\omega_\ell)] = \begin{cases} \frac{1}{2}c_{ij}(\omega_k) + O(n^{-1}), & k = \ell \\ O(n^{-1}), & k \neq \ell. \end{cases} \tag{C.25}$$

$$E[d_{ci}(\omega_k)d_{sj}(\omega_\ell)] = \begin{cases} -\frac{1}{2}q_{ij}(\omega_k) + O(n^{-1}), & k = \ell \\ O(n^{-1}), & k \neq \ell \end{cases} \tag{C.26}$$

$$E[d_{si}(\omega_k)d_{cj}(\omega_\ell)] = \begin{cases} \frac{1}{2}q_{ij}(\omega_k) + O(n^{-1}), & k = \ell \\ O(n^{-1}), & k \neq \ell \end{cases} \tag{C.27}$$

$$E[d_{si}(\omega_k)d_{sj}(\omega_\ell)] = \begin{cases} \frac{1}{2}c_{ij}(\omega_k) + O(n^{-1}), & k = \ell \\ O(n^{-1}), & k \neq \ell. \end{cases} \tag{C.28}$$

Proof. We define

$$S_n^{ij}(\omega_k, \omega_\ell) = \sum_{s=1}^{n}\sum_{t=1}^{n} \gamma_{ij}(s-t)e^{-2\pi i\omega_k s}e^{2\pi i\omega_\ell t}. \tag{C.29}$$

Then, we may verify the theorem with manipulations like

$$\begin{aligned} E[d_{ci}(\omega_k)d_{sj}(\omega_k)] &= \frac{1}{4i}\sum_{s=1}^{n}\sum_{t=1}^{n}\gamma_{ij}(s-t)(e^{2\pi i\omega_k s}+e^{-2\pi i\omega_k s}) \\ &\quad \times(e^{2\pi i\omega_k t}-e^{-2\pi i\omega_k t}) \\ &= \frac{1}{4i}\Big[S_n^{ij}(-\omega_k,\omega_k)+S_n^{ij}(\omega_k,\omega_k) \\ &\quad -S_n^{ij}(\omega_k,\omega_k)-S_n^{ij}(\omega_k,-\omega_k)\Big] \\ &= \frac{1}{4i}\Big[c_{ij}(\omega_k)-iq_{ij}(\omega_k) \\ &\quad -(c_{ij}(\omega_k)+iq_{ij}(\omega_k))+O(n^{-1})\Big] \\ &= -\frac{1}{2}q_{ij}(\omega_k)+O(n^{-1}), \end{aligned}$$

where we have used the fact that the properties given in Lemmas C.1–C.3 can be verified for the cross-spectral density functions $f_{ij}(\omega),\ i,j=1,\ldots,p$. ∎

Now, if the underlying multivariate time series $\boldsymbol{x}_t$ is a normal process, it is clear that the DFTs will be jointly normal and we may define the vector DFT, $\boldsymbol{d}(\omega_k)=(d_1(\omega_k),\ldots,d_p(\omega_k))'$ as

$$\begin{aligned} \boldsymbol{d}(\omega_k) &= n^{-1/2}\sum_{t=1}^{n}\boldsymbol{x}_t\ e^{-2\pi\omega_k t} \\ &= \boldsymbol{d}_c(\omega_k)-i\boldsymbol{d}_s(\omega_k), \end{aligned} \tag{C.30}$$

where

$$\boldsymbol{d}_c(\omega_k)=n^{-1/2}\sum_{t=1}^{n}\boldsymbol{x}_t\cos(2\pi\omega_k t) \tag{C.31}$$

and

$$\boldsymbol{d}_s(\omega_k)=n^{-1/2}\sum_{t=1}^{n}\boldsymbol{x}_t\sin(2\pi\omega_k t) \tag{C.32}$$

are the cosine and sine transforms, respectively, of the observed vector series $\boldsymbol{x}_t$. Then, constructing the vector of real and imaginary parts $(\boldsymbol{d}_c'(\omega_k),\boldsymbol{d}_s'(\omega_k))'$, we may note it has mean zero and $2p\times 2p$ covariance matrix

$$\Sigma(\omega_k)=\frac{1}{2}\begin{pmatrix} C(\omega_k) & -Q(\omega_k) \\ Q(\omega_k) & C(\omega_k)\end{pmatrix} \tag{C.33}$$

to order n^{-1} as long as $\omega_k-\omega=O(n^{-1})$. We have introduced the $p\times p$ matrices $C(\omega_k)=\{c_{ij}(\omega_k)\}$ and $Q=\{q_{ij}(\omega_k)\}$. The complex random variable

$\boldsymbol{d}(\omega_k)$ has covariance

$$\begin{aligned} S(\omega_k) &= E[\boldsymbol{d}(\omega_k)\boldsymbol{d}^*(\omega_k)] \\ &= E\Big[\big(\boldsymbol{d}_c(\omega_k) - i\boldsymbol{d}_s(\omega_k)\big)\big(\boldsymbol{d}_c(\omega_k) - i\boldsymbol{d}_s(\omega_k)\big)^*\Big] \\ &= E[\boldsymbol{d}_c(\omega_k)\boldsymbol{d}_c(\omega_k)'] + E[\boldsymbol{d}_s(\omega_k)\boldsymbol{d}_s(\omega_k)'] \\ &\quad -i\big(E[\boldsymbol{d}_s(\omega_k)\boldsymbol{d}_c(\omega_k)'] - E[\boldsymbol{d}_c(\omega_k)\boldsymbol{d}_s(\omega_k)']\big) \\ &= C(\omega_k) - iQ(\omega_k). \end{aligned} \tag{C.34}$$

If the process $\boldsymbol{x}_t$ has a multivariate normal distribution, the complex vector $\boldsymbol{d}(\omega_k)$ has approximately the complex multivariate normal distribution with mean zero and covariance matrix $S(\omega_k) = C(\omega_k) - iQ(\omega_k)$ if the real and imaginary parts have the covariance structure as specified above. In the next section, we work further with this distribution and show how it adapts to the real case. If we wish to estimate the spectral matrix $S(\omega)$, it is natural to take a band of frequencies of the form $\omega_{k:n} + \ell/n$, for $\ell = -m, \ldots, m$ as before, so that the estimator becomes (4.87) of §4.6. A discussion of further properties of the multivariate complex normal distribution is deferred.

It is also of interest to develop a large sample theory for cases in which the underlying distribution is not necessarily normal. If x_t is not necessarily a normal process, some additional conditions are needed to get asymptotic normality. In particular, introduce the notion of a generalized linear process

$$\boldsymbol{y}_t = \sum_{r=-\infty}^{\infty} A_r \boldsymbol{w}_{t-r}, \tag{C.35}$$

where $\boldsymbol{w}_t$ is a $p \times 1$ vector white noise process with $p \times p$ covariance $E[\boldsymbol{w}_t \boldsymbol{w}_t'] = G$ and the $p \times p$ matrices of filter coefficients A_t satisfy

$$\sum_{t=-\infty}^{\infty} \mathrm{tr}\{A_t A_t'\} = \sum_{t=-\infty}^{\infty} \|A_t\|^2 < \infty. \tag{C.36}$$

In particular, stable vector ARMA processes satisfy these conditions. For generalized linear processes, we state the following general result from Hannan (1970, p.224).

Theorem C.7 *If $\boldsymbol{x}_t$ is generated by a generalized linear process with a continuous spectrum that is not zero at ω and $\omega_{k:n} + \ell/n$ are a set of frequencies within L/n of ω, the joint density of the cosine and sine transforms (C.31) and (C.32) converges to that of L independent $2p \times 1$ normal vectors with covariance matrix $\Sigma(\omega)$ with structure given by (C.33). At $\omega = 0$ or $\omega = 1/2$, the distribution is real with covariance matrix $2\Sigma(\omega)$.*

The above result provides the basis for inference involving the Fourier transforms of stationary series because it justifies approximations to the likelihood

function based on multivariate normal theory. We make extensive use of this result in Chapter 7, but will still need a simple form to justify the distributional result for the sample coherence given in (4.93). The next section gives an elementary introduction to the complex normal distribution.

C.3 The Complex Multivariate Normal Distribution

The multivariate normal distribution will be the fundamental tool for expressing the likelihood function and determining approximate maximum likelihood estimators and their large sample probability distributions. A detailed treatment of the multivariate normal distribution can be found in standard texts such as Anderson (1984). We will use the multivariate normal distribution of the $p \times 1$ vector $x = (x_1, x_2, \ldots, x_p)'$, as defined by its density function

$$p(\boldsymbol{x}) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}, \tag{C.37}$$

which can be shown to have mean vector $E[\boldsymbol{x}] = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and covariance matrix

$$\Sigma = E[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})']. \tag{C.38}$$

We use the notation $\boldsymbol{x} \sim N_p\{\boldsymbol{\mu}, \Sigma\}$ for densities of the form (C.37) and note that linearly transformed multivariate normal variables of the form $\boldsymbol{y} = A\boldsymbol{x}$, with A a $q \times p$ matrix $q \leq p$, will also be multivariate normal with distribution

$$\boldsymbol{y} \sim N_q\{A\boldsymbol{\mu}, A\Sigma A'\}. \tag{C.39}$$

Often, the partitioned multivariate normal , based on the vector $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2')'$, split into to $p_1 \times 1$ and $p_2 \times 1$ components $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, respectively, will be used where $p = p_1 + p_2$. If the mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')'$ and covariance matrices

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{C.40}$$

are also compatibly partitioned, the marginal distribution of any subset of components is multivariate normal, say,

$$\boldsymbol{x}_1 \sim N_{p_1}\{\boldsymbol{\mu}_1, \Sigma_{11}\},$$

and that the conditional distribution $\boldsymbol{x}_2$ given $\boldsymbol{x}_1$ is normal with mean

$$E[\boldsymbol{x}_2|\boldsymbol{x}_1] = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) \tag{C.41}$$

and conditional covariance

$$\mathrm{cov}[\boldsymbol{x}_2|\boldsymbol{x}_1] = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \tag{C.42}$$

In the previous section, the real and imaginary parts of the DFT had a partitioned covariance matrix as given in (C.33), and we use this result to say the complex $p \times 1$ vector

$$\boldsymbol{z} = \boldsymbol{x}_1 - i\boldsymbol{x}_2 \tag{C.43}$$

has a complex multivariate normal distribution, with mean vector $\boldsymbol{\mu}_z = \boldsymbol{\mu}_1 - i\boldsymbol{\mu}_2$ and $p \times p$ covariance matrix

$$\Sigma_z = C - iQ \tag{C.44}$$

if the real multivariate $2p \times 1$ normal vector $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2')'$ has a real multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \boldsymbol{\mu}_2')'$ and covariance matrix

$$\Sigma = \frac{1}{2}\begin{pmatrix} C & -Q \\ Q & C \end{pmatrix}. \tag{C.45}$$

The restrictions $C' = C$ and $Q' = -Q$ are necessary for the matrix Σ to be a covariance matrix, and these conditions then imply $\Sigma_z = \Sigma_z^*$ is Hermitian. The probability density function of the complex multivariate normal vector $\boldsymbol{z}$ can be expressed in the concise form

$$p_{\boldsymbol{z}}(\boldsymbol{z}) = \pi^{-p}|\Sigma_z|^{-1}\exp\{-(\boldsymbol{z} - \boldsymbol{\mu}_z)^*\Sigma_z^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z)\}, \tag{C.46}$$

and this is the form that we will often use in the likelihood. The result follows from showing that $p_{\boldsymbol{x}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = p_{\boldsymbol{z}}(\boldsymbol{z})$ exactly, using the fact that the quadratic and Hermitian forms in the exponent are equal and that $|\Sigma_x| = |\Sigma_z|^2$. The second assertion follows directly from the fact that the matrix Σ_x has repeated eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_p$ corresponding to eigenvectors $(\alpha_1', \alpha_2')'$ and the same set, $\lambda_1, \lambda_2, \ldots, \lambda_p$ corresponding to $(\alpha_2', -\alpha_1')'$. Hence

$$|\Sigma_x| = \prod_{i=1}^{p} \lambda_i^2 = |\Sigma_z|^2.$$

For further material relating to the complex multivariate normal distribution, see Goodman (1963), Giri (1965), or Khatri (1965).

Example C.2 A Bivariate Complex Normal Distribution

Consider the joint distribution of the complex random variables $u_1 = x_1 - ix_2$ and $u_2 = y_1 - iy_2$, where the partitioned vector $(x_1, x_2, y_1, y_2)'$ has a real multivariate normal distribution with mean $(0, 0, 0, 0)'$ and covariance matrix

$$\Sigma = \frac{1}{2}\begin{pmatrix} c_{xx} & 0 & c_{xy} & -q_{xy} \\ 0 & c_{xx} & q_{xy} & c_{xy} \\ c_{xy} & q_{xy} & c_{yy} & 0 \\ -q_{xy} & c_{yx} & 0 & c_{yy} \end{pmatrix}. \tag{C.47}$$

Now, consider the conditional distribution of $\boldsymbol{y} = (y_1, y_2)'$, given $\boldsymbol{x} = (x_1, x_2)'$. Using (C.41), we obtain

$$E(\boldsymbol{y} \mid \boldsymbol{x}) = \begin{pmatrix} x_1 & -x_2 \\ x_2 & x_1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \tag{C.48}$$

where

$$(b_1, b_2) = \left(\frac{c_{yx}}{c_{xx}}, \frac{q_{yx}}{c_{xx}} \right). \tag{C.49}$$

It is natural to identify the cross-spectrum

$$f_{xy} = c_{xy} - iq_{xy}, \tag{C.50}$$

so that the complex variable identified with the pair is just

$$\begin{aligned} b &= b_1 - ib_2 \\ &= \frac{c_{yx} - iq_{yx}}{c_{xx}} \\ &= \frac{f_{yx}}{f_{xx}}, \end{aligned}$$

and we identify it as the complex regression coefficient. The conditional covariance follows from (C.42) and simplifies to

$$\operatorname{cov}(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{2} f_{y \cdot x} \, I_2, \tag{C.51}$$

where I_2 denotes the 2×2 identity matrix and

$$\begin{aligned} f_{y \cdot x} &= c_{yy} - \frac{c_{xy}^2 + q_{xy}^2}{c_{xx}} \\ &= f_{yy} - \frac{|f_{xy}|^2}{f_{xx}} \end{aligned} \tag{C.52}$$

Example C.2 leads to an approach for justifying the distributional results for the function coherence given in (4.93). That equation suggests that the result can be derived using the regression results that lead to the F-statistics in §2.2. Suppose that we consider L values of the sine and cosine transforms of the input x_t and output y_t, which we will denote by $d_{x,c}(\omega_k + \ell/n), d_{x,s}(\omega_k + \ell/n), d_{y,c}(\omega_k + \ell/n), d_{y,s}(\omega_k + \ell/n)$, sampled at $L = 2m + 1$ frequencies, $\ell = -m, \ldots, m$, in the neighborhood of some target frequency ω. Suppose these cosine and sine transforms are re-indexed and denoted by $d_{x,cj}, d_{x,sj}, d_{y,cj}, d_{y,sj}$, for $j = 1, 2, \ldots, L$, producing $2L$ real random variables with a large sample normal distribution that have limiting covariance matrices of the form (C.47) for each j. Then, the conditional normal distribution of the 2×1 vector $d_{y,cj}, d_{y,sj}$ given $d_{x,cj}, d_{x,sj}$, given in Example C.2, shows that we may write, approximately, the regression model

$$\begin{pmatrix} d_{y,cj} \\ d_{y,sj} \end{pmatrix} = \begin{pmatrix} d_{x,cj} & -d_{x,sj} \\ d_{x,sj} & d_{x,cj} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} V_{cj} \\ V_{sj} \end{pmatrix},$$

where V_{cj}, V_{sj} are approximately uncorrelated with approximate variances

$$E[V_{cj}^2] = E[V_{sj}^2] = (1/2) f_{y \cdot x}.$$

Now, construct, by stacking, the $2L \times 1$ vectors $\boldsymbol{y}_c = (d_{y,c1}, \ldots, d_{y,cL})'$, $\boldsymbol{y}_s = (d_{y,s1}, \ldots, d_{y,sL})'$, $\boldsymbol{x}_c = (d_{x,c1}, \ldots, d_{x,cL})'$ and $\boldsymbol{x}_s = (d_{x,s1}, \ldots, d_{x,sL})'$, and rewrite the regression model as

$$\begin{pmatrix} \boldsymbol{y}_c \\ \boldsymbol{y}_s \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_c & -\boldsymbol{x}_s \\ \boldsymbol{x}_s & \boldsymbol{x}_c \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{v}_c \\ \boldsymbol{v}_s \end{pmatrix}$$

where $\boldsymbol{v}_c$ and $\boldsymbol{v}_s$ are the error stacks. Finally, write the overall model as the regression model in Chapter 2, namely,

$$\boldsymbol{y} = Z\boldsymbol{b} + \boldsymbol{v},$$

making the obvious identifications in the previous equation. Conditional on Z, the model becomes exactly the regression model considered in Chapter 2 where there are $q = 2$ regression coefficients and $2L$ observations in the observation vector $\boldsymbol{y}$. To test the hypothesis of no regression for that model, we use an F-Statistic that depends on the difference between the residual sum of squares for the full model, say,

$$RSS = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'Z(Z'Z)^{-1}Z'\boldsymbol{y} \tag{C.53}$$

and the residual sum of squares for the reduced model, $RSS_0 = \boldsymbol{y}'\boldsymbol{y}$. Then,

$$F_{2,2L-2} = (L-1)\frac{RSS_0 - RSS}{RSS} \tag{C.54}$$

has the F-distribution with 2 and $2L-2$ degrees of freedom. Also, it follows by substitution for $\boldsymbol{y}$ that

$$\begin{aligned} RSS_0 &= \boldsymbol{y}'\boldsymbol{y} \\ &= \boldsymbol{y}_c'\boldsymbol{y}_c + \boldsymbol{y}_s'\boldsymbol{y}_s \\ &= \sum_{j=1}^{L} (d_{y,cj}^2 + d_{y,sj}^2) \\ &= L\widehat{f}_y(\omega), \end{aligned}$$

which is just the sample spectrum of the output series. Similarly,

$$Z'Z = \begin{pmatrix} L\widehat{f}_x & 0 \\ 0 & L\widehat{f}_x \end{pmatrix}$$

and

$$\begin{aligned} Z'\boldsymbol{y} &= \begin{pmatrix} (\boldsymbol{x}_c'\boldsymbol{y}_c + \boldsymbol{x}_s'\boldsymbol{y}_s) \\ (\boldsymbol{x}_c'\boldsymbol{y}_s - \boldsymbol{x}_s'\boldsymbol{y}_c) \end{pmatrix} \\ &= \begin{pmatrix} \sum_{j=1}^{L}(d_{x,cj}d_{y,cj} + d_{x,sj}d_{y,sj}) \\ \\ \sum_{j=1}^{L}(d_{x,cj}d_{y,sj} - d_{x,sj}d_{y,cj}) \end{pmatrix} \\ &= \begin{pmatrix} L\widehat{c}_{yx} \\ L\widehat{q}_{yx} \end{pmatrix}. \end{aligned}$$

together imply that

$$\boldsymbol{y}'Z(Z'Z)^{-1}Z'\boldsymbol{y} = L\ |\widehat{f}_{xy}|^2/\widehat{f}_x.$$

Substituting into (C.54) gives

$$F_{2,2L-2} = (L-1)\frac{|\widehat{f}_{xy}|^2/\widehat{f}_x}{\left(\widehat{f}_y - |\widehat{f}_{xy}|^2/\widehat{f}_x\right)},$$

which converts directly into the F-statistic (4.93), using the sample coherence defined in (4.92).

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Stat. Math.*, 21, 243-247.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principal. In *2nd Int. Symp. Inform. Theory*, 267-281. B.N. Petrov and F. Csake, eds. Budapest: Akademia Kiado.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Trans. Automat. Contr.*, AC-19, 716-723.

Alagón, J. (1989). Spectral discrimination for two groups of time series. *J. Time Series Anal.*, 10, 203-214.

Alspach, D.L. and H.W. Sorensen (1972). Nonlinear Bayesian estimation using Gaussian sum approximations. *IEEE Trans. Automat. Contr.*, AC-17, 439-447.

Anderson, B.D.O. and J.B. Moore (1979). *Optimal Filtering.* Englewood Cliffs, NJ: Prentice-Hall.

Anderson, T.W. (1978). Estimation for autoregressive moving average models in the time and frequency domain. *Ann. Stat.*, 5, 842-865.

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis, 2nd ed.* New York: Wiley.

Ansley, C.F. and P. Newbold (1980). Finite sample properties of estimators for autoregressive moving average processes. *J. Econ.*, 13, 159-183.

Antognini, J.F., M.H. Buonocore, E.A. Disbrow, and E. Carstens (1997). Isoflurane anesthesia blunts cerebral responses to noxious and innocuous stimuli: a fMRI study. *Life Sci.*, 61, PL349-PL354.

Bandettini, A., A. Jesmanowicz, E.C. Wong, and J.S. Hyde (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Res. Med.*, 30, 161-173.

Bar-Shalom, Y. (1978). Tracking methods in a multi-target environment. *IEEE Trans. Automat. Contr.*, AC-23, 618-626.

Bar-Shalom, Y. and E. Tse (1975). Tracking in a cluttered environment with probabilistic data association. *Automatica*, 11, 4451-4460.

Bazza, M., R.H. Shumway, and D.R. Nielsen (1988). Two-dimensional spectral analysis of soil surface temperatures. *Hilgardia*, 56, 1-28.

Bedrick, E.J. and C.-L. Tsai (1994). Model selection for multivariate regression in small samples. *Biometrics*, 50, 226-231.

Beran, J. (1994). *Statistics for Long Memory Processes.* New York: Chapman and Hall.

Berk, K.N. (1974). Consistent autoregressive spectral estimates. *Ann. Stat.*, 2, 489-502.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Stat. Soc. B*, 36, 192-236.

Bhat, R.R. (1985). *Modern Probability Theory, 2nd ed.* New York: Wiley.

Bhattacharya, A. (1943). On a measure of divergence between two statistical populations. *Bull. Calcutta Math. Soc.*, 35, 99-109.

Blackman, R.B. and J.W. Tukey (1959). *The Measurement of Power Spectra from the Point of View of Communications Engineering.* New York: Dover.

Blight, B.J.N. (1974). Recursive solutions for the estimation of a stochastic parameter *J. Am. Stat. Assoc.*, 69, 477-481

Bloomfield, P. (1976). *Fourier Analysis of Time Series: An Introduction.* New York: Wiley.

Bloomfield, P. (2000). *Fourier Analysis of Time Series: An Introduction, 2nd ed.* New York: Wiley.

Bloomfield, P. and J.M. Davis (1994). Orthogonal rotation of complex principal components. *Int. J. Climatol.*, 14, 759-775.

Bogart, B. P., M. J. R. Healy, and J.W. Tukey (1962). The Quefrency Analysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking. In *Proc. of the Symposium on Time Series Analysis*, pp. 209-243, Brown University, Providence, USA.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *J. Econ.*, 31, 307- 327.

Box, G.E.P. and G.M. Jenkins (1970). *Time Series Analysis, Forecasting, and Control.* Oakland, CA: Holden-Day.

Box, G.E.P., G.M. Jenkins and G.C. Reinsel (1994). *Time Series Analysis, Forecasting, and Control, 3rd ed.* Englewood Cliffs, NJ: Prentice Hall.

Box, G.E.P. and D.A. Pierce (1970). Distributions of residual autocorrelations in autoregressive integrated moving average models. *J. Am. Stat. Assoc.*, 72, 397-402.

Box, G.E.P. and G.C. Tiao (1973). *Bayesian Inference in Statistical Analysis.* New York: Wiley.

Breiman, L. and J. Friedman (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Am. Stat. Assoc.*,

80, 580-619.

Brillinger, D.R. (1973). The analysis of time series collected in an experimental design. In *Multivariate Analysis-III.*, pp. 241-256. P.R. Krishnaiah ed. New York: Academic Press.

Brillinger, D.R. (1980). Analysis of variance and problems under time series models. In *Handbook of Statistics*, Vol I, pp. 237-278. P.R. Krishnaiah and D.R. Brillinger, eds. Amsterdam: North Holland.

Brillinger, D.R. (1981). *Time Series: Data Analysis and Theory, 2nd ed.* San Francisco: Holden-Day. Republished in 2001 by the Society for Industrial and Applied Mathematics, Philadelphia.

Brockwell, P.J. and R.A. Davis (1991). *Time Series: Theory and Methods, 2nd ed.* New York: Springer-Verlag.

Bruce, A. and H-Y. Gao (1996). *Applied Wavelet Analysis with S-PLUS*. New York: Springer-Verlag.

Caines, P.E. (1988). *Linear Stochastic Systems*. New York: Wiley.

Carlin, B.P., N.G. Polson, and D.S. Stoffer (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. Am. Stat. Assoc.*, 87, 493-500.

Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika*, 81, 541-553.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of the observations. *Ann. Math. Stat.*, 25, 573-578.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, 74, 829-836.

Cochrane, D. and G.H. Orcutt (1949). Applications of least squares regression to relationships containing autocorrelated errors. *J. Am. Stat. Assoc.*, 44, 32-61.

Cooley, J.W. and J.W. Tukey (1965). An algorithm for the machine computation of complex Fourier series. *Math. Comput.*, 19, 297-301.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*. New York: Wiley.

Dahlhaus, R. (1989). Efficient parameter estimation for self-similar processes. *Ann. Stat.*, 17, 1749-1766.

Dargahi-Noubary, G.R. and P.J. Laycock (1981). Spectral ratio discriminants and information theory. *J. Time Series Anal.*, 16, 201-219.

Danielson, J. (1994). Stochastic volatility in asset prices: Estimation with simulated maximum likelihood. *J. Econometrics*, 61, 375-400.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia: CBMS-NSF Regional Conference Series in Applied Mathematics.

Davies, N., C.M. Triggs, and P. Newbold (1977). Significance levels of the

Box-Pierce portmanteau statistic in finite samples. *Biometrika*, 64, 517-522.

Dent, W. and A.-S. Min. (1978). A Monte Carlo study of autoregressive-integrated-moving average processes. *J. Econ.*, 7, 23-55.

Dempster, A.P., N.M. Laird and D.B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, 39, 1-38.

Diggle, P.J., K.-Y. Liang, and S.L. Zeger (1994). *The Analysis of Longitudinal Data*. Oxford: Clarendon Press.

Ding, Z., C.W.J. Granger, and R.F. Engle (1993). A long memory property of stock market returns and a new model. *J. Empirical Finance*, 1, 83-106.

Donoho, D.L. and I.M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 425-455.

Donoho, D.L. and I.M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. of Am. Stat. Assoc.*, 90, 1200-1224.

Durbin, J. (1960). Estimation of parameters in time series regression models. *J. R. Stat. Soc. B*, 22, 139-153.

Durbin, J. and S.J. Koopman (2001). *Time Series Analysis by State Space Methods* Oxford: Oxford University Press.

Efron, B. and R. Tibshirani (1994). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50, 987-1007.

Engle, R.F., D. Nelson, and T. Bollerslev (1994). ARCH Models. In *Handbook of Econometrics*, Vol IV, pp. 2959-3038. R. Engle and D. McFadden, eds. Amsterdam: North Holland.

Fahrmeir, L. and G. Tutz (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer-Verlag.

Fox, R. and M.S. Taqqu (1986). Large sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Ann. Stat.*, 14, 517-532.

Friedman, J.H. (1984). A Variable Span Smoother. Tech. Rep. No. 5, Lab. for Computational Statistics, Dept. Statistics, Stanford Univ., California.

Friedman, J.H. and W. Stuetzle. (1981). Projection pursuit regression. *J. Am. Stat. Assoc.*, 76, 817-823.

Frühwirth-Schnatter, S. (1994). Data Augmentation and Dynamic Linear Models. *J. Time Series Anal.*, 15, 183–202.

Fuller, W.A. (1976). *Introduction to Statistical Time Series*. New York: Wiley.

Fuller, W.A. (1995). *Introduction to Statistical Time Series, 2nd ed.* New York: Wiley.

Gabr, M.M. and T. Subba-Rao (1981). The estimation and prediction of subset bilinear time series models with applications. *J. Time Series Anal.*, 2, 155-171.

Gelfand, A.E. and A.F.M. Smith (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, 85, 398-409.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, 6, 721-741.

Geweke, J.F. (1977). The dynamic factor analysis of economic time series models. In *Latent Variables in Socio-Economic Models*, pp 365-383. D. Aigner and A. Goldberger, eds. Amsterdam: North Holland.

Geweke, J.F. and K.J. Singleton (1981). Latent variable models for time series: A frequency domain approach with an application to the Permanent Income Hypothesis. *J. Econ.*, 17, 287-304.

Geweke, J.F. and S. Porter-Hudak (1983). The estimation and application of long-memory time series models. *J. Time Series Anal.*, 4, 221-238.

Gilks, W.R., S. Richardson, and D.J. Spiegelhalter (eds.) (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Giri, N. (1965). On complex analogues of T^2 and R^2 tests. *Ann. Math. Stat.*, 36, 664-670.

Goldfeld, S.M. and R.E. Quandt (1973). A Markov model for switching regressions. *J. Econ.*, 1, 3-16.

Goodman, N.R. (1963). Statistical analysis based on a certain multivariate complex Gaussian distribution. *Ann. Math. Stat.*, 34, 152-177.

Goodrich, R.L. and P.E. Caines (1979). Linear system identification from nonstationary cross-sectional data. *IEEE Trans. Automat. Contr.*, AC-24, 403-411.

Gordon, K. and A.F.M. Smith (1990). Modeling and monitoring biomedical time series. *J. Am. Stat. Assoc.*, 85, 328-337.

Gouriéroux, C. (1997). *ARCH Models and Financial Applications*. New York: Springer-Verlag.

Granger, C.W. and R. Joyeux (1980). An introduction to long-memory time series models and fractional differencing. *J. Time Series Anal.*, 1, 15-29.

Grether, D.M. and M. Nerlove (1970). Some properties of optimal seasonal adjustment. *Econometrica*, 38, 682-703.

Gupta N.K. and R.K. Mehra (1974). Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations. *IEEE Trans. Automat. Contr.*, AC-19, 774-783.

Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57, 357-384.

Hannan, E.J. (1970). *Multiple Time Series*. New York: Wiley.

Hannan, E.J. (1973). The asymptotic theory of linear time series models. *J. Appl. Prob.*, 10, 130-145.

Hannan, E.J. and M. Deistler (1988). *The Statistical Theory of Linear Systems*. New York: Wiley.

Hansen, J. and S. Lebedeff (1987). Global trends of measured surface air temperature. *J. Geophys. Res.*, 92, 1345-1372.

Hansen, J. and S. Lebedeff (1988). Global surface air temperatures: Update through 1987. *J. Geophys. Lett.*, 15, 323-326.

Harrison, P.J. and C.F. Stevens (1976). Bayesian forecasting (with discussion). *J. R. Stat. Soc. B*, 38, 205-247.

Harvey, A.C. and P.H.J. Todd (1983). Forecasting economic time series with structural and Box-Jenkins models: A case study. *J. Bus. Econ. Stat.*, 1, 299-307.

Harvey, A.C. and R.G. Pierse (1984). Estimating missing observations in economic time series. *J. Am. Stat. Assoc.*, 79, 125-131.

Harvey, A.C. (1991). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, A.C. (1993). *Time Series Models*. Cambridge, MA: MIT Press.

Harvey A.C., E. Ruiz and N. Shephard (1994). Multivariate stochastic volatility models. *Rev. Economic Studies*, 61, 247-264.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109.

Hosking, J.R.M. (1981). Fractional differencing. *Biometrika*, 68, 165-176.

Hurst, H. (1951). Long term storage capacity of reservoirs. *Trans. Am. Soc. Civil Eng.*, 116, 778-808.

Hurvich, C.M. and S. Zeger (1987). Frequency domain bootstrap methods for time series. *Tech. Report 87-115*, Department of Statistics and Operations Research, Stern School of Business, New York University.

Hurvich, C.M and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.

Hurvich, C.M. and K.I. Beltrao (1993). Asymptotics for the low-requency oridnates of the periodogram for a long-memory time series. *J. Time Series Anal.*, 14, 455-472.

Hurvich, C.M., R. Deo and J. Brodsky (1998). The mean squared error of Geweke and Porter-Hudak's estimator of the memory parameter of a long-memory time series. *J. Time Series Anal.*, 19, 19-46.

Jacquier, E., N.G. Polson, and P.E. Rossi (1994). Bayesian analysis of stochastic volatility models. *J. Bus. Econ. Stat.*, 12, 371-417.

Jazwinski, A.H. (1970). *Stochastic Processes and Filtering Theory*. New York: Academic Press.

Jenkins, G.M. and D.G. Watts. (1968). *Spectral Analysis and Its Applications*. San Francisco: Holden-Day.

Johnson, R.A. and D.W. Wichern (1992). *Applied Multivariate Statistical Analysis, 3rd ed.*. Englewood Cliffs, NJ: Prentice-Hall.

Jones, P.D. (1994). Hemispheric surface air temperature variations: A reanalysis and an update to 1993. *J. Clim.*, 7, 1794-1802.

Jones, R.H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics*, 22, 389-395.

Jones, R.H. (1984). Fitting multivariate models to unequally spaced data. In *Time Series Analysis of Irregularly Observed Data*, pp. 158-188. E. Parzen, ed. Lecture Notes in Statistics, 25, New York: Springer-Verlag.

Jones, R.H. (1993). *Longitudinal Data With Serial Correlation : A State-Space Approach*. London: Chapman and Hall.

Journel, A.G. and C.H. Huijbregts (1978). *Mining Geostatistics*. New York: Academic Press.

Juang, B.H. and L.R. Rabiner (1985). Mixture autoregressive hidden Markov models for speech signals, *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-33, 1404-1413.

Kakizawa, Y., R. H. Shumway, and M. Taniguchi (1998). Discrimination and clustering for multivariate time series. *J. Am. Stat. Assoc.*, 93, 328-340.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Trans ASME J. Basic Eng.*, 82, 35-45.

Kalman, R.E. and R.S. Bucy (1961). New results in filtering and prediction theory. *Trans. ASME J. Basic Eng.,* 83, 95-108.

Kay, S.M. (1988). *Modern Spectral Analysis: Theory and Applications*. Englewood Cliffs, NJ: Prentice-Hall.

Kazakos, D. and P. Papantoni-Kazakos (1980). Spectral distance measuring between Gaussian processes. *IEEE Trans. Automat. Contr.*, AC-25, 950-959.

Khatri, C.G. (1965). Classical statistical analysis based on a certain multivariate complex Gaussian distribution. *Ann. Math. Stat.*, 36, 115-119.

Kim S., N. Shephard and S. Chib (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *Rev. Economic Studies*, 65, p.361-393.

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series (with discussion). *J. Am. Stat. Assoc.*, 82, 1032-1041, (C/R: p1041-1063; C/R: V83 p1231).

Kitagawa, G. and W. Gersch (1984). A smoothness priors modeling of time series with trend and seasonality. *J. Am. Stat. Assoc.*, 79, 378-389.

Kitagawa, G. and W. Gersch (1996). *Smoothness Priors Analysis of Time Series*. New York: Springer-Verlag.

Kolmogorov, A.N. (1941). Interpolation and extrapolation von stationären zufälligen folgen. *Bull. Acad. Sci. U.R.S.S.*, 5, 3-14.

Krishnaiah, P.R., J.C. Lee, and T.C. Chang (1976). The distribution of likelihood ratio statistics for tests of certain covariance structures of complex multivariate normal populations. *Biometrika*, 63, 543-549.

Kullback, S. and R.A. Leibler (1951). On information and sufficiency. *Ann. Math. Stat.*, 22, 79-86.

Kullback, S. (1958). *Information Theory and Statistics*. Gloucester, MA: Peter Smith.

Lachenbruch, P.A. and M.R. Mickey (1968). Estimation of error rates in discriminant analysis. *Technometrices*, 10, 1-11.

Laird, N. and J. Ware (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.

Lam, P.S. (1990). The Hamilton model with a general autoregressive component: Estimation and comparison with other modles of economic time series. *J. Monetary Econ.*, 26, 409-432.

Lay, T. (1997). Research required to support comprehensive nuclear test ban treaty monitoring. *National Research Council Report, National Academy Press*, 2101 Constitution Ave., Washington, DC 20055.

Levinson, N. (1947). The Wiener (root mean square) error criterion in filter design and prediction. *J. Math. Phys.*, 25, 262-278.

Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scand. J. Stat.*, 5, 81-91.

Liu, L.M. (1991). Dynamic relationship analysis of U.S. gasoline and crude oil prices. *J. Forecast.*, **10**, 521-547.

Ljung, G.M. and G.E.P. Box (1978). On a measure of lack of fit in time series models. *Biometrika*, 65, 297-303.

Lütkepohl, H. (1985). Comparison of criteria for estimating the order of a vector autoregressive process. *J. Time Series Anal.*, 6, 35-52.

Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis, 2nd ed.* Berlin: Springer-Verlag.

McQuarrie, A.D.R. and R.H. Shumway (1994). *ASTSA for Windows*.

McQuarrie, A.D.R. and C-L. Tsai (1998). *Regression and Time Series Model Selection*, Singapore: World Scientific.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, 15, 661-675.

McBratney, A.B. and R. Webster (1981). Detection of ridge and furrow pattern by spectral analysis of crop yield. *Int. Stat. Rev.*, 49, 45-52.

McCulloch, R.E. and R.S. Tsay (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *J. Am. Stat. Assoc.*, 88, 968-978.

McDougall, A. J., D.S. Stoffer and D.E. Tyler (1997). Optimal transformations and the spectral envelope for real-valued time series. *J. Stat. Plan. Infer.*, **57**, 195-214.

McLeod A.I. (1978). On the distribution of residual autocorrelations in Box-Jenkins models. *J. R. Stat. Soc. B*, 40, 296-302.

McLeod, A.I. and K.W. Hipel (1978). Preservation of the rescaled adusted range, I. A reassessment of the Hurst phenomenon. *Water Resour. Res.*, 14, 491-508.

Meinhold, R.J. and N.D. Singpurwalla (1983). Understanding the Kalman filter. *Am. Stat.*, 37, 123-127.

Meinhold, R.J. and N.D. Singpurwalla (1989). Robustification of Kalman filter models. *J. Am. Stat. Assoc.*, 84, 479-486.

Meng X.L. and Rubin, D.B. (1991). Using EM to obtain asymptotic variance–covariance matrices: The SEM algorithm. *J. Am. Stat. Assoc.*, 86, 899-909.

Metropolis N., A.W. Rosenbluth, M.N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21, 1087-1091.

Mickens, R.E. (1987). *Difference Equations.* New York: Van Nostrand Reinhold.

Newbold, P. and T. Bos (1985). *Stochastic Parameter Regression Models.* Beverly Hills: Sage.

Ogawa, S., T.M. Lee, A. Nayak and P. Glynn (1990). Oxygenation-sensititive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn. Reson. Med.*, 14, 68-78.

Palma, W. and N.H. Chan (1997). Estimation and forecasting of long-memory time series with missing values. *J. Forecast.*, 16, 395-410.

Paparoditis, E. and Politis, D.N. (1999). The local bootstrap for periodogram statistics. *J. Time Series Anal.*, 20, 193-222.

Parker, D.E., P.D. Jones, A. Bevan and C.K. Folland (1994). Interdecadal changes of surface temperature since the late 19th century. *J. Geophysical Research*, 90, 14373-14399.

Parker, D.E., C.K. Folland and M. Jackson (1995). Marine surface temperature: observed variations and data requirements. *Climatic Change*, 31, 559-60

Parzen, E. (1961). Mathematical considerations in the estimation of spectra. *Technometrics*, 3, 167-190.

Parzen, E. (1983). Autoregressive spectral estimation. In *Time Series in the Frequency Domain, Handbook of Statistics*, Vol. 3, pp. 211-243. D.R. Brillinger and P.R. Krishnaiah eds. Amsterdam: North Holland.

Pawitan, Y. and R.H. Shumway (1989). Spectral estimation and deconvolution for a linear time series model. *J. Time Series Anal.*, 10, 115-129.

Peña, D. and I. Guttman (1988). A Bayesian approach to robustifying the Kalman filter. In *Bayesian Analysis of Time Series and Dynamic Linear Models*, pp. 227-254. J.C. Spall, ed. New York: Marcel Dekker.

Percival, D.B. and A.T. Walden (1993). *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques* Cambridge: Cambridge University Press.

Percival, D.B. and A.T. Walden (2000). *Wavelet Methods for Time Series Analysis*. Cambridge: Cambridge University Press.

Pinsker, M.S. (1964). *Information and Information Stability of Random Variables and Processes*, San Francisco: Holden Day.

Pole, P.J. and M. West (1988). Nonnormal and nonlinear dynamic Bayesian modeling. In *Bayesian Analysis of Time Series and Dynamic Linear Models*, pp. 167-198. J.C. Spall, ed. New York: Marcel Dekker.

Press, W.H., S.A. Teukolsky, W. T. Vetterling, and B.P. Flannery (1993). *Numerical Recipes in C: The Art of Scientific Computing, 2nd ed.* Cambridge: Cambridge University Press.

Priestley, M.B., T. Subba-Rao and H. Tong (1974). Applications of principal components analysis and factor analysis in the identification of multivariable systems. *IEEE Trans. Automat. Contr.*, AC-19, 730-734.

Priestley, M.B. and T. Subba-Rao (1975). The estimation of factor scores and Kalman filtering for discrete parameter stationary processes. *Int. J. Contr.*, 21, 971-975.

Priestley, M.B. (1981). *Spectral Analysis and Time Series*. Vol. 1: Univariate Series; Vol 2: Multivariate Series, Prediction and Control. New York: Academic Press.

Priestley, M.B. (1988). *Nonlinear and Nonstationary Time Series Analysis*. London: Academic Press.

Quandt, R.E. (1972). A new approach to estimating switching regressions. *J. Am. Stat. Assoc.*, 67, 306-310.

Rabiner, L.R. and B.H. Juang (1986). An introduction to hidden Markov models, *IEEE Acoust., Speech, Signal Process.*, ASSP-34, 4-16.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: Wiley.

Rauch, H.E., F. Tung, and C.T. Striebel (1965). Maximum likelihood estimation of linear dynamic systems. *J. AIAA*, 3, 1445-1450.

Reinsel, G.C. (1997). *Elements of Multivariate Time Series Analysis, 2nd ed.* New York: Springer-Verlag.

Renyi, A. (1961). On measures of entropy and information. In *Proceedings of 4th Berkeley Symp. Math. Stat. and Probability*, pp. 547-561, Berkeley: Univ. of California Press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14, 465-471.

Robinson, P.M. (1995). Gaussian semiparametric estimation of long range dependence. *Ann. Stat.*, 23, 1630-1661.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci.*, 42, 43-47.

Royston, P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31, 115-124.

Sandmann, G. and S.J. Koopman (1998). Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *J. Econometrics*, 87 , 271-301.

Sargan, J.D. (1964). Wages and prices in the United Kingdom: A study in econometric methodology. In *Econometric Analysis for National Economic Planning*, eds. P. E. Hart, G. Mills and J. K. Whitaker. London: Butterworths. reprinted in *Quantitative Economics and Econometric Analysis*, pp. 275-314, eds. K. F. Wallis and D. F. Hendry (1984). Oxford: Basil Blackwell.

Scheffé, H. (1959). *The Analysis of Variance.* New York: Wiley.

Schuster, A. (1898). On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terrestrial Magnetism,* III, 11-41.

Schuster, A. (1906). On the periodicities of sunspots. *Phil. Trans. R. Soc., Ser. A*, 206, 69-100.

Schwarz, F. (1978). Estimating the dimension of a model. *Ann. Stat.*, 6, 461-464.

Schweppe, F.C. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Trans. Inform. Theory*, IT-4, 294-305.

Seber, G.A.G. (1977). *Linear Regression Analysis.* New York: Wiley.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In *Time Series Models in Econometrics, Finance and Other Fields* , pp 1-100. D.R. Cox, D.V. Hinkley, and O.E. Barndorff-Nielson eds. London: Chapman and Hall.

Shumway, R.H. and W.C. Dean (1968). Best linear unbiased estimation for multivariate stationary processes. *Technometrics*, 10, 523-534.

Shumway, R.H. (1970). Applied regression and analysis of variance for stationary time series. *J. Am. Stat. Assoc.*, 65, 1527-1546.

Shumway, R.H. (1971). On detecting a signal in N stationarily correlated noise series. *Technometrics*, 10, 523-534.

Shumway, R.H. and A.N. Unger (1974). Linear discriminant functions for stationary time series. *J. Am. Stat. Assoc.*, 69, 948-956.

Shumway, R.H. (1982). Discriminant analysis for time series. In *Classification, Pattern Recognition and Reduction of Dimensionality, Handbook of Statistics Vol. 2*, pp. 1-46. P.R. Krishnaiah and L.N. Kanal, eds. Amsterdam: North Holland.

Shumway, R.H. and D.S. Stoffer (1982). An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Series Anal.*, 3, 253-264.

Shumway, R.H. (1983). Replicated time series regression: An approach to signal estimation and detection. In *Time Series in the Frequency Domain, Handbook of Statistics Vol. 3*, pp. 383-408. D.R. Brillinger and P.R. Krishnaiah, eds. Amsterdam: North Holland.

Shumway, R.H. (1988). *Applied Statistical Time Series Analysis.* Englewood Cliffs, NJ: Prentice-Hall.

Shumway, R.H., R.S. Azari, and Y. Pawitan (1988). Modeling mortality fluctuations in Los Angeles as functions of pollution and weather effects. *Environ. Res.*, 45, 224-241.

Shumway, R.H. and D.S. Stoffer (1991). Dynamic linear models with switching. *J. Am. Stat. Assoc.*, 86, 763-769, (Correction: V87 p. 913).

Shumway, R.H. and K.L. Verosub (1992). State space modeling of paleoclimatic time series. In *Pro. 5th Int. Meeting Stat. Climatol.*. Toronto, pp. 22-26, June, 1992.

Shumway, R.H. (1996). Statistical approaches to seismic discrimination. In *Monitoring a Comprehensive Test Ban Treaty*, pp. 791-803. A.M. Dainty and E.S. Husebye eds. Doordrecht, The Netherlands: Kluwer Academic

Shumway, R.H., S.E. Kim and R.R. Blandford (1999). Nonlinear estimation for time series observed on arrays. Chapter 7, S. Ghosh, ed. *Asymptotics, Nonparametrics and Time Series*, pp. 227-258. New York: Marcel Dekker.

Small, C.G. and D.L. McLeish (1994). *Hilbert Space Methods in Probability and Statistical Inference.* New York: Wiley.

Smith, A.F.M. and M. West (1983). Monitoring renal transplants: An application of the multiprocess Kalman filter. *Biometrics*, 39, 867-878.

Spliid, H. (1983). A fast estimation method for the vector autoregressive moving average model with exogenous variables. *J. Am. Stat. Assoc.*, 78, 843-849.

Stoffer, D.S. (1982). Estimation of Parameters in a Linear Dynamic System with Missing Observations. Ph.D. Dissertation. Univ. California, Davis.

Stoffer, D.S. (1987). Walsh-Fourier analysis of discrete-valued time series. *J. Time Series Anal.*, 8, 449-467.

Stoffer, D.S., M. Scher, G. Richardson, N. Day, and P. Coble (1988). A Walsh-Fourier analysis of the effects of moderate maternal alcohol consumption on neonatal sleep-state cycling. *J. Am. Stat. Assoc.*, 83, 954-963.

Stoffer, D.S. and K.D. Wall (1991). Bootstrapping state space models: Gaussian maximum likelihood estimation and the Kalman filter. *J. Am. Stat. Assoc.*, 86, 1024-1033.

Stoffer, D.S., D.E. Tyler, and A.J. McDougall (1993). Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika*, 80, 611-622.

Stoffer, D.S. and D.E. Tyler (1998). Matching sequences: Cross-spectral analysis of categorical time series *Biometrika*, 85, 201-213.

Stoffer, D.S. (1999). Detecting common signals in multiple time series using the spectral envelope. *J. Am. Stat. Assoc.*, 94, 1341-1356.

Stoffer, D.S. and K.D. Wall (2004). Resampling in State Space Models. In *State Space and Unobserved Component Models Theory and Applications*, Chapter 9, pp. 227-258. Andrew Harvey, Siem Jan Koopman, and Neil Shephard, eds. Cambridge: Cambridge University Press.

Subba-Rao, T. (1981). On the theory of bilinear time series models. *J. R. Stat. Soc. B*, 43, 244-255.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections, *Commun. Statist, A, Theory Methods*, 7, 13-26.

Taniguchi, M., M.L. Puri, and M. Kondo (1994). Nonparametric approach for non-Gaussian vector stationary processes. *J. Mult. Anal.*, 56, 259-283.

Tanner, M. and W.H. Wong (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Stat. Assoc.*, 82, 528-554.

Tiao, G.C. and R.S. Tsay (1989). Model specification in multivariate time series (with discussion). *J. Roy. Statist. Soc. B*, 51, 157-213.

Tiao, G. C. and R.S. Tsay (1994). Some advances in nonlinear and adaptive modeling in time series analysis. *J. Forecast.*, 13, 109-131.

Tiao, G.C., R.S. Tsay and T .Wang (1993). Usefulness of linear transformations in multivariate time series analysis. *Empir. Econ.*, 18, 567-593.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Stat.*, 22, 1701-1728.

Tong, H. (1983). *Threshold Models in Nonlinear Time Series Analysis*. Springer Lecture Notes in Statistics, 21. New York: Springer-Verlag.

Tong, H. (1990). *Nonlinear Time Series: A Dynamical System Approach*. Oxford: Oxford Univ. Press.

Tsay, R. (1987). Conditional hetereroscadasticity in time series analysis. *J. Am. Stat. Assoc.*, 82, 590-604.

Venables, W.N. and B.D. Ripley (1994). *Modern Applied Statistics with S-Plus*. New York: Springer-Verlag.

Walker, G. (1931). On periodicity in series of related terms. *Proc. R. Soc. Lond., Ser. A*, 131, 518-532.

Watson, G.S. (1966). Smooth regression analysis. *Sankhya*, 26, 359-378.

Weiss, A.A. (1984). ARMA models with ARCH errors. *J. Time Series Anal.*, 5, 129-143.

West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models 2nd ed*. New York: Springer-Verlag.

Whittle, P. (1961). Gaussian estimation in stationary time series. *Bull. Int. Stat. Inst.*, 33, 1-26.

Wiener, N. (1949). *The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*. New York: Wiley.

Wu, C.F. (1983). On the convergence properties of the EM algorithm. *Ann. Stat.*, 11, 95-103.

Young, P.C. and D.J. Pedregal (1998). Macro-economic relativity: Government spending, private investment and unemployment in the USA. Centre for Research on Environmental Systems and Statistics, Lancaster University, U.K.

Yule, G.U. (1927). On a method of investigating periodicities in disturbed series with special reference to Wolfer's Sunspot Numbers. *Phil. Trans. R. Soc. Lond.*, A226, 267-298.

Index

Springer Texts in Statistics *(continued from page ii)*

Lehmann and Romano: Testing Statistical Hypotheses, Third Edition
Lehmann and Casella: Theory of Point Estimation, Second Edition
Lindman: Analysis of Variance in Experimental Design
Lindsey: Applying Generalized Linear Models
Madansky: Prescriptions for Working Statisticians
McPherson: Applying and Interpreting Statistics: A Comprehensive Guide, Second Edition
Mueller: Basic Principles of Structural Equation Modeling: An Introduction to LISREL and EQS
Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume I: Probability for Statistics
Nguyen and Rogers: Fundamentals of Mathematical Statistics: Volume II: Statistical Inference
Noether: Introduction to Statistics: The Nonparametric Way
Nolan and Speed: Stat Labs: Mathematical Statistics Through Applications
Peters: Counting for Something: Statistical Principles and Personalities
Pfeiffer: Probability for Applications
Pitman: Probability
Rawlings, Pantula and Dickey: Applied Regression Analysis
Robert: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, Second Edition
Robert and Casella: Monte Carlo Statistical Methods
Rose and Smith: Mathematical Statistics with *Mathematica*
Ruppert: Statistics and Finance: An Introduction
Santner and Duffy: The Statistical Analysis of Discrete Data
Saville and Wood: Statistical Methods: The Geometric Approach
Sen and Srivastava: Regression Analysis: Theory, Methods, and Applications
Shao: Mathematical Statistics, Second Edition
Shorack: Probability for Statisticians
Shumway and Stoffer: Time Series Analysis and Its Applications: With R Examples, Second Edition
Simonoff: Analyzing Categorical Data
Terrell: Mathematical Statistics: A Unified Introduction
Timm: Applied Multivariate Analysis
Toutenburg: Statistical Analysis of Designed Experiments, Second Edition
Wasserman: All of Nonparametric Statistics
Wasserman: All of Statistics: A Concise Course in Statistical Inference
Weiss: Modeling Longitudinal Data
Whittle: Probability via Expectation, Fourth Edition
Zacks: Introduction to Reliability Analysis: Probability Models and Statistical Methods