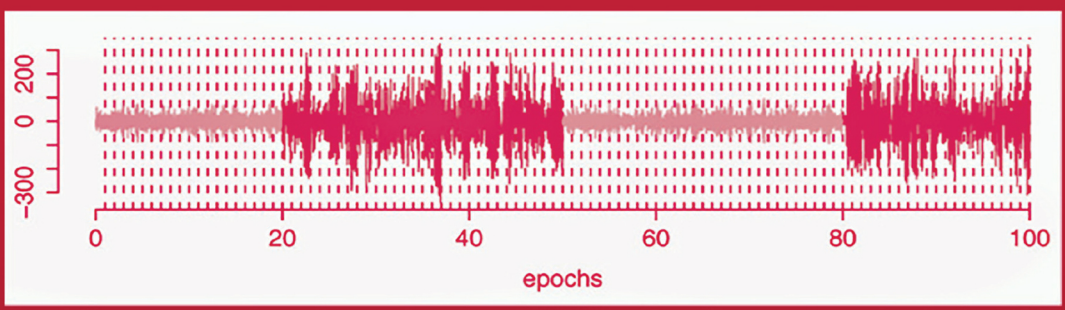
Texts in Statistical Science

Time Series

Modeling, Computation, and Inference

Second Edition



Raquel Prado
Marco A. R. Ferreira
Mike West

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Time Series

CHAPMAN & HALL/CRC
Texts in Statistical Science Series
Joseph K. Blitzstein, *Harvard University, USA*
Julian J. Faraway, *University of Bath, UK*
Martin Tanner, *Northwestern University, USA*
Jim Zidek, *University of British Columbia, Canada*

Recently Published Titles

Statistical Analysis of Financial Data
With Examples in R
James Gentle

Statistical Rethinking
A Bayesian Course with Examples in R and STAN, Second Edition
Richard McElreath

Statistical Machine Learning
A Model-Based Approach
Richard Golden

Randomization, Bootstrap and Monte Carlo Methods in Biology
Fourth Edition
Bryan F. J. Manly, Jorje A. Navarro Alberto

Principles of Uncertainty, Second Edition
Joseph B. Kadane

Beyond Multiple Linear Regression
Applied Generalized Linear Models and Multilevel Models in R
Paul Roback, Julie Legler

Bayesian Thinking in Biostatistics
Gary L. Rosner, Purushottam W. Laud, and Wesley O. Johnson

Linear Models with Python
Julian J. Faraway

Modern Data Science with R, Second Edition
Benjamin S. Baumer, Daniel T. Kaplan, and Nicholas J. Horton

Probability and Statistical Inference
From Basic Principles to Advanced Models
Miltiadis Mavrakakis and Jeremy Penzer

Bayesian Networks
With Examples in R, Second Edition
Marco Scutari and Jean-Baptiste Denis

Times Series
Modeling, Computation, and Inference, Second Edition
Raquel Prado, Marco A. R. Ferreira, and Mike West

For more information about this series, please visit: https://www.crcpress.com/Chapman--Hall/CRC-Texts-in-Statistical-Science/book-series/CHTEXSTASCI

Time Series

Modeling, Computation, and Inference

Second Edition

Raquel Prado

Marco A. R. Ferreira

Mike West

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

A CHAPMAN & HALL BOOK

First edition published 2021
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

Library of Congress Cataloging-in-Publication Data

ISBN: 978-1-498-74702-8 (hbk)
ISBN: 978-1-032-04004-2 (pbk)
ISBN: 978-1-351-25942-2 (ebk)

Typeset in CMR10
by KnowledgeWorks Global Ltd.

Contents

Preface

This book aims to integrate mainstream modeling approaches in time series with a range of significant recent developments in methodology and applications of time series analysis. We present overviews of several classes of models and related methodology for inference, statistical computation for model fitting and assessment, and forecasting. The book focuses mainly on time domain approaches while covering core topics and theory in the frequency domain, and connections between the two are often explored. Statistical analysis and inference involves likelihood and Bayesian methodologies, with a strong emphasis on using modern, simulation-based approaches for statistical parameter estimation, model fitting, and prediction; ranges of models and analyses are developed using Bayesian approaches and tools including Markov chain Monte Carlo and sequential Monte Carlo methods that define nowadays standard methodology.

Time series model theory and methods are illustrated with examples and case studies involving problems and data arising from a variety of applied fields, including signal processing, biomedical studies, finance, econometrics, and the environmental sciences. The book has three major aims: (1) to serve as a graduate textbook on Bayesian time series modeling and analysis; (2) to provide a broad range of references on state-of-the-art approaches to univariate and multivariate time series analysis, serving as an informed guide to the recent literature and a handbook for researchers and practitioners in applied areas that require sophisticated tools for analyzing challenging time series problems; and (3) to contact ranges of traditional as well as new and emerging topics that lie at research frontiers. Most of the material presented in Chapters 1 to 5, as well as selected topics from Chapters 6 to 11, are suitable as the core material for a one-term/semester or a one-quarter graduate course in time series analysis. Alternatively, a course might be structured to cover material on models and methods for univariate time series analysis based on Chapters 1 to 7 at greater depth in

one course, with material and supplements related to the multivariate time series models and methods of Chapters 8 to 11 as a second course. Then, most chapters also contact more advanced topics and link to research areas with open questions.

Contents

The book presents a selective coverage of core and more advanced and recent topics in the very broad field of time series analysis. As one of the oldest and richest areas of statistical science, and a field that contacts applied interests across a huge spectrum of science, social science, and engineering applications, "time series" simply cannot be comprehensively covered in any single text. Our aim, to the contrary, is to present, summarize, and overview core models and methods, complementing the pedagogical development with a selective range of recent research developments and applications that exemplify the growth of time series analysis into new areas based on these core foundations. The flavor of examples and case studies reflects our own interests and experiences in time series research and applications in collaborations with researchers from other fields, and we aim to convey some of the interest in, and utility of, the modeling approaches through these examples. Readers and students with backgrounds in statistical inference and some exposure to applied statistics and computation should find the book accessible.

Chapter 1 offers an introduction and a brief review of Bayesian inference, including Markov chain Monte Carlo (MCMC) methods. Chapter 2 presents autoregressive moving average models (ARMA) from a Bayesian perspective and illustrates these models with several examples. Chapter 3 discusses some theory and methods of frequency domain approaches, including harmonic regression models and their relationships with the periodogram and Bayesian spectral analysis. Some multivariate extensions are explored later in Chapters 8 and 9 in contexts of analyzing multiple and multivariate time series. Chapter 4 reviews dynamic models and methods for inference and forecasting for this broad and flexible class of models. More specifically, this chapter includes a review of the dynamic linear models (DLMs) of West and Harrison (1997), discusses extensions to nonlinear and non-Gaussian dynamic models, and reviews key developments of MCMC for filtering, parameter learning, and smoothing. Chapter 5 concerns issues of model specification and posterior inference in a particular class of DLMs: the broadly useful and widely applied class of time-varying autoregressive models. Theory and methods related to time series decompositions into interpretable latent processes, and examples in which real data sets are analyzed, are included. Chapter 6 covers recent developments of sequential Monte Carlo methods for general state-space models. Chapter 7 reviews a

selection of topics involving statistical mixture models in time series analysis, focusing on multiprocess models and univariate stochastic volatility models. Chapter 8 illustrates the analysis of multiple time series with common underlying structure and motivates some of the multivariate models that are developed later in Chapters 9 and 10. Chapter 9 discusses multivariate ARMA models, focusing on vector autoregressive (VAR) models, time series decompositions within this class of models, and mixtures of VAR models. Chapter 10 discusses a range of multivariate dynamic linear models, models and methods for time-varying, stochastic covariance matrices related to stochastic volatility, and contacts research frontiers in discussion of multivariate dynamic graphical models and other recent developments. The latter include contact with models and perspectives on problems of modeling and forecasting for increasingly large, complex, and hierarchically structured time series in commercial and other areas. Chapter 11 details developments of dynamic modeling with latent factor structures, a central area of time series methodology that has been heavily driven by advances in Bayesian methodology for dynamic models.

A collection of problems is included at the end of each chapter. Some of the chapters also include appendices that provide relevant supplements on statistical distribution theory and other mathematical aspects.

Acknowledgments

We recognize a number of colleagues for their impact on our thinking and eventual contributions to the broad field of time series modeling and forecasting, and directly or indirectly on the evolution of this text. Gabriel Huerta and Giovanni Petris provided material inputs that led to revisions of the core text material, and suggested some of the problems listed at the end of Chapters 1, 2, and 3. We thank Carlos Carvalho, Hedibert Lopes, Abel Rodríguez, and several other anonymous reviewers, as well as many colleagues at the University of California Santa Cruz (UCSC), Virginia Tech, and Duke University, and students from courses at UCSC, Virginia Tech, and Duke over many years, for their continued input as well as just day-to-day interactions that have had impact on the evolution of the core material presented.

Several of the data sets analyzed in the book come from collaborations with researchers in other fields, and such collaborations have been (and, we hope and expect) will continue to be critical to developments in modeling and methodology. Among many others, we are most appreciative of past contributions of collaborators including Dr. Andrew D. Krystal, Dr. Jose M. Quintana, and Dr. Leonard Trejo. We acknowledge the support and facilities at the Department of Statistics and the Baskin School of Engineering at

UCSC, the Department of Statistics at Virginia Tech, and the Department of Statistical Science at Duke University. We would also like to acknowledge the support of the Statistical and Applied Mathematical Science Institute (SAMSI) in North Carolina. In particular, some of the sections in Chapter 6 were written while Raquel Prado was visiting SAMSI as a participant of the 2008–2009 program on sequential Monte Carlo methods. We also acknowledge grants from the National Science Foundation, the National Institutes of Health, and a number of nongovernmental organizations and companies that have, over many years, provided support for our research that has contributed, directly and indirectly, to the development of models and methods presented in this book.

Raquel Prado, Marco A. R. Ferreira, and Mike West
December 2020

Authors

Raquel Prado is professor in the Department of Statistics at the Baskin School of Engineering at the University of California Santa Cruz (UCSC), USA. Her main research areas are time series analysis and Bayesian modeling, with a focus on analysis of large-dimensional nonstationary time series data and applications to biomedical signal processing and brain imaging.

Dr. Prado leads NSF- and NIH-funded projects, including multi-institutional and multi-disciplinary collaborative projects. She has supervised over 20 graduate students at UCSC and other academic institutions. Her former students work in academia, high tech companies, national laboratories, and local government agencies.

Dr. Prado is past president of the International Society for Bayesian Analysis (ISBA). She is an ISBA fellow and a fellow of the American Statistical Association (ASA). She has served on several committees at ASA and ISBA and is currently a member of the Committee on Applied and Theoretical Statistics (CATS) of the National Academies of Sciences, Engineering and Medicine.

Marco A. R. Ferreira is an associate professor in the Department of Statistics at Virginia Tech, where he served from 2016-2020 as the Director of Graduate Programs. Dr. Ferreira has served the statistics profession in editorial boards of multiple scientific journals including the journal *Bayesian Analysis*, in several committees of ISBA and ASA, as well as in scientific committees of numerous domestic and international conferences.

Dr. Ferreira's current research areas include dynamic models for time series and spatiotemporal data, multiscale models, objective Bayesian methods, stochastic search algorithms, and statistical computation. Major areas of application include bioinformatics, finance, and environmental science. His research is, and has been, funded by grants from the National Science Foundation. Marco has advised over 10 PhD students and postdocs and

has published over 50 scientific papers. His former students and postdocs work in academic, industrial, and government positions.

Mike West holds a Duke University Distinguished Chair as the Arts & Sciences Professor of Statistics & Decision Sciences in the Department of Statistical Science, where he led the development of statistics from 1990-2002. A past president of the International Society for Bayesian Analysis (ISBA), Mike has served the international statistics profession in founding roles for ISBA and in other professional organizations and institutions. Dr. West's research and teaching activities are in Bayesian analysis in ranges of interlinked areas: theory and methods of dynamic models in time series analysis, multivariate analysis, latent structure, high-dimensional inference and computation, quantitative and computational decision analysis, stochastic computational methods, and statistical computing, among other topics. Interdisciplinary R&D has ranged across applications in signal processing, finance, econometrics, climatology, systems biology, genomics and neuroscience, among other areas. His main current interests are in macroeconomic forecasting and policy decisions ,financial econometric forecasting and decisions, dynamic network studies in IT/commerce, and large-scale forecasting and decision problems in business and industry.

Dr. West has received a number of international awards for research and professional service, and multiple distinguished speaking awards. He has been, and continues to be, a statistical consultant for various companies, banks, government agencies, and academic centers, co-founder of a biotech company, and past, current advisor, or board member for several financial and IT companies. Dr. West teaches in academia and through short courses, works with and advises many undergraduates and master's students, and has mentored over 60 primary PhD students and postdoctoral associates, most of whom are now in academic, industrial, or government positions involving advanced statistical research.

Chapter 1

Notation, definitions, and basic inference

This chapter discusses key goals of time series analysis with motivating examples from different applied areas. Notation and key concepts related to time series processes are introduced, including the characterization of stationary processes. This is followed by a brief review on likelihood and Bayesian modeling and inference tools, which includes a primer on simulation-based methods for posterior inference within the Bayesian framework. The modeling and inference tools are illustrated for the class of first-order autoregressive processes.

1.1 Problem Areas and Objectives

The expression *time series data*, or *time series*, usually refers to a set of observations collected sequentially in time. These observations could have been collected at equally spaced time points. In this case we use the notation y_t with $(t = \ldots, -1, 0, 1, 2, \ldots)$; i.e., the set of observations is indexed by t, the time at which each observation was taken. If the observations were not taken at equally spaced points, then we use the notation y_{t_i}, with $i = 1, 2, \ldots$.

A *time series process* is a stochastic process or a collection of random variables y_t indexed in time. Note that y_t will be used throughout the book to denote a random variable or an actual realization of the time series process at time t. We use the notation $\{y_t, t \in \mathcal{T}\}$, or simply $\{y_t\}$, to refer to the time series process. If $\mathcal{T}$ is of the form $\{t_i, i \in \mathbb{N}\}$, with $\mathbb{N}$ the natural numbers, then the process is a discrete-time random process, and if $\mathcal{T}$ is an interval in the real line, or a collection of intervals in the real line, then the process is a continuous-time random process. In this framework, a time

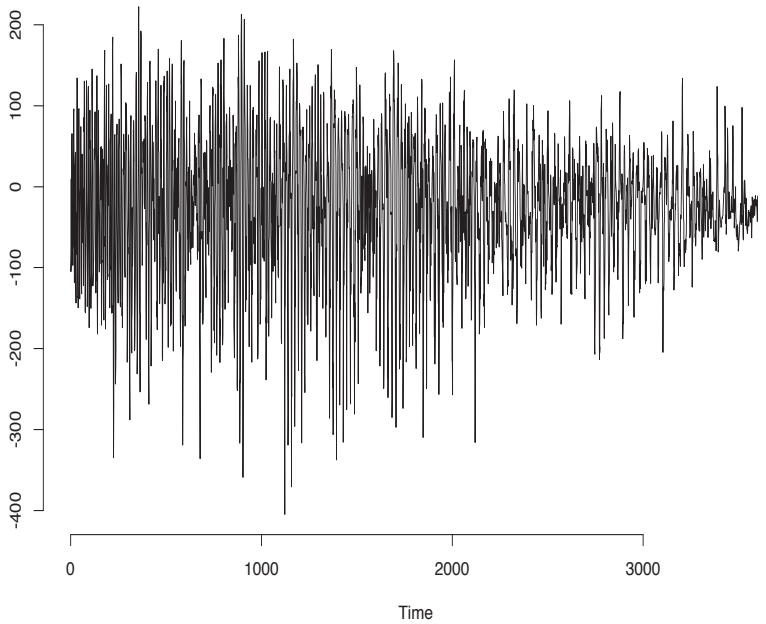


Figure 1.1 *EEG series (units in millivolts). The EEG was recorded at channel F_3 from a subject who received ECT.*

series data set $y_t, (t = 1, \ldots, T)$, also denoted by $y_{1:T}$, is just a collection of T equally spaced realizations of some time series process.

In many statistical models the assumption that the observations are realizations of independent random variables is key. In contrast, time series analysis is concerned with describing the dependence among the elements of a sequence of random variables.

At each time t, y_t can be a scalar quantity, such as the total amount of rainfall collected at a certain location in a given day t, or it can be a k-dimensional vector containing k scalar quantities that were recorded simultaneously. For instance, if the total amount of rainfall and the average temperature at a given location are measured in day t, we have $k = 2$ scalar quantities and a two-dimensional vector of observations $\mathbf{y}_t = (y_{1,t}, y_{2,t})'$. In general, for k scalar quantities recorded at time t, we have a realization $\mathbf{y}_t$ of a vector process $\{\mathbf{y}_t, t \in \mathcal{T}\}$, with $\mathbf{y}_t = (y_{1,t}, \ldots, y_{k,t})'$.

Figure 1.1 displays a portion of an electroencephalogram (EEG) recorded on a patient's scalp under certain electroconvulsive therapy (ECT) conditions. ECT is a treatment for patients under major clinical depression

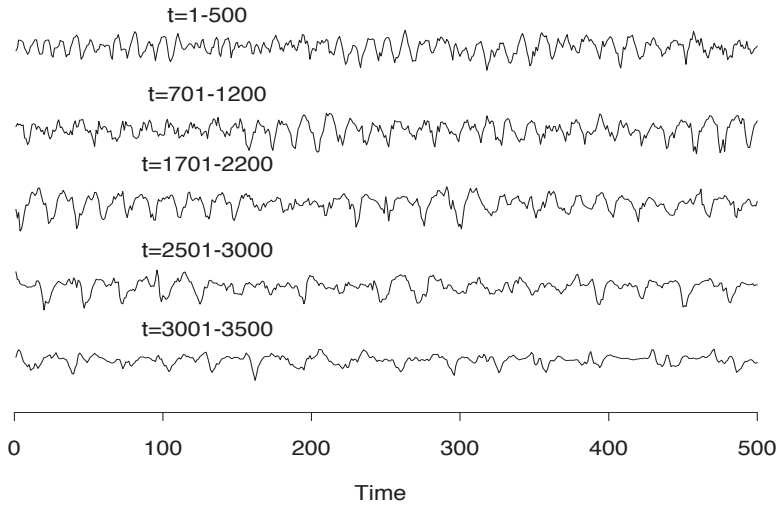


Figure 1.2 *Sections of the EEG trace displayed in Figure 1.1.*

(Krystal, Prado, and West 1999). When ECT is applied to a patient, seizure activity appears and can be recorded via electroencephalograms. The data correspond to one of 19 EEG series recorded simultaneously at different locations over the scalp. The main objective in analyzing these signals is the characterization of the clinical efficacy of ECT in terms of particular features that can be inferred from the recorded EEG traces. The data are fluctuations in electrical potential taken at a sampling rate of 256 Hz (i.e., 256 observations per second). For a more detailed description of these data and a full statistical analysis, see West, Prado, and Krystal (1999), Krystal, Prado, and West (1999), and Prado, West, and Krystal (2001).

From the time series analysis viewpoint, the objective here is modeling the data to provide useful insight about the underlying processes driving the multiple series during a seizure episode. Studying the differences and commonalities among the 19 EEG channels is also key. Univariate time series models for each individual EEG series could be explored and used to investigate relationships across the 19 channels (Chapters 2, 5, and 8). Multivariate time series analyses (Chapters 9 and 10)—in which the observed series, $\mathbf{y}_t$, is a 19-dimensional vector whose elements are the observed voltage levels measured at the 19 scalp locations at each time t—can also be considered. Uncovering the common latent structure that may underlie the 19 EEG time series over time can be achieved by decomposing these observed EEGs into simpler latent non observable components. Such latent

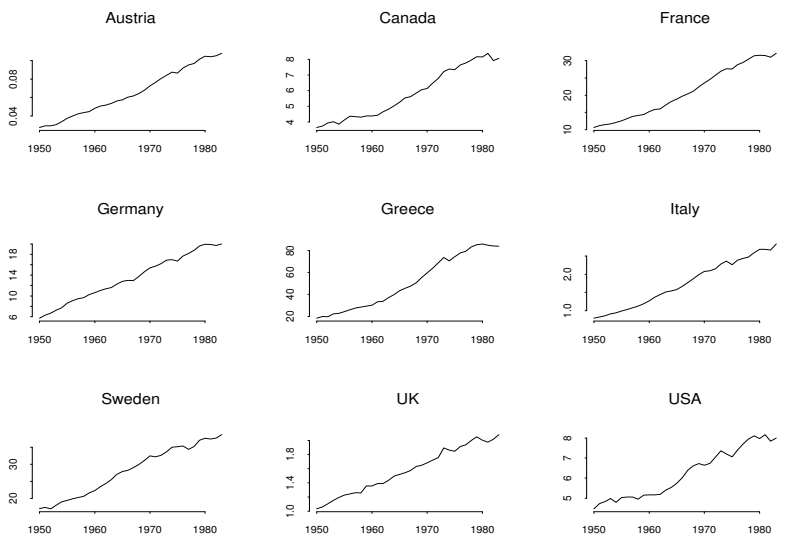


Figure 1.3 *International annual GDP time series.*

components can be obtained via time series decompositions derived from a specific state-space modeling framework (Chapters 5 and 8), or by explicitly modeling them as latent factors in a dynamic factor model (Chapter 11).

These EEG series display a quasiperiodic behavior that changes dynamically in time, as shown in Figure 1.2, where different portions of the EEG trace shown in Figure 1.1 are displayed. In particular, it is clear that the relatively high-frequency components that appear initially are slowly decreasing toward the end of the series. Any time series model used to describe these data should take into account their nonstationary and quasiperiodic structure. We discuss various modeling alternatives for analyzing these data in the subsequent chapters, including the class of time-varying autoregressions and some multichannel models.

Figure 1.3 shows the annual per capita GDP (gross domestic product) time series for Austria, Canada, France, Germany, Greece, Italy, Sweden, UK, and USA from 1950 to 1983. Goals of the analysis include forecasting turning points and comparing characteristics of the series across the national economies. Univariate and multivariate analyses of the GDP data can be considered.

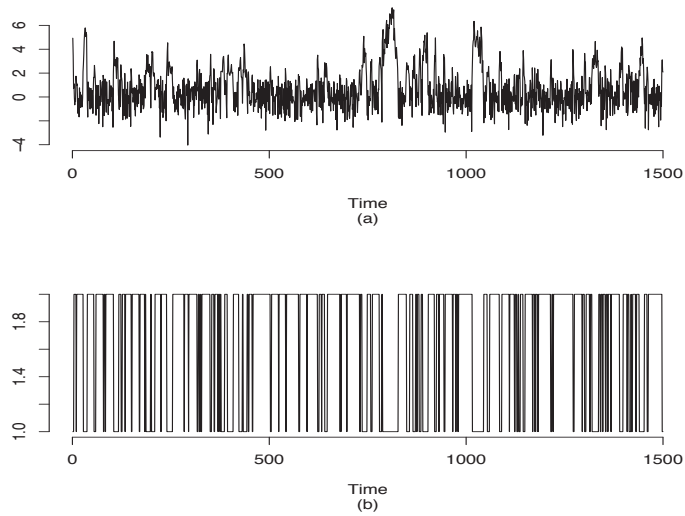


Figure 1.4 *(a): Simulated time series y_t; (b) Indicator variable δ_t with $\delta_t = 1$ if y_t was sampled from $\mathcal{M}_1$ and $\delta_t = 2$ if y_t was sampled from $\mathcal{M}_2$.*

One of the main differences between any time series analysis of the GDP series and any time series analysis of the EEG series, regardless of the type of models used in such analyses, lies in the objectives. As mentioned above, one of the goals in analyzing the GDP data is forecasting future outcomes of the series for the several countries given the observed values. In the EEG study previously described, there is no interest in forecasting future values of the series given the observed traces; instead, the objective is finding an appropriate model that describes the structure of the series and its latent components.

Other objectives of time series analysis include monitoring a time series in order to detect possible "on-line" (real time) changes. This is important for control purposes in engineering, industrial, and medical applications. For instance, consider a time series generated from the process $\{y_t\}$ with

$$y_t = \begin{cases} 0.9y_{t-1} + \epsilon_t^{(1)}, & y_{t-1} > 1.5 \quad (\mathcal{M}_1) \\ -0.3y_{t-1} + \epsilon_t^{(2)}, & y_{t-1} \leq 1.5 \quad (\mathcal{M}_2), \end{cases} \tag{1.1}$$

where $\epsilon_t^{(1)} \sim N(0, v_1)$, $\epsilon_t^{(2)} \sim N(0, v_2)$, and $v_1 = v_2 = 1$. Figure 1.4 (a) shows a time series plot of 1,500 observations simulated according to (1.1). Figure 1.4 (b) displays the values of an indicator variable, δ_t, with $\delta_t = 1$ if y_t was generated from $\mathcal{M}_1$, and $\delta_t = 2$ if y_t was generated from $\mathcal{M}_2$. Model (1.1) is a *threshold autoregressive (TAR) model* with two regimes

that belongs to the broader class of mixture models (see Chapter 7). TAR models were initially developed by H. Tong (Tong 1983; Tong 1990). In particular, (1.1) can be written in the following, more general, form

$$y_t = \begin{cases} \phi^{(1)} y_{t-1} + \epsilon_t^{(1)}, & \theta + y_{t-d} > 0 \quad (\mathcal{M}_1) \\ \phi^{(2)} y_{t-1} + \epsilon_t^{(2)}, & \theta + y_{t-d} \leq 0 \quad (\mathcal{M}_2), \end{cases} \tag{1.2}$$

with $\epsilon_t^{(1)} \sim N(0, v_1)$ and $\epsilon_t^{(2)} \sim N(0, v_2)$. These are nonlinear models and the interest lies in making inferences on d, θ, and the parameters $\phi^{(1)}, \phi^{(2)}, v_1$, and v_2.

The TAR model (1.2) serves the purpose of illustrating, at least for a very simple case, a situation that arises in many engineering applications, particularly in the area of control theory. From a control theory viewpoint, we can think of (1.2) as a bimodal process in which two scenarios of operation are handled by two control modes ($\mathcal{M}_1$ and $\mathcal{M}_2$). In each mode the evolution is governed by a stochastic process. Autoregressions of order one, or AR(1) models (a formal definition of this type of process is given later in this chapter), were chosen in this example, but more sophisticated structures can be considered. The transitions between the modes occur when the series crosses a specific threshold and so, we can talk about an internally triggered mode switch. In an externally triggered mode switch, the moves are defined by external variables. In terms of the goals of time series analysis in this case we can consider two possible scenarios. In many control settings where the transitions between modes occur in response to a controller's actions, the current state is always known, and so, the learning process can be split into two: learning the stochastic models that control each mode conditional on the fact that we know in which mode we are—i.e., inferring $\phi^{(1)}, \phi^{(2)}, v_1$, and v_2—and learning the transition rule, that is, making inferences about d and θ assuming we know the values $\delta_{1:T}$. In other control settings for which the mode transitions do not occur in response to a controller's actions, it is necessary to simultaneously infer the parameters associated to the stochastic models that describe each mode and the transition rule. In this case we want to estimate $\phi^{(1)}, \phi^{(2)}, v_1, v_2, \theta$, and d conditioning only on the observed data $y_{1:T}$. Depending on the application, it may also be necessary to achieve parameter learning from the time series sequentially in time. Methods for sequential state and parameter learning in time series models are discussed throughout this book.

Clustering also arises as the primary goal in many applications. For example, a common scenario is one in which a collection of N time series generated from a relatively small number of processes, say K, with $K << N$, are available. It is not known a priori which time series are generated from which processes, and so the main objective of the analysis consists on grouping the time series into K clusters according to their spectral

characteristics. Some references in this area include Kakizawa, Shumway, and Taniguchi (1998), Huan, Ombao, and Stoffer (2004), Gao, Ombao, and Ho (2009), Pamminger and Frühwirth-Schnatter (2010), and Nieto-Barajas and Contreras-Cristán (2014).

Finally, we may use time series techniques to describe serial dependencies between parameters of a given model with additional structure. For example, we could have a linear regression model of the form $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$, for which ϵ_t does not exhibit the usual independent structure $\epsilon_t \sim N(0, v)$ for all t, but instead, the probability distribution of ϵ_t depends on $\epsilon_{t-1}, \ldots, \epsilon_{t-k}$ for some integer $k > 0$.

1.2 Stochastic Processes and Stationarity

Many time series models are based on the assumption of stationarity. Intuitively, a stationary time series process is a process whose behavior does not depend on when we start to observe it. In other words, different sections of the series will look roughly the same at intervals of the same length. Here we provide two widely used definitions of stationarity.

A time series process $\{y_t, t \in \mathcal{T}\}$ is *completely* or *strongly stationary* if, for any sequence of times $t_1, t_2, \ldots, t_n$, and any lag h with $h = 0, \pm 1, \pm 2, \ldots$, the probability distribution of the vector $(y_{t_1}, \ldots, y_{t_n})'$ is identical to the probability distribution of the vector $(y_{t_1+h}, \ldots, y_{t_n+h})'$.

In practice it is very difficult to verify that a process is strongly stationary and so, the notion of *weak* or *second-order stationarity* arises. A process is said to be weakly stationary, or second-order stationary if, for any sequence of times $t_1, \ldots, t_n$, and any integer lag h, all the first and second joint moments of $(y_{t_1}, \ldots, y_{t_n})'$ exist and are equal to the first and second joint moments of $(y_{t_1+h}, \ldots, y_{t_n+h})'$. If $\{y_t\}$ is second-order stationary, we have that

$$E(y_t) = \mu, \quad V(y_t) = v, \quad Cov(y_t, y_s) = \gamma(t - s), \tag{1.3}$$

where μ, v are constant, independent of t and $\gamma(t-s)$ is also independent of t and s, depending only on the length of the interval between time points. It is also possible to define stationarity up to order m in terms of the m joint moments (see for example Priestley 1994).

If the first two moments exist, complete stationarity implies second-order stationarity, but the converse is not necessarily true. If $\{y_t\}$ is a Gaussian process, i.e., if for any sequence of time points $t_1, \ldots, t_n$ the vector $(y_{t_1}, \ldots, y_{t_n})'$ follows a multivariate normal distribution, strong and weak stationarity are equivalent (see Shumway and Stoffer 2017 for a proof).

1.3 Autocorrelation and Cross-correlation

The first step in a statistical analysis often consists on performing a descriptive study of the data in order to summarize their main features. One of the most widely used descriptive techniques in time series data analysis is that of exploring the correlation patterns displayed by a series, or a couple of series, at different time points. This is done by plotting the sample autocorrelation and cross-correlation values, which are estimates of the autocorrelation and cross-correlation functions.

We begin by defining the concepts of autocovariance, autocorrelation, and cross-correlation functions. We then show how to estimate these functions from data. Let $\{y_t, t \in \mathcal{T}\}$ be a time series process. The autocovariance function of $\{y_t\}$ is defined as follows:

$$\gamma(s,t) = Cov\{y_t, y_s\} = E\{(y_t - \mu_t)(y_s - \mu_s)\}, \tag{1.4}$$

for all s, t, with $\mu_t = E(y_t)$. For stationary processes $\mu_t = \mu$ for all t and the covariance function depends on $|t - s|$ only. In this case we can write the autocovariance as a function of a particular time lag h, i.e.,

$$\gamma(h) = Cov\{y_t, y_{t-h}\}. \tag{1.5}$$

The autocorrelation function (ACF) is then given by

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(t,t)\gamma(s,s)}}. \tag{1.6}$$

For stationary processes, the ACF can be written in terms of a lag h:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}. \tag{1.7}$$

The ACF measures the linear dependence between a value of the time series process at time t and past or future values of such process. It inherits the properties of any correlation function–$\rho(h)$ always takes values in the interval $[-1, 1]$. In addition, $\rho(h) = \rho(-h)$ and, if y_t and y_{t-h} are independent, then $\rho(h) = 0$.

It is also possible to define the cross-covariance and cross-correlation functions of two univariate time series. If $\{y_t\}$ and $\{z_t\}$ are two time series processes, the cross-covariance is defined as

$$\gamma_{y,z}(s,t) = E\{(y_t - \mu_{y_t})(z_s - \mu_{z_s})\}, \tag{1.8}$$

for all s, t, and the cross-correlation is then given by

$$\rho_{y,z}(s,t) = \frac{\gamma_{y,z}(s,t)}{\sqrt{\gamma_{y,y}(t,t)\gamma_{z,z}(s,s)}}. \tag{1.9}$$

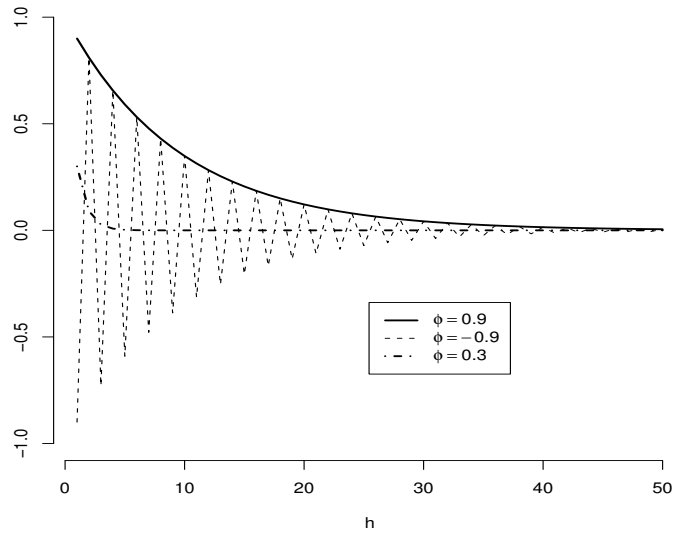


Figure 1.5 *Autocorrelation functions for AR*(1) *processes with parameters* 0.9, −0.9, *and* 0.3.

If both processes are stationary, we can write the cross-covariance and cross-correlation functions in terms of a lag value h. This is

$$\gamma_{y,z}(h) = E\{(y_t - \mu_y)(z_{t-h} - \mu_z)\} \tag{1.10}$$

and

$$\rho_{y,z}(h) = \frac{\gamma_{y,z}(h)}{\sqrt{\gamma_y(0)\gamma_z(0)}}. \tag{1.11}$$

Example 1.1 *White noise.* Consider a process such that $y_t \sim N(0, v)$ for all t, with $Cov(y_t, y_s) = 0$ if $t \neq s$. In this case $\gamma(0) = v$, $\gamma(h) = 0$ for all $h \neq 0$, and so, $\rho(0) = 1$ and $\rho(h) = 0$ for all $h \neq 0$.

Example 1.2 *First-order autoregression or AR*(1). In Chapter 2 we formally define and study the properties of general autoregressions of order p, or AR(p) processes. Here, we illustrate some properties of the simplest AR process, the AR(1). Consider a process such that $y_t = \phi y_{t-1} + \epsilon_t$ with $\epsilon_t \sim N(0, v)$ for all t. It is possible to show (see Problem 1 in this chapter) that, if $|\phi| < 1$, $\gamma(h) = \phi^{|h|}\gamma(0)$ for $h = 0, \pm 1, \pm 2, \ldots$, with $\gamma(0) = \frac{v}{(1-\phi^2)}$, and $\rho(h) = \phi^{|h|}$. Figure 1.5 displays the ACFs of AR(1) processes with parameters $\phi = 0.9, \phi = -0.9$ and $\phi = 0.3$, for lag values $h = 1 : 50$. For negative values of ϕ the ACF has an oscillatory behavior. In addition,

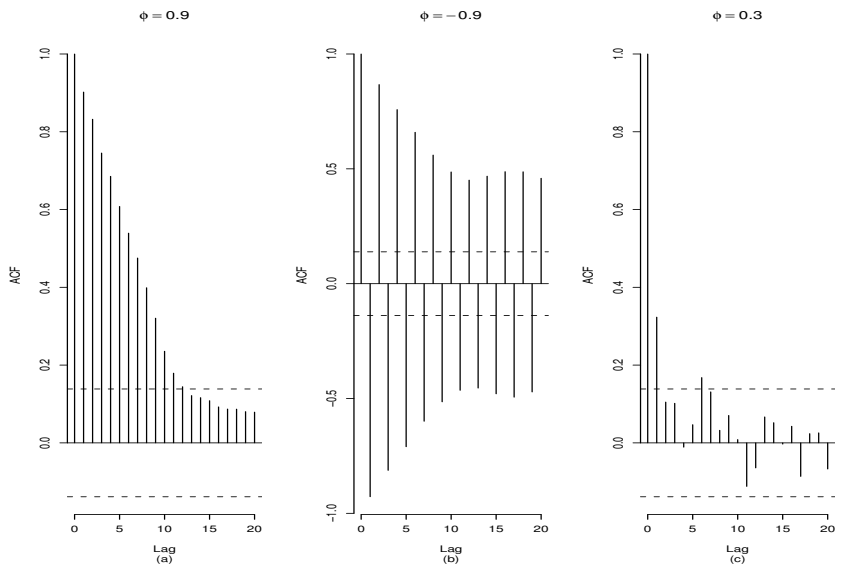


Figure 1.6 *Sample autocorrelations for AR processes with parameters 0.9, −0.9, and 0.3 (graphs (a), (b), and (c), respectively).*

the rate of decay of the ACF is a function of ϕ. The closer $|\phi|$ gets to the unity the lower the rate of decay is (e.g., compare the ACFs for $\phi = 0.9$ and $\phi = 0.3$). This is related to the characterization of stationary AR(1) processes as discussed in Chapter 2. An AR(1) process is stationary if and only if $|\phi| < 1$. This condition can also be written as a function of the characteristic root of the process. An AR(1) is stationary if and only if the root of the characteristic polynomial, $\Phi(u)$ with $\Phi(u) = 1 - \phi u$, lies outside the unit circle. This happens if and only if $|\phi| < 1$.

We now show how to estimate the autocovariance, autocorrelation, cross-covariance, and cross-correlation functions from data. Assume we have data $y_{1:T}$. The usual estimate of the autocovariance function is the sample autocovariance, which, for $h > 0$, is given by

$$\hat{\gamma}(h) = \frac{1}{T} \sum_{t=1}^{T-h} (y_{t+h} - \bar{y})(y_t - \bar{y}), \tag{1.12}$$

where $\bar{y} = \sum_{t=1}^{T} y_t / T$ is the sample mean. We can then obtain estimates of the autocorrelation function as $\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}$, for $h = 0, 1, \ldots$. Similarly,

estimates of the cross-covariance and cross-correlation functions can be obtained. The sample cross-covariance is given by

$$\hat{\gamma}_{y,z}(h) = \frac{1}{T}\sum_{t=1}^{T-h}(y_{t+h} - \bar{y})(z_t - \bar{z}), \tag{1.13}$$

and so, the sample cross-correlation is given by

$$\hat{\rho}_{y,z}(h) = \hat{\gamma}_{y,z}(h) \Big/ \sqrt{\hat{\gamma}_y(0)\hat{\gamma}_z(0)}\ .$$

Example 1.3 *Sample ACFs of AR*(1) *processes.* Figure 1.6 displays the sample autocorrelation functions of simulated AR(1) processes with parameters $\phi = 0.9$, $\phi = -0.9$, and $\phi = 0.3$. The sample ACFs were computed based on a sample of $T = 200$ data points. For $\phi = 0.9$ and $\phi = 0.3$, the corresponding sample ACFs decay with the lag. The oscillatory form of the ACF for the process with $\phi = -0.9$ is captured by the corresponding sample ACF.

The estimates given in (1.12) and (1.13) are not unbiased estimates of the autocovariance and cross-covariance functions. Results related to the distributions of the sample autocorrelation and the sample cross-correlation functions appear, for example, in Shumway and Stoffer (2017).

1.4 Smoothing and Differencing

As mentioned before, many time series models are built under the stationarity assumption. Several descriptive techniques have been developed to study the stationary properties of a time series so that an appropriate model can then be applied to the data. For instance, looking at the sample autocorrelation function may be helpful in identifying some features of the data. However, in many practical scenarios the data are realizations from one or several nonstationary processes. In this case, methods that aim to eliminate the nonstationary components are often used. The idea is to separate the nonstationary components from the stationary ones so that the latter can be carefully studied via traditional time series models such as, for example, the ARMA (autoregressive moving average) models that will be discussed in subsequent chapters.

We review some commonly used methods for extracting nonstationary components from a time series. We do not attempt to provide a comprehensive list of such methods. Instead, we just list and summarize a few of them. We view these techniques as purely descriptive.

Many descriptive time series methods are based on the notion of *smoothing* the data, that is, decomposing the series as a sum of two components: a so called "smooth" component, plus another component that includes all the features of the data that are left unexplained by the smooth component. This is similar to the "signal plus noise" concept used in signal processing. The main difficulty with this approach lies in deciding which features of the data are part of the signal or the smooth component, and which ones are part of the noise.

One way of smoothing a time series is by moving averages (see Kendall, Stuart, and Ord 1983; Kendall and Ord 1990; Chatfield 1996; and Diggle 1990 for detailed discussions and examples). If we have data $y_{1:T}$, we can smooth them by applying an operation of the form

$$z_t = \sum_{j=-q}^{p} a_j y_{t+j}, \quad t = (q+1) : (T-p), \tag{1.14}$$

with p and q nonnegative integers, and where the a_js are weights such that $\sum_{j=-q}^{p} a_j = 1$. It is generally assumed that $p = q$, $a_j \geq 0$ for all j and $a_j = a_{-j}$. The order of the moving average in this case is $2p + 1$. The first question that arises when applying a moving average to a series is how to choose p and the weights. The simplest alternative is choosing a low value of p and equal weights. The higher the value of p, the smoother z_t is going to be. Other alternatives include successively applying a simple moving average with equal weights, or choosing the weights in such a way that a particular feature of the data is highlighted. For example, if a given time series recorded monthly displays a trend plus a yearly cycle, choosing a moving average with $p = 6$, $a_6 = a_{-6} = 1/24$, and $a_j = 1/12$ for $j = 0, \pm 1, \ldots, \pm 5$ would diminish the impact of the periodic component, emphasizing the trend (see Diggle 1990 for an example).

Figure 1.7 (a) shows monthly values of a Southern Oscillation Index (SOI) time series during 1950–1995. This series consists of 540 observations of the SOI computed as the difference of the departure from the long term monthly mean sea level pressures at Tahiti in the South Pacific and Darwin in Northern Australia. The index is one measure of the so called "El Niño-Southern Oscillation"—an event of critical importance and interest in climatological studies in recent decades. The fact that most of the observations in the last part of the series take negative values is related to a recent warming in the tropical Pacific. Figures 1.7 (b) and (c) show two smoothed series obtained via moving averages of orders 3 and 9, respectively, with equal weights. As explained before, we can see that the higher the order of the moving average the smoother the resulting series is.

Other ways to smooth a time series include fitting a linear regression to remove a trend or, more generally, fitting a polynomial regression; fitting a

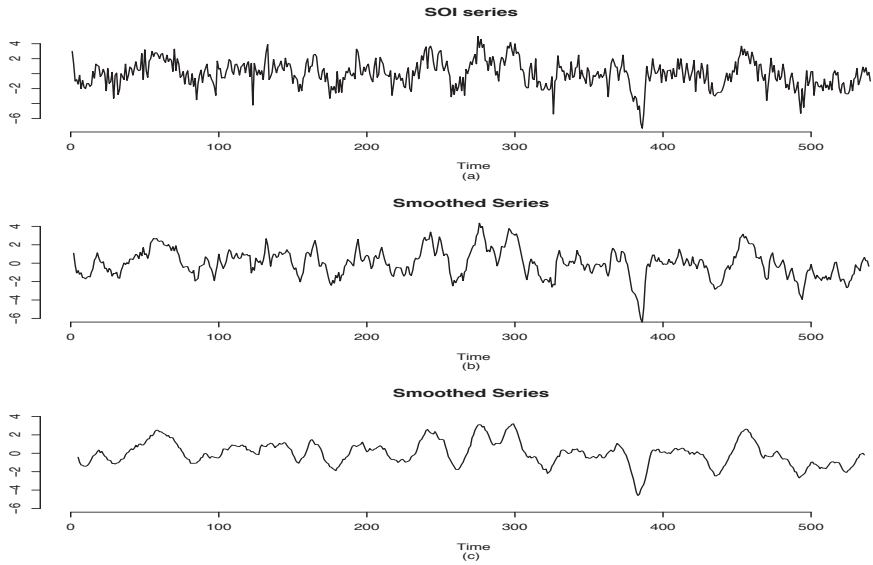


Figure 1.7 *(a): Southern oscillation index (SOI) time series; (b): Smoothed series obtained using a moving average of order 3 with equal weights; (c): Smoothed series obtained using a moving average of order 9 with equal weights.*

harmonic regression to remove periodic components; and performing kernel or spline smoothing.

Smoothing by polynomial regression consists on fitting a polynomial to the series. In other words, we want to estimate the parameters of the model

$$y_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p + \epsilon_t,$$

where ϵ_t is usually assumed as a sequence of zero mean, independent Gaussian random variables. Similarly, fitting harmonic regressions provides a way to remove cycles from a time series. So, if we want to remove periodic components with frequencies $w_1, \ldots, w_p$, we need to estimate $a_1, b_1, \ldots, a_p, b_p$ in the model

$$\begin{aligned} y_t &= a_1 \cos(2\pi w_1 t) + b_1 \sin(2\pi w_1 t) + \cdots \\ &\quad + a_p \cos(2\pi w_p t) + b_p \sin(2\pi w_p t) + \epsilon_t. \end{aligned}$$

In both cases the smoothed series would then be obtained as $\hat{y}_t$, with $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 t + \cdots + \hat{\beta}_p t^p$, and $\hat{y}_t = \hat{a}_1 \cos(2\pi w_1 t) + \hat{b}_1 \sin(2\pi w_1 t) + \cdots + \hat{a}_p \cos(2\pi w_p t) + \hat{b}_p \sin(2\pi w_p t)$, respectively, where $\hat{\beta}_i$, $\hat{a}_i$, and $\hat{b}_i$ are point

estimates of the parameters. Usually $\hat{\beta}_i$ and $\hat{a}_i, \hat{b}_i$ are obtained by least squares estimation.

In kernel smoothing a smoothed version, z_t, of the original series y_t is obtained as follows:

$$z_t = \sum_{i=1}^{T} w_t(i) y_t, \quad w_i(t) = K\left(\frac{t-i}{b}\right) \Big/ \sum_{j=1}^{T} K\left(\frac{t-j}{b}\right),$$

where $K(\cdot)$ is a kernel function, such as a normal kernel. The parameter b is a bandwidth. The larger the value of b, the smoother z_t is.

Cubic and smoothing splines, as well as the *lowess* smoother (Cleveland 1979; Cleveland and Devlin 1988; lowess stands for locally weighted scatterplot smoothing) are also commonly used smoothing techniques. See Shumway and Stoffer (2017) for details and illustrations on these smoothing techniques.

Another way of smoothing a time series is by taking its differences. Differencing provides a way to remove trends. The first difference of a series y_t is defined in terms of an operator D that produces the transformation $Dy_t = y_t - y_{t-1}$. Higher-order differences are defined by successively applying the operator D. Differences can also be defined in terms of the backshift operator B, with $By_t = y_{t-1}$, and so $Dy_t = (1 - B)y_t$. Higher-order differences can be written as $D^d y_t = (1 - B)^d y_t$.

In connection with the methods presented here, it is worth mentioning that wavelet decompositions have been widely used in recent years for smoothing time series. Vidakovic (1999) and Percival and Walden (2006) present statistical approaches to modeling by wavelets. Wavelets are basis functions that are used to represent other functions. They are analogous to the sines and cosines in the Fourier transformation. One of the advantages of using wavelets bases, as opposed to Fourier representations, is that they are localized in frequency and time, and so, they are suitable for dealing with nonstationary signals that display jumps and other abrupt changes.

1.5 A Primer on Likelihood and Bayesian Inference

Assume that we have collected T observations, $y_{1:T}$, of a scalar time series process $\{y_t\}$. Suppose that for each y_t we have a probability distribution that can be written as a function of some parameter, or collection of parameters, namely $\boldsymbol{\theta}$, in such a way that the dependence of y_t on $\boldsymbol{\theta}$ is described in terms of a probability density function $p(y_t|\boldsymbol{\theta})$. If we think of $p(y_t|\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, rather than a function of y_t, we refer to it as the likelihood function. Using Bayes' theorem it is possible to obtain the posterior density function of $\boldsymbol{\theta}$ given y_t, $p(\boldsymbol{\theta}|y_t)$, as the product of the likelihood and the

prior density $p(\boldsymbol{\theta})$, i.e.,

$$p(\boldsymbol{\theta}|y_t) = \frac{p(\boldsymbol{\theta})p(y_t|\boldsymbol{\theta})}{p(y_t)}, \tag{1.15}$$

with $p(y_t) = \int p(\boldsymbol{\theta})p(y_t|\boldsymbol{\theta})d\boldsymbol{\theta}$. $p(y_t)$ defines the so-called predictive density function. The prior distribution offers a way to incorporate our prior beliefs about $\boldsymbol{\theta}$ and Bayes' theorem allows us to update such beliefs after observing the data.

Bayes' theorem can also be used in a sequential way as follows: Before collecting any data, prior beliefs about $\boldsymbol{\theta}$ are expressed in a probabilistic form via $p(\boldsymbol{\theta})$. Assume that we then collect our first observation at time $t = 1$, y_1, and we obtain $p(\boldsymbol{\theta}|y_1)$ using Bayes' theorem. Once y_2 is observed we can obtain $p(\boldsymbol{\theta}|y_{1:2})$ via Bayes' theorem as $p(\boldsymbol{\theta}|y_{1:2}) \propto p(\boldsymbol{\theta})p(y_{1:2}|\boldsymbol{\theta})$. Now, if y_1 and y_2 are conditionally independent on $\boldsymbol{\theta}$, we can write $p(\boldsymbol{\theta}|y_{1:2}) \propto p(\boldsymbol{\theta}|y_1)p(y_2|\boldsymbol{\theta})$, i.e., the posterior of $\boldsymbol{\theta}$ given y_1 becomes a prior distribution before observing y_2. Similarly, $p(\boldsymbol{\theta}|y_{1:T})$ can be obtained in a sequential way, if all the observations are independent. However, in time series analysis the observations are not independent. For example, a common assumption is that each observation at time t depends only on $\boldsymbol{\theta}$ and the observation taken at time $t-1$. In this case we have

$$p(\boldsymbol{\theta}|y_{1:T}) \propto p(\boldsymbol{\theta})p(y_1|\boldsymbol{\theta}) \prod_{t=2}^{T} p(y_t|y_{t-1}, \boldsymbol{\theta}). \tag{1.16}$$

General models in which y_t depends on an arbitrary number of past observations will be studied in subsequent chapters. We now consider an example in which the posterior distribution has the form (1.16).

Example 1.4 *The AR(1) model.* We consider again the AR(1) process. The model parameters in this case are given by $\boldsymbol{\theta} = (\phi, v)'$. Now, for each time $t > 1$, the conditional likelihood is $p(y_t|y_{t-1}, \boldsymbol{\theta}) = N(y_t|\phi y_{t-1}, v)$. In addition, it can be shown that $y_1 \sim N(0, v/(1-\phi^2))$ if the process is stationary (see Problem 1 in Chapter 2) and so, the likelihood is given by

$$p(y_{1:T}|\boldsymbol{\theta}) = \frac{(1-\phi^2)^{1/2}}{(2\pi v)^{T/2}} \exp\left\{-\frac{Q^*(\phi)}{2v}\right\}, \tag{1.17}$$

with

$$Q^*(\phi) = y_1^2(1-\phi^2) + \sum_{t=2}^{T}(y_t - \phi y_{t-1})^2. \tag{1.18}$$

The posterior density is obtained via Bayes' rule and so

$$p(\boldsymbol{\theta}|y_{1:T}) \quad \propto \quad p(\boldsymbol{\theta})\frac{(1-\phi^2)^{1/2}}{(2\pi v)^{T/2}} \exp\left\{\frac{-Q^*(\phi)}{2v}\right\}.$$

We can also use the conditional likelihood $p(y_{2:T}|\boldsymbol{\theta}, y_1)$ as an approximation to the likelihood (see Box, Jenkins, Reinsel, and Ljung 2015 A7.4 for a justification), which leads to the following posterior density,

$$p(\boldsymbol{\theta}|y_{1:T}) \quad \propto \quad p(\boldsymbol{\theta})v^{-(T-1)/2}\exp\left\{\frac{-Q(\phi)}{2v}\right\}, \tag{1.19}$$

with $Q(\phi) = \sum_{t=2}^{T}(y_t - \phi y_{t-1})^2$. Several choices of $p(\boldsymbol{\theta})$ can be considered and will be discussed later. In particular, it is common to assume a prior structure such that $p(\boldsymbol{\theta}) = p(v)p(\phi|v)$, or $p(\boldsymbol{\theta}) = p(v)p(\phi)$.

Another important class of time series models is that in which parameters are indexed in time. In this case each observation is related to a parameter, or a set of parameters, say $\boldsymbol{\theta}_t$, that evolve over time. The so-called class of Dynamic Linear Models (DLMs) considered in Chapter 4 deals with models of this type. In such framework it is necessary to define a process that describes the evolution of $\boldsymbol{\theta}_t$ over time. As an example, consider the time-varying AR model of order one, or TVAR(1), given by

$$\begin{aligned} y_t &= \phi_t y_{t-1} + \epsilon_t, \\ \phi_t &= \phi_{t-1} + \nu_t, \end{aligned}$$

where ϵ_t and ν_t are independent in time and mutually independent, with $\epsilon_t \sim N(0, v)$ and $\nu_t \sim N(0, w)$. Some distributions of interest are the posterior distributions at time t, $p(\phi_t|y_{1:t})$ and $p(v|y_{1:t})$, the backward filtering or smoothing distributions $p(\phi_t|y_{1:T})$, and the h-steps ahead forecast distribution $p(y_{t+h}|y_{1:t})$. Details on how to find these distributions for rather general DLMs are given in Chapter 4.

1.5.1 ML, MAP, and LS Estimation

It is possible to obtain point estimates of the model parameters by maximizing the likelihood function or the full posterior distribution. A variety of methods and algorithms have been developed to achieve this goal. We briefly discuss some of these methods. In addition, we illustrate how these methods work in the simple AR(1) case.

A point estimate of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ can be obtained by maximizing the likelihood function $p(y_{1:T}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. In this case we use the notation $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\text{ML}}$. Similarly, if instead of maximizing the likelihood function we maximize the posterior distribution $p(\boldsymbol{\theta}|y_{1:T})$, we obtain the maximum a posteriori estimate for $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\text{MAP}}$.

Often, the likelihood function and the posterior distribution are complicated nonlinear functions of $\boldsymbol{\theta}$ and so it is necessary to use methods such as

the Newton–Raphson algorithm or the scoring method to obtain the maximum likelihood estimator (MLE) or the maximum a posteriori (MAP) estimator. In general, the Newton–Raphson algorithm can be summarized as follows. Let $g(\boldsymbol{\theta})$ be the function of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ that we want to maximize, and $\hat{\boldsymbol{\theta}}$ be the maximum. At iteration m of the Newton–Raphson algorithm we obtain $\boldsymbol{\theta}^{(m)}$, an approximation to $\hat{\boldsymbol{\theta}}$, as follows:

$$\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)} - \left[g''(\boldsymbol{\theta}^{(m-1)})\right]^{-1} \times \left[g'(\boldsymbol{\theta}^{(m-1)})\right], \tag{1.20}$$

where $g'(\boldsymbol{\theta})$ and $g''(\boldsymbol{\theta})$ denote the first- and second-order partial derivatives of the function g, i.e., $g'(\boldsymbol{\theta})$ is a k-dimensional vector given by $g'(\boldsymbol{\theta}) = \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_k}\right)'$, and $g''(\boldsymbol{\theta})$ is a $k \times k$ matrix of second-order partial derivatives whose ij-th element is given by $\left[\frac{\partial g^2(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right]$, for $i, j = 1 : k$. Under certain conditions this algorithm produces a sequence $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$, that will converge to $\hat{\boldsymbol{\theta}}$. In particular, it is important to begin with a good starting value $\boldsymbol{\theta}^{(0)}$, since the algorithm does not necessarily converge for values in regions where $-g''(\cdot)$ is not positive definite. An alternative method is the scoring method, which involves replacing $g''(\boldsymbol{\theta})$ in (1.20) by the matrix of expected values $E(g''(\boldsymbol{\theta}))$.

In many practical scenarios, especially when dealing with models that have very many parameters, it is not useful to summarize the inferences in terms of the joint posterior mode. Instead, summaries are made in terms of marginal posterior modes, that is, the posterior modes for subsets of model parameters. Let us say that we can partition our model parameters in two sets, $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, so that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, and assume we are interested in $p(\boldsymbol{\theta}_2|y_{1:T})$. The EM (Expectation-Maximization) algorithm proposed in Dempster, Laird, and Rubin (1977) is useful when dealing with models for which $p(\boldsymbol{\theta}_2|y_{1:T})$ is hard to maximize directly, but it is relatively easy to work with $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, y_{1:T})$ and $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, y_{1:T})$. The EM algorithm can be described as follows:

1. Start with some initial value $\boldsymbol{\theta}_2^{(0)}$.
2. For $m = 1, 2, \ldots$
   - Compute $E^{(m-1)}[\log p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|y_{1:T})]$ given by the expression

$$\int \log p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|y_{1:T}) p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(m-1)}, y_{1:T}) d\boldsymbol{\theta}_1. \tag{1.21}$$

   This is the E-step.
   - Set $\boldsymbol{\theta}_2^{(m)}$ to the value that maximizes (1.21). This is the M-step.

At each iteration the algorithm satisfies that $p(\boldsymbol{\theta}_2^{(m)}|y_{1:T}) \geq p(\boldsymbol{\theta}_2^{(m-1)}|y_{1:T})$. There is no guarantee that the EM algorithm converges to the mode; in the case of multimodal distributions the algorithm may converge to a local

mode. Various alternatives have been considered to avoid getting stuck in a local mode, such as running the algorithm with several different random initial points, or using simulated annealing methods. Some extensions of the EM algorithm include the ECM (expectation-conditional-maximization) algorithm, the ECME (expectation-conditional-maximization-either, a variant of the ECM in which either the log-posterior density or the expected log-posterior density is maximized) and the SEM (supplemented EM) algorithms (see Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin 2014 and references therein) and stochastic versions of the EM algorithm such as the MCEM (Monte Carlo EM, see Wei and Tanner 1990).

Example 1.5 *ML, MAP, and LS estimators for the AR(1) model.* Consider the AR(1) model $y_t = \phi y_{t-1} + \epsilon_t$, with $\epsilon_t \sim N(0,1)$. In this case $v = 1$ and $\theta = \phi$. The conditional MLE is found by maximizing $\exp\{-Q(\phi)/2\}$ or, equivalently, by minimizing $Q(\phi)$. Therefore, we obtain $\hat{\phi} = \phi_{\text{ML}} = \sum_{t=2}^{T} y_t y_{t-1} / \sum_{t=2}^{T} y_{t-1}^2$. Similarly, the MLE for the unconditional likelihood function is obtained by maximizing $p(y_{1:T}|\phi)$ or, equivalently, by minimizing the expression

$$-0.5[\log(1-\phi^2) - Q^*(\phi)].$$

Newton–Raphson or scoring methods can be used to find $\hat{\phi}$. As an illustration, the conditional and unconditional ML estimators were found for

100 samples from an AR(1) with $\phi = 0.9$. Figure 1.8 shows a graph with the conditional and unconditional log-likelihood functions (solid and dotted lines respectively). The points correspond to the maximum likelihood estimators with $\hat{\phi} = 0.9069$ and $\hat{\phi} = 0.8979$ being the MLEs for the conditional and unconditional likelihoods, respectively. For the unconditional case, a Newton–Raphson algorithm was used to find the maximum. The algorithm converged after five iterations with a starting value of 0.1.

Figure 1.9 shows the log-posterior densities of ϕ under Gaussian priors of the form $\phi \sim N(\mu, c)$, for $\mu = 0$, $c = 1.0$ (left panel) and $c = 0.01$ (right panel). Note that this prior does not impose any restriction on ϕ and so it gives nonnegative probability to values of ϕ that lie in the nonstationary region. It is possible to choose priors on ϕ whose support is the stationary region. This will be considered in Chapter 2. Figure 1.9 illustrates the effect of the prior on the MAP estimators. For a prior $\phi \sim N(0,1)$, the MAP estimators are $\hat{\phi}_{\text{MAP}} = 0.9051$ and $\hat{\phi}_{\text{MAP}} = 0.8963$ for the conditional and unconditional likelihoods, respectively. When a smaller value of c is considered, or in other words, when the prior distribution is more concentrated around zero, then the MAP estimates shift toward the prior mean. For a prior $\phi \sim N(0, 0.01)$, the MAP estimators are $\hat{\phi}_{\text{MAP}} = 0.7588$ and $\hat{\phi}_{\text{MAP}} = 0.7550$ for the conditional and unconditional likelihoods, respectively.

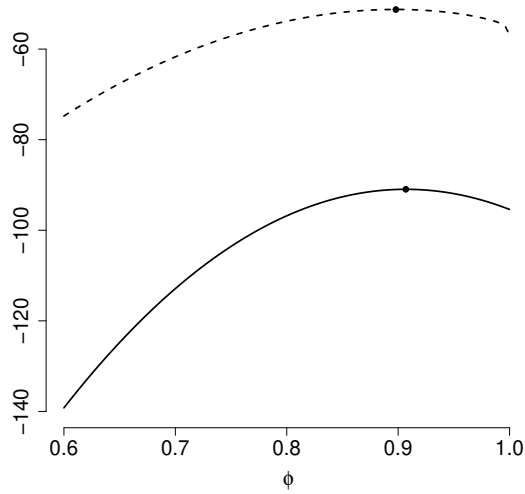


Figure 1.8 *Conditional and unconditional log-likelihoods (solid and dashed lines, respectively) based on 100 observations simulated from an* $AR(1)$ *with* $\phi = 0.9$.

It is also possible to obtain the least squares estimators for the conditional and unconditional likelihoods. For the conditional case, the least squares (LS) estimator is obtained by minimizing the conditional sum of squares $Q(\phi)$, and so in this case $\phi_{\mathrm{ML}} = \phi_{\mathrm{LS}}$. In the unconditional case, the LS estimator is found by minimizing the unconditional sum of squares $Q^*(\phi)$, and so the LS and the ML estimators do not coincide.

1.5.2 Traditional Least Squares

Likelihood and Bayesian approaches for fitting linear autoregressions rely on very standard methods of linear regression analysis. Therefore, some review of the central ideas and results in regression is in order and given here. This introduces notation and terminology that will be used throughout the book.

A linear model with a univariate response variable and $p > 0$ regression variables (otherwise predictors or covariates) has the form

$$y_i = \mathbf{f}_i'\boldsymbol{\beta} + \epsilon_i,$$

for $i = 1, 2, \ldots$, where y_i is the i-th observation on the response variable, and has corresponding values of the regressors in the design vector $\mathbf{f}_i' = (f_{i1}, \ldots, f_{ip})$. The design vectors are assumed known and fixed prior to

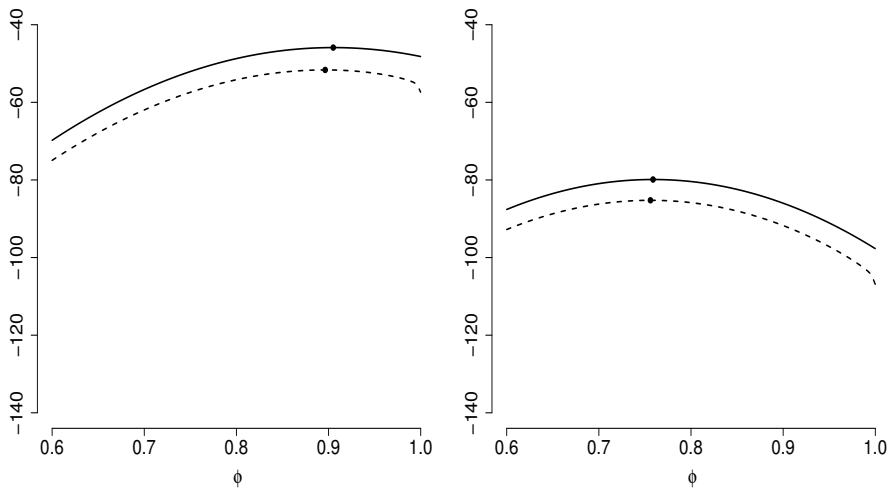


Figure 1.9 *Conditional and unconditional log-posterior densities (solid and dashed lines, respectively) based on 100 observations simulated from an AR(1) with* $\phi = 0.9$. *The posterior densities were obtained with priors of the form* $\phi \sim N(0, c)$, *for* $c = 1$ *(left panel) and* $c = 0.01$ *(right panel).*

observing the corresponding responses. The error terms ϵ_i are assumed independent and normal, distributed as $N(\epsilon_i|0, v)$ with some variance v. The regression parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is to be estimated, along with the error variance. Now assume we have a set of n responses denoted as $\mathbf{y} = (y_1, \ldots, y_n)'$. We note that this notation is general and so, the responses are not necessarily temporally indexed and n is not necessarily equal to T. The model for $\mathbf{y}$ is

$$\mathbf{y} = \mathbf{F}'\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.22}$$

where $\mathbf{F}$ is the known $p \times n$ design matrix with i-th column $\mathbf{f}_i$. In addition, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$, with $\boldsymbol{\epsilon} \sim N(\boldsymbol{\epsilon}|0, v\mathbf{I}_n)$, and $\mathbf{I}_n$ the $n \times n$ identity matrix. The sampling distribution is defined as

$$p(\mathbf{y}|\mathbf{F}, \boldsymbol{\beta}, v) = \prod_{i=1}^{n} N(y_i|\mathbf{f}_i'\boldsymbol{\beta}, v) = (2\pi v)^{-n/2} \exp(-Q(\mathbf{y}, \boldsymbol{\beta})/2v),$$

where $Q(\mathbf{y}, \boldsymbol{\beta}) = (\mathbf{y} - \mathbf{F}'\boldsymbol{\beta})'(\mathbf{y} - \mathbf{F}'\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{f}_i'\boldsymbol{\beta})^2$. This gives a likelihood function for $(\boldsymbol{\beta}, v)$. We can also write $Q(\mathbf{y}, \boldsymbol{\beta})$ as

$$Q(\mathbf{y}, \boldsymbol{\beta}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{F}\mathbf{F}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + R,$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{F}\mathbf{F}')^{-1}\mathbf{F}\mathbf{y}$ and $R = (\mathbf{y} - \mathbf{F}'\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{F}'\hat{\boldsymbol{\beta}})$. This assumes that $\mathbf{F}$ is of full rank p, otherwise an appropriate linear transformation of the design vectors can be used to reduce $\mathbf{F}$ to a full rank matrix and the model decreases in dimension. Here $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ and the residual sum of squares R gives the MLE of v as R/n; a more usual estimate of v is $s^2 = R/(n-p)$, with $n-p$ being the associated degrees of freedom.

1.5.3 Full Bayesian Analysis

We summarize some aspects of various Bayesian approaches for fitting linear models, including reference and conjugate analyses. Nonconjugate analyses may lead to posterior distributions that are not available in closed form. Therefore, nonconjugate inferential approaches often rely on obtaining random draws from the posterior distribution using Markov chain Monte Carlo methods, which will be used a good deal later in this book. Some key references are the books of Box and Tiao (1973) and Zellner (1996). The book of Greenberg (2008) provides an excellent introduction to Bayesian statistics and econometrics using a simulation-based approach.

1.5.3.1 Reference Bayesian Analysis

Reference Bayesian analysis is based on the traditional reference (improper) prior $p(\boldsymbol{\beta}, v) \propto 1/v$. The corresponding posterior density is $p(\boldsymbol{\beta}, v|\mathbf{y}, \mathbf{F}) \propto p(\mathbf{y}|\mathbf{F}, \boldsymbol{\beta}, v)/v$ and has the following features:

- The marginal posterior for $\boldsymbol{\beta}$ is a multivariate Student-t with $n-p$ degrees of freedom. It has mode $\hat{\boldsymbol{\beta}}$, scale matrix $s^2(\mathbf{F}\mathbf{F}')^{-1}$, and density

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{F}) = c(n,p)|\mathbf{F}\mathbf{F}'|^{1/2}\{1 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{F}\mathbf{F}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/(n-p)s^2\}^{-n/2}$$

  with $c(n,p) = \Gamma(n/2)/[\Gamma((n-p)/2)(s^2\pi(n-p))^{p/2}]$, where $\Gamma(\cdot)$ is the gamma function. When n is large, the posterior is approximately normal, $N(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, s^2(\mathbf{F}\mathbf{F}')^{-1})$. Note also that, given v, the conditional posterior for $\boldsymbol{\beta}$ is exactly normal, namely $N(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, v(\mathbf{F}\mathbf{F}')^{-1})$.
- The marginal posterior for v is inverse gamma with parameters $(n-p)/2$ and $(n-p)s^2/2$, or $(v|\mathbf{y}, \mathbf{F}) \sim IG((n-p)/2, (n-p)s^2/2)$.
- The total sum of squares of the responses $\mathbf{y}'\mathbf{y} = \sum_{i=1}^{n} y_i^2$ factorizes as $\mathbf{y}'\mathbf{y} = R + \hat{\boldsymbol{\beta}}'\mathbf{F}\mathbf{F}'\hat{\boldsymbol{\beta}}$. The sum of squares explained by the regression is $\mathbf{y}'\mathbf{y} - R = \hat{\boldsymbol{\beta}}'\mathbf{F}\mathbf{F}'\hat{\boldsymbol{\beta}}$; this is also called the fitted sum of squares, and a larger value implies a smaller residual sum of squares and, in this sense, a closer fit to the data.
- Under a proper prior distribution for $(\boldsymbol{\beta}, v)$ the marginal density of $(\mathbf{y}|\mathbf{F})$

can be obtained as

$$p(\mathbf{y}|\mathbf{F}) = \int p(\mathbf{y}|\mathbf{F}, \boldsymbol{\beta}, v)p(\boldsymbol{\beta}, v)d\boldsymbol{\beta}dv.$$

Note that the reference prior used here is improper, invalidating the calculation of a proper marginal density for $(\mathbf{y}|\mathbf{F})$. However, one can still obtain an expression for $p(\mathbf{y}|\mathbf{F})$ up to a proportionality constant as

$$p(\mathbf{y}|\mathbf{F}) = \int \frac{p(\mathbf{y}|\mathbf{F}, \boldsymbol{\beta}, v)}{v}\, d\boldsymbol{\beta}dv \propto \frac{\Gamma((n-p)/2)}{\pi^{(n-p)/2}}|\mathbf{F}\mathbf{F}'|^{-1/2}R^{-(n-p)/2}.$$

This can also be written as

$$p(\mathbf{y}|\mathbf{F}) \propto \frac{\Gamma((n-p)/2)}{\pi^{(n-p)/2}}|\mathbf{F}\mathbf{F}'|^{-1/2}(\mathbf{y}'\mathbf{y})^{(p-n)/2}\{1-\hat{\boldsymbol{\beta}}'\mathbf{F}\mathbf{F}'\hat{\boldsymbol{\beta}}/(\mathbf{y}'\mathbf{y})\}^{(p-n)/2}.$$

For large n, the term $\{1-\hat{\boldsymbol{\beta}}'\mathbf{F}\mathbf{F}'\hat{\boldsymbol{\beta}}/(\mathbf{y}'\mathbf{y})\}^{(p-n)/2}$ in the above expression is approximately $\exp(\hat{\boldsymbol{\beta}}'\mathbf{F}\mathbf{F}'\hat{\boldsymbol{\beta}}/2r)$ where $r = \mathbf{y}'\mathbf{y}/(n-p)$.

Some additional comments:

- For models with the same number of parameters that differ only through $\mathbf{F}$, the corresponding observed data densities will tend to be larger for those models with larger values of the explained sum of squares $\hat{\boldsymbol{\beta}}'\mathbf{F}\mathbf{F}'\hat{\boldsymbol{\beta}}$ (though the determinant term plays a role too). Otherwise, $p(\mathbf{y}|\mathbf{F})$ also depends on the parameter dimension p.
- *Orthogonal regression.* If $\mathbf{F}\mathbf{F}' = k\mathbf{I}_p$ for some k, then everything simplifies. Write $\mathbf{f}_j^*$ for the j-th column of $\mathbf{F}'$, and β_j for the corresponding component of the parameter vector $\boldsymbol{\beta}$. Then $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ where each $\hat{\beta}_j$ is the individual MLE from a model on $\mathbf{f}_j^*$ alone, i.e., $\mathbf{y} = \mathbf{f}_j^*\beta_j + \boldsymbol{\epsilon}$, and the elements of $\boldsymbol{\beta}$ are uncorrelated under the posterior T distribution. The explained sum of squares partitions into a sum of individual pieces too, namely $\hat{\boldsymbol{\beta}}'\mathbf{F}\mathbf{F}'\hat{\boldsymbol{\beta}} = \sum_{j=1}^p \mathbf{f}_j^{*'}\mathbf{f}_j^*\hat{\beta}_j^2$, and so calculations and interpretations are easy.

Example 1.6 *Reference analysis in the AR(1) model.* For the conditional likelihood using the notation above we have $\mathbf{y} = (y_2, \ldots, y_T)'$, $\mathbf{F} = (y_1, \ldots, y_{T-1})$ and the reference prior $p(\phi, v) \propto 1/v$. The MLE for ϕ is $\phi_{\mathrm{ML}} = \sum_{t=2}^T y_{t-1}y_t / \sum_{t=1}^{T-1} y_t^2$. Under the reference prior $\phi_{\mathrm{MAP}} = \phi_{\mathrm{ML}}$. The residual sum of squares is given by

$$R = \sum_{t=2}^T y_t^2 - \frac{(\sum_{t=2}^T y_t y_{t-1})^2}{\sum_{t=1}^{T-1} y_t^2},$$

and so $s^2 = R/(T-2)$ estimates v. The marginal posterior distribution

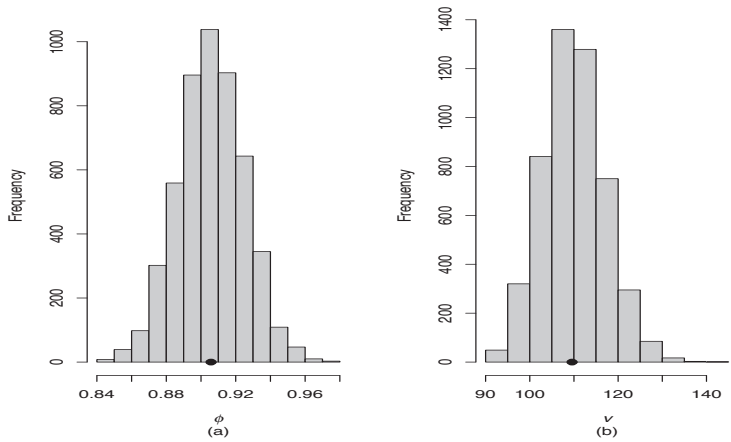


Figure 1.10 *(a)* $p(\phi|\mathbf{y},\mathbf{F})$*; (b)* $p(v|\mathbf{y},\mathbf{F})$.

of ϕ is a univariate Student-t distribution with $T-2$ degrees of freedom, centered at ϕ_{ML} with scale $s^2(\mathbf{F}\mathbf{F}')^{-1}$, i.e.,

$$(\phi|\mathbf{y},\mathbf{F}) \sim t_{(T-2)}\left(m, \frac{C}{T-2}\right),$$

where

$$m = \frac{\sum_{t=2}^{T} y_{t-1}y_t}{\sum_{t=1}^{T-1} y_t^2}$$

and

$$C = \frac{\sum_{t=2}^{T} y_t^2 \sum_{t=2}^{T} y_{t-1}^2 - \left(\sum_{t=2}^{T} y_t y_{t-1}\right)^2}{\left(\sum_{t=1}^{T-1} y_t^2\right)^2}.$$

Finally, the posterior for v is a scaled inverse chi-squared with $T-2$ degrees of freedom and scale s^2, i.e., $Inv\text{-}\chi^2(v|T-2,s^2)$ or, equivalently, an inverse gamma with parameters $(T-2)/2$ and $(T-2)s^2/2$, $IG(v|(T-2)/2,(T-2)s^2/2)$.

As an illustration, a reference analysis was performed for a time series of 500 points simulated from an AR(1) model with $\phi = 0.9$ and $v = 100$. Figures 1.10 (a) and (b) display the marginal posterior densities of $(\phi|\mathbf{y},\mathbf{F})$ and $(v|\mathbf{y},\mathbf{F})$ based on 5,000 samples from the joint posterior of ϕ and v. The circles in the histogram indicate ϕ_{ML} and s^2, respectively.

1.5.3.2 Conjugate Bayesian Analysis

Let $p(y_t|\boldsymbol{\theta})$ be a likelihood function. A class Π of prior distributions forms a *conjugate family* if the posterior $p(\boldsymbol{\theta}|y_t)$ belongs to the class Π for every prior $p(\boldsymbol{\theta})$ in Π.

Consider again the model $\mathbf{y} = \mathbf{F}'\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\mathbf{F}$ a known $p \times n$ design matrix and $\boldsymbol{\epsilon} \sim N(\boldsymbol{\epsilon}|0, v\mathbf{I}_n)$. In a conjugate Bayesian analysis for this model priors of the form

$$p(\boldsymbol{\beta}, v) = p(\boldsymbol{\beta}|v)p(v) = N(\boldsymbol{\beta}|\mathbf{m}_0, v\mathbf{C}_0) \times IG(v|n_0/2, d_0/2) \tag{1.23}$$

are taken with $\mathbf{m}_0$ a vector of dimension p and $\mathbf{C}_0$ a $p \times p$ matrix. Both $\mathbf{m}_0$ and $\mathbf{C}_0$ are known quantities. The corresponding posterior distribution has the following form:

$$\begin{aligned} p(\boldsymbol{\beta}, v|\mathbf{y}, \mathbf{F}) \quad \propto \quad & v^{-[(p+n+n_0)/2+1]} \times \\ & e^{-[(\boldsymbol{\beta}-\mathbf{m}_0)'\mathbf{C}_0^{-1}(\boldsymbol{\beta}-\mathbf{m}_0)+(\mathbf{y}-\mathbf{F}'\boldsymbol{\beta})'(\mathbf{y}-\mathbf{F}'\boldsymbol{\beta})+d_0]/2v}. \end{aligned}$$

This analysis has the following features:

- $(\mathbf{y}|\mathbf{F}, v) \sim N(\mathbf{F}'\mathbf{m}_0, v(\mathbf{F}'\mathbf{C}_0\mathbf{F} + \mathbf{I}_n))$ and $(\mathbf{y}|\mathbf{F})$ follows a multivariate Student-t distribution, i.e., $(\mathbf{y}|\mathbf{F}) \sim T_{n_0}[\mathbf{F}'\mathbf{m}_0, d_0(\mathbf{F}'\mathbf{C}_0\mathbf{F} + \mathbf{I}_n)/n_0]$.
- The posterior distribution of $\boldsymbol{\beta}$ given v is Gaussian, $(\boldsymbol{\beta}|\mathbf{y}, \mathbf{F}, v) \sim N(\mathbf{m}, v\mathbf{C})$, with

$$\begin{aligned} \mathbf{m} &= \mathbf{m}_0 + \mathbf{C}_0\mathbf{F}[\mathbf{F}'\mathbf{C}_0\mathbf{F} + \mathbf{I}_n]^{-1}(\mathbf{y} - \mathbf{F}'\mathbf{m}_0) \\ \mathbf{C} &= \mathbf{C}_0 - \mathbf{C}_0\mathbf{F}[\mathbf{F}'\mathbf{C}_0\mathbf{F} + \mathbf{I}_n]^{-1}\mathbf{F}'\mathbf{C}_0, \end{aligned}$$

  or, defining $\mathbf{e} = \mathbf{y} - \mathbf{F}'\mathbf{m}_0$, $\mathbf{Q} = \mathbf{F}'\mathbf{C}_0\mathbf{F} + \mathbf{I}_n$, and $\mathbf{A} = \mathbf{C}_0\mathbf{F}\mathbf{Q}^{-1}$ we can also write $\mathbf{m} = \mathbf{m}_0 + \mathbf{A}\mathbf{e}$ and $\mathbf{C} = \mathbf{C}_0 - \mathbf{A}\mathbf{Q}\mathbf{A}'$.
- $(v|\mathbf{y}, \mathbf{F}) \sim IG(n^*/2, d^*/2)$ with $n^* = n + n_0$ and

$$d^* = (\mathbf{y} - \mathbf{F}'\mathbf{m}_0)'\mathbf{Q}^{-1}(\mathbf{y} - \mathbf{F}'\mathbf{m}_0) + d_0.$$

- $(\boldsymbol{\beta}|\mathbf{y}, \mathbf{F}) \sim T_{n^*}[\mathbf{m}, d^*\mathbf{C}/n^*]$.

Example 1.7 *Conjugate analysis in the AR(1) model using the conditional likelihood.* Assume we choose a prior of the form $\phi|v \sim N(0, v)$ and $v \sim IG(n_0/2, d_0/2)$, with n_0 and d_0 known. Then, $p(\phi|\mathbf{y}, \mathbf{F}, v) \sim N(m, vC)$ with

$$m = \frac{\sum_{t=1}^{T-1} y_t y_{t+1}}{\sum_{t=1}^{T-1} y_t^2 + 1}, \quad C = \frac{1}{1 + \sum_{t=1}^{T-1} y_t^2},$$

$(v|\mathbf{y}, \mathbf{F}) \sim IG(n^*/2, d^*/2)$ with $n^* = T + n_0 - 1$ and

$$d^* = \sum_{t=2}^{T} y_t^2 - \frac{\left(\sum_{t=1}^{T-1} y_t y_{t+1}\right)^2}{\sum_{t=1}^{T-1} y_t^2 + 1} + d_0.$$

1.5.4 Nonconjugate Bayesian Analysis

For the general regression model, the reference and conjugate priors produce joint posterior distributions that have closed analytical forms. However, in many scenarios it is either not possible or not desirable to work with a conjugate prior or with a prior that leads to a posterior distribution that can be written in analytical form. In these cases it might be possible to use analytical or numerical approximations to the posterior. Another alternative consists on summarizing the inference by obtaining random draws from the posterior distribution. Sometimes it is possible to obtain such draws by direct simulation, but often this is not the case, and so methods such as Markov chain Monte Carlo (MCMC) are used.

Consider again the AR(1) model under the full likelihood (1.17). No conjugate prior is available in this case. Furthermore, a prior of the form $p(\phi, v) \propto 1/v$ does not produce a posterior distribution in closed form. In fact, the joint posterior distribution is such that

$$p(\phi, v|y_{1:T}) \propto v^{-(T/2+1)}(1-\phi^2)^{1/2} \exp\left\{\frac{-Q^*(\phi)}{2v}\right\}. \tag{1.24}$$

Several approaches could be considered to summarize this posterior distribution. For example, we could use a normal approximation to the distribution $p(\phi, v|y_{1:T})$ centered at the ML or MAP estimates of (ϕ, v). In general, the normal approximation to a posterior distribution $p(\boldsymbol{\theta}|y_{1:T})$ is given by

$$p(\boldsymbol{\theta}|y_{1:T}) \approx N(\hat{\boldsymbol{\theta}}, v(\hat{\boldsymbol{\theta}})), \tag{1.25}$$

with $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\text{MAP}}$ and $v(\boldsymbol{\theta})^{-1} = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \log p(\boldsymbol{\theta}|y_{1:T})$.

Alternatively, it is possible to use iterative MCMC methods to obtain samples from $p(\phi, v|y_{1:T})$. We summarize two of the most widely used MCMC methods below: the Metropolis algorithm and the Gibbs sampler. For full consideration of MCMC methods see, for example, Gamerman and Lopes (2006) and Robert and Casella (2005).

1.5.5 Posterior Sampling

1.5.5.1 The Metropolis-Hastings Algorithm

Assume that our target posterior distribution, $p(\boldsymbol{\theta}|y_{1:T})$, can be computed up to a normalizing constant. The Metropolis-Hastings algorithm (Metropolis et al. 1953, Hastings 1970) creates a sequence of random draws $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, ..., whose distributions converge to the target distribution. Each sequence can be considered as a Markov chain whose stationary distribution is $p(\boldsymbol{\theta}|y_{1:T})$. The sampling algorithm can be summarized as follows:

- Draw a starting point $\boldsymbol{\theta}^{(0)}$ with $p(\boldsymbol{\theta}^{(0)}|y_{1:T}) > 0$ from a starting distribution $p_0(\boldsymbol{\theta})$.
- For $m = 1, 2, \ldots$
  1. Sample a candidate $\boldsymbol{\theta}^*$ from a jumping distribution $J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)})$. If the distribution J is symmetric, i.e., if $J(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b) = J(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a)$ for all $\boldsymbol{\theta}_a, \boldsymbol{\theta}_b$, and m, then we refer to the algorithm as the Metropolis algorithm. If J_m is not symmetric, we refer to the algorithm as the Metropolis-Hastings algorithm.
  2. Compute the importance ratio
$$r = \frac{p(\boldsymbol{\theta}^*|y_{1:T})/J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)})}{p(\boldsymbol{\theta}^{(m-1)}|y_{1:n})/J(\boldsymbol{\theta}^{(m-1)}|\boldsymbol{\theta}^*)}.$$
  3. Set
$$\boldsymbol{\theta}^{(m)} = \begin{cases} \boldsymbol{\theta}^* & \text{with probability} = \min(r, 1) \\ \boldsymbol{\theta}^{(m-1)} & \text{otherwise.} \end{cases}$$

An ideal jumping distribution is one that is easy to sample from and makes the evaluation of the importance ratio easy. In addition, the jumping distributions $J(\cdot|\cdot)$ should be such that each jump moves a reasonable distance in the parameter space so that the random walk is not too slow, and also, the jumps should not be rejected too often.

1.5.5.2 Gibbs Sampling

Assume $\boldsymbol{\theta}$ has k components, i.e., $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$. The Gibbs sampler (Geman and Geman 1984) can be viewed as a special case of the Metropolis-Hastings algorithm for which the jumping distribution at each iteration m is a function $p(\boldsymbol{\theta}_j^*|\boldsymbol{\theta}_{-j}^{(m-1)}, y_{1:T})$, where $\boldsymbol{\theta}_{-j}$ denotes a vector with all the components of $\boldsymbol{\theta}$ except for component $\boldsymbol{\theta}_j$. In other words, for each component of $\boldsymbol{\theta}$ we do a Metropolis-Hastings step for which the jumping distribution is given by

$$J_j(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)}) = \begin{cases} p(\boldsymbol{\theta}_j^*|\boldsymbol{\theta}_{-j}^{(m-1)}, y_{1:T}) & \text{if } \boldsymbol{\theta}_{-j}^* = \boldsymbol{\theta}_{-j}^{(m-1)} \\ 0 & \text{otherwise,} \end{cases}$$

and so $r = 1$ and every jump is accepted.

If it is not possible to sample from $p(\boldsymbol{\theta}_j^*|\boldsymbol{\theta}_{-j}^{(m)}, y_{1:T})$ an approximation, say $g(\boldsymbol{\theta}_j^*|\boldsymbol{\theta}_{-j}^{(m-1)})$, can be considered. However, in this case it is necessary to compute the Metropolis acceptance ratio r.

1.5.5.3 Convergence

In theory, a value from the posterior distribution of $(\boldsymbol{\theta}|y_{1:T})$ is obtained by MCMC when the number of iterations of the chain approaches infinity. In practice, a value obtained after a sufficiently large number of iterations is taken as a draw from the target posterior distribution of $(\boldsymbol{\theta}|y_{1:T})$. How can we determine how many MCMC iterations are enough to obtain convergence? As pointed out in Gamerman and Lopes (2006), there are two general approaches to the study of convergence. One is probabilistic and it consists on measuring distances and bounds on distribution functions generated from a chain. So, for example, it is possible to measure the total variation distance between the distribution of the chain at iteration i and the target distribution of $(\boldsymbol{\theta}|y_{1:T})$. An alternative approach consists on studying the convergence of the chain from a statistical perspective. This approach is easier and more practical than the probabilistic one; however, it cannot guarantee convergence.

There are several ways of monitoring convergence from a statistical viewpoint, ranging from graphical displays of the MCMC traces for all or some of the model parameters or functions of such parameters, to sophisticated statistical tests. As mentioned before, one of the two main problems with simulation-based iterative methods is deciding whether the chain has reached convergence, i.e., if the number of iterations is large enough to guarantee that the available samples are drawn from the target posterior distribution. In addition, large within-sequence correlation may lead to inferences that are not precise enough. In other words, if M draws from a chain with very large within-sequence correlation are used to represent the posterior distribution, the "effective" number of draws used in such representation is far smaller than M. Some well-known tests to assess convergence are implemented as R packages (R Core Team 2018), such as Bayesian Output Analysis (BOA, Smith 2007) and Convergence Diagnosis and Output Analysis for MCMC (CODA, Plummer, Best, Cowles, and Vines 2006). Specifically, these packages include convergence diagnostics such as the Brooks, Gelman, and Rubin diagnostic for a list of sequences (Brooks and Gelman 1998; Gelman and Rubin 1992), which monitors the mixing of the simulated sequences by comparing the within and between variance of the sequences; the Geweke diagnostic (1992) and Heidelberger and Welch diagnostic (1983), which are based on sequential testing of portions of the simulated chains to determine if they correspond to samples from the same distribution; and the Raftery and Lewis method (Raftery and Lewis 1992), which considers the problem of how many iterations are needed to estimate a particular posterior quantile from a single MCMC chain. BOA and CODA also provide the user with some descriptive plots