Springer Series in Statistics

Jianqing Fan
Qiwei Yao

Nonlinear Time Series

Nonparametric and Parametric Methods

Springer

Springer Series in Statistics

Springer
New York
Berlin
Heidelberg
Hong Kong
London
Milan
Paris
Tokyo

Jianqing Fan
Qiwei Yao

Nonlinear Time Series

Nonparametric and Parametric Methods

Springer

Jianqing Fan
Department of Operations Research
and Financial Engineering
Princeton University
Princeton, NJ 08544
USA
jqfan@princeton.edu

Qiwei Yao
Department of Statistics
London School of Economics
London WC2A 2AE
UK
q.yao@lse.ac.uk

Library of Congress Cataloging-in-Publication Data
Fan, Jianqing.
Nonlinear time series : nonparametric and parametric methods / Jianqing Fan, Qiwei Yao.
p. cm. — (Springer series in statistics)
Includes bibliographical references and index.
ISBN 0-387-95170-9 (alk. paper)
1. Time-series analysis. 2. Nonlinear theories. I. Yao, Qiwei. II. Title. III. Series.
QA280 .F36 2003
519.2′32—dc21 2002036549

ISBN 0-387-95170-9 Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 10788773

Typesetting: Pages created by the authors using a Springer LaTeX 2e macro package.

www.springer-ny.com

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

To those

Who educate us;

Whom we love; and

With whom we collaborate

Preface

Among many exciting developments in statistics over the last two decades, nonlinear time series and data-analytic nonparametric methods have greatly advanced along seemingly unrelated paths. In spite of the fact that the application of nonparametric techniques in time series can be traced back to the 1940s at least, there still exists healthy and justified skepticism about the capability of nonparametric methods in time series analysis. As enthusiastic explorers of the modern nonparametric toolkit, we feel obliged to assemble together in one place the newly developed relevant techniques. The aim of this book is to advocate those modern nonparametric techniques that have proven useful for analyzing real time series data, and to provoke further research in both methodology and theory for nonparametric time series analysis.

Modern computers and the information age bring us opportunities with challenges. Technological inventions have led to the explosion in data collection (e.g., daily grocery sales, stock market trading, microarray data). The Internet makes big data warehouses readily accessible. Although classic parametric models, which postulate global structures for underlying systems, are still very useful, large data sets prompt the search for more refined structures, which leads to better understanding and approximations of the real world. Beyond postulated parametric models, there are infinite other possibilities. Nonparametric techniques provide useful exploratory tools for this venture, including the suggestion of new parametric models and the validation of existing ones.

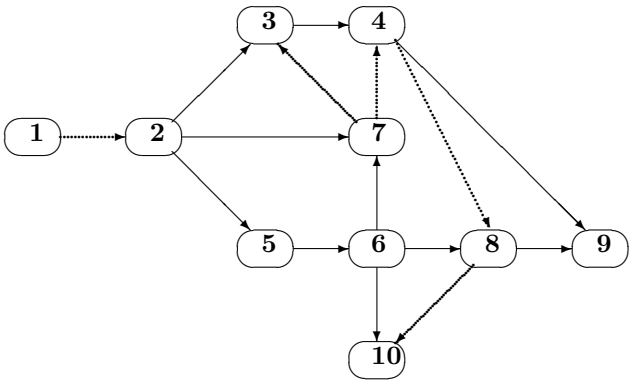
In this book, we present an up-to-date picture of techniques for analyzing time series data. Although we have tried to maintain a good balance

among methodology, theory, and numerical illustration, our primary goal is to present a comprehensive and self-contained account for each of the key methodologies. For practical relevant time series models, we aim for exposure with definition, probability properties (if possible), statistical inference methods, and numerical examples with real data sets. We also indicate where to find our (only our!) favorite computing codes to implement these statistical methods. When soliciting real-data examples, we attempt to maintain a good balance among different disciplines, although our personal interests in quantitative finance, risk management, and biology can be easily seen. It is our hope that readers can apply these techniques to their own data sets.

We trust that the book will be of interest to those coming to the area for the first time and to readers more familiar with the field. Application-oriented time series analysts will also find this book useful, as it focuses on methodology and includes several case studies with real data sets. We believe that nonparametric methods must go hand-in-hand with parametric methods in applications. In particular, parametric models provide explanatory power and concise descriptions of the underlying dynamics, which, when used sensibly, is an advantage over nonparametric models. For this reason, we have also provided a compact view of the parametric methods for both linear and selected nonlinear time series models. This will also give new comers sufficient information on the essence of the more classical approaches. We hope that this book will reflect the power of the integration of nonparametric and parametric approaches in analyzing time series data. The book has been prepared for a broad readership—the prerequisites are merely sound basic courses in probability and statistics. Although advanced mathematics has provided valuable insights into nonlinear time series, the methodological power of both nonparametric and parametric approaches can be understood without sophisticated technical details. Due to the innate nature of the subject, it is inevitable that we occasionally appeal to more advanced mathematics; such sections are marked with a "*". Most technical arguments are collected in a "Complements" section at the end of each chapter, but key ideas are left within the body of the text.

The introduction in Chapter 1 sets the scene for the book. Chapter 2 deals with basic probabilistic properties of time series processes. The highlights include strict stationarity via ergodic Markov chains (§2.1) and mixing properties (§2.6). We also provide a generic central limit theorem for kernel-based nonparametric regression estimation for α-mixing processes. A compact view of linear ARMA models is given in Chapter 3, including Gaussian MLE (§3.3), model selection criteria (§3.4), and linear forecasting with ARIMA models (§3.7). Chapter 4 introduces three types of parametric nonlinear models. An introduction on threshold models that emphasizes developments after Tong (1990) is provided. ARCH and GARCH models are presented in detail, as they are less exposed in statistical literature. The chapter concludes with a brief account of bilinear models. Chapter 5

introduces the nonparametric kernel density estimation. This is arguably the simplest problem for understanding nonparametric techniques. The relation between "localization" for nonparametric problems and "whitening" for time series data is elucidated in §5.3. Applications of nonparametric techniques for estimating time trends and univariate autoregressive functions can be found in Chapter 6. The ideas in Chapter 5 and §6.3 provide a foundation for the nonparametric techniques introduced in the rest of the book. Chapter 7 introduces spectral density estimation and nonparametric procedures for testing whether a series is white noise. Various high-order autoregressive models are highlighted in Chapter 8. In particular, techniques for estimating nonparametric functions in FAR models are introduced in §8.3. The additive autoregressive model is exposed in §8.5, and methods for estimating conditional variance or volatility functions are detailed in §8.7. Chapter 9 outlines approaches to testing a parametric family of models against a family of structured nonparametric models. The wide applicability of the generalized likelihood ratio test is emphasized. Chapter 10 deals with nonlinear prediction. It highlights the features that distinguish nonlinear prediction from linear prediction. It also introduces nonparametric estimation for conditional predictive distribution functions and conditional minimum volume predictive intervals.



The interdependence of the chapters is depicted above, where solid directed lines indicate prerequisites and dotted lines indicate weak associations. For lengthy chapters, the dependence among sections is not very strong. For example, the sections in Chapter 4 are fairly independent, and so are those in Chapter 8 (except that §8.4 depends on §8.3, and §8.7 depends on the rest). They can be read independently. Chapter 5 and §6.3 provide a useful background for nonparametric techniques. With an understanding of this material, readers can jump directly to sections in Chapters 8 and 9. For readers who wish to obtain an overall impression of the book, we suggest reading Chapter 1, §2.1, §2.2, Chapter 3, §4.1, §4.2, Chapter 5,

§6.3, §8.3, §8.5, §8.7, §9.1, §9.2, §9.4, §9.5 and §10.1. These core materials may serve as the text for a graduate course on nonlinear time series.

Although the scope of the book is wide, we have not achieved completeness. The nonparametric methods are mostly centered around kernel/local polynomial based smoothing. Nonparametric hypothesis testing with structured nonparametric alternatives is mainly confined to the generalized likelihood ratio test. In fact, many techniques that are introduced in this book have not been formally explored mathematically. State-space models are only mentioned briefly within the discussion on bilinear models and stochastic volatility models. Multivariate time series analysis is untouched. Another noticeable gap is the lack of exposure of the variety of parametric nonlinear time series models listed in Chapter 3 of Tong (1990). This is undoubtedly a shortcoming. In spite of the important initial progress, we feel that the methods and theory of statistical inference for some of those models are not as well-established as, for example, ARCH/GARCH models or threshold models. Their potential applications should be further explored.

Extensive effort was expended in the composition of the reference list, which, together with the bibliographical notes, should guide readers to a wealth of available materials. Although our reference list is long, it merely reflects our immediate interests. Many important papers that do not fit our presentation have been omitted. Other omissions and discrepancies are inevitable. We apologize for their occurrence.

Although we both share the responsibility for the whole book, Jianqing Fan was the lead author for Chapters 1 and 5–9 and Qiwei Yao for Chapters 2–4 and 10.

Many people have been of great help to our work on this book. In particular, we would like to thank Hong-Zhi An, Peter Bickel, Peter Brockwell, Yuzhi Cai, Zongwu Cai, Kung-Sik Chan, Cees Diks, Rainer Dahlhaus, Liudas Giraitis, Peter Hall, Wai-Keung Li, Jianzhong Lin, Heng Peng, Liang Peng, Stathis Paparoditis, Wolfgang Polonik, John Rice, Peter Robinson, Richard Smith, Howell Tong, Yingcun Xia, Chongqi Zhang, Wenyang Zhang, and anonymous reviewers. Thanks also go to *Biometrika* for permission to reproduce Figure 6.10, to Blackwell Publishers Ltd. for permission to reproduce Figures 8.8, 8.15, 8.16, to *Journal of American Statistical Association* for permission to reproduce Figures 8.2 – 8.5, 9.1, 9.2, 9.5, and 10.4 – 10.12, and to World Scientific Publishing Co, Inc. for permission to reproduce Figures 10.2 and 10.3.

Jianqing Fan's research was partially supported by the National Science Foundation and National Institutes of Health of the USA and the Research Grant Council of the Hong Kong Special Administrative Region. Qiwei Yao's work was partially supported by the Engineering and Physical Sciences Research Council and the Biotechnology and Biological Sciences Research Council of the UK. This book was written while Jianqing Fan was employed by the University of California at Los Angeles, the University of

North Carolina at Chapel Hill, and the Chinese University of Hong Kong, and while Qiwei Yao was employed by the University of Kent at Canterbury and the London School of Economics and Political Science. We acknowledge the generous support and inspiration of our colleagues. Last but not least, we would like to take this opportunity to express our gratitude to all our collaborators for their friendly and stimulating collaboration. Many of their ideas and efforts have been reflected in this book.

December 2002

Jianqing Fan
Qiwei Yao

Contents

1
Introduction

In attempts to understand the world around us, observations are frequently made sequentially over time. Values in the future depend, usually in a stochastic manner, on the observations available at present. Such dependence makes it worthwhile to predict the future from its past. Indeed, we will depict the underlying dynamics from which the observed data are generated and will therefore forecast and possibly control future events. This chapter introduces some examples of time series data and probability models for time series processes. It also gives a brief overview of the fundamental ideas that will be introduced in this book.

1.1 Examples of Time Series

Time series analysis deals with records that are collected over time. The time order of data is important. One distinguishing feature in time series is that the records are usually dependent. The background of time series applications is very diverse. Depending on different applications, data may be collected hourly, daily, weekly, monthly, or yearly, and so on. We use notation such as $\{X_t\}$ or $\{Y_t\}$ $(t = 1, \cdots, T)$ to denote a time series of length T. The unit of the time scale is usually implicit in the notation above . We begin by introducing a few real data sets that are often used in the literature to illustrate time series modeling and forecasting.

Example 1.1 *(Sunspot data)* The recording of sunspots dates back as far as 28 B.C., during the Western Han Dynasty in China (see, e.g., Needham

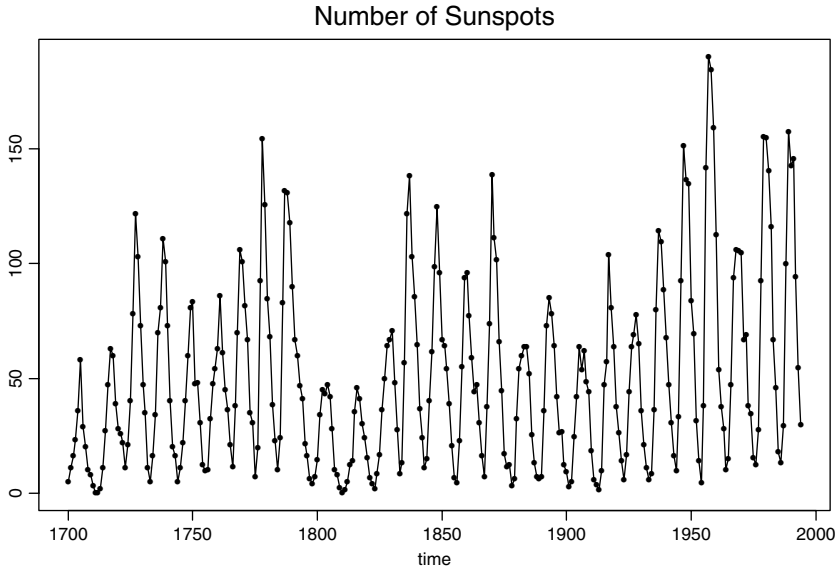


FIGURE 1.1. Annual means of Wolf's sunspot numbers from 1700 to 1994.

1959, p. 435 and Tong, 1990, p. 419). Dark spots on the surface of the Sun have consequences in the overall evolution of its magnetic oscillation. They also relate to the motion of the solar dynamo. The Zurich series of sunspot relative numbers is most commonly analyzed in the literature. Izenman (1983) attributed the origin and subsequent development of the Zurich series to Johann Rudolf Wolf (1816–1893). Let X_t be the annual means of Wolf's sunspot numbers, or simply the sunspot numbers in year $1770 + t$. The sunspot numbers from 1770 to 1994 are plotted against time in Figure 1.1. The horizontal axis is the index of time t, and the vertical axis represents the observed value X_t over time t. Such a plot is called a *time series plot*. It is a simple but useful device for analyzing time series data.

Example 1.2 *(Canadian lynx data)* This data set consists of the annual fur returns of lynx at auction in London by the Hudson Bay Company for the period 1821–1934, as listed by Elton and Nicolson (1942). It is a proxy of the annual numbers of the Canadian lynx trapped in the Mackenzie River district of northwest Canada and reflects to some extent the population size of the lynx in the Mackenzie River district. Hence, it helps us to study the population dynamics of the ecological system in that area. Indeed, if the proportion of the number of lynx being caught to the population size remains approximately constant, after logarithmic transforms, the differences between the observed data and the population sizes remain approximately constant. For further background information on this data

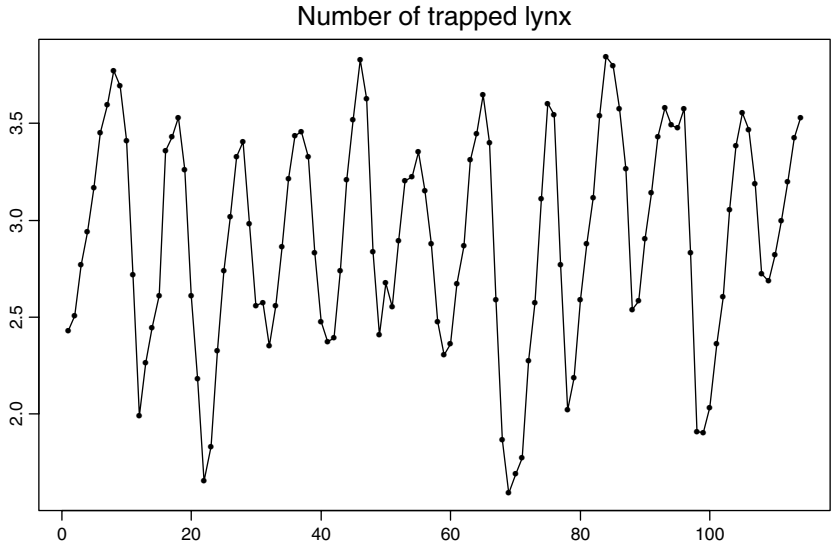


FIGURE 1.2. Time series for the number (on $\log_{10}$ scale) of lynx trapped in the MacKenzie River district over the period 1821–1934.

set, we refer to §7.2 of Tong (1990). Figure 1.2 depicts the time series plot of

$$X_t = \log_{10}(\text{number of lynx trapped in year } 1820 + t), \quad t = 1, 2, \cdots, 114.$$

The periodic fluctuation displayed in this time series has profoundly influenced ecological theory. The data set has been constantly used to examine such concepts as "balance-of-nature", predator and prey interaction, and food web dynamics, for example, see Stenseth et al. (1999) and the references therein.

Example 1.3 *(Interest rate data)* Short-term risk-free interest rates play a fundamental role in financial markets. They are directly related to consumer spending, corporate earnings, asset pricing, inflation, and the overall economy. They are used by financial institutions and individual investors to hedge the risks of portfolios. There is a vast amount of literature on interest rate dynamics, see, for example, Duffie (1996) and Hull (1997). This example concerns the yields of the three-month, six-month, and twelve-month Treasury bills from the secondary market rates (on Fridays). The secondary market rates are annualized using a 360-day year of bank interest and quoted on a discount basis. The data consist of 2,386 weekly observations from July 17, 1959 to September 24, 1999, and are presented in Figure 1.3. The data were previously analyzed by Andersen and Lund

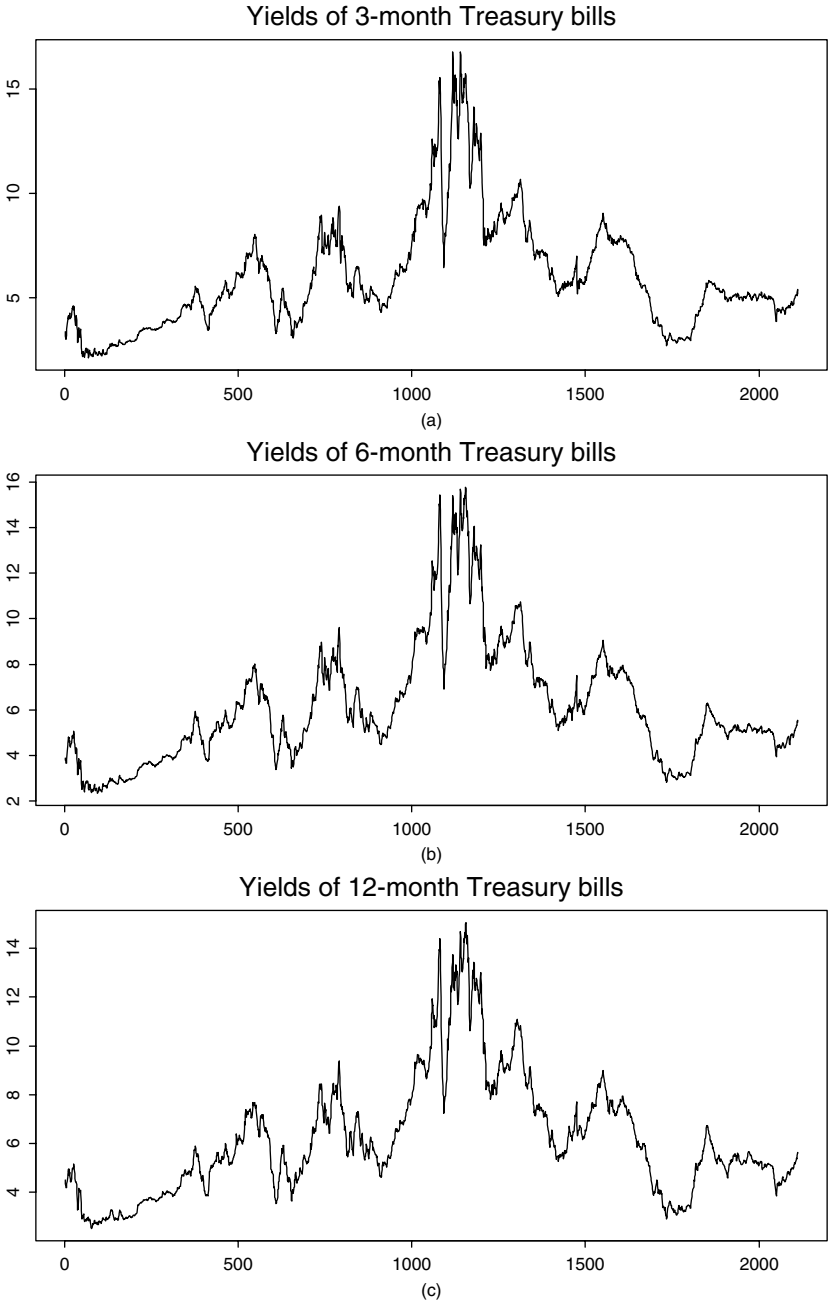


FIGURE 1.3. Yields of Treasury bills from July 17, 1959 to December 31, 1999 (source: Federal Reserve): (a) Yields of three-month Treasury bills; (b) yields of six-month Treasury bills; and (c) yields of twelve-month Treasury bills.

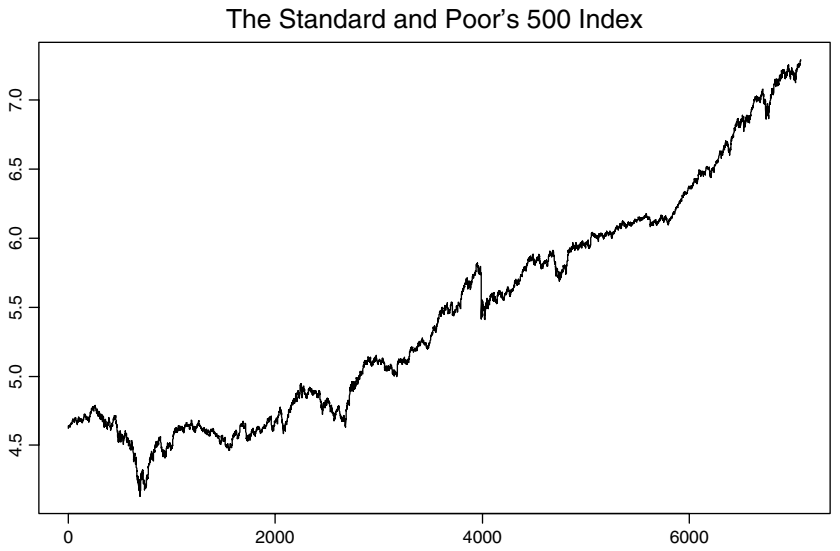


FIGURE 1.4. The Standard and Poor's 500 Index from January 3, 1972 to December 31, 1999 (on the natural logarithm scale).

(1997) and Gallant and Tauchen (1997), among others. This is a multivariate time series. As one can see in Figure 1.3, they exhibit similar structures and are highly correlated. Indeed, the correlation coefficients between the yields of three-month and six-month and three-month and twelve-month Treasury bills are 0.9966 and 0.9879, respectively. The correlation matrix among the three series is as follows:

$$\begin{pmatrix} 1.0000 & 0.9966 & 0.9879 \\ 0.9966 & 1.0000 & 0.9962 \\ 0.9879 & 0.9962 & 1.0000 \end{pmatrix}.$$

Example 1.4 *(The Standard and Poor's 500 Index)* The Standard and Poor's 500 index (S&P 500) is a value-weighted index based on the prices of the 500 stocks that account for approximately 70% of the total U.S. equity market capitalization. The selected companies tend to be the leading companies in leading industries within the U.S. economy. The index is a market capitalization-weighted index (shares outstanding multiplied by stock price)—the weighted average of the stock price of the 500 companies. In 1968, the S&P 500 became a component of the U.S. Department of Commerce's Index of Leading Economic Indicators, which are used to gauge the health of the U.S. economy. It serves as a benchmark of stock market performance against which the performance of many mutual funds is compared. It is also a useful financial instrument for hedging the risks

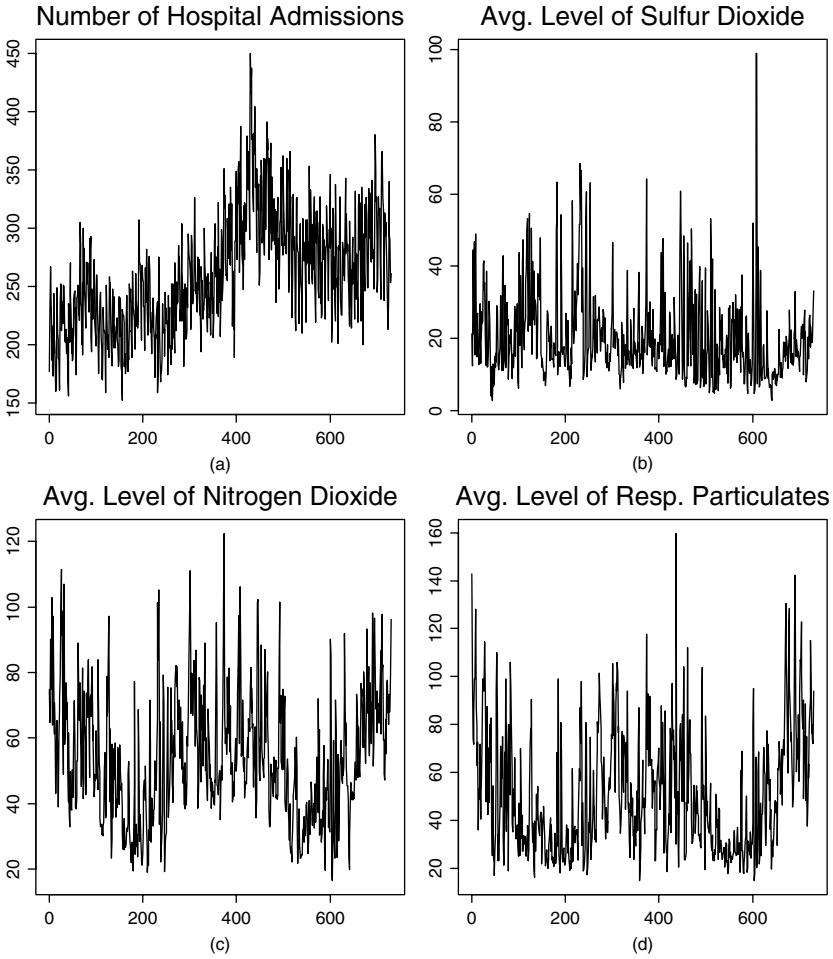


FIGURE 1.5. Time series plots for the environmental data collected in Hong Kong between January 1, 1994 and December 31, 1995: (a) number of hospital admissions for circulatory and respiratory problems; (b) the daily average level of sulfur dioxide; (c) the daily average level of nitrogen dioxide; and (d) the daily average level of respirable suspended particulates.

of market portfolios. The S&P 500 began in 1923 when the Standard and Poor's Company introduced a series of indices, which included 233 companies and covered 26 industries. The current S&P 500 Index was introduced in 1957. Presented in Figure 1.4 are the 7,076 observations of daily closing prices of the S&P 500 Index from January 3, 1972 to December 31, 1999. The logarithm transform has been applied so that the difference is proportional to the percentage of investment return.

Example 1.5 *(An environmental data set)* The environmental condition plays a role in public health. There are many factors that are related to the quality of air that may affect human circulatory and respiratory systems. The data set used here (Figure 1.5) comprises daily measurements of pollutants and other environmental factors in Hong Kong between January 1, 1994 and December 31, 1995 (courtesy of Professor T.S. Lau). We are interested in studying the association between the level of pollutants and other environmental factors and the number of total daily hospital admissions for circulatory and respiratory problems. Among pollutants that were measured are sulfur dioxide, nitrogen dioxide, and respirable suspended particulates (in $\mu g/m^3$). The correlation between the variables nitrogen dioxide and particulates is quite high (0.7820). However, the correlation between sulfur dioxide and nitrogen dioxide is not very high (0.4025). The correlation between sulfur dioxide and respirable particulates is even lower (0.2810). This example distinguishes itself from Example 1.3 in which the interest mainly focuses on the study of cause and effect.

Example 1.6 *(Signal processing—deceleration during car crashes)* Time series often appear in signal processing. As an example, we consider the signals from crashes of vehicles. Airbag deployment during a crash is accomplished by a microprocessor-based controller performing an algorithm on the digitized output of an accelerometer. The accelerometer is typically mounted in the passenger compartment of the vehicle. It experiences decelerations of varying magnitude as the vehicle structure collapses during a crash impact. The observed data in Figure 1.6 (courtesy of Mr. Jiyao Liu) are the time series of the acceleration (relative to the driver) of the vehicle, observed at 1.25 milliseconds per sample. During normal driving, the acceleration readings are very small. When vehicles are crashed or driven on very rough and bumpy roads, the readings are much higher, depending on the severity of the crashes. However, not all such crashes activate airbags. Federal standards define minimum requirements of crash conditions (speed and barrier types) under which an airbag should be deployed. Automobile manufacturers institute additional requirements for the airbag system. Based on empirical experiments using dummies, it is determined whether a crash needs to trigger an airbag, depending on the severity of injuries. Furthermore, for those deployment events, the experiments determine the latest time (required time) to trigger the airbag deployment device. Based on the current and recent readings, dynamical decisions are made on whether or not to deploy airbags.

These examples are, of course, only a few of the multitude of time series data existing in astronomy, biology, economics, finance, environmental studies, engineering, and other areas. More examples will be introduced later. The goal of this book is to highlight useful techniques that have been developed to draw inferences from data, and we focus mainly on non-

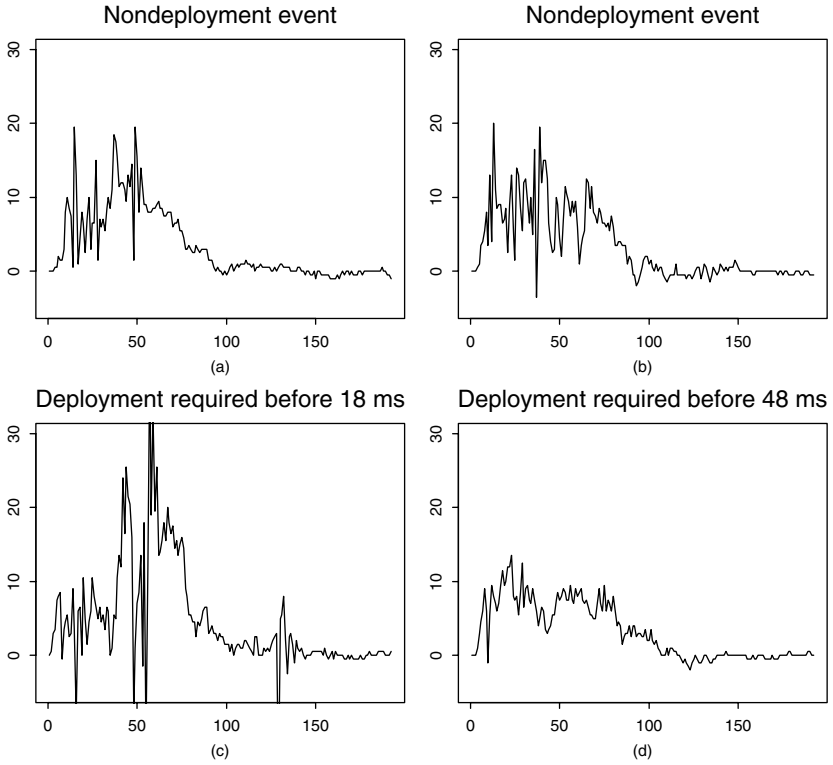


FIGURE 1.6. Time series plots for signals recorded during crashes of four vehicles. The acceleration (in a) is plotted against time (in milliseconds) after crashes. The top panels are the events that require no airbag deployments. The bottom panels are the events that need the airbag triggered before the required time.

parametric and semiparametric techniques that deal with nonlinear time series, although a compact and largely self-contained review of the most frequently used parametric nonlinear and linear models and techniques is also provided. We aim to accomplish a stochastic model that will represent the data well in the sense that the observed time series can be viewed as a realization from the stochastic process. The model should reflect the underlying dynamics and can be used for forecasting and controlling whenever appropriate. The observed time series are typically regarded as a realization from the stochastic process. An important endeavor is to unveil the unknown probability laws that describe well the underlying process. Once such a model has been established, it can be used for various purposes such as understanding and interpreting the mechanisms that generated the data, forecasting, and controlling the future.

1.2 Objectives of Time Series Analysis

The objectives of time series analysis are diverse, depending on the background of applications. Statisticians usually view a time series as a realization from a stochastic process. A fundamental task is to unveil the probability law that governs the observed time series. With such a probability law, we can *understand the underlying dynamics, forecast future events, and control future events via intervention.* Those are the three main objectives of time series analysis.

There are infinitely many stochastic processes that can generate the same observed data, as the number of observations is always finite. However, some of these processes are more plausible and admit better interpretation than others. Without further constraints on the underlying process, it is impossible to identify the process from a *finite* number of observations. A popular approach is to confine the probability law to a specified family and then to select a member in that family that is most plausible. The former is called modeling and the latter is called estimation, or more generally statistical inference. When the form of the probability laws in a family is specified except for some finite-dimensional defining parameters, such a model is referred to as a *parametric model.* When the defining parameters lie in a subset of an infinite dimensional space or the form of probability laws is not completely specified, such a model is often called a *nonparametric model*. We hasten to add that the boundary between parametric models and nonparametric models is not always clear. However, such a distinction helps us in choosing an appropriate estimation method. An analogy is that the boundary between "good" and "bad", "cold" and "hot", "healthy" and "unhealthy" is moot, but such a distinction is helpful to characterize the nature of the situation.

Time series analysis rests on proper statistical modeling. Some of the models will be given in §1.3 and §1.5, and some will be scattered throughout the book. In selecting a model, interpretability, simplicity, and feasibility play important roles. A selected model should reasonably reflect the physical law that governs the data. Everything else being equal, a simple model is usually preferable. The family of probability models should be reasonably large to include the underlying probability law that has generated the data but should not be so large that defining parameters can no longer be estimated with reasonably good accuracy. In choosing a probability model, one first extracts salient features from the observed data and then chooses an appropriate model that possesses such features. After estimating parameters or functions in the model, one verifies whether the model fits the data reasonably well and looks for further improvement whenever possible. Different purposes of the analysis may also dictate the use of different models. For example, a model that provides a good fitting and admits nice interpretation is not necessarily good for forecasting.

It is not our goal to exhaust all of the important aspects of time series analysis. Instead, we focus on some recent exciting developments in modeling and forecasting nonlinear time series, especially those with nonparametric and semiparametric techniques. We also provide a compact and comprehensible view of both linear time series models within the ARMA framework and some frequently used parametric nonlinear models.

1.3 Linear Time Series Models

The most popular class of linear time series models consists of autoregressive moving average (ARMA) models, including purely autoregressive (AR) and purely moving-average (MA) models as special cases. ARMA models are frequently used to model linear dynamic structures, to depict linear relationships among lagged variables, and to serve as vehicles for linear forecasting. A particularly useful class of models contains the so-called autoregressive integrated moving average (ARIMA) models, which includes *stationary* ARMA - processes as a subclass.

1.3.1 White Noise Processes

A stochastic process $\{X_t\}$ is called *white noise*, denoted as $\{X_t\} \sim \mathrm{WN}(0, \sigma^2)$, if

$$EX_t = 0, \quad \mathrm{Var}(X_t) = \sigma^2, \quad \text{and} \quad \mathrm{Cov}(X_i, X_j) = 0, \quad \text{for all } i \neq j.$$

White noise is defined by the properties of its first two moments only. It serves as a building block in defining more complex *linear* time series processes and reflects information that is not directly observable. For this reason, it is often called an *innovation process* in the time series literature. It is easy to see that a sequence of *independent and identically distributed* (i.i.d.) random variables with mean 0 and finite variance σ^2 is a special white noise process. We use the notation $\mathrm{IID}(0, \sigma^2)$ to denote such a sequence.

The probability behavior of a stochastic process is completely determined by all of its finite-dimensional distributions. When all of the finite-dimensional distributions are Gaussian (normal), the process is called a *Gaussian process*. Since uncorrelated normal random variables are also independent, a Gaussian white noise process is, in fact, a sequence of i.i.d. normal random variables.

1.3.2 AR Models

An *autoregressive model* of order $p \geq 1$ is defined as

$$X_t = b_1 X_{t-1} + \cdots + b_p X_{t-p} + \varepsilon_t, \tag{1.1}$$

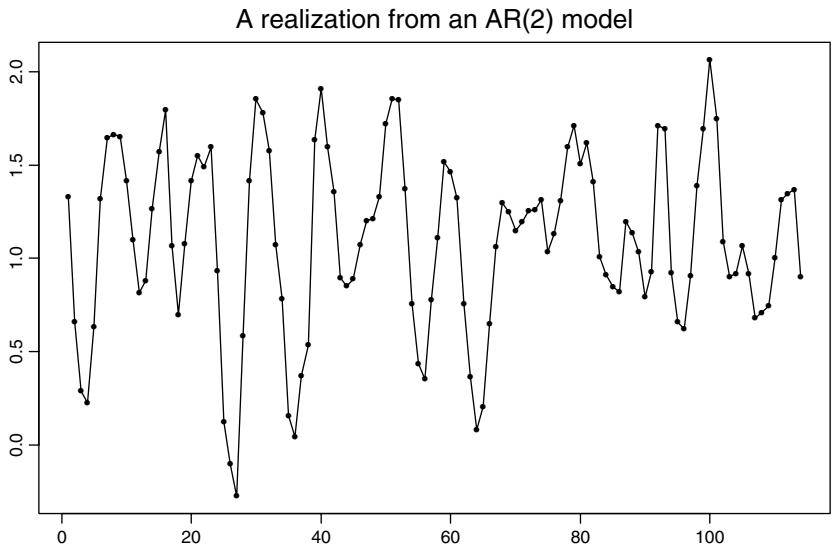


FIGURE 1.7. A length of 114 time series from the AR(2) model $X_t = 1.07 + 1.35X_{t-1} - 0.72X_{t-2} + \varepsilon_t$ with $\{\varepsilon_t\} \sim_{\text{i.i.d.}} N(0, 0.24^2)$. The parameters are taken from the AR(2) fit to the lynx data.

where $\{\varepsilon_t\} \sim \text{WN}(0, \sigma^2)$. We write $\{X_t\} \sim \text{AR}(p)$. The time series $\{X_t\}$ generated from this model is called the AR(p) process.

Model (1.1) represents the current state X_t through its immediate p past values $X_{t-1}, \cdots, X_{t-p}$ in a linear regression form. The model is easy to implement and therefore is arguably the most popular time series model in practice. Comparing it with the usual linear regression models, we exclude the intercept in model (1.1). This can be absorbed by either allowing ε_t to have a nonzero mean or deleting the mean from the observed data before the fitting. The latter is in fact common practice in time series analysis.

Model (1.1) explicitly specifies the relationship between the current value and its past values. This relationship also postulates the way to generate such an AR(p) process. Given a set of initial values $X_{-t_0-1}, \cdots, X_{-t_0-p}$, we can obtain X_t for $t \geq -t_0$ iteratively from (1.1) by generating $\{\varepsilon_t\}$ from, for example, the normal distribution $N(0, \sigma^2)$. Discarding the first $t_0 + 1$ values, we regard $\{X_t, t \geq 1\}$ as a realization of the process defined by (1.1). We choose $t_0 > 0$ sufficiently large to minimize the artifact due to the arbitrarily selected initial values. Figure 1.7 shows a realization of a time series of length 114 from an AR(2)-model.

We will also consider nonlinear autoregressive models in this book. We adopt the convention that the term AR-model always refers to a linear autoregressive model of the form (1.1) unless otherwise specified.

1.3.3 MA Models

A *moving average process* with order $q \geq 1$ is defined as

$$X_t = \varepsilon_t + a_1\varepsilon_{t-1} + \cdots + a_q\varepsilon_{t-q}, \tag{1.2}$$

where $\{\varepsilon_t\} \sim \text{WN}(0, \sigma^2)$. We write $\{X_t\} \sim \text{MA}(q)$.

An MA-model expresses a time series as a moving average of a white noise process. The correlation between X_t and X_{t-h} is due to the fact that they may depend on the same ε_{t-j}'s. Obviously, X_t and X_{t-h} are uncorrelated when $h > q$.

Because the white noise $\{\varepsilon_t\}$ is unobservable, the implementation of an MA-model is more difficult than that of an AR - model. The usefulness of MA models may be viewed from two aspects. First, they provide parsimonious representations for time series exhibiting MA-like correlation structure. As an illustration, we consider a simple MA(1)-model

$$X_t = \varepsilon_t - 0.9\varepsilon_{t-1}.$$

It can be proved that X_t admits the equivalent expression

$$X_t + \sum_{j=1}^{\infty}(0.9)^j X_{t-j} = \varepsilon_t.$$

(The infinite sum above converges in probability.) Note that $0.9^{20} = 0.1216$. Therefore, if we model a data set generated from this MA(1) process in terms of an AR(p) - model, then we need to use high orders such as $p > 20$. This will obscure the dynamic structure and will also render inaccurate estimation of the parameters in the AR(p) model.

The second advantage of MA models lies in their theoretical tractability. It is easy to see from the representation of (1.2) that the exploration of the first two moments of $\{X_t\}$ can be transformed to that of $\{\varepsilon_t\}$. The white noise $\{\varepsilon_t\}$ can be effectively regarded as an "i.i.d." sequence when we confine ourselves to the properties of the first two moments only. We will see that a routine technique in linear time series analysis is to represent a more general time series, including the AR-process, as a moving average process, typically of infinite order (see §2.1).

A moving average series is very easy to generate. One first generates a white noise process $\{\varepsilon_t\} \sim \text{WN}(0, \sigma^2)$ from, for example, normal distribution $N(0, \sigma^2)$ and then computes the observed series $\{X_t\}$ according to (1.2).

1.3.4 ARMA Models

The AR and MA classes can be further enlarged to model more complicated dynamics of time series. Combining AR and MA forms together yields the

popular autoregressive moving average (ARMA) model defined as

$$X_t = b_1 X_{t-1} + \cdots + b_p X_{t-p} + \varepsilon_t + a_1 \varepsilon_{t-1} + \cdots + a_q \varepsilon_{t-q}, \tag{1.3}$$

where $\{\varepsilon_t\} \sim \mathrm{WN}(0, \sigma^2)$, $p, q \geq 0$ are integers, and (p, q) is called the order of the model. We write $\{X_t\} \sim \mathrm{ARMA}(p, q)$. Using the backshift operator, the model can be written as

$$b(B)X_t = a(B)\varepsilon_t,$$

where B denotes the backshift operator, which is defined as

$$B^k X_t = X_{t-k}, \qquad k = \pm 1, \pm 2, \cdots,$$

and $a(\cdot)$ and $b(\cdot)$ are polynomials defined as

$$b(z) = 1 - b_1 z - \cdots - b_p z^p, \quad a(z) = 1 + a_1 z + \cdots + a_q z^q.$$

ARMA models are one of the most frequently used families of parametric models in time series analysis. This is due to their flexibility in approximating many stationary processes. However, there is no universal key that can open every door. The ARMA models do not approximate well the *nonlinear* phenomena described in §1.4 below. enddocument

1.3.5 ARIMA Models

A useful subclass of ARMA models consists of the so-called *stationary* models defined in §2.1. The stationarity reflects certain time-invariant properties of time series and is somehow a necessary condition for making a statistical inference. However, real time series data often exhibit time trend (such as slowly increasing) and/or cyclic features that are beyond the capacity of stationary ARMA models. The common practice is to preprocess the data to remove those *unstable* components. Taking the difference (more than once if necessary) is a convenient and effective way to detrend and deseasonalize. After removing time trends, we can model the new and remaining series by a stationary ARMA model. Because the original series is the integration of the differenced series, we call it an *autoregressive integrated moving average* (ARIMA) process.

A time series $\{Y_t\}$ is called an *autoregressive integrated moving average* (ARIMA) *process* with order p, d, and q, denoted as $\{Y_t\} \sim \mathrm{ARIMA}(p, d, q)$, if its d-order difference $X_t = (1-B)^d Y_t$ is a *stationary* $\mathrm{ARMA}(p, q)$ process, where $d \geq 1$ is an integer, namely, $b(B)(1 - B)^d Y_t = a(B)\varepsilon_t$.

It is easy to see that an $\mathrm{ARIMA}(p, d, q)$ model is a special $\mathrm{ARMA}(p+d, p)$ model that is typically nonstationary since $b(B)(1 - B)^d$ is a polynomial of order $p + d$. As an illustration, we have simulated a time series of length 200 from the $\mathrm{ARIMA}(1, 1, 1)$ model

$$(1 - 0.5B)(1 - B)Y_t = (1 + 0.3B)\varepsilon_t, \quad \{\varepsilon_t\} \sim_{\text{i.i.d}} N(0, 1). \tag{1.4}$$

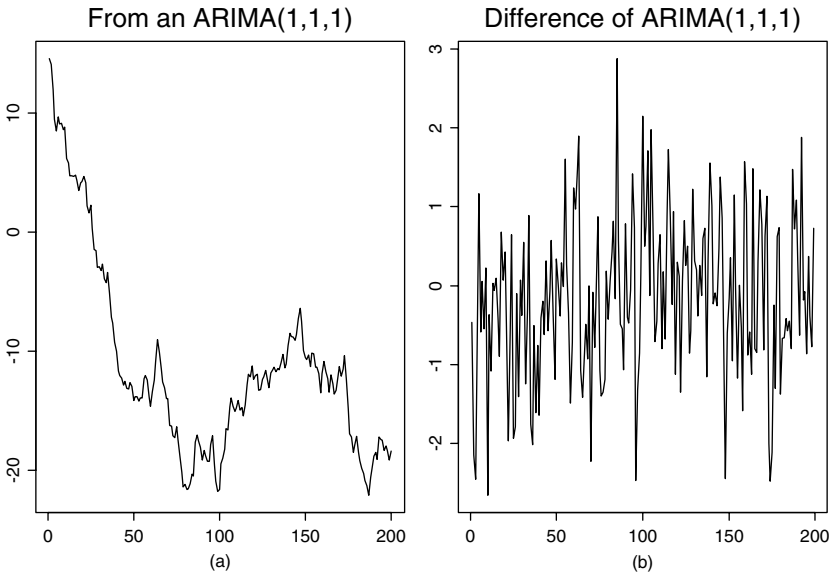


FIGURE 1.8. (a) A realization of a time series from ARIMA$(1, 1, 1)$ given by (1.4). The series exhibits an obvious time trend. (b) The first-order difference of the series.

The original time series is plotted in Figure 1.8(a). The time trend is clearly visible. Figure 1.8(b) presents the differenced series $\{Y_t - Y_{t-1}\}$. The decreasing time trend is now removed, and the new series appears stable.

1.4 What Is a Nonlinear Time Series?

From the pioneering work of Yule (1927) on AR modeling of the sunspot numbers to the work of Box and Jenkins (1970) that marked the maturity of ARMA modeling in terms of theory and methodology, linear Gaussian time series models flourished and dominated both theoretical explorations and practical applications. The last four decades have witnessed the continuous popularity of ARMA modeling, although the original ARMA framework has been enlarged to include long-range dependence with fractionally integrated ARMA (Granger and Joyeux 1980, Hosking 1981), multivariate VARMA and VARMAX models (Hannan and Deistler 1988), and random walk nonstationarity via cointegration (Engle and Granger 1987). It is safe to predict that in the future the ARMA model, including its variations, will continue to play an active role in analyzing time series data due to its simplicity, feasibility, and flexibility.

However, as early as the 1950s, P.A.P. Moran, in his classical paper (i.e., Moran 1953) on the modeling of the Canadian lynx data, hinted at a lim-

itation of linear models. He drew attention to the "curious feature" that the residuals for the sample points greater than the mean were significantly smaller than those for the sample points smaller than the mean. This, as we now know, can be well-explained in terms of the so-called "regime effect" at different stages of population fluctuation (§7.2 of Tong 1990; Stenseth et al.1999). Modeling the regime effect or other *nonstandard features* is beyond the scope of Gaussian time series models. (Note that a stationary purely nondeterministic Gaussian process is always linear; see Proposition 2.1.) Those nonstandard features, which we refer to as *nonlinear features* from now on, include, for example, nonnormality, asymmetric cycles, bimodality, nonlinear relationship between lagged variables, variation of prediction performance over the state-space, time irreversibility, sensitivity to initial conditions, and others. They have been well-observed in many real time series data, including some benchmark sets such as the sunspot, Canadian lynx, and others. See Tong (1990, 1995) and Tjøstheim (1994) for further discussion on this topic.

The endeavors to model the nonlinear features above can be divided into two categories—*implicit* and *explicit*. In the former case, we retain the general ARMA framework and choose the distribution of the white noise appropriately so that the resulting process exhibits a specified nonlinear feature (§1.5 of Tong 1990 and references therein). Although the form of the models is still linear, conditional expectations of the random variables given their lagged values, for example, may well be nonlinear. Thanks to the Wold decomposition theorem (p. 187 of Brockwell and Davis 1991), such a formal linear representation exists for any stationary (see §2.1 below) time series with no deterministic components. Although the modeling capacity of this approach is potentially large (Breidt and Davis 1992), it is difficult in general to identify the "correct" distribution function of the white noise from observed data. It is not surprising that the research in this direction has been surpassed by that on explicit models that typically express a random variable as a nonlinear function of its lagged values. We confine ourselves in this book to explicit nonlinear models.

Beyond the linear domain, there are infinitely many nonlinear forms to be explored. The early development of nonlinear time series analysis focused on various nonlinear parametric forms (Chapter 3 of Tong 1990; Tjøstheim 1994 and the references therein). The successful examples include, among others, the ARCH-modeling of fluctuating *volatility* of financial data (Engle 1982; Bollerslev 1986) and the threshold modeling of biological and economic data (§7.2 of Tong 1990; Tiao and Tsay 1994). On the other hand, recent developments in nonparametric regression techniques provide an alternative to model nonlinear time series (Tjøstheim 1994; Yao and Tong 1995 a, b; Härdle, Lütkepohl, and Chen 1997; Masry and Fan 1997). The immediate advantage of this is that little prior information on model structure is assumed, and it may offer useful insights for further parametric fitting. Furthermore, with increasing computing power in recent years, it

has become commonplace to access and to attempt to analyze time series data of unprecedented size and complexity. With these changes has come an increasing demand for nonparametric and semiparametric data-analytic tools that can identify the underlying structure and forecast the future according to a new standard of accuracy. The validity of a parametric model for a large real data set over a long time span is always questionable. All of these factors have led to a rapid development of computationally intensive methodologies (see, e.g., Chapter 8) that are designed to identify complicated data structures by exploring local lower-dimensional structures.

1.5 Nonlinear Time Series Models

In this section, we introduce some nonlinear time series models that we will use later on. This will give us some flavor for nonlinear time series models. For other parametric models, we refer to Chapter 3 of Tong (1990). We always assume $\{\varepsilon_t\} \sim \text{IID}(0, \sigma^2)$ instead of $\text{WN}(0, \sigma^2)$ when we introduce various nonlinear time series models in this section. Technically, this assumption may be weakened when we proceed with theoretical explorations later on. However, as indicated in a simple example below, a white noise process is no longer a pertinent building block for nonlinear models, as we have to look for measures beyond the second moments to characterize the nonlinear dependence structure.

1.5.1 A Simple Example

We begin with a simple example. We generate a time series of size 200 from the model

$$X_t = 2X_{t-1}/(1 + 0.8X_{t-1}^2) + \varepsilon_t, \tag{1.5}$$

where $\{\varepsilon_t\}$ is a sequence of independent random variables uniformly distributed on $[-1, 1]$. Figure 1.9(a) shows the 200 data points plotted against time. The scatterplot of X_t against X_{t-1} appears clearly nonlinear; see Figure 1.9(b). To examine the dependence structure, we compute the sample correlation coefficient $\rho(k)$ between the variables X_t and X_{t-k} for each k and plot it against k in Figure 1.9(c). It is clear from Figure 1.9(c) that $\rho(k)$ does not appear to die away at least up to lag 50, although the data are generated from a simple nonlinear autoregressive model with order 1. In fact, to reproduce the correlation structure depicted in Figure 1.9(c), we would have to fit an $\text{ARMA}(p, q)$ model with $p + q$ fairly large. This indicates that correlation coefficients are no longer appropriate measures for the dependence of nonlinear time series.

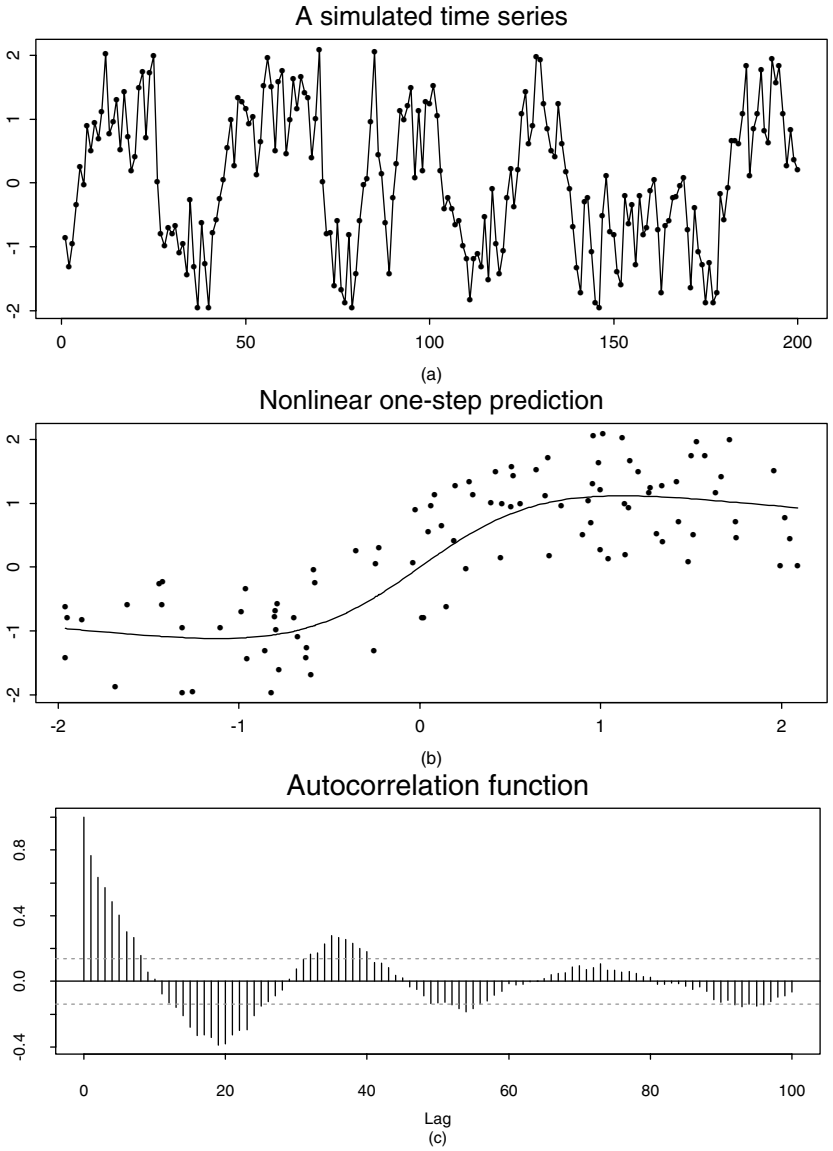


FIGURE 1.9. (a) A realization of a time series from model (1.5). (b) Scatter plot of the variable $\{X_{t-1}\}$ against $\{X_t\}$. (c) The sample autocorrelation function; the two dashed lines are approximate 95%-confidence limits around 0.

1.5.2 ARCH Models

An *autoregressive conditional heteroscedastic (ARCH) model* is defined as

$$X_t = \sigma_t \varepsilon_t \quad \text{and} \quad \sigma_t^2 = a_0 + b_1 X_{t-1}^2 + \cdots + b_q X_{t-q}^2, \tag{1.6}$$

where $a_0 \geq 0$, $b_j \geq 0$, and $\{\varepsilon_t\} \sim \text{IID}(0, 1)$.

ARCH models were introduced by Engle (1982) to model the varying (conditional) variance or volatility of time series. It is often found in economics and finance that the larger values of time series also lead to larger instability (i.e., larger variances), which is termed (*conditional*) *heteroscedasticity*. For example, it is easy to see from Figure 1.3 that the yields of Treasury bills exhibit the largest variation around the peaks. In fact, the conditional heteroscedasticity is also observed in the sunspot numbers in Figure 1.1 and the car crash signals in Figure 1.6.

Bollerslev (1986) introduced a *generalized autoregressive conditional heteroscedastic (GARCH) model* by replacing the second equation in (1.6) with

$$\sigma_t^2 = a_0 + a_1\sigma_{t-1}^2 + \cdots + a_p\sigma_{t-p}^2 + b_1X_{t-1}^2 + \cdots + b_qX_{t-q}^2, \tag{1.7}$$

where $a_j \geq 0$ and $b_j \geq 0$.

1.5.3 Threshold Models

The *threshold autoregressive* (TAR) model initiated by H. Tong assumes different linear forms in different regions of the state-space. The division of the state-space is usually dictated by one *threshold variable*, say, X_{t-d}, for some $d \geq 1$. The model is of the form

$$X_t = b_0^{(i)} + b_1^{(i)}\, X_{t-1} + \cdots + b_p^{(i)}\, X_{t-p} + \varepsilon_t^{(i)}, \quad \text{if} \quad X_{t-d} \in \Omega_i \tag{1.8}$$

for $i = 1, \cdots k$, where $\{\Omega_i\}$ forms a (nonoverlapping) partition of the real line, and $\{\varepsilon_t^{(i)}\} \sim \text{IID}(0, \sigma_i^2)$. We refer the reader to §5.2 and Tong (1990) for more detailed discussion on TAR models.

The simplest thresholding model is the two-regime (i.e. $k = 2$) TAR model with $\Omega_1 = \{X_{t-d} \leq \tau\}$, where the threshold τ is unknown. As an illustration, we simulated a time series from the two-regime TAR(2)-model

$$X_t = \begin{cases} 0.62 + 1.25X_{t-1} - 0.43X_{t-2} + \varepsilon_t, & X_{t-2} \leq 3.25 \\ 2.25 + 1.52X_{t-1} - 1.24X_{t-2} + \varepsilon_t', & X_{t-2} > 3.25, \end{cases} \tag{1.9}$$

where $\varepsilon_t \sim N(0, 0.2^2)$ and $\varepsilon_t' \sim N(0, 0.25^2)$. This model results from a two-regime TAR fit to the lynx data with a prescribed threshold variable X_{t-2}; see §7.2.6 of Tong (1990). Figure 1.10 depicts the simulated data and their associated sample autocorrelation function. Although the form of the model above is simple, it effectively captures many interesting features of the lynx dynamics; see §7.2 of Tong (1990).

1.5.4 Nonparametric Autoregressive Models

Nonlinear time series have infinite possible forms. We cannot entertain the thought that one particular family would fit all data well. A natural

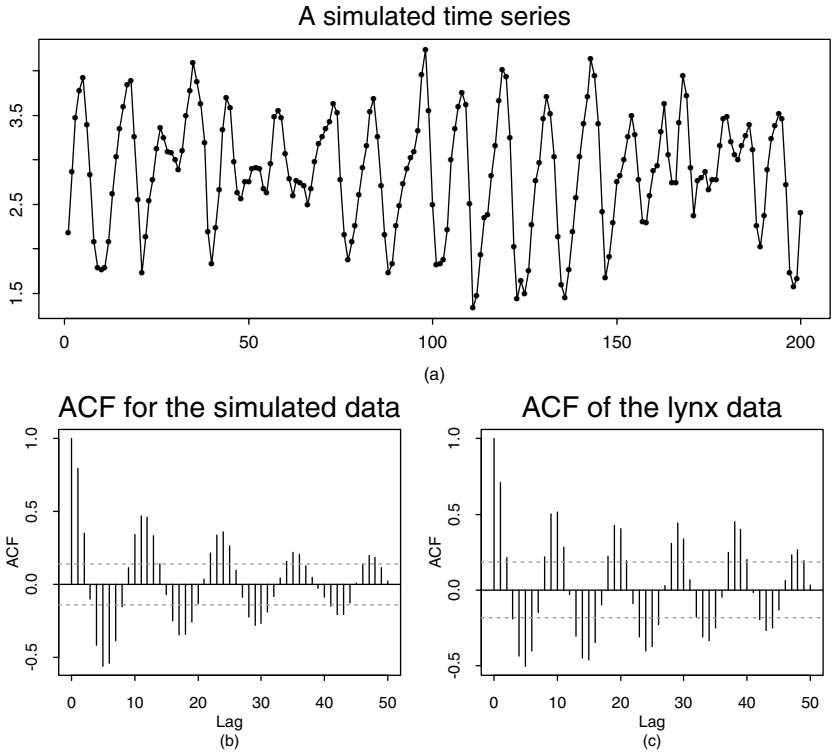


FIGURE 1.10. (a) A realization of a time series of length 200 from model (1.9). (b) and (c) The sample autocorrelation functions for the simulated data and the lynx data: two lines are approximate 95%-confidence limits around 0.

alternative is to adopt a nonparametric approach. In general, we can assume that

$$X_t = f(X_{t-1}, \ldots, X_{t-p}) + \sigma(X_{t-1}, \ldots, X_{t-p})\varepsilon_t, \tag{1.10}$$

where $f(\cdot)$ and $\sigma(\cdot)$ are unknown functions, and $\{\varepsilon_t\} \sim \text{IID}(0, 1)$. Instead of imposing concrete forms on functions f and σ, we only make some qualitative assumptions, such as that the functions f and σ are smooth. Model (1.10) is called a *nonparametric autoregressive conditional heteroscedastic* (NARCH) model or *nonparametric autoregressive* (NAR) model if $\sigma(\cdot)$ is a constant.

Obviously, model (1.10) is very general, making very few assumptions on how the data were generated. It allows heteroscedasticity. However, such a model is only useful when $p = 1$ or 2. For moderately large p, the functions in such a "saturated" nonparametric form are difficult to estimate unless the sample size is astronomically large. The difficulty is intrinsic and is often referred to as the "*curse of dimensionality*" in the nonparametric regression literature; see §7.1 of Fan and Gijbels (1996) for further discussion.

There are many useful models between parametric models and nonparametric models (1.10). For example, an extension of the thresholding model results is the so-called *function-coefficient autoregressive (FAR)* form

$$X_t = f_1(X_{t-d})X_1 + \cdots + f_p(X_{t-d})X_{t-p} + \sigma(X_{t-d})\varepsilon_t, \tag{1.11}$$

where $d > 0$ and $f_1(\cdot), \cdots, f_p(\cdot)$ are unknown coefficient functions. We write $\{X_t\} \sim \text{FAR}(p)$. Obviously, a FAR$(p)$ model is more flexible than a TAR(p) model. The coefficient functions in FAR models can be well-estimated with moderately large samples.

A powerful extension of (1.11) is to replace the "threshold" variable by a linear combination of the lagged variables of X_t with the coefficients determined by the data. This will enlarge the class of models substantially. Furthermore, it is of important practical relevance. For example, in modeling population dynamics it is of great biological interest to detect whether the population abundance or the population growth dominates the nonlinearity. We will discuss such a generalized FAR model in §8.4.

Another useful nonparametric model, which is a natural extension of the AR(p) model, is the following *additive autoregressive model*:

$$X_t = f_1(X_1) + \cdots + f_p(X_{t-p}) + \varepsilon_t. \tag{1.12}$$

Denote it by $\{X_t\} \sim \text{AAR}(p)$. Again, this model enhances the flexibility of AR models greatly. Because all of the unknown functions are one-dimensional, the difficulties associated with the curse of dimensionality can be substantially eased.

1.6 From Linear to Nonlinear Models

Nonlinear functions may well be approximated by either local linearization or global spline approximations. We illustrate these fundamental ideas below in terms of models (1.11) and (1.12). On the other hand, a goodness-of-fit test should be carried out to assess whether a nonparametric model is necessary in contrast to parametric models such as AR or TAR. The *generalized likelihood ratio* statistic provides a useful vehicle for this task. We briefly discuss the basic idea below. These topics will be systematically presented in Chapters 5–9.

1.6.1 Local Linear Modeling

Due to a lack of knowledge of the form of functions $f_1, \cdots, f_p$ in model (1.11), we can only use their qualitative properties: these functions are smooth and hence can be locally approximated by a constant or a linear function. To estimate the functions $f_1, \cdots, f_p$ at a given point x_0, for

simplicity of discussion we approximate them locally by a constant

$$f_j(x) \approx a_j, \qquad \text{for } x \in (x_0 - h, x_0 + h), \tag{1.13}$$

where h is the size of the neighborhood that the constant approximations hold. The local parameter a_j corresponds to $f_j(x_0)$. This leads to the local AR(p) model

$$X_t \approx a_1 X_{t-1} + \cdots + a_p X_{t-p} + \sigma(x_0)\varepsilon_t, \qquad X_{t-d} \in x_0 \pm h.$$

Using only the subset of data

$$\{(X_{t-p}, \cdots, X_t) : X_{t-d} \in x_0 \pm h, t = p+1, \cdots, T\},$$

we can fit an AR(p) model via the *least squares* method by minimizing

$$\sum_{t=p+1}^{T} \big(X_t - a_1 X_{t-1} - \cdots - a_p X_{t-p}\big)^2 I(|X_{t-d} - x_0| \le h), \tag{1.14}$$

where $I(\cdot)$ is the indicator function. The minimizer depends on the point x_0, which is denoted by $(\widehat{a}_1(x_0), \cdots, \widehat{a}_p(x_0))$. This yields an estimator of $f_1, \cdots, f_p$ at the point x_0:

$$\widehat{f}_1(x_0) = \widehat{a}_1(x_0), \cdots, \widehat{f}_p(x_0) = \widehat{a}_p(x_0).$$

Because x_0 runs over an interval $[a, b]$, we obtain estimated functions over $[a, b]$. To plot them, the estimated functions are frequently evaluated on a grid of points on $[a, b]$. Depending on the resolution needed, the number of grid points typically ranges from 100 to 400. Most of the graphs plotted in this book use 101 grid points.

The idea above can be improved in two ways. First, the local constant approximations in (1.13) can be improved by using the local linear approximations:

$$f_j(x) \approx a_j + b_j(x - x_0) \qquad \text{for } x \in x_0 \pm h. \tag{1.15}$$

The local parameter b_j corresponds to the local slope of f_j at the point x_0. This leads to the following approximate model:

$$\begin{aligned} X_t \quad \approx \quad & \{a_1 + b_1(X_{t-d} - x_0)\}X_{t-1} - \cdots - \{a_p + b_p(X_{t-d} - x_0)\}X_{t-p} \\ & + \sigma(x_0)\varepsilon_t \qquad \text{for } X_{t-d} \in x_0 \pm h. \end{aligned}$$

Second, the uniform weights in (1.14) can be replaced by the weighting scheme $K((X_{t-d} - x_0)/h)$ using a nonnegative unimodal function K. This leads to the minimization of the locally weighted squares

$$\begin{aligned} \sum_{t=p+1}^{T} & \big[X_t - \{a_1 + b_1(X_{t-d} - x_0)\}X_{t-1} - \cdots \\ & - \{a_p + b_p(X_{t-d} - x_0)\}X_{t-p}\big]^2 K\left(\frac{X_{t-d} - x_0}{h}\right), \end{aligned} \tag{1.16}$$

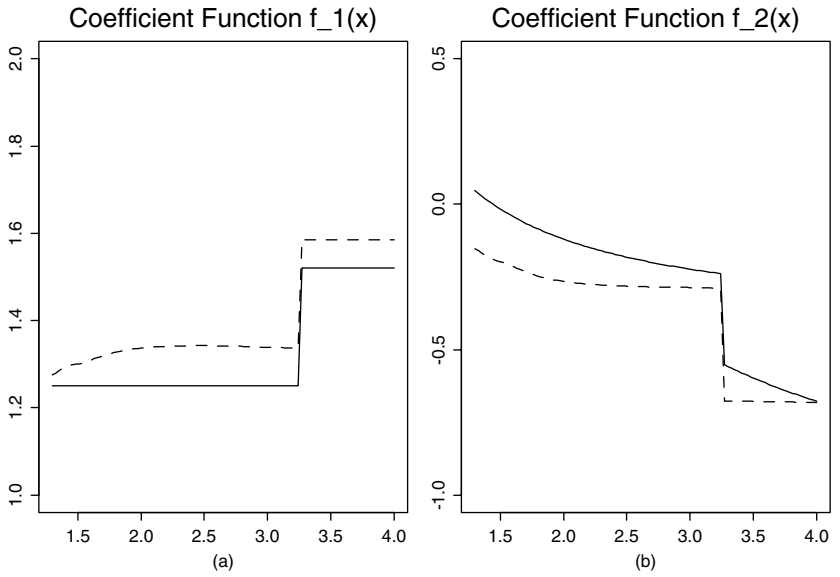


FIGURE 1.11. Fitted coefficient functions for the simulated data given in Figure 1.10(a). Dashed curves are estimated functions and solid curves are true functions.

which attributes the weight for each term according to the distance between X_{t-d} and x_0. When K has a support on $[-1, 1]$, the weighted regression (1.16) uses only the local data points in the neighborhood $X_{t-d} \in x_0 \pm h$. In general, weight functions need not have bounded supports, as long as they have thin tails. The weight function K is called the *kernel function* and the size of the local neighborhood h is called the *bandwidth* in the literature of nonparametric function estimation.

As an illustration, we fit a FAR(2)-model with $d = 2$ to the simulated data presented in Figure 1.10(a). Note that model (1.9) can be written as the FAR(2) model with

$$\widehat{f}_1^0(x) = \begin{cases} 1.25, & x \le 3.25, \\ 1.52, & x > 3.25, \end{cases}$$

and

$$\widehat{f}_2^0(x) = \begin{cases} 0.62/x - 0.43, & x \le 3.25, \\ 2.25/x - 1.24 & x > 3.25. \end{cases}$$

Figure 1.11 depicts the resulting estimates using the *Epanechnikov kernel*

$$K(t) = \frac{3}{4}(1 - t^2)_+$$

and bandwidth $h = 1.5$. Here, x_+ is the positive part of x, taking value x when $x \ge 0$ and 0 otherwise. The discontinuity of the underlying functions

around 3.25 can easily be detected by using nonparametric *change-point* techniques. Thus, we fit FAR(2) for two subsets of data with $X_{t-2} < 3.25$ and $X_{t-2} \geq 3.25$. Even though we do not assume any specific forms of the coefficient functions, the resulting estimates are quite close to the true functions. After inspecting the nonparametric fitting, one can now postulate a parameter model such as a TAR(2) to analyze the data further.

The combination of nonparametric and parametric methods has been proven fruitful. Nonparametric estimates attempt to find a good estimate among a large class of functions. This reduces the risk of modeling biases but at the expense of obtaining crude estimates. These estimates provide us guidance in choosing an appropriate family of parametric models. Parametric methods can be used to refine the fitting, which leads to easily interpretable estimators for the underlying dynamics. This is another reason why we introduce both parametric and nonparametric methods in this book.

1.6.2 Global Spline Approximation

Local linear modeling cannot be directly employed to fit the additive autoregressive model (1.12). To approximate unknown functions $f_1, \cdots, f_p$ locally at the point $(x_1, \cdots, x_p)$, we need to localize simultaneously in the variables $X_{t-1}, \cdots, X_{t-p}$. This yields a p-dimensional hypercube, which contains hardly any data points, unless the local neighborhood is very large. When the local neighborhood is too large to contain enough data points, the errors in the approximation will be large. This is the key problem underlying the curse of dimensionality. As we will see in §8.5, the local linear method can be applied to the AAM models by incorporating the *backfitting algorithm*.

To attenuate the problem, we approximate nonlinear functions by, for example, piecewise linear functions. The positions where piecewise linear functions can possibly change their slopes are called *knots*. Let $t_{j,1}, \cdots, t_{j,m_j}$ be the knots for approximating the unknown function f_j $(j = 1, \cdots, p)$. Then

$$f_j(x) \approx b_{j,0} + b_{j,1}x + b_{j,2}(x - t_{j,1})_+ + \cdots + b_{j,m_j+1}(x - t_{j,k})_+. \quad (1.17)$$

Denote by $f_j(x, \mathbf{b}_j)$ the piecewise linear function on the right-hand side of (1.17). When the knots are fine enough in the interval $[a, b]$, the resulting piecewise linear functions can approximate the smooth function f_j quite well. This is an example of polynomial *spline* modeling. After the spline approximation with the given knots, one can estimate parameters by the least squares method: minimize the following sum-of-square errors with respect to b:

$$\sum_{t=p}^{T} \{X_t - f_1(X_{t-1}, \mathbf{b}_1) - \cdots - f_p(X_{t-p}, \mathbf{b}_p)\}^2. \quad (1.18)$$

The estimated functions are simply

$$\widehat{f}_j(x) = f_j(x, \widehat{\mathbf{b}}_j).$$

The global spline modeling approach solves one large parametric problem (1.8). In contrast, the local modeling approaches solve many small parametric problems.

1.6.3 Goodness-of-Fit Tests

After fitting nonparametric models, we frequently ask whether a parametric model is adequate. Similarly, after fitting a parametric model, one asks whether the parametric model has excessive modeling biases. In the latter case, we can embed the parametric model into a larger family of models, such as nonparametric models. In both situations, we test a parametric hypothesis against a nonparametric alternative.

As an example, we consider different models for the simulated data presented in Figure 1.10(a). To test whether an AR(2) model

$$H_0 : X_t = b_1 X_{t-1} + b_2 X_{t-2} + \varepsilon_t$$

fits the data, we employ the FAR(2) model

$$X_t = f_1(X_{t-2})X_{t-1} + f_2(X_{t-2})X_{t-2} + \varepsilon_t$$

as the alternative hypothesis. One can now compute the residual sum of squares (RSS) under both null and alternative models; namely,

$$\mathrm{RSS}_0 = \sum_{t=3}^{T} \Big\{ X_t - \widehat{b}_1 X_{t-1} - \widehat{b}_2 X_{t-2} \Big\}^2 \tag{1.19}$$

and

$$\mathrm{RSS}_1 = \sum_{t=3}^{T} \Big\{ X_t - \widehat{f}_1(X_{t-2})X_{t-1} - \widehat{f}_2(X_{t-2})X_{t-2} \Big\}^2. \tag{1.20}$$

For these particular data, $\mathrm{RSS}_0 = 13.82$ and $\mathrm{RSS}_1 = 10.60$. Now, define the *generalized likelihood ratio* (GLR) statistic as

$$\mathrm{GLR} = \frac{T-2}{2} \log(\mathrm{RSS}_0/\mathrm{RSS}_1) = 26.25.$$

The null distribution of the GLR statistic can be found either by the generalized likelihood theory developed in Fan, Zhang, and Zhang (2001) or via a *bootstrap* method. By applying the bootstrap approach, we obtain the p-value 0% based on 1,000 bootstrap replications. The method will be detailed in §9.3. This provides strong evidence against the null hypothesis.

The result is again consistent with the fact that the data were simulated from (1.9).

Consider now whether the TAR(2) model (1.9) adequately fits the data. Again, we use the nonparametric FAR(2) model above as the alternative hypothesis. In this case, the RSS under the null model is given by

$$\mathrm{RSS}_0 = \sum_{t=3}^{T} \Big\{ X_t - \widehat{f}_1^0(X_{t-2}) X_{t-1} - \widehat{f}_2^0(X_{t-2}) X_{t-2} \Big\}^2,$$

where $\widehat{f}_1^0$ and $\widehat{f}_2^0$ are simply the (estimated) coefficient functions in model (1.9). For these particular data, $\mathrm{RSS}_0 = 9.260$ and $\mathrm{RSS}_1 = 10.60$. This is possible because the fitting methods under the null and alternative hypotheses are not the same. This leads to the generalized likelihood ratio statistic

$$\mathrm{GLR} = \frac{T-2}{2} \log(\mathrm{RSS}_0/\mathrm{RSS}_1) = -13.41.$$

This means that the null model (1.9) fits even better than the nonparametric alternative model. This is not surprising because the data were drawn from (1.9).

By applying the bootstrap approach, we obtain the p-value 0.523 based on 1,000 bootstrap replications. This provides little evidence against H_0. In other words, both the TAR(2) and FAR(2) models provide indistinguishable fitting to these simulated data.

1.7 Further Reading

This book does not intend to exhaust all aspects of nonlinear time series analysis. Instead, we mainly focus on various commonly-used nonparametric and parametric techniques. The techniques for modeling linear time series within the ARMA framework are presented in a compact and comprehensible manner for the sake of comparison and complement.

There are many excellent books on time series written at different levels for different purposes. Almost all of them are on parametric models. Box and Jenkins (1970) is the first book systematically dealing with time series analysis within the ARMA framework. Many examples used in the book are now classic. It is a good guide into the practical aspects. Brockwell and Davis (1996) is a modern textbook with a comprehensive and user-friendly package ITSM. It also includes state-space models and multivariate models. Shumway and Stoffer (2000) provide an ideal text for graduate courses for nonmathematics/statistics students. It has wide coverage, with numerous interesting real data examples. Chatfield (1996) and Cryer (1986) offer alternatives for more compact courses. Brockwell and Davis (1991) discuss the theory of time series in depth, which should be ideal for serious theorists. Their work contains a lucid discussion of continuous-time AR models

and analysis of heavy-tailed time series. The book by Anderson (1971) has been written specifically to appeal to mathematical statisticians trained in the more classical parts of statistics. Taniguchi and Kakizawa (2000) present a wealth of modern asymptotic theory of inference for various time series models, including (linear) ARMA processes, long-memory processes, nonlinear time series, continuous-time processes, nonergodic processes, diffusion processes, and others. Brillinger (1981) and Priestley (1981) offer wide coverage as well as in-depth accounts of the spectral analysis of time series. Early monographs on nonlinear time series include Priestley (1988). Tong (1990) provides comprehensive coverage of parametric nonlinear time series analysis. It also initiates the link between nonlinear time series and nonlinear dynamic systems (chaos). The state-space modeling of time series data, making judicious use of the celebrated Kalman filters and smoothers, is well-presented by Harvey (1990), Kitagawa and Gersch (1996), and more recently by Durbin, and Koopman (2001). West and Harrison (1989) deal with dynamic models based on Bayesian methods. Golyandina, Nekrutkin and Zhigljavsky (2001) summarize the techniques based on *singular-spectrum analysis*. Analysis of multivariate time series is systematically presented by Hannan (1970), Lütkepohl (1993), and Reinsel (1997). Diggle (1990) specializes in applications to biological and medical time series. Tsay (2002) assembles the techniques for analyzing financial time series. Akaike and Kitagawa (1999) and Xie (1993) collect some interesting case studies for practical problems in diverse fields. Monographs on more specific topics include those by Gouriéroux (1997) on ARCH/GARCH models, Subba-Rao and Gabr (1984) and Terdik (1999) on bilinear models, Tong (1983) on threshold models, Nicholls and Quinn (1982) on random coefficient autoregressive models, and Beran (1995) on long-memory processes. For nonparametric approaches, Györfi, Härdle, Sarda, and Vieu (1989) and Bosq (1998) are concerned with the asymptotic theory of kernel estimation for time series data and provide useful techniques (such as mixing and exponential inequalities) for further exploration of theoretical properties of nonparametric time series models.

Nonparametric modeling is a very large and dynamic field. It keeps expanding due to the demand for nonlinear approaches and the availability of modern computing power. Indeed, most parametric models and techniques have their nonparametric counterparts. Many excellent books have been written in this very dynamic area. There are three basic approaches to nonparametric modeling: *kernel-local polynomial*, *spline*, and *orthogonal series* methods. For kernel density estimation and regression, see Devroye and Györfi (1985), Silverman (1986), Müller (1988), Härdle (1990), Scott (1992), Wand and Jones (1995), and Simonoff (1996). Local polynomial methods are extensively discussed by Fan and Gijbels (1996). Work on spline modeling has been published by Wahba (1990), Green and Silverman (1994), and Eubank (1999). Hastie and Tibshirani (1990) outline nonparametric additive modeling. For orthogonal series methods such as *Fourier*

series and *wavelets*, see Ogden (1997), Efromovich (1999), and Vidakovic (1999), among others. Nonparametric hypothesis testing can be found in the books by Bowman and Azzalini (1997) and Hart (1997). Applications of nonparametric methods to functional data can be found in the work of Ramsay and Silverman (1997).

1.8 Software Implementations

Part of the computation in this book was carried out using the software package S-Plus. A large part of linear modeling was performed using the ITSM package of Brockwell and Davis (1996), estimation for GARCH models was carried out in S+GARCH. The procedures that are computationally more demanding were implemented in the C language. Most of the one-dimensional smoothing described in this book can easily be implemented by using existing software. Local linear smoothing with automatic bandwidth selection was programmed in C-code. Varying-coefficient models (1.11) can be implemented using any package with a least-squares function by introducing weights. Most of the graphics in this book are plotted using S-Plus.

It is our hope that readers will be stimulated to use the methods described in this book for their own applications and research. Our aim is to provide information in sufficient detail so that readers can produce their own implementations. This will be a valuable exercise for students and readers who are new to the area. To assist this endeavor, we have placed all of the data sets and codes used in this book on the following web site.

http://www.stat.unc.edu/faculty/fan/nls.html

This page intentionally blank

2
Characteristics of Time Series

Statistical inference is about learning something that is unknown from the known. Time series analysis is no exception in this aspect. In order to achieve this, it is necessary to assume that at least some features of the underlying probability law are sustained over a time period of interest. This leads to the assumptions of different types of stationarity, depending on the nature of the problem at hand. The dependence in the data marks the fundamental difference between time series analysis and classical statistical analysis. Different measures are employed to describe the dependence at different levels to suit various practical needs. In this chapter, we introduce the most commonly used definitions for stationarity and dependence measures. We also make comments on when those definitions and measures are most relevant in practice.

2.1 Stationarity

2.1.1 Definition

We introduce two types of stationarity, namely (weak) stationarity and strict stationarity, in this section. Both of them require that time series exhibit certain time-invariant behavior.

Definition 2.1 *A time series* $\{X_t, t = 0, \pm 1, \pm 2, \cdots\}$ *is stationary if* $E(X_t^2) < \infty$ *for each* t*, and*

(i) $E(X_t)$ *is a constant, independent of* t*, and*

(ii) $\mathrm{Cov}(X_t, X_{t+k})$ *is independent of* t *for each* k.

Definition 2.2 *A time series* $\{X_t, t = 0, \pm 1, \pm 2, \cdots\}$ *is strictly stationary if* $(X_1, \cdots, X_n)$ *and* $(X_{1+k}, \cdots, X_{n+k})$ *have the same joint distributions for any integer* $n \geq 1$ *and any integer* k.

The stationarity, which is often referred to as the weak stationarity in textbooks, assumes that only the first two moments of time series are time-invariant and is generally weaker than the strict stationarity, provided that the process has finite second moments. Weak stationarity is primarily used for linear time series, such as ARMA processes, where we are mainly concerned with the linear relationships among variables at different times. In fact, the assumption of stationarity suffices for most linear time series analysis, such as in spectral analysis. In contrast, we have to look beyond the first two moments if our focus is on nonlinear relationships. This explains why strict stationarity is often required in the context of nonlinear time series analysis.

2.1.2 Stationary ARMA *Processes*

It is obvious that a white noise process $\mathrm{WN}(0, \sigma^2)$ is stationary but not necessarily strictly stationary; see §1.3.1. In view of the discussion above, it is natural to use $\mathrm{WN}(0, \sigma^2)$ as a building block for general linear time series. For Gaussian time series, we need only to focus on the properties of the first two moments, too. A stationary Gaussian process is also strictly stationary.

First, we consider moving average models. It is easy to see from (1.2) that any $\mathrm{MA}(q)$ process with finite q is stationary. Let us consider an $\mathrm{MA}(\infty)$ model defined as

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} \quad \text{for all } t, \tag{2.1}$$

where $\{\varepsilon_t\} \sim \mathrm{WN}(0, \sigma^2)$ and $\sum_{j=0}^{\infty} |a_j| < \infty$. Therefore

$$E|X_t| \leq E|\varepsilon_1| \sum_{j=0}^{\infty} |a_j| < \infty.$$

This implies that the infinite sum on the right-hand side of (2.1) converges in probability, and also in mean of order 1 as well as order 2, as the condition $\sum_j |a_j| < \infty$ implies $\sum_j a_j^2 < \infty$. (Under the additional condition that $\{\varepsilon_t\}$ is independent, the infinite sum also converges almost surely due to the Loève theorem; see Corollary 3, p. 117 of Chow and Teicher 1997.) Furthermore, $EX_t = 0$, and

$$\mathrm{Cov}(X_t, X_{t+k}) = \sum_{j,l=0}^{\infty} a_j a_k E(\varepsilon_{t-j}, \varepsilon_{t+k-l}) = \sigma^2 \sum_{j=0}^{\infty} a_j a_{j+|k|}, \tag{2.2}$$

which is independent of t. Hence, such an MA(∞) model also defines a stationary process. Obviously, it $\{\varepsilon_t\}$ are i.i.d and $E|\varepsilon_t| < \infty$. The process $\{X_t\}$ defined by (2.1) is strictly stationary.

For a general ARMA (p, q) model defined in (1.3), we may express the process in a compact form in terms of backshift operator B as:

$$b(B)X_t = a(B)\varepsilon_t \quad \text{for all } t, \tag{2.3}$$

where B is the *backshift operator* defined as

$$B^k X_t = X_{t-k} \quad k = 0, \pm 1, \pm 2, \cdots,$$

and $b(\cdot)$ and $a(\cdot)$ are polynomials given by

$$b(z) = 1 - b_1 z - \cdots - b_p z^p, \quad a(z) = 1 + a_1 z + \cdots + a_q z^q. \tag{2.4}$$

Remark 2.1 For ARMA models as defined in (2.3), we always assume that polynomials $b(\cdot)$ and $a(\cdot)$ do not have common factors. Otherwise, a process so defined is effectively equivalent to the process with orders smaller than (p, q) after removing those common factors.

Theorem 2.1 *The process $\{X_t,\, t = 0, \pm 1, \pm 2, \cdots\}$ given by (2.3) is stationary if $b(z) \neq 0$ for all complex numbers z such that $|z| \leq 1$.*

Proof. Let $z_1, \cdots, z_p$ be the roots of $b(z) = 0$. Then $|z_j| > 1$ and $b(z) = \prod_{1 \leq j \leq p}(1 - z/z_j)$. It follows from some simple Taylor expansions that for any $|z| \leq 1$,

$$b(z)^{-1} = \prod_{j=1}^{p} \left(1 - z/z_j\right)^{-1} = \prod_{j=1}^{p} \left\{ \sum_{k=0}^{\infty} (z/z_j)^k \right\} \equiv \sum_{j=0}^{\infty} c_j z^j.$$

Note that

$$\sum_{j=0}^{\infty} |c_j| \leq \prod_{j=1}^{p} \left\{ \sum_{k=0}^{\infty} 1/|z_j|^k \right\} = \prod_{j=1}^{p} \left(1 - 1/|z_j|\right)^{-1} < \infty.$$

Now, write $c(z) = \sum_{j \geq 0} c_j z^j$. Then $c(z)\, b(z) \equiv 1$. Therefore

$$X_t = c(B)b(B)X_t = c(B)a(B)\varepsilon_t = d(B)\varepsilon_t, \tag{2.5}$$

where $d(z) = c(z)a(z) = \sum_{j=0}^{\infty} d_j z^j$ with $\sum_{j=0}^{\infty} |d_j| < \infty$. This indicates that $\{X_t\}$ is effectively an MA(∞) process defined as in (2.1) and is therefore stationary. ∎

Another important concept in time series is causality.

Definition 2.3 *A time series $\{X_t\}$ is causal if for all t*

$$X_t = \sum_{j=0}^{\infty} d_j \varepsilon_{t-j}, \quad \sum_{j=0}^{\infty} |d_j| < \infty,$$

where $\{\varepsilon_t\} \sim \mathrm{WN}(0, \sigma^2)$.

Causality means that X_t is caused by the white noise process (from the past) up to time t and is effectively an MA(∞) process. For the ARMA process defined in (2.3), causality is equivalent to the condition that $b(z) \neq 0$ for all $|z| \leq 1$ (p. 83 of Brockwell and Davis 1996), and therefore it implies stationarity, but the converse is not true. In fact, the model (2.3) admits a unique stationary solution if and only if $b(z) \neq 0$ for all complex numbers z on the unit cycle $|z| = 1$ (p. 82 of Brockwell and Davis 1996). However, it may be shown that under the condition $b(z) \neq 0$ for all $|z| \geq 1$, the stationary solution of (2.3) with $q = 0$, for example, is of the form

$$X_t = \sum_{j=0}^{\infty} d_j \varepsilon_{t+j},$$

which is not causal. One may argue whether such a process should be called a time series since X_t depends on 'future' noise ε_{t+j} for $j \geq 1$. However, any stationary noncausal ARMA process can be represented as a causal ARMA process (with the same orders) in terms of a newly defined white noise, and both processes have identical first two moments (Proposition 3.5.1 of Brockwell and Davis 1991). Therefore, we lose no generality by restricting our attention to the subset of causal processes in the class of stationary ARMA processes. But we should be aware of the fact that even if the original process is defined in terms of an i.i.d. process $\{\varepsilon_t\}$, the white noise in the new representation is no longer i.i.d.

In Theorem 2.1, the condition that the process $\{X_t\}$ is doubly infinite in time is important. For example, the process defined by

$$X_t = 0.5X_{t-1} + \varepsilon_t$$

for $t = 0, \pm 1, \pm 2, \cdots$ is stationary (also strictly stationary), where $\{\varepsilon_t\} \sim_{\text{i.i.d.}} N(0, 1)$. However, the process defined by the equation above for $t = 1, 2, \cdots$ only and initiated at $X_0 \sim U(0, 1)$ is no longer stationary since $EX_t = 0.5^{t+1}$ for all $t \geq 0$. The process $\{X_t, t = 1, 2, \cdots\}$ will be (strictly) stationary if and only if we start the process with $X_0 \sim N(0, 1/0.75)$, which is in fact the stationary distribution of the Markov chain defined by the AR(1) model above (see Theorem 2.2 below).

2.1.3 Stationary Gaussian Processes

A time series $\{X_t\}$ is said to be *Gaussian* if all its finite-dimensional distributions are normal. If $\{\varepsilon_t\} \sim_{\text{i.i.d.}} N(0, \sigma^2)$ and $b(z) \neq 0$ for all $|z| \leq 1$, $\{X_t\}$ defined by (2.3) is a stationary Gaussian process (and therefore also strictly stationary). On the other hand, it follows from the Wold decomposition theorem (p. 187 of Brockwell and Davis 1991) that for any stationary Gaussian process $\{X_t\}$ with mean 0, it holds that

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} + V_t, \tag{2.6}$$

where $\sum_j a_j^2 < \infty$, $\{\varepsilon_t\}$ and $\{V_t\}$ are two independent normal processes, $\{\varepsilon_t\} \sim_{\text{i.i.d.}} N(0, \sigma^2)$, and $\{V_t\}$ is deterministic in the sense that, for any t, V_t is entirely determined by its lagged values $V_{t-1}, V_{t-2}, \cdots$ (i.e., V_t is $\mathcal{F}_{t-1}$-measurable, where $\mathcal{F}_t$ is the σ-algebra generated by $\{V_{t-k},\ k = 1, 2, \cdots\}$). When $V_t \equiv 0$, we call $\{X_t\}$ *purely nondeterministic*. Therefore, a purely nondeterministic stationary Gaussian process is always linear in the sense that it can be written as an MA(∞) process with normal white noise.

A particularly simple case is a q-dependent stationary Gaussian process in the sense that X_t and X_{t+k} are independent for all $k > q$. This implies that $V_t \equiv 0$ and $a_j = 0$ for all $j > q$ in (2.6). Therefore $\{X_t\} \sim \text{MA}(q)$.

On the other hand, if, given $\{X_{t-1}, \cdots, X_{t-p}\}$, X_t is independent of $\{X_{t-k},\ k > p\}$, it is easy to see that

$$\varepsilon_t \equiv X_t - E(X_t | X_{t-1}, \cdots, X_{t-p})$$

is independent of $\{X_{t-k},\ k \geq 1\}$ since $\text{Cov}(\varepsilon_t, X_{t-k}) = 0$ for $k \geq 1$. Therefore ε_t is also independent of $\{\varepsilon_{t-k},\ k \geq 1\}$ since ε_{t-k} is a function of $\{X_{t-k}, X_{t-k-1}, \cdots\}$ only. Hence $\{\varepsilon_t\} \sim_{\text{i.i.d.}} N(0, \sigma^2)$. Due to the normality, $E(X_t | X_{t-1}, \cdots, X_{t-p})$ is a linear function of $X_{t-1}, \cdots, X_{t-p}$:

$$E(X_t | X_{t-1}, \cdots, X_{t-p}) = b_1 X_{t-1} + \cdots + b_p X_{t-p}$$

for some coefficients $b_1, \cdots, b_p$. This implies that $\{X_t\} \sim \text{AR}(p)$ since

$$\begin{aligned} X_t &= E(X_t | X_{t-1}, \cdots, X_{t-p}) + \varepsilon_t \\ &= b_1 X_{t-1} + \cdots + b_p X_{t-p} + \varepsilon_t. \end{aligned}$$

The results above are summarized as follows.

Proposition 2.1 *Let $\{X_t\}$ be a stationary Gaussian time series.*
(i) $\{X_t\} \sim \text{MA}(\infty)$ if it is a purely nondeterministic process.
(ii) $\{X_t\} \sim \text{MA}(q)$ if, it is a q-dependent process.
(iii) $\{X_t\} \sim \text{AR}(p)$ if, given $\{X_{t-1}, \cdots, X_{t-p}\}$, X_t is independent of $\{X_{t-k},\ k > p\}$.

*2.1.4 Ergodic Nonlinear Models**

It is relatively straightforward to check stationarity in linear time series models. However, it is by no means easy to check whether a time series defined by a nonlinear model is strictly stationary. It remains open to prove (or disprove) that some simple nonlinear models (such as quadratic functions) may generate a strictly stationary process. The common practice is to represent a time series as a (usually vector-valued) Markov chain and to establish that the Markov chain is ergodic. Stationarity follows from the fact that an ergodic Markov chain is stationary.

First, we give a brief introduction of Markov chains. A vector-valued stochastic process $\{\mathbf{X}_t\}$ is called a Markov chain if it fulfills the Markovian

property that the conditional distribution of $\mathbf{X}_{t+1}$ given $\{\mathbf{X}_t, \mathbf{X}_{t-1}, \cdots\}$ depends on $\mathbf{X}_t$ only for all t. The Markovian property requires that, given the present and the past, the future depends on the present only. The conditional distribution of $\mathbf{X}_{t+1}$ given $\mathbf{X}_t$ is called the transition distribution at time t. If the transition distribution is independent of time t, the Markov chain is called homogeneous. In this book, we consider homogeneous Markov chains only. Therefore, we simply call them Markov chains.

We consider a general form of nonlinear AR model

$$X_t = f(X_{t-1}, \cdots, X_{t-p}) + \varepsilon_t, \tag{2.7}$$

where $\{\varepsilon_t\}$ is a sequence of i.i.d. random variables. When a time series model is defined with an i.i.d. noise, we always assume implicitly that ε_t is independent of $\{X_{t-k},\, k \geq 1\}$. This condition is natural when the process $\{X_t\}$ is generated from the model in the natural time order.

Define

$$\mathbf{X}_t = (X_t, \cdots, X_{t-p+1})^\tau, \quad \boldsymbol{\varepsilon}_t = (\varepsilon_t, 0, \cdots, 0)^\tau,$$

and for $\mathbf{x} = (x_1, \cdots, x_p)^\tau \in R^p$,

$$\mathbf{f}(\mathbf{x}) = (f(x_1), x_1, \cdots, x_{p-1})^\tau.$$

Then, it follows from (2.7) that $\{\mathbf{X}_t\}$ is a Markov chain defined as

$$\mathbf{X}_t = \mathbf{f}(\mathbf{X}_{t-1}) + \boldsymbol{\varepsilon}_t. \tag{2.8}$$

Let $G(\cdot)$ be the distribution function of $\boldsymbol{\varepsilon}_t$, and let $F_n(\cdot|\mathbf{x})$ be the conditional distribution of $\mathbf{X}_n$ given $\mathbf{X}_0 = \mathbf{x}$. It follows from (2.8) that, for $n \geq 2$,

$$F_n(\mathbf{y}|\mathbf{x}) = \int G\{\mathbf{y} - \mathbf{f}(\mathbf{u})\} F_{n-1}(d\mathbf{u}|\mathbf{x}) \tag{2.9}$$

and $F_1(\mathbf{y}|\mathbf{x}) = G\{\mathbf{y} - \mathbf{f}(\mathbf{x})\}$, which is in fact the transition distribution of the Markov chain.

The (Harris) ergodicity introduced below is defined in terms of the convergence of probability distributions in the norm of total variation. For two probability distributions P_1 and P_2 defined on the same sample space, the *total variation* of $(P_1 - P_2)$ is defined as

$$||P_1 - P_2|| = \sup \sum_j |P_1(A_j) - P_2(A_j)|,$$

where the supremum is taken over all measurable partitions $\{A_j\}$ of the sample space. If P_i has probability density function p_i $(i = 1, 2)$, it may be shown that

$$||P_1 - P_2|| = \int |p_1(\mathbf{x}) - p_2(\mathbf{x})| d\mathbf{x}.$$

Definition 2.4 *If there exists a distribution F and a constant $\rho \in (0, 1]$ such that*

$$\rho^{-n}||F_n(\cdot|\mathbf{x}) - F(\cdot)|| \to 0 \quad \text{for any } \mathbf{x}, \tag{2.10}$$

the Markov model (2.8) is called ergodic *when $\rho = 1$ and* geometrically ergodic *when $\rho < 1$. F is called the* stationary distribution. *In the expression above, $||\cdot||$ denotes the total variation.*

Obviously, geometric ergodicity implies ergodicity. The ergodicity of a Markov chain depends entirely on its transition distribution. If the transition distribution is strictly positive and regular, the process is "weakly" ergodic in the sense that $F_n \to F$ at all continuous points of F, and furthermore the process initiated from F is strictly stationary; see, for example, §8.7 of Feller (1971). Unfortunately, the processes as defined in (2.8) do not fulfill those conditions. The Harris ergodicity adopted here strengthens the convergence in terms of the total variation, which effectively ensures the required stationarity. For further discussion on Harris ergodicity, we refer the reader to Chan (1990a, 1993b).

Theorem 2.2 *Suppose that the Markov model (2.8) is ergodic. Then there exists a stationary (p-dimensional) distribution F such that the time series $\{X_t,\, t = 1, 2, \cdots\}$ defined by (2.7) and initiated at $(X_0, X_{-1}, \cdots, X_{-p+1})^\tau \sim F$ is strictly stationary.*

Proof. Let $n \to \infty$ on both sides of (2.9), it then follows from the fact that the total variation of $(F_n - F)$ converges to 0 that

$$F(\cdot) = \int G(\cdot - \mathbf{f}(\mathbf{y}))F(d\mathbf{y}).$$

Note that $G(\cdot - \mathbf{f}(\mathbf{y}))$ is the conditional distribution of $\mathbf{X}_{t+1}$ given $\mathbf{X}_t = \mathbf{y}$. The equation above indicates that if $\mathbf{X}_t \sim F$, then $\mathbf{X}_{t+1} \sim F$. Therefore, all of the random variables $\{\mathbf{X}_{t+k}$ for $k \geq 2\}$ share the same marginal distribution F. The Markovian property implies that the joint distribution of $(\mathbf{X}_t, \mathbf{X}_{t+1}, \cdots, \mathbf{X}_{t+k})$ is completely determined by the transition density and the marginal distribution of $\mathbf{X}_t$. Hence, the Markov chain $\{\mathbf{X}_t,\, t = 1, 2, \cdots\}$ defined by (2.8) and initiated at $\mathbf{X}_0 \sim F$ is strictly stationary. By considering the first component of $\mathbf{X}_t$'s only, we obtain the theorem. ∎

For ergodic Markov chains, the law of large numbers always holds, irrespective of initial distributions. The theorem below was proved in Chan (1993a).

Theorem 2.3 *Suppose that model (2.8) is ergodic with stationary distribution F. For $\{X_t,\, t = 1, 2, \cdots\}$ defined by (2.7) with any initial variables $(X_0, X_{-1}, \cdots, X_{-p+1})$,*

$$\frac{1}{n}\sum_{t=1}^{n} g(X_t) \xrightarrow{a.s.} E_F\{g(X_t)\}$$

provided $E_F|g(X_t)| < \infty$.

It is by no means easy to derive a general condition under that (2.8) is (Harris) ergodic. We list a few simple criteria below that are often used to check whether a nonlinear model is ergodic. For $\mathbf{x} = (x_1, \cdots, x_p)^\tau$, we write $||\mathbf{x}|| = (x_1^2 + \cdots + x_p^2)^{1/2}$.

Theorem 2.4 *Suppose that in model (2.7)* $f(\cdot)$ *is measurable and* ε_t *has a positive density function and* $E\varepsilon_t = 0$. *The induced Markov model (2.8) is geometrically ergodic if one of the following three conditions holds.*

(i) f *is bounded on bounded sets and*

$$\lim_{||\mathbf{x}||\to\infty} |f(\mathbf{x}) - (b_1 x_1 + \cdots + b_p x_p)|/||\mathbf{x}|| = 0, \tag{2.11}$$

where $b_1, \cdots, b_p$ *are some constants satisfying the condition that* $1 - b_1 z - \cdots - b_p z^p \neq 0$ *for all complex* z *satisfying* $|z| \leq 1$.

(ii) There exist constants $\lambda \in (0, 1)$ *and* c *for which*

$$|f(\mathbf{x})| \leq \lambda \max\{|x_1|, \cdots, |x_p|\} + c.$$

(iii) There exist constants $\rho \in (0, 1)$, c, *and* $a_i \geq 0$, *and* $a_1 + \cdots + a_p = 1$ *such that*

$$|f(\mathbf{x})| \leq \rho(a_1|x_1| + \cdots + a_p|x_p|) + c. \tag{2.12}$$

In the above, (i) and (ii) were obtained by An and Huang (1996), and (iii) was proved by Bhattacharya and Lee (1995). An and Chen (1997) extended the condition (2.12) to the case where $\rho = 1$. An and Huang (1996) also derived a condition for the case where $f(\cdot)$ in (2.7) is continuous. To simplify statements, we call model (2.7) (geometrically) ergodic if the induced Markov model (2.8) is (geometrically) ergodic.

Example 2.1 *(TAR-model)* Consider the TAR model with k regimes [see also (1.8)],

$$\begin{aligned} X_t &= \sum_{i=1}^{k} \{b_{i0} + b_{i1} X_{t-1} + \cdots + b_{i,p_i} X_{t-p_i}\} \\ &\quad \times I(r_{i-1} \leq X_{t-d} < r_i) + \varepsilon_t, \end{aligned} \tag{2.13}$$

where $\{\varepsilon_t\}$ satisfies the condition in Theorem 2.4, $-\infty = r_0 < r_1 < \cdots < r_k = \infty$, and $d, p_1, \cdots, p_k$ are some positive integers. It follows from Theorems 2.4 and 2.2 that there exists a strictly stationary solution $\{X_t\}$ from the model above if either $\max_{1\leq i\leq k} \sum_{j=1}^{p_i} |b_{ij}| < 1$, which entails condition (ii) of Theorem 2.4, or $\max_{1\leq i\leq k} |b_{ij}| < a_j$ and $a_1 + \cdots + a_p = 1$, where $p = \max_{1\leq i\leq k} p_i$ which implies condition (iii). ∎

The conditions imposed above are unfortunately more stringent than necessary. It remains as a challenge to derive the necessary and sufficient condition for model (2.13) to be ergodic. Chan and Tong (1985) proved that the simple TAR model

$$X_t = \begin{cases} \alpha + X_{t-1} + \varepsilon_t, & X_{t-1} \le 0, \\ \beta + X_{t-1} + \varepsilon_t, & X_{t-1} > 0 \end{cases} \tag{2.14}$$

is ergodic if and only if $\alpha < 0 < \beta$. Note that, for this model, condition (2.12) holds with $\rho = 1$.

From Theorems 2.2 and 2.4, we may derive some sufficient conditions for AAR model (1.12) or FAR model (1.11) admitting a strictly stationary solution. In general, (2.7) admits a strictly stationary solution if $f(\mathbf{x})$ grows slower than $\|\mathbf{x}\|$ as $\|\mathbf{x}\| \to \infty$, since (2.11) holds with all $b_i = 0$. On the other hand, if $f(\cdot)$ in (2.7) is a polynomial function with order greater than 1, which is unbounded, the condition that ε_t be compactly supported is necessary for ergodicity when $p = 1$ (Chan and Tong 1994). Finally, we note that a causal AR(p) model with i.i.d. white noise is geometrically ergodic, which can be seen easily from Theorem 2.4(i).

2.1.5 Stationary ARCH *Processes*

We introduce a general form of ARCH(∞) model

$$Y_t = \rho_t \xi_t, \quad \rho_t = a + \sum_{j=1}^{\infty} b_j Y_{t-j}, \tag{2.15}$$

where $\{\xi_t\}$ is a sequence of nonnegative i.i.d. random variables with $E\xi_t = 1$ and $a \ge 0$ and $b_j \ge 0$. Obviously, the model above includes the standard ARCH model (1.6) as a special case if we let $Y_t = X_t^2$ (the standard model allows observing the sign of $\{X_t\}$, which, however, contains no information on the variance of the series). It also contains the GARCH model (1.7) if the coefficients $\{a_i\}$ in (1.7) fulfill certain conditions; for example, all $a_i \ge 0$ and $\sum_{i \ge 1} a_i < 1$. In this case, (1.7) admits the expression $\sigma_t^2 = a_0 + \sum_{j=1}^{\infty} c_j X_{t-j}^2$ with $a_0 \ge 0$ and $c_j \ge 0$.

Theorem 2.5 *(i) Under the condition* $\sum_{j=1}^{\infty} b_j < 1$, *model (2.15) has a unique strictly stationary solution* $\{Y_t,\, t = 0, \pm 1, \pm 2, \cdots\}$ *for which*

$$EY_t = a \Big/ \left\{ 1 - \sum_{j=1}^{\infty} b_j \right\}.$$

Furthermore the unique solution is $Y_t \equiv 0$ *for all* t *if* $a = 0$.

(ii) Suppose that $E\xi_t^2 < \infty$ *and*

$$\max\{1, (E\xi_t^2)^{1/2}\} \sum_{j=1}^{\infty} b_j < 1. \tag{2.16}$$

Then, model (2.15) has a unique strictly stationary solution $\{Y_t\}$ *with* $EY_t^2 < \infty$.

The theorem above was established by Giraitis, Kokoszka, and Leipus (2000) through a Volterra expansion of Y_t in terms of $\{\xi_{t-k},\, k \geq 0\}$. We reproduce its proof for part (i) in §2.7.1 below. Note that an ARCH process is not a linear process in the sense that it cannot be expressed as an MA(∞) process defined in terms of an i.i.d. white noise. In fact, the Volterra expansion contains multiplicative terms of ξ_j, which makes the theoretical investigation more complicated. But, on the other hand, the fact that all of the quantities involved (such as c, b_j, and ξ_j) are nonnegative does bring appreciable convenience to the analytic derivations; see §2.7.1.

It follows from Theorem 2.5 that the ARCH model (1.6) admits a strictly stationary solution if $\sum_{j=1}^{q} b_j < 1$. Giraitis, Kokoszka, and Leipus (2000) also established the central limit theorem below. A stochastic process $W(t)$ is called a *Brownian motion* or *Wiener process* if it is a Gaussian process starting at zero with mean zero and covariance function $EW(t)W(\tau) = \min(t, \tau)$.

Theorem 2.6 *Suppose that* $\{Y_t\}$ *is the strictly stationary process defined by (2.15) for which condition (2.16) holds. Define for* $t \in [0, 1]$

$$S(t) = \frac{1}{\sqrt{n}\sigma} \sum_{j=1}^{[nt]} (Y_j - EY_j),$$

where $\sigma^2 = \sum_{t=-\infty}^{\infty} \mathrm{Cov}(Y_t, Y_0) < \infty$. *Then, for any* $k \geq 1$ *and* $0 \leq t_1 < \cdots < t_k \leq 1$,

$$\{S(t_1), \cdots, S(t_k)\} \xrightarrow{D} \{W(t_1), \cdots, W(t_k)\},$$

where $\{W(t),\, 0 \leq t \leq 1\}$ *is the standard Wiener process with mean 0 and covariance* $E\{W(t)W(s)\} = \min(t, s)$.

The theorem above indicates that the stochastic process $\{S(t),\ 0 < t \leq 1\}$ converges in distribution to the Brownian motion $\{W(t),\ 0 < t \leq 1\}$.

2.2 Autocorrelation

For linear time series $\{X_t\}$, we are interested in the linear relationships among the random variables at different time points t. The autocorrelation coefficient measures the linear dependence between X_{t+k} and X_t. The partial autocorrelation coefficient is the correlation between the residual of X_{t+k} and that of X_t after regressing both linearly on $X_{t+1}, \cdots, X_{t+k-1}$.

2.2.1 Autocovariance and Autocorrelation

For stationary time series $\{X_t\}$, it follows from Definition 2.1 that

$$\mathrm{Cov}(X_{t+k}, X_t) = \mathrm{Cov}(X_k, X_0) \quad \text{for any } k.$$

That means that the correlation between X_t and X_s depends on the absolute time difference $|t-s|$ only.

Definition 2.5 *Let $\{X_t\}$ be a stationary time series. The* autocovariance function (ACVF) *of $\{X_t\}$ is*

$$\gamma(k) = \mathrm{Cov}(X_{t+k}, X_t), \quad k = 0, \pm 1, \pm 2, \cdots .$$

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

$$\rho(k) = \gamma(k)/\gamma(0) = \mathrm{Corr}(X_{t+k}, X_t), \quad k = 0, \pm 1, \pm 2, \cdots .$$

From the definition above, we can see that both $\gamma(\cdot)$ and $\rho(\cdot)$ are even functions, namely

$$\gamma(-k) = \gamma(k) \quad \text{and} \quad \rho(-k) = \rho(k).$$

The theorem below presents the necessary and sufficient condition for a function to be an ACVF of a stationary time series.

Theorem 2.7 (Characterization of ACVF) *A real-valued function $\gamma(\cdot)$ defined on the integers is the ACVF of a stationary time series if and only if it is even and nonnegative definite in the sense that*

$$\sum_{i,j=1}^{n} a_i a_j \gamma(i-j) \geq 0 \tag{2.17}$$

for any integer $n \geq 1$ and arbitrary real numbers $a_1, \cdots, a_n$.

The necessity of the theorem above follows from the fact that the sum in (2.17) is the variance of random variable $\sum_{j=1}^{n} a_j X_j$. Hence, the sum is nonnegative. The proof of the sufficiency uses Kolmogorov's existence theorem; see p. 27 of Brockwell and Davis (1991).

We now examine the properties of ACVFs and ACFs for stationary ARMA processes. First, it is obvious that a process is a white noise if and only if $\rho(k) = 0$ for all $k \neq 0$.

For MA(∞) process

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j},$$

where $\{\varepsilon_t\} \sim \text{WN}(0, \sigma^2)$, $a_0 = 1$ and $\sum_{j=0}^{\infty} |a_j| < \infty$. It is easy to see from (2.2) that

$$\gamma(k) = \sigma^2 \sum_{j=0}^{\infty} a_j a_{j+|k|}, \qquad \rho(k) = \frac{\sum_{j=0}^{\infty} a_j a_{j+|k|}}{\sum_{j=0}^{\infty} a_j^2}. \tag{2.18}$$

Therefore, if $\{X_t\} \sim \text{MA}(q)$ (i.e., $a_j = 0$ for all $j > q$), the formulas above reduce to

$$\gamma(k) = \sigma^2 \sum_{j=0}^{q-|k|} a_j a_{j+|k|} \text{ and } \rho(k) = \frac{\sum_{j=0}^{q-|k|} a_j a_{j+|k|}}{\sum_{j=0}^{q} a_j^2} \quad \text{for } |k| \le q, \tag{2.19}$$

and $\gamma(k)$ and $\rho(k)$ are 0 for all $|k| > q$. We say that the ACF of an MA(q) process cuts off at q. This is a benchmark property for MA processes.

For causal ARMA(p, q) process

$$X_t = b_1 X_{t-1} + \cdots + b_p X_{t-p} + \varepsilon_t + a_1 \varepsilon_{t-1} + \cdots + a_q \varepsilon_{t-q},$$

where $\{\varepsilon_t\} \sim \text{WN}(0, \sigma^2)$, we may calculate the ACVF and ACF through their MA(∞) representation

$$X_t = \sum_{j=0}^{\infty} d_j \varepsilon_{t-j},$$

where d_j's are the coefficients of polynomial $b(z)^{-1}a(z)$ (see (2.5)), which may be evaluated recursively as:

$$\begin{aligned} d_0 &= a_0\,(= 1), \\ d_1 &= a_1 + d_0 b_1, \\ d_2 &= a_2 + d_0 b_2 + d_1 b_1, \\ &\cdots\cdots \end{aligned}$$

and, in general,

$$d_k = a_k + \sum_{j=0}^{k-1} d_j b_{k-j}, \quad k \ge 1. \tag{2.20}$$

We assume that $a_j = 0$ for $j > q$ and $b_i = 0$ for $i > p$ in the recursion above. Now, both the ACVF and ACF are given as in (2.18), with a_j replaced by d_j. It is easy to see from (2.18) and (2.20) that, for causal ARMA processes, the ACF depends on the coefficients $\{b_j\}$ and $\{a_j\}$ only and is independent of the variance of white noise σ^2. (Of course, the ACVF depends on σ^2.) This indicates that the autocorrelation of an ARMA process is dictated by the coefficients in the model and is independent of the amount of white noise injected into the model.

The approach above does not lead to a simple closed-form solution. It provides little information on the asymptotic behavior of $\rho(k)$ as $k \to \infty$, which reflects the "memory" of the ARMA(p, q) process. To investigate this asymptotic behavior, we calculate the covariance of both sides of an ARMA(p, q) model with X_{t-k} for $k > q$. By (2.3),

$$\text{Cov}\{b(B)X_t, X_{t-k}\} = \text{Cov}\{a(B)\varepsilon_t, X_{t-k}\} = 0$$

since $\varepsilon_t, \cdots, \varepsilon_{t-q}$ are independent of X_{t-k} for $k > q$. This leads to the Yule–Walker equation

$$\gamma(k) - b_1\gamma(k-1) - \cdots - b_p\gamma(k-p) = 0, \quad k > q. \tag{2.21}$$

It is easy to see that the general solution of this equation is

$$\gamma(k) = \alpha_1 z_1^{-k} + \cdots + \alpha_p z_p^{-k}, \tag{2.22}$$

where $\alpha_1, \cdots, \alpha_p$ are arbitrary constants and $z_1, \cdots, z_p$ are the p roots of equation

$$1 - b_1 z - \cdots - b_p z^p = 0.$$

The condition for causality implies $|z_j| > 1$ for all j. Therefore, it follows from (2.22) that $\gamma(k)$ converges to 0 at an exponential rate as $|k| \to \infty$.

We summarize the findings above in the proposition below.

Proposition 2.2 *(i) For causal ARMA processes, $\rho(k) \to 0$ at an exponential rate as $|k| \to \infty$.*

(ii) For MA(q) processes, $\rho(k) = 0$ for all $|k| > q$.

2.2.2 Estimation of ACVF *and* ACF

Given a set of observations $\{X_1, \cdots, X_T\}$ from a stationary time series, we may estimate the ACVF by the sample autocovariance function defined as

$$\widehat{\gamma}(k) = \frac{1}{T}\sum_{t=1}^{T-k}(X_t - \bar{X}_T)(X_{t+k} - \bar{X}_T), \quad k = 0, 1, \cdots, T-1, \tag{2.23}$$

where $\bar{X}_T = \frac{1}{T}\sum_{t=1}^{T} X_t$. This also leads to estimating the ACF by the sample autocorrelation function

$$\widehat{\rho}(k) = \widehat{\gamma}(k)/\widehat{\gamma}(0), \quad k = 0, 1, \cdots, T-1.$$

It is impossible to estimate $\gamma(k)$ and $\rho(k)$ for $k \geq T$ from observed data $X_1, \cdots, X_T$. Even for k slightly smaller than T, the estimates $\widehat{\gamma}(k)$ and $\widehat{\rho}(k)$ are unreliable since there are only a few pairs (X_t, X_{t+k}) available. A useful guide proposed by Box and Jenkins (1970 p. 30) requires $T \geq 50$ and $k \leq T/4$.

A natural alternative for estimating ACVF and ACF is to replace the divisor T in (2.23) by $T-k$. The resulting estimators could be substantially different for large k and are less biased. However, it is fair to say that $\widehat{\rho}(k)$ and $\widehat{\gamma}(k)$ defined with the divisor T are more preferable in practice, as reflected by the fact that they have been implemented as default estimators in most time series packages. This may be due to the fact that in time series analysis we are more interested in estimating the ACF as a whole function rather than $\rho(k)$ for some fixed k. It may be shown that $\{\widehat{\gamma}(k)\}$, and therefore also $\{\widehat{\rho}(k)\}$, is a nonnegative-definite function if we define $\widehat{\gamma}(-k) = \widehat{\gamma}(k)$ for $k \geq 1$ and $\widehat{\gamma}(k) = 0$ for $|k| \geq T$. This property may be lost if we replace the divisor T by $T-k$. Furthermore, when k becomes large relative to T, the smaller variance of $\widehat{\gamma}(k)$ compensates for its larger bias.

The sample ACF plays an active role in model identification. For example, the ACF of an MA(q) process cuts off at q. But, its sample ACF will not have a clear cutoff at lag q due to random fluctuations. The proper statistical inference rests on the sampling distributions of the statistics involved. Let $\boldsymbol{\rho}(k) = (\rho(1), \cdots, \rho(k))^\tau$ and $\widehat{\boldsymbol{\rho}}(k)$ be defined in the same way. The theorem below presents the asymptotic normality of sample mean $\bar{X}_T$, sample variance $\widehat{\gamma}(0)$, and sample ACF when the sample size $T \to \infty$. Its proof relies on the central limit theorem for m-dependent sequences; see, for example, Theorem 6.4.2 of Brockwell and Davis (1991). The basic idea is to approximate the double infinite MA process (2.24) by a finite MA process. We refer to §7.3 of Brockwell and Davis (1991) for the detailed technical derivations.

Theorem 2.8 *Let $\{X_t\}$ be a stationary process defined as*

$$X_t = \mu + \sum_{j=-\infty}^{\infty} a_j \varepsilon_{t-j}, \tag{2.24}$$

where $\{\varepsilon_t\} \sim \text{IID}(0, \sigma^2)$ and $\sum_{j=-\infty}^{\infty} |a_j| < \infty$.

(i) If $\sum_{j=-\infty}^{\infty} a_j \neq 0$, $\sqrt{T}(\bar{X}_T - \mu) \xrightarrow{D} N(0, \nu_1^2)$, where

$$\nu_1^2 = \sum_{j=-\infty}^{\infty} \gamma(j) = \sigma^2 \left(\sum_{j=-\infty}^{\infty} a_j \right)^2.$$

(ii) If $E\varepsilon_t^4 < \infty$, $\sqrt{T}\{\widehat{\gamma}(0) - \gamma(0)\} \xrightarrow{D} N(0, \nu_2^2)$, where

$$\nu_2^2 = 2\sigma^2 \sum_{j=-\infty}^{\infty} \rho(j)^2 = 2\sigma^2 \left\{ 1 + 2 \sum_{j=1}^{\infty} \rho(j)^2 \right\}. \tag{2.25}$$

(iii) If $E\varepsilon_t^4 < \infty$, $\sqrt{T}\{\widehat{\boldsymbol{\rho}}(k) - \boldsymbol{\rho}(k)\} \xrightarrow{D} N(0, \mathbf{W})$, *where* $\mathbf{W}$ *is a* $k \times k$ *matrix with its* (i, j)*th element given by Bartlett's formula*

$$\begin{aligned} w_{ij} &= \sum_{t=-\infty}^{\infty} \{\rho(t+i)\rho(t+j) + \rho(t-i)\rho(t+j) + 2\rho(i)\rho(j)\rho(t)^2 \\ &\quad - 2\rho(i)\rho(t)\rho(t+j) - 2\rho(j)\rho(t)\rho(t+i)\}. \end{aligned} \tag{2.26}$$

From (2.25), the sample variance $\widehat{\gamma}(0)$ has the asymptotic variance $2\sigma^2\{1+2\sum_{j\geq 1}\rho(j)^2\}/T$. When $\{X_t\}$ is an i.i.d. sequence (i.e., $a_j = 0$ for all $j \neq 0$ in (2.24)), this asymptotic variance becomes $2\sigma^2/T$. Comparing these two quantities, as far as the estimation of $\gamma(0) = \mathrm{Var}(X_t)$ is concerned, we may call

$$T' = T \Big/ \left\{1 + \sum_{j=1}^{\infty} \rho(j)^2\right\}$$

the equivalent number of independent observations, which reflects the loss of information due to the correlation in the data.

If $\{X_t\}$ is an MA(q) process (i.e., $a_j = 0$ for all $j < 0$ and $j > q$), it follows from Theorem 2.8(iii) that

$$\sqrt{T}\,\widehat{\rho}(j) \xrightarrow{D} N\left(0, 1 + 2\textstyle\sum_{t=1}^{q}\rho(q)^2\right), \quad j > q. \tag{2.27}$$

This is a very useful result for the estimation of the order q for an MA-process. In particular, if $\{X_t\} \sim \mathrm{WN}(0, \sigma^2)$, then

$$\sqrt{T}\,\widehat{\rho}(j) \xrightarrow{D} N(0, 1), \quad \text{for } j \neq 0.$$

Hence, there is an approximately 95% chance that $\widehat{\rho}(j)$ falls in the interval $\pm 1.96T^{-1/2}$.

2.2.3 Partial Autocorrelation

The ACF $\rho(k)$ measures the correlation between X_t and X_{t-k} regardless of their relationship with the intermediate variables $X_{t-1}, \cdots X_{t-k+1}$. The order determination in fitting an AR model relies on the correlation, conditioned on immediate variables; see, for example, Proposition 2.1 (iii). We will only include a further lagged variable X_{t-k} in the model for X_t if X_{t-k} makes a genuine and additional contribution to X_t in addition to those from $X_{t-1}, \cdots, X_{t-k+1}$. The *partial autocorrelation coefficient* (PACF) is used for measuring such a relationship.

Definition 2.6 *Let* $\{X_t\}$ *be a stationary time series with* $EX_t = 0$. *The PACF is defined as* $\pi(1) = \mathrm{Corr}(X_1, X_2) = \rho(1)$ *and*

$$\pi(k) = \mathrm{Corr}(R_{1|2,\cdots,k}, R_{k+1|2,\cdots,k}) \quad \text{for } k \geq 2,$$

where $R_{j|2,\cdots,k}$ is the residual from the linear regression of X_j on $(X_2, \cdots, X_k)$, namely

$$R_{j|2,\cdots,k} = X_j - (\alpha_{j2}X_2 + \cdots + \alpha_{jk}X_k),$$

and

$$(\alpha_{j2}, \cdots, \alpha_{jk}) = \arg \min_{\beta_2,\cdots,\beta_k} E\{X_j - (\beta_2 X_2 + \cdots + \beta_k X_k)\}^2. \qquad (2.28)$$

In the definition above, we assume that $EX_t = 0$ to simplify the notation. For a Gaussian process, the partial autocorrelation is in fact equal to

$$\pi(k) = E\{\mathrm{Corr}(X_1, X_{k+1}|X_2, \cdots, X_k)\}.$$

In general, PACF is introduced in a rather indirect manner and is defined in terms of the least square regression (2.28). Nevertheless, it follows immediately from the definition that the PACF cuts off at p for causal AR(p) processes. In general, the PACF is entirely determined by the ACF; see (2.29) below.

Proposition 2.3 *(i) For any stationary time series $\{X_t\}$,*

$$\pi(k) = \frac{\gamma(k) - \mathrm{Cov}(X_{k+1}, \mathbf{X}_{2,k}^\tau)\boldsymbol{\Sigma}_{2,k}^{-1}\mathrm{Cov}(\mathbf{X}_{2,k}, X_1)}{\gamma(0) - \mathrm{Cov}(X_1, \mathbf{X}_{2,k}^\tau)\boldsymbol{\Sigma}_{2,k}^{-1}\mathrm{Cov}(\mathbf{X}_{2,k}, X_1)}, \quad k \geq 1, \qquad (2.29)$$

where $\gamma(\cdot)$ is the ACVF *of $\{X_t\}$, $\mathbf{X}_{2,k} = (X_k, X_{k-1}, \cdots, X_2)^\tau$, and $\boldsymbol{\Sigma}_{2,k} = \mathrm{Var}(\mathbf{X}_{2,k})$.*

(ii) For causal AR*(p) models, $\pi(k) = 0$ for all $k > p$.*

The following theorem establishes a link between PACF and AR-modeling. It shows that $\pi(k)$ is the last autoregressive coefficient in the autoregressive approximation for X_t by its nearest k lagged variables. The following theorem is proved in §2.7.3.

Theorem 2.9 *Let $\{X_t\}$ be a stationary time series and $EX_t = 0$. Then $\pi(k) = b_{kk}$ for $k \geq 1$, where*

$$(b_{1k}, \cdots, b_{kk}) = \arg \min_{b_1,\cdots,b_k} E(X_t - b_1 X_{t-1} - \cdots - b_k X_{t-k})^2.$$

The theorem above also paves the way for the estimation of PACF—we need to fit a sequence of AR models with order $k = 1, 2, \cdots$ in order to estimate $\pi(k)$ for $k = 1, 2, \cdots$. More precisely, we estimate $\pi(k)$ by $\widehat{\pi}(k) = \widehat{b}_{kk}$ from the sample $(X_1, \cdots, X_T)$, where $(\widehat{b}_{k1}, \cdots, \widehat{b}_{kk})$ minimizes the sum

$$\sum_{t=k+1}^{T} (X_t - b_1 X_{t-1} - \cdots - b_k X_{t-k})^2.$$

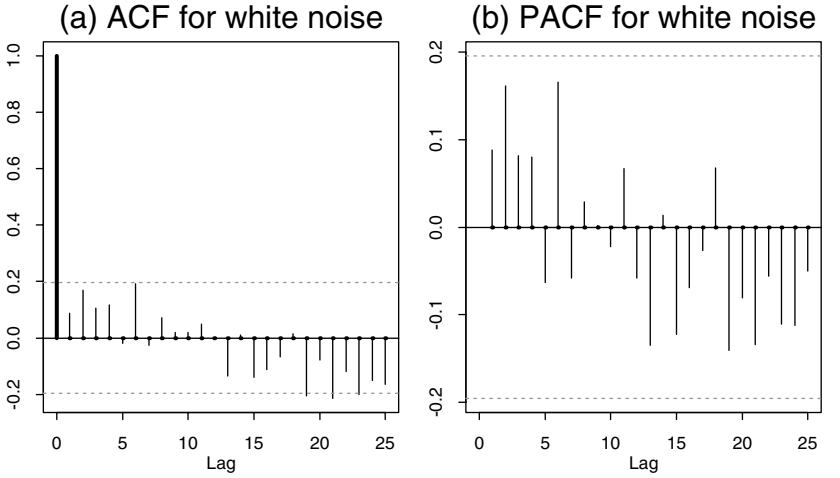


FIGURE 2.1. The sample (thin line) and the true (thick line) ACF and PACF plots for a Gaussian white noise process.

In practice, the estimation is often carried out in terms of some standard algorithms such as the Levinson–Durbin algorithm and the Burg algorithm; see §3.2.3 and §5.1 of Brockwell and Davis (1996). The asymptotic properties of $\widehat{\pi}(k)$ will be discussed in Chapter 3 in conjunction with those of parameter estimation for AR models; see Proposition 3.1.

We present the direct proofs for both Proposition 2.3(i) and Theorem 2.9 in §2.7, as their proofs in textbooks are usually mixed with the algorithms used in determining AR coefficients.

2.2.4 ACF Plots, PACF Plots, and Examples

Both ACF and PACF provide important information on the correlation structure of time series and play active roles in model identification as well as estimation. For example, the ACF cuts off at q for MA(q) processes and the PACF cuts off at p for AR(p) processes. Plotting the estimated ACF and PACF against the time lag is a simple but very useful technique in analyzing time series data. Such an ACF plot is called a *correlogram*.

In Examples 2.2–2.4, we plot some estimated ACFs and PACFs (thin lines) based on samples of size $T = 100$ together with the true ACFs and PACFs (thick lines); see Figures 2.1–2.3. We also superimpose the horizontal lines (dashed lines) at $\pm 1.96/\sqrt{T}$. These intervals give the pointwise acceptance region for testing the null hypothesis $H_0 : \rho(k) = 0$ at the 5% significance level; see (2.27) and its subsequent discussion. They assist us in judging whether a particular $\rho(k)$ is statistically significantly different from zero. We used standard Gaussian white noise $\{\varepsilon_t\} \sim_{\text{i.i.d.}} N(0, 1)$ in the examples.

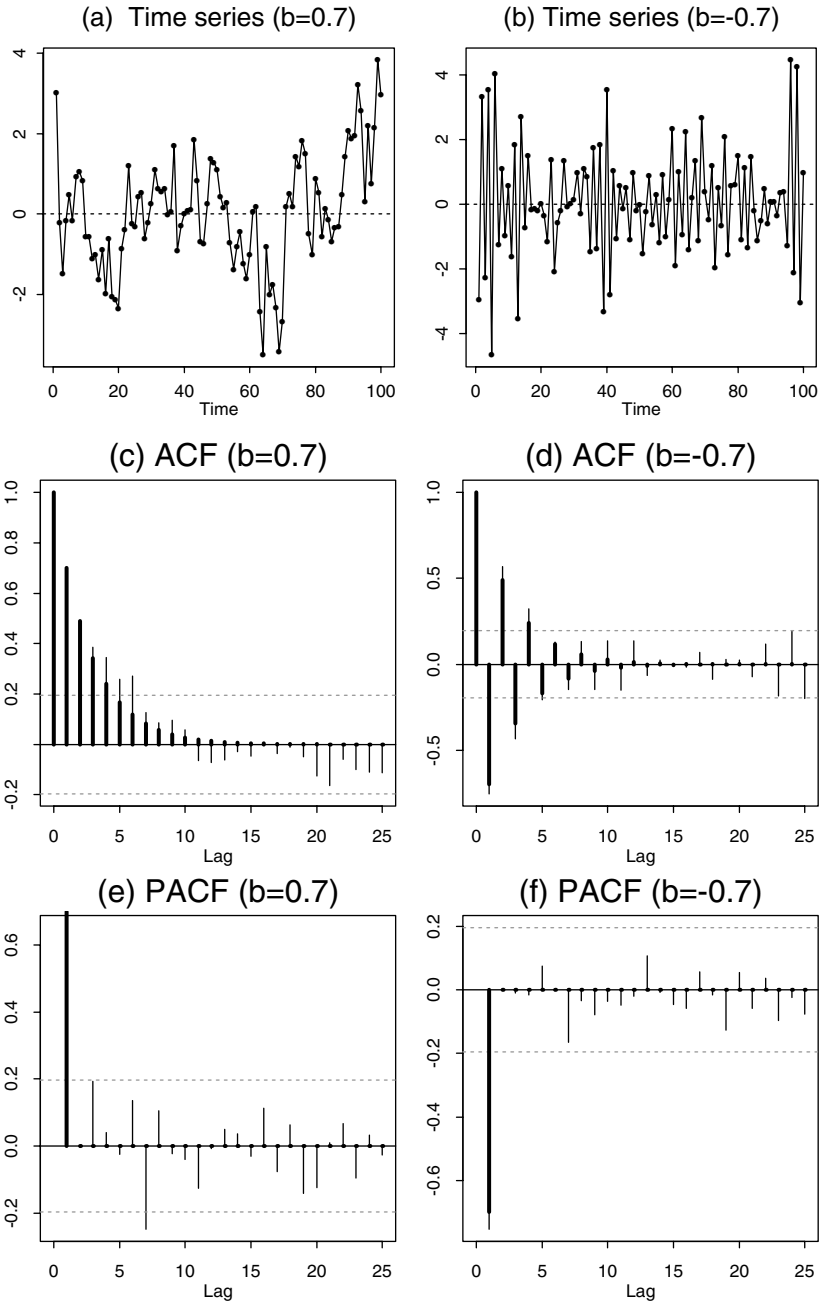


FIGURE 2.2. Time series plots and the sample (thin line) and the true (thick line) ACF and PACF plots for AR(1) models with $b = 0.7$ or -0.7.

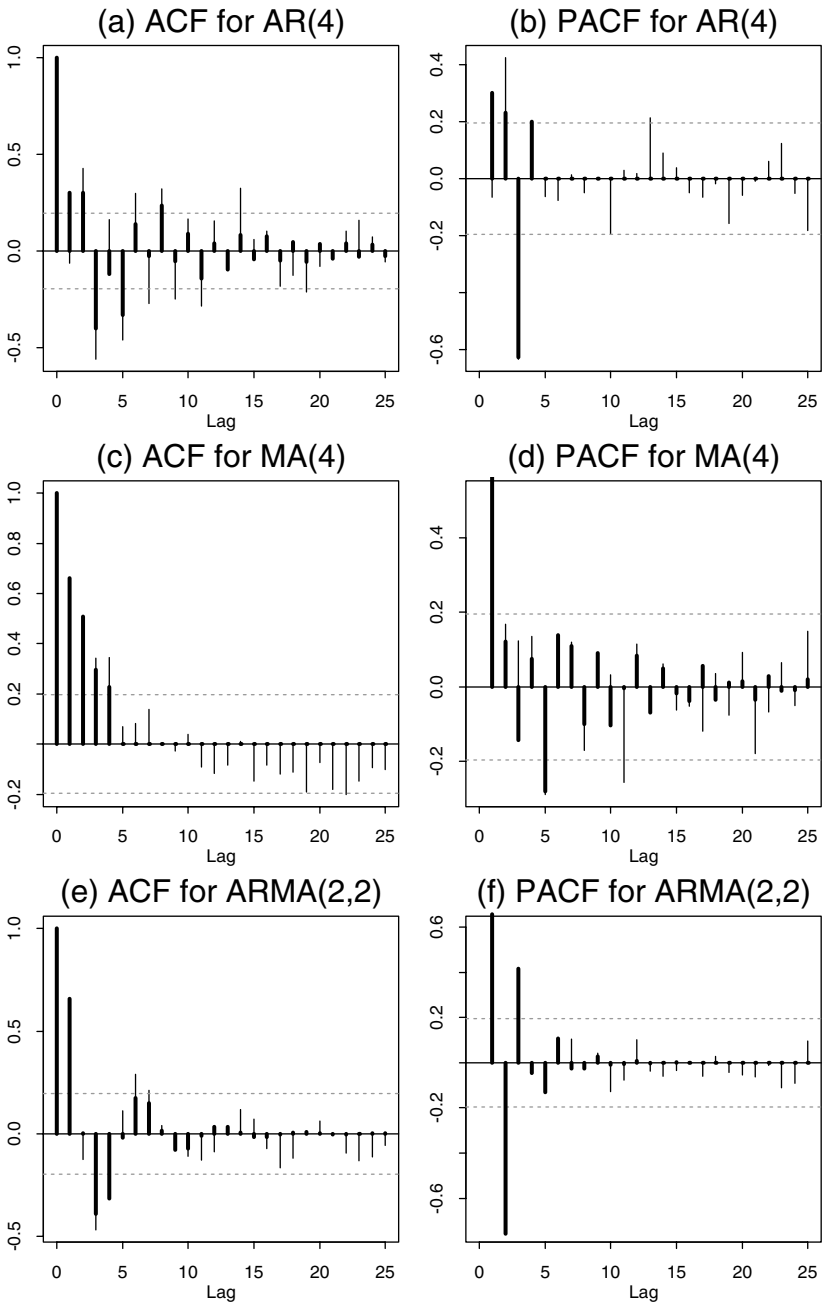


FIGURE 2.3. The sample (thin line) and the true (thick line) ACF and PACF plots for three stationary models defined in Example 2.4.

Example 2.2 *(White noise)* Let $X_t = \varepsilon_t$ for $t = 0, \pm 1, \cdots$. Then $\rho(k) = 0$ and $\pi(k) = 0$ for all $k \geq 1$. A sample of size 100 is generated from standard normal distribution. The estimated ACF and PACF are plotted in Figure 2.1. The estimated ACF and PACF are almost always between the two bounds $\pm 1.96/\sqrt{T} and \pm 0.196$. ∎

Example 2.3 Let us consider AR(1) model

$$X_t = bX_{t-1} + \varepsilon_t,$$

where $|b| < 1$. This process is causal (and therefore also stationary). It is easy to see that X_t depends on its past values through X_{t-1} only. From Yule–Walker equation (2.21), we may derive that $\rho(k) = b^{|k|}$. A simulated series with length 100 is plotted against time in Figures 2.2 (a) for $b = 0.7$ and 2.2 (b) for $b = -0.7$. When $b > 0$, the series is more stable and smoother in the sense that X_t tends to retain the same sign as X_{t-1}. In contrast, when $b < 0$, the series oscillates around its mean value 0. The similar pattern is preserved in its correlogram as shown in Figures 2.2 (c) and (d), although the absolute value of ACF decays fast. For the AR(1) model, $\pi(k) = 0$ for $k \geq 2$. Most estimated values for $\pi(k)$ $(k \geq 2)$ are between $\pm 1.96/\sqrt{T}$ and ± 0.196. ∎

Example 2.4 We consider three causal ARMA models:

$$\begin{aligned}
\text{AR}(4): &\quad X_t = 0.5X_{t-1} + 0.3X_{t-2} - 0.7X_{t-3} + 0.2X_{t-4} + \varepsilon_t, \\
\text{MA}(4): &\quad X_t = \varepsilon_t + 0.6\varepsilon_{t-1} + 0.6\varepsilon_{t-2} + 0.3\varepsilon_{t-3} + 0.7\varepsilon_{t-4}, \\
\text{ARMA}(2,2): &\quad X_t = 0.8X_{t-1} - 0.6X_{t-2} + \varepsilon_t + 0.7\varepsilon_{t-1} + 0.4\varepsilon_{t-2}.
\end{aligned}$$

Now, the correlation structure is no longer as clear-cut as for an AR(1) model, although $\rho(k) = 0$ for the MA(4) model and $\pi(k) = 0$ for the AR(4) model for all $k > 4$. Nevertheless, both ACF and PACF decay to 0 fast; see Figure 2.3. Furthermore, it tends to hold that, for large values of k,

$$|\widehat{\rho}(k)| > |\rho(k)|, \quad |\widehat{\pi}(k)| > |\pi(k)|.$$

This is due to the fact that when the true values of $\rho(k)$ and $\pi(k)$ are close to 0 for large k, the errors in estimation become "overwhelming". This phenomenon is common in the estimation of both ACF and PACF; see also Figures 2.1 and 2.2. ∎

2.3 Spectral Distributions

The techniques used in analyzing stationary time series may be divided into two categories: time domain analysis and frequency domain analysis. The former deals with the observed data directly, as in conventional statistical

analysis with independent observations. The frequency domain analysis, also called *spectral analysis*, applies the *Fourier transform* to the data (or ACVF) first, and the analysis proceeds with the transformed data only. The spectral analysis is in principle equivalent to the time domain analysis based on ACVF. However, it provides an alternative way of viewing a process via decomposing it into a sum of uncorrelated periodic components with different frequencies, which for some applications may be more illuminating. Since the properties beyond the second moments will be lost in spectral distributions, we argue that the spectral analysis, at least in its classical form, is not useful in handling nonlinear features. In this section, we first introduce the concept of spectral distribution via a simple periodic process. Spectral density is defined for stationary processes with "short memory" in the sense that $\sum_k |\gamma(k)| < \infty$. We derive a general form of spectral density functions for stationary ARMA processes via linear filters.

2.3.1 Periodic Processes

We first consider the simple periodic process

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

where both frequency ω and amplitude A are constant while the phase φ is a random variable distributed uniformly on the interval $[-\pi, \pi]$. Then $EX_t = 0$, and

$$\begin{aligned} \mathrm{Cov}(X_t, X_{t+\tau}) &= \frac{A^2}{2\pi}\int_{-\pi}^{\pi} \cos(\omega t + \varphi)\cos(\omega t + \omega\tau + \varphi)d\varphi \\ &= \frac{A^2}{4\pi}\int_{-\pi}^{\pi} \{\cos(2\omega t + 2\varphi + \omega\tau) + \cos(\omega\tau)\}d\varphi = \frac{A^2}{2}\cos(\omega\tau), \end{aligned} \tag{2.30}$$

which depends on τ only. Therefore, $\{X_t\}$ is stationary with $\gamma(\tau) = \frac{A^2}{2}\cos(\omega\tau)$.

Now, we turn to a more general form of *periodic process,*

$$X_t = \sum_{j=-k}^{k} A_j \cos(\omega_j t + \varphi_j), \tag{2.31}$$

where $\{\varphi_j\}$ are independent random variables with the common distribution $U[-\pi, \pi]$, $\{A_j\}$ and $\{\omega_j\}$ are constants, $A_0 = 0$, and

$$0 \le \omega_1 < \ldots < \omega_k \le \pi, \quad \omega_{-j} = -\omega_j.$$

Furthermore for $j = 1, \cdots, k$,

$$\varphi_{-j} = -\varphi_j, \quad A_{-j} = A_j.$$

By treating $\{X_t,\, t = 0, \pm 1, \pm 2, \ldots\}$ as observed values on regular time intervals of a continuous wave, the process is an accumulation of $2k$ periodic waves with frequencies $\omega_{-k}, \ldots, \omega_k$. Note that X_t is equal to

$$X_t = 2\sum_{j=1}^{k} A_j \cos(\omega_j t + \varphi_j). \tag{2.32}$$

Algebraic manipulation similar to (2.30) shows that $\{X_t\}$ is stationary with the ACVF

$$\gamma(\tau) = \sum_{j=-k}^{k} A_j^2 \cos(\omega_j \tau).$$

It is easy to see that any linear combination of sinusoidals can be expressed as in (2.32), and therefore also as in (2.31). We use the symmetric form (2.31) for technical convenience; see (2.33) and (2.34) below. Since we take observations at discrete times $0, \pm 1, \pm 2, \ldots$ only, waves with frequencies higher than π cannot be identified. (For any $X_t = \cos(\omega t + \varphi)$ with $\omega > \pi$, there exist $\omega' \in [0, \pi]$ and φ' such that $X_t = \cos(\omega' t + \varphi')$.) In principle, we may restrict frequencies to the interval $[0, \pi]$ only. We include $[-\pi, 0)$ in the frequency domain entirely for technical convenience.

Note that $\mathrm{Var}(X_t) = \gamma(0) = \sum_{j=-k}^{k} A_j^2$. Define the (unnormalized) spectral distribution function

$$G(\omega) = \sum_{j:\, \omega_j \le \omega} A_j^2, \quad -\pi \le \omega \le \pi,$$

which is a discrete distribution with mass A_j^2 at point ω_j for $j = \pm 1, \cdots \pm k$. In fact, $G(\omega)$ can be viewed as the contribution to $\mathrm{Var}(X_t)$ from the waves with frequencies not greater than ω. Therefore, if we regard $\mathrm{Var}(X_t)$ as the total power (or energy) of the process $\{X_t\}$, $G(\cdot)$ reflects how this total power is distributed over its components at different frequencies. In fact, the ACVF $\gamma(\cdot)$ can be expressed as a Stieltjes integral

$$\gamma(\tau) = \int_{-\pi}^{\pi} \cos(\omega\tau) dG(\omega) = \sum_{j=-k}^{k} \cos(\omega_j \tau) A_j^2.$$

Note that the symmetry of form (2.31) ensures that the distribution of $G(\cdot)$ is symmetric on the interval $[-\pi, \pi]$. Hence, the integral above can be written as

$$\gamma(\tau) = \int_{-\pi}^{\pi} \{\cos(\omega\tau) + i \sin(\omega\tau)\} dG(\omega) = \int_{-\pi}^{\pi} e^{i\omega\tau} dG(\omega), \tag{2.33}$$

where $i = \sqrt{-1}$.

We further normalize G and define the normalized spectral distribution function

$$F(\omega) = G(\omega)/\gamma(0) = G(\omega)/G(\pi).$$

Then F is a proper probability distribution that has probability mass $A_j^2/\gamma(0)$ at ω_j, $j = \pm 1, \cdots, \pm k$. It follows from (2.33) immediately that

$$\rho(\tau) = \int_{-\pi}^{\pi} e^{i\omega\tau} dF(\omega). \tag{2.34}$$

In summary, we have defined the spectral distribution for a time series that is an accumulation of finite periodic waves as defined in (2.31). The spectral distribution depicts the distribution of the total power (i.e. the variance) over the waves at different frequencies. Further, the ACVF and ACF can be expressed as Fourier transforms of the spectral distribution functions in (2.33) and (2.34). This simple model is illustrative, as any stationary time series can be viewed as an accumulation of (usually infinite) periodic waves with different frequencies. The statements above on spectral distributions are still valid in general.

2.3.2 Spectral Densities

We now introduce the spectral distribution or spectral density for a stationary time series through the Wiener–Khintchine theorem below. As we will see, a spectral distribution is defined in terms of an autocovariance function only. Therefore, it is powerless to deal with the properties beyond the second moments of a time series.

Theorem 2.10 (Wiener–Khintchine theorem) *A real-valued function defined at all the integers* $\{\rho(\tau) : \tau = 0, \pm 1, \pm 2, \ldots\}$ *is the ACF of a stationary time series if and only if there exists a symmetric probability distribution on* $[-\pi, \pi]$ *with distribution function* F *for which*

$$\rho(\tau) = \int_{-\pi}^{\pi} e^{i\tau\omega} dF(\omega), \tag{2.35}$$

where F *is called the* normalized spectral distribution function *of the time series. If* F *has a density function* f*,*

$$\rho(\tau) = \int_{-\pi}^{\pi} e^{i\tau\omega} f(\omega) d\omega,$$

and f *is called the* normalized spectral density function.

The theorem is also called Wold's theorem or Herglotz's theorem (in slightly different forms). We give a direct proof in §2.7.4, which is almost

the same as that on pp. 118–119 of Brockwell and Davis (1991), although they dealt with complex-valued processes.

Since $\rho(\cdot)$ is real, it holds that

$$\rho(\tau) = \int_{-\pi}^{\pi} \cos(\omega\tau) dF(\omega) = 2\int_{0}^{\pi} \cos(\omega\tau) dF(\omega).$$

Theorem 2.11 *Suppose that $\{\rho(\tau)\}$ is the ACF of a stationary time series and is absolutely summable in the sense that $\sum_{\tau=1}^{\infty} |\rho(\tau)| < \infty$. Then the normalized spectral density function exists and is a symmetric probability density function on the interval $[-\pi, \pi]$ defined as*

$$f(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \rho(\tau) e^{-i\tau\omega} = \frac{1}{2\pi} \left\{ 1 + 2\sum_{\tau=1}^{\infty} \rho(\tau) \cos(\omega\tau) \right\}. \qquad (2.36)$$

Proof. Let $f(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \rho(\tau) e^{-i\tau\omega}$. First, we show $f \geq 0$. Define $\xi = \sum_{j=1}^{n} e^{-ij\omega} X_j$. Then ξ is a random variable taking complex values, and

$$\mathrm{Var}(\xi) = \mathrm{Cov}(\xi, \bar{\xi}) = \sum_{j,k=1}^{n} \gamma(j-k) e^{-i(j-k)\omega} \geq 0,$$

where $\bar{\xi}$ denotes the conjugate of ξ. Define $f_n(\omega) = \mathrm{Var}(\xi)/\{2\pi n\gamma(0)\} \geq 0$. Then

$$f_n(\omega) = \frac{1}{2\pi n} \sum_{j,k=1}^{n} \rho(j-k) e^{-i(j-k)\omega} = \frac{1}{2\pi} \sum_{|m|<n} (1 - |m|/n) \rho(m) e^{-im\omega}.$$

For any $\varepsilon > 0$, we may choose a large integer $N > 0$ such that

$$\frac{1}{2\pi} \sum_{|m| \geq N} |\rho(m)| < \varepsilon.$$

Then, for any $n > N$,

$$|f_n(\omega) - f(\omega)| \leq \frac{1}{n} \frac{1}{2\pi} \sum_{|m|<N} |m\rho(m)| + 2\varepsilon \to 2\varepsilon \quad \text{as } n \to \infty.$$

This implies that $f_n(\omega) \to f(\omega)$. Therefore $f(\omega) \geq 0$.

Now, it holds for any integer j that

$$\int_{-\pi}^{\pi} e^{ij\omega} f(\omega) d\omega = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \rho(\tau) \int_{-\pi}^{\pi} e^{i(j-\tau)\omega} d\omega = \rho(j).$$

Let $j = 0$ in the expression above, and we have $\int_{-\pi}^{\pi} f(\omega) d\omega = 1$. Hence $f(\cdot)$ is the normalized spectral density. The second equality in (2.36) follows from the fact that $\rho(\cdot)$ is symmetric, which itself implies that $f(\cdot)$ is symmetric. ∎

In some applications such as engineering, spectral decomposition of the total power (i.e., the variance) is of primary interest. For this purpose, we define the nonnormalized spectral distribution and density functions as

$$G(\omega) = \gamma(0)F(\omega), \quad g(\omega) = \gamma(0)f(\omega),$$

which we simply call the *spectral distribution function* and the *spectral density function,* respectively. It follows from Theorems 2.10 and 2.11 immediately that

$$\gamma(\tau) = \int_{-\pi}^{\pi} e^{i\tau\omega} dG(\omega) = \int_{-\pi}^{\pi} \cos(\tau\omega) dG(\omega) \tag{2.37}$$

and

$$g(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma(\tau) e^{-i\tau\omega} = \frac{1}{2\pi} \left\{ \gamma(0) + 2 \sum_{\tau=1}^{\infty} \gamma(\tau) \cos(\tau\omega) \right\}, \tag{2.38}$$

provided that $\sum_\tau |\gamma(\tau)| < \infty$. Note that $G(\pi) = \gamma(0) = \mathrm{Var}(X_t)$. Hence, if we regard $\{X_t\}$ as an accumulation of periodic waves with different frequencies in $[-\pi, \pi]$,

$$G(\omega_2) - G(\omega_1) = \int_{\omega_1}^{\omega_2} g(\omega) d\omega$$

could be viewed as the contributions to the total power from the waves with the frequencies in the range $(\omega_1, \omega_2]$. If g is large at ω_0, the waves with frequencies around ω_0 make a large contribution to the total variation of $\{X_t\}$.

Formulas (2.36) or (2.38) may be used to calculate spectral density functions when ACVFs can be evaluated explicitly. For example, we know, by (2.38), that the spectral density for a white noise process is a constant on $[-\pi, \pi]$. Further, for the MA(q) process

$$X_t = \varepsilon_t + a_1\varepsilon_{t-1} + \cdots + a_q\varepsilon_{t-q}, \quad \{\varepsilon_t\} \sim \mathrm{WN}(0, \sigma^2),$$

the normalized spectral density is

$$f(\omega) = \frac{1}{2\pi} + \frac{1}{\pi} \sum_{k=1}^{q} \frac{\sum_{j=k}^{q} a_j a_{j-k}}{1 + \sum_{j=1}^{q} a_j^2} \cos(k\omega) \tag{2.39}$$

(see (2.19)). However, (2.36) and (2.38) do not lead to simple solutions for general stationary ARMA processes the explicit spectral density functions of which can be derived in terms of a device called the linear filter, discussed in §2.3.3 below.

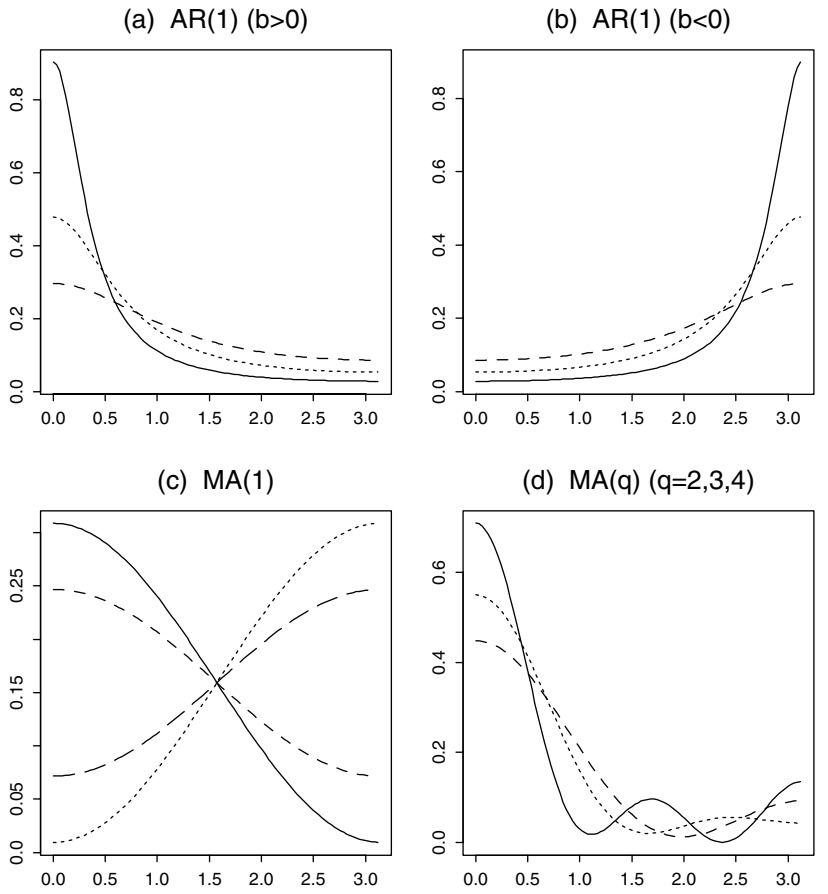


FIGURE 2.4. Spectral density functions for (a) AR(1) with $b = 0.7$ (solid line), 0.5 (dotted line), and 0.3 (dashed line); (b) AR(1) with $b = -0.7$ (solid line), -0.5 (dotted line), and -0.3 (dashed line); (c) MA(1) model with $a = 0.7$ (solid line), -0.7 (dotted line), 0.3 (dashed line), and -0.3 (long-dashed line); and (d) MA(4) process $X_t = \varepsilon_t + 0.6\varepsilon_{t-1} + 0.6\varepsilon_{t-2} + a_3\varepsilon_{t-3} + a_4\varepsilon_{t-4}$ with $(a_3, a_4) = (0.3, 0.7)$ (solid line), (0.3, 0) (dotted line), and (0, 0) (dashed line).

Example 2.5 For the stationary AR(1) process

$$X_t = bX_{t-1} + \varepsilon_t, \quad |b| < 1, \ \ \{\varepsilon_t\} \sim \mathrm{WN}(0, \sigma^2),$$

$\rho(k) = b^{|k|}$ ($|b| < 1$). It follows from (2.36) that the normalized spectral density function is

$$f(\omega) = \frac{1}{2\pi}\left\{1 + 2\sum_{k=1}^{\infty} b^k \cos(k\omega)\right\} = \frac{1}{2\pi}\left\{1 + 2\mathrm{Re}\left[\sum_{k=1}^{\infty}(be^{i\omega})^k\right]\right\}.$$

Note that

$$\sum_{k=1}^{\infty}(be^{i\omega})^k = \frac{be^{i\omega}}{1-be^{i\omega}} = \frac{b\cos\omega - b^2 + ib\sin\omega}{1+b^2-2b\cos\omega}.$$

Taking the real part, we obtain that

$$f(\omega) = \frac{1}{2\pi}\left\{1 + 2\frac{b\cos\omega - b^2}{1+b^2-2b\cos\omega}\right\} = \frac{1}{2\pi}\frac{1-b^2}{1+b^2-2b\cos\omega}. \tag{2.40}$$

■

We plot normalized spectral density functions of some simple stationary processes on the half interval $[0,\pi]$ in Figure 2.4. Note that the normalized spectral density function for MA processes is given by (2.39).

2.3.3 Linear Filters

Definition 2.7 *For two time series $\{X_t\}$ and $\{Y_t\}$, we call $\{X_t\}$ a filtered version of $\{Y_t\}$ if*

$$X_t = \sum_{k=-\infty}^{\infty} \varphi_k Y_{t-k}, \tag{2.41}$$

where the coefficients $\{\varphi_k\}$ are absolutely summable (i.e., $\sum_{k=-\infty}^{\infty}|\varphi_k| < \infty$).

The device (2.41) is often referred to as a *linear filter*, in which $\{Y_t\}$ is the input and $\{X_t\}$ is the output. The filter can be expressed in a more compact form in terms of the backshift operator,

$$X_t = \varphi(B)Y_t, \tag{2.42}$$

where

$$\varphi(z) = \sum_{k=-\infty}^{\infty} \varphi_k z^k.$$

We may purposely design the filter such that it will boost (or suppress) the signals (of the input) within a certain frequency band, producing output with the desired properties. The function

$$\Gamma(\omega) \equiv \sum_{k=-\infty}^{\infty} \varphi_k e^{-ik\omega} = \varphi(e^{-i\omega})$$

is called a *transfer function* of the linear filter. Its squared modulus $|\Gamma(\omega)|^2$ is called a *power transfer function*. The theorem below shows that the signal-boosting (or suppression) is controlled by the power transfer function.

Theorem 2.12 *Let $\{X_t\}$ and $\{Y_t\}$ be two stationary processes satisfying (2.41). Suppose that their ACFs are absolutely summable. Then*

$$g_x(\omega) = g_y(\omega)|\Gamma(\omega)|^2, \qquad -\pi \le \omega \le \pi,$$

where g_x and g_y are the spectral density functions of $\{X_t\}$ and $\{Y_t\}$, respectively.

Proof. Without loss of generality, we let $EX_t = EY_t = 0$. Then

$$\begin{aligned}\gamma_x(\tau) &= E(X_t X_{t+\tau}) = \sum_{j,k=-\infty}^{\infty} \varphi_j \varphi_k E(Y_{t-j} Y_{t+\tau-k}) \\ &= \sum_{j,k=-\infty}^{\infty} \varphi_j \varphi_k \gamma_y(\tau + j - k).\end{aligned}$$

Therefore

$$\begin{aligned}g_x(\omega) &= \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \gamma_x(\tau) e^{-i\tau\omega} = \frac{1}{2\pi} \sum_{j,k,\tau=-\infty}^{\infty} \varphi_j \varphi_k \gamma_y(\tau + j - k) e^{-i\tau\omega} \\ &= \frac{1}{2\pi} \sum_j \varphi_j e^{ij\omega} \sum_k \varphi_k e^{-ik\omega} \sum_\tau \gamma_y(\tau + j - k) e^{-i(\tau+j-k)\omega} \\ &= \frac{1}{2\pi} |\Gamma(\omega)|^2 \sum_l \gamma_y(l) e^{-il\omega} = g_y(\omega)|\Gamma(\omega)|^2.\end{aligned}$$

The proof is completed. ∎

Example 2.6 *(A three-point moving average filter of an AR(1))* Let $\{Y_t\}$ be a stationary AR(1) process defined by

$$Y_t = bY_{t-1} + \varepsilon_t, \quad |b| < 1, \ \ \{\varepsilon_t\} \sim \mathrm{WN}(0, \sigma^2).$$

Then $\mathrm{Var}(Y_t) = \sigma^2/(1 - b^2)$ and $\rho_y(\tau) = b^{|\tau|}$. It follows from (2.40) that

$$g_y(\omega) = \frac{1}{2\pi} \frac{\sigma^2}{1 + b^2 - 2b\cos(\omega)},$$

which is shown in Figure 2.5(a) with $b = 0.5$ and -0.5 (with $\sigma^2 = 0.75$). Define a three-point moving average filter

$$X_t = \frac{1}{3}(Y_{t-1} + Y_t + Y_{t+1}). \tag{2.43}$$

The transfer function is

$$\Gamma(\omega) = (e^{i\omega} + 1 + e^{-i\omega})/3 = \{1 + 2\cos(\omega)\}/3.$$

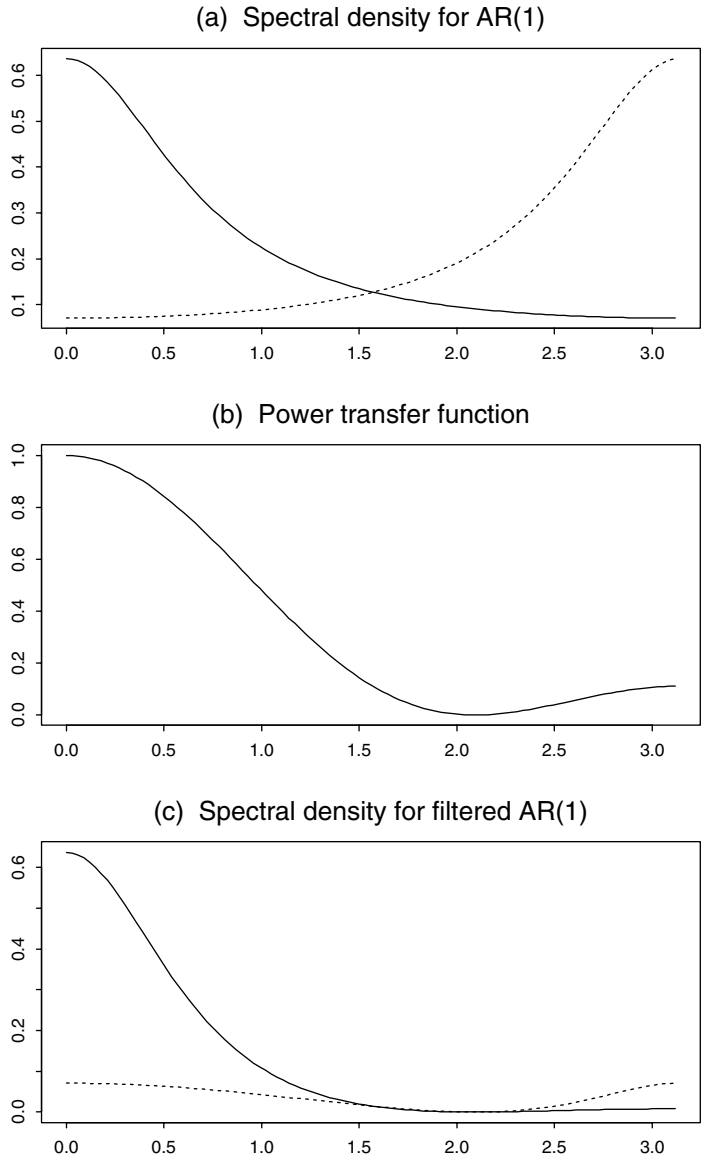


FIGURE 2.5. Example 2.6. (a) Spectral density for an AR(1) process $\{Y_t\}$ with $b = 0.5$ (solid lines) and $b = -0.5$ (dotted lines). (b) Power transfer function. (c) Spectral density for the output $\{X_t\}$ with $b = 0.5$ (solid lines) and $b = -0.5$ (dotted lines).

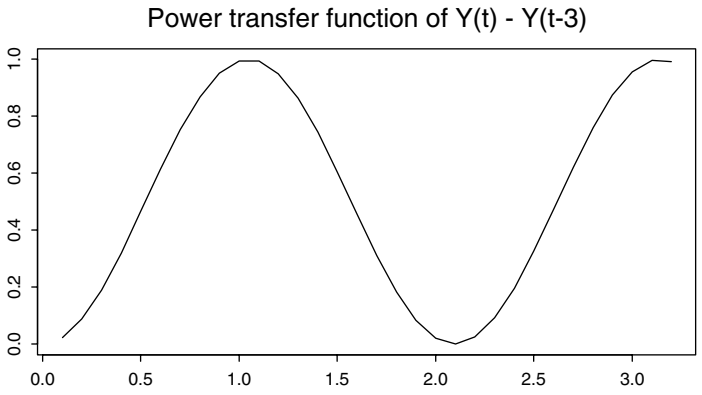


FIGURE 2.6. The power transfer function of the difference filter (2.45).

The power transfer function is

$$|\Gamma(\omega)|^2 = \frac{1 + 4\cos(\omega) + 4\cos^2(\omega)}{9} = \frac{3 + 4\cos(\omega) + 2\cos(2\omega)}{9}. \quad (2.44)$$

Figure 2.5(b) shows that (2.43) is a lower-pass filter since it passes the signals at lower frequencies and suppresses the signals at higher frequencies. It follows from Theorem 2.12 that the spectral density of $\{X_t\}$ is

$$g_x(\omega) = \frac{\sigma^2}{18\pi} \cdot \frac{3 + 4\cos(\omega) + 2\cos(2\omega)}{1 + b^2 - 2b\cos(\omega)},$$

which is plotted in Figure 2.5(c). Note that the AR(1) process with $b = -0.5$ has most power distributed in higher frequencies near π; see Figure 2.5(a). Having passed through a lower-pass filter (2.43), those high-frequency signals are largely suppressed. This deduces a substantial power (i.e., variance) loss in the output process. ∎

Note that the power transfer function (2.44) is equal to 0 at $\omega = \frac{2\pi}{3}$, so the three-point moving average filter removes the periodic components with period $2\pi/\omega = 3$. In practice, we often adopt the difference filter

$$X_t = (Y_t - Y_{t-3})/2 \quad (2.45)$$

to remove those components. The power transfer function of the difference filter above is

$$\{1 - \cos(3\omega)\}/2,$$

which is shown in Figure 2.6. We can see that this difference filter passes the signals around frequencies $\pi/3$ and π and removes signals at frequencies 0 and $\frac{2\pi}{3}$. Therefore, it is no longer a lower-pass filter. In general, we may

use either moving average or difference filters to remove certain periodic components. However we should be aware in the meantime of the different impacts on the filtered series; see, for example, Figures 2.5(b) and 2.6.

Now, we derive the spectral density function for a general ARMA(p, q) process defined by

$$X_t - b_1 X_{t-1} - \cdots - b_p X_{t-p} = \varepsilon_t + a_1 \varepsilon_{t-1} + \cdots + a_q \varepsilon_{t-q}, \tag{2.46}$$

or simply $b(B)X_t = a(B)\varepsilon_t$, where $\{\varepsilon_t\} \sim \text{WN}(0, \sigma^2)$ and $b(z) \neq 0$ for all $|z| \leq 1$. Define

$$\begin{aligned} Y_t &= X_t - b_1 X_{t-1} - \cdots - b_p X_{t-p} \\ &= \varepsilon_t + a_1 \varepsilon_{t-1} + \cdots + a_q \varepsilon_{t-q}. \end{aligned}$$

Then $\{Y_t\}$ is a filtered version of $\{X_t\}$ and also a filtered version of $\{\varepsilon_t\}$. It follows from Theorem 2.12 that

$$g_y(\omega) = g_x(\omega) \left| 1 - \sum_{j=1}^{p} b_j e^{-ij\omega} \right|^2 = g_\varepsilon(\omega) \left| 1 + \sum_{j=1}^{q} a_j e^{-ij\omega} \right|^2.$$

Since $g_\varepsilon(\omega) = \sigma^2/(2\pi)$, it holds that

$$g_x(\omega) = \frac{\sigma^2}{2\pi} \frac{|1 + \sum_{j=1}^{q} a_j e^{-ij\omega}|^2}{|1 - \sum_{j=1}^{p} b_j e^{-ij\omega}|^2} = \frac{\sigma^2}{2\pi} \frac{|a(e^{-i\omega})|^2}{|b(e^{-i\omega})|^2}. \tag{2.47}$$

Letting $b_1 = \cdots = b_p = 0$ in the expression above and comparing it with (2.39), we obtain that

$$\left| 1 + \sum_{j=1}^{q} a_j e^{-ij\omega} \right|^2 \propto 1 + 2 \sum_{k=1}^{q} \frac{\sum_{j=k}^{q} a_j a_{j-k}}{1 + \sum_{j=1}^{q} a_j^2} \cos(k\omega).$$

Combining this with (2.47), we obtain the following proposition showing that the spectral density of a stationary ARMA(p, q) process is of the form

$$\frac{A_0 + A_1 \cos(\omega) + \cdots + A_q \cos(q\omega)}{B_0 + B_1 \cos(\omega) + \cdots + B_q \cos(p\omega)}, \tag{2.48}$$

where $\{A_j\}$ and $\{B_j\}$ are constants. In fact, it is easy to see that this spectral density can be expressed more explicitly as follows.

Proposition 2.4 *For a stationary ARMA(p, q) process defined as in (2.46), the spectral density function is given as in (2.47). Furthermore,*

$$g_x(\omega) = \frac{\sigma^2}{2\pi} \frac{1 + \sum_{j=1}^{q} a_j^2 + 2 \sum_{k=1}^{q} \{\sum_{j=k}^{q} a_j a_{j-k}\} \cos(k\omega)}{1 + \sum_{j=1}^{p} b_j^2 + 2 \sum_{k=1}^{p} \{\sum_{j=k}^{p} b'_j b'_{j-k}\} \cos(k\omega)},$$

where $a_0 = b'_0 = 1$ and $b'_k = -b_k$ for $1 \leq k \leq p$.

Example 2.7 Let $\{X_t\} \sim \text{AR}(1)$ be stationary, and

$$Y_t = X_t + e_t, \quad \{e_t\} \sim \text{WN}(0, \sigma_e^2),$$

and $\{e_t\}$ is uncorrelated with $\{X_t\}$. It is easy to see that $\{Y_t\}$ is stationary and $\gamma_y(\tau) = \gamma_x(\tau)$ for $\tau \neq 0$ and $\gamma_x(\tau) + \sigma_e^2$ for $\tau = 0$. Hence, it follows from (2.48) that

$$g_y(\omega) = g_x(\omega) + \frac{\sigma_e^2}{2\pi} = \frac{A}{B + C\cos\omega} + \frac{\sigma_e^2}{2\pi} = \frac{A' + B'\cos\omega}{C' + D'\cos\omega}.$$

This spectral density looks similar to that of an ARMA(1, 1) process and hence seems to suggest that $\{Y_t\}$ is an ARMA(1, 1) process. Indeed, if we write $X_t - aX_{t-1} = \varepsilon_t$, we have an explicit expression for Y_t as follows

$$Y_t - aY_{t-1} = e_t - ae_{t-1} + \varepsilon_t.$$

Note that the term ε_t is invisible from the form of the spectral density of $\{Y_t\}$. ∎

2.4 Periodogram

The periodogram is a powerful tool for statistical inference for time series in the frequency domain. This is largely due to the fact that the periodogram ordinates for a stationary ARMA process are asymptotically independent and exponentially distributed. The periodogram is defined in terms of the discrete Fourier transform of observed data.

2.4.1 Discrete Fourier Transforms

Let $\{X_1, \cdots, X_T\}$ be T successive observations of a time series. Thinking of the underlying process as being periodic with the period T, we can express those X_t's as linear combinations of sinusoidals. To this end, we define *Fourier frequencies*

$$\omega_k = \frac{2\pi k}{T}, \quad k = -\left[\frac{T-1}{2}\right], \ldots, -1, 0, 1, \ldots, \left[\frac{T}{2}\right],$$

where $[y]$ denotes the integer part of y (i.e., the largest integer not greater than y). Let

$$\mathbf{e}_k = \frac{1}{\sqrt{T}} \begin{pmatrix} e^{i\omega_k} \\ e^{2i\omega_k} \\ \vdots \\ e^{Ti\omega_k} \end{pmatrix}, \quad k = -\left[\frac{T-1}{2}\right], \ldots, -1, 0, 1, \ldots, \left[\frac{T}{2}\right]. \quad (2.49)$$

Then, the T components of $\mathbf{e}_k$ may be viewed as the observed values at T discrete time points of a periodic wave at the frequency ω_k. Note that $\{\mathbf{e}_k\}$ are orthonormal in the sense that

$$\begin{aligned}\bar{\mathbf{e}}_j^\tau \mathbf{e}_k &= T^{-1}\sum_{l=1}^{T}\exp\{il(\omega_k-\omega_j)\} \\ &= T^{-1}\frac{[\exp\{iT(\omega_k-\omega_j)\}-1]\exp\{i(\omega_k-\omega_j)\}}{\exp\{i(\omega_k-\omega_j)\}-1} \\ &= \begin{cases} 1 & \text{if } k=j \\ 0 & \text{if } k\neq j, \end{cases}\end{aligned}$$

where $\bar{\mathbf{e}}_j^\tau = T^{-\frac{1}{2}}(e^{-i\omega_j},\ldots,e^{-Ti\omega_j})$ is the conjugate of $\mathbf{e}_j$. Therefore $\{\mathbf{e}_k\}$ is a base of the T-dimensional complex space in the sense that any T-dimensional complex vector can be expressed as a linear combination of $\mathbf{e}_j$'s. Hence, there exist T (complex) numbers α_k's such that

$$\mathbf{X} \equiv \begin{pmatrix} X_1 \\ \vdots \\ X_T \end{pmatrix} = \sum_{k=-[\frac{T-1}{2}]}^{[\frac{T}{2}]} \alpha_k \mathbf{e}_k. \tag{2.50}$$

This decomposes the series $\{X_t\}$ into linear combinations of periodic waves $\mathbf{e}_k$ with frequency ω_k. The magnitude $|\alpha_k|$ represents the energy of $\{X_t\}$ at the frequency ω_k. Due to the orthonormality of $\{\mathbf{e}_k\}$, it is easy to see, by multiplying $\bar{\mathbf{e}}_k^\tau$ on both sides, that

$$\alpha_k = \bar{\mathbf{e}}_k^\tau \mathbf{X} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} X_t e^{-it\omega_k} \tag{2.51}$$

and

$$\sum_{t=1}^{T} X_t^2 = \sum_{k=-[\frac{T-1}{2}]}^{[\frac{T}{2}]} |\alpha_k|^2. \tag{2.52}$$

We call $\{\alpha_k\}$ the *discrete Fourier transform* of $\{X_t\}$.

Since we only deal with real X_t's, the equation (2.50) reduces to

$$\begin{aligned} X_t &= \frac{1}{\sqrt{T}}\sum_{k=-[\frac{T-1}{2}]}^{[\frac{T}{2}]} \alpha_k e^{it\omega_k} \qquad (2.53) \\ &= \frac{1}{\sqrt{T}}\alpha_0 + \frac{2}{\sqrt{T}}\sum_{k=1}^{[\frac{T-1}{2}]}\{\alpha_{k,1}\cos(\omega_k t)+\alpha_{k,2}\sin(\omega_k t)\} + (-1)^{T/2}\frac{1}{\sqrt{T}}\alpha_{T/2} \end{aligned}$$

for $t = 1, \cdots, T$. In the expression above the last term on the right-hand side is defined to be 0 if T is odd, and

$$\alpha_{k,1} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \cos(\omega_k t), \quad \alpha_{k,2} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t \sin(\omega_k t). \tag{2.54}$$

2.4.2 Periodogram

Definition 2.8 *The* periodogram *of a set of real numbers* $\{X_1, \cdots, X_T\}$ *is defined as*

$$I_T(\omega_k) = \frac{1}{T} \left| \sum_{t=1}^{T} X_t e^{-it\omega_k} \right|^2 = |\alpha_k|^2, \quad k = -\left[\frac{T-1}{2}\right], \ldots, -1, 0, 1, \ldots, \left[\frac{T}{2}\right],$$

where $\omega_k = 2\pi k/T$ *is the Fourier frequency, and* α_k *is given in (2.51).*

Obviously,

$$I_T(\omega_k) = \alpha_{k,1}^2 + \alpha_{k,2}^2,$$

where $\alpha_{k,1}$ and $\alpha_{k,2}$ are defined in (2.54). Further, it follows from (2.52) immediately that

$$\sum_{t=1}^{T} X_t^2 = \sum_{k=-[\frac{T-1}{2}]}^{[\frac{T}{2}]} I_T(\omega_k).$$

The periodic representation (2.53) distributes the total energy $\sum_{t=1}^{T} X_t^2$ of the original data $\{X_1, \cdots, X_T\}$ over T periodic waves $\mathbf{e}_k$ with different frequencies ω_k and energy $I(\omega_k)$. When $I_T(\omega_k)$ is large, the waves at (or around) the frequency ω_k have large energy. The theorem below establishes the link between periodogram and spectral density function. Its proof is given in §2.7.5.

Theorem 2.13 *For* $k = -[\frac{T-1}{2}], \ldots, [\frac{T}{2}]$ *and* $k \neq 0$,

$$I_T(\omega_k) = \sum_{\tau=-(T-1)}^{T-1} \widehat{\gamma}(\tau) e^{-i\tau\omega_k},$$

where $\widehat{\gamma}(\cdot)$ *is the sample ACVF defined as in (2.23).*

The theorem above defines a natural estimator for spectral density function

$$\widehat{g}(\omega) = I_T(\omega)/(2\pi), \quad \omega \in (-\pi, \pi);$$

see (2.36). However, this naive substitution estimator is inconsistent and is therefore not that useful in practice since $\mathrm{Var}\{I_T(\omega)\}$ converges to a nonzero constant; see Theorem 2.14 below.

Theorem 2.14 *Suppose that* $\{X_1, \cdots, X_T\}$ *is a sample from the stationary process defined as*

$$X_t = \sum_{j=-\infty}^{\infty} a_j \varepsilon_{t-j}, \quad \{\varepsilon_t\} \sim \text{IID}(0, \sigma^2), \text{ and } \sum_{j=-\infty}^{\infty} |a_j| < \infty.$$

Let $n = [(T-1)/2]$. *For* $k = 1, \cdots, n$, *define*

$$\xi_{2k-1} = \frac{\sqrt{2}}{\sqrt{T}\,\sigma} \sum_{t=1}^{T} \varepsilon_t \cos(\omega_k t), \quad \xi_{2k} = \frac{\sqrt{2}}{\sqrt{T}\,\sigma} \sum_{t=1}^{T} \varepsilon_t \sin(\omega_k t).$$

(i) Because $T \to \infty$, $\{\xi_k,\ k = 1, \cdots, 2n\}$ *is a sequence of asymptotically independent and standard normal random variables in the sense that for any fixed* $c_1, \cdots, c_r \in R$ *and* $r \geq 1$, $\sum_{j=1}^{r} c_j \xi_{k_j} \xrightarrow{D} N(0, \sum_{j=1}^{r} c_j^2)$ *for any* $1 \leq k_1 < \cdots < k_r \leq 2n$.

(ii) For $k = 1, \cdots, n$,

$$I_T(\omega_k) = 2\pi g(\omega_k)\, \frac{\xi_{2k-1}^2 + \xi_{2k}^2}{2} + R_T(\omega_k),$$

where $g(\cdot)$ *is the spectral density of* $\{X_t\}$ *and* $\max_{1 \leq k \leq n} E|R_T(\omega_k)| \to 0$ *as* $T \to \infty$.

The proof of Theorem 2.14 is given in §2.7.6. It follows from Theorem 2.14(i) that $\xi_{2k-1}^2 + \xi_{2k}^2$ is asymptotically χ^2 with 2 degrees of freedom. Hence, the limit distribution of random variable $(\xi_{2k-1}^2 + \xi_{2k}^2)/2$ is exponential with mean 1. By Theorem 2.14(ii), we could approximately regard

$$I_T(\omega_k)/\{2\pi g(\omega_k)\} \quad \text{for } k = 1, \cdots, n$$

as n i.i.d. standard exponential random variables when the sample size T is large. It can also be proved that under some additional conditions on a_j's and ε_t

$$\text{Var}\{I_T(\omega_k)\} = 4\pi^2 g^2(\omega_k) + O(T^{-1/2}), \quad k = 1, \cdots, n;$$

see Theorem 10.3.2 of Brockwell and Davis (1991). Note that in the theorem above we only consider periodogram ordinates $I_T(\omega_k)$ with $\omega_k \in (0, \pi)$. It is easy to see from Definition 2.8 that

$$I_T(\omega) = I_T(-\omega), \quad \omega \in (0, \pi).$$

The symmetry above is in alignment with the symmetry of spectral density functions. However, in most applications, we use periodogram ordinate $I_T(\omega)$ with positive ω only.

To overcome the inconsistency problem in practice, tapering or other smoothing techniques are often applied in calculating the periodogram; see Brillinger (1981), Dahlhaus (1990b), Chen, Dahlhaus and Wu (2000), and also §7.2 and §7.3.

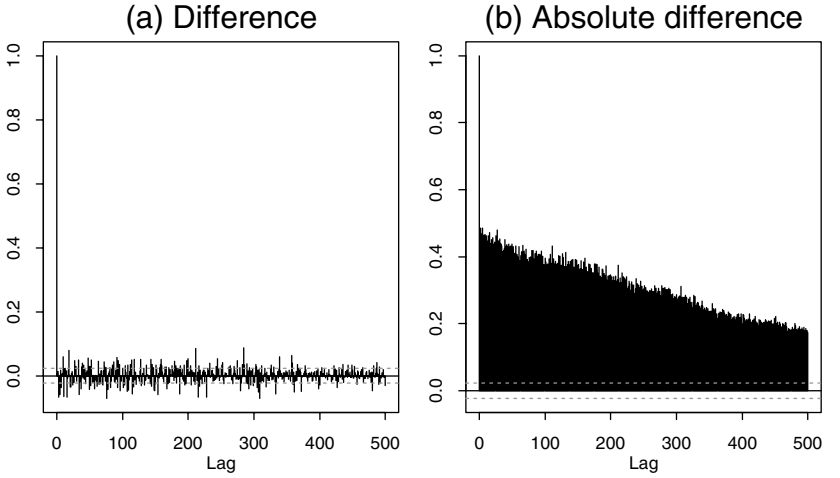


FIGURE 2.7. ACF plots for (a) the differences $\{Y_t - Y_{t-1}\}$ and (b) the absolute differences $\{|Y_t - Y_{t-1}|\}$ of S&P 500 Index data reported in Example 1.4.

2.5 Long-Memory Processes*

It follows from Proposition 2.2(i) (see also (2.22)) that the ACF of a stationary ARMA process satisfies the inequality

$$|\rho(k)| \le Cr^k, \quad k = 1, 2, \cdots,$$

where $C > 0$ and $r \in (0, 1)$ are some constants. Therefore $\sum_{k=0}^{\infty} |\rho(k)| < \infty$. A process with the absolutely summable ACF is often referred to as a short-memory process. There exists another type of stationary process for which the ACF decays to 0 at a much slower rate; for example, it exhibits the asymptotic behavior

$$\rho(k) \sim Ck^{2d-1} \quad \text{as } k \to \infty,$$

where $C \neq 0$ and $d < 0.5$. We refer to the feature above as the long-memory phenomenon. In other words, the ACF of a long-memory process decays to 0 at the slower rate k^{2d-1}, and $\sum_{k=0}^{\infty} |\rho(k)| = \infty$ when $d \in (0, 0.5)$.

The long-memory features have been observed in diverse fields such as hydrology, economics, and finance. Figure 2.7 displays the ACF plots for both differenced series $\{Y_t - Y_{t-1}\}$ and absolutely differenced series $\{|Y_t - Y_{t-1}|\}$, where $\{Y_t\}$ is the S&P 500 Index time series reported in Example 1.4. There seems to exist overwhelming evidence of the long-memory feature in the absolute differences $\{|Y_t - Y_{t-1}|\}$. In this section, we present an introduction to fractionally integrated ARMA processes, which form the most frequently used class of long-memory processes. We refer the reader to Beran (1995) for a systematic treatment of long-memory processes.

2.5.1 Fractionally Integrated Noise

For any real number $d > -1$, define the difference operator by means of the binomial expansion

$$\nabla^d \equiv (1-B)^d = \sum_{j=0}^{\infty} \varphi_j B^j,$$

where $\varphi_0 = 1$, for $j \geq 1$

$$\varphi_j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} = \prod_{0<k\leq j} \frac{k-1-d}{k},$$

and $\Gamma(\cdot)$ is the gamma function defined as

$$\Gamma(x) = \begin{cases} \int_0^\infty t^{x-1}e^{-t}dt, & x > 0, \\ \infty & x = 0, \\ x^{-1}\Gamma(1+x), & -1 < x < 0. \end{cases}$$

Definition 2.9 *A zero-mean stationary process* $\{X_t, t = 0, \pm 1, \pm 2, \cdots\}$ *is said to be an ARIMA*$(0, d, 0)$ *process with* $d \in (-0.5, 0.5)$ *and* $d \neq 0$ *if*

$$\nabla^d X_t = \varepsilon_t, \qquad \{\varepsilon_t\} \sim \text{WN}(0, \sigma^2). \tag{2.55}$$

$\{X_t\}$ *is often called* fractionally integrated noise.

It can be shown that for $d \in (-0.5, 0.5)$, $\sum \varphi_j^2 < \infty$. This ensures that $\nabla^d X_t = \sum_{j=0}^{\infty} \varphi_j X_{t-j}$ converges in mean square. The theorem below guarantees the existence of fractionally integrated noise processes; see §13.2 of Brockwell and Davis (1991) for its proof.

Theorem 2.15 *For* $d \in (-0.5, 0.5)$ *and* $d \neq 0$*, there exists a unique purely nondeterministic, zero-mean, and stationary process*

$$X_t = \nabla^{-d}\varepsilon_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad t = 0, \pm 1, \pm 2, \cdots,$$

which satisfies (2.55), where $\psi_0 = 1$*, and*

$$\psi_j = \frac{\Gamma(j+d)}{\Gamma(j+1)\Gamma(d)} = \prod_{0<k\leq j} \frac{k-1+d}{k}, \quad j = 1, 2, \cdots.$$

Furthermore, $Var(X_t) = \gamma_x(0) = \sigma^2\Gamma(1-2d)/\Gamma^2(1-d)$*, the ACF of* $\{X_t\}$ *is of the form*

$$\rho(k) = \frac{\Gamma(k+d)\Gamma(1-d)}{\Gamma(k-d+1)\Gamma(d)} = \prod_{0<j\leq k} \frac{j-1+d}{j-d}, \quad k = 1, 2, \cdots,$$

the PACF $\pi(k) = d/(k-d)$ $(k \geq 1)$*, and the spectral density function of* $\{X_t\}$ *may be written as*

$$g(\omega) = \frac{\sigma^2}{2\pi}|1 - e^{-i\omega}|^{-2d} = \frac{\sigma^2}{2\pi}|2\sin(\omega/2)|^{-2d}.$$

By Stirling's formula $\Gamma(x) \sim \sqrt{2\pi}e^{-x+1}(x-1)^{x-1/2}$ $(x \to \infty)$, we may see that the ACF of a fractionally integrated noise process admits the asymptotic expression

$$\rho(k) \sim k^{2d-1}\Gamma(1-d)/\Gamma(d) \quad \text{as } k \to \infty.$$

Thus $\sum_k |\rho(k)| = \infty$ for $d > 0$. This tail behavior of the ACF is reflected in its spectral density around the origin as:

$$g(\omega) \sim \frac{\sigma^2}{2\pi}\omega^{-2d} \quad \text{as } \omega \to 0,$$

which has an infinite pole at the origin when $d > 0$. (Note that $\sin(x) \sim x$ as $x \to 0$.) Fractionally integrated noise processes themselves are of limited value in modeling long-memory data since the two parameters d and σ^2 allow little flexibility. They serve as building blocks to generate a much more general class of long-memory processes—fractionally integrated ARMA processes.

2.5.2 Fractionally Integrated ARMA processes

Definition 2.10 *A zero-mean stationary process* $\{X_t, t = 0, \pm 1, \pm 2, \cdots\}$ *is said to be a FARIMA*(p, d, q) *process with* $d \in (-0.5, 0.5)$ *and* $d \neq 0$ *if*

$$\nabla^d X_t \sim \text{ARMA}(p, q).$$

$\{X_t\}$ *is also called a* fractionally integrated ARMA process.

Let $\{X_t\} \sim \text{FARIMA}(p, d, q)$. The definition above implies that

$$b(B)\nabla^d X_t = a(B)\varepsilon_t, \quad \{\varepsilon_t\} \sim \text{WN}(0, \sigma^2), \tag{2.56}$$

where $b(z) = 1 - b_1 z - \cdots - b_p z^p$, $a(z) = 1 + a_1 z + \cdots + a_q z^q$. Note that $a^{-1}(B)$, $b(B)$, and ∇^d are polynomials of operators B. Since the terms in a product of those polynomials are exchangeable, it holds that

$$\nabla^d a^{-1}(B)b(B)X_t = \varepsilon_t.$$

Let $Y_t = a^{-1}(B)b(B)X_t$. Then $\nabla^d Y_t = \varepsilon_t$ (i.e., $\{Y_t\}$ is a fractionally integrated noise). Note that

$$b(B)X_t = a(B)Y_t.$$

Thus, a FARIMA(p, d, q) process can be viewed as an ARMA(p, q) process driven by a fractionally integrated noise FARIMA$(0, d, 0)$. The theorem below guarantees the existence of fractionally integrated ARMA processes; see §13.2 of Brockwell and Davis (1991) for its proof.

Theorem 2.16 *Let $d \in (-0.5, 0.5)$ and $d \neq 0$, and $a(z) = 0$ and $b(z) = 0$ have no common roots. If $b(z) \neq 0$ for all $|z| \leq 1$, equation (2.56) defines a unique nondeterministic stationary solution*

$$X_t = \sum_{j=0}^{\infty} \psi_j \nabla^{-d} \varepsilon_{t-j},$$

where $\psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = a(z)/b(z)$. Furthermore, the ACF and the spectral density function of $\{X_t\}$ exhibit the asymptotic properties

$$\rho(k) \sim Ck^{2d-1} \quad \text{as } k \to \infty,$$

where $C \neq 0$, and

$$g(\omega) = \frac{\sigma^2}{2\pi} \frac{|a(e^{-i\omega})|^2}{|b(e^{-i\omega})|^2} |1 - e^{-i\omega}|^{-2d} \sim \frac{\sigma^2}{2\pi} [a(1)/b(1)]^2 \omega^{-2d} \quad \text{as } \omega \to 0.$$

As we pointed out earlier, a long-memory process FARIMA(p, d, q) is an ARMA process driven by a fractionally integrated noise FARIMA$(0, d, 0)$. Theorem 2.16 indicates that the FARIMA(p, d, q) exhibits the same long-memory behavior as the FARIMA$(0, d, 0)$, reflected by the asymptotic properties of both the ACF (as $k \to \infty$) and spectral density (as $\omega \to 0$); see also Theorem 2.15.

2.6 Mixing*

The classical asymptotic theory in statistics is built on the central limit theorem and the law of large numbers for the sequences of independent random variables. In the study of the asymptotic properties for linear time series that are the sequences of dependent random variables, the conventional approach is to express a time series in terms of an MA process in which the white noise $\{\varepsilon_t\}$ is assumed to be i.i.d.; see, for example, Theorems 2.8 and 2.14. Unfortunately, the MA representation such as (2.24) is no longer relevant in the context of nonlinear time series, where more complicated dependence structures will be encountered. We need to impose certain asymptotic independence in order to appreciate large sample properties of nonlinear time series inferences. A *mixing* time series can be viewed as a sequence of random variables for which the past and distant future are asymptotically independent.

For mixing sequences, both the law of large numbers (i.e., ergodic theorem) and central limit theorem can be established. In this section, we introduce different mixing conditions. Since the α-mixing is the weakest among the most frequently used mixing conditions, we state some limit theorems and probability inequalities for α-mixing processes. They play important roles in the development of asymptotic theory for nonlinear time series. For a more detailed discussion on mixing conditions, we refer the reader to Bradley (1986) and Doukhan (1994). Finally, we present a central limit theorem for a generic form that is constantly encountered in nonparametric regression based on kernel smoothing.

2.6.1 Mixing Conditions

To simplify the notation, we only introduce mixing conditions for strictly stationary processes (in spite of the fact that a mixing process is not necessarily stationary). The idea is to define *mixing coefficients* to measure the strength (in different ways) of dependence for the two segments of a time series that are apart from each other in time. Let $\{X_t,\ t = 0, \pm 1, \pm 2, \cdots\}$ be a strictly stationary time series. For $n = 1, 2, \cdots$, define

$$\begin{aligned}
\alpha(n) &= \sup_{A \in \mathcal{F}_{-\infty}^{0},\ B \in \mathcal{F}_{n}^{\infty}} |P(A)P(B) - P(AB)|, \\
\beta(n) &= E\left\{ \sup_{B \in \mathcal{F}_{n}^{\infty}} |P(B) - P(B|X_0, X_{-1}, X_{-2}, \cdots)| \right\}, \\
\rho(n) &= \sup_{X \in \mathcal{L}^2(\mathcal{F}_{-\infty}^{0}),\ Y \in \mathcal{L}^2(\mathcal{F}_{n}^{\infty})} |\mathrm{Corr}(X, Y)|, \\
\varphi(n) &= \sup_{A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{n}^{\infty}, P(A) > 0} |P(B) - P(B|A)|, \\
\psi(n) &= \sup_{A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{n}^{\infty}, P(A)P(B) > 0} |1 - P(B|A)/P(B)|, \qquad (2.57)
\end{aligned}$$

where $\mathcal{F}_i^j$ denotes the σ-algebra generated by $\{X_t,\ i \le t \le j\}$, and $\mathcal{L}^2(\mathcal{F}_i^j)$ consists of $\mathcal{F}_i^j$-measurable random variables with finite second moment. (Readers are referred to Chapter 1 of Chow and Teicher (1997) for the definitions of σ-algebra and measurable functions.) Intuitively, $\mathcal{F}_i^j$ assembles all information on the time series collected between time i and time j. When at least one of the mixing coefficients converges to 0 as $n \to \infty$, we may say that the process $\{X_t\}$ is asymptotically independent. Note that $\mathcal{F}_{n+1}^{\infty} \subset \mathcal{F}_n^{\infty}$ for any $n \ge 1$. Thus, all of the mixing coefficients defined above are monotonically nonincreasing.

Definition 2.11 *The process $\{X_t\}$ is said to be* α-mixing *if $\alpha(n) \to 0$,*

β-mixing *if* $\beta(n) \to 0$,
ρ-mixing *if* $\rho(n) \to 0$,
φ-mixing *if* $\varphi(n) \to 0$,
and
ψ-mixing *if* $\psi(n) \to 0$ *as* $n \to \infty$.

Some basic facts on mixing conditions are now in order.

(i) The diagram below illustrates the relationships among the five mixing conditions:

$$\psi\text{-mixing} \longrightarrow \varphi\text{-mixing} \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} \beta\text{-mixing} \\ \rho\text{-mixing} \end{array} \begin{array}{c} \searrow \\ \nearrow \end{array} \alpha\text{-mixing};$$

see, for example, Bradley (1986). Further, it is well-known that

$$\alpha(k) \leq \frac{1}{4}\rho(k) \leq \frac{1}{2}\varphi^{1/2}(k).$$

The α-mixing, also called *strong mixing,* is the weakest among the five, which is implied by any one of the four other mixing conditions. On the other hand, ψ-mixing is the strongest. In general, β-mixing (also called *absolute regular*) and ρ-mixing do not imply each other. However, for Gaussian processes, ρ-mixing is equivalent to α-mixing and therefore is weaker than β-mixing. See §1.3.2 of Doukhan (1994) for examples of time series with various mixing properties.

(ii) The mixing properties are hereditary in the sense that, for any measurable function $m(\cdot)$, the process $\{m(X_t)\}$ possesses the mixing property of $\{X_t\}$.

(iii) Consider the MA(∞) process

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j},$$

where $a_j \to 0$ exponentially fast (note that causal ARMA(p, q) processes fulfill this condition), and $\{\varepsilon_t\}$ is an i.i.d. sequence. If the probability density function of ε_t exists (such as normal, Cauchy, exponential, and uniform distributions), then $\{X_t\}$ is β-mixing with $\beta(n) \to 0$ exponentially fast (Pham and Tran 1985). However, this result does not always hold when ε_t is discrete. For example, the process $X_{t+1} = 0.5X_t + \varepsilon_t$ is not α-mixing when ε_t has a binomial distribution; see Andrews (1984).

(iv) If $\{X_t\}$ is a strictly stationary Markov chain, the mixing coefficients are effectively defined with $(\mathcal{F}_{-\infty}^0, \mathcal{F}_n^\infty)$ replaced by $(\sigma(X_0), \sigma(X_n))$, where $\sigma(X_t)$ denotes the σ-algebra generated by the single random variable X_t only (Theorem 4.1 of Bradley 1986). Further, the mixing coefficients decay to 0 exponentially fast if $\{X_t\}$ is ρ-, ϕ-, or ψ-mixing (Theorem 4.2 of Bradley 1986).

(v) It follows from (iv) above and Lemma 1.3 in Bosq (1998) that if $\{X_t\}$ is a strictly stationary and α-mixing Markov chain, the mixing coefficient is bounded by

$$\alpha(n) \leq \frac{1}{2} \int |f_{0,n}(x,y) - f(x)f(y)| dxdy,$$

where f is the marginal density function of X_t, and $f_{0,n}$ is the joint density of (X_0, X_n).

(vi) Davydov (1973) showed that, for a strictly stationary Markov chain $\{X_t\}$,

$$\beta(n) = \int ||F_n(\cdot|x) - F(\cdot)|| F(dx), \tag{2.58}$$

where F is the marginal distribution of X_t, $F_n(\cdot|x)$ is the conditional distribution of X_n given $X_0 = x$, and $||\cdot||$ denotes the total variation. If $\{X_t\}$ is geometrically ergodic satisfying the additional condition that

$$||F_n(\cdot|x) - F(\cdot)|| \leq A(x)\eta^n \tag{2.59}$$

almost surely with respect to the distribution $F(\cdot)$, where $\eta \in (0,1)$ is a constant and $\int A(x)F(dx) < \infty$ (see also (2.10)), it follows immediately from (2.58) that $\{X_t\}$ is β-mixing with exponentially decaying coefficients. Nummelin and Tuominen (1982) provided some sufficient conditions under which (2.59) holds; see also §2.4 of Doukhan (1994).

(vii) We may define a "one-side-infinite" process $\{X_t,\ t \geq 1\}$ to be, for example, α-mixing if $\alpha(n) \to 0$, where $\alpha(n)$ is defined as in (2.57) with "$\sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_n^\infty}$" replaced by "$\max_{k \geq 1} \sup_{A \in \mathcal{F}_1^k, B \in \mathcal{F}_{n+k}^\infty}$". Then, a strictly stationary process $\{X_t, t = 0, \pm 1, \pm 2, \cdots\}$ is α-mixing if and only if its "positive half" $\{X_t, t = 1, 2, \cdots\}$ is α-mixing. This remark also applies to the four other mixing conditions.

(viii) A GARCH(p,q) process defined by (1.7) and (1.6) is α-mixing with exponentially decaying coefficients if (i) $\sum_{1 \leq i \leq p} a_i + \sum_{1 \leq j \leq q} b_j < 1$ and (ii) the density function of ε_t is positive in an interval containing 0; see Theoerm 3.1 and Remark 3.2 of Basrak, Davis and Mikosch (2002).

(ix) Any sequence of independent (or m-dependent) random variables is mixing in all of the types defined in Definition 2.11. On the other hand, a sequence generated by a deterministic equation such as

$$X_{t+1} = m(X_t), \quad t = 0, \pm 1, \pm 2, \cdots$$

is not α-mixing, where $m(\cdot)$ is a nonlinear function. Intuitively, X_{t+n} is completely determined by X_t. Hence, their dependence cannot vanish even when $n \to \infty$. To appreciate this, assume that the process defined above is stable in the sense that it admits an *invariant probability measure*

$$P(X_t \in A) = \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} I(X_j \in A), \quad \text{for all } t \geq 1 \text{ and measurable } A.$$

Take a point x_0 such that $c \equiv P(A) = P(X_1 \le x_0) > 0$ and let

$$A_0 = \{m^{(n)}(X_0) \le x_0\} \in \mathcal{F}_{-\infty}^0, \quad B_0 = \{X_n \le x_0\} \in \mathcal{F}_n^\infty,$$

where $m^{(n)}$ denotes the nth fold of m. Then $A_0 = B_0$ since $X_n = m^{(n)}(X_0)$. Thus

$$\alpha(n) \ge |P(AB) - P(A)P(B)| = P(A)\{1 - P(A)\} = c(1-c) > 0$$

for all $n \ge 1$. Therefore $\alpha(n)$ does not converge to 0.

In the rest of this book, we mainly deal with the α-mixing processes. The discussion above indicates that such a condition is likely to hold for strictly stationary time series, including ARMA processes and geometrically ergodic Markov chains. On the other hand, it is by no means easy in general to check whether a nonlinear time series is, for example, α-mixing. The special properties for Markov chains stated in (iv)–(vi) above certainly make the theoretical investigation easier. In this vein, mixing properties for nonlinear AR and nonlinear ARCH(1) processes have been established; see, for example, §2.4.2 of Doukhan (1994). Unfortunately, the required conditions on underlying distributions are difficult to check in general. This partially explains why it is a common practice to assume a certain mild asymptotic independence (such as α-mixing) as a precondition in the context of asymptotic theory of statistical analysis for nonlinear time series.

2.6.2 Inequalities

We introduce three types of inequalities, namely covariance inequalities, moment inequalities for partial sums, and exponential inequalities for tail probabilities. They will serve as basic tools in the development of asymptotic theory in nonlinear time series analysis; see, for example, the proof of Theorem 2.21 in §2.6.3. The exponential inequalities are required to derive uniform convergence rates for nonparametric estimators.

(a) *Covariance inequalities*

Let X and Y be two real random variables. Define

$$\alpha = \sup_{A \in \sigma(X),\ B \in \sigma(Y)} |P(A)P(B) - P(AB)|.$$

Proposition 2.5 below presents the bound for $\mathrm{Cov}(X, Y)$ in terms of the dependence measure α. Its proof can be found in §1.2.2 of Doukhan (1994).

Proposition 2.5 *(i) If $E\{|X|^p + |Y|^q\} < \infty$ for some $p, q \ge 1$ and $1/p + 1/q < 1$, it holds that*

$$|\mathrm{Cov}(X, Y)| \le 8\,\alpha^{1/r}\{E|X|^p\}^{1/p}\{E|Y|^q\}^{1/q},$$

where $r = (1 - 1/p - 1/q)^{-1}$.

(ii) If $P(|X| \le C_1) = 1$ *and* $P(|Y| \le C_2) = 1$ *for some constants* C_1 *and* C_2*, it holds that*

$$|\mathrm{Cov}(X, Y)| \le 4\,\alpha\, C_1 C_2.$$

Note that if we allow X and Y to be complex-valued random variables, Proposition 2.5(ii) still holds, with the coefficient "4" on the right-hand side of the inequality replaced by "16". Using this modified inequality $(k-1)$ times, we obtain the following proposition, which plays an important role in the proof of central limit theorems for mixing sequences. The result was first proved by Volkonskii and Rozanov (1959).

Proposition 2.6 *Let* $\mathcal{F}_i^j$ *and* $\alpha(\cdot)$ *be the same as in (2.57). Let* $\xi_1, \cdots, \xi_k$ *be complex-valued random variables measurable with respect to the* σ*- algebras* $\mathcal{F}_{i_1}^{j_1}, \cdots, \mathcal{F}_{i_k}^{j_k}$*, respectively. Suppose* $i_{l+1} - j_l \ge n$ *for* $l = 1, \cdots, k-1$*, and* $j_l \ge i_l$ *and* $P(|\xi_l| \le 1) = 1$ *for* $l = 1, \cdots, k$*. Then*

$$|E(\xi_1 \cdots \xi_k) - E(\xi_1) \cdots E(\xi_k)| \le 16(k-1)\alpha(n).$$

(b) *Moment inequalities*

Let $\{X_t\}$ be a sequence of random variables with mean 0. For any integers $r \ge 0$ and $q \ge 2$, define

$$M_{r,q} = \sup |\mathrm{Cov}(X_{t_1} \cdots X_{t_p},\ X_{t_{p+1}} \cdots X_{t_q})|, \tag{2.60}$$

where the supremum is taken over all $1 \le t_1 \le \cdots \le t_q$ and $1 \le p < q$ with $t_{p+1} - t_p = r$. The proposition below provides the bounds for the moments of the partial sum $S_n \equiv X_1 + \cdots + X_n$.

Theorem 2.17 *If, for some fixed* $q \ge 2$*,* $M_{r,q} = O(r^{-q/2})$ *as* $r \to \infty$*, then there exists a positive constant* C *independent of* n *for which*

$$|E(S_n^q)| \le C n^{q/2}. \tag{2.61}$$

The theorem above is Theorem 1 of Doukhan and Louhichi (1999). Proposition 2.7 below specifies some conditions under which (2.61) holds for α-mixing processes. A sharper condition can be found in Lemma 7 of Doukhan and Louhichi (1999).

Proposition 2.7 *Let* $\{X_t\}$ *be a strictly stationary and* α*-mixing process with mean* 0*. Let* $\alpha(\cdot)$ *be the mixing coefficient defined in (2.57) and* $q \ge 2$*. Then (2.61) holds if one of the following two conditions holds:*

(i) $E|X_t|^\delta < \infty$ *for some* $\delta > q$*, and* $\alpha(n) = O(n^{-\frac{\delta q}{2(\delta - q)}})$*,*
(ii) $P(|X_t| < C_1) = 1$ *for some constant* C_1*, and* $\alpha(n) = O(n^{-q/2})$*.*

Proof. We give a proof for case (i) only. It follows from Proposition 2.5(i) that

$$\begin{aligned} & |\mathrm{Cov}(X_{t_1} \cdots X_{t_p},\ X_{t_p+1} \cdots X_{t_q})| \\ \le\ & \alpha(r)^{1-\frac{q}{\delta}} \{E|X_{t_1} \cdots X_{t_p}|^{\frac{\delta}{p}}\}^{\frac{p}{\delta}} \{E|X_{t_p+1} \cdots X_{t_q}|^{\frac{\delta}{q-p}}\}^{\frac{q-p}{\delta}}. \end{aligned}$$

By using the Hölder inequality successively, we have that

$$\begin{aligned}\{E|X_{t_1}\cdots X_{t_p}|^{\frac{\delta}{p}}\}^{\frac{p}{\delta}} &\leq \{E|X_{t_1}|^{\delta}\}^{\frac{1}{\delta}}\{E|X_{t_2}\cdots X_{t_p}|^{\frac{\delta}{p-1}}\}^{\frac{p-1}{\delta}} \\ &\leq \cdots \leq \{E|X_1|^{\delta}\}^{\frac{p}{\delta}}.\end{aligned}$$

Thus

$$|\mathrm{Cov}(X_{t_1}\cdots X_{t_p},\ X_{t_p+1}\cdots X_{t_q})| \leq \alpha(r)^{1-\frac{q}{\delta}}\{E|X_1|^{\delta}\}^{\frac{q}{\delta}}.$$

This implies that $M_{r,q} \leq \alpha(r)^{1-q/\delta}\{E|X_1|^{\delta}\}^{q/\delta}$. Now (2.61) follows from Theorem 2.17, and condition $\alpha(r) = O\left(r^{-\frac{\delta q}{2(\delta-q)}}\right)$. ∎

(c) *Exponential inequalities*

Let $\{X_t\}$ be a strictly stationary process with mean 0 and $S_n = X_1 + \cdots + X_n$. Let $\alpha(\cdot)$ be defined as in (2.57). The two theorems below follow from Theorems 1.3 and 1.4 of Bosq (1998) directly.

Theorem 2.18 *Suppose that* $P(|X_t| \leq b) = 1$. *Then*
(i) For each $q = 1, \cdots, [n/2]$ *and* $\varepsilon > 0$,

$$P(|S_n| > n\varepsilon) \leq 4\exp\left(-\frac{\varepsilon^2 q}{8b^2}\right) + 22\left(1+\frac{4b}{\varepsilon}\right)^{\frac{1}{2}} q\,\alpha\left(\left[\frac{n}{2q}\right]\right).$$

(ii) For each $q = 1, \cdots, [n/2]$ *and* $\varepsilon > 0$,

$$P(|S_n| > n\varepsilon) \leq 4\exp\left(-\frac{\varepsilon^2 q}{8\nu^2(q)}\right) + 22\left(1+\frac{4b}{\varepsilon}\right)^{\frac{1}{2}} q\,\alpha\left(\left[\frac{n}{2q}\right]\right),$$

where $\nu^2(q) = 2\sigma^2(q)/p^2 + b\varepsilon/2$, $p = n/(2q)$, *and*

$$\begin{aligned}\sigma^2(q) &= \max_{0\leq j\leq 2q-1} E\{([jp]+1-jp)X_1 + X_2 + \cdots \\ &+ X_{[(j+1)p]-[jp]} + (jp+p-[jp+p])X_{[(j+1)p]-[jp]+1}\}^2.\end{aligned}$$

Theorem 2.19 *Suppose that Cramer's condition is fulfilled, that is, for some constant* $C > 0$,

$$E|X_t|^k \leq C^{k-2}k!\,EX_t^2 < \infty, \quad k = 3, 4, \cdots. \tag{2.62}$$

Then, for any $n \geq 2$, $k \geq 3$, $q \in [1, n/2]$, *and* $\varepsilon > 0$,

$$\begin{aligned}P(|S_n| > n\varepsilon) &\leq 2\{1 + n/q + \mu(\varepsilon)\}\,e^{-q\mu(\varepsilon)} \\ &+ 11n\{1 + 5\varepsilon^{-1}(EX_t^k)^{\frac{1}{2k+1}}\}\,\alpha\left(\left[\frac{n}{q+1}\right]\right)^{\frac{2k}{2k+1}},\end{aligned}$$

where $\mu(\varepsilon) = \varepsilon^2/(25EX_t^2 + 5C\varepsilon)$.

2.6.3 Limit Theorems for α-Mixing Processes

Let $\{X_t\}$ be a strictly stationary and α-mixing process. Define $S_n = X_1 + \cdots + X_n$. Let $\gamma(\cdot)$ be the ACVF of $\{X_t\}$ whenever it exists.

Proposition 2.8 *Suppose that $E|X_t| < \infty$. Then as $n \to \infty$, $S_n/n \xrightarrow{a.s.} EX_t$.*

Proposition 2.8 is an *ergodic theorem* for α-mixing processes. It follows from the fact that an α-mixing sequence is mixing in the sense of ergodic theory; see Theorem 10.2.1 of Doob (1953) and also Theorem 17.1.1 of Ibragimov and Linnik (1971).

Theorem 2.20 *Suppose that one of the following two conditions holds:*
(i) $E|X_t|^\delta < \infty$ and $\sum_{j\geq 1} \alpha(j)^{1-2/\delta} < \infty$ for some constant $\delta > 2$,
(ii) $P(|X_t| < C) = 1$ for some constant $C > 0$, and $\sum_{j\geq 1} \alpha(j) < \infty$.
Then $\sum_{j\geq 1} |\gamma(j)| < \infty$, and

$$\frac{1}{n}\mathrm{Var}(S_n) \to \gamma(0) + 2\sum_{j=1}^{\infty} \gamma(j). \tag{2.63}$$

The proof for case (i) can be found in §1.5 of Bosq (1998) (where the law of the iterated logarithm for α-mixing processes is also presented). We present the proof for case (ii) below.

Proof of Theorem 2.20(ii). It follows from Proposition 2.5(ii) that $|\gamma(j)| \leq 4\alpha(j)\{E|X_1|\}^2$. This implies that $\sum_j |\gamma(j)| < \infty$. For any $n \geq 2$,

$$\begin{aligned}
\frac{1}{n}\mathrm{Var}(S_n) &= \frac{1}{n}\sum_{j=1}^{n} \mathrm{Var}(X_j) + \frac{2}{n}\sum_{1\leq i<j\leq n} \mathrm{Cov}(X_i, X_j) \\
&= \gamma(0) + 2\sum_{l=1}^{n-1} \left(1 - \frac{l}{n}\right)\gamma(l).
\end{aligned}$$

Now (2.63) follows from the dominated convergence theorem. ∎

Theorem 18.4.1 of Ibragimov and Linnik (1971) specified the necessary and sufficient conditions of the *central limit theorem* (CLT) for α-mixing processes. Peligrad (1986) and §1.5 of Doukhan (1994) provided collections of the CLTs under different conditions for α- and other mixing processes. The result presented in Theorem 2.21 below follows directly from Theorem 1.7 of Peligrad (1986). The key idea in the proof of CLTs for dependent processes is to adopt the standard small-block and large-block arguments due to Bernstein (1926); see the proof for Theorem 18.4.1 of Ibragimov and Linnik (1971). We also attach a proof for part of Theorem 2.21 below to illustrate this key idea.

Theorem 2.21 *Assume that* $EX_t = 0$, *and* $\sigma^2 = \gamma(0) + 2\sum_{j=1}^{\infty}\gamma(j)$ *is positive. Then*

$$S_n/\sqrt{n} \xrightarrow{D} N(0, \sigma^2)$$

if one of the following two conditions holds:
(i) $E|X_t|^\delta < \infty$ *and* $\sum_{j\geq 1}\alpha(j)^{1-2/\delta} < \infty$ *for some constant* $\delta > 2$*;*
(ii) $P(|X_t| < c) = 1$ *for some constant* $c > 0$ *and* $\sum_{j\geq 1}\alpha(j) < \infty$.

Proof. We only present the proof for case (ii). To employ the small-block and large-block arguments, we partition the set $\{1, \cdots, n\}$ into $2k_n + 1$ subsets with large blocks of size l_n and small blocks of size s_n and the last remaining set of size $n - k_n(l_n + s_n)$, where l_n and s_n are selected such that

$$s_n \to \infty, \quad s_n/l_n \to 0, \quad l_n/n \to 0, \quad \text{and} \quad k_n \equiv [n/(l_n + s_n)] = O(s_n).$$

For example, we may choose $l_n = O(n^{\frac{r-1}{r}})$ and $s_n = O(n^{1/r})$ for any $r > 2$. Then $k_n = O(n^{1/r}) = O(s_n)$. For $j = 1, \cdots, k_n$, define

$$\xi_j = \sum_{i=(j-1)(l_n+s_n)+1}^{jl_n+(j-1)s_n} X_i, \qquad \eta_j = \sum_{i=jl_n+(j-1)s_n+1}^{j(l_n+s_n)} X_i,$$

and $\zeta = \sum_{i=k_n(l_n+s_n)+1}^{n} X_i$. Note that $\alpha(n) = o(n^{-1})$ and $k_n s_n/n \to 0$. It follows from Proposition 2.7(ii) that

$$\frac{1}{n}E\left(\sum_{j=1}^{k_n}\eta_j\right)^2 \to 0, \quad \frac{1}{n}E\zeta^2 \to 0.$$

Thus

$$\frac{1}{\sqrt{n}}S_n = \frac{1}{\sqrt{n}}\left\{\sum_{j=1}^{k_n}\xi_j + \sum_{j=1}^{k_n}\eta_j + \zeta\right\} = \frac{1}{\sqrt{n}}\sum_{j=1}^{k_n}\xi_j + o_p(1). \tag{2.64}$$

It follows from Proposition 2.6 that as $n \to \infty$

$$\begin{aligned} & \left| E\left\{\exp\left(\frac{it}{\sqrt{n}}\sum_{j=1}^{k_n}\xi_j\right)\right\} - \prod_{j=1}^{k_n} E\{\exp(it\xi_j/\sqrt{n})\}\right| \\ \leq \;& 16\,(k_n - 1)\,\alpha(s_n) \;\to\; 0. \end{aligned} \tag{2.65}$$

Now, applying Theorem 2.20(ii), we have $l_n^{-1}E\xi_1^2 \to \sigma^2$, which implies the *Feller condition*

$$\frac{1}{n}\sum_{j=1}^{k_n}E\xi_j^2 = \frac{k_n\, l_n}{n}\cdot\frac{1}{l_n}E\xi_1^2 \to \sigma^2.$$

By Proposition 2.7(ii),

$$\begin{aligned} E\{\xi_1^2 I(|\xi_1| \geq \varepsilon\sigma n^{\frac{1}{2}})\} &\leq \{E\xi_1^4\}^{\frac{1}{2}}\, P\{|\xi_1| \geq \varepsilon\sigma n^{\frac{1}{2}}\} \\ &\leq Cl_n \cdot \frac{1}{n\varepsilon^2\sigma^2} E\xi_1^2 = O(l_n^2/n). \end{aligned}$$

Consequently,

$$\frac{1}{n}\sum_{j=1}^{k_n} E\{\xi_j^2 I(|\xi_j| \geq \varepsilon\sigma n^{\frac{1}{2}})\} = O(k_n l_n^2/n^2) = O(l_n/n) \to 0,$$

which is the *Lindberg condition*. Using the standard argument for the proof of CLTs (see, for example, p. 315 of Chow and Teicher 1997), we have

$$\prod_{j=1}^{k_n} E\{\exp(it\xi_j/\sqrt{n})\} \to e^{-t^2\sigma^2/2}.$$

(Note that the convergence above would follow directly from the CLT for the sum of independent random variables if the random variables $\{\xi_j\}$ were independent.) This, together with (2.65) and (2.64), entails the required CLT. ∎

2.6.4 A Central Limit Theorem for Nonparametric Regression

In this section, we present a central limit theorem that can be used directly to derive the asymptotic distributions of nonparametric regression estimators based on kernel smoothing for α-mixing processes.

Let $\{(e_t, X_t)\}$ be a two-dimensional stochastic process. Let x be a fixed real number, $W(\cdot)$ be a given function, and $h = h(n) > 0$ be a constant depending on n. Define the triangular array

$$Y_{t,n} \equiv Y_{t,n}(x) = e_t\, W\Big(\frac{X_t - x}{h}\Big), \quad t = 1, \cdots, n;\ n \geq 1. \tag{2.66}$$

We will establish the central limit theorem for the partial sum

$$S_n(x) = \sum_{t=1}^{n} Y_{t,n} \tag{2.67}$$

under the following regularity conditions.

> (C1) $\{(e_t, X_t)\}$ is a strictly stationary process with $E(e_t|X_t) = 0$, $E(e_t^2|X_t) = \sigma(X_t)^2$, and $E(|e_t|^\delta) < \infty$ for some $\delta > 2$. Furthermore, the function $\sigma(\cdot)^2$ and the marginal density function $p(\cdot)$ of X_t are continuous at x.

(C2) The conditional density function of (X_1, X_j) given (e_1, e_j) is bounded by a positive constant C_0 independent of $j > 1$.

(C3) $\{(e_t, X_t)\}$ is α-mixing with the mixing coefficients satisfying the condition $\sum_{t\geq 1} t^{\lambda}\alpha(t)^{1-2/\delta} < \infty$ for some $\lambda > 1 - 2/\delta$.

(C4) $W(\cdot)$ is a bounded function, and $\int |W(u)|^k du < \infty$ for $k = 1, 2$.

(C5) $n \to \infty$, $h \to 0$, and $nh^{\frac{\lambda+2-2/\delta}{\lambda+2/\delta}} = O(n^{\epsilon_o})$ for some constant $\epsilon_o > 0$.

Conditions (C1)–(C5) are standard in nonparametric regression. In particular, (C5) is implied by the condition that $nh^3 \to \infty$ and $h \to 0$ (since $\lambda > 1 - 2/\delta$). Note that although the mixing conditions were introduced for univariate processes in §2.6.1, they are readily applicable to the case where $X_t = \mathbf{X}_t$ is a vector-valued process.

The partial sums of forms (2.67) and (2.66) are constantly encountered in nonparametric regression estimation. They differ from the conventional partial sums (such as those treated in Theorem 2.21) in two aspects. First, each $Y_{t,n}$ depends on n through $h = h(n)$. Furthermore, due to the localization dictated by $W(\cdot/h)$, only the terms with X_t close to x on the right-hand side of (2.67) are effectively counted asymptotically. This changes the convergence rate from the standard $n^{1/2}$ to $(nh)^{1/2}$. For further discussion on this type of localization, see Chapters 5 and 6.

Due to the differences stated above, the limit theorems such as Proposition 2.8 and Theorem 2.21 are not directly applicable to the partial sum $S_n(x)$, although similar results can be established in a similar manner with additional regularity conditions. Note that the (weak) laws of large numbers may be derived relatively easily from the exponential inequalities in Theorems 2.18 and 2.19. We only present a central limit theorem below.

Theorem 2.22 *Under conditions (C1)–(C5), it holds that*

$$\frac{1}{\sqrt{nh}} S_n(x) \xrightarrow{D} N\Big(0,\ \sigma(x)^2 p(x) \int W(u)^2 du\Big).$$

The proof of the theorem above is presented in §2.7.7. It is similar to the proof of Theorem 2.21 in spirit. The conditions and the proof of Theorem 2.22 have been used in Masry and Fan (1997). See also the proof of Theorem 6.3.

2.7 Complements

2.7.1 Proof of Theorem 2.5(i)

It follows from (2.15) that, for any integer $k \geq 1$,

$$\begin{aligned} Y_t &= a\xi_t + \sum_{i=1}^{\infty} b_i \xi_t \xi_{t-i} \rho_{t-i} = a\xi_t + a\sum_{i=1}^{\infty} b_i \xi_t \xi_{t-i} + \sum_{i,j=1}^{\infty} b_i b_j \xi_t \xi_{t-i} Y_{t-i-j} \\ &= a\xi_t + a\sum_{l=1}^{k} \sum_{1 \leq j_1, \cdots, j_l < \infty} b_{j_1} \cdots b_{j_l} \xi_t \xi_{t-j_1} \cdots \xi_{t-j_1-\cdots-j_l} \\ &+ \sum_{1 \leq j_1, \cdots, j_{k+1} < \infty} b_{j_1} \cdots b_{j_{k+1}} \xi_t \xi_{t-j_1} \cdots \xi_{t-j_1-\cdots-j_k} Y_{t-j_1-\cdots j_{k+1}}. \quad (2.68) \end{aligned}$$

Define

$$Y_t' = a\xi_t + a\sum_{l=1}^{\infty} \sum_{1 \leq j_1, \cdots, j_l < \infty} b_{j_1} \cdots b_{j_l} \xi_t \xi_{t-j_1} \cdots \xi_{t-j_1-\cdots-j_l}.$$

Note that all of the terms on the right-hand side of the expression above are nonnegative, and for any $l \geq 1$,

$$\begin{aligned} & E\left\{ \sum_{1 \leq j_1, \cdots, j_l < \infty} b_{j_1} \cdots b_{j_l} \xi_t \xi_{t-j_1} \cdots \xi_{t-j_1-\cdots-j_l} \right\} \\ = & \sum_{1 \leq j_1, \cdots, j_l < \infty} b_{j_1} \cdots b_{j_l} = \left\{ \sum_{j=1}^{\infty} b_j \right\}^l. \end{aligned}$$

Thus $0 \leq Y_t' < \infty$ a.s., $E(Y_t') = a/(1 - \sum_j b_j)$, and therefore $\{Y_t'\}$ is strictly stationary. It is easy to verify that Y_t' fulfills (2.15).

To prove the uniqueness, let $\{Y_t\}$ be a strictly stationary solution of (2.15) with $|EY_t| < \infty$. We will show below that $Y_t = Y_t'$ a.s. for any fixed t.

Let t be fixed now. It follows from (2.68) that for any $k \geq 1$

$$\begin{aligned} |Y_t - Y_t'| &\leq \sum_{1 \leq j_1, \cdots, j_{k+1} < \infty} b_{j_1} \cdots b_{j_{k+1}} \xi_t \xi_{t-j_1} \cdots \xi_{t-j_1-\cdots-j_k} |Y_{t-j_1-\cdots j_{k+1}}| \\ &+ a\sum_{l=k+1}^{\infty} \sum_{1 \leq j_1, \cdots, j_l < \infty} b_{j_1} \cdots b_{j_l} \xi_t \xi_{t-j_1} \cdots \xi_{t-j_1-\cdots-j_l}. \end{aligned}$$

The expectation of the right-hand side of the above is not greater than

$$\left\{ E|Y_1| + a/\left(1 - \sum_{i=1}^{\infty} b_i\right) \right\} \left(\sum_{j=1}^{\infty} b_j \right)^{k+1}.$$

Let $A_k = \{|Y_t - Y_t'| > 1/k\}$. Then

$$P(A_k) \le kE|Y_t - Y_t'| \le k\left\{E|Y_1| + a/\left(1 - \sum_{i=1}^{\infty} b_i\right)\right\}\left(\sum_{j=1}^{\infty} b_j\right)^{k+1}.$$

Thus $\sum_{k\ge1} P(A_k) < \infty$. It follows from the Borel–Cantelli lemma (see, e.g., Theorem 3.2.1 in Chow and Teicher 1997) that $P\{A_k,\ i.o.\} = 0$. Since $A_k \subset A_{k+1}$, it holds that $P(A_k) = 0$ for any k (i.e., $Y_t = Y_t'$ a.s.). ∎

2.7.2 Proof of Proposition 2.3(i)

Let $\boldsymbol{\gamma}_{2,k;j} = \mathrm{Cov}(\mathbf{X}_{2,k}, X_j)$. First we derive an explicit expression for $\boldsymbol{\alpha} \equiv (\alpha_2, \cdots, \alpha_k)^\tau$ defined in (2.28) for $j = 1$. Note that for any $\boldsymbol{\beta} \equiv (\beta_2, \cdots, \beta_k)^\tau$,

$$\begin{aligned} E(X_1 - \boldsymbol{\beta}^\tau \mathbf{X}_{2,k})^2 &= E(X_1 - \boldsymbol{\alpha}^\tau \mathbf{X}_{2,k})^2 + E\{(\boldsymbol{\alpha} - \boldsymbol{\beta})^\tau \mathbf{X}_{2,k}\}^2 \\ &+ 2E\{(X_1 - \boldsymbol{\alpha}^\tau \mathbf{X}_{2,k})\mathbf{X}_{2,k}^\tau\}(\boldsymbol{\alpha} - \boldsymbol{\beta}). \end{aligned} \tag{2.69}$$

Hence $E(X_1 - \boldsymbol{\beta}^\tau \mathbf{X}_{2,k})^2 \ge E(X_1 - \boldsymbol{\alpha}^\tau \mathbf{X}_{2,k})^2$ for any $\boldsymbol{\beta}$ if and only if the third term on the right-hand side of the expression above is 0 (for any $\boldsymbol{\beta}$). This is equivalent to

$$E\{(X_1 - \boldsymbol{\alpha}^\tau \mathbf{X}_{2,k})\mathbf{X}_{2,k}^\tau\} = 0. \tag{2.70}$$

This normal equation leads to the least squares solution $\boldsymbol{\alpha} = \boldsymbol{\Sigma}_{2,k}^{-1}\boldsymbol{\gamma}_{2,k;1}$. Hence, we have $R_{1|2,\cdots,k} = X_1 - \boldsymbol{\gamma}_{2,k;1}^\tau \boldsymbol{\Sigma}_{2,k}^{-1}\mathbf{X}_{2,k}$. In the same vein, $R_{k+1|2,\cdots,k} = X_{k+1} - \boldsymbol{\gamma}_{2,k;k+1}^\tau \boldsymbol{\Sigma}_{2,k}^{-1}\mathbf{X}_{2,k}$. It follows from (2.70) that

$$\begin{aligned} \mathrm{Cov}(R_{k+1|2,\cdots,k}, R_{1|2,\cdots,k}) &= \mathrm{Cov}(X_{k+1}, R_{1|2,\cdots,k}) \\ &= \gamma(k) - \boldsymbol{\gamma}_{2,k;1}^\tau \boldsymbol{\Sigma}_{2,k}^{-1}\boldsymbol{\gamma}_{2,k;k+1}. \end{aligned}$$

The conclusion follows from the fact that

$$\mathrm{Var}(R_{1|2,\cdots,k}) = \gamma(0) - \boldsymbol{\gamma}_{2,k;1}^\tau \boldsymbol{\Sigma}_{2,k}^{-1}\boldsymbol{\gamma}_{2,k;1}$$

and $\mathrm{Var}(R_{k+1|2,\cdots,k}) = \mathrm{Var}(R_{1|2,\cdots,k})$. (Note that $\{X_t\}$ is time-reversible as far as its first two moments properties are concerned.) ∎

2.7.3 Proof of Theorem 2.9

It may be shown in terms of a decomposition similar to (2.69) that

$$\begin{pmatrix} b_{11} \\ \vdots \\ b_{kk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{2,k} & \boldsymbol{\gamma}_{2,k;1} \\ \boldsymbol{\gamma}_{2,k;1}^\tau & \gamma(0) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\gamma}_{2,k;k+1} \\ \gamma(k) \end{pmatrix}.$$

It follows from the matrix partition inverse formula (p. 33 of Rao 1973) that

$$\begin{pmatrix} \boldsymbol{\Sigma}_{2,k} & \boldsymbol{\gamma}_{2,k;1} \\ \boldsymbol{\gamma}_{2,k;1}^{\tau} & \gamma(0) \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{2,k}^{-1}(I + \boldsymbol{\gamma}_{2,k;1}\boldsymbol{\gamma}_{2,k;1}^{\tau}\boldsymbol{\Sigma}_{2,k}^{-1})/\nu & -\boldsymbol{\Sigma}_{2,k}^{-1}\boldsymbol{\gamma}_{2,k;1}/\nu \\ -\boldsymbol{\gamma}_{2,k;1}^{\tau}\boldsymbol{\Sigma}_{2,k}^{-1}/\nu & \nu^{-1} \end{pmatrix},$$

where $\nu = \gamma(0) - \boldsymbol{\gamma}_{2,k;1}^{\tau}\boldsymbol{\Sigma}_{2,k}^{-1}\boldsymbol{\gamma}_{2,k;1}$. Combining the two expressions above , we have

$$b_{kk} = \{\gamma(k) - \boldsymbol{\gamma}_{2,k;1}^{\tau}\boldsymbol{\Sigma}_{2,k}^{-1}\boldsymbol{\gamma}_{2,k;k+1}\}/\nu,$$

which is the same as the right-hand side of (2.29). ∎

2.7.4 Proof of Theorem 2.10

Suppose that (2.35) holds. Then for any $a_1, \cdots, a_n \in R$,

$$\begin{aligned} \sum_{j,k=1}^{n} a_j a_k \rho(j-k) &= \int_{-\pi}^{\pi} \sum_{j,k=1}^{n} a_j a_k e^{i\omega(j-k)} dF(\omega) \\ &= \int_{-\pi}^{\pi} |\sum_{j}^{n} a_j e^{i\omega j}|^2 dF(\omega) \geq 0. \end{aligned}$$

It follows from Theorem 2.7 that $\{\rho(k)\}$ is the ACF of a stationary time series.

Conversely, suppose that $\{\rho(k)\}$ is the ACF of a stationary time series $\{X_t\}$. Define, for $\omega \in [-\pi, \pi]$,

$$f_n(\omega) = \frac{1}{2\pi n} \sum_{j,k=1}^{n} e^{-i\omega(j-k)} \rho(j-k) = \frac{1}{2\pi n} \sum_{|m|<n} (n - |m|)\rho(m) e^{-i\omega m}.$$

Then $f_n(\omega) = \mathrm{Var}(\xi) = \mathrm{Cov}(\xi, \bar{\xi}) \geq 0$, where $\xi = \sum_{j=1}^{n} e^{-i\omega j} X_j / \sqrt{2\pi\gamma(0)}$ is a complex-valued random variable and $\bar{\xi}$ denotes its conjugate. Let

$$F_n(\omega) = \int_{-\pi}^{\omega} f_n(\omega) d\omega, \quad \omega \in [-\pi, \pi].$$

Then, for any integer j,

$$\int_{-\pi}^{\pi} e^{ij\omega} dF_n(\omega) = \frac{1}{2\pi} \sum_{|m|<n} (1 - |m|/n)\rho(m) \int_{-\pi}^{\pi} e^{i(j-m)\omega} d\omega.$$

Note that the integral on the right-hand side of the expression above is nonzero (i.e., 2π) if and only if $j = m$. Therefore

$$\int_{-\pi}^{\pi} e^{ij\omega} dF_n(\omega) = \begin{cases} (1 - |j|/n)\rho(j), & |j| < n, \\ 0, & \text{otherwise.} \end{cases} \tag{2.71}$$

Since $\{F_n(\cdot)\}$ is a sequence of probability distribution functions defined on the finite interval $[-\pi, \pi]$, it follows from Helly's selection theorem (see, e.g., Lemma 8.2.2 of Chow and Teicher 1997) that there exists a subsequence of $\{F_n\}$ that converges in distribution to a probability distribution function F. Taking the limit as $n \to \infty$ in (2.71) for that subsequence, we conclude from the Helly–Bray theorem (Corollary 8.1.6 of Chow and Teicher 1997) that

$$\int_{-\pi}^{\pi} e^{ij\omega} dF(\omega) = \rho(j).$$

■

2.7.5 Proof of Theorem 2.13

Note that $\mathbf{e}_0$ is a vector with 1 as all of its components. Hence, for $k \neq 0$,

$$\sum_{t=1}^{T} e^{it\omega_k} = \mathbf{e}_0^\tau \mathbf{e}_k = 0, \quad \sum_{t=1}^{T} e^{-it\omega_k} = \mathbf{e}_k^\tau \mathbf{e}_0 = 0.$$

Therefore

$$\begin{aligned} I_T(\omega_k) &= \frac{1}{T} \sum_{t=1}^{T} X_t e^{-it\omega_k} \sum_{s=1}^{T} X_s e^{is\omega_k} \\ &= \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} (X_t - \bar{X}_T)(X_s - \bar{X}_T) e^{-i(t-s)\omega_k}. \end{aligned}$$

By a change of variable, $\tau = t - s$, and then an exchange of summation, we have

$$\begin{aligned} I(\omega_k) &= \frac{1}{T} \sum_{t=1}^{T} \sum_{\tau=t-1}^{t-T} (X_t - \bar{X}_T)(X_{t-\tau} - \bar{X}_T) e^{-i\tau\omega_k} \\ &= \sum_{\tau=-(T-1)}^{T-1} \frac{1}{T} \sum_{t=1}^{T-|\tau|} (X_t - \bar{X}_T)(X_{t+|\tau|} - \bar{X}_T) e^{-i\tau\omega_k} \\ &= \sum_{\tau=-(T-1)}^{T-1} \widehat{\gamma}(\tau) e^{-i\tau\omega_k}. \end{aligned}$$

■

2.7.6 Proof of Theorem 2.14

First, we prove (i). To simplify the notation, we write $\sum_{j=1}^{r} c_j \xi_{k_j} = \sum_{l=1}^{2n} b_l \xi_l$, where $b_l = c_j$ for $l = k_j$ and 0 otherwise. It is easy to see that $E(\sum_{l=1}^{2n} b_l \xi_l) =$

0 and

$$\sum_{j=1}^{r} c_j \xi_{k_j} = \frac{\sqrt{2}}{\sqrt{T}\,\sigma} \sum_{k=1}^{T} \varepsilon_k \sum_{j=1}^{n} \{b_{2j-1} \cos(\omega_j k) + b_{2j} \sin(\omega_j k)\}.$$

This is a sequence of the linear combinations of independent random variables, and the central limit theorem (CLT) will be employed. We now calculate the variance of the sum above . To this end, note that

$$\frac{1}{T} \sum_{k=1}^{T} \cos(\omega_j k) \cos(\omega_l k) = (\mathbf{e}_j + \mathbf{e}_{-j})^\tau (\mathbf{e}_l + \mathbf{e}_{-l})/4 = \delta_{j,l}/2,$$

where $\mathbf{e}_j$ is defined as in (2.49), and $\delta_{j,l} = 1$ if $j = l$ and 0 otherwise. In the same vein, we have

$$\frac{1}{T} \sum_{k=1}^{T} \sin(\omega_j k) \sin(\omega_l k) = \delta_{j,l}/2, \quad \frac{1}{T} \sum_{k=1}^{T} \cos(\omega_j k) \sin(\omega_l k) = 0.$$

Hence

$$\begin{aligned} & \mathrm{Var}\left(\sum_{j=1}^{r} c_j \xi_{k_j} \right) \\ = \; & \frac{2}{T\sigma^2} \mathrm{Var}\left(\sum_{k=1}^{T} \varepsilon_k \sum_{j=1}^{n} \{b_{2j-1} \cos(\omega_j k) + b_{2j} \sin(\omega_j k)\} \right) \\ = \; & \frac{2}{T} \sum_{k=1}^{T} \left(\sum_{j=1}^{n} \{b_{2j-1} \cos(\omega_j k) + b_{2j} \sin(\omega_j k)\} \right)^2 \\ = \; & \sum_{j=1}^{2n} b_j^2 = \sum_{j=1}^{r} c_j^2. \end{aligned}$$

Write $d_k = \sum_{j=1}^{n} \{b_{2j-1} \cos(\omega_j k) + b_{2j} \sin(\omega_j k)\}$. It is easy to see that $|d_k| \le r \max_j |c_j|$ for all $1 \le k \le 2n$. Hence, for any $\eta > 0$,

$$\frac{1}{T} \sum_{k=1}^{T} E\{\varepsilon_k^2 d_k^2 I(|\varepsilon_k d_k| > \eta T^{1/2})\} \le C_2 E\{\varepsilon_1^2 I(|\varepsilon_1| > \eta T^{1/2}/C_1)\} \to 0,$$

where C_1 and C_2 are some positive constants. It follows from the CLT for double arrays of random variables (see, e.g. p. 31 of Serfling 1980) that $\sum_{j=1}^{r} c_j \xi_{k_j} \xrightarrow{D} N(0, \sum_{j=1}^{r} c_j^2)$.

To prove (ii), note that the discrete Fourier transform of $\{X_t\}$ can be expressed as

$$\begin{aligned}
\alpha_k &= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t e^{-i\omega_k t} \\
&= \frac{1}{\sqrt{T}} \sum_{j=-\infty}^{\infty} a_j e^{-i\omega_k j} \sum_{t=1}^{T} \varepsilon_{t-j} e^{-i\omega_k (t-j)} \\
&= \frac{1}{\sqrt{T}} \sum_{j=-\infty}^{\infty} a_j e^{-i\omega_k j} \left(\sum_{t=1}^{T} \varepsilon_t e^{-i\omega_k t} + U_{Tj} \right) \\
&= a(e^{-i\omega_k}) \alpha_{k,\varepsilon} + Y_T(\omega_k),
\end{aligned}$$

where $a(z) = \sum_{j=-\infty}^{\infty} a_j z^j$, $\alpha_{k,\varepsilon} = T^{-1/2} \sum_{t=1}^{T} \varepsilon_t e^{-i\omega_k t}$ is the discrete Fourier transform of $\{\varepsilon_t\}$, and

$$U_{Tj} = \sum_{t=1-j}^{T-j} \varepsilon_t e^{-i\omega_k t} - \sum_{t=1}^{T} \varepsilon_t e^{-i\omega_k t}, \quad Y_T(\omega_k) = T^{-1/2} \sum_{j=-\infty}^{\infty} a_j e^{-i\omega_k j} U_{Tj}.$$

Note that $|\alpha_{k,\varepsilon}|^2 = \sigma^2(\xi_{2k-1}^2 + \xi_{2k}^2)/2$. Hence

$$\begin{aligned}
I_T(\omega_k) &= |\alpha_k|^2 = |a(e^{-iw_k})|^2 |\alpha_{k,\varepsilon}|^2 + R_T(\omega_k) \\
&= 2\pi g(\omega_k)(\xi_{2k-1}^2 + \xi_{2k}^2) + R_T(\omega_k),
\end{aligned}$$

where $g(\omega_k) = |a(e^{-iw_k})|^2 \sigma^2/(2\pi)$ is the spectral density function of $\{X_t\}$ (see Theorem 2.12), and

$$R_T(\omega_k) = |Y_T(\omega_k)|^2 + a(e^{-i\omega_k}) \alpha_{k,\varepsilon} Y_T(-\omega_k) + a(e^{i\omega_k}) \bar{\alpha}_{k,\varepsilon} Y_T(\omega_k). \quad (2.72)$$

Note that if $|j| < T$, U_{Tj} is a sum of $2|j|$ independent random variables, whereas if $|j| \geq T$, U_{Tj} is a sum of $2T$ independent random variables. Thus, $E|U_{Tj}|^2 \leq 2\min(|j|, T)\sigma^2$. Therefore, for any fixed positive integer l and $T > l$,

$$\begin{aligned}
E|Y_T(\omega_k)|^2 &\leq \frac{1}{T} \left(\sum_{j=-\infty}^{\infty} |a_j| (EU_{Tj}^2)^{1/2} \right)^2 \\
&\leq \frac{2\sigma^2}{T} \left(\sum_{j=-\infty}^{\infty} |a_j| \{\min(|j|, T)\}^{1/2} \right)^2 \\
&\leq 2\sigma^2 \left(\frac{1}{\sqrt{T}} \sum_{|j| \leq l} |a_j| \ |j|^{1/2} + \sum_{|j|>l} |a_j| \right)^2 .
\end{aligned}$$

Note that the right-hand side of the expression above is independent of k and can be smaller than any given positive constant (by choosing l large enough accordingly) as $T \to \infty$. Hence, $\max_{1\le k\le n} E|Y_T(\omega)|^2 \to 0$. On the other hand, $|a(e^{\pm i\omega_k})| \le \sum_j |a_j| < \infty$ and $E|\alpha_{k,\varepsilon}|^2 = \sigma^2$. Application of the Cauchy–Schwartz inequality to (2.72) gives $\max_{1\le k\le n} E|R_T(\omega_k)| \to 0$. ∎

2.7.7 Proof of Theorem 2.22

First, we calculate the variance of $S_n(x)$. Let $Z_t = Y_{t,n}/\sqrt{h}$. It is easy to see that $E(Z_t) = 0$, and

$$\frac{1}{nh}\mathrm{Var}\{S_n(x)\} = E(Z_1^2) + 2\sum_{j=1}^{n-1}(1 - j/n)E(Z_1 Z_{j+1}).$$

Condition (C1) implies that

$$\begin{aligned} E(Z_1^2) &= \frac{1}{h}\int E(e_1^2|X_1 = y)p(y)W\Big(\frac{y-x}{h}\Big)^2 dy \\ &= \int \sigma(x+hu)^2 p(x+hu)W(u)^2 du \\ &\to \sigma(x)^2 p(x)\int W(u)^2 du \;\equiv\; \nu(x), \end{aligned} \tag{2.73}$$

as $h \to 0$. By conditioning on (e_1, e_{j+1}), it follows from (C2) that

$$\begin{aligned} |E(Z_1 Z_{j+1})| &= \frac{1}{h}\left|E\left\{e_1 e_{j+1} W\left(\frac{X_1 - x}{h}\right) W\left(\frac{X_{j+1} - x}{h}\right)\right\}\right| \\ &\le C_0 h^{-1} E|e_1 e_{j+1}|\Big\{\int W\left(\frac{y-x}{h}\right) dy\Big\}^2 \\ &\le C_0 h\, E(e_1^2)\Big\{\int W(u)du\Big\}^2 \;=\; O(h). \end{aligned}$$

Therefore

$$\left|\sum_{j=1}^{m_n} E(Z_1 Z_{j+1})\right| = O(m_n h). \tag{2.74}$$

By Proposition 2.5(i),

$$|E(Z_1 Z_{j+1})| \le C\alpha(j)^{1-2/\delta} h^{2/\delta - 1}.$$

Let $m_n = [\frac{1}{h|\log h|}]$. Then $m_n \to \infty$, $m_n h \to 0$, and

$$\sum_{j=m_n+1}^{n-1} |E(Z_1 Z_{j+1})| \le C\frac{h^{2/\delta-1}}{m_n^\lambda} \sum_{j=m_n+1}^{n} j^\lambda \alpha(j)^{1-2/\delta} \to 0;$$

see condition (C3). Combining this with (2.74), we have

$$\sum_{j=1}^{n-1} E(Z_1 Z_{j+1}) \to 0. \tag{2.75}$$

Now, it follows from (2.73) and (2.75) that

$$\frac{1}{nh}\mathrm{Var}\{S_n(x)\} = \nu(x)\{1 + o(1)\}. \tag{2.76}$$

To prove the CLT, we employ the small-block and large-block arguments as follows. We partition the set $\{1, \cdots, n\}$ into $2k_n + 1$ subsets with large blocks of size l_n, small blocks of size s_n, and the last remaining set of size $n - k_n(l_n + s_n)$ and write accordingly

$$S_n(x) = \sum_{j=1}^{k_n} \xi_j + \sum_{j=1}^{k_n} \eta_j + \zeta, \tag{2.77}$$

where

$$\xi_j = \sum_{i=(j-1)(l_n+s_n)+1}^{jl_n+(j-1)s_n} Y_{i,n}, \qquad \eta_j = \sum_{i=jl_n+(j-1)s_n+1}^{j(l_n+s_n)} Y_{i,n},$$

and $\zeta = \sum_{i=k_n(l_n+s_n)+1}^{n} Y_{i,n}$. Put

$$l_n = \big[\sqrt{nh}/\log n\big], \qquad s_n = \big[\big(\sqrt{n/h}\,\log n\big)^{\frac{1-2/\delta}{\lambda+1}}\big].$$

It follows from condition (C5) that $s_n/l_n \to 0$. Therefore

$$k_n = [n/(l_n + s_n)] = O(\sqrt{n/h}\,\log n).$$

Note that condition (C3) implies $n^{\frac{\lambda+1}{1-2/\delta}}\alpha(n) \to 0$. Hence

$$k_n \alpha(s_n) \to 0. \tag{2.78}$$

It follows from (2.76) that

$$\frac{1}{nh}E(\zeta^2) \le \frac{l_n + s_n}{n} \cdot \frac{1}{(l_n + s_n)h}\mathrm{Var}(\zeta) \to 0,$$

and $E(\eta_j^2) = s_n h \nu(x)\{1 + o(1)\}$. Hence

$$\begin{aligned}
\frac{1}{nh}E\left(\sum_{j=1}^{k_n}\eta_j\right)^2 &= \frac{k_n s_n}{n}\nu(x)\{1 + o(1)\} + \frac{1}{h}\left|\sum_{j=1}^{k_n-1}\mathrm{Cov}(\eta_1, \eta_{j+1})\right| \\
&\le \frac{k_n s_n}{n}\nu(x)\{1 + o(1)\} + \sum_{j=1}^{n-1}|E(Z_1 Z_{j+1})| \\
&\to 0.
\end{aligned} \tag{2.79}$$

The limit on the right-hand side of (2.79) makes use of (2.75). Now, by (2.77),

$$\frac{1}{\sqrt{nh}} S_n(x) = \frac{1}{\sqrt{nh}} \sum_{j=1}^{k_n} \xi_j + o_p(1) \equiv \frac{1}{\sqrt{nh}} Q_n + o_p(1). \tag{2.80}$$

Similar to (2.79), it holds that

$$\frac{1}{nh} \mathrm{Var}(Q_n) = \frac{1}{nh} E(Q_n^2) = \frac{k_n l_n}{n} \nu(x) + o(1) \to \nu(x). \tag{2.81}$$

We employ a truncation argument now. Write $e_t^L = e_t I(|e_t| \le L)$ and $e_t^R = e_t I(|e_t| > L)$ for a fixed constant $L > 0$. Write

$$Q_n^L = \sum_{j=1}^{k_n} \xi_j^L, \qquad Q_n^R = \sum_{j=1}^{k_n} \xi_j^R,$$

where ξ_j^L and ξ_j^R are defined in the same manner as ξ_j with e_t replaced, respectively, by e_t^L and e_t^R. Similar to (2.81), we have that

$$\frac{1}{nh} \mathrm{Var}(Q_n^L) \to \mathrm{Var}\{e_1 I(|e_1| \le L) | X_1 = x\} p(x) \int W(u)^2 du \equiv \nu_L(x) \tag{2.82}$$

and

$$\frac{1}{nh} \mathrm{Var}(Q_n^R) \to \mathrm{Var}\{e_1 I(|e_1| > L) | X_1 = x\} p(x) \int W(u)^2 du. \tag{2.83}$$

Define

$$M_n = \left| E \exp(itQ_n/\sqrt{nh}) - \exp\{-t^2 \nu(x)/2\} \right|,$$

where $i = \sqrt{-1}$ now. Then, the required result follows from the statement that

$$\lim_{n \to \infty} M_n < \epsilon \tag{2.84}$$

for any given $\epsilon > 0$. Note that

$$\begin{aligned} M_n &\le E \left| \exp(itQ_n^L/\sqrt{nh}) \{\exp(itQ_n^R/\sqrt{nh}) - 1\} \right| \\ &+ \left| E \exp(itQ_n^L/\sqrt{nh}) - \prod_{j=1}^{k_n} E(it\xi_j^L/\sqrt{nh}) \right| \\ &+ \left| \prod_{j=1}^{k_n} E(it\xi_j^L/\sqrt{nh}) - \exp\{-t^2 \nu_L(x)/2\} \right| \\ &+ \left| \exp\{-t^2 \nu_L(x)/2\} - \exp\{-t^2 \nu(x)/2\} \right|. \end{aligned}$$

Note that the first term on the right-hand side of the expression above is bounded by

$$E\big|\exp(itQ_n^R/\sqrt{nh}) - 1\big| = O\{\mathrm{Var}(Q_n^R)/(nh)\}$$

which may be smaller than $\epsilon/2$ by choosing large L ; see (2.83). The last term may also be smaller than $\epsilon/2$ by choosing large L as well; see (2.82). By Proposition 2.6, the second term is bounded by $16(k_n - 1)\alpha(s_n)$, which converges to 0 due to (2.78). To prove that the third term converges to 0 is equivalent to proving that

$$\frac{1}{\sqrt{nh}}\, Q_n^L \xrightarrow{D} N\big(0, \nu_L(x)\big)$$

while treating $\{\xi_j^L\}$ as a sequence of independent random variables. The latter is implied by the Lindberg condition

$$\frac{1}{nh}\sum_{j=1}^{k_n} E\big[(\xi_j^L)^2 I\{|\xi_j^L| > \omega\nu_L(x)\sqrt{nh}\}\big] \to 0$$

for any $\omega > 0$; see, for example, p. 315 of Chow and Teicher (1997). Note that $l_n/\sqrt{nh} \to 0$. When n is large enough, $\{|\xi_j^L| > \omega\nu^L(x)\sqrt{nh}\}$ is an empty set for all j. Hence the limit above holds. Therefore, we have shown that (2.83) holds for any $\epsilon > 0$. The proof is completed. ∎

2.8 Additional Bibliographical Notes

The literature on strict stationarity and ergodicity of nonlinear time series (2.7) may be divided into two categories: general cases and special cases. In addition to those presented in Theorem 2.4, Tweedie (1975, 1976), Nummelin (1984), Chan and Tong (1985), Chan (1990a), Tjøstheim (1990), Meyn and Tweedie (1993, 1994) derived various useful tools to identify the (geometric) ergodicity of model (2.7). The research on ergodicity for some individual models includes Petruccelli and Woolford (1984), Chan, Petruccelli, Tong and Woolford (1985), Chen and Tsay (1991) and Guo and Petruccelli (1991)on various TAR models and Chen and Tsay (1993) for FAR models.

The general framework leading to model (2.15) was introduced by Robinson (1991b). Work on stationarity on ARCH and GARCH models includes, among others, Nelson (1990), Bougerol and Picard(1992a, b), Nelson and Cao (1992), and Kokoszka and Leipus (2000).

Dahlhaus (1997) introduced the class of *local stationary time series* in terms of a spectral representation. The class provides, for example, an arbitrarily close approximation to an AR model with the coefficients varying

with respect to time; see also Neumann and von Sachs (1997) and Adak (1998). The time-domain approach for local stationarity can be traced back at least to Ozaki and Tong (1975) and Kitagawa and Akaike (1978). They proposed the concept of *interval-wise* stationarity, which divides a time series into several time intervals and fits a stationary model on each interval.

Withers (1981) introduced the l-mixing condition, which is weaker than α-mixing. Unfortunately, the l-mixing property is not hereditary in the sense that $\{X_t\}$ being l-mixing does not guarantee $\{g(X_t)\}$ being l-mixing for nonlinear $g(\cdot)$. Doukhan and Louhichi (1999) provide a unifying approach dealing with mixing, association, Gaussian sequences and Bernoulli shifts. The limit theorems and various inequalities were established under a unifying weak dependence condition. Yoshihara (1976) and Denker and Keller (1983) provided asymptotic theory of U-statistics for β-mixing processes. A central limit theorem on degenerate U-statistics under the β-mixing condition can be found in Hjellvik, Yao, and Tjøstheim (1998).

3
ARMA Modeling and Forecasting

Fitting an appropriate ARMA(p, q) model to an observed time series data set involves two interrelated problems, namely determining the order (p, q) (which is usually referred to as model identification) and estimating parameters in the model. Further, the postfitting diagnostic checking on the validity of the fitted model is equally important.

In this chapter, we first present a comprehensive account on the (Gaussian) maximum likelihood approach for parameter estimation, which covers the methodology, the algorithms and the asymptotic properties. Then we outline some routine procedures for model identification and diagnostic checking, paying particular attention to the Akaike information criterion and its variants. Although fitting a time series model is always in the order of model identification, parameter estimation, and diagnostic checking, we deal with the problem of estimation first since almost all identification methods involve estimating parameters. Finally, we briefly discuss the forecasting methods based on nonstationary ARMA models. The methods presented there are practically applicable for ARIMA models.

3.1 Models and Background

Let $X_1, \cdots, X_T$ be observations from a causal ARMA(p, q) process defined by

$$X_t - b_1 X_{t-1} - \cdots - b_p X_{t-p} = \varepsilon_t + a_1 \varepsilon_{t-1} + \cdots + a_q \varepsilon_{t-q}, \tag{3.1}$$

where $\{\varepsilon_t\} \sim \text{WN}(0, \sigma^2)$. Our goal is to determine the AR-order p and the MA-order q and to estimate the AR coefficients $\{b_j\}$, the MA coefficients $\{a_j\}$, and the variance of the white noise σ^2.

In the setting above, we let $EX_t = 0$. In practice, we subtract the sample mean from the data before the fitting; see Theorem 2.8(i) for the asymptotic property of the sample mean.

We assume that model (3.1) is causal (see Definition 2.3). This is equivalent to the condition $b(z) \equiv 1 - b_1 z - \cdots - b_p z^p \neq 0$ for all $|z| \leq 1$. To avoid ambiguity, we also assume that $\{a_i\}$ and σ^2 have been adjusted (without changing the ACVF of the model) to ensure that

$$a(z) \equiv 1 + a_1 z + \cdots + a_q z^q \neq 0 \quad \text{for all } |z| < 1. \tag{3.2}$$

This condition rules out the possibility that two different causal ARMA models share the same ACVF; see Proposition 4.4.2 of Brockwell and Davis (1991). In practice, the assumption above implies that whenever we encounter more than one set of solutions, we always pick those estimated AR coefficients such that the fitted model is causal and those estimated MA coefficients such that condition (3.2) holds.

In the case $q = 0$, model (3.1) reduces to a pure AR model, which is in the form of a linear regressive model. Therefore, standard procedures for linear regression estimation, such as the *least squares method*, are readily applicable. An alternative approach is to replace the ACVF in (2.21) by its sample version. Then, the estimators for the b_j's are obtained by solving the equations with $k = 1, \cdots, p$. This leads to the well-known *Yule–Walker estimators*; see §8.1 of Brockwell and Davis (1991). Both methods admit simple and closed-form solutions. However, they are not directly applicable for MA and ARMA models. The modified Yule–Walker estimators for ARMA models are typically inefficient. Therefore, we focus on the (Gaussian) maximum likelihood method, which in principle is applicable to any stationary time series. In fact, Theorem 3.2 below shows that Gaussian maximum likelihood estimators for ARMA models with i.i.d. white noise $\{\varepsilon_t\}$ are always asymptotically normal, with the variance independent of the distribution of ε_t.

Due to the dependence in the data, the Gaussian likelihood function for an ARMA model involves the inverse of a $T \times T$ covariance matrix and does not admit an explicit maximum likelihood estimator. This posed difficulties in implementing the method in practice at early stages. Various ad hoc methods, aiming for approximating the exact maximum likelihood approach, have been proposed to ease the computational burden; see Section 5.4 of Priestley (1981) and the references therein. For example, for an AR(p) model with Gaussian white noise, the conditional distribution function of $X_{p+1}, \cdots, X_T$ given the first p observations $X_1, \cdots, X_p$ is

$$(2\pi\sigma^2)^{-(T-p)/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{t=p+1}^{T} (X_t - b_1 X_{t-1} - \cdots - b_p X_{t-p})^2 \right\},$$

from which we can easily derive the *conditional maximum likelihood estimators* for b_i's and σ^2. By doing this, we effectively reduce the sample size from T to $T-p$. The full likelihood is the product of the conditional density of $X_{p+1}, \cdots, X_T$ given $X_1, \cdots, X_p$ and the density of $X_1, \cdots, X_p$. Thus, the conditional likelihood function contains nearly all of the information in the data except that contained in the density function of $X_1, \cdots, X_p$. On the other hand, with modern computer power coupled with efficient algorithms, we argue that it is ready now for us to use the exact maximum likelihood (either the full likelihood or, more conveniently, the conditional likelihood) estimation as a benchmark procedure for the estimation of ARMA models. The estimation is in fact implemented in most modern time series packages such as the ITSM of Brockwell and Davis (1996).

3.2 The Best Linear Prediction—Prewhitening

In order to avoid computing the inverse of large matrices in likelihood functions, we prewhiten the data first, which is effectively equivalent to evaluating the best linear predictor for X_t based on $X_{t-1}, \cdots, X_1$ for $t \geq 2$.

Definition 3.1 *Let $\{X_t\}$ be a stationary process with mean zero. We call*

$$\widehat{X}_{k+1} = \varphi_{k1} X_k + \cdots + \varphi_{kk} X_1 \tag{3.3}$$

the best linear predictor *for X_{k+1} based on $X_k, \cdots, X_1$ if*

$$E(X_{k+1} - \widehat{X}_{k+1})^2 = \min_{\{\psi_j\}} E\left(X_{k+1} - \sum_{j=1}^{k} \psi_j X_{k-j+1}\right)^2. \tag{3.4}$$

Taking the derivatives with respect to ψ_j and setting them to zero, we obtain a system of equations:

$$E\left(X_{k+1} - \sum_{j=1}^{k} \varphi_{kj} X_{k-j+1}\right) X_{k-i+1} = 0.$$

This yields the following theorem.

Theorem 3.1 *A set of coefficients $\{\varphi_{kj}\}$ satisfies (3.3) and (3.4) if and only if*

$$\sum_{j=1}^{k} \varphi_{kj} \gamma(i-j) = \gamma(i), \quad i = 1, \cdots, k, \tag{3.5}$$

where $\gamma(\cdot)$ is the ACVF of $\{X_t\}$.

Proof. For any $\{\psi_j\}$,

$$E\left(X_{k+1} - \sum_{j=1}^{k} \psi_j X_{k-j+1}\right)^2 \tag{3.6}$$

$$= E(X_{k+1} - \widehat{X}_{k+1})^2 + E\left\{\sum_{j=1}^{k} (\varphi_{kj} - \psi_j) X_{k-j+1}\right\}^2 + 2B,$$

where

$$\begin{aligned} B &= E\left\{\left(X_{k+1} - \sum_{j=1}^{k} \varphi_{kj} X_{k+1-j}\right) \sum_{i=1}^{k} (\varphi_{ki} - \psi_i) X_{k+1-i}\right\} \\ &= \sum_{i=1}^{k} (\varphi_{ki} - \psi_i) \left\{\gamma(i) - \sum_{j=1}^{k} \varphi_{kj}\gamma(i-j)\right\}. \end{aligned}$$

It is easy to see from (3.6) that $E(X_{k+1} - \sum_{j=1}^{k} \psi_j X_{k-j+1})^2 \geq E(X_{k+1} - \widehat{X}_{k+1})^2$ for any $\{\psi_j\}$ if $B = 0$. The latter is equivalent to condition (3.5). On the other hand, suppose that there exists an i $(1 \leq i \leq k)$ for which $C_1 \equiv \gamma(i) - \sum_{j=1}^{k} \varphi_{kj}\gamma(i-j) \neq 0$. Let $\psi_i = \varphi_{ki} + C_2$ and $\psi_j = \varphi_{kj}$ for all $j \neq i$. Then

$$E(X_{k+1} - \sum_{j=1}^{k} \psi_j X_{k-j+1})^2 = E(X_{k+1} - \widehat{X}_{k+1})^2 + C_2^2 \mathrm{Var}(X_t) - 2C_2C_1.$$

Choosing C_2 such that $C_2C_1 > 0$ and $|C_2| < 2|C_1|/\mathrm{Var}(X_t)$ entails that $E(X_{k+1} - \sum_{j=1}^{k} \psi_j X_{k-j+1})^2 < E(X_{k+1} - \widehat{X}_{k+1})^2$, which contradicts the definition of the best linear predictor. Therefore (3.5) is also a necessary condition for (3.4). ∎

From the proof above we can see that

$$\mathrm{Cov}(\widehat{X}_{k+1} - X_{k+1}, X_i) = 0, \quad i = 1, \cdots, k.$$

Since $X_i - \widehat{X}_i$ is a linear combination of $X_i, \cdots, X_1$ only, we conclude that $\{X_t - \widehat{X}_t,\ t = 1, \cdots, T\}$ is a sequence of uncorrelated random variables, where we define $\widehat{X}_1 \equiv 0$. Transforming the original data $\{X_t,\ t = 1, \cdots, T\}$ to the uncorrelated sequence $\{X_t - \widehat{X}_t,\ t = 1, \cdots, T\}$ is called *prewhitening*. It is easy to see that $E(X_t - \widehat{X}_t) = 0$ and

$$\begin{aligned} \nu_t &\equiv \mathrm{Var}(X_{t+1} - \widehat{X}_{t+1}) = E\{(X_{t+1} - \widehat{X}_{t+1})X_{t+1}\} \\ &= \gamma(0) - \sum_{j=1}^{t} \varphi_{tj}\gamma(j). \end{aligned} \tag{3.7}$$

Note that, for causal ARMA processes, ACVF can be easily evaluated numerically based on its MA(∞)-representation; see (2.20). Based on the ACVF, the predictive errors $\{X_t - \widehat{X}_t\}$ and their variances $\{\nu_t\}$ can be calculated through the *innovation algorithm* described below. For its proof, we refer to Proposition 5.2.2 of Brockwell and Davis (1991).

Innovation algorithm: Set $\nu_0 = \gamma(0)$. Based on the cross-recursive equations

$$\begin{aligned} \theta_{k,k-j} &= \nu_j^{-1}\left\{\gamma(k-j) - \sum_{i=0}^{j-1}\theta_{j,j-i}\theta_{k,k-i}\nu_i\right\}, \\ \nu_k &= \gamma(0) - \sum_{j=0}^{k-1}\theta_{k,k-j}^2\nu_j, \end{aligned}$$

compute the values of $\{\theta_{ij}\}$ and $\{\nu_j\}$ in the order

$$\begin{aligned} &\theta_{11}, \nu_1, \\ &\theta_{22}, \theta_{21}, \nu_2, \\ &\theta_{33}, \theta_{32}, \theta_{31}, \nu_3, \\ &\cdots\ \cdots \\ &\theta_{T-1,T-1}, \theta_{T-1,T-2}, \cdots, \theta_{T-1,1}, \nu_{T-1}. \end{aligned}$$

The best linear predictors are given by $\widehat{X}_1 = 0$ and

$$\widehat{X}_{k+1} = \sum_{j=1}^{k}\theta_{kj}(X_{k+1-j} - \widehat{X}_{k+1-j}), \quad k = 1, \cdots, T-1. \tag{3.8}$$

3.3 Maximum Likelihood Estimation

3.3.1 Estimators

Let $\mathbf{X}_T = (X_1, \cdots, X_T)^\tau$ and $\widehat{\mathbf{X}}_T = (\widehat{X}_1, \cdots, \widehat{X}_T)^\tau$. It follows from (3.8) that $\widehat{\mathbf{X}}_T = \mathbf{\Theta}(\mathbf{X}_T - \widehat{\mathbf{X}}_T)$, where

$$\mathbf{\Theta} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \theta_{11} & 0 & 0 & \cdots & 0 \\ \theta_{22} & \theta_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ \theta_{T-1,T-1} & \theta_{T-1,T-2} & \theta_{T-1,T-3} & \cdots & 0 \end{pmatrix}.$$

Hence, we may write $\mathbf{X}_T = \mathbf{C}(\mathbf{X}_T - \widehat{\mathbf{X}}_T)$, where $\mathbf{C} = \mathbf{\Theta} + \mathbf{I}_T$ is a lower-triangular matrix with all main diagonal elements 1, and $\mathbf{I}_T$ is the $T \times T$

identity matrix. Let $\mathbf{D}$ be the diagonal matrix $\mathbf{D} = \text{diag}(\nu_0, \cdots, \nu_{T-1})$. Since $E\mathbf{X}_T = 0$, it follows from (3.7) that

$$\mathbf{\Sigma} \equiv \text{Var}(\mathbf{X}_T) = \mathbf{C}\mathbf{D}\mathbf{C}^\tau \quad \text{and} \quad |\mathbf{\Sigma}| = |\mathbf{D}| = \prod_{j=0}^{T-1} \nu_j.$$

Note that the matrices $\mathbf{\Sigma}$, $\mathbf{C}$, and $\mathbf{D}$ depend on the parameters $\mathbf{a}$ and $\mathbf{b}$ (see the innovation algorithm). Hence, if $\{X_t\}$ is a Gaussian causal ARMA process defined by (3.1), the likelihood function is of the form (the density of multivariate normal distribution)

$$\begin{aligned} L(\mathbf{b}, \mathbf{a}, \sigma^2) &\propto |\mathbf{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{X}_T^\tau \mathbf{\Sigma}^{-1}\mathbf{X}_T\right\} \\ &= (\nu_0 \cdots \nu_{T-1})^{-1/2} \exp\left\{-\frac{1}{2}\sum_{j=1}^{T}(X_j - \widehat{X}_j)^2/\nu_{j-1}\right\} \\ &= \sigma^{-T}(r_0 \cdots r_{T-1})^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{T}(X_j - \widehat{X}_j)^2/r_{j-1}\right\}, \quad (3.9) \end{aligned}$$

where $\mathbf{b} = (b_1, \cdots, b_p)^\tau$, $\mathbf{a} = (a_1, \ldots, a_q)^\tau$, and $r_j = \nu_j/\sigma^2$. Maximizing this likelihood function, we obtain the *maximum likelihood estimator*

$$(\widehat{\mathbf{b}}, \widehat{\mathbf{a}}, \widehat{\sigma}^2) = \arg \min_{(\mathbf{b},\mathbf{a})\in\mathcal{B}, \sigma^2>0} L(\mathbf{b}, \mathbf{a}, \sigma^2), \quad (3.10)$$

where

$$\mathcal{B} = \{(\mathbf{b}, \mathbf{a}) \,:\, b(z) \cdot a(z) \neq 0 \text{ for all } |z| \leq 1\}. \quad (3.11)$$

In the definition above, we require the estimator to be in the set $\mathcal{B}$ to ensure that the fitted model is causal and *invertible*. We call the ARMA(p, q) model (3.1) invertible if $a(z) \neq 0$ for all complex numbers z with $|z| \leq 1$. From the proof of Theorem 2.1, we can see that the invertibility implies that $\{X_t\}$ can be expressed as an AR(∞) process.

We can see from (2.19), (3.5), and (3.7) that $\{r_i\}$ and $\{\varphi_{ji}\}$ do not depend on σ^2. Maximizing over σ first, by (3.9), the maximum likelihood estimators can be expressed as

$$(\widehat{\mathbf{b}}, \widehat{\mathbf{a}}) = \arg \min_{(\mathbf{b},\mathbf{a})\in\mathcal{B}} \left(\log\{S(\mathbf{b}, \mathbf{a})\} + T^{-1}\sum_{j=1}^{T}\log r_{j-1}\right), \quad \widehat{\sigma}^2 = S(\widehat{\mathbf{b}}, \widehat{\mathbf{a}})/T, \quad (3.12)$$

where

$$S(\mathbf{b}, \mathbf{a}) = \sum_{j=1}^{T}(X_j - \widehat{X}_j)^2/r_{j-1}. \quad (3.13)$$

Equality (3.9) shows that we can avoid calculating the inverse of the covariance matrix $\mathbf{\Sigma}$ through prewhitening, which reduces a great deal of the computational burden in searching for $(\widehat{\mathbf{b}}, \widehat{\mathbf{a}})$. In numerical implementation, we often drop the constraint $(\mathbf{b}, \mathbf{a}) \in \mathcal{B}$ in (3.12) and solve the unconstrained minimization problem first. Let $\widetilde{b}_j$'s and $\widetilde{a}_i$'s be the unconstrained minimizers. As long as they do not entertain a unit root in the sense that $(1 - \sum_{j=1}^{p} \widetilde{b}_j z^j)(1 + \sum_{i=1}^{q} \widetilde{a}_i z^i) \neq 0$ for all $|z| = 1$, the constrained minimizer $(\widehat{\mathbf{b}}, \widehat{\mathbf{a}}) \in \mathcal{B}$ can be obtained as follows. Let $z_1, \cdots, z_p$ be the roots of $1 - \sum_{j=1}^{p} \widetilde{b}_j z^j = 0$, that is,

$$1 - \sum_{j=1}^{p} \widetilde{b}_j z^j = \prod_{j=1}^{p} (1 - z/z_j).$$

(See §9.5 of Press et al. (1992) for the algorithms to find the roots' z_i's.) Without loss of generality, we assume that $|z_j| < 1$ for $1 \leq j \leq k$ and $|z_j| > 1$ for $k < j \leq p$. Then, the desired estimators' $\widehat{b}_j$'s are defined by the equation

$$1 - \sum_{j=1}^{p} \widehat{b}_j z^j = \prod_{j=1}^{k} (1 - z_j z) \prod_{i=k+1}^{p} (1 - z/z_i);$$

see Proposition 4.4.2 of Brockwell and Davis (1991) and the discussion on causality below Definition 2.3. The estimators' $\widehat{a}_i$'s may be obtained in a similar manner. Furthermore, the estimator $\widehat{\sigma}^2$ defined in (3.12) should be calculated based on $(\widehat{\mathbf{b}}, \widehat{\mathbf{a}})$ instead of $(\widetilde{\mathbf{b}}, \widetilde{\mathbf{a}})$. Note that $(\widehat{\mathbf{b}}, \widehat{\mathbf{a}})$ and $(\widetilde{\mathbf{b}}, \widetilde{\mathbf{a}})$ share the same ACF. Thus, it holds that $S(\widehat{\mathbf{b}}, \widehat{\mathbf{a}}) = S(\widetilde{\mathbf{b}}, \widetilde{\mathbf{a}})$. From (3.12), we can see that the likelihood function admits the same values at $(\widehat{\mathbf{b}}, \widehat{\mathbf{a}})$ and $(\widetilde{\mathbf{b}}, \widetilde{\mathbf{a}})$. Hence $(\widehat{\mathbf{b}}, \widehat{\mathbf{a}})$ obtained above is the genuine constrained maximum likelihood estimator within the set $\mathcal{B}$.

The maximum likelihood estimation has been implemented in most modern time series packages, such as the ITSM of package Brockwell and Davis (1996). In general, some nonlinear optimization programs will be used in conjunction with the innovation algorithm in the search. One common optimization routine for this purpose is the Newton–Raphson procedure, which computes the estimator in an iterative manner. To illustrate the basic idea of the procedure, we write $\boldsymbol{\beta} = (\mathbf{b}^\tau, \mathbf{a}^\tau)^\tau$ and

$$\ell(\boldsymbol{\beta}) = \log\{S(\mathbf{b}, \mathbf{a})\} + T^{-1} \sum_{j=1}^{T} \log r_{j-1}.$$

Let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood estimator for $\boldsymbol{\beta}$. It is easy to see from (3.12) that $\dot{\ell}(\widehat{\boldsymbol{\beta}}) = 0$, where $\dot{\ell}$ denotes the derivative of ℓ with respect to $\boldsymbol{\beta}$. When the sample size T is large, it is reasonable to expect that $\widehat{\boldsymbol{\beta}}$ is close

to the true value $\boldsymbol{\beta}_0$. Hence, a simple Taylor expansion entails

$$-\dot{\ell}(\boldsymbol{\beta}_0) \approx \ddot{\ell}(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

where $\ddot{\ell}$ denotes the Hessian matrix. Based on this, we may define the iterative estimators

$$\widehat{\boldsymbol{\beta}}_{k+1} = \widehat{\boldsymbol{\beta}}_k - \{\ddot{\ell}(\widehat{\boldsymbol{\beta}}_k)\}^{-1}\dot{\ell}(\widehat{\boldsymbol{\beta}}_k), \qquad k = 0, 1, \cdots .$$

With a carefully selected initial value $\widehat{\boldsymbol{\beta}}_0$, the iterative estimators $\widehat{\boldsymbol{\beta}}_k$ may converge to a limit that is taken as the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ since $\dot{\ell}(\widehat{\boldsymbol{\beta}}) \approx 0$. Although this idea is very simple, the actual implementation is much more complicated and involves quite a few fine technical details; see, for example, §9.4 and §9.6 of Press et al. (1992). A good initial estimator plays an important role in ensuring a secure and fast convergence. On the other hand, more sophisticated optimization algorithms, such as those presented in Chapter 10 of Press et al. (1992), may also be used for this purpose.

Although we advocate the maximum likelihood estimation method, some preliminary estimation based on relatively simple or ad hoc methods provides good initial values for the algorithms searching for the maximum likelihood estimators, which practically constitute an important part of the maximum likelihood estimation procedure. We refer the reader to §8.1–§8.4 of Brockwell and Davis (1991) for detailed discussion on various preliminary estimation methods. Those methods have also been incorporated in the package ITSM.

In fact, the method described above may be applied to compute the maximum likelihood estimators for any Gaussian processes. On the other hand, when $\{X_t\}$ is not Gaussian, we may still regard (3.9) as a measure of goodness of fit to the data and choose the parameters that maximize this measure. We will always refer to (3.9) as the *Gaussian likelihood function* and estimators derived from maximizing the Gaussian likelihood as maximum pseudolikelihood estimators or simply maximum likelihood estimators, regardless of the underlying distribution. Theorem 3.2 below shows that the maximum likelihood estimators so defined are asymptotically distribution-free as long as $\{\varepsilon_t\} \sim \text{IID}(0, \sigma^2)$. However, when ε_t is not Gaussian, the maximum pseudolikelihood estimators $(\widehat{\mathbf{b}}, \widehat{\mathbf{a}})$ are typically inefficient. Furthermore, when ε_t has heavy tails in the sense that $\text{Var}(\varepsilon_t) = \infty$, Gaussian likelihood (as well as the least squares approach) may lead to inconsistent estimators. Some robust methods are more adaptive to heavy-tailed data; see, for example, Davis, Knight, and Liu (1992) and Hall, Peng, and Yao (2002).

3.3.2 Asymptotic Properties

We first introduce some notation. Let $\{W_t\} \sim \text{WN}(0,1)$. Define

$$b(B)U_t = W_t \qquad \text{and} \qquad a(B)V_t = W_t.$$

Namely, $\{U_t\}$ is an AR(p) process defined in terms of the AR-coefficients in model (3.1) and $\{V_t\}$ is an AR(q) process defined in terms of the MA-coefficients in model (3.1), and the two processes are correlated with each other since they are defined in terms of the same white noise process $\{W_t\}$. Let $\mathbf{Z} = (U_{-1}, \cdots, U_{-p}, V_{-1}, \cdots, V_{-q})^\tau$, and

$$\mathbf{W}(\mathbf{b}, \mathbf{a}) = \{\text{Var}(\mathbf{Z})\}^{-1}. \tag{3.14}$$

Theorem 3.2 *Let $\{X_t\}$ be the ARMA process defined in (3.1) in which $\{\varepsilon_t\} \sim \text{IID}(0, \sigma^2)$ with $\sigma^2 > 0$ and the true value $(\mathbf{b}_0, \mathbf{a}_0) \in \mathcal{B}$ defined in (3.11). Then, as $T \to \infty$,*

$$T^{\frac{1}{2}} \begin{pmatrix} \widehat{\mathbf{b}} - \mathbf{b}_0 \\ \widehat{\mathbf{a}} - \mathbf{a}_0 \end{pmatrix} \xrightarrow{D} N\big(\mathbf{0}, \mathbf{W}(\mathbf{b}_0, \mathbf{a}_0)\big)$$

and $\widehat{\sigma}^2 \xrightarrow{P} \sigma^2$, where $(\widehat{\mathbf{b}}, \widehat{\mathbf{a}})$ and $\widehat{\sigma}^2$ are defined in (3.12), and $\mathbf{W}(\cdot)$ is defined in (3.14).

The theorem above shows that the asymptotic distributions of the estimators are independent of σ^2. In this sense, the quality of the estimators for causal and invertible ARMA models is not affected by the magnitude of the white noise. This is due to the fact that the ratio of signal to noise in a causal ARMA model is independent of σ^2. For example, for the AR(1) model $X_t = bX_{t-1} + \varepsilon_t$ with $\{\varepsilon_t\} \sim \text{WN}(0, \sigma^2)$, we have

$$\text{Var}(X_t) = b^2 \text{Var}(X_{t-1}) + \sigma^2.$$

Due to stationarity,

$$\text{Var}(X_t) = \text{Var}(X_{t-1}) \quad \text{and} \quad \text{Var}(X_t) = \sigma^2/(1-b^2).$$

Hence, the ratio of signal to noise is

$$\{\text{Var}(X_t)/\text{Var}(\varepsilon_t)\}^{1/2} = (1-b^2)^{-1/2},$$

which does not depend on σ^2. In fact, this conclusion holds for general causal ARMA processes.

The theorem above was first obtained by Hannan (1973) based on some sophisticated frequency-domain arguments; see §10.8 of Brockwell and Davis (1991). A time-domain proof was given in Yao and Brockwell (2001). Note that we do not even need the condition that $E(\varepsilon_t)^4 < \infty$ (see Theorem 2.8).

This is due to the fact that under the condition $(\mathbf{b}_0, \mathbf{a}_0) \in \mathcal{B}$, $\{X_t\}$ is effectively an AR(∞) process. Therefore, the asymptotic behavior can be established similarly to that of the estimator in a linear regression model. In fact, the condition that $\{\varepsilon_t\} \sim \text{IID}(0, \sigma^2)$ in the theorem above can be replaced by $\{\varepsilon_t\}$ being merely a sequence of martingale differences with constant conditional variance $\sigma^2 < \infty$; see Hannan (1973).

The covariance matrix $\mathbf{W}(\cdot)$ is dictated by two correlated AR processes. We list below concrete forms of $\mathbf{W}(\cdot)$ for some simple models to illustrate the usefulness of this asymptotic result.

(i) *AR(p) models*

For autoregressive models with order p, $\mathbf{Z} = (U_p, \cdots, U_1)^\tau$. Since the scale of $\mathbf{Z}$ is a factor of σ as large as $(X_p, \cdots, X_1)$, $\text{Var}(\mathbf{Z}) = \mathbf{\Gamma}_p/\sigma^2$, where $\mathbf{\Gamma}_p$ is a $p \times p$ matrix with $\gamma(i-j)$ as its (i,j)th element. Thus, the asymptotic variance of $\widehat{\mathbf{b}} = (\widehat{b}_1, \cdots, \widehat{b}_q)^\tau$ is $\sigma^2 \mathbf{\Gamma}_p^{-1}/T$. In the special cases $p = 1$ and 2, we have

$$\begin{aligned} &\text{AR}(1): \quad \text{Var}(\widehat{b}_1) \approx (1 - b_1^2)/T, \\ &\text{AR}(2): \quad \text{Var}\begin{pmatrix} \widehat{b}_1 \\ \widehat{b}_2 \end{pmatrix} \approx \frac{1}{T}\begin{pmatrix} 1 - b_2^2 & -b_1(1+b_2) \\ -b_1(1+b_2) & 1 - b_2^2 \end{pmatrix}. \end{aligned}$$

For an AR(p) model, the least squares estimator, the Yule–Walker estimator, and the conditional maximum likelihood estimator share the same asymptotic distribution as the maximum likelihood estimator $\widehat{\mathbf{b}}$; see Theorem 10.8.2 of Brockwell and Davis (1991). This can be understood as follows. Note that an AR(p) model may be regarded as an AR(k) model for any $k > p$ with $b_j = 0$ for $p < j \le k$. It can be proved that, for any causal AR(p) model and $k > p$, the (k,k)th element of $\mathbf{\Gamma}_k^{-1}$ is $|\mathbf{\Gamma}_{k-1}|/|\mathbf{\Gamma}_k| = \sigma^{-2}$. Thus $\widehat{b}_k$ is asymptotically normal with mean 0 and variance $1/T$. Applying this result in the context of the estimation of partial autocorrelation function $\pi(\cdot)$ (see §2.2.3), we obtain the following proposition, which will play an important role in identifying purely autoregressive models.

Proposition 3.1 *Suppose that $\{X_1, \cdots, X_T\}$ is a sample from a causal AR(p) model defined with* IID$(0, \sigma^2)$ *white noise and $\sigma^2 > 0$. Then, as $T \to \infty$,*

$$T^{-1/2}\,\widehat{\pi}(k) \xrightarrow{D} N(0,1) \quad \text{for any } k > p,$$

where $\widehat{\pi}(k) \equiv \widehat{b}_k$ is an estimator for b_k in fitting an AR(k) model to the data $\{X_1, \cdots, X_T\}$ using the (conditional) maximum likelihood method, the least squares method, or the Yule–Walker estimation method.

(ii) *MA(q) models*

For moving average models with order q, the asymptotic variance of $\widehat{\mathbf{a}} = (\widehat{a}_1, \cdots, \widehat{a}_q)^\tau$ is $(\mathbf{\Gamma}_q^*)^{-1}/T$, and $\mathbf{\Gamma}_q^*$ is a $q \times q$ matrix with $\gamma^*(i-j)$ as its

(i,j)th element, where $\gamma^*(\cdot)$ is the ACVF of the AR(q) process $a(B)Y_t = e_t$ and $\{e_t\} \sim \text{WN}(0,1)$. In the special cases of $q = 1$ and 2, we have

$$\begin{aligned} &\text{MA(1)}: \quad \text{Var}(\widehat{a}_1) \approx (1-a_1^2)/T, \\ &\text{MA(2)}: \quad \text{Var}\begin{pmatrix} \widehat{a}_1 \\ \widehat{a}_2 \end{pmatrix} \approx \frac{1}{T}\begin{pmatrix} 1-a_2^2 & a_1(1-a_2) \\ a_1(1-a_2) & 1-a_2^2 \end{pmatrix}. \end{aligned}$$

(iii) *ARMA(1,1) models*

For model $X_t - bX_{t-1} = \varepsilon_t + a\varepsilon_{t-1}$, it can be computed that

$$\text{Var}\begin{pmatrix} \widehat{b} \\ \widehat{a} \end{pmatrix} \approx \frac{1+ab}{T\,(a+b)^2}\begin{pmatrix} (1-b^2)(1+ab) & (b^2-1)(1-a^2) \\ (b^2-1)(1-a^2) & (1-a^2)(1+ab) \end{pmatrix}.$$

3.3.3 Confidence Intervals

The asymptotic variance matrix $\mathbf{W}(\mathbf{b},\mathbf{a})$ given in (3.14) may be used to calculate the *standard errors* of the maximum likelihood estimators for parameters in causal and invertible ARMA models. For example, the standard errors for $\widehat{b}_j$ and $\widehat{a}_i$ are $(w_{jj}/T)^{1/2}$ and $(w_{p+i,p+i}/T)^{1/2}$, respectively, where w_{ii} is the (i,i)th element of $\mathbf{W}(\widehat{\mathbf{b}},\widehat{\mathbf{a}})$. Most time series packages automatically provide the values of standard errors when calculating estimates.

On the other hand, approximate confidence regions for parameters in ARMA models can be easily constructed in terms of the limit distribution in Theorem 3.2. For example, an approximate $(1-\alpha)$ confidence region for the AR coefficient vector $\mathbf{b}$ is obtained as

$$\{\mathbf{b} = (b_1, \cdots, b_p)^\tau : \ (\widehat{\mathbf{b}} - \mathbf{b})^\tau \widehat{\mathbf{W}}_1^{-1} (\widehat{\mathbf{b}} - \mathbf{b}) \le \chi^2_{1-\alpha}(p)/T\},$$

where $\widehat{\mathbf{W}}_1$ is the $p \times p$ upper-left submatrix of $\mathbf{W}(\widehat{\mathbf{b}},\widehat{\mathbf{a}})$, and $\chi^2_p(1-\alpha)$ is the 100αth percentile of the χ^2-distribution with p degrees of freedom. An approximate $(1-\alpha)$ confidence interval for the single parameter b_j is given as

$$\{b_j : \ |\widehat{b}_j - b_j| \le T^{-1/2} w_{jj}^{1/2} z_{1-\alpha/2}\},$$

where z_α denotes the 100αth percentile of the standard normal distribution.

3.4 Order Determination

In this section, we first introduce general principles for determining the order (p,q) in ARMA modeling, namely AIC, BIC, and FPE. In the end, we outline a general routine procedure for model identification.

3.4.1 Akaike Information Criterion

Akaike's information criterion (AIC) (Akaike 1973, 1974) is used to select the optimum parametric model based on observed data. It has been regarded as one of the important breakthroughs in statistics in the twentieth century. The basic idea of the AIC can be described as follows. Suppose that we use a probability density function f to approximate an unknown density g. The *Kullback–Leibler information*

$$I(g;f) = \int g(x)\log g(x)dx - \int g(x)\log f(x)dx \tag{3.15}$$

provides a measure for the lack of the approximation. It is easy to see that

$$I(g;f) = E\log\{g(X)/f(X)\}, \quad \text{with} \quad X \sim g.$$

By Jensen's inequality, we have

$$\begin{aligned} I(g;f) &= -E\log\big\{f(X)/g(X)\big\} \geq -\log\Big(E\big\{f(X)/g(X)\big\}\Big) \\ &= -\log\Big(\int f(x)/g(x)g(x)dx\Big) = 0, \end{aligned}$$

with equality holding if and only if $f = g$. A good approximation should make the measure $I(g;f)$ as small as possible. Note that the first term on the right-hand side of (3.15) does not depend on f. Hence, we should choose f that minimizes

$$-\int g(x)\log f(x)dx = -E_g\{\log f(X)\}.$$

Since we do not know g, we have only a set of observations $\{X_1, \cdots, X_T\}$ from g. Naturally, we will replace the expectation above by its unbiased estimator

$$-\frac{1}{T}\sum_{j=1}^{T}\log f(X_j).$$

Typically, we would choose f from among a set of parametric family $\{f_m(\cdot|\theta_m)\}$ indexed by m. The form of f_m is typically given for each m. For example, in the context of time series analysis f_m may stand for an ARMA family with the order $m \equiv (p,q)$, and $\theta_m = (b_1, \cdots, b_p, a_1, \cdots, a_q)$. The best approximation would minimize

$$-\frac{1}{T}\sum_{j=1}^{T}\log f_m(X_j|\theta_m). \tag{3.16}$$

Note that this is a two-step optimization: searching for the minimizer of θ_m for fixed m and then searching for the global minimum over different values

of m. Obviously, the minimizer in the first step is the maximum likelihood estimator $\widehat{\theta}_m$. The second step involves searching for m that minimizes

$$-\frac{1}{T}\sum_{j=1}^{T}\log f_m(X_j|\widehat{\theta}_m).$$

However, there is a serious drawback in this approach: the expression above is no longer an unbiased estimator for $-E_g\{\log f_m(X|\theta_m)\}$ due to the overfitting caused by the double use of the same data for the estimation of the expected log-likelihood and the estimation of the parameter θ_m. Akaike (1973) proposed to rectify this problem by adding the bias

$$-E_g\{\log f_m(X|\theta_m)\} + \frac{1}{T}\sum_{j=1}^{T} E_g\{\log f_m(X_j|\widehat{\theta}_m)\}$$

to the sample likelihood function. He showed that the bias can asymptotically be approximated as

$$-E_g\{\log f_m(X|\theta_m)\} + \frac{1}{T}\sum_{j=1}^{T} E_g\{\log f_m(X_j|\widehat{\theta}_m)\} \approx p_m/T,$$

where p_m denotes the number of estimated parameters; see also §2.1.3 of Kitagawa and Gersch (1996). Thus, a term p_m/T should be added to (3.16) in order to correct the bias, leading to

$$-\frac{1}{T}\sum_{j=1}^{T}\log f_m(X_j|\widehat{\theta}_m) + p_m/T.$$

Multiplying by a factor of $2T$, which does not affect the choice of m, we define the following *Akaike information criterion* (AIC):

$$\begin{aligned}\mathrm{AIC}(m) &= -2\sum_{j=1}^{T}\log f_m(X_j|\widehat{\theta}_m) + 2\,p_m \qquad (3.17)\\ &= -2(\text{maximized log likelihood}) + 2(\text{No. of estimated parameters}).\end{aligned}$$

On the right-hand side of the expression above, the first term reflects the lack of fit; increasing the complexity (e.g., number of parameters p_m) of the model is likely to make this term decrease. However, the model complexity is penalized by the second term. The optimum model that minimizes the AIC is a trade-off between the two terms, that is similar to the bias and variance trade-off in nonparametric estimation (see Chapter 5).

In the context of fitting an ARMA model to time series data, if we regard the Gaussian likelihood (3.9) as the true likelihood function, the AIC is of the form (after discarding some constants)

$$\mathrm{AIC}(p,q) = -2\log\{L(\widehat{\mathbf{b}},\widehat{\mathbf{a}},S(\widehat{\mathbf{b}},\widehat{\mathbf{a}})/T)\} + 2(p+q+1), \qquad (3.18)$$

where $\widehat{\mathbf{b}}$ and $\widehat{\mathbf{a}}$ are the maximum likelihood estimators for $(b_1, \cdots, b_p)^\tau$ and $(a_1, \cdots, a_q)^\tau$ defined in (3.10) and $S(\cdot, \cdot)$ is defined in (3.13). Hurvich and Tsai (1989) argued that a better bias correction could be obtained if we replaced $(p + q + 1)$ by an asymptotically equivalent quantity $T(p + q + 1)/(T-p-q-2)$; see also pp. 303–304 of Brockwell and Davis (1991). This leads to a modified criterion

$$\text{AICC}(p, q) = -2\log\{L(\widehat{\mathbf{b}}, \widehat{\mathbf{a}}, S(\widehat{\mathbf{b}}, \widehat{\mathbf{a}})/T)\} + \frac{2(p + q + 1)T}{T - p - q - 2}. \tag{3.19}$$

In view of the fact that the AIC tends to overestimate the orders (Akaike 1970, Jones 1975; Shibata 1980), AICC places a heavier penalty for large values of p and q to counteract the overfitting tendency of the AIC.

Suppose now that we fit a pure AR model with order $1 \le p \le L$ with $L \ge 1$ prescribed. Let $T' = T - L$. Based on the conditional Gaussian likelihood function

$$(2\pi)^{T'}\sigma^{-T'} \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=L+1}^{T}(X_t - b_1X_{t-1} - \cdots - b_pX_{t-p})^2\right\},$$

which is the full likelihood function divided by the density function of $X_1, \cdots, X_p$, we may define the following simpler versions of the AIC and AICC:

$$\text{AIC}(p) = T'\log(\widehat{\sigma}_p^2) + 2(p + 1), \tag{3.20}$$

$$\text{AICC}(p) = T'\log(\widehat{\sigma}_p^2) + \frac{2(p + 1)T'}{T' - p - 2}, \tag{3.21}$$

where

$$\widehat{\sigma}_p^2 = \frac{1}{T'}\sum_{t=L+1}^{T}(X_t - \widehat{b}_1X_{t-1} - \cdots - \widehat{b}_pX_{t-p})^2. \tag{3.22}$$

We select the order p that minimizes AIC(p) or AICC(p) defined above. The conditional argument above effectively reduces the sample size from T to $T' = T - L$. When T is large relative to L, the selected orders differ little from those derived from (3.18) and (3.19).

3.4.2 FPE Criterion for AR Modeling

An alternative procedure for the order determination in AR modeling is the *final prediction error criterion* due to Akaike (1969). The basic idea is very simple. We have a hypothetical set of observations $\{\widetilde{X}_1, \cdots, \widetilde{X}_T\}$ that are from the same underlying process as the real observations $\{X_1, \cdots, X_T\}$. Then, we estimate the AR coefficients from $\{X_t\}$ and select the order such that the prediction errors on the fictitious data $\{\widetilde{X}_t\}$ obtain the minimum.

Namely, we choose p that minimizes

$$\widetilde{\sigma}_p^2 \equiv \frac{1}{T'} \sum_{t=L+1}^{T} (\widetilde{X}_t - \widehat{b}_1 \widetilde{X}_{t-1} - \cdots - \widehat{b}_p \widetilde{X}_{t-p})^2,$$

where $\widehat{b}_j$'s are the MLE of b_j's based on $\{X_1, \cdots, X_T\}$.

If the underlying process is a stationary AR(p) process with IID$(0, \sigma^2)$ white noise, the asymptotic approximations

$$E(\widetilde{\sigma}_p^2) \approx \sigma^2(1 + p/T'), \qquad E(\widehat{\sigma}_p^2) \approx \sigma^2(1 - p/T'),$$

hold,where $\widehat{\sigma}_p^2$ is defined in (3.22). Therefore, we may use $\widehat{\sigma}_p^2(T'+p)/(T'-p)$ as an approximation for the unobservable $\widetilde{\sigma}_p^2$. Now, the final prediction error is defined as

$$\text{FPE}(p) \;=\; \widehat{\sigma}_p^2 \frac{T' + p}{T' - p}.$$

The FPE criterion selects p that minimizes FPE(p).

Note that

$$\begin{aligned} T' \log\{\text{FPE}(p)\} &= T' \log(\widehat{\sigma}_p^2) + T' \log\left(1 + \frac{2p}{T' - p}\right) \\ &= T' \log(\widehat{\sigma}_p^2) + T' \frac{2p}{T' - p} + O(1/T'). \end{aligned}$$

Comparing this with (3.20) and (3.21), we have

$$\text{AIC}(p) \;=\; \text{AICC}(p) + O\left(\frac{1}{T'}\right) \;=\; T' \log\{\text{FPE}(p)\} + O\left(\frac{1}{T'}\right).$$

In this sense, the AIC, AICC, and FPE are asymptotically equivalent.

3.4.3 Bayesian Information Criterion

Since the AIC (also AICC and FPE) does not lead to a consistent order selection (Akaike 1970; Shibata 1980; Woodroofe 1982), various procedures have been proposed to modify the criterion in order to obtain consistent estimators. As a popular alternative to AIC, the *Bayesian information criterion* (BIC) defines the optimum model that minimizes

$$-2(\text{maximized log likelihood}) + \log T \times (\text{No. of estimated parameters}).$$

Comparing this with (3.17), BIC increases the penalty for the model complexity by replacing the factor 2 by $\log(T)$. This ensures that the estimated order is consistent (Hannan 1980). In the context of ARMA models, we have

$$\text{BIC}(p, q) = -2 \log\{L(\widehat{\mathbf{b}}, \widehat{\mathbf{a}}, S(\widehat{\mathbf{b}}, \widehat{\mathbf{a}})/T)\} + (p + q + 1) \log T. \tag{3.23}$$

In the same vein as (3.20) and (3.21), we may define the BIC for fitting AR models as

$$\mathrm{BIC}(p) = T' \log(\widehat{\sigma}_p^2) + (p+1) \log T'. \tag{3.24}$$

In the expressions above, all of the estimators are derived from the maximum likelihood method or its asymptotic equivalents.

The name BIC is due to the fact that the criterion was derived from diverse Bayesian arguments, see for example, Akaike (1977), Kashyap (1977), and Schwarz (1978). In fact, it can also be derived from a non-Bayesian argument such as in Rissanen (1980).

3.4.4 Model Identification

It is a general philosophy in model identification to allow modelers certain flexibility in exercising their subjective judgment. It is a fact of life that there rarely exists a true model in practice. A good practice of model identification should end with a selected model that is statistically sound and practically meaningful. A parsimonious model is always preferable when two candidate models appear about equally good. Below, we list a routine guideline from a purely data-analytic point of view.

Step 1. Examination of time-series plot
The first step is to produce a time-series plot; namely, to plot X_t against t and examine the plot to identify obvious trends, seasonal components, and outliers. These components should be removed through differencing, moving-averaging, or other appropriate methods (see §6.2).

Step 2. Examination of correlogram
Trend and seasonal components may show up in a correlogram (i.e., the plot of sample ACF $\widehat{\rho}(k)$ against k). A slowly damping correlogram is indicative of a slowly varying trend component. A periodic fluctuating correlogram is indicative of a periodic component (with the same period). Taking the difference at appropriate time lags may remove those nonstationary components.

Step 3. Determining the MA-order from the ACF and the AR-order from the PACF
If the data appear stationary in both the time-series plot and correlogram, we may try to identify the order (p, q) from the sample ACF $\{\widehat{\rho}(k)\}$ and the sample PACF $\{\widehat{\pi}(k)\}$ first. As a rule of thumb, we fit an AR(p) model to the data if

$$|\widehat{\pi}_k| \leq 1.96/\sqrt{T} \tag{3.25}$$

for about 95% of k's among all $k > p$ (see Proposition 3.1), and we fit an MA(q) model to the data if

$$|\widehat{\rho}(k)| \le 1.96 \left\{ 1 + 2 \sum_{j=1}^{q} \widehat{\rho}(j)^2 \right\}^{1/2} / \sqrt{T} \tag{3.26}$$

for about 95% of k's among all $k > q$ (see (2.27)). (We ignore the correlation among $\widehat{\pi}(k)$'s and $\widehat{\rho}(k)$'s for different k in the heuristic argument above.) Unfortunately, the simple patterns above are seldom observed in real data analysis. On the other hand, it is always recommended to estimate (or double-check) the order using the formal procedures in Step 4 below (see, for example, Example 3.2 below).

Step 4. Determining the orders using AIC or other information criteria
Since Akaike's pioneering work on AIC, various information criteria have been developed; see Choi (1992) for a survey. Each method has its own merit. A practically relevant question is when to use what, although a general answer to this question is inconceivable. The choice should depend on the nature and the aim of the data analysis. Empirical experience suggests that AIC is a good starting point. If we prefer a simple model that reflects the main and interpretable features, we may also try BIC, for example. On the other hand, forecasting based on an AR model with a slightly overestimated order does little harm. Shibata (1980) and Hurvich and Tsai (1989) showed that AIC, AICC, and FPE are asymptotically efficient, while BIC is not. The asymptotic efficiency is a desirable property defined in terms of the one-step mean square prediction error achieved by the fitted model. The AIC was not designed to be consistent, nor is its inconsistency necessarily a defect (Hannan 1986).

It is easy to see from (3.5) and (3.7) that the likelihood function (3.9) depends on the coefficients $\{b_j\}$ and $\{a_j\}$ only through the ACF. There may exist quite a few different ARMA models that provide almost equally good approximations to the sample ACF of the observed data set; see Example 3.3 below. Therefore, we may consider the models with AIC values within a small distance from the minimum AIC value as competitive candidates. Selection among the competitive models may be based on interpretation, simplicity, or diagnostic checking using the techniques described in §3.5 and/or §7.4. Formal statistical tests may also be employed if a choice has to be made, for example, between two candidate models.

To gain insights on the various criteria, we illustrate the methods through some simulated examples below. Estimation was carried out using the package ITSM of Brockwell and Davis (1996). ITSM evaluates the maximum likelihood estimate based on a 'preliminary estimate' that is obtained through the Yule–Walker method or other methods. Note that the ITSM calculates the BIC based on a different formula (p. 171 of Brockwell and

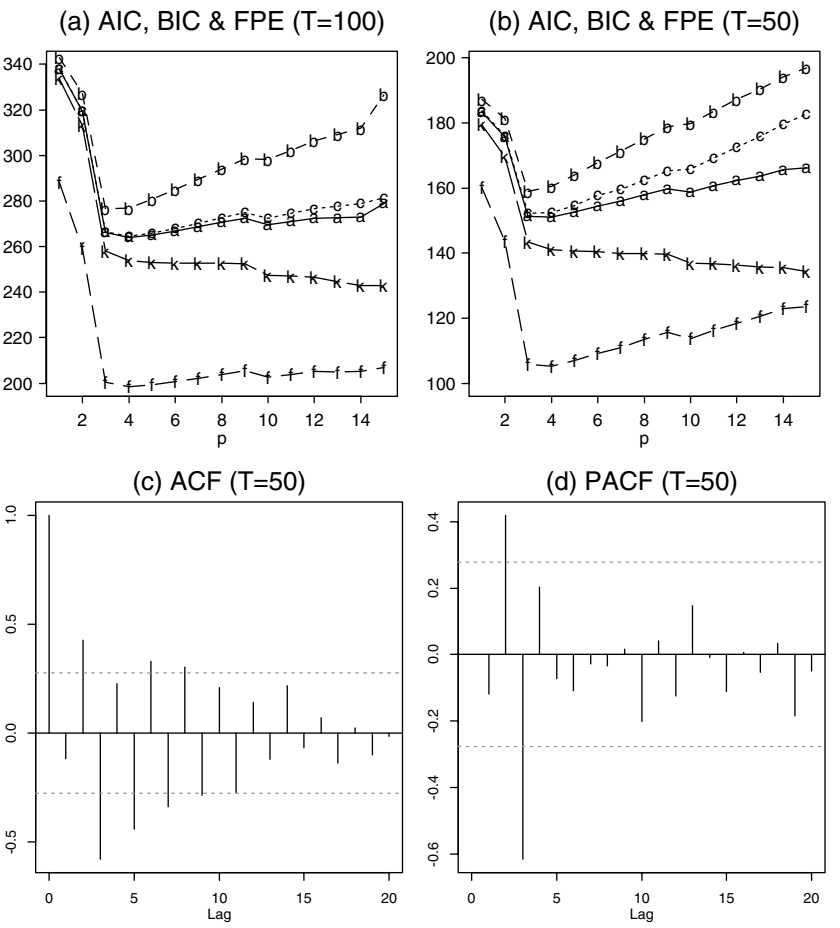


FIGURE 3.1. Example 3.1—Fitting an AR(p) model. Plots of AIC(p) (labeled "a"), AICC(p) (labeled "c"), BIC(p) (labeled "b"), and αFPE(p)+β (labeled "f") against p for sample size (a) $T = 100$ and (b) $T = 50$. [$(\alpha, \beta) = (100, 120)$ in (a) and (50, 50) in (b)]. Lines labeled with k are $-2\log$(maximum likelihood). (c) and (d) are ACF and PACF plots for a sample of size 50.

Davis 1996). The BIC values reported below were calculated based on (3.23) directly.

Example 3.1 *(Fitting AR models)* We draw a sample of 100 observations from the model

$$X_t = 0.5X_{t-1} + 0.3X_{t-2} - 0.7X_{t-3} + 0.2X_{t-4} + \varepsilon_t, \quad \{\varepsilon_t\} \sim_{\text{i.i.d.}} N(0, \sigma^2).$$

Assuming that the data are drawn from an AR model with an unknown order p, we fit a causal AR model with the order determined by the data. The sample PACF plotted in Figure 2.3(b) shows that $\widehat{\pi}(k)$'s for $k > 4$

are almost always between the bounds $\pm 1.96/\sqrt{T}$ (see (3.25)). This suggests that AR(p) with $p = 4$ might be a suitable candidate model. On the other hand, the sample ACF plotted in Figure 2.3(a) does not appear to have a clear cutoff. We apply AIC, AICC, BIC, and FPE to estimate the order p. The results are displayed in Figure 3.1(a). We also plot the $-2\log$(maximum likelihood) (i.e., the first term on the right hand side of (3.18), (3.19), and (3.23)), which decreases as the order p increases, although the decrease became slow and steady after the model reached the order 3. Nevertheless, this shows that the penalty for the model complexity is necessary for model selection. The difference between AIC and AICC is small for small values of p and only shows up when p is large. Due to the larger penalty on complex models, BIC increases faster than both AIC and AICC, as p increases. AIC, AICC and FPE chose the correct order 4 for the given sample, whereas BIC prefers orders 3 to 4 by a narrow numerical margin of 0.21. We repeated the exercise with a sample of size 50 and obtained similar results; see Figure 3.1(b). The PACF plot in Figure 3.1(d) suggests the order $p = 3$. Both AIC and FPE choose the correct order 4 for the given sample, whereas both AICC and BIC prefer orders 3 to 4 by numerical margins 0.18 and 1.62, respectively. The models with orders 3 and 4 could be regarded as competitive models. With sample size $T = 50$, the maximum likelihood estimates for the AR coefficients in the AR(4) model are

$$0.36,\ 0.29,\ -0.69,\ \text{and}\ 0.22,$$

with the standard errors 0.14, 0.11, 0.11, and 0.14, respectively. The estimate for the variance of the white noise is 0.94.

Example 3.2 *(Fitting MA models)* We generate a sample of 100 from the model

$$X_t = \varepsilon_t + 0.6\varepsilon_{t-1} + 0.6\varepsilon_{t-2} + 0.3\varepsilon_{t-3} + 0.7\varepsilon_{t-4}, \quad \{\varepsilon_t\} \sim_{\text{i.i.d.}} N(0,1)$$

with $\sigma = 1$. Assuming that the data are drawn from an MA model with an unknown order q, we determine the order and estimate MA coefficients from the data. The sample ACF plotted in Figure 2.3(c) shows that $\widehat{\rho}(k)$'s for $k > 4$ are almost always within the bounds $\pm 1.96/\sqrt{T}$ and therefore also within the bounds $\pm 1.96\{1 + 2\sum_{j=1}^{4} \widehat{\rho}^2(j)\}^{1/2}/\sqrt{T}$ (see (3.26)). This suggests that MA(4) might be a suitable candidate model. We apply AIC, AICC, BIC, and FPE to estimate q. The results are displayed in Figure 3.2(a). In general, we see a similar pattern as in the previous example. All of the AIC, AICC, and BIC select the correct order $q = 4$. We repeat the exercise with a sample of size $T = 50$. The order $q = 4$ is still selected by all of the three information criteria; see Figure 3.2(b). With $T = 50$, the maximum likelihood estimates for the MA coefficients are

$$0.56,\ 0.54,\ 0.17,\ \text{and}\ 0.71,$$

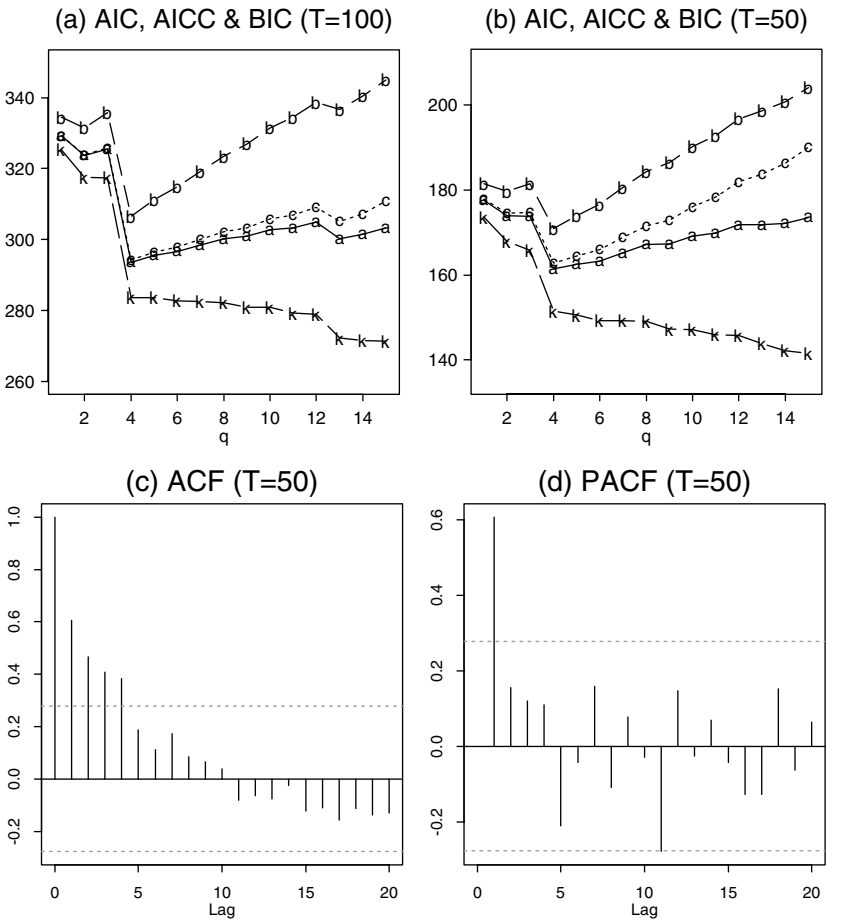


FIGURE 3.2. Example 3.2—Fitting an MA(q) model. Plots of AIC(q) (labeled "a"), AICC(q) (labeled "c") and BIC(q) (labeled "b") against q for sample size (a) $T = 100$ and (b) $T = 50$. Lines labeled with k are $-2\log$(maximum likelihood). (c) and (d) are ACF and PACF plots for a sample of size 50.

with the standard errors 0.12, 0.14, 0.20, and 0.18, respectively. The estimate for the variance of the white noise is 1.08. Note that with the sample size 50 the PACF plot in Figure 3.2(d) seems to suggest that the AR(1) model would be a reasonable alternative. However, the corresponding AIC value is 169.09, which is substantially larger than 161.43—the AIC value corresponding to the fitted MA(4) model. The difference is in the same order of magnitude as the difference of the AIC from order 3 to order 4. In fact, AR(1) is the best AR-model for the data according to AIC. However, it is a poor fitting overall. For example, the estimated variance for the white noise is 1.58, greatly exceeding the true value 1. This example indicates

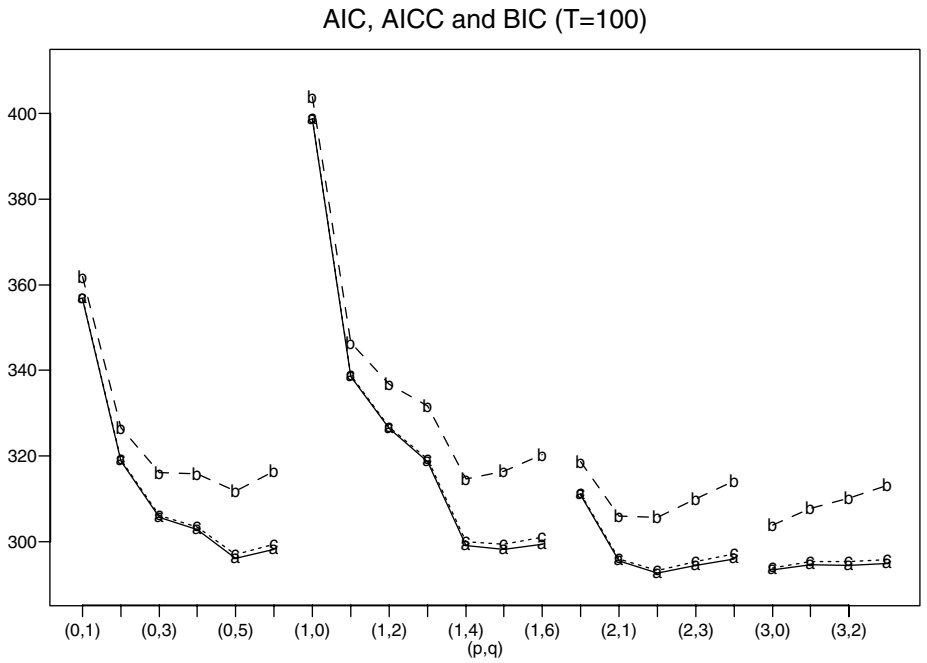


FIGURE 3.3. Example 3.3—Fitting an ARMA(p, q) model. Plots of AIC(p, q) (labeled "a"), AICC(p, q) (labeled "c") and BIC(p, q) (labeled "b") against (p, q). The four segments of curves correspond to $p = 0$ and $1 \le q \le 6$, $p = 1$ and $0 \le q \le 6$, $p = 2$ and $0 \le q \le 3$, and $p = 3$ and $0 \le q \le 3$, respectively.

that heuristic order selection based on PACF and ACF could sometimes be misleading.

Example 3.3 *(Fitting ARMA models)* We generate a sample of 100 from the model

$$X_t = 0.8X_{t-1} - 0.6X_{t-2} + \varepsilon_t + 0.7\varepsilon_{t-1} + 0.4\varepsilon_{t-2}, \quad \{\varepsilon_t\} \sim_{\text{i.i.d.}} N(0, 1).$$

Assuming that the true model is unknown, we will fit the data with an appropriate causal and invertible ARMA model. The sample ACF and PACF plotted in Figures 2.3 (e) and (f) show that $\widehat{\rho}(k)$ for $k > 6$ and $\widehat{\pi}(j)$ for $j > 3$ are not significantly different from 0. Thus, we search for the optimum ARMA(p, q) model with $0 \le p \le 3$ and $0 \le q \le 6$. The results are displayed in Figure 3.3. Both AIC and AICC select the true model with $p = q = 2$ whereas, BIC favors AR(3). Note that the values of AIC, AICC, or BIC for those two models are very close; see Figure 3.3. We regard both as "competitive models". The fitted ARMA(2, 2) model from the maximum likelihood method is

$$X_t = 0.72X_{t-1} - 0.64X_{t-2} + \varepsilon_t + 0.74\varepsilon_{t-1} + 0.41\varepsilon_{t-2}, \tag{3.27}$$

with $\mathrm{Var}(\varepsilon_t) = 0.96$. The fitted AR(3) model is

$$X_t = 1.29X_{t-1} - 1.10X_{t-2} + 0.34X_{t-3} + \varepsilon_t, \tag{3.28}$$

with $\mathrm{Var}(\varepsilon_t) = 1.01$. We will revisit this example in §3.5.

3.5 Diagnostic Checking

In view of the fact that a statistical model is only an approximation to reality, it is important to conduct postfitting diagnostic checking to see whether the fitted model explains the data well. In this section, we outline some standard methods for model diagnostics. The most frequently used techniques are residual-based methods that are designed to test whether the residuals derived from the fitted model behave like a white noise process. Some nonparametric tests to be introduced in §7.4 are also designed for this purpose. It is useful to bear in mind that the residual-based methods usually have little power to detect overfitting. Therefore, it is important to select an appropriate order, using criteria that penalize the model complexity.

3.5.1 Standardized Residuals

First, we define the *standardized residuals* from a fitted ARMA(p, q) model as follows. Based on the form of the likelihood function (3.9), the standardized residuals should be the estimates of the WN$(0, 1)$ random variables

$$R_j = (X_j - \widehat{X}_j)/(\sigma^2 r_{j-1})^{1/2}, \quad j = 1, \cdots, T.$$

Note that both $\widehat{X}_j$ and r_{j-1} depend on the unknown parameters $\mathbf{b} = (b_1, \cdots, b_p)^\tau$ and $\mathbf{a} = (a_1, \cdots, a_q)^\tau$. Replacing them (as well as σ^2) by the maximum likelihood estimators defined in (3.12), we obtain the standardized residuals

$$\widehat{R}_j = \{X_j - \widehat{X}_j(\widehat{\mathbf{b}}, \widehat{\mathbf{a}})\}/\{\widehat{\sigma}^2\, r_{j-1}(\widehat{\mathbf{b}}, \widehat{\mathbf{a}})\}^{1/2}, \quad j = 1, \cdots, T. \tag{3.29}$$

In the expression above, we write both $\widehat{X}_j$ and r_{j-1} explicitly as functions of $\widehat{\mathbf{b}}$ and $\widehat{\mathbf{a}}$ to indicate that they are approximated from the fitted model. If the model is correct, $\{\widehat{R}_j\}$ should resemble $\{R_j\}$, which is a WN$(0, 1)$ process. Furthermore $\{\widehat{R}_j\}$ should resemble an i.i.d. $N(0, 1)$ sequence if $\{\varepsilon_t\}$ in the model is Gaussian.

3.5.2 Visual Diagnostic

A simple and powerful diagnostic tool is to look at the time-series plot of $\{\widehat{R}_j\}$, and the plots of $\widehat{R}_j$ against the regressors (one in each time). Those

plots should resemble "purely random" pattern of the WN(0, 1) process if the fitted model is adequate. The inadequacy may be indicated by a systematic pattern such as deviation of the mean from zero, changing variation over time or over the regions of regressors, the existence of trend and cyclic components, and so on. We may superimpose the horizontal lines at ± 1.96 in the plots, and expect that about 95% of the residuals are within the two lines if the model fits the data.

The correlogram of $\{\widehat{R}_j\}$ may also be revealing. If the model is correct, we expect the sample ACF of $\{\widehat{R}_j\}$ to fall within the bounds $\pm 1.96/\sqrt{T}$ at about 95% of time lags. Box and Pierce (1970) modified the bounds $\pm 1.96/\sqrt{T}$ to take into account the dependence of the sample ACF of $\{\widehat{R}_j\}$ at different time lags; see also §9.4 of Brockwell and Davis (1991).

3.5.3 Tests for Whiteness

There are an abundance of tests for whiteness that can be applied to test whether $\{\widehat{R}_j\}$ is a white noise process. The tests presented in §7.4 are designed to test the whiteness based on the spectral density of the residuals. All of the tests are approximately valid, giving approximately correct levels of tests, since the parameters under the null hypothesis are typically estimated with rate $O_P(T^{-1/2})$.

Example 3.3 *(Continued)* We conduct a diagnostic check for both the fitted ARMA(2, 2) model (3.27) and AR(3) model (3.28). The standardized residuals calculated according to the formula (3.29) from both models are plotted in Figures 3.4 (a) and (b), and their ACFs and PACFs are plotted in Figures 3.4 (c)–(f). All of the plots provide stark evidence to support that the residuals from both models behave like a white noise process.

By applying the tests in §7.4, the p-values for the Fisher test (7.33), the generalized likelihood ratio test (7.37), the adaptive Neyman test (7.43), and the χ^2-test (7.40) are 0.66, 0.89 (with 8 d.f.), 0.89, and 0.36 (with 46 d.f.) for model (3.27) and .78, 0.84 (with 8 d.f.), 0.93, and 0.57 (with 47 d.f.) for model (3.28). Those tests lend further support to both fitted models.

Note that in principle two seemingly different ARMA models could effectively be the same if their MA(∞)- or AR(∞)-representations were almost the same. However, the two fitted models above are really different. Because model (3.27) is invertible, it can be represented as

$$X_t = \varepsilon_t + \sum_{j=1}^{\infty} d_j X_{t-j},$$

in which $|d_j| \geq 0.15$ for $1 \leq j \leq 6$, and the first three d_j's are -1.46, 1.31 and -0.38, respectively. This indicates that model (3.27) is substantially different from (3.28). This example illustrates that there could be two or

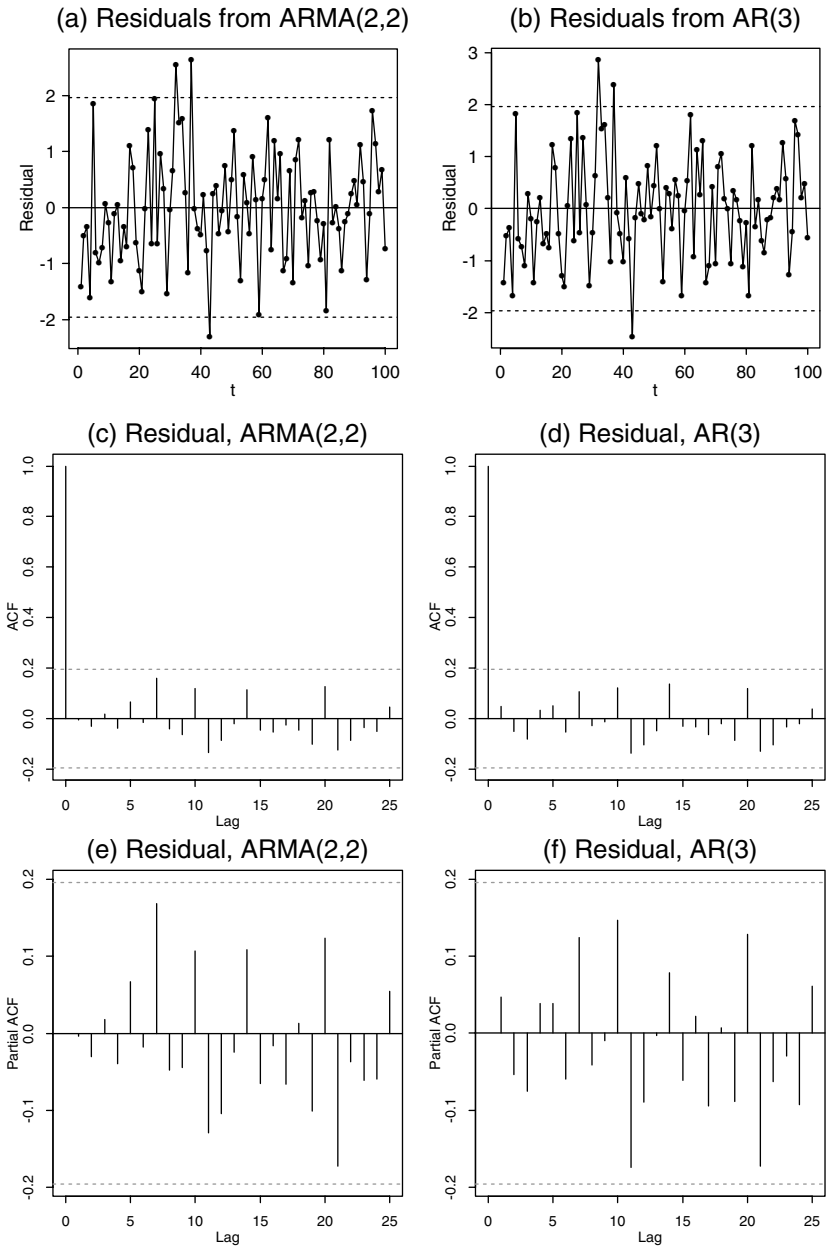


FIGURE 3.4. Example 3.3—analyzing residuals. Time-series plot, correlogram, and PACF plot for the standardized residuals from the fitted ARMA(2, 2) model (3.27) are displayed in (a), (c), and (e), respectively. Those for the standardized residuals from the fitted AR(3) model (3.28) are displayed in (b), (d), and (f).

more fundamentally different models that fit a given finite series almost equally well.

3.6 A Real Data Example—Analyzing German Egg Prices

Figure 3.5(a) displays the weekly egg prices at a German agricultural market between April 1967 and May 1990. The series is of length 300 and is the first quarter of a longer series extensively analyzed in Finkenstädt (1995). The sample mean and variance are 12.38 and 6.77, respectively. Since the data exhibit a clear nonstationary feature (see also its correlogram in Figure 3.5(c)), we take the first-order difference of the series. The differenced series are plotted in Figure 3.5(b), which looks more stationary-like. A scrutiny of Figures 3.5 (d) and (f) suggests that we may fit an ARMA(p, q) model with $p \leq 7$ and $q \leq 7$. Note $|\widehat{\rho}(k)|$ or $|\widehat{\pi}(k)| \geq 1.96/\sqrt{T}$ for $k = 18, 22, 25$, and a few other larger values. But with sample size 300, we prefer to fit the data with some small-order models first. We subtract the sample mean -0.015 from the data before the fitting. We select the model based on AICC simply because it is implemented in ITSM.

The optimum AR model based on AICC is AR(7) with the AICC-value 698.24. The estimated AR-coefficients $\widehat{b}_1, \cdots, \widehat{b}_7$ are

$$0.322,\ -0.159,\ 0.021,\ -0.004,\ -0.055,\ -0.023 \text{ and } -0.163.$$

The ratios $\widehat{b}_j/\{\mathrm{SE}(\widehat{b}_j)\}$ for $j = 1, \cdots, 7$ are

$$5.651,\ -2.666,\ 0.035,\ -0.071,\ -0.906,\ -0.378, \text{ and } -2.869,$$

where $\mathrm{SE}(\widehat{b}_j)$ stands for the standard error of $\widehat{b}_j$ (see §3.3.3). The small values ($<< 1.96$) of $|\widehat{b}_j/\mathrm{SE}(\widehat{b}_j)|$ are the significant supporting evidence to the hypothesis $b_j = 0$. Therefore, we fit the data with the same model again with the constraints $b_3 = b_4 = 0$. The fitted model becomes

$$X_t = 0.321X_{t-1} - 0.160X_{t-2} - 0.057X_{t-5} - 0.023X_{t-6} - 0.165X_{t-7} + \varepsilon_t, \tag{3.30}$$

where $\{\varepsilon_t\} \sim N(0, 0.567)$. By leaving out the terms X_{t-3} and X_{t-4}, the AICC-value has been reduced to 694.04.

On the other hand, we fit an MA(7) model to the data. The estimated MA-coefficients $\widehat{a}_1, \cdots, \widehat{a}_7$ are

$$0.320,\ -0.038,\ -0.054,\ -0.023,\ -0.048,\ -0.046, \text{ and } -0.195.$$

The ratios $\widehat{a}_j/\mathrm{SE}(\widehat{a}_j)$ for $j = 1, \cdots, 8$ are

$$5.541,\ -0.629,\ -0.896,\ -0.386,\ -0.790,\ -0.757, \text{ and } -3.210.$$

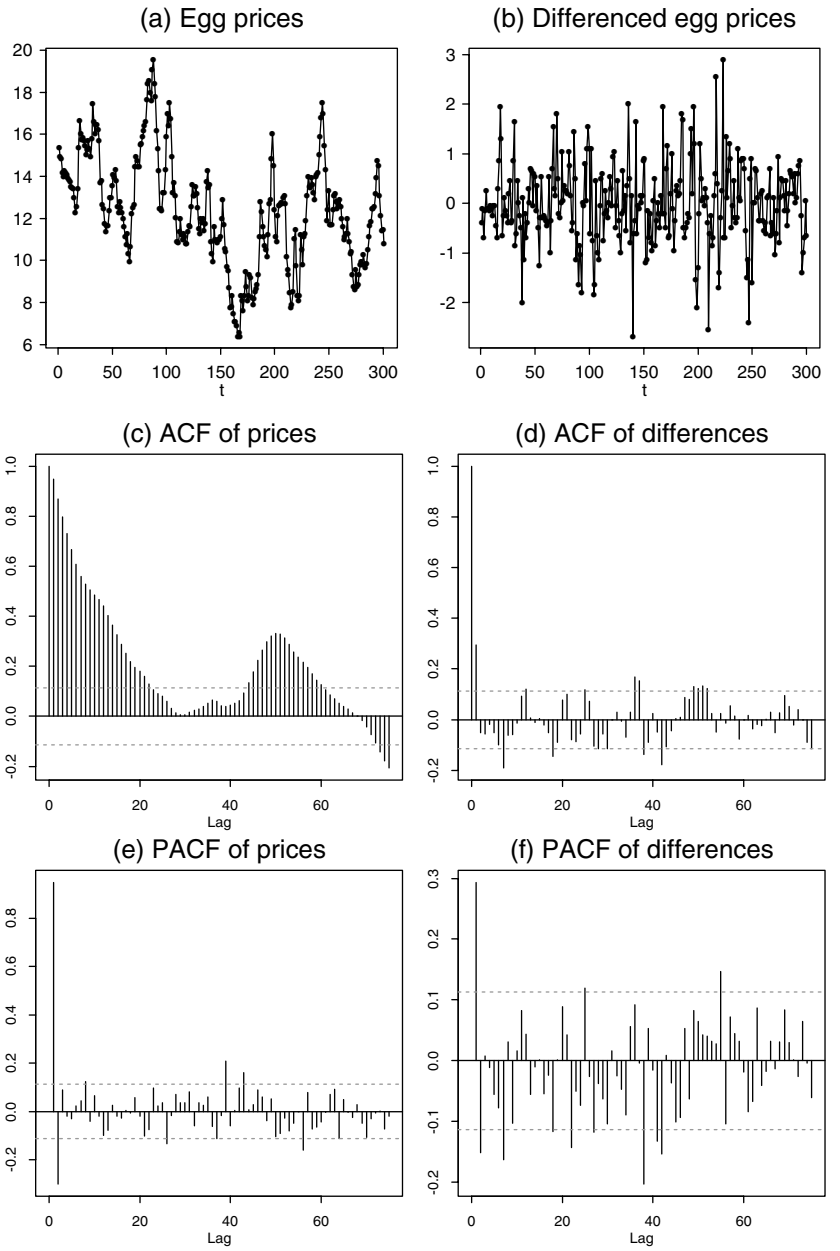


FIGURE 3.5. (a) Time plot of weekly German egg prices over a period of 300 weeks. (b) Lag-1 differenced series. (c) and (e) are ACF and PACF plots of the series displayed in (a). (d) and (f) are ACF and PACF plots of the differenced series displayed in (b).

Note that five out of seven of these ratios are smaller than 1 in absolute value. We use AICC to select the optimum model among the MA(7) family with at least one of a_2, a_3, a_4, a_5, and a_6 equal to 0. The selected model is

$$X_t = \varepsilon_t + 0.345\varepsilon_{t-1} - 0.173\varepsilon_{t-7} \tag{3.31}$$

with $\widehat{\sigma}^2 = 0.570$ and the standard errors of the two coefficients in the model above 0.054 and 0.051 in order. The corresponding AICC-value is 689.34, which is noticeably smaller than that of model (3.30).

Looking at the AR-coefficients in model (3.30), we search for an optimum ARMA(p, q) model with $p = 1$ or 2 and $1 \leq q \leq 7$. The model selected by AICC is the ARMA(1, 2)

$$X_t = 0.906X_{t-1} + \varepsilon_t - 0.619\varepsilon_{t-1} - 0.381\varepsilon_{t-2}, \tag{3.32}$$

with $\widehat{\sigma}^2 = 0.563$ and AICC= 690.58. The standard errors for the three coefficients in the model are 0.022, 0.053, and 0.052 in that order.

According to AICC, both models (3.31) and (3.32) are comparable with each other. We conduct diagnostic checking on both of them. The standardized residuals and their ACF and PACF plots are depicted in Figure 3.6. Slightly more than 5% (but $\leq 6\%$) of residuals from both models are beyond the bounds ± 1.96. But both ACF and PACF plots show that there still exists weak but significant autocorrelation in the residuals at some discrete lags. To see whether there is a genuine lack of fitting, we fit AR models to both residuals. In both cases AICC picks the optimum order $p = 0$, which indicates that the residuals are fairly white. By applying the tests in §7.4, the p-values for the Fisher test (7.33), the generalized likelihood ratio test (7.37), the adaptive Neyman test (7.43) and the χ^2-test (7.40) are 0.22, 0.82 (with 8 d.f.), 0.04, and 0.00 (with 50 d.f.) for model (3.31) and 0.22, 0.53 (with 9 d.f.), 0.01, and 0.00 (with 50 d.f.) for model(3.32). While both fittings passed the Fisher test and the generalized likelihood ratio test comfortably, they failed in the χ^2-test and model(3.32) also failed the adaptive Neyman test. This reflects the fact that there still exists some significant autocorrelation in the residuals; see Figures 3.5 (c) and (d). In fact, we set a_T equal to 50 in the adaptive Neyman test (7.43). The maximum value of T^*_{AN} was obtained at $m = 42$ for both models (3.31) and(3.32). One possible remedy is to include variables at the lags at which (partial) autocorrelation is significant. However it is in general difficult to interpret the resulting model. For example, it is hard to argue why the egg price at present depends more on the price 42 weeks ago rather than those in the last couple of weeks. Therefore, we decide to leave our fitted models unchanged.

Converting models (3.31) and (3.32) to the original egg price data $\{Y_t\}$, we obtain the two competitive ARIMA models

$$Y_t = Y_{t-1} + \varepsilon_t + 0.345\varepsilon_{t-1} - 0.173\varepsilon_{t-7}, \quad \{\varepsilon_t\} \sim \text{WN}(0, 0.570),$$

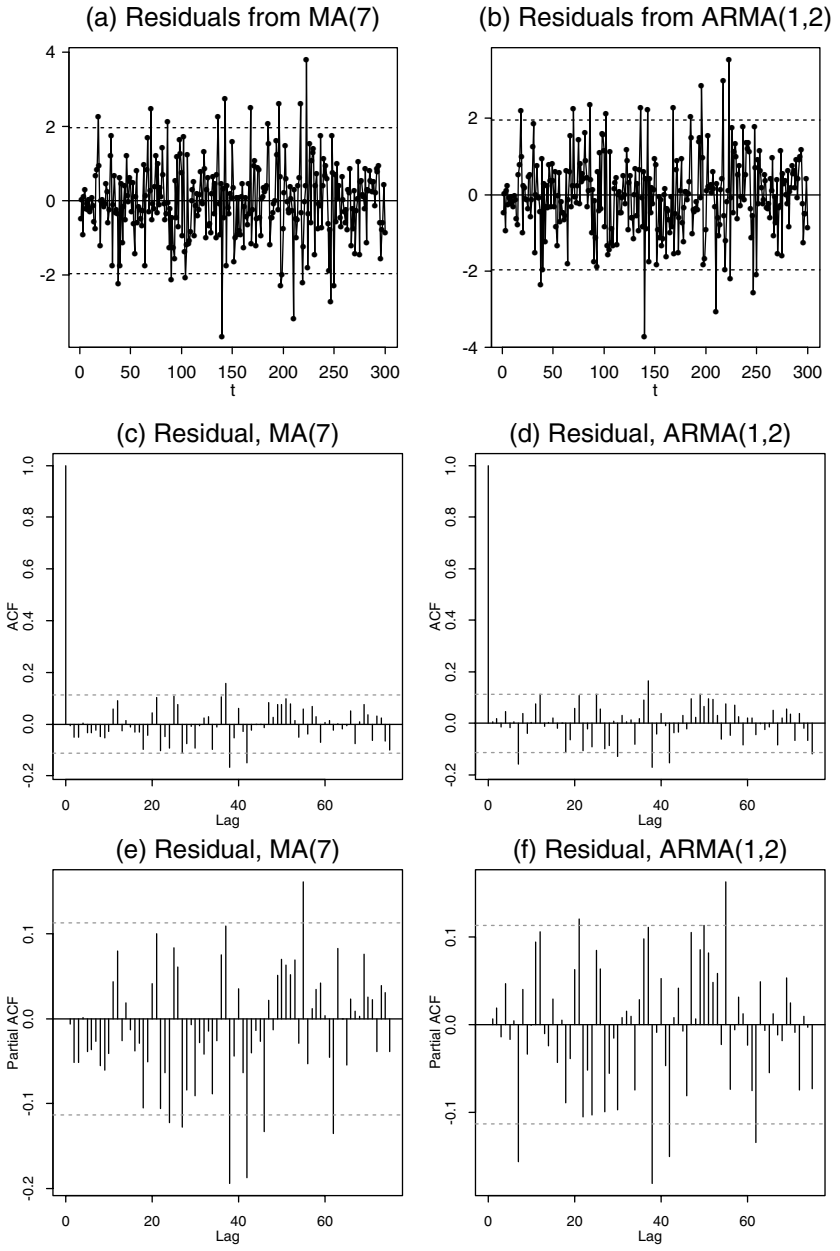


FIGURE 3.6. Fitting German egg price data. (a) Standardized residuals from fitted MA(7) model (3.31). (b) Standardized residuals from fitted ARMA(1, 2) model (3.32). (c) and (e) are the ACF-plot and PACF-plot of residuals from the MA(7) model. (d) and (f) are the ACF-plot and PACF-plot of residuals from the ARMA(1, 2) model.

and

$$\begin{aligned} Y_t &= -0.001 + 1.906Y_{t-1} - 0.906Y_{t-2} \\ &+ \varepsilon_t - 0.619\varepsilon_{t-1} - 0.381\varepsilon_{t-2}, \quad \{\varepsilon_t\} \sim \text{WN}(0, 0.563). \end{aligned}$$

3.7 Linear Forecasting

In this section, we discuss the forecasting for nonstationary ARMA (such as ARIMA) time series. We assume that the time series has mean 0 over all time. The practical implication of this assumption is that the mean function can be dealt with easily such that either the time series has a constant mean or we have fairly substantial prior knowledge on the way in which the mean function varies. It is intuitively clear that we can only forecast the future if the underlying process sustains certain stability over time.

First, we present a definition for the least squares m-step-ahead predictor, which is typically a linear predictor for linear time series. When the time series follows an AR equation (such as ARIMA$(p, 0, 0)$ processes), the m-step-ahead predictor can be recursively computed based on the AR equation (see (3.35) below). The mean squared predictive error can also be calculated in a recursive manner. The stationarity is not required.

For a general ARMA(p, q) process with $q > 0$, we need to assume that the process is invertible (although not necessarily stationary). This is an essential assumption that enables us to recover white noise signals $\{\varepsilon_t, t \leq T\}$ from observations $\{X_t, t \leq T\}$.

Although we will proceed with a general form of the ARMA model without the assumption of stationarity, we assume that both the form of the model and the coefficients in the model are known and remain unchanged, so we can predict the future based on the stable form of the model (instead of stationarity). The techniques presented are practically applicable to ARIMA models for which the parameters can be replaced by their estimators obtained from the differenced observations.

3.7.1 The Least Squares Predictors

Suppose that we have observations $X_1, \cdots, X_T$ from a time series $\{X_t\}$. We *forecast* a future observation X_{T+m} for some $m \geq 1$ based on the observations $X_T, \cdots, X_1$. The time series is not necessarily stationary. But we assume that $EX_t = 0$ and $E(X_t^2) < \infty$ for all t.

Definition 3.2 *For $m > -T$, we call $X_T(m)$, a (measurable) function of $X_T, \cdots, X_1$, the* least squares predictor *for X_{T+m} based on $X_T, \cdots, X_1$ if*

$$X_T(m) = \arg\inf_f E(X_{T+m} - f)^2,$$

where the infimum is taken over all the (measurable) functions of $X_T, \cdots, X_1$.

In the definition above, we allow m to be nonpositive for technical convenience. It is easy to see that $X_T(m) = X_{T-|m|}$ when $-T < m \leq 0$.

Proposition 3.2 $X_T(m) = E(X_{T+m}|X_T, \cdots, X_1)$.

Proof. Let $\mathbf{X}_T = (X_T, \cdots, X_1)^\tau$. For any measurable $f = f(\mathbf{X}_T)$,

$$\begin{aligned} E(X_{T+m} - f)^2 &= E\{X_{T+m} - E(X_{T+m}|\mathbf{X}_T)\}^2 \\ &+ E\{E(X_{T+m}|\mathbf{X}_T) - f\}^2 + 2B, \end{aligned}$$

where

$$\begin{aligned} B &= E[\{X_{T+m} - E(X_{T+m}|\mathbf{X}_T)\}\{E(X_{T+m}|\mathbf{X}_T) - f\}] \\ &= EE[\{X_{T+m} - E(X_{T+m}|\mathbf{X}_T)\}\{E(X_{T+m}|\mathbf{X}_T) - f\}|\mathbf{X}_T] \\ &= E[\{E(X_{T+m}|\mathbf{X}_T) - f\}E\{X_{T+m} - E(X_{T+m}|\mathbf{X}_T)|\mathbf{X}_T\}] \\ &= E[\{E(X_{T+m}|\mathbf{X}_T) - f\}\{E(X_{T+m}|\mathbf{X}_T) - E(X_{T+m}|\mathbf{X}_T)\}] \\ &= 0. \end{aligned}$$

Thus $E(X_{T+m} - f)^2 \geq E\{X_{T+m} - E(X_{T+m}|\mathbf{X}_T)\}^2$. ∎

3.7.2 Forecasting in AR Processes

Suppose that $\{X_t\}$ is defined by an autoregressive model (not necessarily stationary)

$$X_t = b_1 X_{t-1} + \cdots + b_p X_{t-p} + \varepsilon_t, \tag{3.33}$$

where $\{\varepsilon_t\} \sim \text{WN}(0, \sigma^2)$ and

$$E(\varepsilon_t|X_{t-1}, X_{t-2}, \cdots) = 0 \quad \text{for all } t. \tag{3.34}$$

The proposition below shows that $X_T(m)$ is a linear function of $X_T, \cdots, X_1$. Therefore $X_T(m)$ is also the best linear predictor which is discussed, for example, in §9.5 of Brockwell and Davis (1991).

Proposition 3.3 *Let* $\{X_t\}$ *be a process defined by (3.33) and* $T \geq p$. *Then*

$$X_T(m) = b_1 X_T(m-1) + \cdots + b_p X_T(m-p), \quad m = 1, 2, \cdots. \tag{3.35}$$

Furthermore, $X_T(m)$ *is a linear function*

$$X_T(m) = \varphi_1^{(m)} X_T + \cdots + \varphi_T^{(m)} X_1, \quad m = 1, 2, \cdots,$$

where $\{\varphi_j^{(m)}\}$ *are some constants.*

The proof of the proposition is trivial, as (3.35) follows from (3.33), (3.34) and Proposition 3.2, and the linearity follows from (3.35) and an obvious induction argument.

Let $\{X_t\}$ be an ARIMA$(p', d, 0)$ process now, that is

$$Y_t \equiv \nabla^d X_t = \sum_{j=0}^{d} \frac{d!}{j!(d-j)!}(-1)^j X_{t-j}, \quad t = 0, \pm 1, \pm 2, \cdots$$

is a stationary AR(p') process, where d is an integer. Therefore $\{X_t\}$ follows formally the AR equation with order $p = p' + d$,

$$\sum_{j=0}^{d} \frac{d!}{j!(d-j)!}(-1)^j X_{t-j} = \varepsilon_t + \sum_{k=1}^{p'} b_k \sum_{j=0}^{d} \frac{d!}{j!(d-j)!}(-1)^j X_{t-k-j},$$

where $\{\varepsilon_t\} \sim \text{WN}(0, \sigma^2)$. Thus, X_t can be expressed in an autoregressive form. With the observations $\{X_1, \cdots, X_T\}$, $X_T(m)$ for $m \geq 1$ can be evaluated recursively based on the equation above; see (3.35). In practice we replace $b_1, \cdots, b_{p'}$ by their maximum likelihood estimators, which are obtained based on the $(T-d)$ "observations" $\{\nabla^d X_j, j = d+1, \cdots, T\}$.

3.7.3 Mean Squared Predictive Errors for AR Processes

For the process $\{X_t\}$ defined by the autoregressive equation (3.33), it can be proved by induction that

$$X_t = \sum_{j=0}^{k} d_j \varepsilon_{t-j} + \sum_{j=k+1}^{k+p} \sum_{i=0}^{k} d_i b_{j-i} X_{t-j}, \quad k = 0, 1, 2, \cdots, \tag{3.36}$$

where d_j's are obtained recursively by (2.20) with $d_0 = 1$ and $a_j = b_{p+j} = 0$ for all $j \geq 1$. Let $t = T + m$ and $k = m - 1$ in the expression above, and we obtain that, for $T \geq p$,

$$X_{T+m} = \sum_{j=0}^{m-1} d_j \varepsilon_{T+m-j} + \sum_{l=0}^{p-1} \sum_{i=0}^{m-1} d_i b_{m+l-i} X_{T-l}.$$

By Proposition 3.2, $X_T(m)$ is equal to the second sum (i.e., the double sum) on the right-hand side of the expression above:

$$X_T(m) = \sum_{l=0}^{p-1} \sum_{i=0}^{m-1} d_i b_{m+l-i} X_{T-l}.$$

Hence, the residual is $X_{T+m} - X_T(m) = \sum_{1 \leq j < m} d_j \varepsilon_{T+m-j}$. This entails the proposition below immediately.

Proposition 3.4 *Let $\{X_t\}$ be a process defined by (3.33) and $T \geq p$. Then, the* mean squared predictive error *of $X_T(m)$ for $m \geq 1$ is given by*

$$\sigma_T^2(m) \equiv E\{X_{T+m} - X_T(m)\}^2 = \sigma^2 \sum_{j=0}^{m-1} d_j^2,$$

where d_j's are obtained recursively by (2.20) with $d_0 = 1$ and $a_j = b_{p+j} = 0$ for all $j \geq 1$. Furthermore, if for any $i, j \geq 1$ and any integer t,

$$E(\varepsilon_{t+i}\varepsilon_{t+j}|X_t, X_{t-1}, \cdots) = E(\varepsilon_{t+i}\varepsilon_{t+j}), \tag{3.37}$$

$\sigma_T^2(m)$ is also equal to the conditional mean squared prediction error, namely

$$\sigma_T^2(m) = E\big[\{X_{T+m} - X_T(m)\}^2 \big| X_T, X_{T-1}, \cdots\big].$$

When $\{X_t\}$ is causal, $\sum_j |d_j| < \infty$. In this case,

$$\sigma_T^2(m) \to \sigma^2 \sum_{j=0}^{\infty} d_j^2 = \mathrm{Var}(X_t) \quad \text{as } m \to \infty$$

(see §2.2.1). This is the same as the noise level and implies that $X_T(m) \to E(X_{T+m}) = 0$, which indicates that a long-term forecasting is nearly impossible.

Condition (3.37) is very mild. For example, it holds if, in addition, $\{\varepsilon_t\}$ is a sequence of independent random variables and ε_t is independent of $\{X_{t-k},\, k \geq 1\}$ for any t. Hence, the proposition above also shows that the conditional predictive error for those processes is independent of the values of $X_T, \cdots, X_1$. This illustrates the fact that the forecast based on linear time series models does not reflect the common knowledge that the risk of a prediction depends on the current state (i.e., $X_T, \cdots, X_1$). We will see in Chapter 10 that the dependence of prediction on its initial condition may be naturally captured in nonlinear time series models.

3.7.4 Forecasting in ARMA Processes

Suppose that $\{X_t\}$ is defined by an ARMA equation (not necessarily stationary)

$$X_t = \sum_{j=1}^{p} b_j X_{t-j} + \varepsilon_t + \sum_{j=1}^{q} a_j \varepsilon_{t-j}, \tag{3.38}$$

where $\{\varepsilon_t\} \sim \mathrm{WN}(0, \sigma^2)$ and condition (3.34) holds. It is easy to see that when $T \geq \max\{p, q\}$,

$$X_T(m) = \sum_{j=1}^{p} b_j X_T(m-j) + \sum_{j=1}^{q} a_j \varepsilon_T(m-j), \tag{3.39}$$

where $\varepsilon_T(i) = E(\varepsilon_{T+i}|X_T, \cdots, X_1)$. Obviously, $\varepsilon_T(i) = 0$ for all $i \geq 1$.

In order to evaluate those $\varepsilon_T(i)$'s in (3.38) with $i \leq 0$, we make two further assumptions. First, we assume that model (3.38) is invertible in the sense that

$$1 + a_1 z + \cdots + a_q z^q \neq 0 \quad \text{for all } |z| \leq 1.$$

Under this condition, model (3.38) admits an AR(∞) expression

$$X_t = \varepsilon_t + \sum_{j=1}^{\infty} c_j X_{t-j}, \quad \sum_{j=1}^{\infty} |c_j| < \infty, \tag{3.40}$$

where $c_0 = 1$ and $c_k = b_k - \sum_{j=0}^{k-1} c_j a_{k-j}$ for $k \geq 1$ (see (2.20)). In the recursive calculation of c_k's, we let $b_{p+j} = a_{q+j} = 0$ for all $j \geq 1$.

We also assume that we have all of the observations from time T back to negative infinite. Therefore, the least squares predictor based on all observations is

$$X_T(m) = E(X_{T+m}|X_T, X_{T-1} \cdots),$$

and $\varepsilon_T(j)$ is defined accordingly as

$$\varepsilon_T(j) = E(\varepsilon_{T+j}|X_T, X_{T-1} \cdots) = \begin{cases} 0 & j \geq 1, \\ \varepsilon_{T+j} & j \leq 0. \end{cases}$$

In practice, we assume that $X_j = 0$ for all $j \leq 0$. We expect that the modification has little impact as long as T is large relative to p and q since then the dependence between X_{T+j} ($j > 0$) and its remote past would be weak or very weak. Summarizing the findings above, we obtain the following propositions. Note that (3.41) can be established in the same manner as Proposition 3.4.

Proposition 3.5 *Let $\{X_t\}$ be defined by model (3.38) which is invertible. Then the least squares predictor for X_{T+m} based on $X_T, X_{T-1}, \cdots$ is defined recursively by (3.39), in which $\varepsilon_T(j) = 0$ for $j \geq 1$ and ε_{T+j} given by (3.40) for $j \leq 0$. Furthermore, the mean-square prediction error is given by*

$$\sigma_T^2(m) \equiv E\{X_{T+m} - X_T(m)\}^2 = \sigma^2 \sum_{j=0}^{m-1} d_j^2, \tag{3.41}$$

where d_j's are obtained recursively by (2.20) with $a_{q+j} = b_{p+j} = 0$ for all $j \geq 1$. Under the additional condition (3.37), it holds that

$$\sigma_T^2(m) = E\big[\{X_{T+m} - X_T(m)\}^2 \big| X_T, X_{T-1}, \cdots\big].$$

Example 3.4 Let us consider the invertible ARIMA(0, 1, 1) model

$$X_t - X_{t-1} = \varepsilon_t - a\varepsilon_{t-1}, \quad \{\varepsilon_t\} \sim \mathrm{WN}(0, \sigma^2),$$

where $|a| < 1$ and $E(\varepsilon_t | X_{t-1}, X_{t-2}, \cdots) = 0$ for any t. Since $|a| < 1$, we have the infinite series expansion

$$(1 - az)^{-1} = \sum_{j=0}^{\infty} a^j z^j, \quad |z| < 1/a.$$

Note that the model can be written as $(1 - B)X_t = (1 - aB)\varepsilon_t$. Hence

$$\begin{aligned} \varepsilon_t &= (1 - aB)^{-1}(1 - B)X_t = \sum_{j=0}^{\infty} a^j B^j (1 - B)X_t \\ &= X_t - (1 - a)\sum_{j=1}^{\infty} a^{j-1} X_{t-j}. \end{aligned}$$

It follows from Proposition 3.5 that

$$\widehat{X}_{T+1} \equiv X_T(1) = (1 - a)\sum_{j=0}^{\infty} a^j X_{T-j}, \tag{3.42}$$

and for $m \geq 2$

$$\begin{aligned} X_T(m) &= (1 - a)\sum_{j=0}^{\infty} a^j X_T(m - j - 1) \\ &= X_T(m-1) - a\left\{ X_T(m-1) - (1-a)\sum_{j=1}^{\infty} a^{j-1} X_T(m-j-1) \right\} \\ &= X_T(m-1) - a\{X_T(m-1) - X_T(m-1)\} = X_T(m-1). \end{aligned}$$

The latter follows also directly from (3.39).

Note that the predictor $\widehat{X}_{T+1}$ defined in (3.42) is a moving average of all of its lagged values with the coefficients exponentially decaying. If we define $\widehat{X}_t = E(X_t | X_{t-1}, X_{t-2}, \cdots)$ for all t, it follows from (3.42) that

$$\begin{aligned} \widehat{X}_{t+1} &= (1 - a)\sum_{j=0}^{\infty} a^j X_{t-j} = (1 - a)X_t + a\widehat{X}_t \\ &= X_t + a(\widehat{X}_t - X_t). \end{aligned}$$

The predictor $\widehat{X}_{t+1}$, giving weight $(1 - a)$ to the most recent observation X_t and weight a to its predicted value, is referred to as the *exponential*

smoothing, which is one of the most frequently used methods in forecasting. The example above shows that it is optimal if $\{X_t\}$ is an invertible ARIMA$(0, 1, 1)$ process. However, as a heuristic algorithm, the exponential smoothing (with $0 < a < 1$) has been widely used in practical forecasting. For example, it plays an important role in volatility forecasting for financial time series; see (8.54). In fact, the exponential smoothing may be viewed as a special type of kernel smoothing (see §6.2.5). This indicates that it is robust against model misspecification. For further discussion on exponential smoothing, see Gardner (1985).

All of the forecasting techniques described in this section are model-based in the sense that they are the best if the assumed model is correct and the parameters in the model are known. Practical experience shows that some heuristic forecasting algorithms, such as the exponential smoothing mentioned above, are well worth serious consideration; see Chapter 9 of Brockwell and Davis (1996).

This page intentionally blank

4

Parametric Nonlinear Time Series Models

The long-lasting popularity of ARMA models convincingly justifies the usefulness of linear models for analyzing time series data. Nevertheless, in view of the fact that any statistical model is an approximation to the real world, a linear model is merely a first step in representing an unknown dynamic relationship in terms of a mathematical formula. The truth is that the world is nonlinear! Therefore, it is not surprising that there exists an abundance of empirical evidence indicating the limitation of the linear ARMA family. To model a number of nonlinear features such as dependence beyond linear correlation, we need to appeal to nonlinear models. In this chapter, we present some parametric nonlinear time series models and their statistical inferences. §4.1 provides an introduction to the threshold modeling for conditional mean functions. §4.2 is devoted to ARCH modeling of nonconstant conditional variance functions—a phenomenon called *conditional heteroscedasticity*. A brief account on bilinear models is given in §4.3.

4.1 Threshold Models

Linear approximation serves as a powerful tool in quantitative scientific investigation in almost all disciplines. However, when we tackle nonlinear problems such as modeling nonlinear dynamics, a global linear law is often inappropriate. For example, it seems naive to assume that, in an economy or an animal population, the expanding phase is governed by the same linear dynamics as the contracting phase. Since a global quadratic (or a

higher-order) autoregressive form is typically unstable, a natural alternative would be to break a global linear approximation into several, each on a subset of the state-space. Under the umbrella of the threshold principle (§3.3 of Tong 1990), there is a class of nonlinear time series models that models nonlinear dynamics based on a "piecewise" linear approximation via partitioning a state-space into several subspaces. The partition is typically dictated by a so-called "threshold" variable. In this section, we present a simple but frequently used form—the threshold autoregressive model; focusing on the developments after Tong (1990). We introduce the techniques for estimation, testing, and model identification and illustrate those techniques through a real data example.

4.1.1 Threshold Autoregressive Models

Definition 4.1 *A* threshold autoregressive (TAR) model *with* k $(k \geq 2)$ *regimes is defined as*

$$X_t = \sum_{i=1}^{k} \{b_{i0} + b_{i1}X_{t-1} + \cdots + b_{i,p_i}X_{t-p_i} + \sigma_i \varepsilon_t\} I(X_{t-d} \in A_i), \quad (4.1)$$

where $\{\varepsilon_t\} \sim \text{IID}(0,1)$*,* $d, p_1, \cdots, p_k$ *are some unknown positive integers,* $\sigma_i > 0$ *and* b_{ij} *are unknown parameters, and* $\{A_i\}$ *forms a partition of* $(-\infty, \infty)$ *in the sense that* $\cup_{i=1}^{k} A_i = (-\infty, \infty)$ *and* $A_i \cap A_j = \varnothing$ *for all* $i \neq j$*.*

In the model above, we fit on each A_i a linear form. The partition is dictated by the threshold variable X_{t-d}, and d is called a delay parameter. It is often (but not always) the case that $A_i = (r_{i-1}, r_i]$ with $-\infty = r_0 < r_1 < \cdots < r_k = \infty$. In this case, r_i's are called thresholds. This model, first introduced by H. Tong in 1977, is in fact a special type of threshold model called self-exciting threshold model; see Tong and Lim (1980) and Tong (1990). It has been widely used to model nonlinear phenomena in diverse areas, including economics (Tiao and Tsay 1994; Hansen 1999), environmental sciences (Mélard and Roy 1988), neural science (Brillinger and Segundo 1979), finance (Li and Lam 1995), hydrology (Tong and Lim 1980), physics (Pemberton 1985), and population dynamics (Stenseth et al. 1999). Its success partially lies in its simplicity in terms of both model-fitting and, perhaps more importantly, model-interpretation. By modeling the nonlinearity via partitioning the state-space, the stationarity may be preserved. This is in marked contrast to change-point models for which the regime-switch happens according to time, resulting in nonstationary processes. Unfortunately, our knowledge of the TAR model is still developing. We do not have a comprehensive theory and methodology as we do for linear ARMA models.

It is easy to see from Theorem 2.4 that model (4.1) admits a strictly stationary solution if (a) $\sigma_1 = \cdots = \sigma_p$ and (b) either $\max_{1 \leq i \leq k} \sum_{j=1}^{p_i} |b_{ij}| < 1$

or $\sum_{j=1}^{p} \max_{1\le i\le k} |b_{ij}| \le 1$ with $p = \max_{1\le i\le k} p_i$. Note that these conditions are sufficient but not necessary; see model (2.14).

The linear correlation of two random variables is explicitly defined. Consequently, autocorrelation of a time series is well-captured by its autocorrelation function (ACF), which, as we have witnessed in Chapter 3, plays a key role in modeling the autolinear relationship. (Note that a PACF is a function of the corresponding ACF; see Proposition 2.3.) Unfortunately, there exists no analog of the ACF to represent nonlinear dependence in general. Various attempts have been made to define appropriate measures for nonlinear dependence/association, either localized or global. But none of them are as simple and illustrative as the ACF and PACF in analyzing linear relationships. We have in fact a paradox here; a nonlinear phenomenon is typically more complex and more difficult to model than a linear one, and the available tools are much less comprehensive and less effective. Therefore data-exploratory and data-analytic techniques such as various plots (§5.2 of Tong 1990), background information, and nonparametric and semi-parametric techniques play important roles in identifying an appropriate (parametric) form in nonlinear modeling. A statistical test for linearity is a routine practice to testify to nonlinearity. We will illustrate some of those ideas in case studies in §4.1.4 below.

For fitting a low-dimensional structure, the scatter plots that plot a time series variable against its lagged values are almost as insightful as any more sophisticated tools. To illustrate this idea, we consider a simple TAR model with two regimes as follows:

$$X_t = \begin{cases} -0.7X_{t-1} + \varepsilon_t, & X_{t-1} \ge r, \\ 0.7X_{t-1} + \varepsilon_t, & X_{t-1} < r, \end{cases} \tag{4.2}$$

where ε_t's are independent $N(0, 0.5^2)$ variables. We generate four sample series (of length 500 for each) from the model above with r equal to, respectively, $-\infty, -1, -0.5$, and 0. Figure 4.1 presents the scatter plots of those four sample series. For $r = -\infty$, the model reduces to a linear AR(1). For all three other cases, the nonlinearity is clearly displayed in those plots. On the other hand, ACF and PACF plots, although still useful, cannot be taken as ultimate measures for the dependence. For example, Figures 4.2 (a) and (d) indicate that there is hardly any significant autocorrelation when $r = -1$ in spite of the intimate dependence between X_t and X_{t-1} defined by (4.2).

It is worth pointing out that the usefulness of TAR models is due to the fact that the class of piecewise linear functions may typically provide a simple and easy-to-handle approximation to a more sophisticated nonlinear function. Figure 4.3 displays some examples for which different nonlinear functions can be approximated by piecewise linear functions with two, three or four regimes, which may be viewed as linear splines with two, three or four knots; see §6.4. In practice, a good fitting from a TAR model does

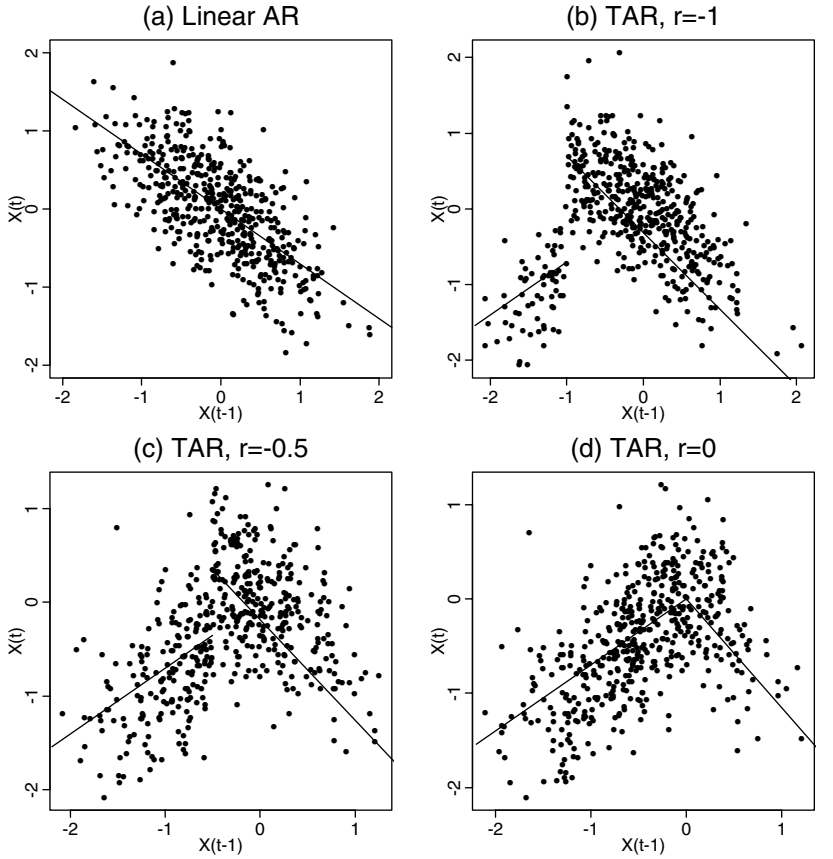


FIGURE 4.1. Scatter plots of the samples generated from TAR model (4.2) with (a) $r = -\infty$, (b) $r = -1$, (c) $r = -0.5$, and (d) $r = 0$. The solid lines are the true regression functions.

not necessarily imply that the underlying process is exactly piecewise linear. But a practically meaningful interpretation is often entertained when each regime in a fitted model represents a different characteristic of the underlying nonlinear dynamic. To illustrate this point, we report below the fitted TAR model for the quarterly U.S. real GNP data due to Tiao and Tsay (1994) in which the regimes are defined in terms of two (instead of one) threshold variables.

Let $Y_0, Y_1, \cdots, Y_{176}$ denote the quarterly US real GNP from February 1947 to January 1991, a total of 177 observations. Figure 4.4 is the time series plot of the growth rate series

$$X_t \equiv \log(Y_t) - \log(Y_{t-1}), \quad t = 1, \cdots, 176.$$

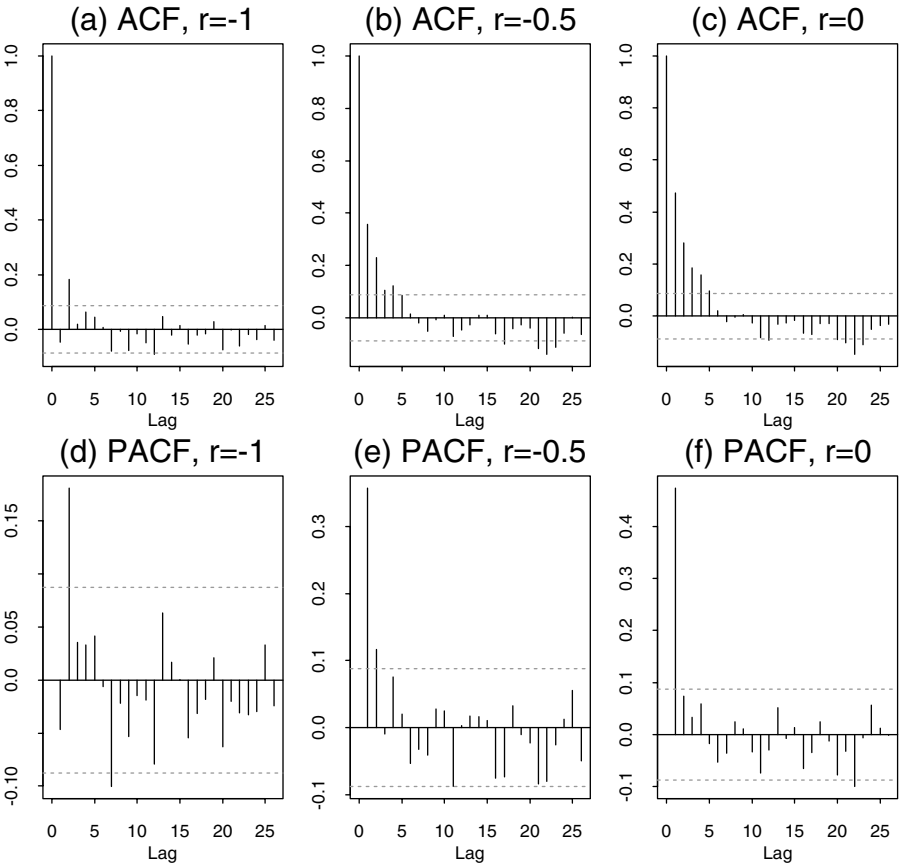


FIGURE 4.2. Sample ACF and PACF plots for time series generated from the TAR model (4.2).

Tiao and Tsay (1994) fitted the growth rates with the following TAR models with four regimes (discarding the two insignificant intercepts):

$$X_t = \begin{cases} -0.015 - 1.076X_{t-1} + \varepsilon_{1,t}, & X_{t-1} \le X_{t-2} \le 0, \\ 0.630X_{t-1} - 0.756X_{t-2} + \varepsilon_{2,t} & X_{t-1} > X_{t-2},\ X_{t-2} \le 0, \\ 0.006 + 0.438X_{t-1} + \varepsilon_{3,t} & X_{t-1} \le X_{t-2},\ X_{t-2} > 0, \\ 0.443X_{t-1} + \varepsilon_{4,t}, & X_{t-1} > X_{t-2} > 0. \end{cases} \tag{4.3}$$

This model can be interpreted as follows.

> The first regime (i.e., $X_{t-1} \le X_{t-2} \le 0$) denotes a recession period in which the economy changed from a contraction to an even worse one. Only six observations were from this recession phase. Furthermore it is reassuring to see the negative explosive nature of the regression function in this regime, indicating that

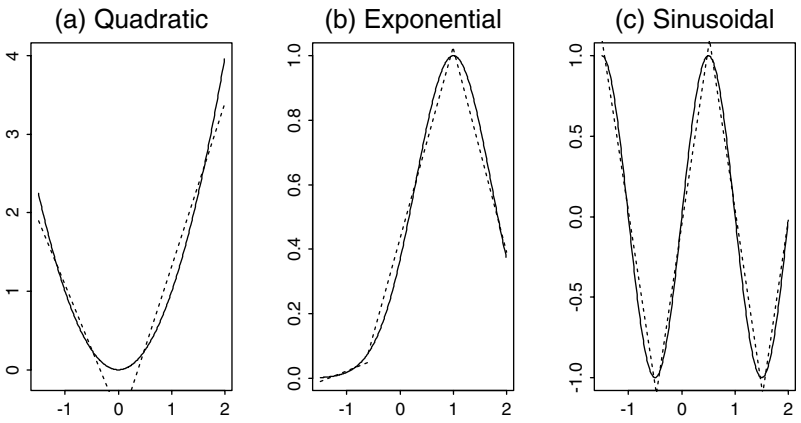


FIGURE 4.3. Piecewise linear approximations (dotted lines) to nonlinear functions (solid curves): (a) x^2, (b) $\exp\{-(x-1)^2\}$, and (c) $\sin(\pi x)$.

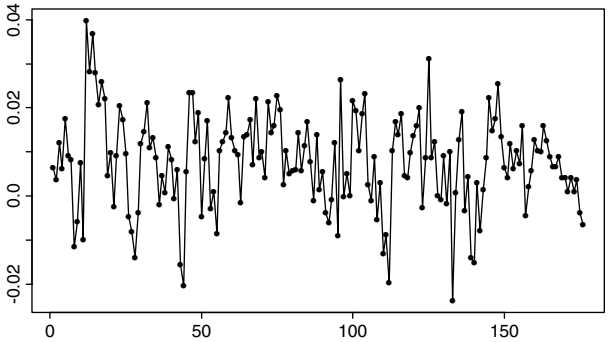


FIGURE 4.4. Time plot of growth of U.S. quarterly real GNP (February 1947–January 1991).

the economy usually recovers quickly from the recession period. In fact, within given series there were only three occasions in which two consecutive negative growth periods were observed.

The second regime (i.e., $X_{t-1} > X_{t-2}$ and $X_{t-2} \leq 0$) corresponds to a period in which the economy was in contraction but also improving. In this phase, the regression function tends to be positive, suggesting that the economy is more likely to grow continuously out of recession once a recovery has started.

The third regime (i.e., $X_{t-1} \leq X_{t-2}$ and $X_{t-2} > 0$) denotes a period in which the economy was reasonable but the growth declined. The fourth regime (i.e., $X_{t-1} > X_{t-2} > 0$) denotes an

expansion period in which the economy became stronger. The fitted linear forms in these two regimes are similar, both with an autoregressive coefficient around 0.44.

All of the coefficients in model (4.3) are statistically significant. For a detailed statistical analysis of this model, see Tiao and Tsay (1994).

4.1.2 Estimation and Model Identification

Suppose that $X_1, \cdots, X_T$ are observed values from model (4.1) with k given. Based on these observations, we estimate the parameters' b_{ij}'s, σ_i's, and d and determine the orders' p_i's and the partition $\{A_i\}$.

First, we assume that the partition $\{A_i\}$ and the orders p_i's are known. To simplify the notation, we assume $d \le p \equiv \max_{1\le i\le k} p_i$. Then, the least squares estimators for the autoregressive coefficients $\mathbf{b}_i \equiv (b_{i0}, b_{i1}, \cdots, b_{i,p_i})^\tau$, $i = 1, \cdots, k$, are defined as $\widetilde{\mathbf{b}}_i$'s, where $\widetilde{\mathbf{b}}_1, \cdots, \widetilde{\mathbf{b}}_k$ and $\widetilde{d}$ minimize

$$\sum_{i=1}^{k} L(\mathbf{b}_i, d; A_i) \tag{4.4}$$

over all possible real values of b_{ij}'s and integer values $1 \le d \le p$, in which

$$L(\mathbf{b}_i, d; A_i) \equiv \sum_{\substack{X_{t-d}\in A_i \\ p<t\le T}} \{X_t - (b_{i0} + b_{i1}X_{t-1} + \cdots + b_{i,p_i}X_{t-p_i})\}^2. \tag{4.5}$$

The minimization above may be viewed as a two-step process: for each fixed d, we first minimize (4.5) for $i = 1, \cdots, k$ and then choose $\widetilde{d}$ to minimize (4.4). Note that, for a fixed d, the minimizer $\widetilde{\mathbf{b}}_i(d)$ of (4.5) is an ordinary least squares estimator of a linear regression model, and it therefore can be obtained explicitly. In case there exists more than one minimizer, we always choose the smallest d as our estimator for the delay parameter. Now, an estimator for the variance σ_i^2 is defined as

$$\widetilde{\sigma}_i^2 = \frac{1}{T_i} L(\widetilde{\mathbf{b}}_i, \widetilde{d}; A_i), \tag{4.6}$$

where T_i is the number of elements in the set $\{t : \ p < t \le T \text{ and } X_{t-\tilde{d}} \in A_i\}$, $i = 1, \cdots k$.

If we assume that ε_t is Gaussian, the least squares estimation derived above is not necessarily asymptotically equivalent to the conditional maximum likelihood estimation. In fact, the conditional maximum likelihood estimators for $\mathbf{b}_i$'s, σ_i's and d can be obtained from maximizing

$$-\frac{1}{2}\sum_{i=1}^{k} L(\mathbf{b}_i, d; A_i)/\sigma_i^2 - \frac{1}{2}\sum_{i=1}^{k} T_i \log \sigma_i.$$

For a given d, b_i is obtained by minimizing $L(\mathbf{b}_i, d; A_i)$, as in (4.4). However, d is obtained by minimizing

$$\sum_{i=1}^{k} L(\mathbf{b}_i, d; A_i)/\sigma_i^2$$

instead of (4.4). For a given $\widetilde{\mathbf{b}}_i$ and $\widetilde{d}$, the conditional MLE for σ_i is obtained by (4.6). The schematic algorithm is as follows. For each given d, via the least-squares method, we obtain $\widetilde{\mathbf{b}}_i$ and $\widetilde{\sigma}_i^2$ from (4.6). This yields a sequence of the conditional likelihood

$$\sum_{i=1}^{k} L(\widetilde{\mathbf{b}}_i, d; A_i)/\widetilde{\sigma}_i^2,$$

indexed by d, from which an estimate of d can be obtained. We do not pursue this further since the efficiency-gain over the least squares estimation is significant only if the discrepancy among σ_i^2's is large.

In practice, the partition $\{A_i\}$ is often unknown and is often assumed to be of the form $A_i = (r_{i-1}, r_i]$ with $-\infty = r_0 < r_1 < \cdots < r_k = \infty$. In theory, we may determine the partition in the manner of an exhausting search as follows: for a given collection of partitions $\{A_i\}$, let $L(\{A_i\}) = \sum_{1\le i\le k} L(\widetilde{\mathbf{b}}_i, \widetilde{d}; A_i)$, the minimal value of (4.4); we search for the partition $\{\widehat{A}_i\}$ that minimizes $L(\{A_i\})$. In practice, k often takes a small value such as 2, 3, or 4, and threshold r_i's are searched within certain inner sample ranges. For example when $k = 2$, we may search for r_1 within, for example, the 60% inner sample range.

Now, we define the *least squares estimators* that minimize (4.4) with $A_i = \widehat{A}_i$ as $\widehat{\mathbf{b}}_i$ and $\widehat{d}$ and define

$$\widehat{\sigma}_i^2 = \frac{1}{T_i} L(\widehat{\mathbf{b}}_i, \widehat{d}; \widehat{A}_i), \quad i = 1, \cdots, k. \tag{4.7}$$

To determine the autoregressive order p_i's, we may define a generalized AIC as:

$$\mathrm{AIC}(\{p_i\}) = \sum_{i=1}^{k} [T_i \log\{\widehat{\sigma}_i^2(p_i)\} + 2(p_i + 1)],$$

where $\widehat{\sigma}_i^2(p_i) \equiv \widehat{\sigma}_i^2$ is given by (4.7). We choose $\{p_i\}$ such that the corresponding AIC-value obtains the minimum. The penalty for the number of regimes is reflected in the sum over p_i in the expression above. Of course, the criteria such as BIC or AICC can be adopted in the same manner in this context; see §3.4.

To consider the asymptotic properties of the estimators, we assume that $\{X_t\}$ generated by (4.1) is strictly stationary and ergodic with finite second moment. It can be shown that if the partition $\{A_i\}$ and the delay

parameter d are given in model (4.1), the least squares estimator for $\mathbf{b}_i$ is asymptotically normal in the sense that

$$T_i^{1/2}\{\widetilde{\mathbf{b}}_i(d) - \mathbf{b}_i\} \xrightarrow{D} N(0, \sigma_i^2 \mathbf{W}_i^{-1}), \tag{4.8}$$

where

$$\mathbf{W}_i = \begin{pmatrix} 1 & \mu \mathbf{1}^\tau \\ \mu \mathbf{1} & E(\boldsymbol{\xi}_i \boldsymbol{\xi}_i^\tau) \end{pmatrix}, \quad \boldsymbol{\xi}_i = (\xi_1, \cdots, \xi_{p_i})^\tau,$$

$\mathbf{1}$ is the $p_i \times 1$ vector with common components 1, $\mu = E\xi_t$, and

$$\xi_t = b_{i0} + b_{i1}\xi_{t-1} + \cdots + b_{i,p_i}\xi_{t-p_i} + e_t, \quad \{e_t\} \sim \text{WN}(0,1)$$

(see also Theorem 3.2). Unfortunately, $\{A_i\}$ and d are typically unknown in practice. The asymptotic properties of the estimators are more complicated, depending on whether the regression function $E(X_t | X_{t-k} = x_{t-k}, k \geq 1)$ is continuous (such as in Figure 4.1(d)) or not (such as in Figures 4.1 (b) and (c)). Intuitively, the discontinuity displayed in Figure 4.1(b) makes the estimation of the threshold r easier than that in a continuous case such as in Figure 4.1(d). Theorem 4.2 below, due to Chan (1993a), shows that the estimator for the threshold converges at the rate T^{-1} (in contrast to the conventional rate $T^{-1/2}$) when the regression function is discontinuous. For the asymptotic properties for continuous cases, see Chan and Tsay (1998).

The two theorems below concern a special case of model (4.1) with $k = 2$ and $p_1 = p_2 = p$ given. In this case the estimation of the partition $\{A_i\}$ reduces to the estimation of the single threshold $r \equiv r_1$. We denote its estimator as $\widehat{r}$.

Theorem 4.1 (Chan 1993a) *Suppose that $\{X_t\}$ satisfies (4.1) with $k = 2$ and $p_1 = p_2 = p$ is ergodic and strictly stationary with finite second moments. Suppose that the joint density function $(X_1, \cdots, X_p)$ is positive everywhere. Then, all of the estimators $\widehat{\mathbf{b}}_1$, $\widehat{\mathbf{b}}_2$, $\widehat{\sigma}_1^2$, $\widehat{\sigma}_2^2$, $\widehat{r}$, and $\widehat{d}$ are strongly consistent.*

Theorem 4.2 (Chan 1993a) *In additional to the condition of Theorem 4.1, we assume that:*

(i) The Markov chain $\mathbf{X}_t = (X_t, X_{t-1}, \cdots, X_{t-p+1})^\tau$ is geometrically ergodic.

(ii) ε_t has a positive and uniformly continuous density function, and $E(\varepsilon_t^4 + X_t^4) < \infty$.

(iii) The autoregressive function is discontinuous (i.e., there exists $\mathbf{z} = (1, z_{p-1}, z_{p-2}, \cdots, z_0)^\tau$ with $z_{p-d} = r$ such that $\mathbf{z}^\tau(\mathbf{b}_1 - \mathbf{b}_2) \neq 0$).

Then $T(\widehat{r} - r) = O_p(1)$, and $(\widehat{r} - r)$ is asymptotically independent of $(\widehat{\mathbf{b}}_1 - \mathbf{b}_1, \widehat{\mathbf{b}}_2 - \mathbf{b}_2)$. Furthermore $\sqrt{T_i}(\widehat{\mathbf{b}}_i - \mathbf{b}_i)$ is asymptotically normal with mean 0 and variance $\sigma_i^2 \mathbf{W}_i^{-1}$ defined in (4.8), $i = 1, 2$.

Note that $\widehat{d}$ takes only integer values. Therefore, it holds almost surely that $\widehat{d}$ is equal to d eventually for all large T; see Theorem 4.1. Theorem 4.2 shows that due to the discontinuity, $\widehat{r} - r$ converges to 0 faster by a factor $T^{-1/2}$ than $\widehat{\mathbf{b}}_i - \mathbf{b}$ does. (The asymptotic distribution of $T(\widehat{r} - r)$ is given in Chan 1993a.) Therefore, both d and r may be viewed as known as far as the asymptotic distributions of $\widehat{\mathbf{b}}_i$'s are concerned; see Theorem 4.2 and (4.8). Thus, approximate confidence intervals for the coefficients b_{ij} may be easily constructed based on (4.8). On the other hand, the required geometrical ergodicity condition in Theorem 4.2 holds if (a) $\sigma_1 = \sigma_2$ and (b) either $\max_{1 \le i \le 2} \sum_{j=1}^{p_i} |b_{ij}| < 1$ or $\sum_{j=1}^{2} \max_{1 \le i \le k} |b_{ij}| \le 1$ with $p = \max_{1 \le i \le k} p_i$; see Example 2.1 and Theorem 2.4.

4.1.3 Tests for Linearity

Since there is a lack of a general measure of nonlinear dependence, testing for linearity becomes a routing exercise to check nonlinearity in fitting nonlinear models. There are now more than a dozen of such tests available (§5.3 of Tong 1990), which may be divided into two categories: portmanteau tests, which test for departure from linear models without specifying alternative models, and the tests designed for some specific alternatives. More recently, the tests that make use of nonparametric and semiparametric fitting have received considerable attention; see Chapter 9. We introduce below the likelihood ratio test for a linear model against a TAR alternative with two regimes due to Chan and Tong (1990) and Chan (1990b). Although the test is designed for a specified alternative, it may be applied to test for a departure to a general smooth nonlinear function since a piecewise linear function will provide a better approximation than that from a (global) linear function. This is in the same spirit as Cox (1981), who suggested the use of quadratic or cubic regression for testing nonlinearity.

Let $X_1, \cdots, X_T$ be observations from a strictly stationary process. We test the null hypothesis that $\{X_t\}$ is from a linear AR(p) model,

$$H_0 : \ X_t = \theta_0 + \sum_{j=1}^{p} \theta_j X_{t-j} + \varepsilon_t,$$

against the alternative,

$$H_1 : \ X_t = \theta_0 + \sum_{j=1}^{p} \theta_j X_{t-j} + I(X_{t-d} \le r) \left\{ \varphi_0 + \sum_{j=1}^{p} \varphi_j X_{t-j} \right\} + \varepsilon_t,$$

which specifies a TAR model with two regimes. In the expression above $\{\varepsilon_t\} \sim_{\text{i.i.d.}} N(0, \sigma^2)$ with $\sigma^2 \in (0, \infty)$, and p and d are known positive integers. Further, we assume that the threshold r lies inside a known bounded closed interval $\mathcal{I}_r$.

Since $\{\varepsilon_t\}$ is Gaussian, the *likelihood ratio test* will reject H_0 for large values of the test statistic

$$\{T - \max(p,d)\}\log(\widehat{\sigma}_0^2/\widehat{\sigma}^2),$$

which is equivalent to the F-test

$$S_T = \{T - \max(p,d)\}(\widehat{\sigma}^2 - \widehat{\sigma}_0^2)/\widehat{\sigma}^2, \tag{4.9}$$

where the factor $T' = \{T - \max(p,d)\}$ is a normalizing constant,

$$\begin{aligned}
\widehat{\sigma}^2 &= \inf_{r\in\mathcal{I}_r,\{\theta_i\},\{\varphi_i\}} \frac{1}{T'} \sum_{t=\max(p,d)+1}^{T} \left\{ X_t - \theta_0 - \sum_{j=1}^{p} \theta_j X_{t-j} \right. \\
&\qquad \left. - I(X_{t-d} \le r)\left(\varphi_0 + \sum_{j=1}^{p} \varphi_j X_{t-j}\right)\right\}^2, \\
\widehat{\sigma}_0^2 &= \inf_{\{\theta_i\}} \frac{1}{T'} \sum_{t=\max(p,d)+1}^{T} \left\{ X_t - \theta_0 - \sum_{j=1}^{p} \theta_j X_{t-j} \right\}^2.
\end{aligned}$$

Based on a Poisson clumping heuristic, Chan (1991) developed the following approximation for the significance levels (when y is large) of the above test:

$$P\{S_T > y|H_0\} \approx 1 - \exp\left\{-2\,\chi_{p+1}^2(y)\left(\frac{y}{p+1} - 1\right)\sum_{i=1}^{p+1}\int_{\mathcal{I}_r} h_i(x)dx\right\}, \tag{4.10}$$

where $\chi_j^2(\cdot)$ denotes the probability density function of the χ^2-distribution with j degrees of freedom, $h_i(x) = dJ_i(x)/dx$, $m_2 = E_0(X_t^2)$ and

$$J_i(x) = \frac{1}{2}\log\left\{\frac{P_0(X_t \le x)}{P_0(X_t > x)}\right\}, \qquad 1 \le i < p,$$

$J_p(x)$ and $J_{p+1}(x)$ are the roots of the equation $y^2 - by + c = 0$ with $b = E_0\{(1 + X_t^2/m_2)I(X_t \le x)\}$, and

$$c = \frac{1}{m_2}[P_0(X_t \le x)E_0\{X_t^2 I(X_t \le x)\} - \{E_0(X_t I(X_t \le x))\}^2].$$

In the expressions above, P_0 and E_0 denote, respectively, the probability law and the expectation under H_0.

Based on (4.10), we can tabulate the approximate upper percentage points for the null distribution of S_T. Tables 4.1 and 4.2 were extracted from Chan (1991). Percentage points for some other values of the order p may be approximately obtained by interpolation.

The approximation for the significance levels of the likelihood ratio tests can also be obtained via the bootstrap; see §9.2.3.

TABLE 4.1. Upper percentage points for the asymptotic null distribution of S_T, with $\mathcal{I}_r$ being the 50% inner sample range.

order p	10%	5%	2.5%	1%	0.1%
0	6.12	7.75	9.33	11.36	16.33
1	9.27	11.18	12.94	15.16	20.45
2	11.34	13.38	15.26	17.60	23.13
3	13.25	15.42	17.39	19.83	25.57
4	15.07	17.33	19.39	21.93	27.87
5	16.80	19.16	21.30	23.93	30.04
6	18.48	20.93	23.14	25.58	32.13
9	23.28	25.96	28.36	31.29	38.01
12	27.83	30.70	33.27	36.39	43.50
15	32.20	35.25	37.97	41.26	48.72
18	36.45	39.67	42.52	45.96	53.74

TABLE 4.2. Upper percentage points for the asymptotic null distribution of S_T, with $\mathcal{I}_r$ being the 80% inner sample range.

order p	10%	5%	2.5%	1%	0.1%
0	7.61	9.21	10.77	12.80	17.75
1	11.05	12.85	14.55	16.72	21.94
2	13.26	15.18	16.98	19.25	24.69
3	15.30	17.31	19.19	21.57	27.20
4	17.22	19.32	21.28	23.73	29.54
5	19.05	21.23	23.26	25.79	31.77
6	20.82	23.07	25.16	27.77	33.90
9	25.84	28.30	30.55	33.36	39.90
12	30.58	33.20	35.61	38.59	45.49
15	35.13	37.91	40.44	43.58	50.81
18	39.54	42.45	45.11	48.39	55.92

4.1.4 Case Studies with Canadian Lynx Data

The annual record of the numbers of the Canadian lynx trapped in the Mackenzie River district of northwest Canada plotted in Figure 1.2 has been featured in several textbooks on time series. The periodic fluctuation displayed in this series has profoundly influenced ecological theory. It has also become a benchmark series to test new statistical methodology for time series analysis. The first time series model built for this particular data set was probably that of Moran (1953). Moran fitted the following linear AR(2) model to the logarithm of the lynx data:

$$X_t = 1.05 + 1.41X_{t-1} - 0.77X_{t-2} + \varepsilon_t, \tag{4.11}$$

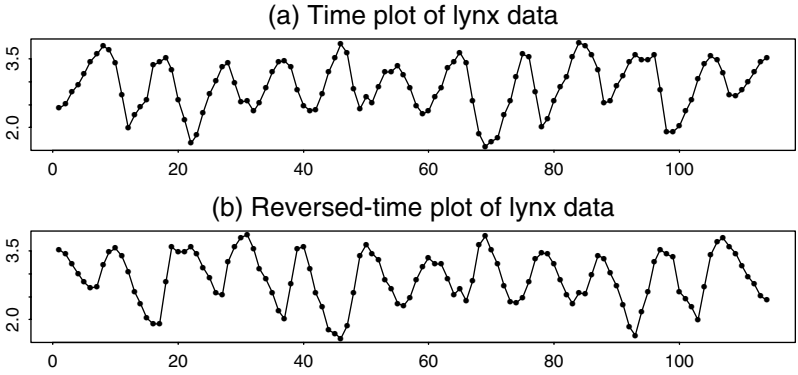


FIGURE 4.5. (a) Time plot of Canadian lynx data; (b) reversed time series plot of lynx data.

where $\{\varepsilon_t\} \sim \text{IID}(0, 0.04591)$. In fact, Moran immediately realized the limitation of the linear fitting, as he pointed out in the same paper a "curious feature"—the sum of squares of residuals corresponding to values of X_t greater than the mean is 1.781, whereas the sum of squares of residuals corresponding to values of X_t smaller than the mean is 4.007. The ratio of the two sums is 2.250, which would be judged significant at the 1% level (F-test) against the null hypothesis that the two sets of residuals are random samples from the same normal population. Later, we will demonstrate how to fit a TAR model to this data set step-by-step, including exploratory analysis using various plots and nonparametric smoothing and statistical tests for linearity. This part of the analysis is partially extracted from §7.2 of Tong (1990). We always refer the lynx data to the $\log_{10}$-transformed data displayed in Figure 1.2.

(a) *Plots and nonparametric smoothing*

As in fitting linear time series models, a judiciously constructed data-plot can be very informative and revealing. Figure 4.5 plots the lynx time series in the conventional as well as the reversed time order. It is clear that the lynx population exhibits a periodic-like fluctuation with most cycles around nine or ten years. It is also clear that there exists some characteristic in this series that is not time-reversible. For example, the population cycle is asymmetric; it took about six years to reach a peak from a trough and took only three or four years to drop from a peak to a trough.

Is the lack of time-reversibility suggestive of nonlinearity? The answer is not necessarily affirmative in general. However, if we look for a statistical model that will reproduce a time-irreversible characteristic, we may appeal for nonlinear modeling. The (linear) ARMA models discussed in Chapter 3 focus on linear autocorrelation, which, by virtue of its nature, is time-reversible. Furthermore, if an ARMA process is defined in terms of a

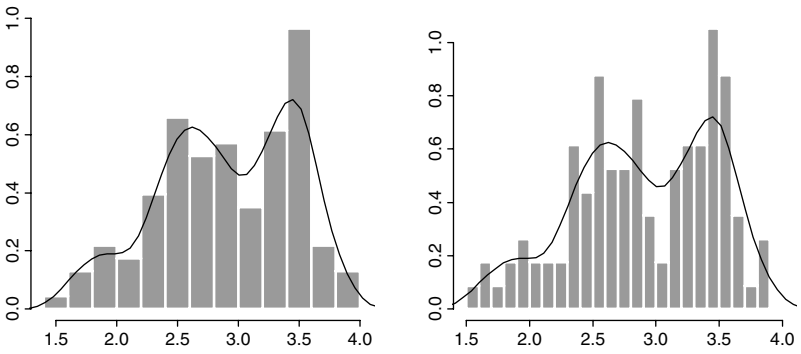


FIGURE 4.6. Two histograms of Canadian lynx data with different bin-sizes, together with the estimated density function (solid curve).

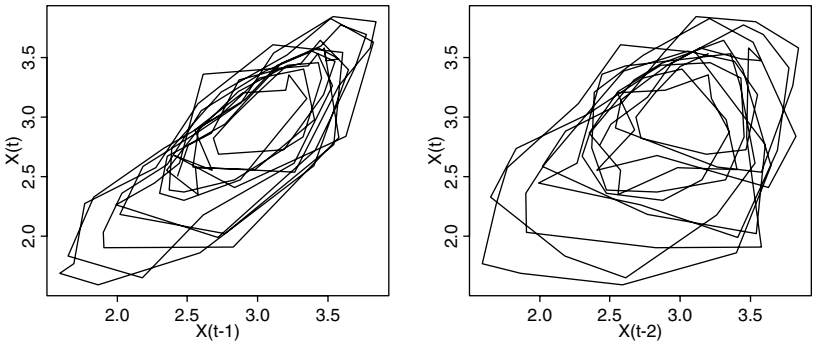


FIGURE 4.7. Directed scatter diagrams at lags 1 and 2 for Canadian lynx data.

Gaussian white noise, its entire probability distribution is time-reversible. Although we do not impose the normality explicitly in an ARMA model, we treat it implicitly as a Gaussian model, as we often look into its first two moment properties only. The spectral analysis is a typical example in point. Such a treatment is entirely legitimate only for Gaussian processes. Therefore, once we have identified some non-Gaussian properties, we may be prepared to entertain a nonlinear model. In this sense, nonnormality may be viewed as nonlinearity (see also Proposition 2.1).

Figure 4.6 presents two histograms of the lynx data with different bin-sizes, which clearly indicate that the marginal distribution is at least bimodal. The nonparametric estimator for the probability density function (see §5.2) produced by the standard S-Plus function "density" with the default setting reinforces this non-Gaussian property.

A *directed scatter diagram* at lag k plots X_t against X_{t-k} with adjacent points (such as (X_{t-k}, X_t) and (X_{t-k+1}, X_{t+1})) linked by straight lines. It is basically a scatter plot presented in a more informative manner. A direct scatter diagram is another powerful graphical tool in analyzing nonlinear time series. For the lynx data, Figure 4.7 shows that there is clearly a void in the center of the diagram at both lags 1 and 2. This indicates convincingly that the joint distributions of (X_{t-k}, X_t), for $k = 1, 2$, are not Gaussian since a two-dimensional normal distribution cannot have a hole in the center of its sample space. This is also consistent with the marginal density shown in Figure 4.6. If $\{X_t\}$ is a Gaussian process, its marginal distribution of X_t is also Gaussian.

The scatter plots in Figures 4.8(a)–(d) display X_t against X_{t-k} for $k = 1, 2, 3$, and 4, together with nonparametric estimators for the lag regression $E(X_t|X_{t-k} = x)$, produced by the standard S-Plus function "ksmooth" with the default setting. (The estimated curves are undersmoothed. However, we decide not to increase the amount of smoothness since the estimators merely serve as explanatory devices at this stage. For a comprehensive account on nonparametric smoothing, see §6.3.) Like most real data, the lag regression at lag 1 is pretty linear. However, $E(X_t|X_{t-k} = x)$ for $k = 2, 3$, and 4 are unlikely to be linear (in x), lending further support that $\{X_t\}$ is not a Gaussian process. Inspired by the linearity portrayed in Figure 4.8(a), we fitted a linear regression of X_t on X_{t-1}, leading to the model $\widehat{X}_t = 0.620 + 0.788X_{t-1}$. We plot the residuals $X_t - \widehat{X}_t$ against X_{t-1} and X_{t-2}, respectively, in Figures 4.8 (e) and (f). As expected, X_{t-1} contains little information on the residuals as the regression curve in Figure 4.8(e) is virtually zero. However, X_{t-2} does contain some additional information. In Figure 4.8(f), except for a few "outliers" the residual points spread almost evenly on both sides of the regression curve, which is clearly nonlinear. This indicates the nonlinear dependence of X_t on its lagged value X_{t-2}.

In summary, by plotting the data in various manners coupled with nonparametric smoothing, we have identified some nonlinear features such as time-irreversibility, nonnormality, and nonlinear autodependence.

(b) *Testing for linearity*

We apply the likelihood ratio test (4.9) for the null hypothesis-H_0: $\{X_t\}$ is a linear AR(2) process-against the alternative-H_1: $\{X_t\}$ follows TAR model (4.1) with two regimes and $p_1 = p_2 = d = 2$. Now $T = 114$ and $\widehat{\sigma}_0^2 = 0.0586$. Setting $\mathcal{I}_r$ equal to the 90% inner sample range, we have $\widehat{\sigma}^2 = 0.0441$. Thus $S_T = (T-2)(\widehat{\sigma}_0^2 - \widehat{\sigma}^2)/\widehat{\sigma}^2 = 36.825$. According to Table 4.2, we reject the null hypothesis of a linear AR(2) model even at the level 0.1%.

(c) *A simple TAR model with biological interpretation*

For many biological populations, birth rates depend on population sizes-for example, due to competition for the resources of habitat, the limitation

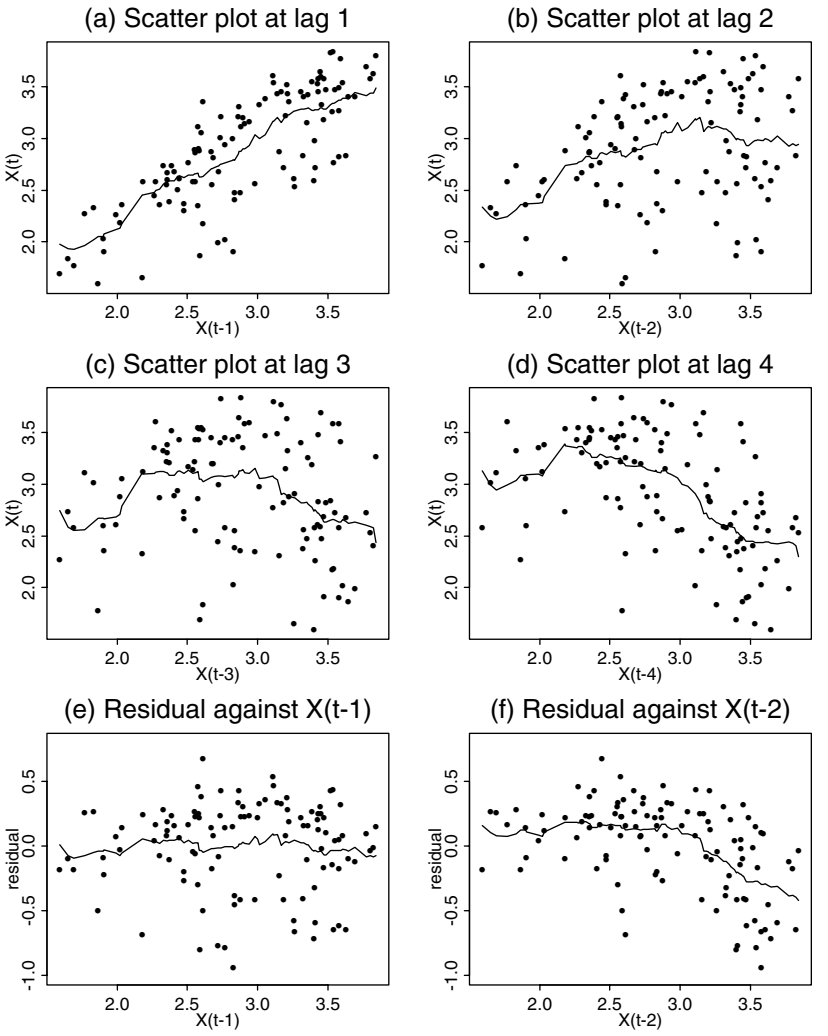


FIGURE 4.8. Scatter plots of X_t against (a) X_{t-1}, (b) X_{t-2}, (c) X_{t-3}, and (d) X_{t-4} for Canadian lynx data together with kernel regression estimators (solid curves) for $E(X_t|X_{t-k} = x)$ and scatter plots of the residual from the linear regression $\widehat{X}_t = 0.620 + 0.788X_{t-1}$ against (e) X_{t-1} and (f) X_{t-2}. Solid curves are nonparametric regression estimators.

of food, the predator–prey interaction and other factors. Typically, the birth rate will increase in the early stage, called an increasing phase, in a population cycle, and it will decrease when the population is oversized in the latter stage, leading to a decreasing phase. A population decrease for one species will lead, in due course, to a population decrease of its predators and the population increase of its prey and also the abundance

of resources. This in turn will lead to a new increasing phase. Therefore, it seems very appealing to model population dynamics in terms of a threshold model in which different regimes would reflect different phases or stages in population cycles. Having incorporated the biological evidence, H. Tong fitted the following TAR model with two regimes with delay variable $d = 2$ to the lynx data:

$$X_t = \begin{cases} 0.62 + 1.25X_{t-1} - 0.43X_{t-2} + \varepsilon_t^{(1)}, & X_{t-2} \le 3.25, \\ 2.25 + 1.52X_{t-1} - 1.24X_{t-2} + \varepsilon_t^{(2)}, & X_{t-2} > 3.25; \end{cases} \tag{4.12}$$

see Tong (1990, p. 377). Let us rewrite the model above, discarding the error terms, as follows:

$$X_t - X_{t-1} = \begin{cases} 0.62 + 0.25X_{t-1} - 0.43X_{t-2}, & X_{t-2} \le 3.25, \\ -(1.24X_{t-2} - 2.25) + 0.52X_{t-1}, & X_{t-2} > 3.25. \end{cases} \tag{4.13}$$

In the upper regime (i.e., $X_{t-2} > 3.25$), $X_t - X_{t-1}$ tends to be negative, implying a population decrease. In the lower regime (i.e., $X_{t-2} \le 3.25$), $X_t - X_{t-1}$ tends to be marginally positive, implying slow population growth. In fact, a sequence $\{X_t\}$ generated by (4.13) will converge to a stable limit cycle of period 9 consisting of an ascent phase of length 6 and a descent phase of length 3. This is in agreement with the observed asymmetric cycles in Figure 4.5.

Stenseth et al. (1999) gave a nice interpretation of model (4.12) in terms of the well-known predator (lynx) and prey (hare) interaction model in ecology. As we pointed out above, the lower regime corresponds roughly to the population increase phase, and the upper regime corresponds to the population decrease phase. Note that the coefficients of X_{t-1} in (4.12) are significantly positive but less so during the increase phase. The coefficients of X_{t-2} are significantly negative and more so during the decline phase. The signs of those coefficients reflect that lynx and hares relate with each other in a specified prey–predator interactive manner. The difference of the coefficients in increasing and decreasing phases represents the so-called *phase-dependence* and *density-dependence* in ecology, which can only be reflected in a nonlinear model. The phase-dependence means that the both lynx and the hare behave differently (in hunting or escaping) when the lynx population increases or decreases. The density-dependence implies that the reproduction rates of animals as well as their behavior depend on the abundance of the population. For further discussion on the biological meaning of TAR fitting for the lynx data, we refer the reader to Stenseth et al. (1999).

(d) *The model selected by AIC*

Setting $k = 2$, $1 \le p_1, p_2 \le 10$, and $1 \le d \le 6$, the AIC selected for the lynx data the following TAR model

$$X_t = \begin{cases} 0.546 + 1.032X_{t-1} - 0.173X_{t-2} + 0.171X_{t-3} - 0.431X_{t-4} \\ +0.332X_{t-5} - 0.284X_{t-6} + 0.210X_{t-7} + \varepsilon_t^{(1)}, & X_{t-2} \le 3.116, \\ 2.632 + 1.492X_{t-1} - 1.324X_{t-2} + \varepsilon_t^{(2)}, & X_{t-2} > 3.116; \end{cases} \tag{4.14}$$

see Tong (1990, p. 387). The estimated threshold is $\widehat{r} = 3.116$, which is *the turning point* of the regression estimator in Figure 4.8(f). The estimated variances for $\varepsilon_t^{(1)}$ and $\varepsilon_t^{(2)}$ are 0.0259 and 0.0505, respectively. The standard errors of the eight estimated coefficients in the lower regime, calculated based on (4.8) (see also Theorem 4.2), are, respectively, 0.275, 0.094, 0.156, 0.149, 0.153, 0.170, 0.167, and 0.101 in order of their appearance in the model. The standard errors of the three estimated coefficients in the upper regime are, respectively, 0.655, 0.102, and 0.195.

Model (4.14) preserves the basic dynamics of the simpler model (4.12). For example, the sequence $\{X_t\}$ generated by (4.14) (discarding the error terms) also converges to a limit cycle of period 9 with an increase phase of length 6 and a decrease phase of length 3. In terms of statistical fitting, model (4.14) represents an improvement over (4.12). However its more complex form also makes biological interpretation less clear. The choice between the two models rests on the purpose of the analysis. Obviously, model (4.12) would be preferable if we aim to model lynx population fluctuation and reflect different characteristics at different phases of the population cycles. On the other hand, model (4.14) entertains better statistical properties, providing better fitting to the original data. Furthermore, it may provide a better forecasting for further values.

It is a good practice to look into several models selected by different criteria such as AIC, BIC, and others, or, say, the best three models selected by the same criterion. The choice of the final model depends on the statistical and/or physical properties of the models, dictated by the purpose of the data analysis. Table 7.6 of Tong (1990, p. 386) listed six selected models for lynx data. Model (4.14) was singled out as the one with both good statistical fitting and adequate resemblance to lynx population fluctuation.

(e) *Diagnostic checking*

Like all statistical fitting, it is important to conduct a diagnostic check for a fitted nonlinear time series model, although some diagnostic ideas have already been incorporated into the modern modeling techniques. For example, a model selected by BIC is usually free from overfitting.

Similar to fitting linear models, the residual-based methods remain as the most frequently used diagnostic tools; see §3.5. The residuals from model (4.14) passed most of those tests comfortably. It is also helpful to compare

some characteristics of the original data with those of the simulated data from a fitted model. For example, as we mentioned above, both models (4.12) and (4.14) can reproduce the asymmetric population cycle successfully.

4.2 ARCH and GARCH Models

In contrast to traditional time series analysis, which focuses on modeling the conditional first moment, ARCH and GARCH models specifically take the dependency of the conditional second moments into modeling consideration. This, hopefully, would accommodate the increasingly important demand to explain and to model risk and uncertainty in, for example, financial time series. In this section, we first present basic probabilistic properties of ARCH and GARCH models. The most frequently used statistical inference methods for ARCH/GARCH modeling will also be introduced. We also briefly mention the application of ARCH/GARCH modeling with financial time series. Further, we illustrate the methodology of GARCH modeling through a real data set. These methods have been implemented in S+GARCH, an add-on module to the S-Plus system. Finally, we give a brief introduction on stochastic volatility models. For a comprehensive account of ARCH and GARCH modeling, see Gouriéroux (1997).

4.2.1 Basic Properties of ARCH Processes

Definition 4.2 *An* autoregressive conditional heteroscedastic (ARCH) *model with order* $p\,(\geq 1)$ *is defined as*

$$X_t = \sigma_t \varepsilon_t \quad \text{and} \quad \sigma_t^2 = c_0 + b_1 X_{t-1}^2 + \cdots + b_p X_{t-p}^2, \tag{4.15}$$

where $c_0 \geq 0$, $b_j \geq 0$ *are constants,* $\{\varepsilon_t\} \sim \text{IID}(0,1)$*, and* ε_t *is independent of* $\{X_{t-k}, k \geq 1\}$ *for all* t*. A stochastic process* $\{X_t\}$ *defined by the equations above is called an* ARCH(p) *process.*

The ARCH model was first introduced by Engle (1982) for modeling the predictive variance for U.K. inflation rates. Since then, it has been widely used to model volatility of financial and economic time series. The basic idea behind the construction of (4.15) is quite intuitive: the predictive distribution of X_t based on its past is a scale-transform of the distribution of ε_t, with the scaling constant σ_t depending on the past of the process. Therefore, conditional quantiles of X_t given its past, which play important roles in financial risk management (see Part (e) of §4.2.8 and §8.5.6 below), can also be evaluated easily. For example, if $\varepsilon_t \sim N(0,1)$, the predictive distribution is $N(0, \sigma_t^2)$, with the variance σ_t^2 depending on the conditions on which the prediction was made. Further, a large predictive variance will

be caused by the large absolute values of observations in the immediate past. This is in marked contrast to the prediction based on linear models for which the conditional mean squared predictive errors are constants; see Proposition 3.4.

Theorem 4.3 *(i) The necessary and sufficient condition for (4.15) defining a unique strictly stationary process* $\{X_t,\ t = 0, \pm 1, \pm 2, \cdots\}$ *with* $EX_t^2 < \infty$ *is* $\sum_{j=1}^p b_j < 1$. *Furthermore,*

$$EX_t = 0 \quad \text{and} \quad EX_t^2 = c_0 / \left\{ 1 - \sum_{j=1}^p b_j \right\},$$

and $X_t \equiv 0$ *for all* t *if* $c_0 = 0$.

(ii) If $E\varepsilon_t^4 < \infty$ *and*

$$\max\{1, (E\varepsilon_t^4)^{1/2}\} \sum_{j=1}^p b_j < 1, \tag{4.16}$$

the strictly stationary solution of (4.15) has the finite fourth moment, namely $EX_t^4 < \infty$.

Proof. The sufficiency of (i) and (ii) follows from Theorem 2.5 immediately. By (4.15) and stationarity, $EX_t = 0$ and

$$EX_t^2 = c_0 + b_1 EX_{t-1}^2 + \cdots + b_p EX_{t-p}^2 = c_0 + \sum_{j=1}^p b_j EX_t^2$$

or

$$EX_t^2 = \frac{c_0}{1 - \sum_{j=1}^p b_j}.$$

The necessity of (i) follows from Theorem 1 of Bollerslev (1986), which shows that the condition $\sum_j b_j < 1$ is also necessary for (4.15) having a (weakly) stationary solution. (Note that Theorem 1 in Bollerslev's paper does not depend on the assumed normality.) Indeed, $EX_t^2 > 0$ entails that $\sum_{j=1}^p b_j < 1$. ∎

It is easy to see from (4.15) that any stationary ARCH process $\{X_t\}$ is also a white noise $\mathrm{WN}(0, c_0/(1 - \sum_{j=1}^p b_j))$; see also Theorem 4.3(i). Furthermore, we may write

$$X_t^2 = c_0 + b_1 X_{t-1}^2 + \cdots + b_p X_{t-p}^2 + e_t, \tag{4.17}$$

where $e_t = (\varepsilon_t^2 - 1)\{c_0 + \sum_{j=1}^p b_j X_{t-j}^2\}$. It is easy to see that

$$E(e_t | X_{t-k}, X_{t-k-1}, \cdots) = 0 \quad \text{for any } k \geq 1. \tag{4.18}$$

Hence, for any $k > p$, by (4.17) and (4.18)

$$E(X_{t+k}^2|X_{t-m}, m \geq 0) = c_0 + \sum_{j=1}^{p} b_j E(X_{t+k-j}^2|X_{t-m}, m \geq 0)$$

or

$$\mathrm{Var}(X_{t+k}|X_{t-m},\, m \geq 0) = c_0 + \sum_{j=1}^{p} b_j \mathrm{Var}(X_{t+k-j}|X_{t-m},\, m \geq 0). \quad (4.19)$$

More generally, for $k \geq 1$,

$$\begin{aligned} \mathrm{Var}(X_{t+k}|X_{t-m},\, m \geq 0) &= c_0 + \sum_{j=1}^{k-1} b_j \mathrm{Var}(X_{t+k-j}|X_{t-m},\, m \geq 0) \\ &+ \sum_{j=k}^{p} b_j X_{t+k-j}^2. \end{aligned} \quad (4.20)$$

The two equations above reflect the fact that the high risk in forecasting will be sustained over a period before it dies away; a phenomenon called *volatility clustering* in financial time series analysis.

It follows from Theorem 4.3(ii) that under the additional condition (4.16), $\{e_t\} \sim \mathrm{WN}(0, \sigma_e^2)$ with

$$\sigma_e^2 = \mathrm{Var}(\varepsilon_t^2)\, E\left\{c_0 + \sum_{j=1}^{p} b_j X_{t-j}^2\right\}^2 < \infty.$$

Note that under the condition $\sum_{j=1}^{p} b_j < 1$,

$$\left|\sum_{j=1}^{p} b_j z^j\right| \leq \sum_{j=1}^{p} b_j |z^j| \leq \sum_{j=1}^{p} b_j < 1 \quad \text{for all } |z| \leq 1.$$

Thus $1 - \sum_{j=1}^{p} b_j z^j \neq 0$ for all $|z| \leq 1$. This means that $\{X_t^2\}$ is a causal AR(p) process. Therefore, the ACF (and also ACVF) of the process $\{X_t^2\}$ can be easily calculated in terms of (2.20) and (2.18). Furthermore, it is easy to see from those formulas that, for all τ, $\mathrm{Corr}(X_t^2, X_{t+\tau}^2) > 0$ if $\sum_{j=1}^{p} b_j > 0$, although $\mathrm{Corr}(X_t, X_{t+\tau}) = 0$.

If we adopt *kurtosis* as a measure for *heavy tails* of distribution, the ARCH process $\{X_t\}$ has heavier tails than those of the white noise $\{\varepsilon_t\}$ on which $\{X_t\}$ is defined. To this end, denote by $\kappa_\varepsilon = E(\varepsilon_t^4)/(E\varepsilon_t^2)^2$ the kurtosis of the distribution of ε_t. Then

$$\begin{aligned} E(X_t^4|X_{t-1}, \cdots, X_{t-p}) &= \sigma_t^4 E\varepsilon_t^4 = \kappa_\varepsilon \sigma_t^4 (E\varepsilon_t^2)^2 \\ &= \kappa_\varepsilon \{E(X_t^2|X_{t-1}, \cdots, X_{t-p})\}^2. \end{aligned}$$

Now, it follows from Jensen's inequality that

$$E(X_t^4) = \kappa_\varepsilon E\{E(X_t^2|X_{t-1}, \cdots, X_{t-p})\}^2 \geq \kappa_\varepsilon (EX_t^2)^2. \tag{4.21}$$

Hence $\kappa_x \equiv E(X_t^4)/(EX_t^2)^2 \geq \kappa_\varepsilon$. In the case where ε_t is normal and $\kappa_x > \kappa_\varepsilon = 3$, X_t has leptokurtosis (i.e., fat tails).

We summarize the findings above in the proposition below.

Proposition 4.1 *Let $\{X_t\}$ be the strictly stationary ARCH(p) process defined by (4.15) with $c_0 > 0$ and $\sum_{j=1}^p b_j < 1$. Then*

(i) $\{X_t\} \sim \mathrm{WN}(0, c_0/(1 - \sum_{j=1}^p b_j))$, and the conditional variance function fulfills equation (4.19).

Under the additional condition (4.16),

(ii) $\{X_t^2\}$ is a (linear) causal AR(p) process, and its ACF is always positive if $\sum_{j=1}^p b_j > 0$, and

(iii) X_t exhibits heavier tails than those of ε_t in the sense that $\kappa_x \geq \kappa_\varepsilon$.

Example 4.1 Consider the strictly stationary ARCH(1) process

$$X_t = \sigma_t \varepsilon_t, \quad \text{and} \quad \sigma_t^2 = c_0 + b_1 X_{t-1}^2, \tag{4.22}$$

where $\{\varepsilon_t\} \sim \mathrm{IID}(0, 1)$, $c_0 > 0$, and $b_1 \in (0, 1)$. Then $EX_t^2 = c_0/(1 - b_1)$, and for $k > 1$,

$$\begin{aligned} \mathrm{Var}(X_{t+k}|X_{t-j},\, j \geq 0) &= \mathrm{Var}(X_{t+k}|X_t) \\ = \; c_0 + b_1 \mathrm{Var}(X_{t+k-1}|X_t). & \end{aligned} \tag{4.23}$$

Iteratively, using (4.23), we have

$$\mathrm{Var}(X_{t+k}|X_{t-j},\, j \geq 0) = \frac{c_0(1 - b_1^k)}{1 - b_1} + b_1^k X_t^2,$$

which indicates that a large value of $|X_t|$ will lead to large predictive risk (i.e., conditional variance) and that the risk will be sustained for a while in the immediate future.

Suppose that $\varepsilon_t \sim N(0, 1)$. Then, the condition (4.16) reduces to $3b_1^2 < 1$ (i.e., $b_1 < 0.577$). Under this condition, $\mathrm{Corr}(X_t^2, X_{t+\tau}^2) = b_1^{|\tau|}$, and $\{X_t^2\}$ follows a causal AR(1) equation

$$X_t^2 = c_0 + b_1 X_{t-1}^2 + e_t,$$

where $e_t = (\varepsilon_t^2 - 1)(c_0 + b_1 X_{t-1})$. Hence, by multiplying the term X_t^2 and taking the expectation on the both sides of the equation above, we have

$$\begin{aligned} EX_t^4 &= c_0 EX_t^2 + b_1 E(X_t^2 X_{t-1}^2) + E(X_t^2 e_t) \\ &= c_0 EX_t^2 + b_1\{b_1 \mathrm{Var}(X_t^2) + (EX_t^2)^2\} + E(e_t^2) \\ &= (1 - b_1)(EX_t^2)^2 + b_1\{b_1 EX_t^4 + (1 - b_1)(EX_t^2)^2\} + \frac{2}{3} EX_t^4. \end{aligned}$$

The last equality makes use of the fact that $c_0 = (1 - b_1)EX_t^2$. It is easy to see now that

$$\frac{EX_t^4}{(EX_t^2)^2} = \frac{3(1 - b_1^2)}{(1 - 3b_1^2)} > 3.$$

Hence X_t has leptokurtosis (fat tails).

We generate a series of length 1000 from (4.22) with $c_0 = 1.5$, $b_1 = 0.9$ and normal ε_t. The first 250 sample points X_t are shown in Figure 4.9(a). Figure 4.9(c) is the plot of corresponding conditional standard deviations σ_t, which clearly indicates that the large values of $|X_t|$ and σ_t are lined together. Both histograms of the sample in Figure 4.9(b) and the plot (against normal) in Figure 4.9(d) show that the tails of the distribution of X_t are heavier than a normal distribution. Figures 4.9 (e) and (f) are the sample ACFs of $\{X_t\}$ and $\{X_t^2\}$.

We repeat the exercise above in Figure 4.10 with reduced value $b_1 = 0.4$. Comparing Figures 4.10 (a) and (b) with Figures 4.9 (a) and (b), the volatility is more prominent for larger values of b_1. Figure 4.9(d) shows that the tails of the distribution of X_t are still heavier than a normal distribution, although not as much so as in the case of $b_1 = 0.9$. Note now that $\{X_t^2\}$ is a causal AR(1) model with the ACF $\rho(k) = 0.4^k$; see Figure 4.10(f). In contrast, in Figure 4.9(f) with $b_1 = 0.9$, there is a substantial discrepancy between $\widehat{\rho}(k)$ and 0.9^k for $k = 1, 2, \cdots$. This is due to the fact that $EX_t^4 = \infty$ when $b_1 = 0.9$. Therefore, the ACVF is not well-defined. ∎

4.2.2 Basic Properties of GARCH Processes

The ARCH model has been extended in a number of directions, some dictated by economic consideration, others by broadly statistical ideas. The most important of these is the extension to include moving average parts, namely the generalized ARCH (GARCH) model due to Bollerslev (1986) and Taylor (1986).

Definition 4.3 *A generalized autoregressive conditional heteroscedastic (GARCH) model with order $p(\geq 1)$ and $q(\geq 0)$ is defined as*

$$X_t = \sigma_t \varepsilon_t \quad \text{and} \quad \sigma_t^2 = c_0 + \sum_{i=1}^{p} b_i X_{t-i}^2 + \sum_{j=1}^{q} a_j \sigma_{t-j}^2, \tag{4.24}$$

where $c_0 \geq 0$, $b_i \geq 0$, and $a_j \geq 0$ are constants, $\{\varepsilon_t\} \sim \text{IID}(0, 1)$, and ε_t is independent of $\{X_{t-k}, k \geq 1\}$ for all t. A stochastic process $\{X_t\}$ defined by the equations above is called a GARCH *(p, q) process.*

Empirical work has shown that the simple ARCH(p) model defined in (4.15) will provide a reasonable fit to financial time series only if the order p is large. Since the rationale for the definition of (4.15) is to take a weighted average of the past squared observations as an approximation to

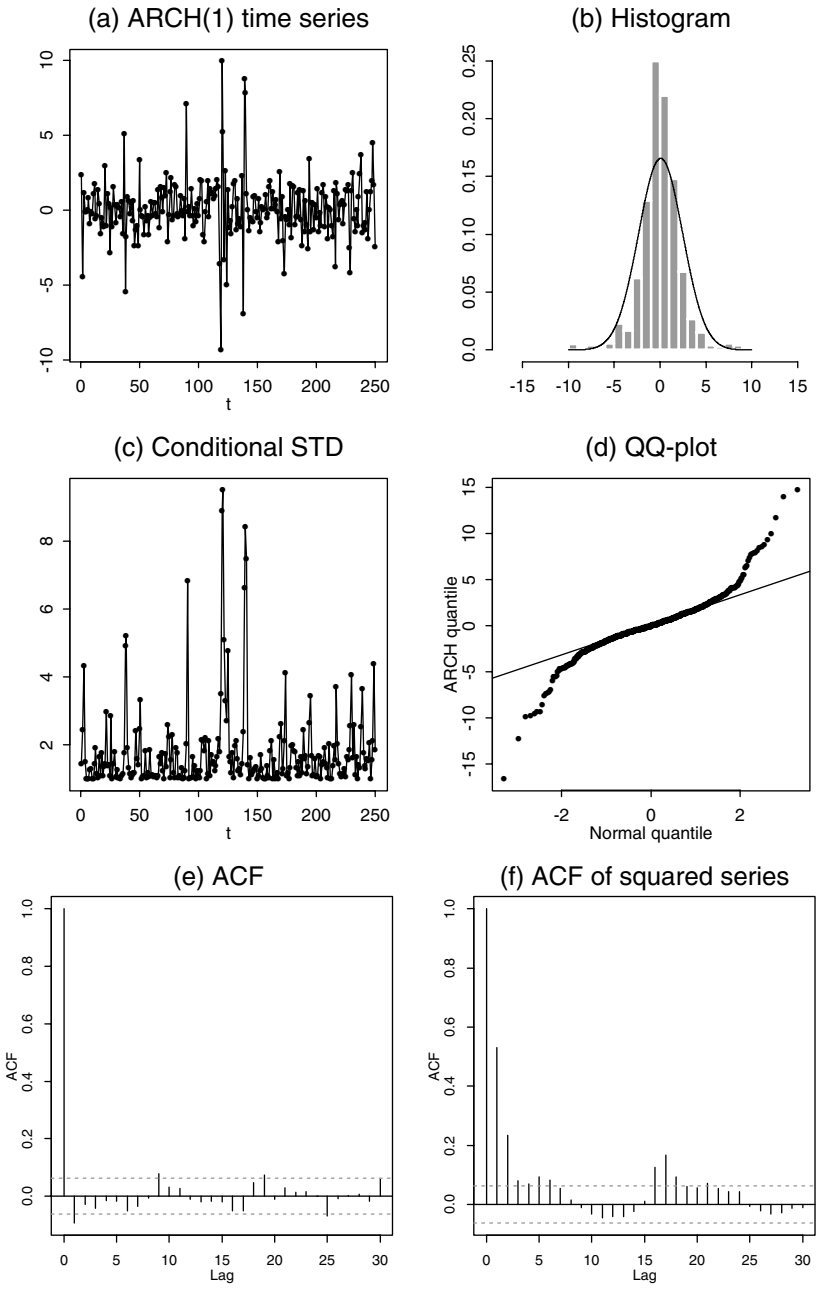


FIGURE 4.9. Example 4.1 — A sample of 1,000 was generated from the ARCH(1) model with $b_1 = 0.9$: (a) and (c) time series plots of the first 250 X_t and σ_t; (b) normalized histogram and the normal density function with the same mean and variance; (d) plot: the sample quantiles versus the quantiles of $N(0, 1)$; (e) and (f) sample ACFs of $\{X_t\}$ and $\{X_t^2\}$, respectively.

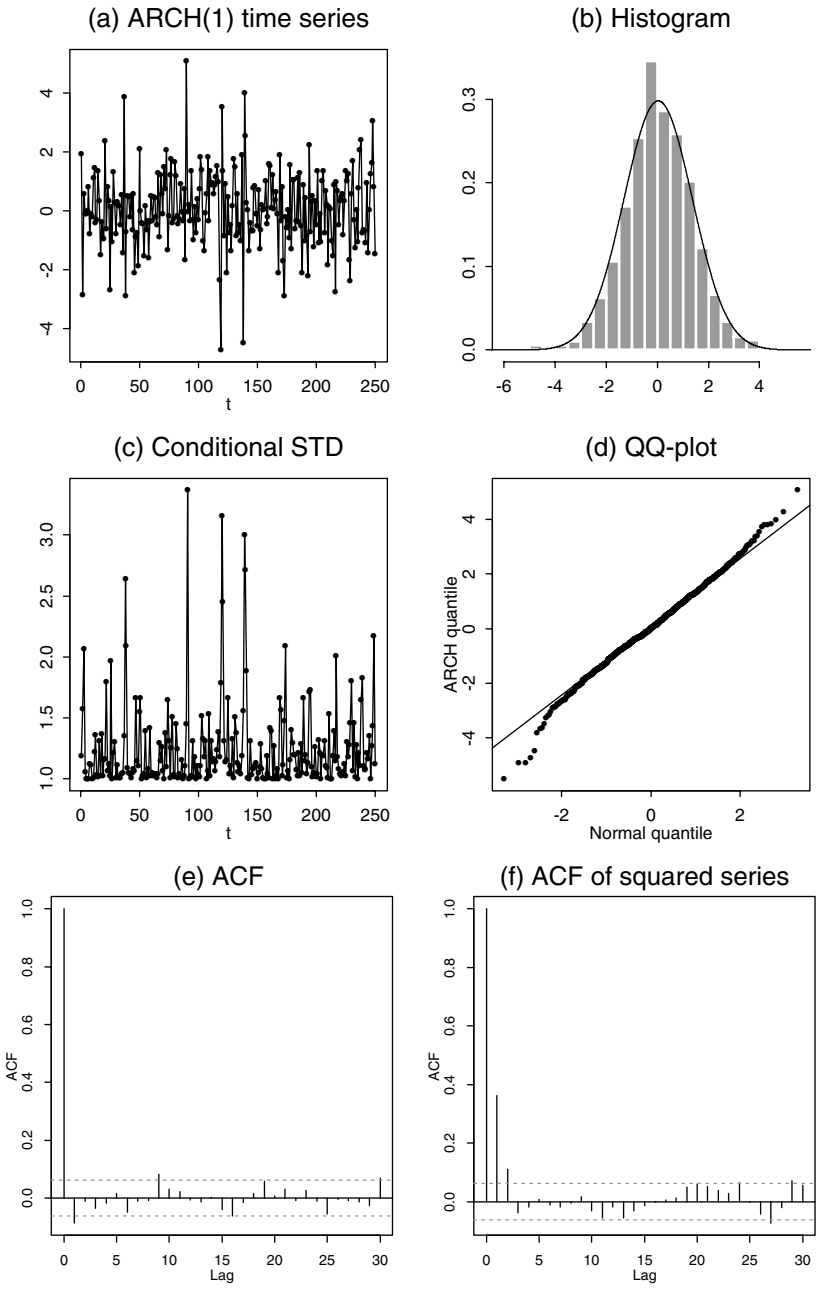


FIGURE 4.10. Example 4.1 — A sample of 1,000 was generated from the ARCH(1) model with $b_1 = 0.4$: (a) and (c) time series plots of the first 250 X_t and σ_t; (b) normalized histogram and the normal density function with the same mean and variance; (d) plot: the sample quantiles versus the quantiles of $N(0, 1)$; (e) and (f) sample ACFs of $\{X_t\}$ and $\{X_t^2\}$, respectively.

the conditional variance σ_t^2, it is quite natural to define σ_t^2 as a weighted average of not only past X_j^2's but also past σ_j^2's. This leads to the GARCH model (4.24), which in fact entertains an interesting link to ARMA models,

$$\begin{aligned} X_t^2 &= c_0 + \sum_{i=1}^{p} b_i X_{t-i}^2 + \sum_{j=1}^{q} a_j \sigma_{t-j}^2 + e_t \\ &= c_0 + \sum_{i=1}^{p\vee q} (b_i + a_i) X_{t-i}^2 + e_t - \sum_{j=1}^{q} a_j e_{t-j}, \end{aligned} \tag{4.25}$$

where $b_{p+j} = a_{q+j} = 0$ for $j \geq 1$, $p \vee q = \max\{p, q\}$, and

$$e_t = X_t^2 - \sigma_t^2 = (\varepsilon_t^2 - 1)\left(c_0 + \sum_{i=1}^{p} b_i X_{t-i}^2 + \sum_{j=1}^{q} a_j \sigma_{t-j}^2 \right). \tag{4.26}$$

Thus, formally $\{X_t^2\}$ follows an ARMA($p\vee q$, q) model. Note that an invertible ARMA(p, q) model with finite p and q is effectively an AR(∞) model. This explains why simple GARCH models, such as GARCH(1, 1), may provide a parsimonious representation for some complex autodependence structure of $\{X_t^2\}$, that can only be accommodated by an ARCH(p) model with large p; see also (4.17). In fact, the GARCH(1, 1) model has been tremendously successful in empirical work and is regarded as the benchmark model by many econometricians.

Theorem 4.4 *The necessary and sufficient condition for (4.24) defining a unique strictly stationary process* $\{X_t, t = 0, \pm 1, \pm 2, \cdots\}$ *with* $EX_t^2 < \infty$ *is*

$$\sum_{i=1}^{p} b_i + \sum_{j=1}^{q} a_j < 1. \tag{4.27}$$

Furthermore, $EX_t = 0$ *and*

$$\mathrm{var}(X_t) = \frac{c_0}{1 - \sum_{i=1}^{p} b_i - \sum_{j=1}^{q} a_j}, \quad \mathrm{Cov}(X_t, X_{t-k}) = 0 \ \textit{for any } k \neq 0.$$

In addition, $EX_t^4 < \infty$*, provided*

$$\max\{1, (E\varepsilon_t^4)^{1/2}\} \frac{\sum_{i=1}^{p} b_i}{1 - \sum_{j=1}^{q} a_j} < 1. \tag{4.28}$$

Proof. Note that the second equation in (4.24) can be formally written as

$$\left(1 - \sum_{j=1}^{q} a_j B^j \right) \sigma_t^2 = c_0 + \sum_{i=1}^{p} b_i B^i X_t^2.$$

Condition (4.27) implies that $1 - \sum_{j=1}^{q} a_j z^j \neq 0$ for all $|z| \leq 1$. Hence

$$\begin{aligned}\sigma_t^2 &= \left(1 - \sum_{j=1}^{q} a_j B^j\right)^{-1} \left\{c_0 + \sum_{i=1}^{p} b_i B^i X_t^2\right\} \\ &= c_0 / \left(1 - \sum_{j=1}^{q} a_j\right) + \sum_{i=1}^{\infty} d_i X_{t-i}^2,\end{aligned}$$

where d_i's are determined by the equation $\sum_{i=1}^{\infty} d_i z^i = \sum_{i=1}^{p} b_i z^i / (1 - \sum_{j=1}^{q} a_j z^j)$. Hence, by taking $z = 1$,

$$\sum_{i=1}^{\infty} d_i = \sum_{i=1}^{p} b_i / \left(1 - \sum_{j=1}^{q} a_j\right).$$

Similar to (2.20), d_i's can be calculated recursively where $d_1 = b_1$ and, for $i \geq 2$,

$$d_i = b_i + \sum_{k=1}^{i-1} a_k d_{i-k}.$$

In the expression above, we assume that $b_{p+j} = a_{q+j} = 0$ for $j \geq 1$. By an inductive argument, we may show $d_i \geq 0$ for all $i > 0$. Now, the theorem follows from Theorem 2.5 and Bollerslev (1986); see the proof of Theorem 4.3. Under stationarity, the variance and covariance can be calculated as

$$EX_t^2 = E\sigma_t^2 = c_0 + \sum_{i=1}^{p} b_i EX_t^2 + \sum_{j=1}^{q} a_j E\sigma_t^2.$$

This implies that

$$EX_t^2 = c_0 + \left(\sum_{i=1}^{p} b_i + \sum_{j=1}^{q} a_j\right) EX_t^2;$$

that is,

$$EX_t^2 = \frac{c_0}{1 - \sum_{i=1}^{p} b_i - \sum_{j=1}^{q} a_j}.$$

Furthermore, for $k > 0$, using the double expectation formula,

$$EX_t X_{t-k} = E\{X_{t-k} E(X_t | X_{t-1}, X_{t-2}, \cdots)\} = 0.$$

This completes the proof. ∎

Theorem 4.4 presents a necessary and sufficient condition for model (4.24) defining a strictly stationary process with a finite second moment.

Bougerol and Picard (1992b) established a necessary and sufficient condition for the existence of a strictly stationary solution that does not necessarily have a finite second moment; see also Kazakevičius and Leipus (2001). The condition is defined in terms of Lyapunov exponents for some random matrices associated with the model and is in general difficult to check in practice; therefore, it is not presented here.

Under condition (4.27), $\{X_t\} \sim \mathrm{WN}(0, c_0/(1-\sum_{i=1}^p b_i - \sum_{j=1}^q a_j))$, and the ARMA representation (4.25) is causal and invertible (although Ee_t^2 is not necessarily finite). Thus $EX_t^2 = E\varepsilon_t^2\, E\sigma_t^2 = E\sigma_t^2$ and

$$E(X_t|X_{t-1}, X_{t-2}, \cdots) = 0.$$

From (4.26), it holds that

$$Ee_t = E(e_t|X_{t-1}, X_{t-2}, \cdots) = 0.$$

Consequently,

$$\begin{aligned}&\mathrm{Var}(X_t|X_{t-1}, X_{t-2}, \cdots) = E(X_t^2|X_{t-1}, X_{t-2}, \cdots)\\ &= c_0 + \sum_{i=1}^p b_i X_{t-i}^2 + \sum_{j=1}^q a_j \sigma_{t-j}^2 = \sigma_t^2.\end{aligned}$$

Thus σ_t^2 defined in (4.24) is the conditional variance of X_t given its infinite past.

If $\{X_t\}$ is a strictly stationary GARCH(p, q) process and condition (4.28) holds, $E\sigma_t^4 = EX_t^4/E\varepsilon_t^4 < \infty$. Hence $Ee_t^4 < \infty$. In this case, $\{X_t^2\}$ is a causal and invertible ARMA$(p \vee q, q)$ process defined in (5.8). In contrast to ARCH processes, the ACF of $\{X_t^2\}$ is not necessarily always positive.

Note that the kurtosis inequality (4.21) still holds if we use the conditional expectations given the whole lagged values instead of only p lagged values. Hence, the following proposition holds.

Proposition 4.2 *(i) A stationary GARCH(p, q) process $\{X_t\}$ defined in (4.24) is a white noise, and σ_t^2 is the conditional variance of X_t given its infinite past.*

(ii) If $\{X_t\}$ is a strictly stationary GARCH(p, q) defined in (4.24) for which condition (4.28) holds, $\{X_t^2\}$ is a causal and invertible ARMA$(p \vee q, q)$ process. Furthermore X_t exhibits heavier tails than those of ε_t in the sense that $\kappa_x \geq \kappa_\varepsilon$.

Example 4.2 Consider the stationary GARCH(1, 1) process

$$X_t = \sigma_t \varepsilon_t \quad \text{and} \quad \sigma_t^2 = c_0 + b_1 X_{t-1}^2 + a_1 \sigma_{t-1}^2, \tag{4.29}$$

where c_0, b_1, and a_1 are positive, $b_1 + a_1 < 1$, and $\{\varepsilon_t\} \sim \mathrm{IID}(0, 1)$. Then $EX_t^2 = c_0/(1 - b_1 - a_1)$. Since $(1 - a_1 B)\sigma_t^2 = c_0 + b_1 X_{t-1}^2$, it holds that

$$\sigma_t^2 = \sum_{j=0}^\infty a_1^j B^j (c_0 + b_1 X_{t-1}^2) = \frac{c_0}{1-a_1} + b_1 \sum_{j=0}^\infty a_1^j X_{t-j-1}^2.$$

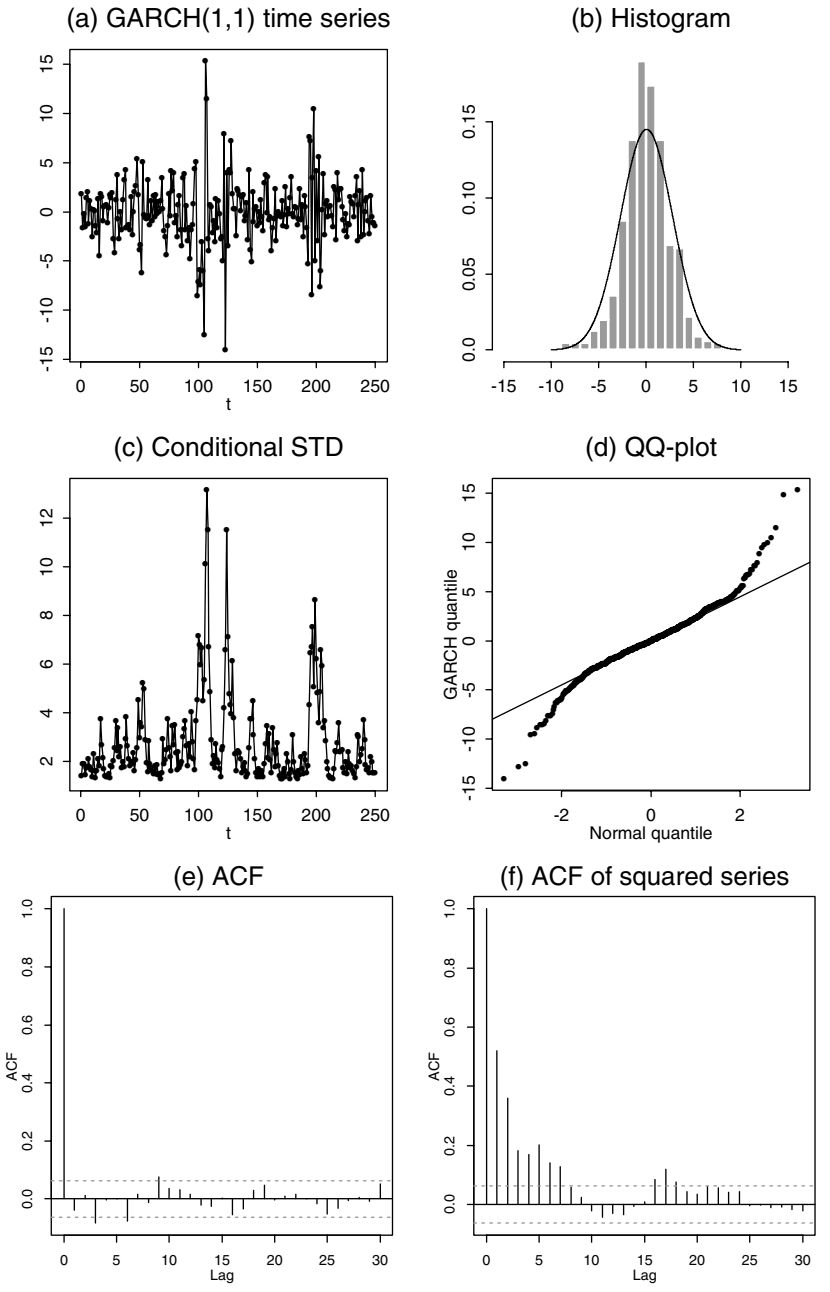


FIGURE 4.11. Example 4.2. A sample of 1,000 was generated from the GARCH(1, 1) model with $b_1 = 0.6$ and $a_1 = 0.3$: (a) and (c) time series plots of the first 250 X_t and σ_t; (b) normalized histogram and the normal density function with the same mean and variance; (d) plot: the sample quantiles versus the quantiles of $N(0, 1)$; (e) and (f) sample ACFs of $\{X_t\}$ and $\{X_t^2\}$, respectively.

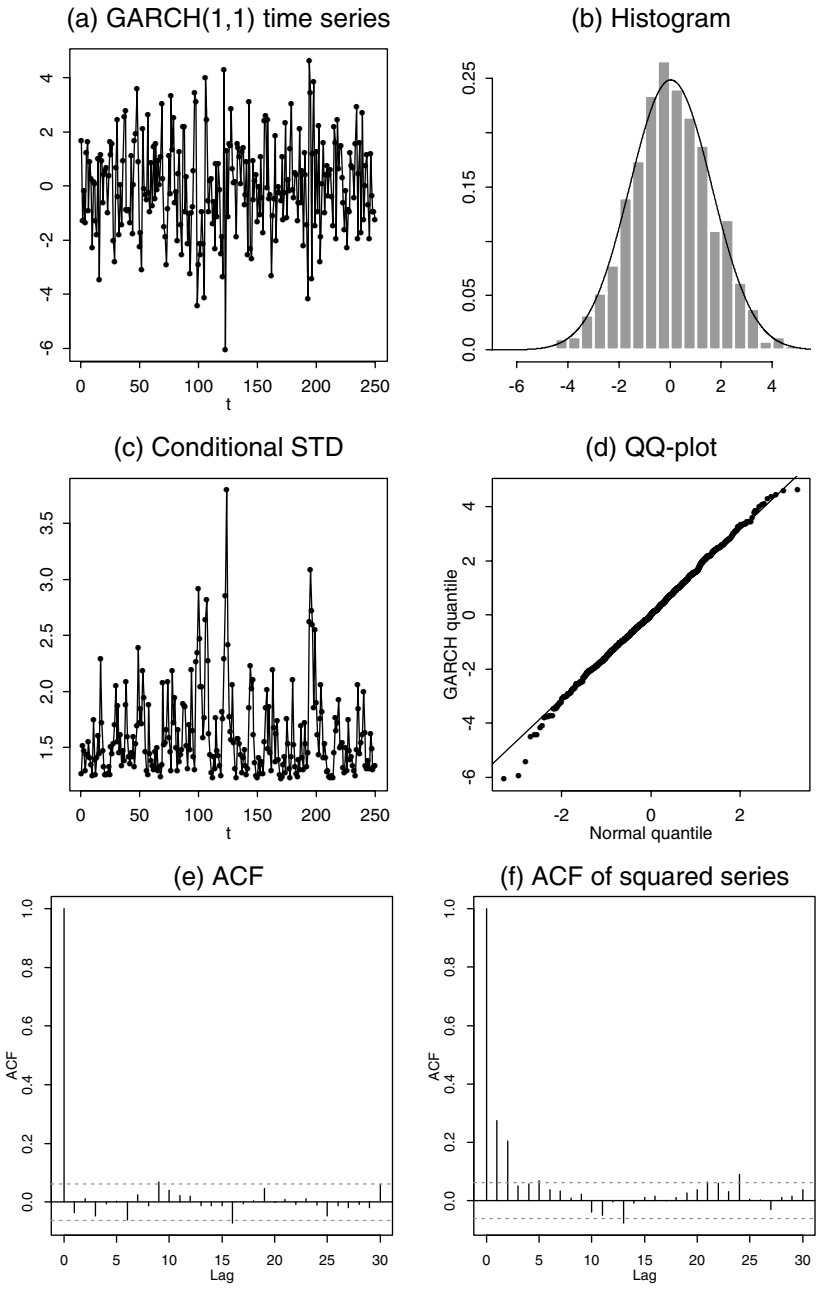


FIGURE 4.12. Example 4.2. A sample of 1,000 was generated from the GARCH(1, 1) model with $b_1 = 0.3$ and $a_1 = 0.3$: (a) and (c) time series plots of the first 250 X_t and σ_t; (b) normalized histogram and the normal density function with the same mean and variance; (d) plot: the sample quantiles versus the quantiles of $N(0, 1)$; (e) and (f) sample ACFs of $\{X_t\}$ and $\{X_t^2\}$, respectively.

Hence

$$\mathrm{Var}(X_t|X_{t-k}, k \geq 1) = \frac{c_0}{1-a_1} + b_1 \sum_{j=0}^{\infty} a_1^j X_{t-j-1}^2,$$

which depends on the infinite past of X_t. This is a marked difference from the ARCH(1) process for which $\mathrm{Var}(X_t|X_{t-k}, k \geq 1)$ depends on X_{t-1} only; see (4.23). This also indicates that the volatility cluster is more persistent for GARCH processes than for ARCH processes.

Nelson (1990) showed that the necessary and sufficient condition for existence of a strictly stationary GARCH(1, 1) process (with possible infinite second moment) is

$$E\{\log(b_1\varepsilon_t^2 + a_1)\} < 0.$$

By Jensen's inequality, $E\{\log(b_1\varepsilon_t^2 + a_1)\} \leq \log E(b_1\varepsilon_t^2 + a_1) = \log(b_1 + a_1)$. Hence, the condition $b_1 + a_1 < 1$ is sufficient to ensure that $\{X_t, t = 0, \pm 1, \cdots\}$ defined in (4.29) is strictly stationary. It becomes also a necessary condition if we require the strictly stationary solution to be also weakly stationary (having a finite second moment). Under the additional condition (4.28), $EX_t^4 < \infty$. Now, define $e_t = (\varepsilon_t^2 - 1)(c_0 + b_1 X_{t-1}^2 + a_1\sigma_{t-1}^2)$. It follows from (4.25) that

$$X_t^2 = c_0 + (b_1 + a_1)X_{t-1}^2 + e_t - a_1 e_{t-1};$$

that is, $\{X_t^2\}$ is a causal and invertible ARMA(1, 1) process. Its ACF is

$$\mathrm{Corr}(X_t^2, X_{t+k}^2) = \frac{(1 - a_1^2 - a_1 b_1)b_1}{1 - a_1^2 - 2a_1 b_1}(b_1 + a_1)^{k-1}, \quad k \geq 1. \tag{4.30}$$

We generated a series of length 1,000 from (4.29) with $c_0 = 1.5$, $b_1 = 0.6$, $a_1 = 0.3$, and $\varepsilon_t \sim N(0, 1)$. The first 250 sample points X_t together with their conditional standard deviation σ_t are shown in Figures 4.11 (a) and (c), respectively. Note that the GARCH(1, 1) process generated here has the same (unconditional) variance as the ARCH(1) process presented in Figure 4.9. However, the conditional variance of the GARCH process is much more volatile; compare Figures 4.9 (a) and (c) and Figure 4.11 (a) and (c). Furthermore, the GARCH process has more persistent volatility clusters. Figures 4.11 (b) and (d) show that the marginal distribution of X_t has leptokurtosis. The sample ACF indicates some significant autocorrelation in the squared X_t but not X_t itself.

We repeat the exercise above with a reduced $b_1 = 0.3$. Now $EX_t^4 < \infty$. The ACF of $\{X_t^2\}$ is well-defined. The sample ACF of $\{X_t^2\}$ plotted in Figure 4.12(f) provides a reasonable estimator for the true ACF, which is 0.337, 0.202, and 0.120 for $k = 1, 2$, and 3, respectively, obtained from (4.30). In contrast, when $b_1 = 0.6$, the ACVF of $\{X_t^2\}$ is not well-defined. The sample ACF plotted in Figure 4.11(f) does not resemble the function defined in (4.30). We also notice that, for the smaller $b_1 = 0.3$, the heavier tail property is no longer pronounced; see Figures 4.12 (b) and (d).

Although the kurtosis κ_x is a simple and intuitive measure to use in practice, it does not give a direct description of the heaviness of the tails of a distribution. A much more pertinent measure would be the tail index introduced by Kesten (1973). For the GARCH(1, 1) model defined in (4.29) with $a_1 + b_1 < 1$, we assume that ε_t^2 has a probability density function with unbounded support, and there exists a constant $\beta_0 \leq \infty$ such that

$$E(a_1 + b_1\varepsilon_t^2)^\beta < \infty \text{ for all } \beta < \beta_0 \text{ and } E(a_1 + b_1\varepsilon_t^2)^{\beta_0} = \infty.$$

Then, the equation

$$E(a_1 + b_1\varepsilon_t^2)^{\gamma/2} = 1 \tag{4.31}$$

defines a unique positive constant γ that is called a *tail index* in the sense that, as $x \to \infty$,

$$P\{|X_t| > x\} \sim E|\varepsilon_t|^\gamma P(\sigma_t > x) \sim Cx^{-\gamma}, \tag{4.32}$$

where $\{X_t\}$ is a strictly stationary solution of (4.29) and $\sigma_t = X_t/\varepsilon_t$. In the expression above, the sign "$\sim$" implies that the ratio of the two quantities on both sides has the limit 1, and $C > 0$ is a constant. Obviously, the estimation of γ is a difficult task. For further discussion on the tail index, see Kesten (1973), Goldie (1991), and de Haan, Resnick, Rootzen and de Vries (1989).

Before the end of this section, we point out that the condition (4.27) is not necessary for the existence of a strictly stationary solution for GARCH(p, q) model (4.24). In fact, Bougerol and Picard (1992b) proved that if the distribution of ε_t has unbounded support and has no atom at zero, and

$$\sum_{i=1}^{p} b_i + \sum_{j=1}^{q} a_j = 1, \tag{4.33}$$

there exists a unique strictly stationary process $\{X_t\}$ satisfying (4.24) and $EX_t^2 = \infty$. In analogy with integrated ARMA (ARIMA) processes (i.e., processes with unit roots), Engle and Bollerslev (1986) coined the name integrated GARCH(p, q) (IGARCH) for the GARCH(p, q) processes for which (4.33) holds. It is perhaps worth mentioning the possible confusion here: ARIMA processes are always nonstationary, whereas, as we have seen above, an IGARCH process may be strictly stationary.

4.2.3 Estimation

We always assume that $\{X_t\}$ is a strictly stationary solution of the GARCH model

$$X_t = \sigma_t\varepsilon_t \quad \text{and} \quad \sigma_t^2 = c_0 + \sum_{i=1}^{p} b_i X_{t-i}^2 + \sum_{j=1}^{q} a_j\sigma_{t-j}^2, \tag{4.34}$$

where $p \geq 1$, $q \geq 0$, $c_0, b_i, a_j > 0$, $\sum_{i=1}^{p} b_i + \sum_{j=1}^{q} a_j < 1$, and $\{\varepsilon_t\} \sim \text{IID}(0,1)$. Based on the observations $X_1, \cdots, X_T$, we discuss various methods for estimating parameters in the models. Our task here is to estimate conditional second moments, which, by virtue of their nature, are more difficult to estimate than conditional means. The likelihood functions, even when correctly specified, tend to be rather flat. Large sample sizes are often required in order to obtain reliable estimates.

We will introduce three types of estimators for parameters c_0, b_i, and a_j. They are *the conditional maximum likelihood estimator, Whittle's estimator*, and *the least absolute deviation estimator*. The first one is the benchmark estimator that has been widely used in the banking industry. The last one is appealing for the models with heavy-tailed errors.

(a) *Conditional maximum likelihood estimators*

Similar to the estimation for ARMA models (see §3.3.1), the most frequently-used estimators for ARCH/GARCH models are those derived from a (conditional) Gaussian likelihood function. For example, if ε_t is normal in model (4.34) and $q = 0$ (i.e., a pure ARCH model), the negative logarithm of the (conditional) likelihood function based on observations $X_1, \cdots, X_T$, ignoring constants, is

$$\sum_{t=p+1}^{T} \left(\log \sigma_t^2 + X_t^2 / \sigma_t^2 \right),$$

where $\sigma_t^2 = c_0 + \sum_{j=1}^{p} b_j X_{t-j}^2$. The (Gaussian) maximum likelihood estimators are defined as the minimizers of the function above. Note that this likelihood function is based on the conditional probability density function of $X_{p+1}, \cdots, X_T$, given $X_p, \cdots, X_1$, since the unconditional probability density function, which involves the joint density of $X_1, \cdots, X_p$, is unattainable.

For a general GARCH model (i.e., $q > 0$ in model (4.34)) the conditional variance σ_t^2 cannot be expressed in terms of a finite number of the past observations $X_{t-1}, X_{t-2}, \cdots$. Some truncation is inevitable. By induction, we may derive that

$$\begin{aligned} \sigma_t^2 &= \frac{c_0}{1 - \sum_{j=1}^{q} a_j} + \sum_{i=1}^{p} b_i X_{t-i}^2 \\ &+ \sum_{i=1}^{p} b_i \sum_{k=1}^{\infty} \sum_{j_1=1}^{q} \cdots \sum_{j_k=1}^{q} a_{j_1} \cdots a_{j_k} X_{t-i-j_1-\cdots-j_k}^2, \end{aligned} \tag{4.35}$$

where the multiple sum vanishes if $q = 0$. Note that the multiple sum above converges with probability 1 since each b_i and a_j is nonnegative, and since the expected value of the multiple series is finite. In practice, we replace

the expression above by a truncated version

$$\widetilde{\sigma}_t^2 = \frac{c_0}{1-\sum_{j=1}^q a_j} + \sum_{i=1}^p b_i X_{t-i}^2 + \sum_{i=1}^p b_i \sum_{k=1}^{\infty} \sum_{j_1=1}^q \cdots \sum_{j_k=1}^q a_{j_1} \cdots a_{j_k} \quad (4.36)$$
$$\times X_{t-i-j_1-\cdots-j_k}^2 I(t-i-j_1-\cdots-j_k \geq 1), \qquad t > p.$$

Note that when $q = 0$, $\widetilde{\sigma}_t^2 = \sigma_t^2 = c_0 + \sum_{i=1}^p b_i X_{t-i}^2$. Let $\mathbf{b} = (b_1, \cdots, b_p)^2$ and $\mathbf{a} = (a_1, \cdots, a_q)^\tau$. The (conditional) maximum likelihood estimator $(\widehat{\mathbf{b}}, \widehat{\mathbf{a}}, \widehat{c}_0)$ is defined by minimizing

$$l_\nu(c_0, \mathbf{b}, \mathbf{a}) = \sum_{t=\nu}^T \left(\log \widetilde{\sigma}_t^2 + X_t^2 / \widetilde{\sigma}_t^2 \right), \quad (4.37)$$

where $\nu > p$ is an integer.

The numerical calculation of the conditional maximum likelihood estimators above may be carried out by using the S+GARCH function "garch", which in fact can also compute the estimators derived from t-distributions and generalized Gaussian distributions. In general, suppose that $f(\cdot)$ is the probability density function of ε_t and is known. Then, the maximum likelihood estimators will be derived from minimizing

$$l_\nu(c_0, \mathbf{b}, \mathbf{a}) = \sum_{t=\nu}^T \{ \log \widetilde{\sigma}_t - \log f(X_t / \widetilde{\sigma}_t) \} \quad (4.38)$$

instead of (4.37). Apart from the normal distribution, some frequently used forms of $f(\cdot)$ are:

- *t-distribution* with v degrees of freedom:

$$f(x) = \frac{\Gamma((v+1)/2)}{(\pi v)^{1/2} \Gamma(v/2)} \left(\frac{v}{v-2} \right)^{1/2} \left(1 + \frac{x^2}{v-2} \right)^{-\frac{(v+1)}{2}},$$

  where $v > 2$ may be treated as a continuous parameter.

- *Generalized Gaussian distribution*:

$$f(x) = v \{ \lambda \, 2^{1+1/v} \Gamma(1/v) \}^{-1} \exp \left\{ -\frac{1}{2} \left| \frac{x}{\lambda} \right|^v \right\},$$

  where $\lambda = \{ 2^{-\frac{2}{v}} \Gamma(\frac{1}{v}) / \Gamma(\frac{3}{v}) \}^{\frac{1}{2}}$ and $0 < v < 2$.

When $v = 1$, the generalized Gaussian distribution reduces to the double exponential distribution $f(x) = \exp\{-\sqrt{2}|x|\}/\sqrt{2}$. All of the distributions above have been normalized to have mean 0 and variance 1, and all of them have heavier tails than normal distributions. For example, $E(|\varepsilon_t|^v) = \infty$ if $\varepsilon_t \sim t_v$.

S+GARCH may also compare two or more fitted models using the function "compare", which will print out the AIC and BIC values for all of the models. For model (4.34),

$$\text{AIC} = l_\nu(\widehat{c}_0, \widehat{\mathbf{b}}, \widehat{\mathbf{a}}) + 2(p+q+1), \tag{4.39}$$

and

$$\text{BIC} = l_\nu(\widehat{c}_0, \widehat{\mathbf{b}}, \widehat{\mathbf{a}}) + (p+q+1)\log(T-\nu+1), \tag{4.40}$$

where $l_\nu(\cdot)$ is defined in (4.38) (see §3.4.1 and §3.4.3).

(b) *Whittle's estimator*

For GARCH(p,q) defined in (4.34), the conditional variance can be written as

$$\sigma_t^2 = c_0 \Big/ \left(1 - \sum_{j=1}^{q} a_j\right) + \sum_{j=1}^{\infty} d_j X_{t-j}^2,$$

where $d_j \geq 0$ and $\sum_{j=1}^{\infty} d_j = \sum_{i=1}^{p} b_i/(1-\sum_{j=1}^{q} a_j)$; see the proof of Theorem 4.4. Suppose that $\{X_t\}$ is fourth-order stationary in the sense that its first four moments are all time-invariant (see the condition (4.28) in Theorem 4.4). Let $Y_t = X_t^2$. Then $\{Y_t\}$ is a stationary AR(∞) process satisfying

$$Y_t = c_0 \Big/ \left(1 - \sum_{j=1}^{q} a_j\right) + \sum_{j=1}^{\infty} d_j Y_{t-j} + e_t, \tag{4.41}$$

where e_t is a martingale difference

$$e_t = (\varepsilon_t^2 - 1)\left\{ c_0 \Big/ \left(1 - \sum_{j=1}^{q} a_j\right) + \sum_{j=1}^{\infty} d_j Y_{t-j} \right\}$$

with $\sigma_e^2 \equiv \text{Var}(e_t) < \infty$. Therefore, the spectral density of the process $\{Y_t\}$ is

$$g(\omega) = \frac{\sigma_e^2}{2\pi} \left| 1 - \sum_{j=1}^{\infty} d_j e^{ij\omega} \right|^{-2};$$

see Theorem 2.12. Based on (4.41), Giraitis and Robinson (2001) proposed the Whittle's estimators for b_i and a_j by minimizing

$$\sum_{j=1}^{T-1} I_T(\omega_j)/g(\omega_j),$$

where $I_T(\cdot)$ is the periodogram of $\{Y_t\}$ (see Definition 2.8), and $\omega_j = 2\pi j/T$; see also Theorem 2.14. Giraitis and Robinson (2001) also established the asymptotic normality for the estimators. Mikosch and Straumann (2000) investigated the Whittle estimation for a heavy-tailed GARCH(1, 1) model.

Whittle's estimators suffer from the lack of efficiency, as e_t is unlikely to be normal even when ε_t is normal. Furthermore the condition (4.28) is hardly fulfilled in financial applications.

(c) *L_1-estimation*

Both estimators discussed in (a) and (b) above are derived from maximizing a Gaussian likelihood or an approximate Gaussian likelihood. In this sense, they are L_2-estimators. It is well-known that L_1-estimators are more robust with respect to heavy-tailed distributions than L_2-estimators. Empirical evidence suggests that some financial time series exhibit heavy-tailed behavior and that the models based on distributions with heavier tails than those of a normal distribution would be more appropriate; see Mandelbrot (1963), Fama (1965), Mittnik, Rachev, and Paolella (1998), and Mittnik and Rachev (2000)

Based on this consideration, Peng and Yao (2002) proposed *least absolute deviations estimators* for c_0, b_i, and a_j in model (4.34) that minimize

$$\sum_{t=\nu}^{T} \left|\log(X_t^2) - \log(\widetilde{\sigma}_t^2)\right|, \tag{4.42}$$

where $\widetilde{\sigma}_t^2$ is defined in (4.36) and $\nu = p+1$ if $q = 0$ and $\nu > p+1$ if $q > 0$.

The idea behind (4.42) implies implicitly a reparameterization of model (4.34) such that $E\varepsilon_t = 0$ and the median (instead of variance) of ε_t^2 is equal to 1. Now, under this new setting, the parameters c_0 and the b_i's differ from those in the old setting by a common constant factor, whereas a_j's are unchanged. Note that in the regression model

$$\log(X_t^2) = \log(\sigma_t^2) + \log(\varepsilon_t^2),$$

the errors $\log(\varepsilon_t^2)$ are i.i.d. with median 0. This naturally leads to the estimators derived from minimizing (4.42). In fact, Peng and Yao (2002) showed that under very mild conditions, the least absolute deviations estimators are asymptotically normal with the standard convergence rate $T^{1/2}$ regardless of whether the distribution of ε_t has heavy tails or not. This is in marked contrast to the conditional maximum likelihood estimators derived from (4.37), which will suffer from slow convergence when ε_t is heavy-tailed; see Theorem 4.6 in §4.2.4.

Simulation comparisons between the least absolute deviations estimator and the conditional Gaussian maximum likelihood estimator were conducted for ARCH(2) and GARCH(1, 1) models, with ε_t being normal and t-distributed with 3 and 4 degrees of freedom in Peng and Yao (2002). The numerical results suggested that for models with very heavy-tailed errors (i.e., $\varepsilon_t \sim t_3$), the least absolute deviations estimator performed much better than the Gaussian maximum likelihood estimator. In contrast, when the error ε_t was normal, the Gaussian maximum likelihood estimator was

preferred. When $\varepsilon_t \sim t_4$, the performance of the two methods was comparable. In fact, the performance of the Gaussian maximum likelihood estimator decreases when the heaviness of the tails increases. However, this is not always the case for the least absolute deviations estimator, as it is robust against heavy tails.

*4.2.4 Asymptotic Properties of Conditional MLEs**

As discussed in the previous section, conditional maximum likelihood estimation remains as one of the most frequently-used methods in fitting GARCH models. In practice, the distribution of ε_t is typically unknown. The estimator derived from a Gaussian likelihood is often employed. Hall and Yao (2003) established the asymptotic properties of this estimator, ranging from nonheavy-tailed to heavy-tailed errors. The results will be presented below in a compact manner. For further mathematical rigors, see Hall and Yao (2003).

Let $\{X_t\}$ be the strictly stationary solution from GARCH(p, q) model (4.34) in which ε_t may not be normal. We assume that $p \geq 1$, $q \geq 0$, $c_0 > 0$, $b_j > 0$ for $j = 1, \cdots, p$, and $a_i > 0$ for $i = 1, \cdots, q$ when $q > 0$. Let $(\widehat{c}_0, \widehat{\mathbf{b}}, \widehat{\mathbf{a}})$ be the estimator derived from minimizing (4.37), which should be viewed as a (conditional) quasimaximum likelihood estimator. We also assume that in (4.37) $\nu = \nu(T)$ diverges to infinite at a rate sufficiently slow to ensure $\nu/T \to 0$ as $T \to \infty$. Theorems 4.5 and 4.6 below present, respectively, its asymptotic distributions for the case of nonheavy-tailed errors and the case of heavy-tailed errors. To this end, we introduce some notation first.

Let $\boldsymbol{\theta} = (c_0, \mathbf{b}^\tau, \mathbf{a}^\tau)^\tau$, $\widehat{\boldsymbol{\theta}} = (\widehat{c}_0, \widehat{\mathbf{b}}^\tau, \widehat{\mathbf{a}}^\tau)^\tau$, and $\mathbf{U}_t = \frac{\partial \sigma_t^2}{\partial \boldsymbol{\theta}}$. It may be shown that $\mathbf{U}_t/\sigma_t^2$ has all of its moments finite. We assume that the matrix

$$\mathbf{M} \equiv E\left(\mathbf{U}_t \mathbf{U}_t^\tau / \sigma_t^4\right) > 0$$

is positive-definite. Let $\{V_t\}$ be a sequence of independent random variables with the same distribution as $\mathbf{M}^{-1}\mathbf{U}_1/\sigma_1^2$. Let $Y_1, Y_2, \ldots$ represent the infinite extension of the multivariate joint extreme-value distribution of the first type, with exponent α. In other words, for each k, the distribution of $(Y_1, \ldots, Y_k)$ is the limiting joint distribution, as $n \to \infty$, of the k largest values of a sample of size n drawn from a distribution in the domain of attraction of the first type of extreme-value distribution, after appropriate normalization for scale. More precisely, we assume that the normalization is chosen such that Y_1 has distribution function $\exp(-y^{-\alpha})$ for $y > 0$. Hall (1978) formulated a representation of the distribution of the full process $Y_1, Y_2, \cdots$. We assume that $\{V_k\}$ and $\{Y_k\}$ are independent of each other. Put

$$W_0 = \sum_{k=1}^{\infty} \{Y_k V_k - E(Y_k)E(V_1)\}.$$

Now, Theorems 4.5 and 4.6 hold under the conditions assumed above. For their proofs, see Hall and Yao (2003). Part (i) of Theorem 4.5 below shows that the finite fourth moment of ε_t will ensure asymptotic normality. Part (ii) shows that the condition can be relaxed so that ε_t^2 is in the domain of attraction of the normal distribution. In general, a distribution G is said to be in *the domain of attraction* of a distribution F if

$$a_n^{-1}(S_n - b_n) \xrightarrow{D} F, \quad \text{as } n \to \infty,$$

where $S_n = \sum_{i=1}^n \xi_i$, $\{\xi_i\} \sim_{\text{i.i.d.}} G$, and $a_n > 0$ and b_n are some constants. See §6.1 of Feller (1971).

Theorem 4.5 *(Hall and Yao 2003)*
(i) If $E(\varepsilon_t^4) < \infty$,

$$\frac{T^{1/2}}{\{E(\varepsilon_t^4) - 1\}^{1/2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N(0, \mathbf{M}^{-1}).$$

(ii) If $E(\varepsilon_t^4) = \infty$ *and the distribution of* ε_t^2 *is in the domain of attraction of the normal distribution, then*

$$\frac{T}{\lambda_T}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N(0, \mathbf{M}^{-1}),$$

where

$$\lambda_T = \inf[\, \lambda > 0 : \; E\{\varepsilon_1^4 I(\varepsilon_1^2 \le \lambda)\} \le \lambda^2 / T \,].$$

Intimately related to the concept of domain of attraction is the distribution family of stable laws. Let $\{\eta_i\} \sim_{\text{i.i.d.}} F$ and $S_n = \sum_{i=1}^n \eta_i$. A nondegenerate distribution F is *stable* if, for each $n \ge 2$, there exist some constants $\alpha > 0, \gamma_n$ such that the distributions of S_n and $n^\alpha \eta_1 + \gamma_n$ are the same, where α is called the exponent of the stable law. It measures the tail-heaviness of a distribution; the smaller the value of α, the heavier the tails. It is known that $\alpha \in (0, 2]$ and $\alpha = 2$ corresponds to normal distributions. Furthermore, a distribution is stable if and only if it has a non-empty domain of attraction; see also §17.5 of Feller (1971). Now, we are ready to state the asymptotic properties of $\widehat{\boldsymbol{\theta}}$ when ε_t^2 has a heavy-tailed distribution.

Theorem 4.6 *(Hall and Yao 2003)*
If the distribution of ε_t^2 *is in the domain of attraction of a stable law with the exponent* $\alpha \in (1, 2)$*, then*

$$\frac{T}{\lambda_T}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} W_0,$$

where

$$\lambda_T = \inf\{\lambda > 0 : \; TP(\varepsilon_1^2 > \lambda) \le 1\}.$$

Note that if the distribution of a random variable ξ belongs to the domain of attraction of a stable law with exponent α, $E|\xi|^{\alpha+\epsilon} = \infty$ and $E|\xi|^{\alpha-\epsilon} < \infty$ for any $\epsilon > 0$. Furthermore, the λ_T, defined in Theorem 4.6, increases as α increases. Therefore, the convergence rate of the estimator $\widehat{\boldsymbol{\theta}}$ is dictated by the distribution tails of ε_t^2; the heavier the tails, the slower the convergence. When a classic central limit theorem holds and the limiting distribution is normal, all of the terms in the partial sums are equally important and none of them dominates the others. This is no longer the case for a heavy-tailed the distribution with, typically, infinite second moment (i.e., $E\{(\varepsilon_t^2)^2\} = \infty$ in the current setting). For heavy-tailed distributions, the partial sums are dominated by a few extremely large values from the tails; see the definition of $\{Y_t\}$ above. Therefore, it will take considerably longer before the partial sums settle, resulting in slower convergence rates.

4.2.5 Bootstrap Confidence Intervals

Theorems 4.5 and 4.6 indicate that the range of possible limit distributions for a (conditional) Gaussian maximum likelihood estimator is extraordinarily vast. In particular, the limit laws are not restricted to a class that can be described by a finite number of parameters. Rather, they depend intimately on the error distribution in its entirety. This makes it impossible (in the heavy-tailed cases) to perform statistical tests or interval estimation based on asymptotic distributions in any conventional sense. Bootstrap methods seem the best option for tackling these problems.

However, it is well known that in the settings where the limiting distribution of a statistic is not normal, standard bootstrap methods are generally not consistent when used to approximate the distribution of the statistic; see, for example, Mammen (1992). To some extent, subsampling methods can be used to overcome the problem of inconsistency; see Bickel, Götze, and van Zwet (1995). However, although this approach consistently approximates the distribution of a statistic, it does so only for a value of sample size that has to be an order of magnitude less than the true sample size. As a result, a confidence interval based on the subsample bootstrap can be very conservative. In the absence of an accurate method for adjusting scale, which typically depends on the convergence rate, subsampling can be unattractive. In this section, we introduce a percentile-t form of the *subsampling method* introduced by Hall and Yao (2003), which gives consistent confidence intervals for parameters in GARCH models. Note that the percentile-t method is usually employed in order to attain a high order of accuracy in approximations where the limiting distribution is normal. That is not the main goal in this context. Instead, it is used to avoid the rescaling of the distribution from subsamplings.

By Theorems 4.5 and 4.6, it holds that

$$\frac{T}{\lambda_T}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} W, \tag{4.43}$$

provided that the distribution of ε_t^2 is in the domain of attraction of a stable law with exponent $\alpha \in (1, 2]$, where W has a proper nondegenerate distribution, and the convergence rate $\frac{T}{\lambda_T}$ is unknown and it may depend on the underlying distribution intimately. To get rid of the influence of this unknown convergence rate, we define

$$\widehat{\tau}^2 = \frac{1}{T}\sum_{t=1}^{T} \widetilde{\varepsilon}_t^4 - \left(\frac{1}{T}\sum_{t=1}^{T} \widetilde{\varepsilon}_t^2\right)^2, \tag{4.44}$$

the sample standard deviation of $\{\widetilde{\varepsilon}_t^2\}$, where $\widetilde{\varepsilon}_t = X_t/\widetilde{\sigma}_t(\widehat{\boldsymbol{\theta}})$. It may be proved (Hall and Yao 2003) that

$$\lambda_T^{-1}\{T(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}), T^{1/2}\widehat{\tau}\} \xrightarrow{D} (W, S),$$

where S is a random variable with $P(0 < S < \infty) = 1$. The studentized statistic is defined as the ratio of the two statistics on the left-hand side of the expression above, which admits the asymptotic distribution

$$T^{\frac{1}{2}} \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{\widehat{\tau}} \xrightarrow{D} W/S \tag{4.45}$$

with the known convergence rate $T^{\frac{1}{2}}$.

Let $\widetilde{\varepsilon}_t = X_t/\widetilde{\sigma}(\widehat{\boldsymbol{\theta}})$ for $t = \nu, \cdots, T$, and let $\widehat{\varepsilon}_\nu, \cdots, \widehat{\varepsilon}_T$ be the standardized version of $\{\widetilde{\varepsilon}_t\}$ such that the sample mean is zero and the sample variance is 1. Now, we draw ε_t^* with replacement from $\{\widehat{\varepsilon}_t\}$ and define $X_t^* = \sigma_t^* \varepsilon_t^*$ for $t = \nu, \cdots, m$ with

$$(\sigma_t^*)^2 = \widehat{c}_0 + \sum_{i=1}^{p} \widehat{b}_i (X_{t-i}^*)^2 + \sum_{j=1}^{q} \widehat{a}_j (\sigma_{t-j}^*)^2$$

and form the statistic $(\widehat{\boldsymbol{\theta}}^*, \widehat{\tau}^*)$ based on $\{X_\nu^*, \cdots, X_m^*\}$ in the same way as $(\widehat{\boldsymbol{\theta}}, \widehat{\tau})$ based on $\{X_\nu, \cdots, X_T\}$. Hall and Yao (2003) proved that as $T \to \infty$, $m \to \infty$, and $m/T \to 0$, it holds for any convex set C that

$$\left| P\left\{ \sqrt{m}\ \frac{(\widehat{\boldsymbol{\theta}}^* - \widehat{\boldsymbol{\theta}})}{\widehat{\tau}^*} \in C \middle| X_1, \cdots X_T \right\} - P\left\{ \sqrt{T}\ \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{\widehat{\tau}} \in C \right\} \right| \to 0.$$

Based on this, a one-sided *bootstrap confidence interval* for, for example, θ_k, the kth component of $\boldsymbol{\theta}$ with the confidence level $\pi \in (0, 1)$ is defined as

$$\big[\, \widehat{\theta}_k - T^{-\frac{1}{2}}\, \widehat{\tau}\, \widehat{u}_\pi,\ \infty\big), \tag{4.46}$$

where

$$\widehat{u}_\pi = \inf\left\{u:\ P\big[m^{1/2}(\widehat{\tau}^*)^{-1}(\widehat{\theta}_k^* - \widehat{\theta}_k) \le u \big| X_1, \cdots, X_T\big] \ge \pi\right\}. \quad (4.47)$$

Both left-sided intervals and two-sided intervals may be constructed in a similar manner. The simulation results reported in Hall and Yao (2003) indicated that the procedure worked reasonably well for ARCH(2) and GARCH(1, 1) models with heavy or nonheavy-tailed errors and is insensitive to the choice of the values of m in the range of 50%–80% of the original sample size T.

In the heavy-tailed case, the limit properties of the maximum likelihood estimators are dictated by the behavior of extreme order statistics. The reason that the full-sample (i.e., T-out-of-T) bootstrap fails to be consistent is that it does not accurately model relationships among extreme order statistics in the sample. For example, for each fixed $k \ge 2$, the probability that the k largest values in a resample are equal does not converge to 0 in the T-out-of-T bootstrap. The probability does converge to 0 for the m-out-of-T bootstrap, provided $m/T \to 0$, and of course it converges to 0 for the sample itself.

In principle, we may construct confidence intervals in terms of a bootstrap approximation for the distribution of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ directly. Since we have to use a subsampling bootstrap, the intervals so constructed will be for the sample size m. Those intervals would have to be converted to those for the sample size T according to the unknown convergence rate T/λ_T, which is practically infeasible. The statistic $\widehat{\tau}$ defined in (4.44) was introduced to studentize $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$. Note that the studentized statistic $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})/\widehat{\tau}$ has a known convergence rate $T^{\frac{1}{2}}$; see (4.45). Therefore, the conversion can be done with ease; see (4.46) and (4.47). Of course, we lose some precision in the estimation due to the introduction of random quantity $\widehat{\tau}$.

4.2.6 Testing for the ARCH Effect

It becomes a routine practice in analyzing financial data, for example, to test the existence of conditional heteroscedasticity. Neglect of the conditional heteroscedasticity may lead to a loss in asymptotic efficiency of parameter estimation (Engle 1982) and can result in overparameterization of an ARMA model (Weiss 1984). It may also cause overrejection of conventional tests, such as (7.29), for serial correlation in mean (Milhoj 1985; Taylor 1986). In principle, we may test the hypotheses of the parameters a_i and b_j based on the bootstrap confidence intervals developed in the last section. We introduce in this section some methods based on more traditional parametric methods that model conditional heteroscedasticity in terms of an ARCH specification. Those methods may be more efficient when additional information on the distribution of ε_t is available and ε_t has finite high order (≥ 4) moments.

Suppose that $\{X_t\}$ is a strictly stationary process defined by

$$X_t = \sigma_t \varepsilon_t, \qquad \sigma_t^2 = c_0 + \sum_{i=1}^{p} b_j X_{t-i}^2, \tag{4.48}$$

where $c_0 > 0$ and $a_j \geq 0$. We are interested in testing the null hypothesis of no ARCH effects, which can be formulated as

$$H_0 : \ b_1 = \cdots = b_p = 0, \tag{4.49}$$

against the alternative hypothesis

$$H_1 : \ b_j \neq 0 \text{ for at least one } j.$$

It is easy to see that, under H_0, the conditional variance $\sigma_t^2 \equiv c_0$ is a constant. If the density function $f(\cdot)$ of ε_t in (4.48) is known, a natural approach is to use the *(conditional) likelihood ratio test* based on the test statistic

$$S_{T,1} = \prod_{t=p+1}^{T} \frac{\sigma_t(\widehat{c}_0, \widehat{\mathbf{b}})^{-1} f\{X_t/\sigma_t(\widehat{c}_0, \widehat{\mathbf{b}})\}}{\sigma_t(\widetilde{c}_0, 0)^{-1} f\{X_t/\sigma_t(\widetilde{c}_0, 0)\}}, \tag{4.50}$$

where $(\widehat{c}_0, \widehat{\mathbf{b}})$ is the maximum (conditional) likelihood estimator for model (4.48) and $\widetilde{c}_0$ is the constrained maximum (conditional) likelihood estimator under the null hypothesis (4.49). It is well-known that under the null hypothesis H_0

$$2\log(S_{T,1}) \xrightarrow{D} \chi_p^2,$$

provided that the density function $f(\cdot)$ is smooth enough and the Fisher information matrix

$$\mathbf{I}(c_0, \mathbf{b}) \equiv \begin{pmatrix} I_{11}(c_0, \mathbf{b}) & \mathbf{I}_{12}(c_0, \mathbf{b}) \\ \mathbf{I}_{21}(c_0, \mathbf{b}) & \mathbf{I}_{22}(c_0, \mathbf{b}) \end{pmatrix} = E\{\dot{\ell}(X_t; c_0, \mathbf{b})\dot{\ell}(X_t; c_0, \mathbf{b})^\tau\} \tag{4.51}$$

exists and is positive-definite; see §4.4.4 of Serfling (1980). (Those conditions are fulfilled for ARCH models with $\varepsilon_t \sim N(0,1)$.) In the expression above,

$$\dot{\ell}(X_t; c_0, \mathbf{b}) = \begin{pmatrix} \dot{\ell}_1(X_t; c_0, \mathbf{b}) \\ \dot{\ell}_2(X_t; c_0, \mathbf{b}) \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial c_0} \log[\sigma_t^{-1} f(X_t/\sigma_t)] \\ \frac{\partial}{\partial \mathbf{b}} \log[\sigma_t^{-1} f(X_t/\sigma_t)] \end{pmatrix}.$$

Furthermore, the asymptotic distribution under the null hypothesis is also shared by both the score test (4.52) advocated by Engle (1982) and the Wald test (4.54) below.

The score test is also called the Rao test or the Lagrange multiplier test. It is based on the fact that the gradient of a log-likelihood function should

be close to 0 under a null hypothesis. More precisely, it can be shown that under H_0,

$$\frac{1}{\sqrt{T-p}} \sum_{t=p+1}^{T} \dot{\ell}_2(X_t; \widetilde{c}_0, \mathbf{0}) \xrightarrow{D} N\big(\mathbf{0}, \mathbf{I}_{22}(c_0, \mathbf{0})\big),$$

where $\mathbf{I}_{22}$ is defined in (4.51). The *score statistic* is defined as

$$S_{T,2} = \frac{1}{T-p} \left\{ \sum_{t=p+1} \dot{\ell}_2(X_t; \widetilde{c}_0, \mathbf{0}) \right\}^{\tau} \big\{ \mathbf{I}_{22}(\widetilde{c}_0, \mathbf{0}) \big\}^{-1} \sum_{t=p+1} \dot{\ell}_2(X_t; \widetilde{c}_0, \mathbf{0}). \tag{4.52}$$

An advantage of the score test is that it does not require computation of $\widehat{c}_0$ and $\widehat{\mathbf{b}}$.

If ε_t in (4.48) is normal, Engle (1982) has shown that the score test may be performed in terms of the statistic TR^2, which is asymptotically equivalent to $S_{T,2}$, where R^2 is the multiple correlation coefficient of X_t^2 and $(X_{t-1}^2, \cdots, X_{t-p}^2)$, namely

$$R^2 = \mathbf{X}_{p+1}^{\tau} \mathcal{X} (\mathcal{X}^{\tau} \mathcal{X})^{-1} \mathcal{X}^{\tau} \mathbf{X}_{p+1} / (\mathbf{X}_{p+1}^{\tau} \mathbf{X}_{p+1}), \tag{4.53}$$

where $\mathbf{X}_k = (X_k^2, X_{k+1}^2, \cdots, X_{T-p-1+k}^2)^{\tau}$, $\mathcal{X} = (\mathbf{1}, \mathbf{X}_p, \mathbf{X}_{p-1}, \cdots, \mathbf{X}_1)$, and $\mathbf{1}$ is a vector with all components 1. Although the asymptotic equivalence is justified for the models with normal errors only, the statistic TR^2 has been used also for nonnormal cases. Since R^2 is the percentage of the part of the variation of X_t^2 that can be explained in terms of its p lagged values, the large values of R^2 are indicative of a linear dependence of X_t^2 on $X_{t-1}^2, \cdots, X_{t-p}^2$. Therefore, it is a sound test statistic for testing the ARCH effect. However, its asymptotic properties are less clear when ε_t is not normal.

The Wald test directly compares the MLE $\widehat{\mathbf{b}}$ with $\mathbf{b} = \mathbf{0}$, the parameter value under the null hypothesis H_0. Note that under H_0, $(T-p)^{\frac{1}{2}} \widehat{\mathbf{b}} \xrightarrow{D} N(\mathbf{0}, \mathbf{I}^{22}(c_0, \mathbf{0}))$, where $\mathbf{I}^{22}$ is the lower $p \times p$ submatrix of $\{\mathbf{I}(c_0, \mathbf{b})\}^{-1}$, which equals to

$$\mathbf{I}^{22}(c_0, \mathbf{b}) = \{\mathbf{I}_{22}(c_0, \mathbf{b}) - \mathbf{I}_{21}(c_0, \mathbf{b}) I_{11}(c_0, \mathbf{b})^{-1} \mathbf{I}_{12}(c_0, \mathbf{b})\}^{-1}.$$

The *Wald statistic* is defined as

$$S_{T,3} = (T-p)\, \widehat{\mathbf{b}}^{\tau} \{\mathbf{I}^{22}(\widehat{c}_0, \widehat{\mathbf{b}})\}^{-1} \widehat{\mathbf{b}}. \tag{4.54}$$

Section 4.4.4 of Serfling (1980) discussed the asymptotic properties of the three tests defined above. Although all three statistics $2\log(S_{T,1})$, $S_{T,2}$, and $S_{T,3}$ share the same asymptotic distribution χ_p^2 under H_0, there are some practical differences in the use of these tests. For example, it is anticipated that they will not have the same power at fixed alternatives. Likelihood

ratio tests are invariant to one-to-one transformations for both random variables and parameters, a property not shared by score tests and Wald tests. On the other hand, the score statistic is potentially simpler from a computational point of view since it depends on the estimator $\widetilde{c}_0$ only; see (4.52).

Based on truncation (4.36), we may extend the test above for testing H_0 against a GARCH(p, q) alternative. This is effectively to test H_0 against an ARCH(∞) alternative. Since our primary interest in this practice is to detect the existence of conditional heteroscedasticity, we may simply test H_0 against an ARCH(p) alternative in which the choice of p is not so critical. Once the null hypothesis H_0 is rejected, we may select an adequate model in terms of AIC or BIC; see (4.39) and (4.40).

Note that in forming the test statistic (4.50) we did not make use of the information that $b_j \geq 0$. (If the estimator $\widehat{b}_j$'s are restricted to be nonnegative, the χ^2-asymptotic approximation stated above is not necessarily valid since the parameter value $b_j = 0$ is at the boundary rather than the interior of the parameter space). Literature concerning testing for H_0 against a one-sided alternative (i.e., $H_1 : b_j > 0$ for some j) includes Lee and King (1993) and Hong (1997).

4.2.7 ARCH Modeling of Financial Data

The direct motivation for introducing ARCH models is to evaluate and/or to forecast risk in financial time series in the form of conditional heteroscedasticity. Standard examples of financial time series are the prices of company-shares quoted at major stock exchanges, interest rates set by governments or major national banks, and foreign exchange rates among different currencies. For stock prices, data sets can be intra-daily "tick-by-tick" trade data. This means that every trade in a specific stock is recorded together with the time when the trade took place. However, most data analyzed in terms of statistical models (such as GARCH) are daily data for which only a single number is recorded for each day. Most commonly analyzed daily stock prices contain only daily closing prices.

Since financial data typically have the autocorrelation coefficient close to 1 at lag 1 (e.g., the exchange rate between the U.S. dollar and pounds sterling hardly changes from today to tomorrow), it is much more interesting and also practically more relevant to model the returns of a financial series rather than the series itself. Let $\{Y_t\}$ be a stock price series, for example. The *returns* are typically defined as

$$X_t = \log Y_t - \log Y_{t-1} \quad \text{or} \quad X_t = \frac{Y_t - Y_{t-1}}{Y_{t-1}},$$

which measure the relative changes of price. Note that the two forms above are approximately the same as

$$\log Y_t - \log Y_{t-1} = \log\left(1 + \frac{Y_t - Y_{t-1}}{Y_{t-1}}\right) \approx (Y_t - Y_{t-1})/Y_{t-1}.$$

Rydberg (2000) summarizes some important *stylized features* of financial return series, which have been repeatedly observed in all kinds of assets including stock prices, interest rates, and foreign exchange rates. We list below some of those features for daily data.

(i) *Heavy tails.* It has been generally accepted that the distribution of the return X_t has tails heavier than the tails of a normal distribution. Typically, it is assumed that X_t only has a finite number of finite moments, although it is still an ongoing debate how many moments actually exist. Nevertheless, it seems a general agreement nowadays to assume that the daily return has a finite second moment (i.e. $EX_t^2 < \infty$). This also serves as a prerequisite for ARCH/GARCH modeling.

(ii) *Volatility clustering.* The term volatility clustering refers to the fact that large price changes occur in clusters. Indeed, large volatility changes tend to be followed by large volatility changes, and periods of tranquillity alternate with periods of high volatility; see, for example, Figure 4.16(a) below.

(iii) *Asymmetry.* There is evidence that the distribution of stock returns is slightly negatively skewed. One possible explanation could be that traders react more strongly to negative information than positive information.

(iv) *Aggregational Gaussianity*. When the sampling frequency decreases, the central limit law sets in and the distribution of the returns over a long time-horizon tends toward a normal distribution. Note that a return over k days is simply the aggregation of k daily returns:

$$\log Y_k - \log Y_0 = \sum_{t=1}^{k} (\log Y_t - \log Y_{t-1}) = \sum_{t=1}^{k} X_t.$$

(v) *Long range dependence*. The returns themselves of all kinds of assets hardly show any serial correlation, which, however, does not mean that they are independent. In fact, both squared returns and absolute returns often exhibit persistent autocorrelations, indicating possible long-memory dependence in those transformed return series; see, for example, Figure 2.7.

ARCH and GARCH models may catch three out of the five stylized features listed above, namely (i), (ii), and (iv). First, ARCH and GARCH processes defined in terms of normal errors are innately heavy-tailed; see Propositions 4.1(iii) and 4.2(ii). The models with heavy-tailed errors such as $\varepsilon_t \sim t_k$ for $k = 4$ or 3 have been used to model very heavy-tailed data in practice. However we should not overlook the fact that a GARCH model with normal errors may be very heavy-tailed as well; see (4.31) and (4.32). Further, the volatility clustering is also portrayed naturally in ARCH and GARCH models; see (4.20). Note that a GARCH(p, q) process is effectively an ARCH(∞) process (see (4.35)). Therefore (4.20) also holds for a GARCH process with $p = \infty$, indicating even more persistent volatility clustering. Finally, a strictly stationary GARCH process $\{X_t\}$ with $EX_t^2 < \infty$ is also a sequence of martingale differences. Therefore, it may hold that $T^{1/2} \sum_{t=1}^{T} X_t$ is asymptotically normal; see Theorem 4 on p. 511 of Shiryayev (1984). Therefore, the aggregational Gaussianity holds.

However, GARCH models, in their classic form, fail to catch the stylized features (iii) asymmetry and (v) long-range dependence. Extension of the classic GARCH form to model these, and also other, stylized features received ample attention in the literature. We list below a few extended GARCH models (in their simplest forms) that are often used in practice. Shephard (1996) provides a more comprehensive survey on extended GARCH models.

(a) *EGARCH*

Nelson (1991) introduced an *exponential GARCH* (EGARCH) model that specifies the model

$$X_t = \varepsilon_t \exp(h_t/2), \quad h_t = \gamma_0 + \gamma_1 h_{t-1} + g(\varepsilon_{t-1}),$$

where

$$g(x) = \omega x + \lambda(|x| - E|x|). \tag{4.55}$$

In contrast to the form of σ_t in the GARCH model, the value of the function $g(\cdot)$ depends on both the size and the sign of its argument. As a result, EGARCH responds nonsymmetrically to random shocks ε_t. Although (4.55) looks somewhat complicated, it is rather straightforward to identify the properties of the process $\{h_t\}$ and therefore also those of $\{X_t\}$. Note that $\{g(\varepsilon_t)\}$ is i.i.d. provided $\{\varepsilon_t\}$ is i.i.d. Therefore $\{h_t\}$ is a causal linear AR(1) process if $|\gamma_1| < 1$. For further discussion on EGARCH models, see Bollerslev, Engle, and Nelson (1994).

(b) *FIARCH*

In order to model the persistent correlations in squared returns, attempts have been made to construct long-memory ARCH type models. Similar

to FARIMA models defined in §2.5.2, we define a fractionally integrated ARCH (FIARCH) model as

$$X_t = \sigma_t \varepsilon_t, \quad \sigma_t^2 = c_0 + \{1 - (1 - B)^d\} X_{t-1}^2 = c_0 + b(B) X_{t-1}^2,$$

where $d \in (-0.5, 0.5)$, and $b(\cdot)$ is a polynomial with hyperbolically (rather than exponentially) decaying coefficients. Although a FIARCH model is in the form of ARCH(∞), the slow-decaying coefficients cause the long-term autocorrelation in the series $\{X_t^2\}$. For further discussion on FIARCH processes, see Baillie, Bollerslev and Mikkelsen (1996), Ding and Granger (1996), Robinson and Zaffaroni (1998), and Mikosch and Stărică (1999).

(c) *ARCH-M*

In finance theory, the relationship between risk and return plays a predominant role. If we take conditional deviation as a measure for risk, it is possible to use risk as a regressor in modeling returns. Engle, Lilien, and Robins (1987) proposed the ARCH in mean (ARCH-M) model

$$X_t = g(\sigma_t^2, \boldsymbol{\theta}) + \varepsilon_t \sigma_t, \quad \sigma_t^2 = c_0 + b\{X_{t-1} - g(\sigma_{t-1}^2, \boldsymbol{\theta})\}^2.$$

A commonly used parameterization is the linear one: $g(y, \boldsymbol{\theta}) = \theta_0 + \theta_1 y$. See Hong (1991) for its statistical properties.

Finally, we point out that many different types of models have been proposed for the modeling of financial data, including the ARCH/GARCH model discussed in this section. The stochastic volatility models (Shephard 1996; also §4.3.3 below) form another class of popular statistical models. See also Rydberg (2000) for references on various models in the category of mathematical finance.

4.2.8 A Numerical Example: Modeling S&P 500 Index Returns

We illustrate the GARCH modeling techniques in terms of the daily S&P 500 index data from January 3, 1972 to December 31, 1999 introduced in Example 1.4. We define the returns

$$X_t = 100(\log Y_t - \log Y_{t-1}),$$

where Y_t is the index at time t. The sample size is $T = 7075$. Numerical fitting of GARCH models was performed using the S+GARCH function garch.

(a) *Graphical investigation*

The S&P 500 returns are plotted in Figure 4.16(a), in which the large sparks around $t = 4,000$ correspond to the stock market crash in October

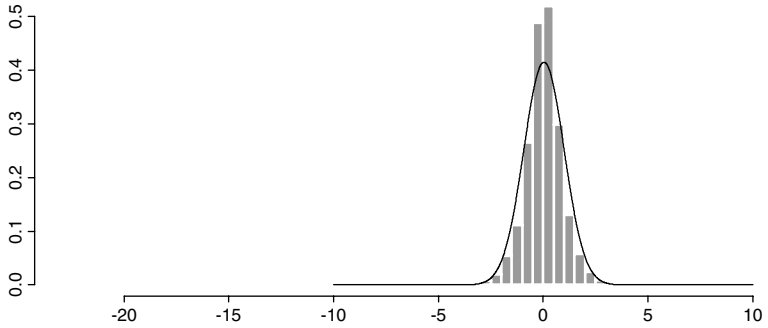


FIGURE 4.13. Histogram of the S&P 500 returns and a normal density function with the same mean and variance.

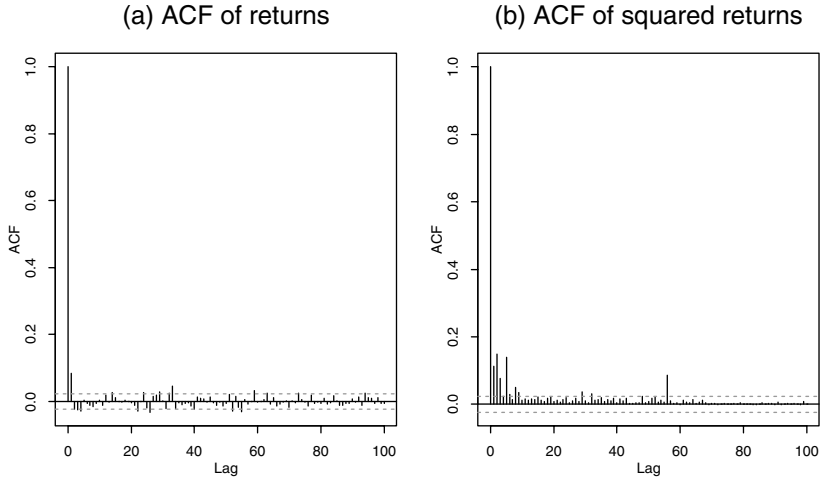


FIGURE 4.14. Correlogram of (a) the S&P 500 returns, and (b) the squared returns.

1987. The histogram in Figure 4.13 has a long stretch on its left due to the single large negative return. However, if we discard this single "outlier", the marginal distribution seems fairly symmetric but not normal. The correlogram in Figure 4.14 shows that there is almost no significant autocorrelation in the return series $\{X_t\}$ itself, but such an autocorrelation does exist in the squared series $\{X_t^2\}$. Figure 4.15 presents the plots of the returns versus the normal distribution and Student's t-distributions with degrees of freedom ranging from 7 to 3. (The plots against t-distributions were produced by the S+GARCH function "aqqplot"). Those plots are informative for identifying the moment condition. For example, both tails of the empirical distribution of the returns are heavier than those of the nor-

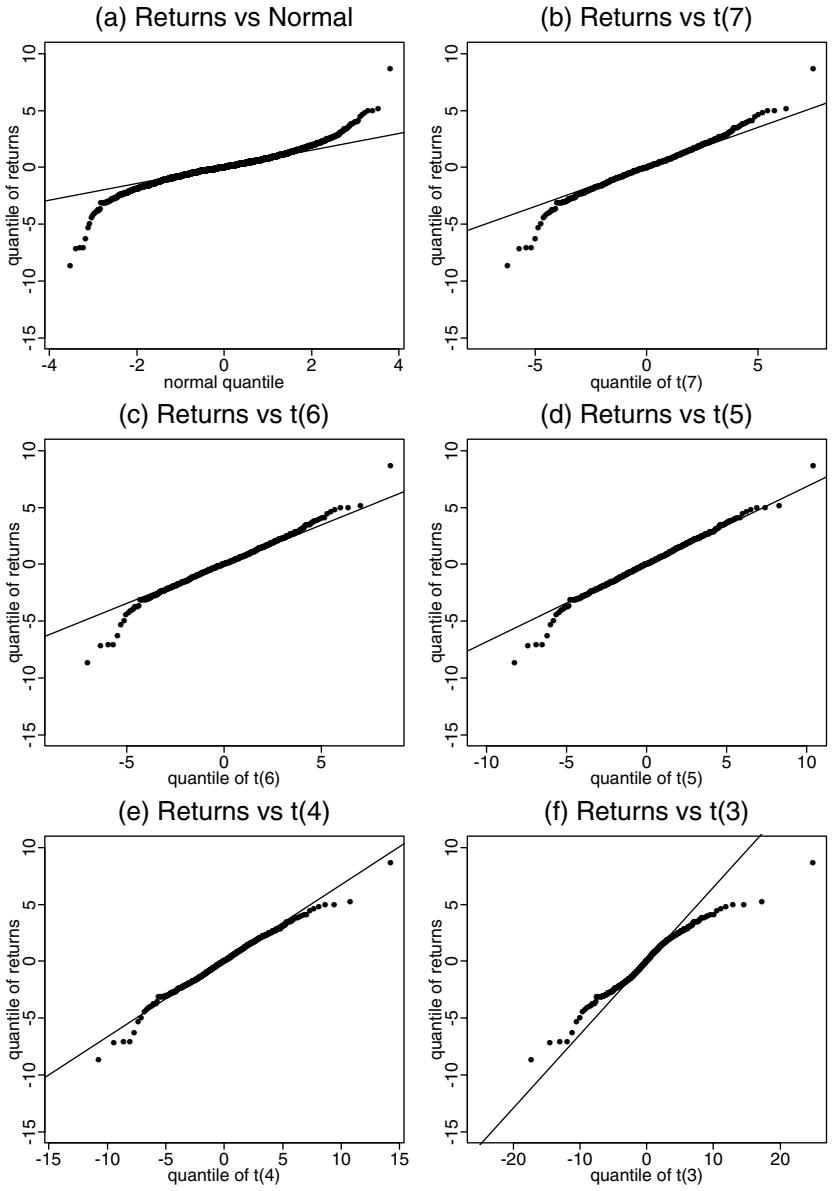


FIGURE 4.15. Sample quantiles of the S&P 500 returns are plotted against, respectively, quantiles of $N(0,1)$ and t-distributions with degrees of freedom between 7 and 3.

mal distribution, and t-distributions with 7 and 6 degrees of freedom are considerably lighter than t-distribution with 3 degrees of freedom. Therefore, it seems reasonable to assume that $E(X_t^6) = \infty$ and $E(|X_t|^{3-\epsilon}) < \infty$ for any $\epsilon > 0$.

(b) *Testing for conditional heteroscedasticity*

Following the argument in J.P. Morgan's *RiskMetrics* (J.P. Morgan 1996, p. 92), we fixed the conditional mean for the daily returns at 0. This leads to the fitting of the data with a GARCH specification (4.48). Figure 4.14(b) indicates a clear autocorrelation in the series $\{X_t^2\}$. To reinforce this observation, we applied the likelihood ratio test (4.50) to test for the existence of the ARCH-effect (i.e., conditional heteroscedasticity). The test has been implemented in the S+GARCH code "archtest.s", which simultaneously carries out the tests with normal ε_t and t_k-distributed ε_t, respectively, for $3 \le k \le 8$. For the S&P 500 returns, the null hypothesis (4.49) was always rejected with p-value virtually 0 for all of the assumed error distributions with order p between 1 and 4. As we pointed out in §4.2.6, the choice of the order p is not important. Since we tend to reject a null hypothesis when the sample size is extremely large, we also applied the test for different sections of original series with length varying between 200 and 1,000. The evidence for rejecting the homoscedasticity hypothesis (4.49) was still overwhelming.

(c) *Fitting a GARCH model with Gaussian error*

To model the conditional heteroscedasticity, we fitted a GARCH(p, q) model with Gaussian error [i.e., $\{\varepsilon_t\} \sim_{\text{i.i.d.}} N(0, 1)$] based on the conditional maximum likelihood method presented in §4.2.3. Among the candidate models with $p \ge 1$ and $q \ge 0$, both AIC (4.39) and BIC (4.40) selected a GARCH(1, 3) model with the estimated conditional standard deviation

$$\widehat{\sigma}_t^2 = 0.015 + 0.112X_{t-1}^2 + 0.492\sigma_{t-1}^2 - 0.034\sigma_{t-2}^2 + 0.420\sigma_{t-3}^2. \quad (4.56)$$

For this selected model, AIC= 17792.2 and BIC= 17826.5. The standard errors of the five estimated coefficients on the right-hand side of the equations above are 0.002, 0.004, 0.070, 0.083, and 0.055, respectively, and were calculated based on the asymptotic normal distribution of the estimator; see Theorem 4.5(i). The coefficient -0.034 in the model above is not significant since the corresponding p-value of the t-test is 0.341. Therefore, the term containing σ_{t-2}^2 may be removed from the model.

Figure 4.16(b) plots the estimated standard deviations $\widehat{\sigma}_t^2$ given in (4.56). Compared with Figure 4.16(a), (4.56) models the volatility in the original return series very well. Figure 4.16(c) shows that the "residuals" $\widehat{\varepsilon}_t$, defined as $\widehat{\varepsilon}_t \equiv X_t/\widehat{\sigma}_t$, are not necessarily always smaller than the original data X_t, but they certainly look much less volatile. Indeed, apart from a few large downward sparks, the variation of the residuals seems fairly homogeneous.

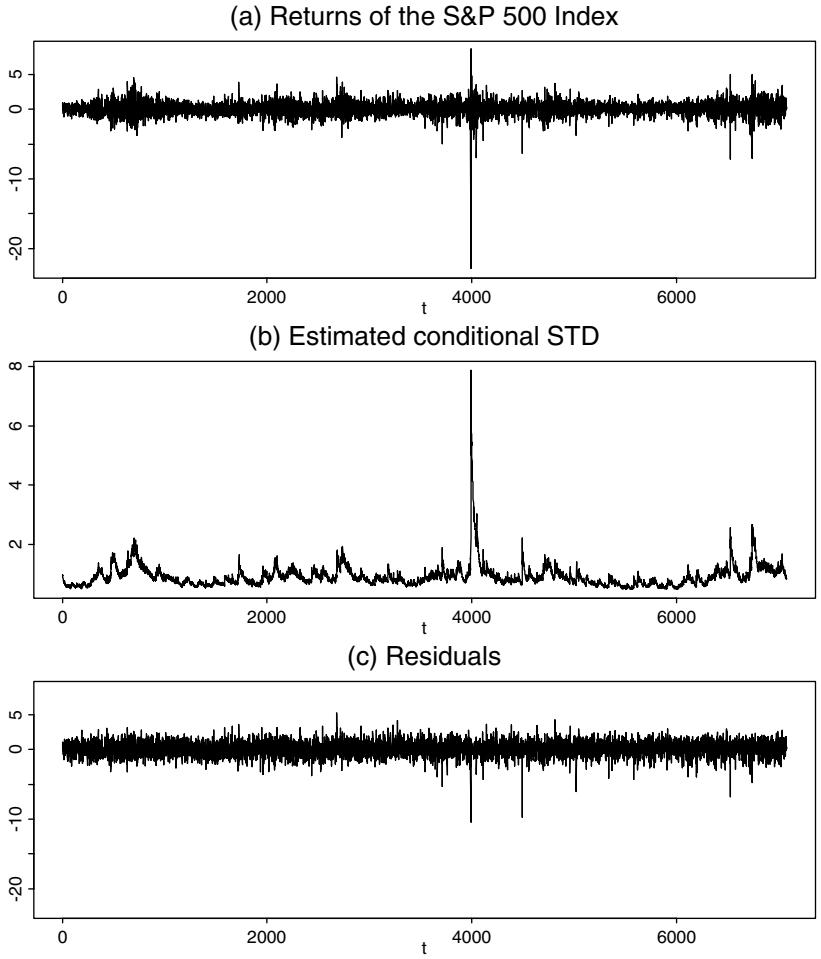


FIGURE 4.16. Time plots of (a) the S&P 500 returns $\{X_t\}$, (b) the estimated conditional standard deviations $\{\widehat{\sigma}_t\}$ given in (4.56), and (c) the residual $\{\widehat{\varepsilon}_t = X_t/\widehat{\sigma}_t\}$. The estimates were derived based on a GARCH(1, 3) model with $\varepsilon_t \sim N(0, 1)$.

Figures 4.17 (a) and (b) indicate that there seems to be no significant autocorrelation in both the residual sequence and its squared sequence. In fact, the residuals pass the likelihood ratio test (4.50) with normal conditional density for the null hypothesis (4.49) comfortably for all attempted values of p (i.e., $1 \leq p \leq 4$), indicating that there exists no significant conditional heteroscedasticity in the residual series. We also applied both Fisher's test (7.33) and the adaptive Neyman test (7.43) for testing the hypothesis that the residuals are from a white noise process. Fisher's test was passed with the p-value 0.364, whereas the adaptive Neyman test was failed with the p-

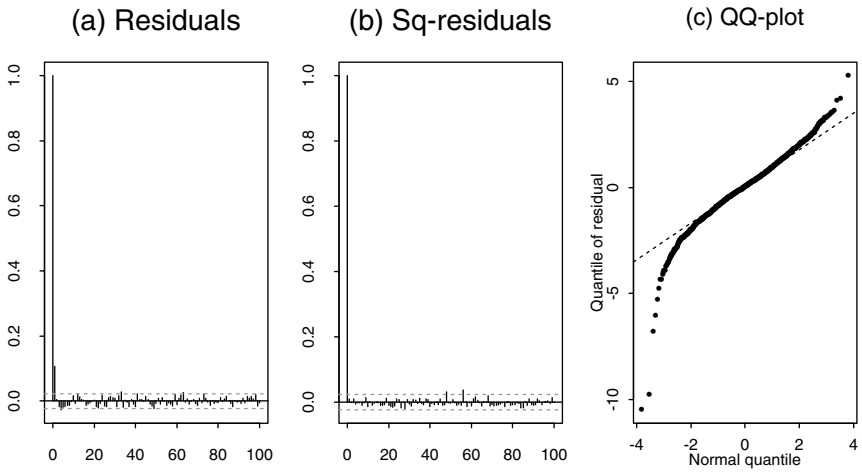


FIGURE 4.17. Correlogram of (a) the residuals from fitted model (4.56) and (b) the squared residuals. (c) plot of the residuals versus a normal distribution.

value virtually equal to 0. Note that if the fitting is perfectly adequate, the residuals should behave like Gaussian white noise. However, Figure 4.17(c) clearly indicates that both tails of the empirical distribution of the residuals are heavier than those of a normal distribution. This suggests that we may explore the possibility of fitting a GARCH model with heavy-tailed ε_t to this data set.

(d) *Fitting a GARCH model with t-distributed error*

Based on the analysis in (c) above, we fitted a GARCH model with $\{\varepsilon_t\} \sim_{\text{i.i.d.}} t_d$ with degree of freedom d, together with other parameters in the model, estimated by the (conditional) maximum likelihood method. Now, both AIC and BIC selected a GARCH(1, 1) model with estimated conditional standard deviation

$$\widehat{\sigma}_t^2 = 0.007 + 0.047X_{t-1}^2 + 0.945\sigma_{t-1}^2. \tag{4.57}$$

For this selected model, AIC= 17411.7 and BIC= 17439.2. The estimated degrees of freedom is $\widehat{d} = 7.41$ with the standard error 0.487. By treating d as a continuous parameter, the estimator is asymptotically normal. The standard errors for the three parameters on the right-hand side of (4.57) are 0.001, 0.005, and 0.005. All the three coefficients are significantly away from zero according to the t-tests. Figure 4.18(b) shows that the conditional standard deviation $\widehat{\sigma}_t$ catches the heteroscedasticity in the original data series, plotted again in Figure 4.18(a), very well. The residuals $\widehat{\varepsilon}_t \equiv X_t/\widehat{\sigma}_t$ depicted in Figure 4.18(c) are obviously less volatile than the original returns. We applied both Fisher's test (7.33) and the adaptive Neyman test

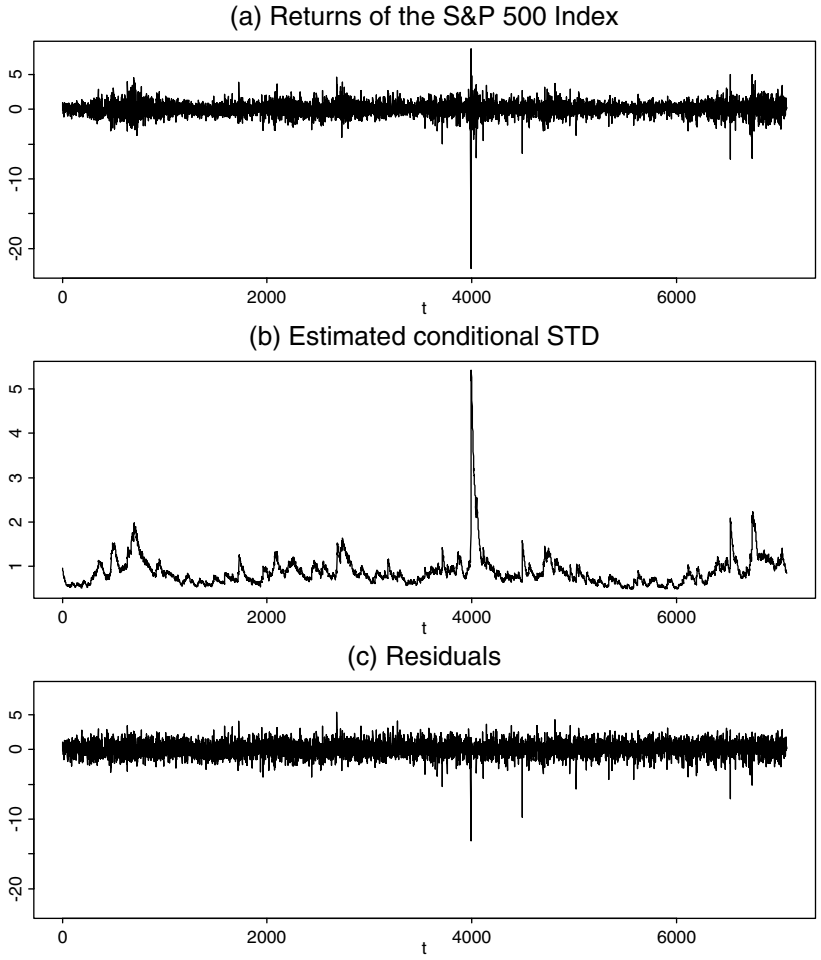


FIGURE 4.18. Time plots of (a) the S&P 500 returns $\{X_t\}$, (b) the estimated conditional standard deviations $\{\widehat{\sigma}_t\}$ given in (4.57), and (c) the residual $\{\widehat{\varepsilon}_t = X_t/\widehat{\sigma}_t\}$. The estimates were derived based on a GARCH(1, 1) model with t-distributed ε_t with the estimated degrees of freedom 7.41.

(7.43) for testing the hypothesis that the residuals are from white noise. Fisher's test was passed with the p-value 0.346, whereas the adaptive Neyman test was failed with the p-value virtually equal to 0. Figures 4.19 (a) and (b) indicate that there is hardly any significant autocorrelation in both the residual series and its squared series. Now, the residuals seem more agreeable to the distribution specified by the model, although the left-tail of its empirical distribution is still heavier than that of the t-distribution with 7.41 degrees of freedom; see Figure 4.19(c).

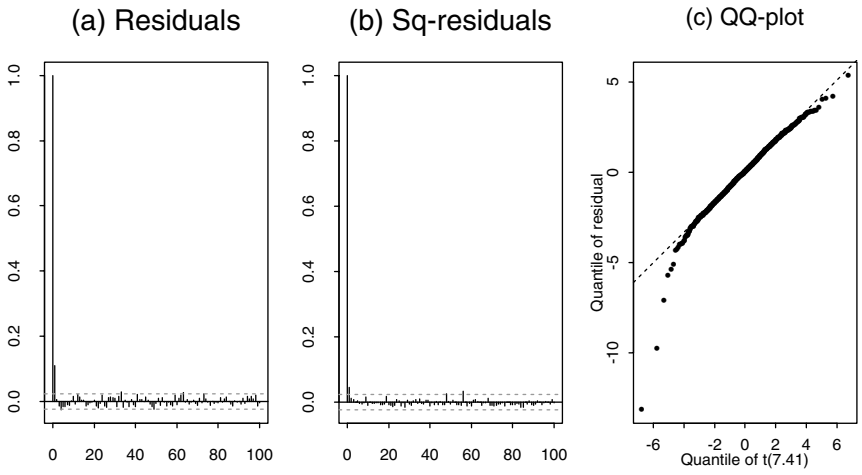


FIGURE 4.19. Correlogram of (a) residuals from fitted model (4.57) and (b) the squared residuals. (c) $q^N - q$ plot of the residuals versus the t-distribution with d.f.=7.41.

(e) *Estimation of* VaR – *value at risk*

As we pointed out earlier, one of direct motivation for ARCH/GARCH modeling is to estimate the predictive distribution for X_t given its lagged values $\{X_{t-k}, k \geq 1\}$. If we assume that X_t follows a specific GARCH model with a known $f(\cdot)$ as the density function of error ε_t, the required predictive density function is just $\sigma_t^{-1} f(\cdot/\sigma_t)$. Then, the task boils down to the estimation of the parameters in the function σ_t^2. Often, in financial applications, we are interested in extreme quantiles of this distribution, which are called *value at risk* (VaR). For $\alpha \in (0,1)$, the $100\alpha\%$ (conditional) quantile is defined as

$$x_\alpha = \inf\{x : \ P(X_t \leq x | X_{t-k}, \, k \geq 1) \geq \alpha\}. \tag{4.58}$$

(See Example 8.14 for further discussion.) An extreme high quantile x_α with α very close to 1 (or an extreme low quantile x_α with α very close to 0) represents the potential loss at the probability α (or $1-\alpha$). The VaR is arguably the most frequently used measure for risk management in finance. For a GARCH process, an estimator for x_α can be easily constructed as

$$\widehat{x}_\alpha = \widehat{\sigma}_t \, x_{\alpha,0}, \tag{4.59}$$

where $x_{\alpha,0}$ is the $100\alpha\%$ quantile of ε_t, namely

$$\int_{x_{\alpha,0}}^{\infty} f(x)dx = \alpha.$$

Obviously, a misspecification of the distribution of ε_t may lead to a considerable error in the estimator (4.59).

For the S&P 500 return data, the fitted GARCH(1, 3) model (4.56) with Gaussian error is inappropriate for a VaR estimation since it seriously misspecifies the tail behavior at both ends of the error distribution; see Figure 4.17(c). The GARCH(1, 1) model (4.57) with t-distributed error would be better for estimating x_α when α is close to 1 since the t-distribution with 7.41 degrees of freedom models the right tail of the error fairly well; see Figure 4.19(c). For $\alpha = 0.95, 0.99, 0.995$, and 0.999, $x_{\alpha,0} = 1.879, 2.952, 3.434$, and 4.656, respectively, for the t-distribution with 7.41 degrees of freedom. Due to the symmetry, $-1.879, -2.952, -3.434$, and -4.656 are the low quantiles with $\alpha = 0.05, 0.01, 0.005$, and 0.001, respectively. However, the method (4.59) may underestimate the low quantiles of the returns, as the left tail of the error distribution may be heavier than that of the t-distribution; see Figure 4.19(c) again. This means that although a carefully selected GARCH model may well catch the conditional heteroscedasticity behavior of the S&P 500 returns, and may well forecast the risk associated with high quantiles, it will unfortunately underforecast the loss due to market crashes of the scale similar to that in October 1987. An EGARCH model will accommodate the asymmetric tail-behavior into the model, but it still cannot forecast the extremely large losses in financial markets, which remains a gigantic challenge to all time series modelers.

For further discussion on the VaR estimation in terms of nonparametric methods, including the methods prescribed by the *RiskMetrics Technical Document* of J.P. Morgan (1996), see Example 8.14.

4.2.9 Stochastic Volatility Models

In this subsection, we give a brief account of stochastic volatility models. This class of models is not within the ARCH/GARCH family, but it offers an alternative for modeling conditional heteroscedasticity of financial returns. An excellent survey on stochastic volatility models is available in Shephard (1996).

A general form of stochastic volatility model may be written as

$$X_t = \varepsilon_t g(h_t) \quad \text{and} \quad h_t = a_0 + a_1 h_{t-1} + e_t,$$

where $\{\varepsilon_t\} \sim \text{IID}(0, 1)$, $\{e_t\} \sim \text{IID}(0, \sigma_e^2)$, $\{\varepsilon_t\}$ and $\{e_t\}$ are independent, and $g(\cdot) > 0$ is a known function. In contrast to a GARCH model, the heteroscedastic variation of X_t is expressed in terms of function $g(h_t)^2$, which depends on a (unobservable) latent process $\{h_t\}$ instead of a lagged value of X_t. The basic idea is that the latent h_t may represent the random and uneven flow of new information that is too complex to be modeled as a function of the lagged values $X_{t-1}, X_{t-2}, \cdots$ only. (Unfortunately, this statement itself is often true in the real world!) It is easy to see that when

$|a_1| < 1$, the latent process $\{h_t\}$ is strictly stationary with

$$\mu_h \equiv E(h_t) = \frac{a_0}{1-a_1} \quad \text{and} \quad \gamma_h(k) \equiv \text{Cov}(h_t, h_{t-k}) = \frac{\sigma_e^2}{1-a_1^2} a_1^{|k|}. \quad (4.60)$$

This, in turn, also ensures that $\{X_t\}$ is a strict stationarity process.

The most popular form of stochastic volatility model is due to Taylor (1986), which specifies

$$X_t = \varepsilon_t \exp(h_t/2), \quad h_t = a_0 + a_1 h_{t-1} + e_t, \quad (4.61)$$

where both $\{\varepsilon_t\}$ and $\{e_t\}$ are Gaussian white noise. Now $\{h_t\}$ is a Gaussian AR(1) process. Therefore, we may derive that, for all $k \geq 1$,

$$E(X_t^{2k}) = E(\varepsilon_t^{2k}) E\{\exp(kh_t)\} = \frac{(2k)! \exp\{k\mu_h + k^2\sigma_h^2/2\}}{2^k k!},$$

where μ_h and $\sigma_h^2 = \gamma_h(0)$ are given in (4.60). Consequently, the kurtosis of X_t fulfills the inequality

$$\kappa_x = E(X_t^4)/\{E(X_t^2)\}^2 = 3\exp\{\sigma_h^2\} > \kappa_\varepsilon = 3.$$

This indicates that the distribution of X_t has heavier tails than that of ε_t —a property also shared by ARCH/GARCH processes.

It follows from (4.61) that

$$\log X_t^2 = h_t + \log \varepsilon_t^2. \quad (4.62)$$

Since $\{h_t\}$ is an AR(1) process and $\{\log \varepsilon_t^2\}$ is white noise, $\{\log X_t^2\}$ is an ARMA(1, 1) process as far as its first two moment properties are concerned; see Example 2.7. Note that the causality of $\{h_t\}$ implies that $\{h_t\}$ and $\{\varepsilon_t\}$ are independent of each other. Based on the normality of $\{h_t\}$, it holds that

$$\begin{aligned} \text{Cov}(X_t^2, X_{t-k}^2) &= E\{\exp(h_t + h_{t-k})\} - E(e^{h_t}) E(e^{h_{t-k}}) \\ &= \exp(2\mu_h + \sigma_h^2)\{\exp(\sigma_h^2 a_1^{|k|}) - 1\}. \end{aligned}$$

Therefore

$$\text{Corr}(X_t^2, X_{t-k}^2) = \frac{\exp(\sigma_h^2 a_1^{|k|}) - 1}{\exp(\sigma_h^2) - 1} \approx \frac{\sigma_h^2}{\exp(\sigma_h^2) - 1} a_1^{|k|}.$$

The approximation holds for large $|k|$, which may be justified by a Taylor expansion. Note that the term on the right-hand side of the expression above is an ACF for an ARMA(1, 1) process. In this sense a stochastic volatility process behaves in a manner similar to a GARCH(1, 1) process; see (4.30).

In spite of the simple theoretical properties stated above, stochastic volatility models, unfortunately, do not facilitate a straightforward statistical estimation and inferences. The main difficulty is that, unlike ARCH and GARCH models, it is not immediately clear how to evaluate the likelihood, as the conditional distribution of X_t given its lagged values is specified implicitly only, due to the presence of latent variable h_t. Simple estimators for parameters may be derived in terms of generalized method-of-moments (Hamilton, 1989, Chapter 14). In order to estimate the latent process $\{h_t\}$, which is necessary for modeling conditional heteroscedasticity, a Kalman filter based on a linear state-space representation for the non-Gaussian process $\{\log X_t^2\}$, given in (4.62), may be employed; see Melino and Turnbull (1990) and Harvey, Ruiz, and Shephard (1994). Some approximate likelihood methods coupled with Markov chain Monte Carlo methods have also been developed for the estimation of stochastic volatility models; see Shephard (1996) and references within.

4.3 Bilinear Models

Perhaps the most natural way to introduce nonlinearity into a linear ARMA model is to add product terms. By restricting to products of time series variable X_{t-j} and innovation ε_{t-i}, we end with a model of the form

$$X_t = \sum_{j=1}^{p} b_j X_{t-j} + \varepsilon_t + \sum_{k=1}^{q} a_k \varepsilon_{t-k} + \sum_{j=1}^{P} \sum_{k=1}^{Q} c_{jk} X_{t-j} \varepsilon_{t-k}, \tag{4.63}$$

where $\varepsilon_t \sim \text{IID}(0, \sigma^2)$, and b_j, a_k and c_{jk} are unknown parameters. This model is called a *bilinear model* with order (p, q, P, Q). For the process $\{X_t\}$ defined by the model above, we write $\{X_t\} \sim BL(p, q, P, Q)$. Bilinear time series models were introduced by Granger and Anderson (1978a). The name of "bilinear" came from the fact that the model is linear in X_j as well as in ε_i. The appeal of this class is at least partially due to the fact that a bilinear model goes beyond a simple linear form and yet retains much of the simple structure of linear ARMA models. Indeed, we may argue that we understand the probabilistic properties and are able to carry out analytic computations for bilinear models more than any other nonlinear time series models. In terms of potential applications, bilinear models are known to be able to model occasional outbursts in time series (see Figure 4.20), which might be useful for modeling seismological data such as records for explosions and earthquakes. However, successful applications are still rare. Furthermore, the performance of statistical inference is less well-understood. The asymptotic distribution of the maximum likelihood estimators is still unknown. Invertibility is essential for understanding the asymptotic properties of the estimators and yet it is almost uncheckable.

4.3.1 A Simple Example

Although we are able to explore a fair amount of the analytical properties of bilinear models, the calculation is typically clouded with cumbersome notation. To illustrate some basic ideas and methods associated with bilinear models, we start with a simple BL(1, 0, 1, 1) model

$$X_t = bX_{t-1} + \varepsilon_t + cX_{t-1}\varepsilon_{t-1}, \tag{4.64}$$

where $\{\varepsilon_t\} \sim \text{IID}(0, \sigma^2)$. Note that the model is not in the form of the general autoregressive model (2.7). Therefore, we cannot use Theorem 2.4 to deduce the conditions for existence of a strictly stationary solution. However, since the model is so close to a linear form, we may express it in a kind of "moving average" with infinite order as we do for linear AR or ARMA processes; see (2.5). Indeed, by iterating (4.64) n times, we have

$$X_t = \left\{\prod_{k=1}^{n}(b + c\varepsilon_{t-k})\right\} X_{t-n} + \varepsilon_t + \sum_{j=1}^{n-1}\left\{\prod_{k=1}^{j}(b + c\varepsilon_{t-k})\right\}\varepsilon_{t-j}. \tag{4.65}$$

If the sum on the right-hand side of the expression above converges in probability as $n \to \infty$, it must also hold that $\prod_{k=1}^{n}(b + c\varepsilon_{t-k}) \xrightarrow{P} 0$. In this case, X_t may be expressed as

$$X_t = \varepsilon_t + \sum_{j=1}^{\infty}\left\{\prod_{k=1}^{j}(b + c\varepsilon_{t-k})\right\}\varepsilon_{t-j}, \tag{4.66}$$

which is in the form of MA(∞) with "random coefficients" given in the curly brackets. Since $\{\varepsilon_t\}$ is i.i.d., $\{X_t\}$ given in (4.66) is a strictly stationary solution of model (4.65). Pham and Tran (1991) showed that under the condition that $E(\varepsilon_t^4) < \infty$, the sum on the right-hand side of (4.65) converges in mean squares if and only if $b^2 + c^2\sigma^2 < 1$. This is also the necessary and sufficient condition for model (4.65) to define a unique strictly stationary solution with $E(X_t^2) < \infty$, provided that $E(\varepsilon_t^4) < \infty$.

It follows from (4.66) that

$$\mu_x = EX_t = \sum_{j=1}^{\infty} b^{j-1}cE(\varepsilon_{t-j}^2) = \frac{\sigma^2 c}{1-b}.$$

Furthermore, the variance $\text{Var}(X_t)$ may be derived from (4.66) as well, although the expression is cumbersome and not particularly inspiring. Nevertheless, it is clear that the condition $E(\varepsilon_t^4) < \infty$ is necessary for $E(X_t^2) < \infty$. By (4.64), it holds that $E(X_t\varepsilon_t) = \sigma^2$. Centering all of the terms in (4.64), we have

$$X_t - \mu_x = b(X_{t-1} - \mu_x) + \varepsilon_t + c(X_{t-1}\varepsilon_{t-1} - \sigma^2).$$

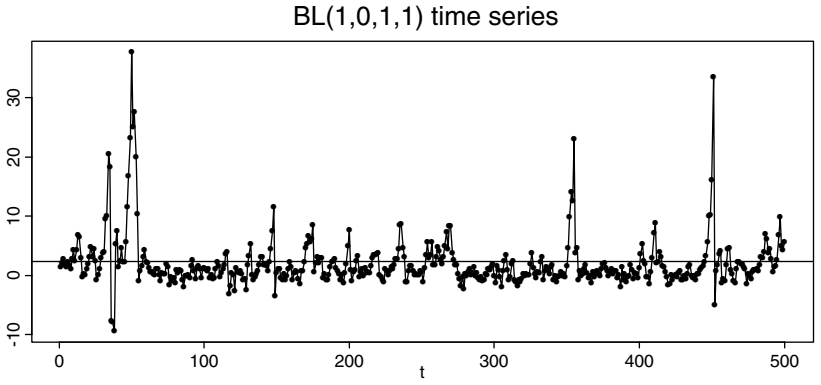


FIGURE 4.20. A time series of length 500 generated from bilinear model (4.64) with $b = 0.75, c = 0.6$, and $\varepsilon_t \sim N(0, 1)$. The horizontal line indicates the mean (2.4) of the process.

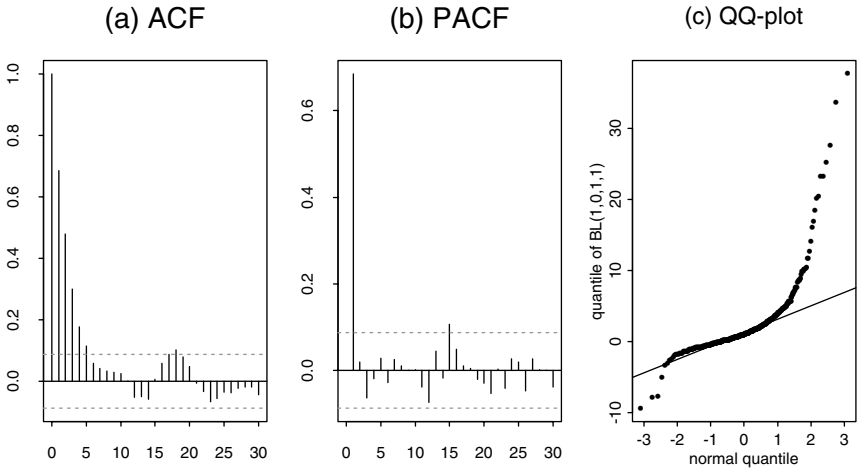


FIGURE 4.21. (a) ACF plot, (b) PACF plot, and (c) $q^N - q$ plot versus normal distribution for bilinear time series displayed in Figure 4.20.

Multiplying both sides by X_{t-k} for $k \geq 2$ and taking the expectation, we derive a Yule–Walker equation

$$\gamma(k) = b\gamma(k-1), \quad k \geq 2,$$

where $\gamma(\cdot)$ denotes the ACVF of $\{X_t\}$. Thus, we may define an ARMA(1, 1) model with the autoregressive coefficient b and the moving average coefficient and the variance of white noise selected such that the model's ACVF is the same as the ACVF of $\{X_t\}$ at both lags 0 and 1. This indicates that a BL(1, 0, 1, 1) process is effectively an ARMA(1, 1) process as far as its

first two moment properties are concerned. In fact, this is true for general bilinear models; see Part (b) of §4.3.3 below.

Figure 4.20 depicts a time series of length 500 generated from the BL(1, 0, 1, 1) model (4.64) with $b = 0.75, c = 0.6$, and standard normal innovations. The series exhibits occasional sharp spikes. Figures 4.21 (a) and (b) indicate that the sample ACF decays fairly fast and that the sample PACF is virtually only significant at lag 1, resembling the properties of an ARMA(1, 1) process. Due to those occasional bursts, the marginal distribution exhibits heavy tails; see the $q^N - q$ plot versus normal distribution in Figure 4.21(c). In general, a bilinear process does not necessarily have all moments finite.

4.3.2 Markovian Representation

The algebraic complication in manipulating bilinear models can be suppressed under their state-space representation in which state vectors are defined by random-coefficient autoregressive models with order 1 and are therefore Markov chains. Accordingly, the representation is also called Markovian representation. The stationarity of bilinear processes and their probabilistic properties can then be deduced from those of state vector processes.

It is instructive to consider first the *subdiagonal* models for which $c_{jk} = 0$ for all $j < k$ in (4.63). We reparameterize a subdiagonal bilinear model as

$$X_t = \sum_{j=1}^{p} b_j X_{t-j} + \varepsilon_t + \sum_{k=1}^{q} a_k \varepsilon_{t-k} + \sum_{j=0}^{P} \sum_{k=1}^{Q} c_{jk} X_{t-j-k} \varepsilon_{t-k}. \tag{4.67}$$

Now, parameters P, Q, and c_{jk} are different from those in (4.63). Let $n = \max\{p, P+q, P+Q\}$, $m = n - \max\{q, Q\}$, and $b_{p+j} = a_{q+j} = c_{P+i,Q+j} = 0$ for all $i, j \geq 1$. It has been established by Pham (1985, 1993) that X_t defined by (4.67) has *the state-space representation*

$$X_t = \mathbf{h}^\tau \mathbf{Z}_{t-1} + \varepsilon_t, \tag{4.68}$$

and

$$\mathbf{Z}_t = (\mathbf{A} + \mathbf{B}\varepsilon_t)\mathbf{Z}_{t-1} + \mathbf{c}\varepsilon_t + \mathbf{d}\varepsilon_t^2, \tag{4.69}$$

where the state-space variable $\mathbf{Z}_t$ is an $n \times 1$ vector with X_{t-m+i} as its ith component for $i = 1, \cdots, m$ and

$$\sum_{k=j}^{m} b_k X_{t+j-k} + \sum_{k=j}^{n-m} \left\{ a_k + \sum_{l=0}^{P} c_{lk} X_{t+j-k-l} \right\} \varepsilon_{t+j-k}$$

as its $(m + j)$th element for $j = 1, \cdots, n - m$, $\mathbf{h}$ is an $n \times 1$ vector with the $(m + 1)$-th element 1 and all others 0, $\mathbf{c}$ is an $n \times 1$ vector with the first $m - 1$ elements 0 followed by $1, b_1 + a_1, \cdots, b_{n-m} + a_{n-m}$, $\mathbf{d}$ is an

$n \times 1$ vector with the first m elements 0 followed by $c_{01}, \cdots, c_{0,n-m}$, $\mathbf{B}$ is an $n \times n$ matrix with

$$\begin{pmatrix} c_{m1} & \cdots & c_{01} \\ \vdots & \vdots & \vdots \\ c_{m,n-m} & \cdots & c_{0,n-m} \end{pmatrix}$$

as the $(n-m) \times (m+1)$ submatrix at the bottom-left corner and all of the other elements 0, and $\mathbf{A}$ is an $n \times n$ matrix with 1 as its $(i, i+1)$ element for $i = 1, \cdots, n-1$, b_j as its $(m+j, m+1)$ element for $j = 1, \cdots, n-m$, and b_{n-1+k} as its (n, k)th element for $k = 1, \cdots, m+1$ and 0 as all of the other elements.

The representation (4.68) and (4.69) can be checked by direct computation. The state-variable equation (4.69) is in the form of the AR(1) model with a random coefficient. Note that $\mathbf{Z}_t$ consists of the lagged values X_{t-1}, $X_{t-2}, \cdots$. Therefore, in (4.69) "regressor" $\mathbf{Z}_{t-1}$ is independent of both "coefficient" $(\mathbf{A} + \mathbf{B}\varepsilon_t)$ and "noise" $(\mathbf{c}\varepsilon_t + \mathbf{d}\varepsilon_t^2)$ if the bilinear process $\{X_t\}$ is *causal* in the sense that X_t is determined by $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \cdots\}$ only. Furthermore $\{\mathbf{Z}_t\}$ is a Markov chain. In fact, the Markovian representation of this nature is also available for the general bilinear model (4.63) with much more added complexity in notation; see Pham (1985, 1993). In general, for $\{X_t\}$ defined in (4.63), it holds that

$$X_t = \mathbf{h}^\tau \mathbf{Z}_{t-1} + \varepsilon_t \tag{4.70}$$

and

$$\mathbf{Z}_t = \mathbf{A}(\varepsilon_t)\mathbf{Z}_{t-1} + \mathbf{c}(\varepsilon_t), \tag{4.71}$$

where $\mathbf{Z}_t$ is an appropriately defined state-space random vector, $\mathbf{h}$ is a constant vector, and $\mathbf{A}(\cdot)$ and $\mathbf{c}(\cdot)$ are constant matrix and vector functions, respectively.

*4.3.3 Probabilistic Properties**

(a) *Stationarity*

For a bilinear process $\{X_t\}$ defined in (4.63), if the state-space process $\{\mathbf{Z}_t\}$ specified in (4.71) is strictly stationary, $\{X_t\}$ is also strictly stationary because of (4.70) and the fact that $\{\varepsilon_t\}$ is i.i.d. Therefore, we only need to derive the conditions under which the state-space equation (4.71) admits a strictly stationary solution. Actually, the form of $\mathbf{A}(\cdot)$ and $\mathbf{c}(\cdot)$ is not important. We are thus led to consider a general random coefficient model

$$\mathbf{Z}_t = \mathbf{A}_t \mathbf{Z}_{t-1} + \mathbf{c}_t, \tag{4.72}$$

where $\mathbf{A}_t$ is a random matrix and $\mathbf{c}_t$ is a random vector, and $\{(\mathbf{A}_t, \mathbf{c}_t)\}$ i.i.d. In comparison with the random-coefficient autoregressive models considered by Nicholls and Quinn (1982), $\mathbf{A}_t$ is not assumed to be independent

of $\mathbf{c}_t$ in the model above. For this reason, (4.72) is called *a generalized random coefficient autoregressive model* (Pham 1986).

Since (4.72) is in the form of AR(1), it holds that for any $n \geq 1$

$$\mathbf{Z}_t = \mathbf{A}_t \cdots \mathbf{A}_{t+1-n}\mathbf{Z}_{t-n} + \mathbf{c}_t + \sum_{j=1}^{n-1} \mathbf{A}_t \cdots \mathbf{A}_{t+1-j}\mathbf{c}_{t-j}. \tag{4.73}$$

Thus, a sufficient condition for the existence of a stationary solution of (4.73) is that the series $\sum_{j=1}^{n-1} \mathbf{A}_t \cdots \mathbf{A}_{t+1-j}\mathbf{c}_{t-j}$ converge in probability. Since $\{\mathbf{A}_t\}$ is i.i.d., it holds that

$$\frac{1}{2j} \log\{\lambda_{\max}(\mathbf{A}_{t+1-j}^{\tau} \cdots \mathbf{A}_t^{\tau}\mathbf{A}_t \cdots \mathbf{A}_{t+1-j})\} \xrightarrow{a.s.} \lambda_0 \in [-\infty, \infty]$$

as $j \to \infty$, where $\lambda_{\max}(\mathbf{A})$ denotes the maximal eigenvalue of matrix $\mathbf{A}$; the limit λ_0, which may take infinite values, is called the *upper Lyapunov exponent* of the sequence $\{\mathbf{A}_t\}$; see, for example, Cohen, Kesten, and Newman (1986). When $\lambda_0 < 0$, it holds that for some fixed $\rho > 0$ and all sufficiently large j,

$$|\lambda_{\max}(\mathbf{A}_{t+1-j}^{\tau} \cdots \mathbf{A}_t^{\tau}\mathbf{A}_t \cdots \mathbf{A}_{t+1-j})| \leq e^{-2j\rho}.$$

Consequently,

$$||\mathbf{A}_t \cdots \mathbf{A}_{t+1-j}\mathbf{c}_{t-j}|| \leq e^{-j\rho}||\mathbf{c}_{t-j}||$$

for all large j, where the matrix norm $||\cdot||$ is defined as

$$||\mathbf{A}||^2 = \sup_{\mathbf{x}\neq 0} \mathbf{x}^{\tau}\mathbf{A}^{\tau}\mathbf{A}\mathbf{x}/\mathbf{x}^{\tau}\mathbf{x},$$

which reduces to the conventional Euclidean norm when $\mathbf{A}$ is a vector. Note that

$$\left\| \sum_{j=1}^{n-1} \mathbf{A}_t \cdots \mathbf{A}_{t+1-j}\mathbf{c}_{t-j} \right\| \leq \sum_{j=1}^{n-1} ||\mathbf{A}_t \cdots \mathbf{A}_{t+1-j}\mathbf{c}_{t-j}||$$

$$\leq \quad C \sum_{j=1}^{n-1} e^{-j\rho}||\mathbf{c}_{t-j}|| \leq C \left\{ \sum_{j=1}^{n-1} e^{-2j\rho} \right\}^{1/2} \left\{ \sum_{j=1}^{n-1} ||\mathbf{c}_{t-j}||^2 \right\}^{1/2},$$

where $C > 0$ is a constant. Hence (4.73) has the unique strictly stationary solution

$$X_t = \mathbf{c}_t + \sum_{j=1}^{\infty} \mathbf{A}_t \cdots \mathbf{A}_{t+1-j}\mathbf{c}_{t-j}, \tag{4.74}$$

provided

(i) the upper Lyapunov exponent of the sequence $\{\mathbf{A}_t\}$ is negative, and

(ii) $E\{||\mathbf{c}_t||^2\} < \infty$.

The infinite sum on the right-hand side of (4.74) converges in mean square. This result was first obtained by Pham (1986) and Brandt (1986). In fact, under some mild additional conditions on the underlying distribution, condition (i) above is also necessary for the existence of a stationary solution; see Theorem 2.5 of Bougerol and Picard (1992a). Theorem 2.1 of Pham (1993) presented the necessary and sufficient condition for the existence of a causal, strictly stationary solution $\{\mathbf{Z}_t\}$ of (4.71) for which $E\{||\mathbf{Z}_t||^2\} < \infty$. This requires, among other things, the condition $E(\varepsilon_t^4) < \infty$.

(b) *Moment properties*

Suppose that $\{X_t\}$ is a strictly stationary solution of the BL(p, q, P, Q) model (4.63) that admits the state-space representation (4.70) and (4.71) where the state-space process $\{\mathbf{Z}_t\}$ can be written as

$$\mathbf{Z}_t = \mathbf{c}(\varepsilon_t) + \sum_{j=1}^{\infty} \mathbf{A}(\varepsilon_t) \cdots \mathbf{A}(\varepsilon_{t+1-j})\mathbf{c}(\varepsilon_{t-j});$$

see (4.74). The equation above indicates that $\mathbf{Z}_t$ is causal, as it depends on $\{\varepsilon_{t-k}, k \geq 0\}$ only. Therefore, in the AR model (4.71), the "regressor" $\mathbf{Z}_{t-1}$ is independent of both the "coefficient" $\mathbf{A}(\varepsilon_t)$ and the "noise" $\mathbf{c}(\varepsilon_t)$. The moments of $\mathbf{Z}_t$ can be evaluated based on (4.71). Based on the moments $\mathbf{Z}_t$, the moments of X_t can be easily obtained through (4.70). Note that X_t is also causal in the sense that X_t is a function of $\{\varepsilon_t, \varepsilon_{t-1}, \cdots\}$ only.

An important feature of the bilinear model is that not all moments exist. It is easy to see from (4.69) and (4.68) that the condition $E(\varepsilon_t^4) < \infty$ is necessary for a subdiagonal (and also general) bilinear linear process having finite second moment. The required conditions become more stringent when the order of moment increases. See Pham (1993) for further discussions on those conditions.

Now, let $E\{X_t^2\} < \infty$. Then, as far as only the first two moment properties are concerned, $\{X_t\}$ is in fact an ARMA(p, q') process with the same b_js as in (4.63) as its autoregressive coefficients, where $q' = \max\{q, Q\}$. To this end, define

$$\begin{aligned} Y_t &\equiv X_t - \sum_{j=1}^{p} b_j X_{t-j} - \mu \\ &= \varepsilon_t + \sum_{k=1}^{q} a_k \varepsilon_{t-k} + \sum_{j=1}^{P}\sum_{k=1}^{Q} c_{jk} X_{t-j}\varepsilon_{t-k} - \mu, \end{aligned} \tag{4.75}$$

where $\mu = (1 - \sum_{j=1}^{p} b_j)EX_t$. The last equality in the expression above follows from (4.63). Then $\{Y_t\}$ is stationary with $EY_t = 0$ and

$$\text{Cov}(Y_t, Y_{t-k}) = 0 \qquad \text{for all } k > q' = \max\{q, Q\}. \tag{4.76}$$

For each t, let

$$\widehat{Y}_t = \sum_{j=1}^{\infty} \phi_j Y_{t-j} \tag{4.77}$$

be the best linear predictor for Y_t based on $Y_{t-1}, Y_{t-2}, \cdots$ in the sense that

$$E(Y_t - \widehat{Y}_t)^2 = \min E \left\{ Y_t - \sum_{j=1}^{\infty} \psi_j Y_{t-j} \right\}^2,$$

where the minimum is taken over all $\{\psi_j\}$ for which the infinite sum on the right-hand side of the expression above converges in mean square. Write $e_t = Y_t - \widehat{Y}_t$. Then, the least square property above implies that $\{e_t\}$ is a sequence of uncorrelated random variables; that is, $\{e_t\} \sim \text{WN}(0, \sigma_e^2)$. Similar to (3.8), it holds that for each t

$$\widehat{Y}_t = \sum_{i=1}^{\infty} \theta_i e_{t-i}.$$

Consequently,

$$Y_t = e_t + \widehat{Y}_t = e_t + \sum_{i=1}^{\infty} \theta_i e_{t-i}.$$

It is easy to see from (4.75) and (4.76) that $\text{Cov}(Y_t, e_{t-k}) = 0$ for any $k > q'$. Hence

$$\theta_k = \text{Cov}(Y_t, e_{t-k})/\sigma_e^2 = 0 \quad \text{for all } k > q'.$$

Therefore $\{Y_t\}$ is an MA(q') process. By (4.75), it holds now that

$$X_t - \sum_{j=1}^{p} b_j X_{t-j} - \mu = e_t + \sum_{i=1}^{q'} a_i e_{t-i} \tag{4.78}$$

(i.e., $\{X_t\}$ is an ARMA(p, q') process).

(c) *Mixing*

The mixing properties of bilinear processes may be established in terms of their Markovian representation (4.70) and (4.71). Since $\{\varepsilon_t\}$ is i.i.d., the bilinear process $\{X_t\}$ shares the mixing properties possessed by the Markov chain $\{\mathbf{Z}_t\}$. Therefore, the ergodicity of $\{\mathbf{Z}_t\}$ ensures that $\{X_t\}$

is β-mixing; see (2.58). Furthermore, the geometric ergodicity implies the β-mixing with exponentially decaying mixing-coefficients; see (2.59).

There exists a fairly large literature dealing with ergodicity of Markov chains; see, for example, Nummelin and Tuominen (1982) and Tweedie (1983). Unfortunately conditions for the ergodicity are not always easy to check in practice. But those conditions are typically very mild. Therefore, we may hope that they will be satisfied in most practical situations.

4.3.4 Maximum Likelihood Estimation

Fitting a bilinear model consists of at least two aspects: determination of the order (p, q, P, Q) and estimation of the parameters b_j, a_k, c_{jk}, and σ^2. The order determination is usually carried out in terms of some well-known model selection criteria such as AIC and BIC. However, the performance of those procedures in the context of bilinear models is not well-understood. This is due to the lack of asymptotic theory for maximum likelihood estimation for bilinear models.

However, if the order (p, q, P, Q) is given, the standard method for approximating a Gaussian likelihood function for time series may be applied to derive approximate maximum likelihood estimators. Let $X_1, \cdots, X_T$ be observations from a strictly stationary BL(p, q, P, Q) process defined by (4.63) in which $\varepsilon_t \sim N(0, \sigma^2)$. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\tau, \boldsymbol{\theta}_2^\tau)^\tau$, where

$$\boldsymbol{\theta}_1 = (b_1, \cdots, b_p, a_1, \cdots, a_q)^\tau, \quad \boldsymbol{\theta}_2 = (c_{11}, \cdots, c_{1Q}, c_{21} \cdots, c_{PQ})^\tau.$$

The (conditional) log likelihood function may be approximated by

$$l(\boldsymbol{\theta}, \sigma^2) = -\frac{N - p'}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=p'+1}^{T} \widehat{\varepsilon}_t(\boldsymbol{\theta})^2,$$

where $p' = \max\{p, P\}$, and $\widehat{\varepsilon}_{p'}(\boldsymbol{\theta}), \widehat{\varepsilon}_{p'+1}(\boldsymbol{\theta}), \cdots$ are computed recursively from model (4.63) with some (arbitrarily) specified initial values for $\varepsilon_{p'-1}$, $\cdots$, $\varepsilon_{p'-q'}$, and $q' = \max\{q, Q\}$.

4.3.5 Bispectrum

Suppose that $\{X_t\}$ is a stationary process with mean 0. Furthermore, we assume that its third moments are also time-invariable in the sense that

$$C(j, k) \equiv E(X_t X_{t+j} X_{t+k})$$

is independent of t for any j and k. Such a process may be referred to as third-order stationary.

The *bispectral density function* of $\{X_t\}$ is defined as

$$g(\omega_1, \omega_2) = \frac{1}{4\pi^2} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} C(j, k) \exp\{-i(j\omega_1 + k\omega_2)\}, \quad \omega_1, \omega_2 \in [-\pi, \pi].$$

It is easy to see that g is well-defined if

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}|C(j,k)|<\infty.$$

On the other hand, it holds by inversion that

$$C(j,k)=\int_{[-\pi,\pi]^2}\exp\{i(j\omega_1+k\omega_2)\}g(\omega_1,\omega_2)d\omega_1 d\omega_2.$$

Hence,

$$E(X_t^3)=C(0,0)=\int_{[-\pi,\pi]^2}g(\omega_1,\omega_2)d\omega_1 d\omega_2.$$

Note that a spectral density depends only on the second moments of the process; see §2.3.2. Likewise, the bispectral density defined above depends on the third moments of $\{X_t\}$ only. Based on the fact that a stationary BL(p,q,P,Q) process is also a stationary ARMA$(p,\max\{q,Q\})$ as far as the covariance structure is concerned, its nonlinearity will only show up in its bispectral density but not in its spectral density (see (4.78)). Although in principle for bilinear processes bispectral density functions may be evaluated explicitly, the derivation is typically tedious and the formulas always appear cumbersome; see §2.6 of Subba Rao and Gabr (1984) for an example with simple BL(1, 0, 1, 1) processes.

For two third-order stationary processes $\{X_t\}$ and $\{Y_t\}$, if $\{X_t\}$ is a filtered version of $\{Y_t\}$, namely

$$X_t=\sum_{k=-\infty}^{\infty}\varphi_k Y_{t-k},\quad \sum_{k=-\infty}^{\infty}|\varphi_k|<\infty,$$

we may show, in the same manner as the proof of Theorem 2.12, that

$$g_x(\omega_1,\omega_2)=g_y(\omega_1,\omega_2)\varphi(e^{-i\omega_1})\varphi(e^{-i\omega_2})\varphi(e^{i(\omega_1+\omega_2)}),\tag{4.79}$$

where g_x, g_y denote, respectively, the bispectral densities of $\{X_t\}$ and $\{Y_t\}$, and $\varphi(z)=\sum_k\varphi_k z^k$.

Suppose now that $\{X_t\}$ is a purely nondeterministic and zero-mean stationary process. The Wold decomposition entails

$$X_t=\varepsilon_t+\sum_{j=1}^{\infty}\varphi_k\varepsilon_{t-k},\quad \{\varepsilon_t\}\sim \mathrm{WN}(0,\sigma^2).\tag{4.80}$$

The process $\{X_t\}$ is called *linear* if $\{\varepsilon_t\}\sim\mathrm{IID}(0,\sigma^2)$. For example, a stationary bilinear process entertains the expression (4.80), but $\{\varepsilon_t\}$ is not an i.i.d. sequence; see (4.78). The spectral density of $\{X_t\}$ in (4.80) is equal to

$$g(\omega)=\frac{\sigma^2}{2\pi}\varphi(e^{-i\omega})\varphi(e^{i\omega}).\tag{4.81}$$

When $\{X_t\}$ is linear (i.e., $\{\varepsilon_t\} \sim \text{IID}(0, \sigma^2)$), it follows from (4.79) that the bispectral density of $\{X_t\}$ is equal to

$$g(\omega_1, \omega_2) = \frac{\mu_3}{4\pi^2}\varphi(e^{-i\omega_1})\varphi(e^{-i\omega_2})\varphi(e^{i(\omega_1+\omega_2)}), \tag{4.82}$$

where $\mu_3 = E(\varepsilon_t^3)$. Combining this with (4.81), we have

$$K(\omega_1, \omega_2) \equiv \frac{2\pi\,|g(\omega_1, \omega_2)|^2}{g(\omega_1)g(\omega_2)g(\omega_1 + \omega_2)} = \mu_3^2/\sigma^6.$$

Thus, we may test for the linearity in terms of a test for the hypothesis that the function $K(\cdot, \cdot)$ is a constant; see §4.4 of Subba Rao and Gabr (1984), Hinich(1982), and Subba Rao (1983) for the development of the tests for linearity based on this idea.

4.4 Additional Bibliographical notes

The developments on threshold models up to the late 1980s were systematically presented in Tong (1990), which also dealt with other parametric nonlinear models not covered in this book. Tsay (1989) proposed an alternative strategy for TAR modeling that selected the delay parameter, number of regimes, and thresholds based on some F- and t-statistics. Threshold models with continuous regression functions were studied in Chan and Tsay (1998). Stramer, Tweedie, and Brockwell (1996) dealt with continuous-time threshold models. Double threshold models that impose threshold structure on both conditional means and conditional variances were proposed by Li and Li (1996).

Lawrance and Lewis (1980, 1985) introduced a class of exponential ARMA models in which coefficients change according to a sequence of independent random variables. Autoregressive models with regime-switch controlled by a Markov chain mechanism were suggested in Tong and Lim (1980, p.285 line −12), and studied by Tyssedal and Tjøstheim (1988) and Hamilton (1989).

Asymptotic properties of maximum likelihood estimators for stationary Markov chains can be found in Billingsley (1961), Basawa and Prakasa Rao (1980), and Hall and Heyde (1980).

Strict stationarity was first established for GARCH(1, 1) models by Nelson (1991), and for GARCH(p, q) processes by Bougerol and Picard (1992b). The conditions for the existence of a strictly stationary GARCH process that is also (weakly) stationary is much simpler; see Giraitis, Kokoszka and Leipus (2000). The extremal behavior of ARCH(1) processes was presented in §8.4.3 of Embrechts, Klüppelberg, and Mikosch (1997), and see also Zhang and Tong (2001).

Under the assumption that $E(\varepsilon_t^4) < \infty$ but ε_t may be non-Gaussian, the asymptotic normality for Gaussian conditional maximum likelihood estimators was established for the ARCH(p) models by Weiss (1986), and for GARCH(1,1) models by Lee and Hansen (1994) and Lumsdaine (1996). Hall and Yao (2003) established the comprehensive asymptotic theory for Gaussian conditional maximum likelihood estimators for general GARCH (p, q) models including heavy-tailed cases. On the other hand, estimation for ARCH(p) models adaptive to unknown error distributions was considered by Linton (1993). Whittle estimation for a general ARCH(∞) process was studied by Giraitis and Robinson (2001).

The most popular ARCH test in the literature is Engle's (1982) Lagrange multiplier test (LMT) TR^2; see (4.53). Lee (1991) showed that a modified LMT for GARCH(p, q) is the same as the LMT for ARCH(p). McLeod and Li (1983) applied the portmanteau tests of (7.29) due to Box and Pierce (1970) and Ljung and Box (1978) in the context of ARCH/GARCH models, which are asymptotically equivalent to the LMT (Granger and Terasvirta 1993, pp.93–94). Other ARCH tests include those of Weiss (1986), Robinson (1991b), and Bera and Higgins (1992). The literature on nonparametric tests for the ARCH effect includes Chen and An (1997) and Laïb (2002); see also Koul and Stute (1999).

The early developments on bilinear models are summarized in Subba Rao and Gabr (1984). An excellent survey on both basic properties and statistical inference for bilinear models is available in Pham (1993). Terdik (1999) is a modern account on the frequency-domain approach for bilinear models based on chaotic Wiener–Itô spectral representation.

The stationarity of bilinear processes was also studied by, among others, Hannan (1982), Liu and Brockwell (1982),Quinn (1982), Bhaskara Rao, Subba Rao and Walker (1983), Pham (1986), and Liu (1990, 1992). The stationarity for random coefficient autoregressive models was studied by Pham (1986) and Bougerol and Picard (1992a). The invertibility of bilinear models was discussed, for example, in Granger and Andersen (1978b), Subba Rao (1981), Quinn (1982), Guegan and Pham (1989), and Pham (1993).

Method-of-moments estimation for simple bilinear models can be found in Kim, Billard, and Nasawa (1990) and Liu and Chen (1991). Sesay and Subba Rao (1992) proposed a Whittle-like estimator for bilinear models. Various statistical tests for linearity were assembled in §5.3 of Tong (1990). Saikkonen and Luukkonen (1991) and Guegan and Pham (1992) studied score tests for bilinear models.

5
Nonparametric Density Estimation

5.1 Introduction

Smoothing is one of the most fundamental techniques in nonparametric function estimation. It usually refers to one-dimensional scatterplot smoothing and density estimation. It serves as a useful building block for nonparametric estimation in a multidimensional setting. Smoothing arose first from spectral density estimation in time series. In a discussion of the seminal paper by Bartlett (1946), Henry E. Daniels suggested that a possible improvement on spectral density estimation could be made by smoothed periodograms. The theory and techniques were then systematically developed by Bartlett (1948, 1950). Thus, smoothing techniques were already prominently featured in time series analysis over half a century ago.

Smoothing problems arise frequently from various aspects of time series analysis. Smoothing techniques provide useful graphic tools for summarizing the marginal distribution of a given time series. They can also be applied to estimate and a remove slowly varying time trend. This results in time domain smoothing. The need to study the associations between a time series and its lagged series leads to state domain smoothing. These techniques can easily be extended to estimate the conditional variance (volatility) of a time series. To examine cyclic patterns and other features, such as the power spectrum in a time series, smoothing techniques are frequently employed to estimate spectral density. An important question in fitting time series data is whether or not the residuals of a fitted model behave like white noise. Nonparametric function estimation provides useful tools

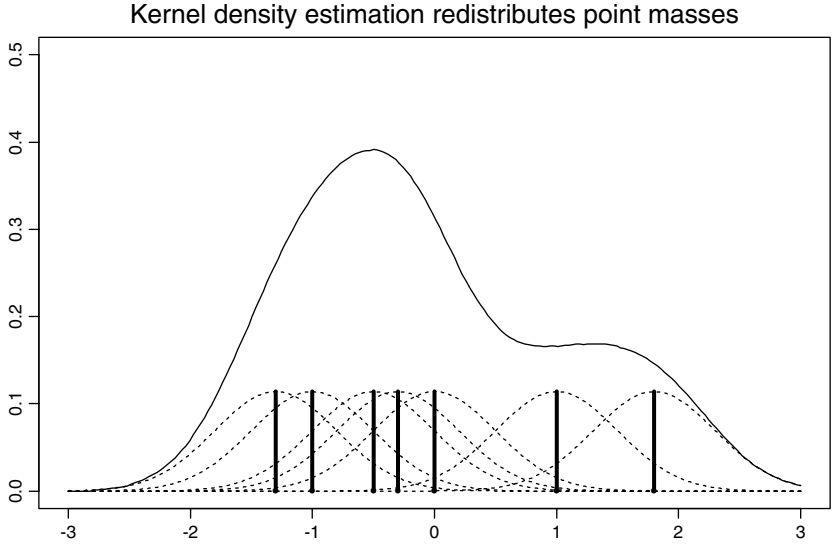


FIGURE 5.1. Kernel density estimation redistributes the point mass, depicted by a solid vertical bar, at each datum point and adds the redistributed masses together to get the final estimate.

for this kind of nonparametric goodness-of-fit test. These subjects will be discussed in this and the next two chapters.

The simplest nonparametric function estimation problem is probably the density estimation. This simple setup provides useful ingredients for our understanding of more complicated problems in nonparametric modeling and inferences. This motivates us to devote this chapter to nonparametric density estimation.

5.2 Kernel Density Estimation

What is the distribution of the yields of Treasury bills? Use of a histogram is a classical method of answering this question. An improvement of the histogram method is the *kernel density estimation*. It is used to examine the overall distribution of a data set. This includes the number and locations of peaks and troughs as well as the symmetry of a density. It is the simplest setting to reveal the basic ingredients of nonparametric function estimation. The comprehensive account of density estimation and its applications is given in Devroye and Györfi (1985), Silverman (1986), and Scott (1992).

Given T data points $X_1, \cdots, X_T$, their *empirical distribution function* is obtained by putting mass $1/T$ at each observed datum:

$$\widehat{F}(x) = \frac{1}{T} \sum_{t=1}^{T} I(X_t \leq x).$$

This cumulative distribution function is nondecreasing and is not that useful for examining the overall structure of the underlying distribution. When one refers to distributions, one often has density functions in mind. However, the density of the empirical distribution does not exist. An improvement over the empirical distribution function is to smoothly redistribute the mass $1/T$ at each datum point to its vicinity (see Figure 5.1). This is usually accomplished by introducing a *kernel function* K, which is usually a nonnegative symmetric, unimodal probability density function. Let h be a *bandwidth* parameter representing the window size in Figure 5.1 (indeed, it is the standard deviation of density functions plotted in dashed lines). Then, the kernel density estimate is defined by

$$\widehat{f}_h(x) = T^{-1} \sum_{t=1}^{T} \frac{1}{h} K\left(\frac{X_t - x}{h}\right) = \int K_h(u - x) d\widehat{F}(u), \tag{5.1}$$

where $K_h(\cdot) = K(\cdot/h)/h$.

Commonly used *kernel functions* include the Gaussian kernel

$$K(u) = (\sqrt{2\pi})^{-1} \exp(-u^2/2)$$

and the symmetric Beta family

$$K_\gamma(u) = \frac{1}{\text{Beta}(1/2, \gamma + 1)} (1 - u^2)^\gamma I(|u| \leq 1).$$

The choices $\gamma = 0, 1, 2$, and 3 correspond to the *uniform,* the *Epanechnikov,* the *biweight*, and the *triweight* kernel functions, respectively. When γ is large, by appropriate rescaling, the symmetric gamma kernel is approximately the same as the Gaussian kernel function. Note that different kernel functions have different support. For example, the uniform kernel has effective support $[-1, 1]$, while the triweight kernel has much shorter effective support (due to a smaller weight at tails) and the Gaussian kernel has much longer effective support (see Figure 5.2). Thus, even with the same bandwidth, different kernels use different amounts of information provided by the local data points around x. Formula (5.7) below attempts to relate the equivalent amount of smoothing using two different kernels. The concept of canonical kernels introduced by Marron and Nolan (1988) attenuates this problem.

To employ the kernel density estimator, one needs to choose the kernel function and the bandwidth. It is well-known both empirically and theoretically that the choice of kernel functions is not very important to the

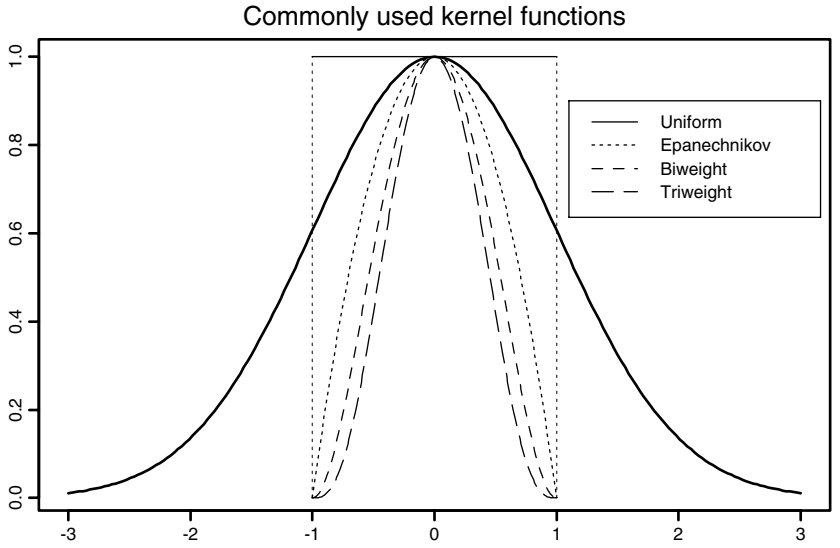


FIGURE 5.2. Commonly used kernel functions normalized to have maximum height 1 to facilitate the presentation. The thick curve is the Gaussian kernel, which has a much longer effective support than other kernels.

kernel density estimator. As long as they are symmetric and unimodal, the resulting kernel density estimator performs nearly the same when the bandwidth h is optimally chosen (see Table 5.1 in §5.4). Thus, as demonstrated in Figure 5.3, a large bandwidth h produces an oversmooth estimate, leaving out possible details such as multimodalities and underestimating the density at peaks. In other words, the estimate can create large biases when a large bandwidth is used. When a small bandwidth is applied, there are not many local data points available to reduce the variance of the estimate. This can result in a wiggly curve. Trial-and-error is needed in order to produce satisfactory results. A data-driven choice of bandwidth can assist us in determining the optimal amount of smoothing (see §5.4 for more details).

As an illustration, Figure 5.3 depicts the estimated distributions for the yields of 3-month Treasury bills using the Gaussian kernel with bandwidths $h = 0.61/3, 0.61$, and 3×0.61. The S-Plus function "density" was used to compute the kernel density estimator. The bandwidth $h = 0.61$ was determined by the normal reference bandwidth selector (5.9) below. It is clear that a small bandwidth results in an undersmoothed estimator, creating a wiggly density function with artificial modes, while a large bandwidth leads to an oversmoothed curve, obscuring fine structure of the underlying distribution. The simple reference bandwidth $h = 0.61$, which is often viewed as an initial choice for h, gives a reasonable amount of smoothing for this example, although the resulting curve appears somewhat oversmoothed.

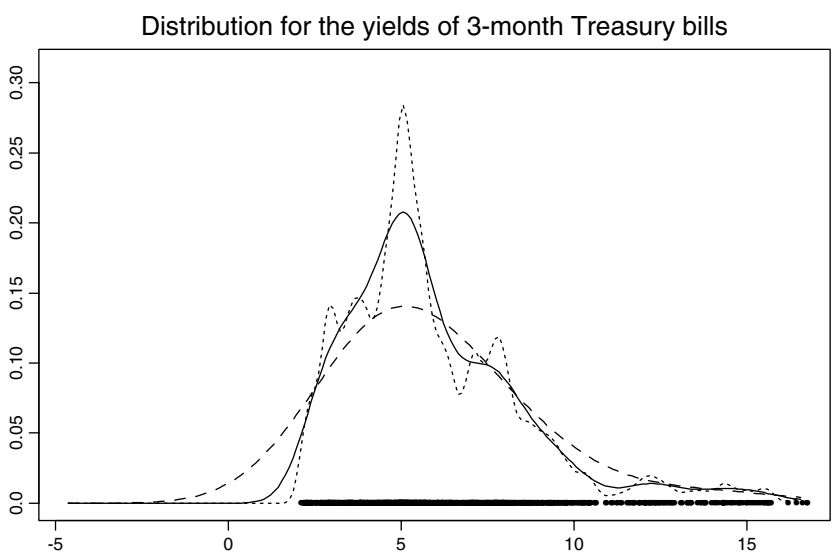


FIGURE 5.3. Estimated densities for the yields (in percent) of 3-month Treasury bills using the Gaussian kernel with bandwidths $h = 0.61/3$ (short-dashed curve), 0.61 (solid curve), and 3×0.61(long-dashed curve). A factor of 3 is intentionally used to illustrate the effects of undersmoothing and oversmoothing.

As shown in Figure 5.3, the distribution of the interest rates has a long right tail. The median and mode are about 5.34%, while the mean is 5.97%. The interest rate at the beginning of the 1980s was as high as over 15%.

5.3 Windowing and Whitening

If the data $\{X_t\}_{t=1}^T$ are a realization from a stationary process with marginal density f, then by a change of variable,

$$E\widehat{f}_h(x) = EK_h(X_t - x) = \int_{-\infty}^{+\infty} K(u)f(x + hu)du. \tag{5.2}$$

Thus, the *bias* of the estimator, defined as $E\widehat{f}_h(x) - f(x)$, does not depend on the dependent structure of the data. It is the same as that for the independent sample. The variance of the estimator can, however, be affected by the dependent structure.

To gain further insights, let us consider the case where K has a bounded support $[-1, 1]$. Then, the kernel density estimator (5.1) uses only the local

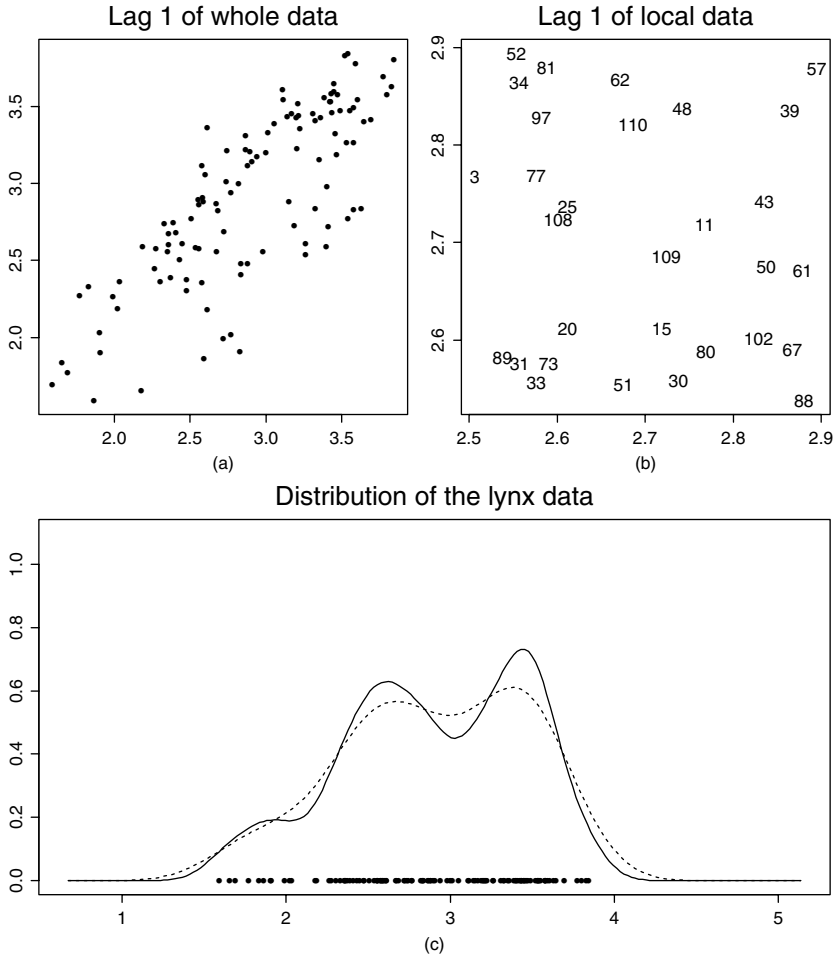


FIGURE 5.4. (a) Lag 1 scatterplot of the lynx data. (b) Lag 1 scatterplot of those data $\{Y_j, j = 1, \cdots, J\}$ falling in the local neighborhood 2.7 ± 0.2. The point $X_{t(j)}$ is plotted against $X_{t(j-1)}$ using the number $t(j)$ to indicate the point $(X_{t(j-1)}, X_{t(j)})$. (c) Kernel density estimate for the lynx data using the bandwidth $h = 0.14$ (solid) and 0.23 (dashed).

data points within the local window $x \pm h$:

$$\widehat{f}_h(x) = T^{-1} \sum_{j=1}^{J} K_h(X_{t(j)} - x),$$

where $t(j)$ is the jth data point falling in the interval $x \pm h$ and J is the total number of local data points. Although the data in the original sequence can be highly correlated, the dependence for the new series $\{Y_j =$

$X_{t(j)}, j = 1, \cdots, J\}$ in the local window around x can be much weaker. This is due to the fact that the time sequence $\{t(j), j = 1, \cdots, J\}$ is quite far apart for small bandwidth h (see Figure 5.4 for an illustration). The lag 1 autocorrelation for the lynx data is much stronger than that for the data in the local window, 2.7 ± 0.2. Indeed, the local data look like those from an independent sample. Hence, one would expect that the asymptotic variance for the kernel density estimator is the same as that for the independent observations when certain mixing conditions are imposed. This intuition is elucidated by Hart (1996). Because of the whitening property by local windowing in the state domain, the kernel density estimators for mixing processes behave very much like those for independent samples. Hence, all techniques for independent samples can be extended to mixing stationary processes. We will develop some of the basic theory in §5.6. The effect of dependence structure on the kernel density estimation was thoroughly studied recently by Claeskens and Hall (2002).

5.4 Bandwidth Selection

When the data $\{X_t\}$ are a realization from a stationary process, by Theorem 5.1 below, the *mean square error* (MSE) of the kernel density estimator can be expressed as

$$\begin{aligned} &\mathrm{MSE}(x) \equiv E\{\widehat{f}_h(x) - f(x)\}^2 \\ \approx\ & \frac{1}{4}\left\{\int_{-\infty}^{+\infty} u^2 K(u)du\right\}^2 \{f''(x)\}^2 h^4 + \int_{-\infty}^{+\infty} K^2(u)du \frac{f(x)}{Th} \end{aligned} \tag{5.3}$$

for x in the interior of the support of f. Here and hereafter, "$\approx$" means that both sides have the same leading terms. This is a pointwise measure. A global measure can be obtained by using a mean integrated square error (MISE):

$$\begin{aligned} &\mathrm{MISE} \equiv E\int_{-\infty}^{+\infty} \{\widehat{f}_h(x) - f(x)\}^2 dx \\ \approx\ & \frac{1}{4}\left\{\int_{-\infty}^{+\infty} u^2 K(u)du\right\}^2 \int_{-\infty}^{+\infty} \{f''(x)\}^2 dx h^4 + \int_{-\infty}^{+\infty} K^2(u)du \frac{1}{Th}. \end{aligned} \tag{5.4}$$

Minimizing the asymptotic MISE with respect to the bandwidth parameter h results in a bandwidth, called the *asymptotically optimal bandwidth* or simply the *optimal bandwidth*, which is given by

$$h_{\mathrm{opt}} = \alpha(K)\|f''\|_2^{-2/5} T^{-1/5}, \tag{5.5}$$

where $\|g\|_2^2 = \int_{-\infty}^{+\infty} g(u)^2 du$ is the L_2-norm, $\mu_2(K) = \int_{-\infty}^{+\infty} u^2 K(u)du$ is the variance of K, and

$$\alpha(K) = \mu_2(K)^{-2/5}\|K\|_2^{2/5}$$

TABLE 5.1. Some useful constants related to the kernel functions.

Functional	Gaussian	Uniform	Epanechnikov	Biweight	Triweight
$\mu_2(K)$	1	0.3333	0.2000	0.1429	0.1111
$\|K\|_2^2$	0.2821	0.5000	0.0600	0.7143	0.8159
$\alpha(K)$	0.7764	1.3501	1.7188	2.0362	2.3122
$\beta(K)$	0.3633	0.3701	0.3491	0.3508	0.3529

is a known constant. With this asymptotically optimal bandwidth, the optimal MISE is given by

$$\frac{5}{4}\beta(K)\|f''(x)\|_2^{2/5}T^{-4/5}, \tag{5.6}$$

where

$$\beta(K) = \mu_2(K)^{2/5}\|K\|_2^{8/5}.$$

It follows from (5.5) that the optimal bandwidths for the two different kernel functions K_1 and K_2 satisfy

$$h_{\text{opt}}(K_1) = \frac{\alpha(K_1)}{\alpha(K_2)}h_{\text{opt}}(K_2),$$

where $h_{\text{opt}}(K_1)$ and $h_{\text{opt}}(K_2)$ are, respectively, the optimal bandwidths associated with the kernel functions K_1 and K_2. Table 5.1 below tabulates the values of these useful functions for a few commonly-used kernel functions. From this table, different choices of kernels using their optimal bandwidth perform nearly the same (see the row with $\beta(K)$). Therefore, the kernel K_2 using the bandwidth h_2 performs nearly the same as the kernel K_1 using the bandwidth

$$h_1 = \frac{\alpha(K_1)}{\alpha(K_2)}h_2. \tag{5.7}$$

This is the idea behind the concept of the canonical kernel (Marron and Nolan 1988). It allows two investigators to compare the amount of smoothing even though they used two different kernels.

The optimal bandwidth (5.5) is not directly usable since it depends on the unknown parameter $\|f''\|_2$. When f is a Gaussian density with standard deviation σ, one can easily deduce from (5.5) that

$$h_{\text{opt},T} = (8\sqrt{\pi}/3)^{1/5}\alpha(K)\sigma T^{-1/5}. \tag{5.8}$$

The *normal reference bandwidth selector* (see, for example, Bickel and Doksum 1977; Silverman 1986) is the one obtained by replacing the unknown parameter σ in (5.8) by the sample standard deviation s. In particular, after

calculating the constant $\alpha(K)$ numerically, we have the following normal reference bandwidth selector:

$$\widehat{h}_{\text{opt},n} = \begin{cases} 1.06sT^{-1/5} & \text{for the Gaussian kernel} \\ 2.34sT^{-1/5} & \text{for the Epanechnikov kernel} \end{cases}. \tag{5.9}$$

An improved rule can be obtained by writing an Edgeworth expansion for f around the Gaussian density. Such a rule is provided in Hjort and Jones (1996b) and is given by

$$\widehat{h}^*_{\text{opt},T} = \widehat{h}_{\text{opt},T}\left(1 + \frac{35}{48}\widehat{\gamma}_4 + \frac{35}{32}\widehat{\gamma}_3^2 + \frac{385}{1024}\widehat{\gamma}_4^2\right)^{-1/5},$$

where $\widehat{\gamma}_3$ and $\widehat{\gamma}_4$ are, respectively, the sample *skewness* and *kurtosis*, defined by

$$\begin{aligned} \widehat{\gamma}_3 &= (T-1)^{-1}\sum_{t=1}^{T}(X_t - \bar{X})^3/s^3, \\ \widehat{\gamma}_4 &= (T-1)^{-1}\sum_{t=1}^{T}(X_t - \bar{X})^4/s^4 - 3. \end{aligned}$$

The normal reference bandwidth selector is only a simple rule of thumb. It is a good selector when the data are nearly Gaussian-distributed and is often reasonable in many applications. However, it can lead to oversmoothing when the underlying distribution is asymmetric or multimodal. In that case, one can either subjectively tune the bandwidth or select the bandwidth by more sophisticated bandwidth selectors. One can also transform data first to make their distribution closer to normal, then estimate the density using the normal reference bandwidth selector, and then apply the inverse transform to obtain an estimated density for the original data. Such a method is called the transformation method; see (5.12) below. For the asymmetric distribution suggested by Figure 5.3, the normal reference gives a somewhat oversmooth estimate. For the bimodal data in Figure 5.4(c), the normal reference bandwidth selector gives $\widehat{h} = 0.23$ and results in an oversmooth estimate. We hence reduce the amount of smoothing until a reasonable estimate (solid curve in Figure 5.4) is obtained.

There are quite a few important techniques for selecting the bandwidth, such as *cross-validation* (CV) and plug-in bandwidth selectors. A conceptually simple technique, with theoretical justification and good empirical performance, is the *plug-in technique*. This technique relies on finding an estimate of the functional $\|f''\|_2^2$ in (5.5). A good implementation of this approach is proposed by Sheather and Jones (1991). An overview on the progress of bandwidth selection can be found in Jones, Marron, and Sheather (1996); see also §6.3.5.

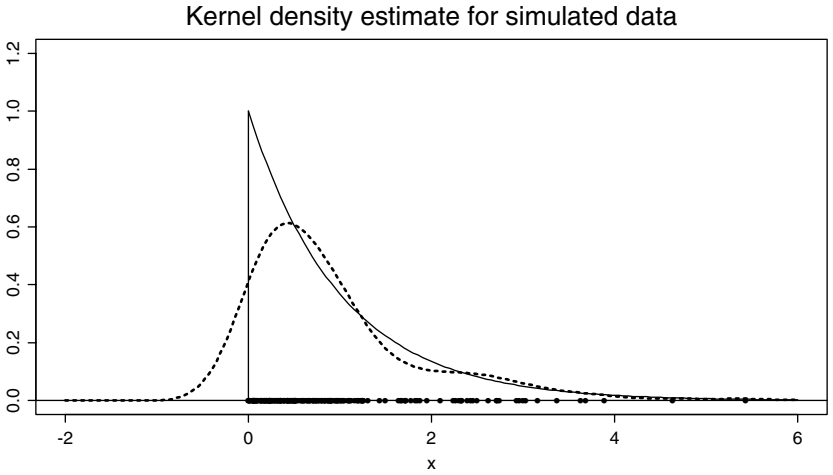


FIGURE 5.5. Kernel density estimate for a random sample of size $n = 200$ drawn from the standard exponential distribution; the solid curve is the true curve, and the dashed curve is the estimated curve. The boundary effect can easily be seen.

5.5 Boundary Correction

In many situations, the density f is known to have a bounded support. For example, the interest rate cannot be less than zero. It is natural to assume that the interest rate has support $[0, \infty)$. In fact, over the last forty years, the lowest short-term interest rate is 2.11% and the highest interest rate is 16.76%, so it is not unreasonable to assume that the short-term interest rate has a support interval [2%, 17%]. However, because a kernel density estimator spreads point masses smoothly around the observed data points, some of those near the boundary of the support are distributed outside the support of the density (see Figure 5.3). As a result, the kernel density estimator underestimates the density in the boundary regions. As shown in Figure 5.3, the problem is more severe for large bandwidths and for the left boundary, where the density is high. Therefore, some adjustments are needed.

To gain some further insights, let us assume without loss of generality that the density function f has a bounded support $[0, 1]$ and we deal with the density estimate at the left boundary. For simplicity, suppose that K has a support $[-1, 1]$. Then, the point $x = ch (0 \le c < 1)$ is a left boundary point. It can easily be seen that as $h \to 0$,

$$E\widehat{f}_h(ch) = \int_{-c}^{\infty} f(ch + hu)K(u)du = f(0) \int_{-c}^{\infty} K(u)du + o(1). \quad (5.10)$$

In particular, $E\widehat{f}_h(0) = f(0)/2+o(1)$ for symmetric kernels. In other words, the estimator at the left boundary point estimates only half of its true density. To illustrate this point, a random sample of size 200 was drawn from the standard exponential distribution $f(x) = \exp(-x)I(x \geq 0)$, and the density is estimated based on the kernel density estimator using the Gaussian kernel with bandwidth 0.344 obtained from (5.9). It is apparent that the estimate at point $x = 0$ is only about half of the true value.

There are several methods to deal with the density estimation at boundary points. Possible approaches include the *boundary kernel, reflection, transformation,* and *local polynomial fitting*. Here, we introduce two simple approaches: reflection and transformation methods.

The reflection method is to construct the kernel density estimate based on the "reflected" data $\{-X_t, t = 1, \cdots, T\}$ and the original data $\{X_t, t = 1, \cdots, T\}$. This results in the estimate

$$\widehat{f}_h^{\star}(x) = \frac{1}{T}\left\{\sum_{t=1}^{T} K_h(X_t - x) + \sum_{t=1}^{T} K_h(-X_t - x)\right\} \quad \text{for } x \geq 0. \tag{5.11}$$

Note that when x is away from the boundary, the second term in (5.11) is negligible. Hence, it only corrects the estimate in the boundary region; see Schuster (1985) and Hall and Wehrly (1991). This estimator is twice the kernel density estimate based on the synthetic data $\{\pm X_t, t = 1, \cdots, T\}$. In general, if the left boundary point is x_0 (instead of 0), the synthetic data are

$$\{-(X_t - x_0), X_t, t = 1, \cdots, T\},$$

leading to the estimate

$$\widehat{f}_h^{\star}(x) = \frac{1}{T}\left\{\sum_{t=1}^{T} K_h(X_t - x) + \sum_{t=1}^{T} K_h(x_0 - X_t - x)\right\}, \quad \text{for } x \geq x_0.$$

For the simulated data given in Figure 5.5, Figure 5.6(a) depicts the estimate based on this method. The Gaussian kernel and bandwidth 0.344 were used.

Another simple method is first to transform the data by

$$Y_i = g(X_i), i = 1, \cdots, n,$$

where g is a given monotone increasing function ranging from $-\infty$ to ∞. Now, apply the kernel density estimator (5.1) to this transformed data set to obtain the estimate $\widehat{f}_Y(y)$, and apply the inverse transform to obtain the density of X. This results in

$$\widehat{f}_X(x) = g'(x)\widehat{f}_Y(g(x)) = g'(x)T^{-1}\sum_{t=1}^{T} K_h(g(x) - g(X_t)), \tag{5.12}$$

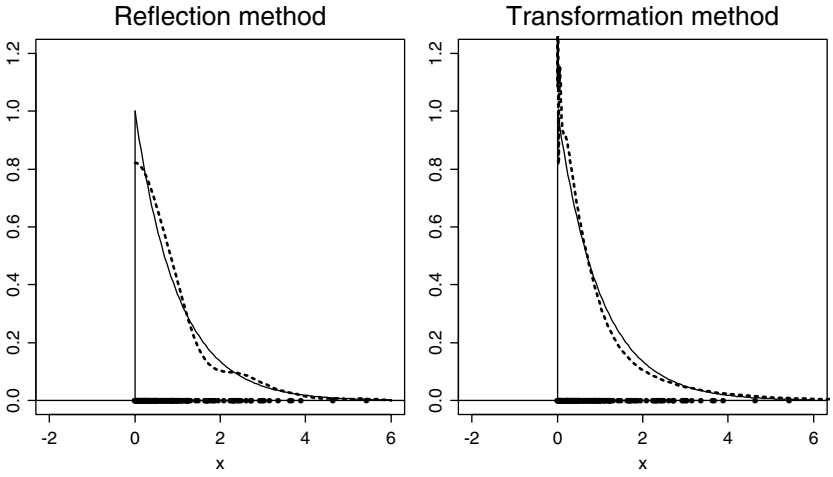


FIGURE 5.6. (a) Kernel density estimation using the reflection method. (b) Kernel estimate using the transformation method; the solid curve is the true curve, and the dashed curve is the estimated curve.

where $g'(\cdot)$ is the derivative function of $g(\cdot)$. Figure 5.6(b) illustrates this idea using the logarithmic transform to the data given in Figure 5.5. We first apply the kernel density to the transformed data $\{-\log(X_t), t = 1, \cdots, 200\}$ to obtain $\widehat{f}_Y(y)$. The normal reference bandwidth selector gives $h = 0.344$ for the transformed data using the Gaussian kernel. Hence, the estimated density is $\widehat{f}_X(x) = \widehat{f}_Y(\log x)/x$, or $\widehat{f}_X(\exp(x)) = \exp(-x)\widehat{f}_Y(x)$. Thus, the estimated density can be obtained by plotting $\exp(x)$ against $\exp(-x)\widehat{f}_Y(x)$. The density at $x = 0$ corresponds to the tail density of the transformed data since $\log(0) = -\infty$, which cannot usually be estimated well due to the lack of data at tails. Except at this point, the transformation method does a fairly good job.

5.6 Asymptotic Results*

We now derive the asymptotic bias and variance of the kernel density estimator as the sample size $T \to \infty$. Necessarily, the bandwidth h depends on T and tends to zero. The idea used here can be extended to more sophisticated settings such as nonparametric regression. We begin with a simple lemma that is useful for deriving asymptotic bias.

Lemma 5.1 *Let f have the pth bounded derivative that is continuous at an interior point x of the support of f. Assume that K is a function such*

that $\int_{-\infty}^{+\infty} |u^p K(u)|du < \infty$. *Then, as* $h \to 0$, *we have*

$$\int_{-\infty}^{+\infty} f(x+hu)K(u)du = \sum_{i=0}^{p} \mu_i(K) f^{(i)}(x)h^i/i! + o(h^p),$$

where $\mu_i(K) = \int_{-\infty}^{+\infty} u^i K(u)du$.

Proof. Let $D = \int_{-\infty}^{+\infty} f(x+hu)K(u)du - \sum_{i=0}^{p} \mu_i(K) f^{(i)}(x)h^i/i!$. Then

$$D = \int_{-\infty}^{+\infty} \left\{ f(x+hu) - \sum_{i=0}^{p} f^{(i)}(x)(hu)^i/i! \right\} K(u)du.$$

By the Taylor expansion,

$$D = \frac{h^p}{p!} \int_{-\infty}^{+\infty} \{f^{(p)}(x+\xi_T) - f^{(p)}(x)\} u^p K(u)du,$$

where ξ_T lies between 0 and hu. By a simple application of the Lebesgue dominated convergence theorem, we have

$$D/h^p \to 0 \qquad \text{as } h \to 0.$$

This completes the proof. ∎

Note that when the kernel function K has a bounded support, the integration above takes place only around a neighborhood of x. Hence, it suffices to assume that the density f has a pth continuous derivative at the point x. For this reason, the bounded support of K is frequently imposed for the sake of simplicity. It can be removed at the cost of lengthier arguments. In particular, the Gaussian kernel is allowed.

By using the lemma above and (5.2) for the kernel function satisfying

$$\int_{-\infty}^{+\infty} K(u)du = 1, \quad \int_{-\infty}^{+\infty} uK(u)du = 0,$$

we obtain immediately that the bias of the kernel density estimator is

$$\frac{\mu_2(K)}{2} f''(x)h^2 + o(h^2),$$

provided that f has a continuous second derivative. If f has a higher-order derivative, a bias of order $O(h^p)$ can be obtained by requiring

$$\mu_0(K) = 1, \mu_j(K) = 0, j = 1, \cdots, p-1, \tag{5.13}$$

but the gain usually is not substantial for practical sample sizes. A kernel satisfying (5.13) is called a *pth order kernel*. When $p > 2$, K can no longer be nonnegative since $\mu_2(K) = 0$.

We now turn to computing the variance component. For this, we assume that the process $\{X_t\}$ is a stationary process with α-mixing coefficient $\alpha(k)$. Furthermore, let $g_\ell(x, y)$ be the joint density between X_1 and $X_{\ell+1}$.

Theorem 5.1 *Let $\{X_t\}$ be an α-mixing process with the mixing coefficient $|\alpha(\ell)| \le C\ell^{-\beta}$ for some $c > 0$ and $\beta > 2$. Assume further that $\|g_\ell\|_\infty = \sup_{(x,y)} g_\ell(x,y)$ is bounded. Suppose that K is a bounded kernel function with a bounded support and $\mu_1(K) = 0$ and that $h \to 0$ in such a way that $Th \to \infty$. If f has the continuous second derivative at an interior point x of the support of f, then*

$$E\widehat{f}_h(x) = f(x) + \frac{\mu_2(K)}{2} f''(x)h^2 + o(h^2)$$

and

$$\mathrm{Var}\{\widehat{f}_h(x)\} = \frac{f(x)}{Th}\|K\|_2^2 + o\left(\frac{1}{Th}\right).$$

Proof. The bias expression follows directly from Lemma 5.1. Thus, we only need to derive the asymptotic expression for the variance term. Let $Z_t = K_h(X_t - x)$. Then, by the stationarity of $\{X_t\}$, we have

$$\mathrm{Var}(\widehat{f}_h(x)) = \frac{1}{T}\mathrm{Var}(Z_1) + \frac{2}{T}\sum_{\ell=1}^{T-1}(1 - \ell/T)\mathrm{Cov}(Z_1, Z_{\ell+1}).$$

Note that $EZ_1 = E\widehat{f}_h(x) = O(1)$. By a change of variables and Lemma 5.1, we have

$$\begin{aligned}
\mathrm{Var}(Z_1) &= EK_h^2(X_t - x) - (EZ_1)^2 \\
&= h^{-1}\int_{-\infty}^{+\infty} K^2(u)f(x + hu)dx - (EZ_1)^2 \\
&= h^{-1}f(x)\|K\|_2^2 + o(h^{-1}).
\end{aligned}$$

Thus, we need only to show that

$$\sum_{\ell=1}^{T-1} |\mathrm{Cov}(Z_1, Z_{\ell+1})| = o(h^{-1}). \tag{5.14}$$

By using Billingsley's inequality (Proposition 2.5 (ii)), we have

$$|\mathrm{Cov}(Z_1, Z_{\ell+1})| \le 4\alpha(\ell)\|Z_1\|_\infty\|Z_{\ell+1}\|_\infty \le 4\alpha(\ell)\|K\|_\infty^2/h^2. \tag{5.15}$$

On the other hand,

$$\begin{aligned}
|\mathrm{Cov}(Z_1, Z_{\ell+1})| &= |EZ_1Z_{\ell+1} - (EZ_1)^2| \\
&\le \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} K_h(u - x)K_h(v - x)g_\ell(u, v)dudv \\
&\quad +(EZ_1)^2 \\
&\le \|g_\ell\|_\infty + (EZ_1)^2.
\end{aligned} \tag{5.16}$$

Hence, the covariance is bounded by a constant C.

We now verify (5.14). Let $d_T \to \infty$ be a sequence of integers. Then, by (5.16),

$$\sum_{\ell=1}^{d_T-1} |\mathrm{Cov}(Z_1, Z_{\ell+1})| \le C d_T.$$

Using (5.15) and the assumption on the mixing coefficient, we have

$$\sum_{\ell=d_T}^{T-1} |\mathrm{Cov}(Z_1, Z_{\ell+1})| \le D \sum_{\ell=d_T}^{\infty} \ell^{-\beta}/h^2 = O(d_T^{-\beta+1}/h^2),$$

for some constant D. By taking $d_T = h^{-2/\beta}$, we have

$$\sum_{\ell=1}^{T-1} |\mathrm{Cov}(Z_1, Z_{\ell+1})| = O(h^{-2/\beta}) = o(1/h)$$

for $\beta > 2$. Hence (5.14) follows. This completes the proof. ∎

The pointwise *mean square error* admits the following bias and variance decomposition:

$$\begin{aligned} \mathrm{MSE}(x) &= E\{\widehat{f}_h(x) - f(x)\}^2 \\ &= \{E\widehat{f}_h(x) - f(x)\}^2 + \mathrm{Var}\{\widehat{f}_h(x)\}. \end{aligned}$$

By Theorem 5.1, an approximation of the MSE is given by (5.3). Minimizing the right-hand side of (5.3) with respect to h yields the asymptotic pointwise optimal bandwidth

$$h_{\mathrm{opt}}(x) = \alpha(K)\{f''(x)\}^{-2/5} f(x)^{1/5} T^{-1/5},$$

provided that $f''(x) \neq 0$. The minimum (ideal) risk, which is the minimizer of the main order approximation of $\mathrm{MSE}(x)$, is given by

$$\frac{5}{4}\beta(K) f''(x)^{2/5} f(x)^{4/5} T^{-4/5}, \tag{5.17}$$

where $\beta(K)$ is given in (5.6). Similarly, minimizing the right-hand side of (5.4) gives the optimal bandwidth (5.5) and the minimum (ideal) MISE in (5.6). Therefore, by taking the square-root, the kernel density estimator can achieve the rate of convergence $T^{-2/5}$. This rate is achievable as long as $h = cT^{-1/5}$ for some $c > 0$. This is the best possible rate for estimating the density function among the class of functions with the second bounded derivative, according to Farrell (1972), Hasminskii (1978), and Stone (1980).

The asymptotic normality of the kernel density estimator also holds. We state the theorem without a proof. The proof is very similar to that of Theorem 6.3. We leave it as an exercise to the reader. The condition $Th^{5/3} \to \infty$ can be relaxed if the mixing condition $\alpha(\ell) \le c|\ell|^{-\beta}$ with $\beta > 2$ is strengthened.

Theorem 5.2 *Under the conditions of Theorem 5.1, if* $Th^{5/3} \to \infty$, *then we have*

$$\sqrt{Th}\left\{\widehat{f}_h(x) - f(x) - \frac{\mu_2(K)}{2}f''(x)h^2 + o(h^2)\right\} \xrightarrow{D} N\left(0, f(x)\|K\|_2^2\right).$$

The kernel density estimator possesses various nice properties. See Chapter 2 of Bosq (1998) for some of them. We prove a result that is similar to Theorem 2.2 of Bosq (1998), but the geometric mixing condition there is significantly relaxed.

Theorem 5.3 *Assume that the mixing coefficient of the process* $\{X_t\}$ *satisfies* $\alpha(\ell) \le c\ell^{-\beta}$ *with* $\beta > 5/2$. *Suppose that the density* f *of* X_t *is bounded on an interval* $[a, b]$ *and that* K *satisfies a Lipschitz condition. Then*

$$\sup_{x\in[a,b]} |\widehat{f}_h(x) - E\widehat{f}_h(x)| = O_P\left\{\left(\frac{\log T}{Th}\right)^{-1/2}\right\},$$

provided that $h \to 0$ *in such a way that*

$$T^{2\beta-5}h^{2\beta+5}(\log T)^{-(2\beta+1)/4} \to \infty.$$

As a corollary of Theorem 5.3, when the process is *geometrically mixing* $\alpha(\ell) < c\rho^\ell$ for some $c > 0$ and some $\rho \in [0, 1)$, Theorem 5.3 holds for $h = dT^{-\gamma}$ for any $\gamma \in (0, 1)$ and $d > 0$.

Theorem 5.3 controls uniformly the stochastic errors of the kernel density estimator. The bias term $E\widehat{f}_h(x) - f(x)$ is deterministic and can easily be bounded uniformly by using Lemma 5.1. It is of order $O(h^2)$. By choosing $h = O((\log T/T)^{1/5})$, one obtains

$$\begin{aligned}\sup_{x\in[a,b]} |\widehat{f}_h(x) - f(x)| &\le \sup_{x\in[a,b]} |E\widehat{f}_h(x) - f(x)| + \sup_{x\in[a,b]} |\widehat{f}_h(x) - E\widehat{f}_h(x)| \\ &= O_P\{(\log T/T)^{2/5}\}\end{aligned}$$

when $\alpha(\ell) \le c\ell^{-\beta}$ with $\beta > 15/4$. This rate is optimal according to Hasminskii (1978).

A more precise description of the uniform convergence is given by Bickel and Rosenblatt (1973), who derived the asymptotic distribution of the normalized statistic

$$M_T = \sup_{0\le x\le 1} [f(x)\|K\|_2^2/(Th)]^{-1/2}(\widehat{f}_h(x) - E\widehat{f}_h(x))$$

for independent samples. We would expect the result to hold for a stationary process under certain mixing conditions. Here, the interval $[0, 1]$ is used for convenience. It can be replaced by any other intervals in the support of f. We require that the density f be continuous and positive on $[0, 1]$

and that $f'(x)/f^{1/2}(x)$ and $f''(x)$ be bounded on $[0,1]$. Assume that K is bounded and symmetric about 0. Moreover, K either vanishes outside an interval $[-A, A]$ and is absolutely continuous on $[-A, A]$ with derivative K' or is absolutely continuous on $(-\infty, \infty)$ such that $\mu_2(K)$, $\mu_2(K')$, and $\|K'\|_2$ are finite.

Theorem 5.4 *Under the conditions above, if $X_1, \cdots, X_T$ are independently and identically distributed, we have*

$$P\left\{(-2\log h)^{1/2}(M_T - d_T) < x\right\} \to \exp(-2\exp(-x)),$$

provided that $h = cT^{-\delta}$ for $0 < \delta < 1/2$ and $c > 0$, where

$$d_T = (-2\log h)^{1/2} + (-2\log h)^{-1/2}\{\log c(K)/\pi^{1/2} - 0.5\log\log h\}$$

if $c(K) = K^2(A)/\|K\|_2^2 > 0$ and otherwise

$$d_T = (-2\log h)^{1/2} + (-2\log h)^{-1/2}\log\frac{\|K'\|_2^2}{4\pi\|K\|_2^2}.$$

A corollary of Theorem 5.4 is that $M_T = O_p\{(-\log h)^{1/2}\}$. This entails that

$$\sup_{0\le x\le 1}\{\widehat{f}_h(x) - E\widehat{f}_h(x)\} = O_P[\{-(\log h)/(Th)\}^{1/2}],$$

the same order as that given in Theorem 5.3. An application of Theorem 5.4 is to construct *simultaneous confidence intervals* for all $\{f(x), x \in [0,1]\}$. Indeed, by Theorem 5.4, with approximate probability $1-\alpha$,

$$M_T \le d_T - (-2\log h)^{-1/2}\log\left\{-\frac{1}{2}\log(1-\alpha)\right\} \equiv c(\alpha, h).$$

The expression above is equivalent to

$$E\widehat{f}_h(x) \in \widehat{f}_h(x) \pm c(\alpha, h)[f(x)\|K\|_2^2/(Th)]^{1/2}, \quad \forall x \in [0,1].$$

Using Theorem 5.1, $E\widehat{f}_h(x) = f(x) + O(h^2)$. Substituting this into the last expression and replacing $f(x)$ by its estimate, we have an approximate level $(1-\alpha)$ confidence interval

$$f(x) \in \widehat{f}_h(x) \pm c(\alpha, h)[\widehat{f}_h(x)\|K\|_2^2/(Th)]^{1/2}, \quad \forall x \in [0,1],$$

if $h = o\{(T\log T)^{-1/5}\}$.

Finally, we would like to make some notes on the optimal kernel. Optimal theory on the choices of kernel functions can be found in Gasser, Müller, and Mammitzsch (1985) and Müller(1991, 1993). The ideal pointwise MSE (5.17) and the ideal MISE (5.6) depend on K through $\beta(K)$ given in (5.6). Theorem 5.5 shows that the optimal kernel is the Epanechnikov kernel.

The result is due to Epanechnikov (1969) for the kernel density estimation. But this result has already appeared in the robust regression literature (see p. 384 of Lehmann 1983, and references therein). Such a kernel has been used as a converging factor in the Fourier transform by Bochner (1936) and in spectral density estimation by Parzen (1961). With the known optimal kernel, it can easily be computed that the values of $\beta(K)$ for other commonly-used kernels are close to optimal. See row $\beta(K)$ of Table 5.1.

Theorem 5.5 *The nonnegative probability density function K that minimizes $\beta(K)$ is a rescaling of the Epanechnikov kernel:*

$$K_{\text{opt}}(u) = \frac{3}{4a}(1 - u^2/a^2)_+ \quad \textit{for any } a > 0.$$

Proof. First, we note that $\beta(K_h) = \beta(K)$ for any $h > 0$. Let K_0 be the Epanechnikov kernel. For any other nonnegative K, by rescaling if necessary, we assume that $\mu_2(K) = \mu_2(K_0)$. Thus, we need only to show that $\|K_0\| \le \|K\|$. Let $\delta = K - K_0$. Then

$$\int_{-\infty}^{+\infty} \delta(u)du = 0, \qquad \int_{-\infty}^{+\infty} u^2\delta(u)du = 0,$$

which implies that

$$\int_{-\infty}^{+\infty} (1 - u^2)\delta(u)du = 0.$$

Using this and the fact that K_0 has the support $[-1, 1]$, we have

$$\begin{aligned}
\int_{-\infty}^{+\infty} \delta(u)K_0(u)du &= \int_{|u|\le 1} \delta(u)(1 - u^2)du \\
&= -\int_{|u|>1} \delta(u)(1 - u^2)du \\
&= \int_{|u|>1} K(u)(u^2 - 1)du.
\end{aligned}$$

Since K is nonnegative, so is the last term. Therefore

$$\begin{aligned}
\int_{-\infty}^{+\infty} K^2(u)du &= \int_{-\infty}^{+\infty} K_0^2(u)du + 2\int_{-\infty}^{+\infty} K_0(u)\delta(u)du + \int_{-\infty}^{+\infty} \delta^2(u)du \\
&\ge \int_{-\infty}^{+\infty} K_0^2(u)du,
\end{aligned}$$

which proves that K_0 is the optimal kernel. ∎

5.7 Complements—Proof of Theorem 5.3

Throughout this section, we use C to denote a generic constant, which may vary from line to line.

We first reduce the problem from the supremum over the interval $[a, b]$ to the maximum over a grid of points on that interval. To this end, partition the interval $[a, b]$ into N subintervals $\{I_j\}$ of equal length. Let $\{x_j\}$ be the centers of I_j. By the Lipschitz condition on K, we have with probability tending to 1

$$|\widehat{f}_h(x) - \widehat{f}_h(x')| \le T^{-1} \sum_{t=1}^{T} |K_h(x - X_t) - K_h(x' - X_t)| \le Ch^{-1}|x - x'|.$$

This entails that

$$|E\widehat{f}_h(x) - E\widehat{f}_h(x')| \le E|\widehat{f}_h(x) - \widehat{f}_h(x')| \le Ch^{-1}|x - x'|.$$

Using these, we have

$$\sup_{x \in I_j} |\widehat{f}_h(x) - E\widehat{f}_h(x)| \le |\widehat{f}_h(x_j) - E\widehat{f}_h(x_j)| + C(Nh)^{-1}.$$

Thus

$$\sup_{x \in [a,b]} |\widehat{f}_h(x) - E\widehat{f}_h(x)| \le \max_{1 \le j \le N} |\widehat{f}_h(x_j) - E\widehat{f}_h(x_j)| + C(Nh)^{-1}.$$

By taking $N = (T/h)^{1/2}$, we have

$$\sup_{x \in [a,b]} |\widehat{f}_h(x) - E\widehat{f}_h(x)| \le \max_{1 \le j \le N} |\widehat{f}_h(x_j) - E\widehat{f}_h(x_j)| + C(Th)^{-1/2}. \quad (5.18)$$

We now bound the tail probability for $\widehat{f}_h(x) - E\widehat{f}_h(x)$. Let $Y_t = K_h(x - X_t) - EK_h(x - X_t)$. Then $\|Y_t\|_\infty < Ch^{-1}$. By using the exponential inequality (Theorem 2.18), we have for any $\varepsilon > 0$, and each integer $q \in [1, T/2]$,

$$\begin{aligned} P\{|\widehat{f}_h(x) - E\widehat{f}_h(x)| > \epsilon\} &\le 4 \exp\left(-\frac{\varepsilon^2 q}{8v^2(q)}\right) \\ &+ 22\left\{1 + \frac{4C}{h\varepsilon}\right\}^{1/2} q\alpha\left(\left[\frac{T}{2q}\right]\right), \quad (5.19) \end{aligned}$$

where

$$v^2(q) = 2\sigma^2(q)/p^2 + C\varepsilon/(2h)$$

with $p = [T/(2q)]$ and

$$\sigma^2(q) = \max_{0 \le j \le 2q-1} \mathrm{Var}\{Y_{jp+1} + \cdots + Y_{(j+1)p+1}\}.$$

By Theorem 5.1, $\sigma^2(q) \le Cph^{-1}$. Thus, by taking $q = T\epsilon$,

$$v^2(q) \le C(ph)^{-1} + C\epsilon h^{-1} \le C\varepsilon h^{-1}.$$

This and (5.19) imply that

$$P\{|\widehat{f}_h(x) - E\widehat{f}_h(x)| > \epsilon\} \le 4\exp(-CTh\varepsilon^2) + CTh^{-1/2}\varepsilon^{\beta+0.5}. \quad (5.20)$$

We now prove the theorem. By taking $\varepsilon^2 = a\log T/(CTh)$ for a sufficiently large a, the right-hand side of (5.20) is bounded by

$$4T^{-a} + CT^{-\beta/2+0.75}h^{-\beta/2-0.75}(\log T)^{\beta/2+0.25}.$$

Consequently,

$$\begin{aligned} & P\left(\max_{1\le j\le N}|\widehat{f}_h(x_j) - E\widehat{f}_h(x_j)| > \varepsilon\right) \\ \le\ & N\{4T^{-a} + T^{-\beta/2+0.75}h^{-\beta/2-0.75}(\log T)^{\beta/2+0.25}\} \\ =\ & o(1) + O\big\{(Th)^{-(2\beta-5)/4}h^{-5/2}(\log T)^{(2\beta+1)/4}\big\}, \end{aligned}$$

which tends to zero. This entails that

$$\max_{1\le j\le N}|\widehat{f}_h(x_j) - E\widehat{f}_h(x_j)| = O_p\left\{\left(\frac{\log T}{Th}\right)^{1/2}\right\},$$

which together with (5.18) proves the theorem. ∎

5.8 Bibliographical Notes

The literature on nonparametric smoothing is vast. It includes kernel density estimation, nonparametric regression, time-domain smoothing, spectral density estimation, and applications to other statistical estimations. Indeed, most parametric problems have their nonparametric counterpart. Most nonparametric results can be generalized from independent data to dependent data. Nonparametric function estimation has been one of the most active areas over the last three decades. Many new techniques have been invented, and many new phenomena have been unveiled. It is impossible to give a complete survey of this vast area. Rather, we sample only a small fraction of references from this active area. They are not even representative of the many important contributions in the field. In this section, we mainly outline the key developments for dependent data. Books on nonparametric function estimation listed in §1.7 give more detailed accounts of the work on independent data. Related literature can be found in §6.7 and §7.6.

Extensive treatments of nonparametric function estimation for dependent data can be found in the monographs by Györfi, Härdle, Sarda, and Vieu (1989), Rosenblatt (1991), and Bosq (1998). They mainly focus on the theoretical developments in univariate nonparametric smoothing.

Density estimation for independent data

There is much literature on kernel density estimation. Most of the work focuses on independent random samples. The basic idea of kernel density estimation appeared in a technical report by Fix and Hodges (1951). The asymptotic mean square errors and mean integrated square errors were studied by Rosenblatt (1956), Parzen (1962), and Watson and Leadbetter (1963). There are a number of books on kernel density estimation. These include Devroye and Györfi (1985), Silverman (1986), Scott (1992) and Wand and Jones (1995). Various properties of kernel density estimators can be found in the books by Prakasa Rao (1983) and Nadaraya (1989).

The properties of kernel density estimation have been widely studied. The idea of using higher-order kernels for bias reduction dates back to Parzen (1962) and Bartlett (1963). Davis (1975) used the sinc kernel to obtain a near root-n consistent estimator for supersmooth densities. Theory on optimal kernels has been extensively developed by Gasser, Müller, and Mammitzsch (1985), Granovsky and Müller (1991), and Müller (1993).

Optimal rates of convergence for density estimation were studied by Farrell (1972). They were further investigated by Hasminskii (1978) and Stone (1980, 1982). Ibragimov and Hasminskii (1984), Donoho and Liu (1991a, b), Fan (1993b), and Low (1993) expanded the scope of the minimax study. Sharp asymptotic minimax risks over Sobolov spaces were established by Pinsker (1980), Efromovich and Pinsker (1982), and Nussbaum (1985). These optimal rates of convergence depend on the smoothness of unknown functions. Adaptive procedures have been constructed so that they are nearly optimal for each given class of functions; see, for example, Efromovich (1985), Lepski (1991, 1992), Donoho and Johnstone (1995, 1996, 1998), Donoho, Johnstone, Kerkyacharian, and Picard (1995), Brown and Low (1996), and Tsybakov (1998). Adaptive estimation based on penalized least-squares can be found in Barron, Birgé, and Massart (1999) and Antoniadis and Fan (2001). Minimax results on nonlinear functionals can be found in Bickel and Ritov (1988), Fan (1991), and Birgé and Massart (1995), among others.

There are a number of variations and modifications of kernel density estimators. Local likelihood estimation of a density can be found in Loader (1996) and Hjort and Jones (1996a). Parametric guided nonparametric density and regression estimation was proposed in Hjort and Glad (1995), Efron and Tibshirani (1996), and Glad (1998). Transformation methods for kernel density estimation were studied by Wand, Marron, and Ruppert (1991) and Yang and Marron (1999). The idea of using variable band-

widths can be found in Breiman, Meisel, and Purcell (1977), Abramson (1982), Hall and Marron (1988), and Hall (1990), among others. Many papers in the literature deal with possible approaches for reducing boundary biases. Boundary kernel methods were introduced and studied by Gasser and Müller (1979) and Gasser, Müller, and Mammitzsch (1985). Schucany and Sommers (1977) and Rice (1984a) suggested a linear combination of two kernel estimators with different bandwidths to reduce biases. Boundary correction methods for smoothing splines have been studied by Rice and Rosenblatt (1981) and Eubank and Speckman (1991), among others.

Density estimation for dependent data

Early references on kernel density estimation for dependent data are Roussas (1967, 1969) and Rosenblatt (1970), where the local asymptotic normality is established. Strong consistency for estimating transition probability densities was established in Yakowitz (1979). Ahmad (1979, 1982) studied consistent properties for estimating the density of an α-mixing process using an orthogonal theory method. Masry (1983) derived asymptotic expressions for the bias and covariance of discrete-time estimates for the marginal probability density function of continuous-time processes; see also Robinson (1983). Density estimation for time series residuals was investigated by Robinson (1987). Cheng and Robinson (1991) established various properties of density estimation for strongly dependent data. The uniform strong consistent rate was established by Pham and Tran (1991) and Cai and Roussas (1992). Györfi and Masry (1990) proved the strong consistency of recursive density estimation for dependent data. Kim and Cox (1995) gave useful moment bounds for mixing random variables. Density estimation for random fields was investigated by Roussas (1995), Carbon, Hallin, and Tran (1996), and Bradley and Tran (1999), among others. Györfi and Lugosi (1992) gave an interesting example where the kernel density estimate is inconsistent. Adams and Nobel (1998) studied density estimation of ergodic processes. The impact of dependence on the MISE, ISE, and optimal bandwidths of the kernel density estimation has been thoroughly studied by Claeskens and Hall (2002).

6
Smoothing in Time Series

6.1 Introduction

Having introduced the basic concept of nonparametric function estimation in the last chapter, we are now ready to apply it to other important smoothing problems in time series. Smoothing techniques are useful graphic tools for estimating slowly-varying time trends, resulting in time domain smoothing (§6.2). Nonparametric inferences on the associations between future events and their associated present and past variables lead to state domain smoothing in §6.3. Spline methods, introduced in §6.4, are useful alternatives to the local polynomial techniques in §6.3. These techniques can easily be extended to estimate the conditional variance (volatility) of a time series and even the whole conditional distribution; see §6.5.

6.2 Smoothing in the Time Domain

6.2.1 Trend and Seasonal Components

The first step in the analysis of time series is to plot the data. This allows one to inspect visually whether a series resembles a realization of a stationary stochastic process. Should a trend or seasonal pattern be observed, it is usually removed before the analysis of the series.

Suppose that a time series $\{Y_t\}$ can be decomposed as

$$Y_t = f_t + s_t + X_t, \tag{6.1}$$

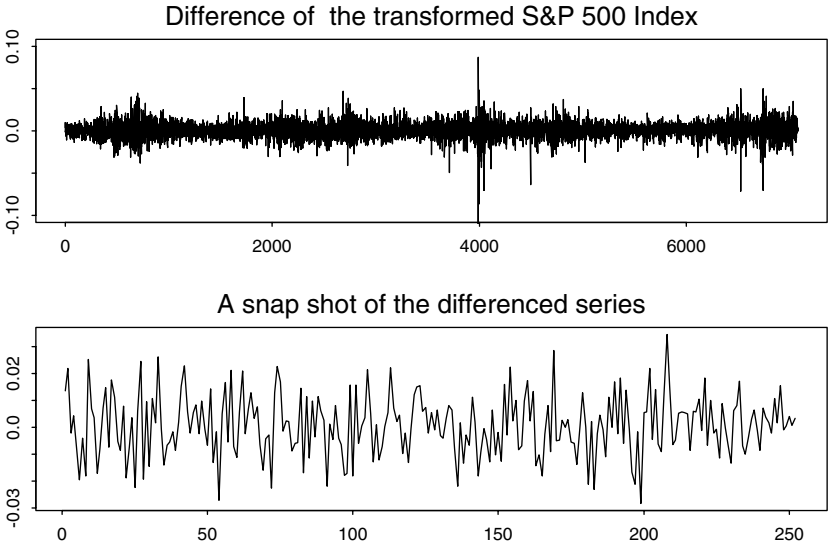


FIGURE 6.1. The difference of the logarithmic transform of the S&P 500 Index from January 3, 1972 to December 31, 1999 (top panel) and from January 4, 1999 to December 31, 1999 (bottom panel).

where f_t represents a slowly varying function known as a trend component, s_t is a periodic function referred to as a "seasonal component" and X_t is a stochastic component, which is assumed to be stationary with mean zero. A variance-stabilizing transformation or the *Box–Cox transform* may be applied before using the decomposition. This family of power transform admits the form

$$g(u) = \begin{cases} u^\lambda, & \text{for } \lambda \neq 0 \\ \log(u), & \text{for } \lambda = 0 \end{cases} \tag{6.2}$$

indexed by the parameter λ, or in the form having continuity at $\lambda = 0$,

$$g(u) = (u^\lambda - 1)/\lambda.$$

This class of transformation was considered by Box and Cox (1964). Note that a translation transform might be needed before using the power transform since the data in the power transform must be nonnegative.

Our objective is to estimate and extract the deterministic components f_t and s_t. It is hoped that the residual component X_t will be stationary and can be further analyzed by using linear and nonlinear time series techniques. An alternative approach, developed extensively by Box and Jenkins (1970), is to repeatedly apply difference operators to the time series $\{Y_t\}$ until the differenced series appears stationary. The differenced series is then processed further by using stationary time series techniques. As an illustration of the Box and Jenkins approach, we took the logarithmic transform

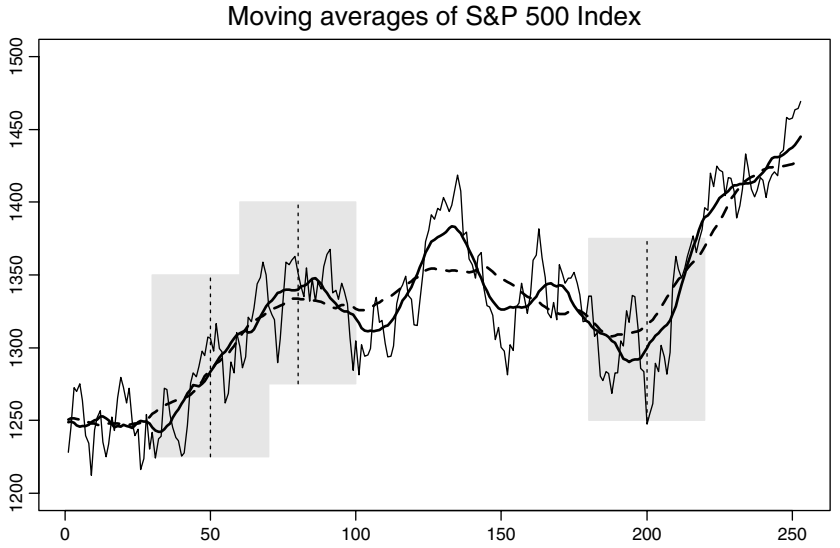


FIGURE 6.2. The S&P 500 Index from January 4, 1999 to December 31, 1999 and its 21 (thick curve) and 41-(dashed curve) trading day moving averages.

of the S&P 500 Index and then computed the first-order difference. Figure 6.1 presents this preprocessed series. The resulting series is basically the percentage of daily price changes in the index. It appears stationary except for a few outliers (e.g., 20.47% market corruption on October 19, 1987, called "Black Monday" in the financial markets). This transform is related to discretization of the geometric Brownian motion model popularly employed for asset pricing in the financial industry.

We first focus on the situation without the seasonal component, namely

$$Y_t = f_t + X_t, \quad EX_t = 0. \tag{6.3}$$

We then return to estimate the trend and seasonal components in §6.3.8.

6.2.2 Moving Averages

Averaging is the most commonly-used technique to reduce stochastic noise. Assume that the trend is slowly-varying so that it can be approximated by a constant in a local time window of size h, namely

$$Y_{t+i} \approx f_t + X_{t+i} \qquad \text{for } -h \le i \le h. \tag{6.4}$$

Then f_t can be estimated by the local average around this window:

$$\widehat{f_t} = (2h+1)^{-1} \sum_{i=-h}^{h} Y_{t+i}. \tag{6.5}$$

As the center t changes, the local time window moves. For example, the estimate at $t = 50$ with $h = 20$ is the average of the data in that first window depicted in Figure 6.2. The centers of the windows are moved to new points to form estimates at these points. As the local window slides from the left to the right, it traces a moving average curve. This is the simplest form of the *moving average smoothing*. It is frequently used to examine the trend of a time series. Figure 6.2 depicts the one-month and two-month moving averages for the S&P 500 Index from January 4, 1999 to December 31, 1999.

A convention for the moving average estimator at the boundary is to ignore the data beyond the observed time range. For example, f_2 is simply estimated by using the average of data $Y_1, \cdots, Y_{2+h}$ (more data to the right of the time point 2 than to the left). This asymmetric average may create an unappealing boundary bias. This boundary effect is more pronounced when the trend at the boundary is steep and the window size is large. As shown in Figure 6.2, the moving average underestimates the trend at the right boundary. The problem can be attenuated by using the *local linear smoothing* (see §6.2.6) or other boundary correction methods, such as the *boundary kernel* method (Gasser and Müller 1979; Müller 1993) and the data-sharpening method (Choi, Hall, and Rousson 2000).

The moving average series (6.5) utilizes both sides of data around the time t. It depends also on the data after time t. To facilitate prediction, the one-sided moving average series

$$\widehat{f}_t^* = h^{-1} \sum_{i=1}^{h} Y_{t-i} \tag{6.6}$$

is also frequently used to examine the time trend. The series employs only the past data up to time $t - 1$.

6.2.3 Kernel Smoothing

An improved version of the moving average estimator is to introduce a weighting scheme. This allows data near the given time point to receive larger weights. This leads to the *kernel regression estimator*, defined by

$$\widehat{f}_{t_0} = \frac{\sum_{t=1}^{T} Y_t K\left(\frac{t-t_0}{h}\right)}{\sum_{t=1}^{T} K\left(\frac{t-t_0}{h}\right)}. \tag{6.7}$$

This estimator is also called the Nadaraya–Watson estimator; see Nadaraya (1964) and Watson (1964). When the uniform kernel $K(u) = 0.5I(|u| \leq 1)$ is employed, the kernel estimator above becomes the moving average estimator (6.5). When the kernel function has a bounded support $[-1, 1]$, the kernel regression estimator is a weighted average of local $(2h + 1)$ data points around the time point t_0. When the kernel $K(t)$ is unimodal with

the mode at zero, the data points near t_0 receive more weight. In general, the kernel function is not required to have a bounded support as long as its tails are thin (e.g., a density function that has a second moment). The nonnegativity requirement of K can also be dropped. The bandwidth h does not need to be an integer.

Note that the normalization constants in the definition of the Gaussian kernel and the symmetric Beta family of kernels are merely used to make the function K a probability density function. They play no roles in kernel regression estimation. In computation, we often normalize the various kernel functions such that they have the same maximum value 1 as in Figure 5.2. With this normalization, (6.7) can be intuitively understood as the effective average of $\sum_{t=1}^{T} K\{(t-t_0)/h\}$ data points. When the kernel function has a support in $(-\infty, 0)$ (such a kernel is also referred to as a one-sided kernel), the kernel regression estimator uses only the data up to time $t_0 - 1$. This is an extension of the one-sided moving average (6.6).

As in the kernel density estimation, the bandwidth h is a critical parameter in kernel regression estimation. As demonstrated in Figure 6.2, a large bandwidth h produces an oversmooth estimate, leaving out possible details of the trend and underestimating the magnitude of peaks and troughs. Specifically, the estimator can create large biases when a large bandwidth is used. When a small bandwidth is applied, there are only a few local data points available to reduce the variance of the estimator. This results in a wiggly curve. For example, with $h = 0$, the moving average estimator (6.5) simply reproduces the original series. Trial-and-error is needed in order to produce satisfactory results. A data-driven choice of bandwidth can assist us in determining the amount of smoothness required . As shown in §6.2.9, the asymptotic variance depends critically on the correlation structure of the underlying process. Hence, the data-driven bandwidth selectors designed for independent data perform poorly in time-domain smoothing. Indeed, Altman (1990), Chu and Marron (1991a), and Hart (1991) reported that the ordinary leave-one-out *cross-validation* method performs poorly for the dependent data. Several modifications were proposed by these authors. The plug-in approaches for bandwidth selection were proposed by Ray and Tsay (1997) and Beran and Feng (2001).

The observation above can also be understood by calculating the bias and variance of the kernel regression estimator. Following direct calculation, under model (6.3), the bias of the kernel estimator is

$$E\widehat{f}_{t_0} - f_{t_0} = \frac{\sum_{t=1}^{T}(f_t - f_{t_0})K\left(\frac{t-t_0}{h}\right)}{\sum_{t=1}^{T} K\left(\frac{t-t_0}{h}\right)}.$$

This does not depend on the error process. It is purely an approximation error. When the bandwidth is small, the approximation errors $f_t - f_{t_0}$ are small and so is the bias term. On the other hand, when h is large, many of the approximation errors $f_t - f_{t_0}$ can be large due to the large distance

between t and t_0, and hence the bias can be large. The variance of this linear estimator,

$$\widehat{f}_{t_0} = \sum_{t=1}^{T} w_t Y_t, \quad w_t = \frac{K\left(\frac{t-t_0}{h}\right)}{\sum_{t=1}^{T} K\left(\frac{t-t_0}{h}\right)},$$

can also be computed. Let $\gamma_X(t)$ be the autocovariance function of the process $X(t)$. Then

$$\mathrm{Var}(\widehat{f}_{t_0}) = \sum_{i=1}^{T}\sum_{j=1}^{T} \gamma_X(|i-j|) w_i w_j. \tag{6.8}$$

The variance depends on the autocorrelation function. Further simplification needs asymptotic analysis. We will discuss this in §6.2.9. It will be shown that the asymptotic variance depends on the behavior of $\gamma_X(k)$ as $k \to \infty$. Suffice it to say that when the bandwidth is small, the variance of the kernel smoothing is large due to the limited amount of the local data point.

6.2.4 Variations of Kernel Smoothers

There are a number of variations of the kernel smoothers. The denominator in (6.7) is not convenient for taking derivatives with respect to t and for mathematical analysis. Instead, assigning the heights of a kernel function as weights, we can also use the areas under the kernel function as weights. Since the total area under the kernel function is one, no denominator is needed. This is the basic idea behind the Gasser–Müller estimator.

In the current context, let $s_t = (2t+1)/2 (t = 1, \cdots, T-1)$ with $s_0 = -\infty$ and $s_T = \infty$. Gasser and Müller (1979) proposed the following estimator:

$$\widehat{f}_{t_0} = \sum_{t=1}^{T} \int_{s_{t-1}}^{s_t} K_h(u - t_0) du Y_t.$$

No denominator is needed since the total weight is

$$\sum_{t=1}^{T} \int_{s_{t-1}}^{s_t} K_h(u - t_0) = \int_{-\infty}^{\infty} K_h(u - t_0) du = 1.$$

The Gasser–Müller estimator is a modification of an earlier version of Priestley and Chao (1972), which is defined as

$$\widehat{f}_{t_0} = \sum_{t=1}^{T} K_h(t - t_0) Y_t.$$

This estimator simply drops the denominator of the Nadaraya–Watson estimator. Approximating the Riemann sum by an integral and by a change of the variable, we have the total weight

$$\sum_{t=1}^{T} K_h(t-t_0) \approx \int_1^T K_h(t-t_0)dt = \int_{-(t_0-1)/h}^{(T-t_0)/h} K(u)du$$

for proper choices of h. If t_0 is not too close to the boundaries and h is small relative to T so that $(t_0-1)/h$ and $(T-t_0)/h$ are large, the integral above is approximately the same as

$$\int_{-\infty}^{\infty} K(u)du = 1.$$

In fact, this holds exactly as long as the support of K is contained in the interval $[-(t_0-1)/h, (T-t_0)/h]$. In other words, for t_0 that is not in the boundary region, the total weight is approximately 1. The argument above relies on the fact that the design points are equispaced. In fact, the Priestley and Chao estimator can only be applied to the equispace setting. It will not be applicable to the state-domain smoothing in §6.3.

6.2.5 Filtering

The kernel regression is a special convolution filter used by engineers. In general, a linear *filter* of length $2h+1$ is defined by

$$\widehat{f}_t = \sum_{i=-h}^{h} w_i Y_{t+i}. \tag{6.9}$$

The kernel regression corresponds to $w_i = K(i/h)/\sum_{j=-h}^{h} K(j/h)$ when K has the support $[-1,1]$. Filters $\{w_i\}$ can be designed to possess various properties. For example, they can be designed to remove high-frequency signals (*low-pass filter*) or low-frequency signals (*high-pass filter*) or signals outside a certain range of frequencies *(bandpass filter);* see §2.3.3. The kernel smoothing is a low-pass filter.

A linear filter can also be defined via a recursion. For example, a one-sided moving average $\widehat{f}_t$ can also be defined via

$$\widehat{f}_t = bY_t + (1-b)\widehat{f}_{t-1}, \quad t = 2, \cdots, T,$$

for some $b < 1$. This is equivalent to using the following weighted moving average of $Y_1, \cdots, Y_t$:

$$\widehat{f}_t = bY_t + b(1-b)Y_{t-1} + \cdots + b(1-b)^{t-2}Y_2 + b(1-b)^{t-1}Y_1.$$

Since the weights decrease exponentially fast, the filter above effectively uses only the local data near time t. The effective size of the smoothing

depends on the parameter b. This method is referred to as *exponential smoothing*.

The exponential smoothing is a special case of kernel smoothing using $K_h(x) = \lambda^x I(x \geq 0)$ with $\lambda^{1/h} = 1-b$. This is a one-sided kernel smoothing. It uses only the data up to the current time t. Further discussion on this subject can be found in Gijbels, Pope, and Wand (1999).

The recursive and convolution filterings can be combined to yield a much richer family of linear filters used in the engineering literature. The idea is very similar to combining AR and MA processes to enlarge the scope of linear processes.

6.2.6 Local Linear Smoothing

The local constant approximation (6.4) can be improved if the local linear approximation is used. Let us approximate the trend f_i as a function of i locally by a linear function

$$Y_i \approx f_t + f_t'(i-t) + X_i \qquad \text{for } |i-t| \leq h.$$

Thus f_t is approximately the intercept of the locally linear model above. See Figure 6.3 for an illustration at $t = 200$. The data inside the window are fitted by a linear regression. Using the least squares method for the data around the local window, we can estimate the local intercept via minimizing

$$\sum_{i=1}^{T} \{Y_i - a - b(i-t)\}^2 K_h(i-t)$$

with respect to a and b. Here, the kernel weights are introduced to weigh down the contributions of the data that are remote from the given time point t. Let $\widehat{a}_t$ and $\widehat{b}_t$ be the least-squares solutions. Here, the subscript t is used to indicate the fact that the solution depends on the given time point t. Then f_t is estimated by the local intercept $\widehat{a}_t$, which admits the explicit expression

$$\widehat{f}_t = \widehat{a} = \sum_{i=1}^{T} w_{t,i} Y_i / \sum_{i=1}^{T} w_{t,i}, \quad w_{t,i} = K_h(i-t)\{S_{T,2}(t) - (i-t)S_{T,1}(t)\}, \tag{6.10}$$

where $S_{T,j}(t) = \sum_{i=1}^{T} K_h(i-t)(i-t)^j$. The whole trend function is estimated when t runs from 1 to T. Thus, the *local linear smoother* is really a running linear regression method. As illustrated in Figure 6.3, the estimate at $t = 80$ is found by forming a new local least-squares problem. The linear fit in each data window is shown as a solid line. The local intercepts-the values of the estimate-are the intersections between the dashed vertical lines and the local linear lines. The local slopes are estimates of the derivatives of the time trend. Further, these local windows can also overlap with each

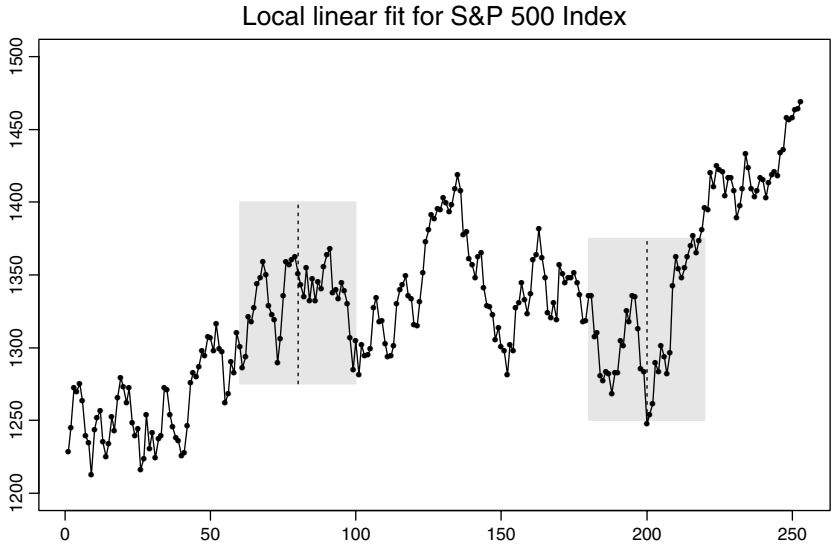


FIGURE 6.3. Local linear fit for the S&P 500 Index from January 4, 1999 to December 31, 1999, using the Epanechnikov kernel and bandwidth $h = 20$. The dashed parabola in each window indicates the weight that each local data point receives.

other (see Figure 6.2). The S-Plus function " lls.s" was programmed and used to compute the smoothed curve in Figure 6.3. This S-Plus function can be obtained from the Web site of this book.

The local linear smoothing can easily be extended to the local polynomial smoothing. A thorough treatment of local polynomial fitting and its applications can be found in Fan and Gijbels (1996). The merits of the local polynomial fitting will be summarized in §6.3.3. Note that weights $w_{t,i}$ in (6.11) satisfy

$$\sum_{i=1}^{T} w_{t,i}(i-t) = S_{T,1}(t)S_{T,2}(t) - S_{T,2}(t)S_{T,1}(t) = 0. \tag{6.11}$$

This implies that if the trend is linear, $f_t = \alpha t + \beta$, the local linear smoother is unbiased:

$$E\widehat{f}_t = \sum_{i=1}^{T} w_{t,i}(\alpha i + \beta) / \sum_{i=1}^{T} w_{t,i} = \alpha t + \beta.$$

In other words, the local linear smoother is unbiased for estimating linear trends, no matter how steep they are. This holds for t in the interior as well as near the boundary. In other words, the local linear estimator would have a small bias for estimating a steep trend. Kernel smoothers, on the other hand, would have large biases for estimating steep trends near

boundary regions because the equation similar to (6.11) does not hold, even approximately.

6.2.7 Other Smoothing Methods

There are many other variations of the kernel local linear smoother. For example, Gasser and Müller (1979) use different weighting schemes from the kernel and local linear weights, and Jones (1997) introduces variations to the local linear smoothing. Chapter 2 of Fan and Gijbels (1996) gives an overview on various smoothing techniques, including splines and orthogonal series methods.

The kernel regression and the local polynomial modeling are based on local approximations at many grid points. Global approximation methods such as splines can also be applied to the time domain smoothing. These ideas will be introduced in the state-domain smoothing in §6.4.

For equispaced designs such as the time domain smoothing, *orthogonal series methods* are also very handy to use. The basic idea is first to transform data using an orthogonal matrix and then selectively set coefficients at high frequencies to zero (or shrink them toward zero). The smoothed estimate can be obtained by the inverse transform of the tapered coefficients. Commonly used orthogonal transforms include the Fourier transform and the wavelet transform. For their statistical applications, see recent books by Ogden (1997), Efromovich (1999), and Vidakovic (1999).

6.2.8 Seasonal Adjustments

There are many ad hoc procedures for seasonal adjustments. We just outline one here to indicate the flavor.

Suppose that the period of the seasonal component in (6.1) is p; namely,

$$s_{k+jp} = s_k, \quad \sum_{k=1}^{p} s_k = 0. \tag{6.12}$$

The last constraint is an identifiability condition. Without this constraint, one can add a constant to the *trend component* f_t and subtract the same constant on the *seasonal component.* Due to the constraint (6.12), the trend can be conveniently estimated by using the moving average (6.5) with $h = (p-1)/2$ when p is an odd number. The seasonal component is averaged out in (6.5) and hence does not contribute to the trend estimate. When the period p is even, one can estimate the trend with a slight modification:

$$\widehat{f_t} = (0.5Y_{t-d} + Y_{t-d+1} + \cdots + Y_{t+d-1} + 0.5Y_{t+d})/p, \qquad d = p/2.$$

The seasonal component can be estimated as follows. For the sake of argument, we assume that we deal with monthly data and that the seasonal

component has period $p = 12$. The value of the seasonal component in, say, March can be well-approximated by the average of all of the observations made in March, after removing the trend component. This leads to the estimate

$$\widehat{s}_k^* = \sum_{j=[(d-k)/p]+1}^{[(T-d-k)/p]} (Y_{k+jp} - \widehat{f}_{k+jp})/\{[(T-d-k)/p] - [(d-k)/p] + 1\},$$

where $[a]$ indicates the integer part of a and $d = [p/2]$. The limits in the summation above are imposed so that the data are not too close to the boundary so as to minimize the boundary effect in trend estimation. This preliminary estimate may not exactly satisfy the constraint (6.12). This can easily be modified by using

$$\widehat{s}_k = \widehat{s}_k^* - d^{-1} \sum_{i=1}^{d} \widehat{s}_i^*, \quad k = 1, \cdots, p$$

to estimate the seasonal component $\{s_k\}$.

The technique above is also applicable in the absence of the trend component f_t. In this case, one does not need to remove the trend–namely, setting $\widehat{f}_t = 0$.

6.2.9 Theoretical Aspects*

The theoretical formulation of problem (6.3) should be made with care. One simple way is to think of the observed time series $\{Y_t\}$ as a discretized sample path from a continuous process

$$Y(t) = f(t) + X(t).$$

This formulation is frequently used in financial time series modeling. The time unit is usually years and weekly data (say) are regarded as the data sampled from a continuous process at the rate $\Delta = 1/52$. The formulation is very powerful for option pricing and risk management in finance. However, it has some drawbacks in the time domain smoothing. First, to be able to estimate $f(t)$ consistently, we need to localize the data around a given time t_0 with window size $h \to 0$. However, as long as the process $X(t)$ is continuous, all local data $\{Y(t) : t \in t_0 \pm h\}$ are highly correlated, with the correlation tending toward 1 as $h \to 0$. This implies that local data do not vary much, and hence local smoothing is not needed. As shown in Figure 6.2, local data do vary substantially, and the local smoothing does improve the trend estimation. Thus, the formulation above seems pathological from the theoretical point of view. Secondly, under the formulation above, the trend $f(t)$ and the stochastic error $X(t)$ have similar degrees

of smoothness (both of them are continuous). Hence, there is no hope of separating the trend part from the stochastic part in $Y(t)$.

An alternative formulation involves extending the nonparametric regression model for equispace design to the time series setup. One assumes that the observed time series is a realization from the model

$$Y_t = g(t/T) + X_t, \quad t = 1, \cdots, T \tag{6.13}$$

for a smooth time trend function g and a stationary process $\{X_t\}$ with $EX_t = 0$. Under this formulation, we can now separate the smooth trend from the noisy stochastic error via smoothing techniques. One minor drawback is that the smooth trend $f(t) = g(t/T)$ depends on the number of observations T. This problem appears already in the literature of nonparametric regression with fixed designs. It is not really a serious issue. After all, the asymptotic theory is only a means of providing a simplified structure for our understanding of theoretical properties. Modeling the trend as $g(t/T)$ is a simple technical device for capturing the feature that the trend is much more slowly varying than the noise.

The selection between the two formulations above depends on the problems under study. In longitudinal data analysis and functional data analysis, Hart and Wehrly (1986) and Silverman (1996) basically used the first formulation: one observes many independent series from the model $Y(t) = f(t) + X(t)$. This formulation is suitable for their problems. For time domain smoothing, model (6.13) is frequently assumed; see, for example, Hall and Hart (1990), Johnstone and Silverman (1997), and Robinson(1997). This enables one to capture the feature that the time trend is much smoother than the underlying stochastic noise. Furthermore, it enables one to consistently estimate the time trend.

With the formulation (6.13), the asymptotic properties for the kernel and the local linear smoothers can be obtained. The bias for estimating g is the same as that for the independent sample with a uniform design. The variances for the kernel and the local linear estimators can also be computed with extra effort. They depend on the covariance structure of the noise process $\{X_t\}$. In general, we assume that the autocorrelation function of $\{X_t\}$ behaves as

$$\gamma_X(k) \equiv \mathrm{Cov}(X_t, X_{t+k}) \sim C_X k^{-\alpha}, \quad \text{as } k \to \infty, \tag{6.14}$$

for some $\alpha > 0$ and some constant C_X. Fractional ARIMA processes defined in §2.5.2 satisfy (6.14).

We now consider the bias and variance of the local linear estimator (6.10) under the model (6.13). We rewrite the estimator (6.10) as $\widehat{g}(t/T)$. For any $u = t/T \in (0, 1)$, using $EY_i = g(i/T)$ and (6.11), we have the bias

$$E\widehat{g}(u) - g(u) = \frac{\sum_{i=1}^{T} w_{Tu,i}\{g(i/T) - g(u) - g'(u)(i/T - u)\}}{\sum_{i=1}^{T} w_{Tu,i}}. \tag{6.15}$$

Note that this bias does not depend on the error process $\{X(t)\}$. It is purely the approximation error of the local linear fit.

For simplicity of technical arguments, we assume that K has a bounded support. This assumption can be weakened at the expense of lengthier arguments. In particular, light-tail kernels such as the Gaussian kernel are allowed. Denote $\int_{-\infty}^{+\infty} v^j K(v)dv$ by μ_j.

We summarize the asymptotic bias and variance in the following Theorem, which will be proved in §6.6.1. Note that because of our scale of time unit, h/T is the same as the bandwidth used for conventional nonparametric regression.

Theorem 6.1 *Suppose that K has a bounded support, satisfying $\mu_0(K) = 1$ and $\mu_1(K) = 0$, and the bandwidth $h \to \infty$ in such a way that $h/T \to 0$.*

(a) If $g''(\cdot)$ exists and is continuous at the point u, then

$$E\widehat{g}(u) - g(u) = \frac{1}{2}\mu_2(K)g''(x)(h/T)^2 + o\{(h/T)^2\}.$$

(b) If the autocovariance γ_X satisfies (6.14), we have

$$\mathrm{Var}\{\widehat{g}(u)\} = \begin{cases} C_X \int\int K(x)K(y)|x-y|^{-\alpha}dxdyh^{-\alpha}, & 0 < \alpha < 1 \\ 2C_X\|K\|_2^2 h^{-1}\log(h), & \text{when } \alpha = 1 \\ \sum_{j=-\infty}^{\infty}\gamma_X(j)\|K\|_2^2 h^{-1}, & \text{when } \alpha > 1 \end{cases} \tag{6.16}$$

Theorem 6.1 shows that the asymptotic variance is strongly influenced by the covariance structure of the process of $\{X_t\}$. This in turn affects the asymptotic optimal bandwidth and explains why data-driven bandwidth selectors for independent data cannot be applied directly to the dependent data.

The result similar to Theorem 6.1 for the kernel estimator was proved by Hall and Hart (1990). It was recently extended to local polynomial fitting by Beran and Feng (2001) using technical arguments different from those given in §6.6.1. It was also shown there that the asymptotic variance is of order h^{-1-2d} for antipersistent processes.

The asymptotic normality for the local linear estimator can also be established. If the error process $\{X_t\}$ is Gaussian, then its weighted average estimator (6.10) is also Gaussian. Thus, the asymptotic normality of the local linear estimator follows directly from Theorem 6.1. Furthermore, under the normality assumption, Csörgö and Mielniczuk (1995) established the asymptotic distribution for the maximum deviation that is similar to Theorem 5.4. However, the normality assumption on $\{X_t\}$ is not critical. It can be removed as demonstrated in Robinson (1997). Here we outline the technical device used in that paper.

Let $\{\varepsilon_t\}$ be a *martingale* difference with respect to its natural σ-fields, namely

$$E(\varepsilon_t|\{\varepsilon_j, j < t\}) = 0, \quad \text{a.s.}$$

Assume that $\{X_t\}$ is a doubly infinite-order moving average process

$$X_t = \sum_{j=-\infty}^{\infty} a_j \varepsilon_{t-j}, \quad \text{with} \sum_{j=-\infty}^{\infty} a_j^2 < \infty,$$

and that the $\{\varepsilon_t^2\}$ are uniformly integrable, satisfying

$$E(\varepsilon_t^2 | \{\varepsilon_j, j < t\}) = 1, \quad \text{a.s.}$$

Fractional ARIMA processes satisfy these assumptions. Consider the weighted sum

$$S_T = \sum_{t=1}^{T} w_{T,t} X_t = \sum_{j=-\infty}^{\infty} \left(\sum_{t=1}^{T} w_{T,t} a_{t-j} \right) \varepsilon_j,$$

which is the sum of the martingale difference. Using the martingale property,

$$\text{Var}(S_T) = \sum_{j=-\infty}^{\infty} \left(\sum_{t=1}^{T} w_{T,t} a_{t-j} \right)^2,$$

which is assumed to exist. The following result is due to Robinson (1997). A similar theorem can be found in Ibragimov and Linnik (1971).

Theorem 6.2 *Under the conditions just stated,*

$$Var(S_T)^{-1/2} S_T \xrightarrow{D} N(0,1),$$

provided that

$$\max_j \left| \sum_{t=1}^{T} w_{T,t} a_{t-j} \right| = o\left(\text{Var}(S_T)^{-1/2} \right).$$

Now, for the local linear estimator (6.10), one can easily see that

$$\widehat{f}_t - E\widehat{f}_t = \sum_{i=1}^{T} w_{t,i} X_i / \sum_{i=1}^{T} w_{t,i}.$$

The asymptotic normality becomes a matter of checking the conditions stated in Theorem 6.2. We omit the details.

6.3 Smoothing in the State Domain

6.3.1 Nonparametric Autoregression

Smoothing in the state domain is strongly related to nonparametric prediction. Consider a stationary time series $\{X_t\}$. For simplicity, we consider

the prediction based on the variable X_{t-1} only. The best prediction of X_t based on $X_{t-1} = x$ is the conditional expectation of X_t given $X_{t-1} = x$,

$$m(x) = E(X_t | X_{t-1} = x),$$

which minimizes the MSE

$$E\{X_t - g(X_{t-1})\}^2$$

among all prediction rules g. This function is also called the *autoregression function* of order 1. When $\{X_t\}$ is a stationary Gaussian process with mean 0, this conditional mean is linear $m(x) = ax$ and the conditional variance is constant. This leads to an AR(1)-model

$$X_t = aX_{t-1} + \varepsilon_t.$$

In general, the function $m(x)$ is not necessarily linear and the conditional variance is not necessarily homoscedastic. However, we can always express the data in the form

$$X_t = m(X_{t-1}) + \sigma(X_{t-1})\varepsilon_t, \tag{6.17}$$

where $\sigma^2(x) = \mathrm{Var}(X_t | X_{t-1} = x)$. Here ε_t has conditional zero mean and unit variance

$$E(\varepsilon_t | X_{t-1}) = 0, \qquad \mathrm{Var}(\varepsilon_t | X_{t-1}) = 1.$$

Nonparametric smoothing techniques can be applied beyond the estimation of the autoregression function. Consider a bivariate sequence $\{(X_t, Y_t) : t = 1, \cdots, T\}$ that can be regarded as a realization from a stationary process. We are interested in estimating the regression function $m(x) = E(Y_t | X_t = x)$. To facilitate comprehension, we write

$$Y_t = m(X_t) + \sigma(X_t)\varepsilon_t, \tag{6.18}$$

where $\sigma^2(x) = \mathrm{Var}(Y_t | X_t = x)$ and ε_t satisfies

$$E(\varepsilon_t | X_t) = 0, \qquad \mathrm{Var}(\varepsilon_t | X_t) = 1.$$

Clearly, this setup includes estimating the autoregression function as a specific example by taking $Y_t = X_{t+1}$. Here are three useful examples.

Example 6.1 Consider a stationary time series $\{Z_t\}$. One takes $Y_t = (Z_t)^k$ and $X_t = Z_{t-1}$ for a given k. Then, the target function becomes

$$m_k(x) = E(Z_t^k | Z_{t-1} = x).$$

The conditional variance can be estimated by using $\widehat{m}_2(x) - \widehat{m}_1(x)^2$. In particular, when $m_1(x)$ is small, such as the difference among the interest

rate data given in Example 1.1, $m_2(x)$ is basically the same as the conditional variance function. In other words, the mean regression function for the data given in Figure 6.4 below is the square of the volatility function

$$\sigma(x) = \sqrt{\text{Var}(X_t|X_{t-1} = x)}.$$

This forms the basis of the volatility estimator given in Stanton (1997) and Fan and Yao (1998).

Example 6.2 Consider again the stationary time series $\{Z_t\}$. One takes $Y_t = I(a < Z_t \leq b)$, the indicator function on the interval $(a, b]$, and $X_t = Z_{t-1}$. Then, the target function becomes

$$m(x) = P(a < Z_t \leq b|Z_{t-1} = x).$$

In particular, if $a = -\infty$, we are estimating the conditional distribution. Furthermore, if $a = y - \delta$ and $b = y + \delta$, then $m(x)/(2\delta)$ is basically the same as the conditional density of Z_t given $Z_{t-1} = x$ when δ is small. This conditional density function is very useful for the summary of the distribution of Z_t given $Z_{t-1} = x$. In particular, the autoregression function is the center of this distribution, and the volatility function is the spread of this distribution. The idea forms the genesis of the methods used by Fan, Yao, and Tong (1996) for estimating conditional densities (§6.5) and their related functionals (§10.3), by Hall, Wolff, and Yao (1999) for estimating conditional distribution functions (§10.3), and by Polonik and Yao (2000) for estimating minimum-volume predictive regions (§10.4).

Example 6.3 For a given time series $\{Z_t\}$, *multistep forecasting* can be accomplished by setting $Y_t = Z_{t+d}$ and $X_t = Z_t$, where d is the number of steps. In this case, we estimate nonparametrically

$$m(x) = E(Z_{t+d}|Z_t = x),$$

the best d-step predictor based on the variable Z_t. Figure 6.6 below depicts the one-step and two-step predictions for the lynx data. By combining this method with the techniques in Examples 6.1 and 6.2, we can estimate conditional variance and conditional density for multistep forecasting.

6.3.2 Local Polynomial Fitting

Local polynomial fitting is a widely used nonparametric technique. It possesses various nice statistical properties. For a detailed account on the subject, see Fan and Gijbels (1996).

Let $m^{(\nu)}(x)$ be the νth derivative of the regression function defined in (6.18). The local polynomial technique is very convenient to use for estimating $m^{(\nu)}(x)$, including the regression function itself, $m(x) = m^{(0)}(x)$. Since the form of the function $m(\cdot)$ is not specified, a remote data point from x_0

provides very little information about $m(x_0)$. Hence, we can only use the local data points around x_0. Assume that $m(x)$ has the $(p+1)$ derivative at the point x_0. By Taylor's expansion, for x in the local neighborhood of x_0, we have

$$\begin{aligned} m(x) &= m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2!}(x - x_0)^2 \\ &\quad + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p + O\left\{(x - x_0)^{p+1}\right\}. \end{aligned} \tag{6.19}$$

In terms of statistical modeling, locally around x_0, we model $m(x)$ as

$$m(x) \approx \sum_{j=0}^{p} \beta_j \, (x - x_0)^j \, . \tag{6.20}$$

The parameters $\{\beta_j\}$ depend on x_0 and are called *local parameters*. Clearly, the local parameter $\beta_\nu = m^{(\nu)}(x_0)/\nu!$. Fitting the *local model* (6.20) using the local data, one minimizes

$$\sum_{t=1}^{T} \left\{ Y_t - \sum_{j=0}^{p} \beta_j \, (X_t - x_0)^j \right\}^2 K_h(X_t - x_0), \tag{6.21}$$

where h is a bandwidth controlling the size of the local neighborhood.

As an illustration, we took $Y_t = (X_t - X_{t-1})^2$, where X_t is the yield of the 12-month Treasury bill. The bandwidth $h = 3.06$ was used, which was selected by the preasymptotic substitution method (see §6.3.5) using the C-code "lls.c". At the point $x_0 = 12$ (percent), a line ($p = 1$) was fitted for the local data in the shaded area $x_0 \pm h$, with weights for each data point indicated by the dashed curve (corresponding to the Epanechnikov kernel). The local intercept β_0 at the point x_0 is the intersection between the fitted line and the vertical line. This forms an estimate of the regression function ($\nu = 0$) at the point $x_0 = 12$. Sliding this window along the horizontal axis, we obtain an estimated curve on the interval $[3, 14]$. The conditional standard deviation is shown in Figure 6.4(b). The residual-based method for estimating the conditional variance, proposed by Fan and Yao (1998) and computed by the C-code "autovar.c" (see also §8.7.2), is shown in the short-dashed curve for comparison. The parametric model $m(x) = \alpha x^\beta$ is frequently used to model the volatility of interest rate dynamics, which is shown in the long-dashed curve. As one can see, there are still substantial differences between the parametric and the nonparametric methods, and the question of adequacy of the parametric fitting arises. The preasymptotic substitution method of Fan and Gijbels (1995) was employed to select bandwidths; see §6.3.5.

Denote by $\widehat{\beta}_j$, $j = 0, \ldots, p$, the solution to the least squares problem (6.21). The *local polynomial estimator* for $m^{(\nu)}(x_0)$ is $\widehat{m}_\nu(x_0) = \nu!\widehat{\beta}_\nu$,

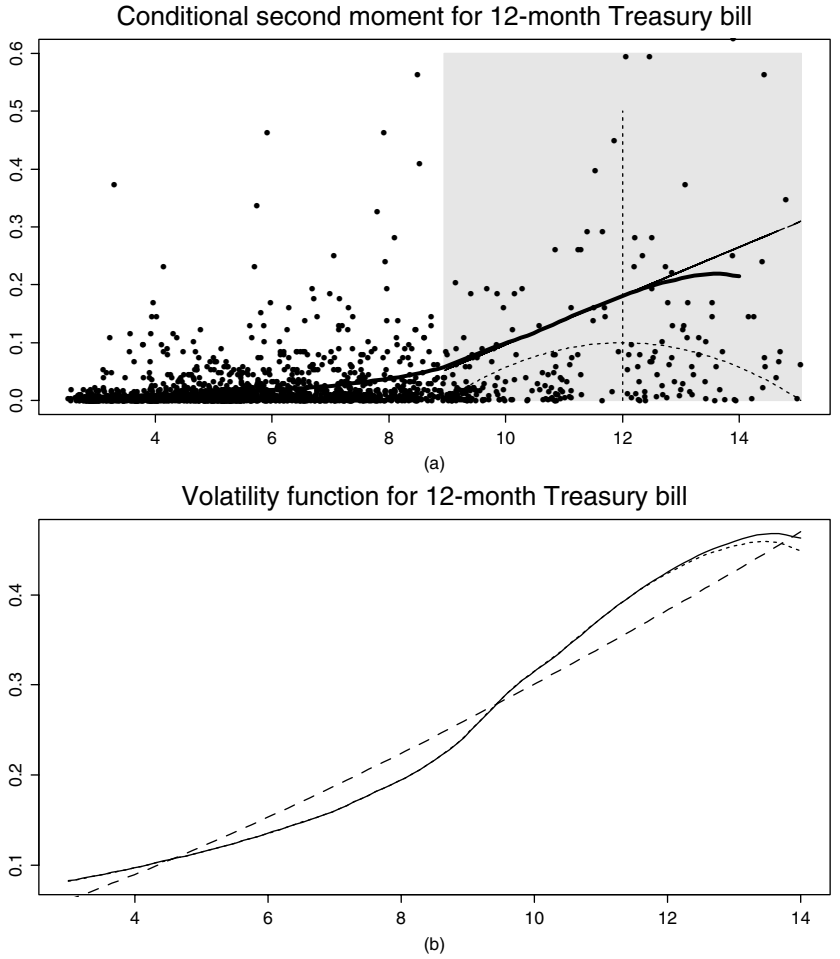


FIGURE 6.4. Local linear fit for estimating conditional variance for the yields of the 12-month Treasury bill. (a) Illustration of the local linear fit with the Epanechnikov kernel and bandwidth $h = 3.06$; (b) estimated conditional standard deviation by using the local linear fit (solid curve), the residual-based method of Fan and Yao (1998) (short-dashed curve), and the parametric model $\sigma(x) = \alpha x^{\beta}$ (long-dashed curve) with $\alpha = 0.143$ and $\beta = 1.324$.

$(\nu = 0, 1, \cdots, p)$. Here, we do not use the notation $\widehat{m}^{(\nu)}(x_0)$ in order to avoid confusion with the νth derivative function of the estimated regression $\widehat{m}(x_0)$. In fact, the derivative $m'(x)$ is estimated by the local slope rather than the derivative of the estimated regression function.

When $p = 0$, the local polynomial fit reduces to the kernel regression estimator

$$m(x) = \frac{\sum_{t=1}^{T} Y_t K_h(X_t - x)}{\sum_{t=1}^{T} K_h(X_t - x)},$$

which is also called the Nadaraya–Watson estimator. Hence, from the local approximation point of view, the kernel regression estimator is based on the local constant approximation; see (6.19).

It is more convenient to work with matrix notation. Denote by $\mathbf{X}$ the design matrix of problem (6.21),

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_T - x_0) & \cdots & (X_T - x_0)^p \end{pmatrix},$$

and put

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_T \end{pmatrix} \qquad \text{and} \qquad \widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \vdots \\ \widehat{\beta}_p \end{pmatrix}.$$

Then, the weighted least squares problem (6.21) can be written as

$$\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{y} - \mathbf{X}\beta), \tag{6.22}$$

with $\beta = (\beta_0, \cdots, \beta_p)^T$, where $\mathbf{W}$ is the diagonal matrix whose ith element is $K_h(X_i - x_0)$. The solution vector is given by

$$\widehat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}. \tag{6.23}$$

To implement the local polynomial estimator, one needs to choose the order p, the bandwidth h, and the kernel K. These parameters are of course related each other. When $h = \infty$, the local polynomial fitting becomes a global polynomial fitting and the order p determines the model complexity. Unlike in the parametric models, the complexity of local polynomial fits is primarily controlled by the bandwidth. Hence p is usually small, and the issue of choosing p becomes less critical. If the objective is to estimate $m^{(\nu)}$, the local polynomial fitting automatically corrects the boundary bias when $p - \nu$ is odd. Furthermore, when $p - \nu$ is odd, compared with the order $p - 1$ fit (so that $p - \nu - 1$ is even), the order p fit contains one extra parameter without increasing the variance for estimating $m^{(\nu)}$. But this extra parameter creates opportunities for bias reduction, particularly in the boundary regions; see Fan (1992), Fan and Gijbels (1992), Hastie and Loader (1993), and Ruppert and Wand (1994). For these reasons, the odd order fits (the order p is chosen so that $p - \nu$ is odd) outperforms the even order fits (the order $(p - 1)$ fit so that $p - \nu$ is even). Based on theoretical and practical

considerations, the order $p = \nu + 1$ is recommended in Fan and Gijbels (1996). If the primary objective is to estimate the regression function, one uses the local linear fit, and if the target function is the first-order derivative, one uses the local quadratic fit, and so on. On the other hand, the choice of the bandwidth h plays an important role in the local polynomial fitting. Too large a bandwidth causes oversmoothing, creating excessive modeling bias, whereas too small a bandwidth results in undersmoothing, obtaining noisy estimates. The bandwidth can be subjectively chosen by users by visually inspecting resulting estimates or automatically chosen by data by minimizing an estimated theoretical risk (see §6.3.5). Since the estimate is based on the local regression (6.21), it is reasonable to require a nonnegative weight function K. It is shown by Fan et al.(1996) that, for all choices of p and ν, the optimal weight function is $K(z) = \frac{3}{4}(1 - z^2)_+$, the Epanechnikov kernel. Thus, it is a universal weighting scheme and provides a useful benchmark to compare with other kernels. As shown in §5.5, other kernels have nearly the same efficiency for practical use of p and ν. Hence, the choice of the kernel function is not critical.

The local polynomial estimator compares favorably with other estimators, including the Nadaraya–Watson estimator, the Gasser and Müller estimator, and the Priestley and Chao estimator. Indeed, it was shown by Fan (1993a) that the local linear fitting is asymptotically minimax among all linear estimators and is nearly minimax among all possible estimators. This minimax property is extended by Fan et al. (1996) to more general local polynomial fitting.

6.3.3 Properties of the Local Polynomial Estimator

Throughout this section, we assume that $(X_1, Y_1), \cdots, (X_T, Y_T)$ are a stationary sequence. Let $\mathcal{F}_i^k$ be the σ-algebra of events generated by the random variables $\{(X_j, Y_j), i \le j \le k\}$. Let $\alpha(k)$ and $\rho(k)$ be their corresponding α- and ρ-mixing coefficients. Denote by $e_{\nu+1}$ the unit vector with 1 at the $(\nu + 1)$ position. Let

$$S_{T,j} = \sum_{t=1}^{T} K_h(X_t - x_0)(X_t - x_0)^j \tag{6.24}$$

and $\mathbf{S}_T = \mathbf{X}^T\mathbf{W}\mathbf{X}$ be the $(p + 1) \times (p + 1)$ matrix, whose (i, j)th element is $S_{T,i+j-2}$.

First, one can easily show that the estimator $\widehat{\beta}_\nu$ can be written as

$$\widehat{\beta}_\nu = e_{\nu+1}^T\widehat{\beta} = \sum_{t=1}^{T} W_\nu^T \left(\frac{X_t - x_0}{h}\right) Y_t, \tag{6.25}$$

where the *effective kernel* W_ν^T is the multiplication of the kernel K with a polynomial function, defined as

$$W_\nu^T(t) = e_{\nu+1}^T \mathbf{S}_T^{-1}\{1, th, \cdots, (th)^p\}^T K(t)/h. \tag{6.26}$$

The expression above reveals that the estimator $\widehat{\beta}_\nu$ looks like a conventional kernel estimator except that the "kernel" W_ν^T depends on the design points $\{X_1, \cdots, X_T\}$ and locations x_0. This explains why the local polynomial fit can adapt automatically to various designs and to boundary estimation. Figure 6.5 presents the effective kernel functions for the local constant fit ($p = 0$) and the local linear fit ($p = 1$) at $x_0 = 0.05$ and $x_0 = 0.5$ for the Epanechnikov kernel K. They satisfy the following moment property:

Proposition 6.1 *The effective kernel W_ν^T satisfies the following finite moment properties*

$$\sum_{t=1}^T (X_t - x_0)^q W_\nu^T\left(\frac{X_t - x_0}{h}\right) = \delta_{\nu,q} \quad 0 \le \nu, q \le p,$$

where $\delta_{\nu,q} = 0$ if $\nu \neq q$ and 1, otherwise.

Proof. By the definition of $\mathbf{S}_T$,

$$\begin{aligned}
&\sum_{t=1}^T (X_t - x_0)^q W_\nu^T\left(\frac{X_t - x_0}{h}\right) \\
&= e_{\nu+1}^T \mathbf{S}_T^{-1} \sum_{t=1}^T (X_t - x_0)^q \begin{pmatrix} 1 \\ X_t - x_0 \\ \vdots \\ (X_t - x_0)^p \end{pmatrix} K_h(X_t - x_0) \\
&= e_{\nu+1}^T \mathbf{S}_T^{-1} \mathbf{S}_T e_{q+1} = \delta_{\nu,q}.
\end{aligned}$$

The conclusion follows. ∎

As a consequence of Proposition 6.1, the local polynomial estimator is unbiased for estimating β_ν when the true regression function $m(x)$ is a polynomial of order p. To gain more insights about the effective kernel, we provide its asymptotic form. We first introduce some notation. Let $\mathbf{S}$ be the $(p+1) \times (p+1)$ matrix whose (i, j) element is μ_{i+j-2}, where $\mu_j = \int_{-\infty}^{+\infty} u^j K(u) du$. Define the *equivalent kernel* by

$$K_\nu^*(t) = e_{\nu+1}^T \mathbf{S}^{-1}(1, t, \cdots, t^p)^T K(t) = \left(\sum_{\ell=0}^p \mathbf{S}^{\nu\ell} t^\ell\right) K(t), \tag{6.27}$$

where $\mathbf{S}^{\nu\ell}$ is the $(\nu+1, \ell+1)$-element of $\mathbf{S}^{-1}$.

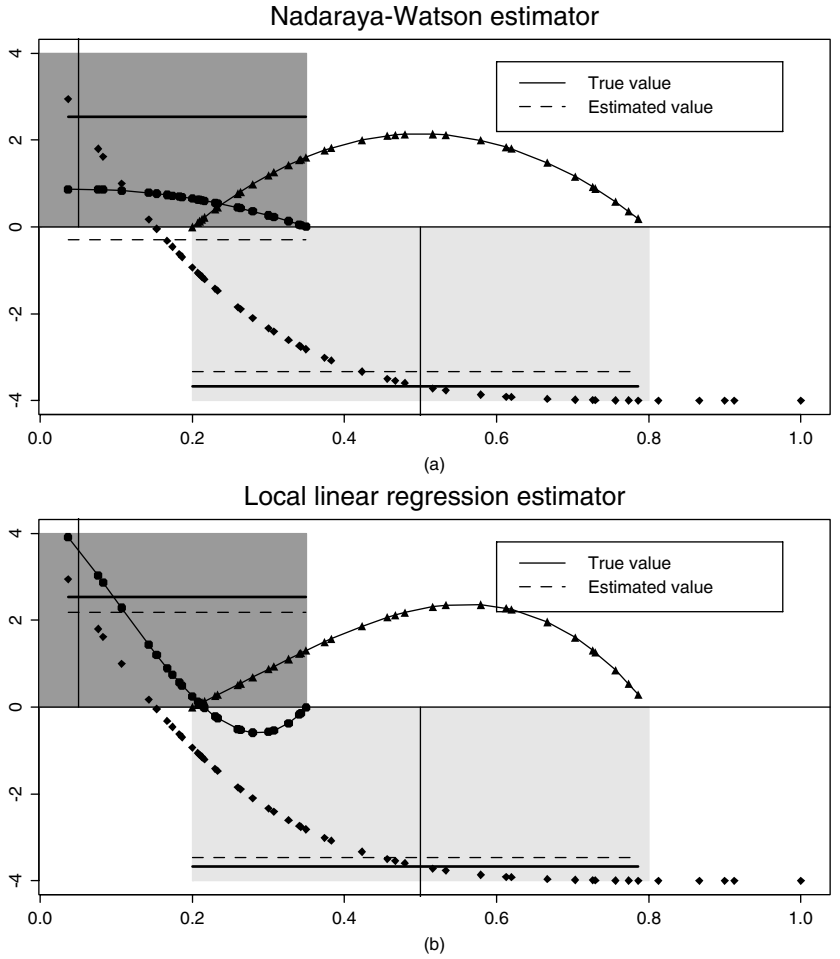


FIGURE 6.5. Effective weights assigned to local data points at an interior point $x_0 = 0.5$ (weights denoted by $\triangle$) and a boundary point $x_0 = 0.05$ (weights denoted by $\circ$) for the local constant fit ($p = 0$) and the local linear fit ($p = 1$), with K being the Epanechnikov kernel. The horizontal solid and dashed lines are the heights of true and estimated functions at $x_0 = 0.05$ and $x_0 = 0.5$, respectively. Their differences are biases at these two points. (a) The Nadaraya–Watson estimator; (b) the local linear fit. For clarity, the data ($\Diamond$) contain no noise.

Proposition 6.2 *Under the conditions of Theorem 5.3, if the marginal density f of X has a continuous derivative at point x_0, then*

$$W_\nu^T(t) = \frac{1}{Th^{\nu+1}f(x_0)} K_\nu^*(t)\{1 + O_P(a_T)\}$$

uniformly in $x_0 \in [a, b]$ *and* t, *where* $a_T = h + (\log T/Th)^{1/2}$. *The equivalent kernel satisfies the following moment condition for a higher-order kernel:*

$$\int_{-\infty}^{+\infty} u^q K_\nu^*(u)du = \delta_{\nu,q} \quad 0 \le \nu, q \le p.$$

Proof. Note that $S_{T,j}/(Th^j)$ is basically the same as the kernel density estimator with the induced kernel $K^*(x) = x^j K(x)$. Hence, by Theorem 5.3,

$$(Th^j)^{-1} S_{T,j} = f(x_0)\mu_j + O_P(a_T) \tag{6.28}$$

uniformly in $x_0 \in [a, b]$. From this, one obtains immediately by substituting (6.28) into each element of S_T that

$$T^{-1}H^{-1}\mathbf{S}_T H^{-1} = f(x_0)\mathbf{S}\{1 + O_P(a_T)\},$$

or equivalently,

$$\mathbf{S}_T = Tf(x_0)H\mathbf{S}H\{1 + O_P(a_T)\},$$

where $H = \text{diag}(1, h, \cdots, h^p)$. Hence, substituting this into the definition of W_T^ν, we find that

$$W_\nu^T(t) = \frac{1}{Th^{\nu+1}f(x_0)} e_{\nu+1}^T \mathbf{S}^{-1}(1, t, \cdots, t^p)^T K(t)\{1 + o_P(a_T)\}.$$

This proves the first conclusion. The second conclusion follows from the same proof as that of Proposition 6.1. ∎

From (6.25) and Proposition 6.2,

$$\widehat{\beta}_\nu = \frac{1}{Th^{\nu+1}f(x_0)} \sum_{t=1}^{T} K_\nu^* \left(\frac{X_t - x_0}{h} \right) Y_t\{1 + O_P(a_T)\}. \tag{6.29}$$

Hence, the local polynomial estimator works like a kernel regression estimator with a known design density f. This explains why the local polynomial fit adapts to various design densities. In contrast, the kernel regression estimator has a large bias at the region where the derivative of f is large; namely, it cannot adapt to highly-skewed designs. To see this, imagine that the true regression function has large slope in this region. Since the derivative of the design density is large, for a given x_0, there are more points on one side of x_0 than the other. When the local average is taken, the Nadaraya–Watson estimate is biased toward the side with more local data points because the local data are asymmetrically distributed. This issue is more pronounced at the boundary regions since the local data are even more asymmetric (see Figure 6.5). On the other hand, the local polynomial fit creates asymmetric weights, if needed, to compensate for this kind of design bias (Figure 6.5 (b)). Hence, it is adaptive to various design densities and to the boundary regions.

We now give the asymptotic bias and variance expression for local polynomial estimators. For independent data, we can obtain the bias and variance expression via conditioning on the design matrix $\mathbf{X}$. However, for time series data such as those in Examples 6.1–6.3, conditioning on $\mathbf{X}$ would mean conditioning on nearly the entire series. Hence, we derive the asymptotic bias and variance using the asymptotic normality rather than conditional expectation. As explained in §5.3, localizing in the state domain weakens the dependent structure for the local data. Hence, one would expect that the result for the independent data continues to hold for the stationary process with certain mixing conditions. The mixing condition and the window size should be related. A rigorous statement of this is given in Condition 1(iv) in §6.6.2. The proof of the following theorem, due to Masry and Fan (1997), will be outlined in §6.6.2.

Theorem 6.3 *Under Condition 1 in §6.6.2, if* $h = O\big(T^{1/(2p+3)}\big)$ *and* $m^{(p+1)}(\cdot)$ *is continuous at the point* x*, then as* $T \to \infty$*,*

$$\sqrt{Th}\left[\mathrm{diag}(1, \cdots, h^p)\{\widehat{\beta}(x) - \beta_0(x)\} - \frac{h^{p+1} m^{(p+1)}(x)}{(p+1)!}\mathbf{S}^{-1}\mathbf{c}_p\right]$$
$$\xrightarrow{D} N\{0, \sigma^2(x)\mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1}/f(x)\},$$

where $\beta_0(x) = \big(m(x), \cdots, m^{(p)}(x)/p!\big)^T$*,* $\mathbf{S}^*$ *is a* $(p+1) \times (p+1)$ *matrix whose element* (i, j) *is* $\nu_{i+j-2} = \int_{-\infty}^{+\infty} t^{i+j-2} K^2(t)dt$*, and* $\mathbf{c}_p$ *is a* $(p+1)$*-dimensional vector with* i *element* μ_{p+2-i}*.*

Note that from the definition of the equivalent kernel, one can easily see that

$$\int_{-\infty}^{+\infty} t^{p+1} K_\nu^*(t)dt = e_{\nu+1}^T \mathbf{S}^{-1}\mathbf{c}_p$$

and

$$\int_{-\infty}^{+\infty} K_\nu^*(t)^2 dt = e_{\nu+1}^T \mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1} e_{\nu+1}.$$

Hence, an immediate consequence of Theorem 6.1 is that the derivative estimator $\widehat{m}_\nu(x)$ is asymptotically normal:

$$\sqrt{Th^{2\nu+1}}\left\{\widehat{m}_\nu(x) - m^{(\nu)}(x) - \int t^{p+1} K_\nu^*(t)dt \frac{\nu! m^{(p+1)}(x)}{(p+1)!} h^{p+1-\nu}\right\}$$
$$\xrightarrow{D} N\left\{0, \frac{(\nu!)^2 \sigma^2(x) \int K_\nu^{*2}(t)dt}{f(x)}\right\}. \tag{6.30}$$

When $\nu = 0$, (6.30) gives the asymptotic normality of $\widehat{m}(x)$ itself.

The asymptotic bias and the asymptotic variance for the local polynomial estimator are naturally defined as

$$\mathrm{AB}(x) = \int t^{p+1} K_\nu^*(t)dt \frac{\nu! m^{(p+1)}(x)}{(p+1)!} h^{p+1-\nu}, \tag{6.31}$$

$$\mathrm{AV}(x) = \frac{(\nu!)^2 \sigma^2(x) \int K_\nu^{*2}(t)dt}{Th^{2\nu+1} f(x)}. \tag{6.32}$$

The ideal choice of bandwidth is the one that minimizes with respect to h

$$\int_{-\infty}^{+\infty} \{\mathrm{AB}^2(x) + \mathrm{AV}(x)\} w(x)dx$$

for a given weight function w. This leads to the asymptotic optimal bandwidth

$$h_{\mathrm{opt}} = C_{\nu,p}(K) \left[\frac{\int \sigma^2(x) w(x)/f(x)dx}{\int \{m^{(p+1)}(x)\}^2 w(x)dx} \right]^{1/(2p+3)} T^{-1/(2p+3)}, \tag{6.33}$$

where

$$C_{\nu,p}(K) = \left[\frac{(p+1)!^2 (2\nu+1) \int K_\nu^{*2}(t)dt}{2(p+1-\nu)\{\int t^{p+1} K_\nu^*(t)dt\}^2} \right]^{1/(2p+3)}.$$

However, this ideal bandwidth is not directly usable since it depends on unknown functions. We will propose methods to estimate this in §6.3.5.

As mentioned in the last section, local polynomial fits adapt automatically to boundary regions when $p - \nu$ is odd. To demonstrate this, we follow the formulation of Gasser and Müller (1979). Suppose that X_t has a bounded support, say, [0, 1]. Then $x = ch (0 \le c < 1)$ is a right boundary point when the kernel K has a bounded support $[0, 1]$. We now consider the behavior of $\widehat{m}_\nu(x)$ at the boundary point $x = ch$. To this end, let

$$\mu_{j,c} = \int_{-c}^{\infty} u^j K(u)du \quad \text{and} \quad \nu_{j,c} = \int_{-c}^{\infty} u^j K^2(u)du.$$

Define $\mathbf{S}_c$, $\mathbf{S}_c^*$, and $\mathbf{c}_{p,c}$ similarly to $\mathbf{S}$, $\mathbf{S}^*$, and $\mathbf{c}_p$, with μ_j and ν_j replaced by $\mu_{j,c}$ and $\nu_{j,c}$, respectively. Similarly, one defines the equivalent kernel at the boundary by

$$K_{\nu,c}^*(t) = e_{\nu+1}^T S_c^{-1} (1, t, \ldots, t^p)^T K(t).$$

Then, we have the following result, whose proof is very analogous to that of Theorem 6.3.

Theorem 6.4 *Suppose that Condition 1 in §6.6.2 holds and $f(0) > 0$. If $h = O\big(T^{1/(2p+3)}\big)$ and $m^{(p+1)}$ and $\sigma^2 f$ are right-continuous at the point 0,*

then as $T \to \infty$,

$$\sqrt{Th}\left[\text{diag}(1, \cdots, h^p)\{\widehat{\beta}(ch) - \beta_0(0)\} - \frac{h^{p+1}m^{(p+1)}(0)}{(p+1)!}\mathbf{S}_c^{-1}\mathbf{c}_{p,c}\right]$$
$$\xrightarrow{D} N\{0, \sigma^2(0)\mathbf{S}_c^{-1}\mathbf{S}_c^*\mathbf{S}_c^{-1}/f(0+)\},$$

where $\beta_0(0) = \big(m(0), \cdots, m^{(p)}(0)/p!\big)^T$.

As a consequence of Theorem 6.4, we have the asymptotic bias and variance at the boundary point $x = ch$ as

$$\text{AB}(x) = \int_{-c}^{\infty} t^{p+1}K_{\nu,c}^*(t)dt\frac{\nu!m^{(p+1)}(0+)}{(p+1)!}h^{p+1-\nu}$$

and

$$\text{AV}(x) = \frac{(\nu!)^2\sigma^2(0+)\int_{-c}^{\infty}K_{\nu,c}^{*2}(t)dt}{Th^{2\nu+1}f(0+)}.$$

Compare them with (6.31) and (6.32). Note that when K is symmetric and $p - \nu$ is even, it can be shown (Ruppert and Wand 1994) that the coefficient in (6.31) is zero. Hence, the bias is of smaller order at an interior point than that at a boundary point. This is referred to as a boundary effect. When $p - \nu$ is odd, the biases at interior and boundary points are of the same order. Indeed, they are even continuous at the point $c = 1$, the boundary between interior and boundary points. Hence, the local polynomial fit does not create excessive boundary bias when $p - \nu$ is odd. Assume that $p - \nu$ is odd and K is symmetric. It can be shown that the asymptotic variance for the local polynomial fit of order $p-1$ has the same asymptotic variance as that for the order p fit (see §3.3 of Fan and Gijbels, 1996). However, the latter has one more parameter, which reduces modeling biases, particularly at boundary regions. This is the theoretical background for our recommendation to use odd order fits. It is indeed an odd world!

The following lemma is very useful for deriving uniform convergence of the local polynomial estimator. It is an extension of a result due to Mack and Silverman (1982).

Lemma 6.1 *Let* $(X_1, Y_1), \cdots, (X_T, Y_T)$ *be a stationary sequence satisfying the mixing condition* $|\alpha(\ell)| \le c\ell^{-\beta}$ *for some* $c > 0$ *and* $\beta > 5/2$. *Assume further that for some* $s > 2$ *and interval* $[a, b]$,

$$E|Y|^s < \infty \quad \textit{and} \quad \sup_{x\in[a,b]}\int |y|^s f(x, y)dy < \infty,$$

where f *denotes the joint density of* (X, Y). *In addition, we assume that Conditions 1 (ii) and (iii) in §6.6.2 hold. Let* K *be a bounded function with*

a bounded support, satisfying the Lipschitz condition. Then

$$\sup_{x\in[a,b]} |T^{-1}\sum_{t=1}^{T}\{K_h(X_t-x)Y_t-E[K_h(X_t-x)Y_t]\}| = O_P[\{Th/\log(T)\}^{-1/2}],$$

provided that $h \to 0$, *for some* $\delta > 0$, $T^{1-2s^{-1}-2\delta}h \longrightarrow \infty$ *and*

$$T^{(\beta+1.5)(s^{-1}+\delta)-\beta/2+5/4}h^{-\beta/2-5/4} \to 0.$$

Note that since $T^{1-2s^{-1}-2\delta}h \longrightarrow \infty$, when the mixing coefficient is exponentially decays, the last condition of Lemma 6.1 holds automatically. In general, when β is sufficiently large, the last condition in the lemma above will hold.

We now state and prove the uniform convergence result for the local polynomial estimator.

Theorem 6.5 *Suppose that the conditions of Lemma 6.1 hold and the design density* f *is uniformly continuous on* $[a,b]$ *with* $\inf_{x\in[a,b]} f(x) > 0$. *Then*

$$\begin{aligned} &\sup_{x\in[a,b]}\left[\text{diag}(1,\cdots,h^p)\{\widehat{\beta}(x)-\beta_0(x)\}-\frac{h^{p+1}m^{(p+1)}(x)}{(p+1)!}\mathbf{S}^{-1}\mathbf{c}_p\right] \\ = \;& O_P[\{Th/\log(1/h)\}^{-1/2}]. \end{aligned}$$

By taking the $(\nu+1)$th element from Theorem 6.5, we have

$$\begin{aligned} &\sup_{x\in[a,b]}\left|\widehat{m}_\nu(x)-m^{(\nu)}(x)-\int t^{p+1}K_\nu^*(t)dt\frac{\nu!m^{(p+1)}(x)}{(p+1)!}h^{p+1-\nu}\right| \\ = \;& O_P[\{Th^{2\nu+1}/\log(T)\}^{-1/2}]. \end{aligned}$$

In particular, the local polynomial estimator has the following uniform convergence:

$$\sup_{x\in[a,b]}|\widehat{m}(x)-m(x)| = O_P[h^{p+1}+\{Th/\log(1/h)\}^{-1/2}].$$

6.3.4 Standard Errors and Estimated Bias

The standard errors for local polynomial estimators are useful for constructing confidence intervals. To derive them, let us temporarily assume that $\{(X_i,Y_i)\}$ are an independent sample from a population. Then, from (6.23),

$$\text{Var}(\widehat{\beta}|\mathbf{X}) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\text{Var}(\mathbf{y}|\mathbf{X})\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}.$$

Note that $\mathrm{Var}(Y_i|X_i) = \sigma^2(X_i)$. Since all operations are conducted locally for $X_i \approx x_0$, the conditional variance above is nearly constant, $\sigma^2(x_0)$. Using this local homoscedasticity, we have

$$\mathrm{Var}(\mathbf{y}|\mathbf{X}) = \mathrm{diag}\Big(\sigma^2(X_1), \cdots, \sigma^2(X_n)\Big) \approx \sigma^2(x_0) I_n.$$

This approximation holds of course only for those $X_i \approx x_0$, but those are exactly the data points involved in calculating the variance. Using this, we have

$$\mathrm{Var}(\widehat{\beta}|\mathbf{X}) \approx \sigma^2(x_0)(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}.$$

The conditional variance $\sigma^2(x_0)$ can be estimated by smoothing using a pilot bandwidth h^* and the square residuals $\{(X_t, \widehat{\varepsilon}_t^2)\}$, where $\widehat{\varepsilon}_t = Y_t - \widehat{m}(X_t)$. This results in an estimate of the covariance matrix

$$\widehat{\Sigma}(x_0) = \widehat{\sigma}^2(x_0)(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}. \tag{6.34}$$

This is a *preasymptotic substitution* method for estimating conditional variance, proposed in Fan and Gijbels (1995). In contrast, many authors use an *asymptotic substitution* method, which substitutes estimates into asymptotic expressions such as (6.31) and (6.32). This not only creates more unknown functions to estimate but also decreases the accuracy of the estimate.

Recall the definition of β_0 in Theorem 6.3. Following the same argument as above, the bias of the local polynomial estimator for an independent sample is

$$E(\beta|\mathbf{X}) - \beta_0 = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{r},$$

where $\mathbf{r} = \mathbf{m} - \mathbf{X}\beta_0$, with the ith element given by

$$\begin{aligned} r_i &= m(X_i) - \sum_{j=1}^{p} \frac{m^{(j)}(x_0)}{j!}(X_i - x_0)^j \\ &= \frac{m^{(p+1)}(x_0)}{(p+1)!}(X_i - x_0)^{p+1} + \frac{m^{(p+2)}(x_0)}{(p+2)!}(X_i - x_0)^{p+2}. \end{aligned}$$

The *preasymptotic substitution* method of Fan and Gijbels (1995) is to estimate $m^{(p+1)}(x_0)$ and $m^{(p+2)}(x_0)$ first by using a local polynomial fit with order $p+2$ and the pilot bandwidth h^*. This gives an estimate of $\mathbf{r}$ and the estimated bias vector

$$\widehat{\mathrm{Bias}}(x_0) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\widehat{\mathbf{r}}. \tag{6.35}$$

For dependent data, the arguments above do not hold. However, as demonstrated in §5.3, the local data behave very much like local independent data. Thus, the estimates (6.34) and (6.35) give a consistent estimate

for the asymptotic bias and asymptotic variance under some mixing conditions. Indeed, by using (6.28) and a similar expression for the kernel K^2, one can easily show that the bias and variance estimators are consistent.

The bias of $\widehat{m}^{(\nu)}(x_0)$ is estimated by the $(\nu+1)$-element of $\widehat{\text{Bias}}(x_0)$, denoted by $\widehat{B}_\nu(x_0)$. Similarly, the $(\nu+1)$ diagonal element of $\widehat{\Sigma}(x_0)$, denoted by $\widehat{V}_\nu(x_0)$, is the estimated variance of $\widehat{m}^{(\nu)}(x_0)$. By using Theorem 6.3, an approximate $(1-\alpha)$ pointwise confidence interval for $m^{(\nu)}(x_0)$ is

$$\widehat{m}_\nu(x_0) - \widehat{B}_\nu(x_0) \pm z_{1-\alpha/2}\widehat{V}_\nu(x_0)^{1/2}, \tag{6.36}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

The estimated bias involves estimation of higher-order derivatives, which usually cannot be estimated well with moderate sample sizes. For this reason, the bias is often ignored in the construction of the confidence intervals. Some even argue that the confidence intervals in parametric models also ignore biases since parametric models are at best approximately correct. For simplicity, we will still call intervals (6.36) with $\widehat{B}(x_0) = 0$ pointwise confidence intervals. As an illustration, Figure 6.6 depicts the estimated regression functions $m_1(x_0) = E(X_{t+1}|X_t = x_0)$ and $m_2(x_0) = E(X_{t+2}|X_t = x_0)$ and their associated pointwise confidence intervals.

6.3.5 Bandwidth Selection

As explained in §5.3, for stationary sequences of data under certain mixing conditions, state-domain smoothing performs very much like nonparametric regression for independent data because windowing reduces dependency among local data. Partially because of this, there are not many studies on bandwidth selection for state-domain smoothing problems. However, it is reasonable to expect the bandwidth selectors for independent data to continue to work for dependent data with certain mixing conditions. Below, we summarize a few useful approaches. When data do not have strong enough mixing, the general strategy is to increase bandwidth in order to reduce the variance.

Cross-validation is very useful for assessing the performance of an estimator via estimating its prediction error. The basic idea is to set one of the data points aside for validation of a model and use the remaining data to build the model. It is defined as

$$\text{CV}(h) = T^{-1}\sum_{i=1}^{T}\{Y_i - \widehat{m}_{h,-i}(X_i)\}^2, \tag{6.37}$$

where $\widehat{m}_{h,-i}$ is the local polynomial estimator (6.25) with $\nu = 0$ and bandwidth h, but without using the ith observation. The summand in (6.37) is a squared-prediction error of the ith data point using the training set $\{(X_j, Y_j) : j \neq i\}$. This cross-validation method uses ideas of Allen (1974)

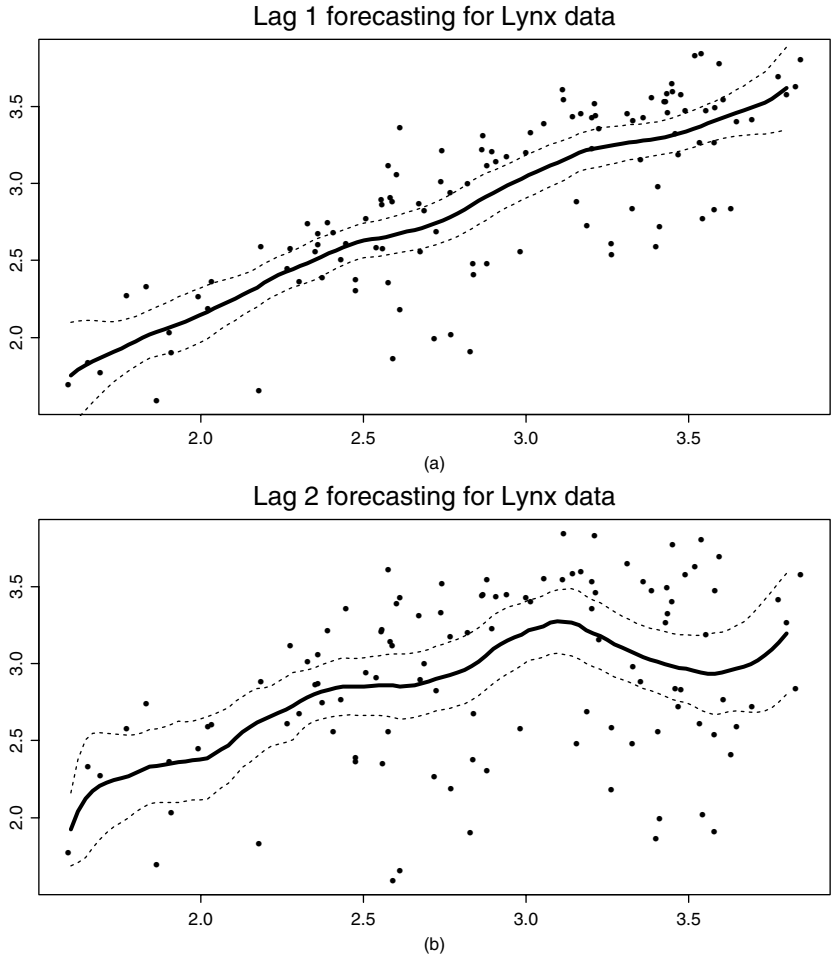


FIGURE 6.6. Local linear fits for the lynx data. (a) One-step prediction; (b) two-step forecasting. The dashed curves are the pointwise 95% confidence intervals.

and Stone (1974) and is computationally intensive. An improved version, in terms of computation, is the *generalized cross-validation* (GCV), proposed by Wahba (1977) and Craven and Wahba (1979). This criterion can be described as follows. By (6.25), the fitted values can be expressed as

$$(\widehat{m}(X_1), \cdots, \widehat{m}(X_T))^T = H(h)Y,$$

where $H(h)$ is a $T \times T$ hat matrix, depending on the X-variate and bandwidth h, and $Y = (Y_1, \cdots, Y_T)^T$. $H(h)$ is called a *smoothing matrix*. Then, the GCV approach selects the bandwidth h that minimizes

$$\text{GCV}(h) = [T^{-1}\text{tr}\{I - H(h)\}]^{-2}\,\text{MASE}(h), \tag{6.38}$$

where $\text{MASE}(h) = T^{-1}\sum_{i=1}^{T}\{Y_i - \widehat{m}(X_i)\}^2$ is the average of squared residuals.

A drawback of the cross-validation method is its inherent variability (see Hall and Johnstone 1992). Furthermore, it cannot be directly applied to select bandwidths for estimating derivative curves. Plug-in methods avoid these problems.

The basic idea is to find a bandwidth h that minimizes the estimated mean integrated square error (MISE). For the preasymptotic substitution method, the MISE is defined as

$$\widehat{\text{MISE}}(h) = \int \left\{ \widehat{B}_\nu(x)^2 + \widehat{V}_\nu(x) \right\} w(x)dx \tag{6.39}$$

for a given weight function w, where $\widehat{B}_\nu(x)$ and $\widehat{V}_\nu(x)$ are given in (6.36). This procedure was proposed by Fan and Gijbels (1995) and depends on the pilot bandwidth h^*. The pilot bandwidth may be selected by a *residual squares criterion* (RSC) proposed by Fan and Gijbels (1995). All automatic bandwidths in this book are selected by this method and implemented by using the C-code "lls.c". This includes bandwidth selection for spectral density estimation (§7.3) and estimation of conditional variance (§8.7).

The residual squares criterion is an automatic method for selecting a bandwidth. Suppose that we wish to select a bandwidth for estimating $m^{(\nu)}(\cdot)$ on an interval, based on the local polynomial fit of order p with odd $p - \nu$. Define

$$\text{RSC}(x_0; h) = \widehat{\sigma}^2(x_0)\{1 + (p+1)V\},$$

where V is the first diagonal element of the matrix

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{W}^2\mathbf{X})(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}.$$

By (6.34), V is the variance reduction of estimator $\widehat{m}(x_0)$, and hence V^{-1} is the effective number of local data points. When h is small, V is expected to be large, and when h is large, $\widehat{\sigma}^2(x_0)$ is large if the bias of the local fit is large. Thus, the RSC compromises these two contradictory demands. Let

$$\text{IRSC}(h) = \int_{[a,b]} \text{RSC}(x; h)dx$$

be the integrated version of the RSC over the interval $[a, b]$. In practice, the integration is replaced by summation over a fine grid of points. Denote the minimizer of $\text{IRSC}(h)$ by $\widehat{h}$. This bandwidth works reasonably in practice. To obtain the optimal bandwidth, some adjustments are needed. Set $C_p = \mu_{2p+2} - c_p^T S^{-1} c_p$, and let

$$\text{adj}_{\nu,p} = \left[\frac{(2\nu+1)C_p \int K_\nu^{*2}(t)dt}{(p+1-\nu)\{\int t^{p+1}K_\nu^*(t)dt\}^2 \int K_0^{*2}(t)dt} \right]^{1/(2p+3)}.$$

This adjusting constant depends on the kernel K and is slightly smaller than 1. The residual squares criterion selects the bandwidth

$$\widehat{h}_{\nu,p}^{\mathrm{RSC}} = \mathrm{adj}_{\nu,p}\widehat{h}.$$

More details can be found in Fan and Gijbels (1995).

The plug-in method of Ruppert, Sheather and Wand (1995) is an asymptotic substitution method. It estimates the derivative function $m^{(p+1)}(x)$, conditional variance $\sigma^2(x)$, and design density $f(x)$ first and then substitutes them into the asymptotic bias and variance expressions. The bandwidth is selected to minimize the estimated MISE. Pilot bandwidths are also needed for this procedure.

The *empirical bias* method, proposed by Ruppert (1997), relies on a different estimation of bias. The bias is estimated empirically by calculating $\widehat{m}_\nu(x_0;h)$ at a grid of h values and then modeling it as a function of h. Let $J_b > 1$ be an integer, and let $h_0^1, h_0^2, \cdots, h_0^{J_b}$ be in a neighborhood of h_0. Calculate $\widehat{m}_\nu(x_0;h_0^\ell)$ for $\ell = 1, \cdots, J_b$. Then, for some integer $a \geq 1$, fit the model

$$\mathrm{d}_0(x_0) + \mathrm{d}_{p+1-\nu}(x_0)h^{p+1-\nu} + \cdots + \mathrm{d}_{p+a-\nu}(x_0)h^{p+a-\nu} \tag{6.40}$$

to the synthetic data $\{(h_0^\ell, \widehat{m}_\nu(x_0;h_0^\ell)) : \ell = 1, \cdots, J_b\}$ using ordinary least-squares. The expression (6.40) is the asymptotic "expected value" of $\widehat{m}_\nu(x_0;h_0^\ell)$ and hence is a natural model to use. An estimator for the bias for estimating $m^{(\nu)}(x_0)$ is then

$$\widehat{\mathrm{d}}_{p+1-\nu}(x_0)h^{p+1-\nu} + \cdots + \widehat{\mathrm{d}}_{p+a-\nu}(x_0)h^{p+a-\nu}. \tag{6.41}$$

More details on bandwidth selections can be found in the references cited above. They can also be found in Chapter 4 of Fan and Gijbels (1996) and in Fan and Gijbels (2000).

6.4 Spline Methods

Spline methods are very useful for nonparametric modeling. They are based on global approximation and are useful extensions of polynomial regression techniques. A polynomial function, possessing all derivatives at all locations, is not very flexible for approximating functions with different degrees of smoothness at different locations. For example, the functions in Figure 6.3 and Figure 6.9(a) cannot be approximated very effectively by a polynomial function. One way to enhance the flexibility of the approximations is to allow the derivatives of the approximating functions to have discontinuities at certain locations. This results in piecewise polynomials called *splines.* The locations where the derivatives of the approximating functions may have discontinuities are called *knots*. Useful reference books on spline applications in statistics include Wahba (1990), Green and Silverman (1994), and Eubank (1999).

6.4.1 Polynomial Splines

As a brief introduction to spline methods, we use the state domain smoothing as the backlog. Let $t_1, \cdots, t_J$ be a sequence of given knots such that $-\infty < t_1 < \cdots < t_J < +\infty$. These knots can be chosen either by data analysts or data themselves. A spline function of order p is a $(p-1)$ continuously differentiable function such that its restriction to each of the intervals $(-\infty, t_1]$, $[t_1, t_2]$, $\cdots$, $[t_{J-1}, t_J]$, $[t_J, +\infty)$ is a polynomial function of order p. Any spline function $s(x)$ of order p with knots $t_1, \cdots, t_J$ can be represented as

$$s(x) = \sum_{j=1}^{J+p+1} \beta_j S_j(x), \tag{6.42}$$

where

$$\begin{cases} S_j(x) = (x - t_j)_+^p, & j = 1, \cdots, J, \\ S_{J+j}(x) = x^{j-1}, & j = 1, \cdots, p+1. \end{cases} \tag{6.43}$$

In other words, the space of all spline functions with knots $t_1, \cdots, t_J$ is a $(J + p + 1)$-dimensional linear space, and the functions $\{S_j(x)\}$ are a *basis* of this linear space, called the *power basis*. The power spline basis has the advantage that deleting a term $S_j(x)(j \le J)$ in (6.42) is the same as deleting a knot. However, as shown in Figure 6.7, the power spline basis may have large multiple correlation coefficients and could result in a nearly degenerate design matrix. Another commonly used basis is the *B-spline* basis (see p. 108 of de Boor (1978) for the definition), which is usually numerically more stable (see Figure 6.7(b)). However, deleting a term from this basis does not correspond to deleting a knot.

Frequently, cubic splines are used in practice. To facilitate the presentation, from now on, we focus on the cubic spline approximation. Substituting (6.42) into (6.18), we have

$$Y_t \approx \sum_{j=1}^{J+4} \beta_j S_j(X_t) + \sigma(X_t)\varepsilon_t.$$

Ignoring the heteroscedasticity, we estimate unknown parameters $\{\beta_j\}$ by minimizing

$$\min_{\beta} \sum_{t=1}^{T} \left\{ Y_t - \sum_{j=1}^{J+4} \beta_j S_j(X_t) \right\}^2. \tag{6.44}$$

Let $\widehat{\beta}_j$ $(j = 1, \cdots, J+4)$ be the least squares estimate. Then, the regression function is estimated by the spline function $\widehat{m}(x) = \sum_{j=1}^{J+4} \widehat{\beta}_j S_j(x)$. This is a cubic spline function since it is a linear combination of the cubic spline basis (6.43).

The polynomial spline method above is sensitive to the choice of knots $\{t_j\}$. One method of choosing the knots automatically is to initially place

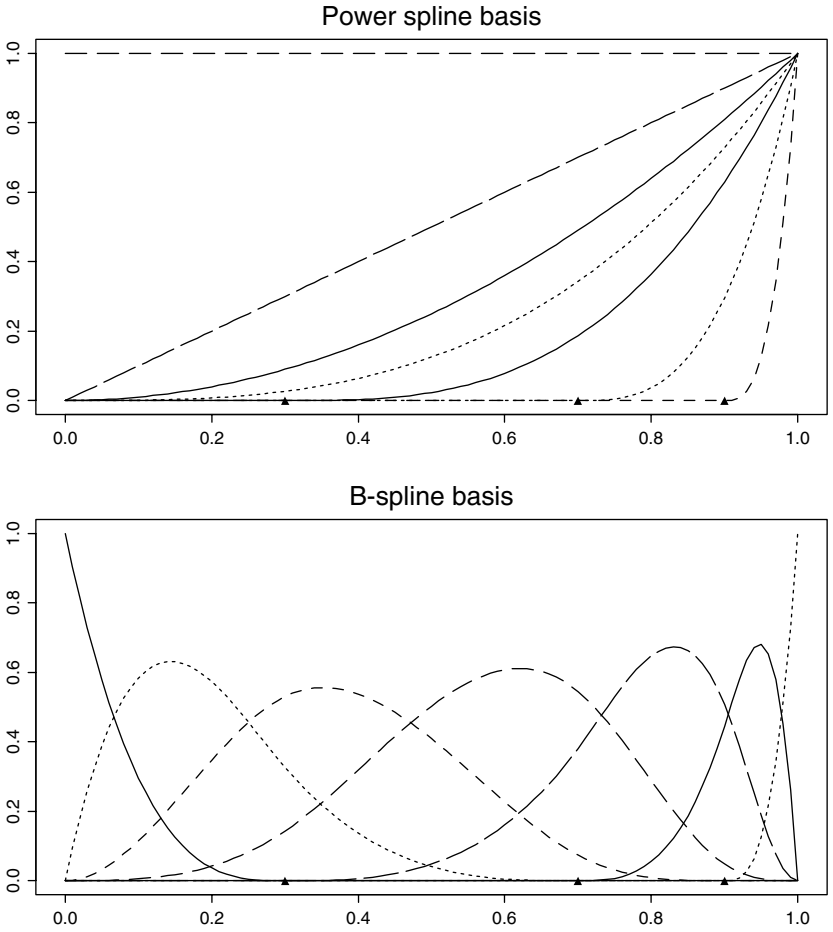


FIGURE 6.7. The power spline basis and B-spline basis for cubic splines with knots 0.3, 0.7 and 0.9. Any cubic spline functions with knots 0.3, 0.7, and 0.9 must be a linear combination of these basis functions.

many knots that might be deleted in the *knot selection* process. These knots are often placed at the order statistics of the X-variable. An example of the initial knots is $t_j = X_{(3j)}, j = 1, \cdots, [T/3]$. One can now treat problem (6.44) as an ordinary least-squares problem and apply linear regression techniques to select "significant variables" among the basis functions $\{S_j(x)\}$. Hence, the knots are selected.

We now briefly describe the *stepwise deletion* method. Let $\widehat{\beta}_j$ be the least-squares estimate resulting from (6.44) and $SE(\widehat{\beta}_j)$ be its estimated standard error. Then, delete the j_0th knot ($1 \leq j_0 \leq J$) having the smallest absolute t-statistic: $|\widehat{\beta}_j|/SE(\widehat{\beta}_j)$ $(1 \leq j \leq J)$. Repeat the process above

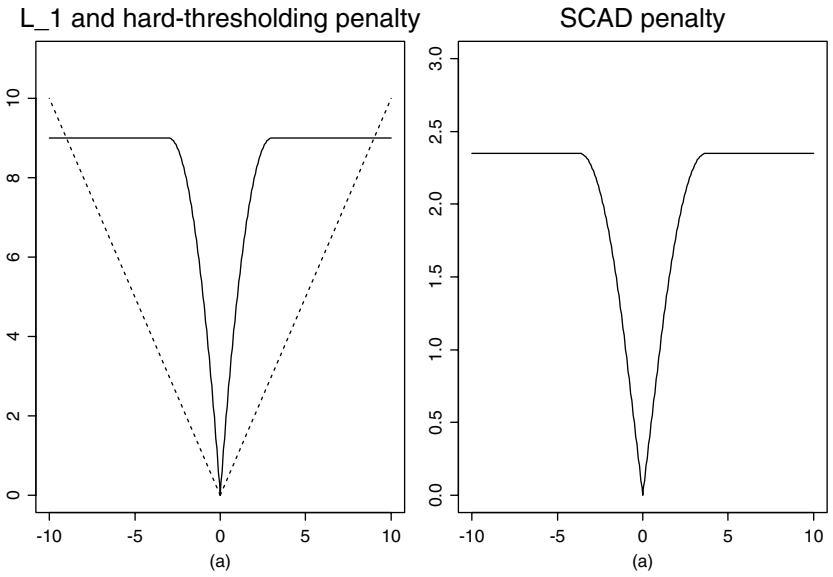


FIGURE 6.8. Some commonly used penalty functions. (a) L_1-penalty function (dashed curve) with $\omega = 1$ and the hard-thresholding penalty with $\omega = 3$; (b) smoothly clipped absolute deviation (SCAD) penalty with $\omega = 1$.

(delete one knot at each step). We obtain a sequence of models indexed by j $(0 \le j \le J)$: the jth model contains $J + 4 - j$ free parameters with residual sum of squares RSS_j. Choose the model $\widehat{j}$ that minimizes the modified Mallows C_p criterion (see Mallows 1973),

$$C_j = \mathrm{RSS}_j + \alpha(J + 4 - j)\widehat{\sigma}^2, \tag{6.45}$$

where $\widehat{\sigma}$ is the estimated standard deviation of the initial model (full model) and α is a smoothing parameter. Kooperberg and Stone (1991) recommend using $\alpha = 3$ instead of the more traditional value $\alpha = 2$ in *Akaike's information criterion* (AIC) Akaike 1970), while Schwarz (1978) recommends using $\alpha = \log T$ in a different context. These kinds of knot selection ideas are often employed by Stone and his collaborators; see, for example, Stone, Hansen, Kooperberg, and Truong (1997) and the references therein.

6.4.2 Nonquadratic Penalized Splines

One drawback of the stepwise approach in knot selection above is the stochastic noise accumulation in the variable selection process. The selected model is not chosen from all possible submodels, and the sampling properties of the procedures are hard to understand. To overcome these drawbacks, Fan and Li (2001) propose the following *penalized least-squares* method. Let s_j be the standard deviation of $\{S_j(X_t) : t = 1, \cdots, T\}$. This

represents the scale of $S_j(\cdot)$. Let $p_\omega(|\theta|)$ be a penalty function with singularities at 0 (Figure 6.8), where ω is a smoothing parameter. Examples of *penalty functions* include

$$\begin{array}{lcll} p_\omega(|\theta|) & = & \omega|\theta|, & L_1\text{-penalty}, \\ p_\omega(|\theta|) & = & \omega^2 - \{(|\theta| - \omega)_+\}^2, & \text{HT-penalty}, \\ p'_\omega(\theta) & = & \omega\{I(\theta \le \omega) + \frac{(a\omega-\theta)_+}{(a-1)\omega} I(\theta > \omega)\}, & \\ & & \text{with } a = 3.7, \text{ for } \theta > 0 & \text{SCAD-penalty}, \end{array} \tag{6.46}$$

which are called L_1-, hard thresholding (HT), and the smoothly clipped absolute deviation (SCAD) penalties, respectively. Note that the SCAD penalty is smoother than the hard-thresholding function. The latter produces discontinuous solutions, resulting in unstable models. The procedure with the L_1-penalty is called LASSO by Tibshirani (1996). It creates unnecessary bias for large coefficients $\widehat{\beta}$. See Antoniadis and Fan (2001) for a more detailed discussion, where necessary conditions are given for the penalized least-squares estimator to have certain properties such as sparsity, continuity, and unbiasedness. Our favorable choice of penalty function is the SCAD, which was derived to overcome the drawbacks of the hard and L_1-penalty functions.

To account for different scales of the basis function $\{S_j(X_t)\}$ for different j, we normalize it by its standard deviation. The *penalized least-squares* method minimizes

$$\sum_{t=1}^{T} \{Y_t - \sum_{j=1}^{J+4} \beta_j S_j(X_t)/s_j\}^2 + \sum_{j=1}^{J+4} p_\omega(|\beta_j|) \tag{6.47}$$

with respect to β. Fan and Li (2001) give an extension of the Newton–Raphson algorithm for optimizing (6.47), and Tibshirani (1996) and Fu (1998) propose two different algorithms for the LASSO. The estimated regression function is

$$\widehat{m}(x) = \sum_{j=1}^{J+4} \widehat{\beta}_j S_j(X_t)/s_j, \tag{6.48}$$

where $\{\widehat{\beta}_j\}$ are the estimated coefficients. Many of these estimated coefficients will be zero, according to Fan and Li (2001). Hence, only a subset of the basis functions will be selected. Due to penalty functions in (6.47), the choice of J can be very large and the problem is still not degenerate.

For the finite-dimensional problems, Fan and Li (2001) show that the penalized least-squares estimator possesses the following *oracle properties.* Suppose that there are $(p + q)$ unknown parameters, p of them zero and q nonzero unknown parameters. The penalized least-squares estimator will estimate those p zero coefficients as zero with probability tending to 1 and those q nonzero coefficients as efficiently as the q-dimensional submodel. In

other words, the penalized least-squares estimator performs as well as an oracle who knows in advance which coefficients are zero and which are not.

6.4.3 Smoothing Splines

A different approach to the regression spline and penalized least-squares method above is the *smoothing spline* method. The basic idea is to find a smooth function that minimizes the residual sum of squares. A popular measure of roughness of a function m is $\|m''(x)\|_2^2$. By the Lagrange multiplier method, minimizing the RSS subject to the roughness constraint is equivalent to the following penalized least-squares problem: minimizing with respect to m,

$$\sum_{t=1}^{T}\{Y_t - m(X_t)\}^2 + \omega \int \{m''(x)\}^2 dx, \tag{6.49}$$

where ω is a smoothing parameter (Lagrange's multiplier). It is clear that $\omega = 0$ corresponds to interpolation, whereas $\omega = +\infty$ results in a linear regression $m(x) = \alpha + \beta x$. As ω ranges from zero to infinity, the estimate ranges from the most complex model (interpolation) to the simplest model (linear model). Thus, the model complexity of the smoothing spline approach is effectively controlled by the smoothing parameter ω. The estimator $\widehat{m}_\omega$ is a spline function and is referred to as a smoothing spline estimator. It admits a Bayesian interpretation (Good and Gaskins 1971; Wahba 1978). Figure 6.9 illustrates the method using the environmental data given in Example 1.5. The method is first applied to the time domain smoothing. It is clear that the cross-validation method gives too small a bandwidth for the time domain smoothing (Altman 1990; Chu and Marron 1991a). The smoothing method is then applied to study the association between the pollutant NO_2 and the number of hospital admissions. The cross-validation method in this state-domain smoothing gives about the right amount of smoothing.

It is well-known that a solution to the minimization of (6.49) is a cubic spline. All possible knots are the data points $\{X_1, \cdots, X_T\}$. By using a spline basis expansion (e.g., B-spline basis),

$$m(x) = \sum_{j=1}^{T+4} \beta_j S_j(x)$$

and $\|m''\|_2^2$ is a quadratic function in $\{\beta_j\}$. Hence, the problem (6.49) is really the same as the penalized least-squares estimator with a quadratic penalty. As a result, many estimated parameters are shrunk toward zero but are not exactly zero. Moreover, since $\{\widehat{\beta}_j\}$ are linearly in the responses,

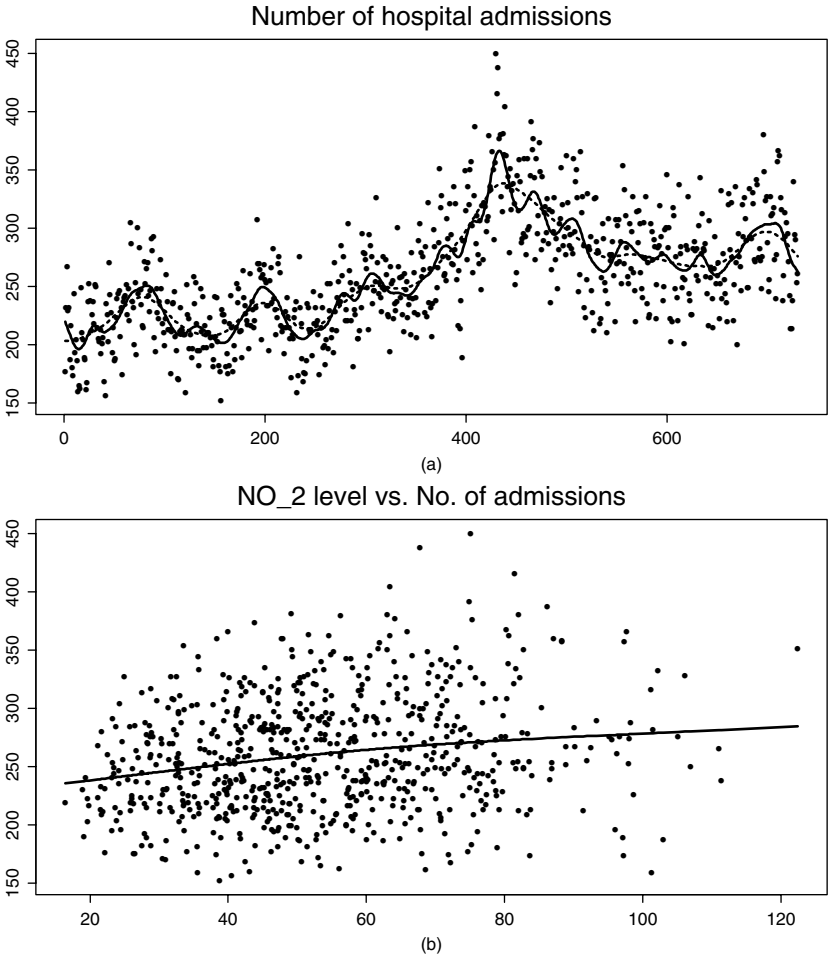


FIGURE 6.9. Nonparametric regression using the smoothing spline method. (a) Estimated trend with smoothing parameter chosen by the cross-validation method (solid curve) and $\omega = 0.0001$ (dashed curve). (b) Estimated regression function between the pollutant NO_2 and the number of hospital admissions.

so is $\widehat{m}_\omega = \sum_{j=1}^{T+4} \widehat{\beta}_j S_j(x)$. Hence, it can be expressed as

$$\widehat{m}_\omega(x) = n^{-1} \sum_{i=1}^{n} W_i(x, \omega; X_1, \cdots, X_n) Y_i, \tag{6.50}$$

where the weight W_i does not depend on the response $\{Y_i\}$. Hence, one can use the GCV (6.38) to select the smoothing parameter ω.

There are strong connections between kernel regression and smoothing splines; see Speckman (1981), Cox (1984) and Silverman (1984, 1985). In

particular, for independent data, Silverman (1984) shows that the smoothing spline is basically a local kernel average with a variable bandwidth. For X_i away from the boundary, and for T large and ω relatively small compared with T,

$$W_i(x, \omega; X_1, \cdots, X_n) \approx f(X_i)^{-1} h(X_i)^{-1} K_s\{(X_i - x)/h(X_i)\}, \quad (6.51)$$

where $h(X_i) = [\omega/\{nf(X_i)\}]^{1/4}$ and

$$K_s(t) = 0.5 \exp(-|t|/\sqrt{2}) \sin(|t|/\sqrt{2} + \pi/4).$$

This approximation is also valid for calculating the mean and variance of the smoothing spline estimator (see Messer 1991).

6.5 Estimation of Conditional Densities

Conditional densities provide a very informative summary of a random variable Y given $X = x$. The mean regression $m(x) = E(Y|X = x)$ is the "center" of this distribution. The conditional standard deviation $\sigma(x)$ provides the likely size with which the random variable Y would deviate away from the conditional mean. Furthermore, the *conditional density* allows us to examine the overall shape of the conditional distribution. In the context of the k-step forecasting, one takes $X = X_t$ and $Y = Y_{t+k}$. The "center" of the conditional density, the mean regression function $m(x)$, provides the predicted value, and the spread of the conditional distribution, the conditional standard deviation $\sigma(x)$, indicates the likely size of the prediction error.

6.5.1 Methods of Estimation

As indicated in Example 6.2, we can estimate the conditional density $f(y|x)$ in the same way as the conditional mean. This can be seen via

$$\begin{aligned} & E\{(2h_2)^{-1} I(|Y - y| \le h_2)|X = x\} \\ = \; & (2h_2)^{-1}\{F(y + h_2|x) - F(y - h_2|x)\} \approx f(y|x) \end{aligned} \quad (6.52)$$

as $h_2 \to 0$, where $F(y|x)$ is the cumulative conditional distribution of Y given $X = x$. Expression (6.52) utilizes the uniform kernel. In general, by Lemma 5.1,

$$E\{K_{h_2}(Y - y)|X = x\} \approx f(y|x) \quad \text{as } h_2 \to 0 \quad (6.53)$$

for a given probability density function K. This leads to the nonparametric regression of the synthetic data $K_{h_2}(Y - y)$ on X.

Let $(X_1, Y_1), \cdots, (X_T, Y_T)$ be a stationary sequence. Applying the local polynomial technique to the synthetic data $\{(X_t, K_{h_2}(Y_t - y))\}$ leads to

$$\sum_{t=1}^{T} \left\{ K_{h_2}(Y_t - y) - \sum_{j=0}^{p} \beta_j (X_t - x)^j \right\}^2 W_{h_1}(X_t - x), \tag{6.54}$$

where $W(\cdot)$ is a kernel function. Let $\widehat{\beta}_j(x, y)$ be the solution to the least squares problem (6.54). Then, from §6.3.2, it is clear that $f^{(\nu)}(y|x) = \frac{\partial^\nu f(y|x)}{\partial x^\nu}$ can be estimated as

$$\widehat{g}_\nu(y|x) = \frac{\widehat{\partial^\nu g(y|x)}}{\partial x^\nu} = \nu! \widehat{\beta}_\nu(x, y). \tag{6.55}$$

By (6.25), the estimator can be expressed as

$$\widehat{f}_\nu(y|x) = \nu! \sum_{t=1}^{T} W_\nu^T\{(X_t - x)/h_1\} K_{h_2}(Y_t - y). \tag{6.56}$$

We rewrite $\widehat{f}_0(\cdot|\cdot)$ as $\widehat{f}(\cdot|\cdot)$. This idea was due to Fan, Yao, and Tong (1996).

By (6.56), when $K(\cdot)$ has zero mean, one can easily see that

$$\int_{-\infty}^{+\infty} y \widehat{f}_\nu(y|x) dy = \nu! \sum_{t=1}^{T} W_\nu^T\{(X_t - x)/h_1\} Y_t = \widehat{m}_\nu(x).$$

Thus, the local polynomial estimation of the mean regression function is simply the mean of the estimated conditional density $\widehat{f}_\nu(y|x)$.

Estimating a bivariate conditional density is computationally intensive. The simultaneous choice of bandwidths h_1 and h_2 can be difficult. A simple rule of thumb is as follows. Choose h_2 by the normal reference method (5.8) and (5.9). Once h_2 is chosen, the problem (6.54) is a standard local polynomial regression problem. Thus, one can use a bandwidth selection method outlined in §6.3.5 to choose an h_1. In the implementation below, the preasymptotic substitution method of Fan and Gijbels (1995) will be used.

As an illustration, we draw a random sample from

$$X_t = 0.23 X_{t-1}(16 - X_{t-1}) + 0.4\varepsilon_t \quad t \geq 1, \tag{6.57}$$

where $\{\varepsilon_t\}$ are independent random variables having the same distribution as the sum of 48 independent random variables, each distributed uniformly on $[-0.25, 0.25]$. By the central limit theorem, ε_t can effectively be treated as a standard normal variable. However, it has bounded support, which is necessary for the stationarity of the time series (see Chan and Tong 1994). The skeleton of model (6.57) appears chaotic; see the top panel of

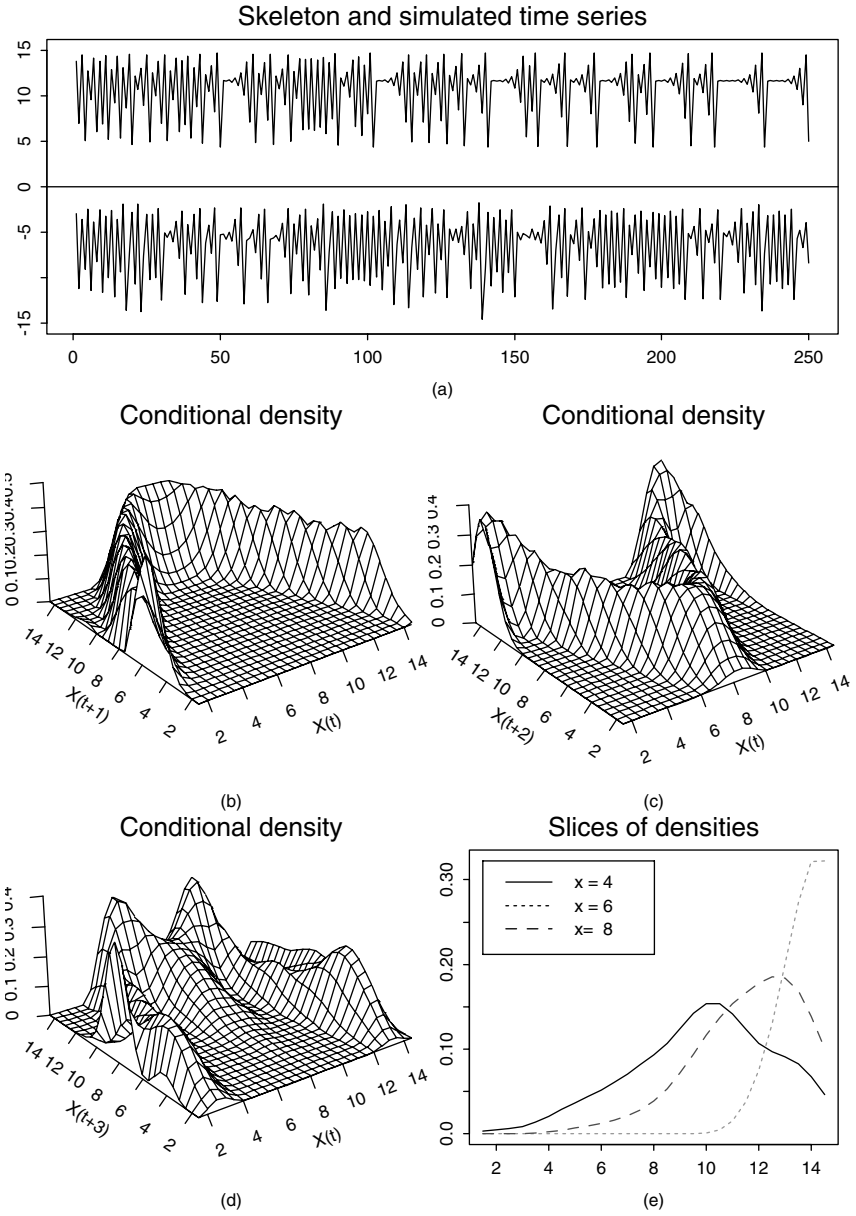


FIGURE 6.10. (a) Time series plot of the skeleton $x_t = 0.23x_{t-1}(16 - x_{t-1})$ with $x_0 = 10$ (top panel) and a simulated series from model (6.57) (bottom panel). (b)–(d) are, respectively, the one-step, two-step, and three-step forecasting conditional densities. (e) Conditional densities of X_{t+3} given $X_t = 4, 6$, and 8. Adapted from Fan, Yao, and Tong (1996).

Figure 6.10(a). The bottom panel of Figure 6.10(a) shows a typical simulated time series. The conditional densities for the one-step-, two-step-, and three-step-ahead predictions are depicted in Figures 6.10 (b)–(e).

For the one-step ahead prediction, the conditional density is approximately normal with constant variance (see (6.57)). This is consistent with the shape shown in Figure 6.10(b). The quadratic ridge is the conditional mean regression function. For the two-step and the three-step forecasts, the true conditional density is hard to derive. Nevertheless, our method is able to estimate their conditional densities. The two-step forecasting densities appear unimodal for all x, whereas the shape of three-step forecasting densities is hard to determine. To help us examine the conditional density, we plot a few slices from Figure 6.10(d). It appears that these conditional densities are unimodal, but their variances are quite different. This noise amplification will be further discussed in Chapter 10. An advantage of the conditional densities approach is that it is more informative, indicating not only the predicted value but also the likely size of prediction errors.

6.5.2 Asymptotic Properties*

We now summarize some asymptotic theory derived in Fan, Yao, and Tong (1996). The technical device is similar to that used in Theorem 6.3. For simplicity, we only discuss the two most useful cases: $p = 1, \nu = 0$ (estimating the conditional density function) and $p = 2, \nu = 1$ (estimating the partial derivative function). Let $\mu_K = \int t^2 K(t)dt$, $\nu_K = \int \{K(t)\}^2 dt$, $\mu_j = \int t^j W(t)dt$, and $\nu_j = \int t^j \{W(t)\}^2 dt$.

Theorem 6.6 *For the local linear fit, under Condition 2 in §6.6.5, we have*

$$\sqrt{Th_1h_2}\{\widehat{f}(y|x) - f(y|x) - \vartheta_{T,0}\} \xrightarrow{D} N\left(0,\ \sigma_0^2\right),$$

provided that the bandwidths h_1 and h_2 converge to zero in such a way that $Th_1h_2 \to \infty$, where

$$\begin{aligned} \vartheta_{T,0}(x,y) &= \frac{h_1^2\mu_2}{2}\frac{\partial^2 f(y|x)}{\partial x^2} + \frac{h_2^2\mu_K}{2}\frac{\partial^2 f(y|x)}{\partial y^2} + o(h_1^2 + h_2^2), \\ \sigma_0^2(x,y) &= \nu_K\nu_0\frac{f(y|x)}{f(x)}, \end{aligned}$$

where $f(x)$ is the marginal density of X_1.

Theorem 6.7 implies that the rate of convergence of the conditional density is of order

$$O_P\left(h_1^2 + h_2^2 + \frac{1}{\sqrt{Th_1h_2}}\right).$$

By taking $h_1 = h_2 = O(T^{-1/6})$, one obtains the optimal rate $O_P(T^{-1/3})$. To state the results for the local quadratic fit (mainly used for derivative

estimation), we denote

$$\begin{aligned}
\vartheta_{T,1}(x,y) &= \frac{1}{2}\mu_K \frac{\partial^2 f(y|x)}{\partial y^2} h_2^2 + o(h_1^3 + h_2^2), \\
\sigma_1^2(x,y) &= \frac{f(y|x)\nu_K}{f(x)} \frac{\mu_4^2\nu_0 - 2\mu_2\mu_4\nu_2 + \mu_2^2\nu_4}{(\mu_4 - \mu_2^2)^2}, \\
\vartheta_{T,2}(x,y) &= \frac{\mu_4}{6\mu_2} \frac{\partial^3 f(y|x)}{\partial x^3} h_1^2 + \frac{1}{2}\mu_K \frac{\partial^3 f(y|x)}{\partial x \partial y^2} h_2^2 + o(h_1^2 + h_2^2), \\
\sigma_2^2(x,y) &= \frac{f(y|x)\nu_K}{f(x)} \frac{\nu_0\nu_2}{\mu_2^2}.
\end{aligned}$$

Theorem 6.7 *Suppose that the bandwidths h_1 and h_2 converge to zero, that $Th_1^3h_2 \to \infty$, and that Condition 2 in §6.6.5 holds. For the local quadratic fit, we have*

$$\sqrt{Th_1h_2/2}\{\widehat{f}(y|x) - f(y|x) - \vartheta_{T,1}\} \xrightarrow{D} N(0, \sigma_1^2)$$

and

$$\sqrt{Th_1^3h_2/2}\left\{\widehat{f}_1(y|x) - \frac{\partial}{\partial x} f(y|x) - \vartheta_{T,2}\right\} \xrightarrow{D} N(0, \sigma_2^2).$$

Moreover, they are asymptotically jointly normal with covariance 0.

6.6 Complements

Throughout this section, we use C to denote a generic constant, which may vary from line to line.

6.6.1 Proof of Theorem 6.1

We first establish the bias expression. Since the kernel K has a bounded support, say $[-1, 1]$, the weight $w_{Tu,i}$ does not vanish only when $|i - Tu| \le h$. Since $h/T \to 0$, by (6.15) and Taylor's expansion, we have

$$E\widehat{g}(u) - g(u) = \frac{\sum_{i=1}^T w_{Tu,i} g''(\xi_i)(i/T - u)^2}{2\sum_{i=1}^T w_{Tu,i}},$$

where ξ_i lies between u and i/T. Hence, we have

$$\max_i |\xi_i - u| \le h/T \to 0. \tag{6.58}$$

Let

$$s_{T,j}(t) = \sum_{i=1}^T K_h(i-t)(i-t)^j g''(\xi_i).$$

Then, by the definition of the weight function $w_{t,i}$, we have

$$E\widehat{g}(u) - g(u) = T^{-2}\frac{S_{T,2}(Tu)s_{T,2}(Tu) - S_{T,1}(Tu)s_{T,3}(Tu)}{S_{T,0}(Tu)S_{T,2}(Tu) - S_{T,1}(Tu)^2}. \tag{6.59}$$

We now show that for all j

$$S_{T,j}(Tu) = h^j\mu_j + O(h^{j-1}) \tag{6.60}$$

and

$$s_{T,j}(Tu) = h^j g''(u)\mu_j + O(h^{j-1}). \tag{6.61}$$

Suppose that (6.60) and (6.61) hold. Substituting these two results into (6.59), the bias result follows from the fact that $\mu_1 = 0$. It remains to prove (6.60) and (6.61). Their arguments are quite similar, and hence we only prove (6.61). By approximating the discrete sum by its integral, using (6.58), we have

$$\begin{aligned} s_{T,j}(Tu) &= h^{j-1}\sum_{i=1}^{T} K(i/h - Tu/h)(i/h - Tu/h)^j g''(\xi_i) \\ &= h^j g''(u)\int_{-\infty}^{+\infty} K(v - Tu/h)(v - Tu/h)^j dv + O(h^{j-1}) \\ &= h^j g''(u)\mu_j + O(h^{j-1}). \end{aligned}$$

This completes the proof for part (a).

For part (b), denote

$$s^*_{T,i}(t) = \sum_{k=1}^{T} K_h(k-t)(k-t)^i\epsilon_k.$$

Then, we have

$$\widehat{g}(u) - E\widehat{g}(u) = \frac{S_{T,2}(Tu)s^*_{T,0}(Tu) - S_{T,1}(Tu)s^*_{T,1}(Tu)}{S_{T,0}(Tu)S_{T,2}(Tu) - S_{T,1}(Tu)^2}, \tag{6.62}$$

If we can show that

$$\begin{aligned} &\mathrm{Var}\{s^*_{T,i}(Tu)\} \\ &= \begin{cases} C_X\int\int G(x)G(y)|x-y|^{-\alpha}dxdyh^{2i-\alpha}, & \text{when } 0<\alpha<1 \\ 2C_X\|G\|_2^2 h^{2i-1}\log(h), & \text{when } \alpha = 1 \\ \sum_{j=-\infty}^{\infty}\gamma_X(j)\|G\|_2^2 h^{2i-1}, & \text{when } \alpha > 1, \end{cases} \end{aligned} \tag{6.63}$$

where $G(v) = v^iK(v)(i = 0, 1)$, then both terms in the numerator of (6.62) are of the same order. Since $\mu_1 = 0$, the second term is indeed of smaller

order. Similarly, by (6.60), the first term in the denominator of (6.62) dominates. Hence,

$$\mathrm{Var}\{\widehat{g}(u)\} = \mathrm{Var}\{s^*_{T,0}(Tu)\}\{1 + o(1)\},$$

and the result follows from (6.63).

To prove (6.63), let $V_T = \mathrm{Var}\{s^*_{T,i}(Tu)/h^i\}$. Since G has a bounded support, $G_h(j)$ vanishes when $|j| \geq h$. Using this, we have for $0 < u < 1$

$$V_T = \sum_{j,k} G_h(j - Tu)G_h(k - Tu)\gamma_X(j - k) = \sum_{j,k} G_h(j)G_h(k)\gamma_X(j - k).$$

By (6.14), for any $\varepsilon > 0$, there exists a large M such that for all $\ell \geq M$

$$(1 - \varepsilon)C_X\ell^{-\alpha} \leq \gamma_X(\ell) \leq (1 + \varepsilon)C_X\ell^{-\alpha}.$$

Write

$$V_T = \sum_k \left\{ \sum_{|\ell| \leq M} + \sum_{|\ell| > M} \right\} G_h(k)G_h(k + \ell)\gamma_X(\ell) \equiv I_1 + I_2.$$

Since $G_h(k)$ vanishes when $|k| \geq h$,

$$\begin{aligned} |I_1| &\leq \sum_k \sum_{|\ell| \leq M} G_h(k)G_h(k + \ell) \\ &\leq (2M + 1)(2h + 1)h^{-2} \max_v |G(v)|^2 \\ &\leq Ch^{-1}. \end{aligned}$$

By approximating the discrete sum by the continuous integral, we have for $0 < \alpha < 1$,

$$\begin{aligned} I_2 &\leq (1 + \varepsilon) \sum_k \sum_{|\ell| > M} G_h(k)G_h(k + \ell)C_X|\ell|^{-\alpha} \\ &= (1 + \varepsilon)C_X \int_{-\infty}^{+\infty} \int_{|y| \geq M} G_h(x)G_h(x + y)|y|^{-\alpha} dxdy\{1 + o(1)\}. \end{aligned}$$

A change of variable leads to

$$I_2 \leq (1 + \varepsilon)C_X h^{-\alpha} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} G(x)G(y)|x - y|^{-\alpha} dxdy\{1 + o(1)\}.$$

Therefore, letting $T \to \infty$ and then $\varepsilon \to 0$, we have

$$\limsup_{T \to \infty} h^\alpha V_T \leq C_X \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} G(x)G(y)|x - y|^{-\alpha} dxdy.$$

Using the same argument, we have

$$\liminf_{T\to\infty} h^{\alpha} V_T \geq C_X \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} G(x)G(y)|x-y|^{-\alpha} dxdy.$$

This proves (6.63) for the case $0 < \alpha < 1$.

When $\alpha = 1$, using arguments similar to those in the paragraph above, we have for each $c > 0$

$$\begin{aligned} V_T &= C_X \int_{-\infty}^{+\infty} \int_{|y|>c/h} G_h(x)G_h(x+y)|y|^{-1} dxdy\{1+o(1)\} \\ &= 2C_X \|G\|_2^2 h^{-1} \log(h)\{1+o(1)\}. \end{aligned}$$

This proves (6.63).

For the case $\alpha > 1$, write

$$V_T = \sum_{\ell=-\infty}^{\infty} \gamma_X(j) \sum_k G_h(k)G_h(k+\ell).$$

By approximating the discrete sum by an integral, we have for each given j

$$\sum_k G_h(k)G_h(k+j) = h^{-1}\|G\|_2^2\{1+o(1)\}.$$

Since $\sum_j |\gamma_X(j)| < \infty$, we have

$$V_T = h^{-1} \sum_{\ell=-\infty}^{\infty} \gamma_X(\ell)\|G\|_2^2 + o(h^{-1}).$$

This completes the proof. ∎

6.6.2 Conditions and Proof of Theorem 6.3

As explained in Figure 5.4, the conditions imposed on the mixing coefficient and bandwidth h should be related. This is more precisely described in Condition 1(iv) below.

Condition 1:

(i) The kernel K is bounded with a bounded support.

(ii) The conditional density $f_{X_0,X_\ell|Y_0,Y_\ell}(x_0, x_\ell|y_0, y_\ell) \leq A_1 < \infty, \forall \ell \geq 1$.

(iii) For ρ-mixing processes, we assume that

$$\sum_{\ell} \rho(\ell) < \infty, \quad EY_0^2 < \infty;$$

for α-mixing processes, we assume that for some $\delta > 2$ and $a > 1 - 2/\delta$,

$$\sum_{\ell} \ell^a [\alpha(\ell)]^{1-2/\delta} < \infty,\ E|Y_0|^\delta < \infty,\ f_{X_0|Y_0}(x|y) \le A_2 < \infty.$$

(iv) For ρ-mixing and strongly mixing processes, we assume, respectively, that there exists a sequence of positive integers satisfying $s_T \to \infty$ and $s_T = o\{(nh_T)^{1/2}\}$ such that

$$(n/h_T)^{1/2}\rho(s_T) \to 0 \quad \text{and} \quad (n/h_T)^{1/2}\alpha(s_T) \to 0 \text{ as } T \to \infty.$$

(v) $\sigma^2(\cdot)$ and $f(\cdot)$ are continuous at the point x and $f(x) > 0$.

Proof. Let $\mathbf{m} = \{m(X_1), \cdots, m(X_T)\}^T$ and $\beta_j = m^{(j)}(x)/j!$. Write

$$\begin{aligned} \widehat{\beta}(x) - \beta_0(x) &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\{\mathbf{m} - \mathbf{X}\beta_0(x)\} \\ &\quad +(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(\mathbf{y} - \mathbf{m}) \\ &\equiv \mathbf{b} + \mathbf{t}. \end{aligned} \tag{6.64}$$

The main idea is to show that the bias vector $\mathbf{b}$ converges in probability to a vector and that the centralized vector $\mathbf{t}$ is asymptotically normal.

We first establish the asymptotic behavior of the bias vector $\mathbf{b}$. By Taylor's expansion of $m(X_i)$ around the point x, we have

$$\mathbf{b} = \mathbf{S}_T^{-1}\{\beta_{p+1}(S_{T,p+1}, \cdots, S_{T,2p+1})^T + o_P(h^{p+1})\}, \tag{6.65}$$

where $\mathbf{S}_T = \mathbf{X}^T\mathbf{W}\mathbf{X}$ and $S_{T,j}$ is defined in (6.24). By (6.28), we have

$$\mathbf{b} = \beta_{p+1}(\mathbf{H}\mathbf{S}\mathbf{H})^{-1}\mathbf{H}\mathbf{c}_p h^{p+1}\{1 + o_P(1)\}, \tag{6.66}$$

with $\mathbf{H} = \text{diag}(1, h, \cdots, h^p)$.

We next consider the joint asymptotic normality of $\mathbf{t}$. By (6.28),

$$\mathbf{t} = f^{-1}(x)\mathbf{H}^{-1}\mathbf{S}^{-1}\mathbf{u}\{1 + o_P(1)\}, \tag{6.67}$$

where $\mathbf{u} = T^{-1}\mathbf{H}^{-1}\mathbf{X}^T\mathbf{W}(\mathbf{y} - \mathbf{m})$. Thus, we need to establish the asymptotic normality of $\mathbf{u}$. Consider an arbitrary linear combination $\mathbf{c}^T\mathbf{u}$. Simple algebra shows that

$$Q_T \equiv \mathbf{c}^T\mathbf{u} = \frac{1}{T}\sum_{i=1}^{T} Z_i, \tag{6.68}$$

where with $C(u) = \sum_{j=0}^{p} c_j u^j K(u)$ and $C_h(u) = C(u/h)/h$,

$$Z_i = \{Y_i - m(X_i)\}C_h(X_i - x).$$

The problem reduces to proving the asymptotic normality of Q_T.

We will show that

$$\sqrt{Th}Q_T \xrightarrow{D} N\left\{0, \theta^2(x)\right\}, \tag{6.69}$$

where

$$\theta^2(x) = \sigma^2(x)f(x)\int_{-\infty}^{+\infty} C^2(x)dx = \sigma^2(x)f(x)\mathbf{c}^T\mathbf{S}^*\mathbf{c}.$$

From this, it follows that

$$\sqrt{Th}\mathbf{u} \xrightarrow{D} N\{0, \sigma^2(x)f(x)\mathbf{S}^*\}.$$

Hence

$$\sqrt{Th}\mathbf{Ht} \xrightarrow{D} N\{0, \sigma^2(x)f^{-1}(x)\mathbf{S}^{-1}\mathbf{S}^*\mathbf{S}^{-1}\}.$$

Using this and (6.66), we obtain Theorem 6.3.

The proof of (6.69) requires some extra work. We divide the proof into two steps: computation of the variance of Q_T and showing the asymptotic normality of Q_T.

Computation of the variance of Q_T

Note that

$$\mathrm{Var}(Z_i) = \frac{1}{h}\left\{\theta^2(x) + o(1)\right\}. \tag{6.70}$$

By stationarity, we have

$$\mathrm{Var}(Q_T) = \frac{1}{T}\mathrm{Var}(Z_1) + \frac{2}{T}\sum_{\ell=1}^{T-1}(1-\ell/T)\mathrm{Cov}(Z_1, Z_{\ell+1}).$$

Let $d_T \to \infty$ be a sequence of integers such that $d_T h_T \to 0$. Define

$$J_1 = \sum_{\ell=1}^{d_T-1}|\mathrm{Cov}(Z_1, Z_{\ell+1})|, \quad J_2 = \sum_{\ell=d_T}^{T-1}|\mathrm{Cov}(Z_1, Z_{\ell+1})|.$$

Let $B = \max_{X\in x\pm h} m(X)$. By conditioning on (Y_1, Y_ℓ) and using Condition 1(ii), we obtain

$$\begin{aligned} & |\mathrm{Cov}(Z_1, Z_\ell)| \\ = \; & |E[\{Y_1 - m(X_1)\}\{Y_\ell - m(X_\ell)\}C_h(X_1 - x)C_h(X_\ell - x)]| \\ \leq \; & A_1 E\{(|Y_1| + B)(|Y_\ell| + B)\}\left(\int_{-\infty}^{+\infty}|C_h(u-x)|du\right)^2 \\ \leq \; & D, \end{aligned}$$

for some $D > 0$. It follows that $J_1 \leq d_T D = o(1/h_T)$. We now consider the contribution of J_2. For ρ-mixing processes, we have from (6.70) that

$$J_2 \leq \mathrm{Var}(Z_1)\sum_{\ell=d_T}^{\infty}\rho(\ell) = o(1/h_T).$$

For strongly mixing processes, we use Davydov's lemma (see Hall and Heyde 1980, Corollary A2) and obtain

$$|\mathrm{Cov}(Z_1, Z_{\ell+1})| \le 8[\alpha(\ell)]^{1-2/\delta}[E|Z_1|^\delta]^{2/\delta}.$$

By conditioning on Y_1 and using Condition 1(iii), we have

$$E|Z_1|^\delta \le A_2 E(|Y_1| + B)^\delta \int_{-\infty}^{+\infty} |C_h(x-u)|^\delta \le D h_T^{-\delta+1}$$

for some $D > 0$. Combining the last two inequalities leads to

$$\begin{aligned} J_2 &\le \delta D^{2/\delta} h_T^{2/\delta-2} \sum_{\ell=d_T}^{\infty} [\alpha(\ell)]^{1-2/\delta} \\ &\le \delta D^{2/\delta} h_T^{2/\delta-2} d_T^{-a} \sum_{\ell=d_T}^{\infty} \ell^a [\alpha(\ell)]^{1-2/\delta} \\ &= o(1/h_T) \end{aligned}$$

by taking $h^{1-2/\delta} d_T^a = 1$. This choice of d_T satisfies the requirement that $d_T h_T \to 0$. Using the properties of J_1 and J_2, we conclude that

$$\sum_{\ell=1}^{T-1} |\mathrm{Cov}(Z_1, Z_{\ell+1})| = o(1/h_T) \tag{6.71}$$

and that $Th_T \,\mathrm{Var}(Q_T) \to \theta^2(x)$.

Asymptotic normality of Q_T

The proof of this step is the same for ρ-mixing and strongly mixing processes. We only concentrate on ρ-mixing processes.

We employ so-called small-block and large-block arguments. Partition the set $\{1, \cdots, T\}$ into subsets with large blocks of size $r = r_T$ and small blocks of size $s = s_T$. A large block is followed by a smaller block. Let $k = k_T = [\frac{T}{r_T + s_T}]$ be the number of blocks. Let $Z_{T,t} = \sqrt{h} Z_{t+1}$. Then $\sqrt{Th} Q_T = T^{-1/2} \sum_{t=0}^{T-1} Z_{T,t}$. By (6.70) and (6.71),

$$\mathrm{Var}(Z_{T,0}) = \theta^2(x)\{1 + o(1)\}, \quad \sum_{t=1}^{T-1} |\mathrm{Cov}(Z_{T,0}, Z_{T,t})| = o(1). \tag{6.72}$$

Let the random variables η_j and ξ_j be the sum over the jth large-block and the jth small-block, respectively; that is,

$$\eta_j = \sum_{t=j(r+s)}^{j(r+s)+r-1} Z_{T,t}, \quad \xi_j = \sum_{t=j(r+s)+r}^{(j+1)(r+s)-1} Z_{T,t},$$

and $\zeta_k = \sum_{t=k(r+s)}^{T-1} Z_{T,t}$ be the sum over the residual block. Then

$$\begin{aligned}\sqrt{Th}Q_T &= \frac{1}{\sqrt{T}}\left\{\sum_{j=0}^{k-1}\eta_j + \sum_{j=0}^{k-1}\xi_j + \zeta_k\right\} \\ &\equiv \frac{1}{\sqrt{T}}\left\{Q'_T + Q''_T + Q'''_T\right\}.\end{aligned}$$

We will show that as $T \to \infty$,

$$\frac{1}{T}E(Q''_T)^2 \to 0, \quad \frac{1}{T}E(Q'''_T)^2 \to 0, \tag{6.73}$$

$$\left| E\left[\exp(itQ'_T)\right] - \prod_{j=0}^{k-1} E\left[\exp(it\eta_j)\right] \right| \to 0, \tag{6.74}$$

$$\frac{1}{T}\sum_{j=0}^{k-1} E\left(\eta_j^2\right) \to \theta^2(x), \tag{6.75}$$

$$\frac{1}{T}\sum_{j=0}^{k-1} E\left[\eta_j^2 I\{|\eta_j| \geq \varepsilon\theta(x)\sqrt{T}\}\right] \to 0, \tag{6.76}$$

for every $\varepsilon > 0$. Statement (6.73) implies that the sums over small and residual blocks Q''_T and Q'''_T are asymptotically negligible. Result (6.74) reveals that the summands in the large blocks $\{\eta_j\}$ in Q'_T are asymptotically independent, and (6.75) and (6.76) are the standard Lindberg–Feller conditions for the asymptotic normality of Q'_T under the independence assumption. Expressions (6.73)–(6.76) entail the asymptotic normality (6.69).

We now establish (6.73)–(6.76). We first choose the block sizes. Condition 1(iv) implies that there exist constants $q_T \to \infty$ such that

$$q_T s_T = o(\sqrt{Th}); \quad q_T(T/h)^{1/2}\alpha(s_T) \to 0.$$

Define the large block size $r_T = [(Th_T)^{1/2}/q_T]$. Then, it can easily be shown that

$$s_T/r_T \to 0, \quad r_T/T \to 0, \quad r_T/(Th_T)^{1/2} \to 0, \quad \frac{T}{r_T}\alpha(s_T) \to 0. \tag{6.77}$$

We now establish (6.73) and (6.74). First, by stationarity and (6.72), we find

$$\mathrm{Var}(\xi_j) = s\theta^2(x)\{1 + o(1)\}.$$

By (6.72) and (6.77), we have

$$E(Q''_T)^2 = k\mathrm{Var}(\xi_j) + O\left(T\sum_{t=0}^{T-1}|\mathrm{Cov}(Z_{T,0}, Z_{T,t})|\right) = o(T).$$

The same argument leads to the second part of (6.73) and (6.74).

Note that the indices in η_j and η_{j+1} are at least s_T apart. Hence, applying Proposition 2.6 with $V_j = \exp(it\eta_j)$, we find

$$\left| E\exp(itQ'_T) - \prod_{j=0}^{k-1} E[\exp(it\eta_j)] \right| \leq 16k\alpha(s_T) \sim 16\frac{T}{r_T}\alpha(s_T),$$

which tends to zero by (6.77). This proves (6.74).

It remains to establish (6.76). We employ a truncation argument as follows. Let $Y_{L,t} = Y_t I\{|Y_t| \leq L\}$, where L is a fixed truncation point. Correspondingly, let us add the superscript L to indicate the quantities that involve $\{Y_{L,t}\}$ instead of $\{Y_t\}$. Then $Q_T = Q_T^L + \tilde{Q}_T^L$, where

$$\tilde{Q}_T^L = T^{-1}\sum_{t=1}^{T}(Z_t - Z_t^L).$$

Using the fact that $C(\cdot)$ is bounded (since K is bounded with a compact support), we have $|Z_{T,t}^L| \leq D/h^{1/2}$ for some constant D. Then, using (6.77), it follows that

$$\max_{0\leq j\leq k-1} |\eta_j^L|/\sqrt{T} \leq Dr_T/\sqrt{Th_T} \to 0.$$

Hence, when T is large, the set $\{|\eta_j^L| \geq \theta_L(x)\varepsilon\sqrt{T}\}$ becomes an empty set, and hence (6.76) holds. Consequently, we have the following asymptotic normality:

$$\sqrt{Th_T}Q_T^L \xrightarrow{D} N\{0, \theta_L^2(x)\}. \tag{6.78}$$

In order to complete the proof (i.e., to establish (6.69)), it suffices to show that as first $T \to \infty$ and then $L \to \infty$, we have

$$Th\mathrm{Var}\left(\tilde{Q}_T^L\right) \to 0. \tag{6.79}$$

Indeed, from this, we proceed as follows:

$$\begin{aligned} & \left|E\exp(it\sqrt{Th}Q_T) - \exp\{-t^2\theta^2(x)/2\}\right| \\ \leq\ & E|\exp(it\sqrt{Th}Q_T^L)\{\exp(it\sqrt{Th}\tilde{Q}_T^L) - 1\}| \\ & + \left|E\exp(it\sqrt{Th}Q_T^L) - \exp\{-t^2\theta_L^2(x)/2\}\right| \\ & + \left|\exp\{-t^2\theta_L^2(x)/2\} - \exp\{-t^2\theta^2(x)/2\}\right|. \end{aligned}$$

The first term is bounded by

$$E|\exp(it\sqrt{Th}\tilde{Q}_T^L) - 1| = O\{\mathrm{Var}(\sqrt{Th}\tilde{Q}_T^L)\}.$$

Letting $T \to \infty$, the first term converges to zero by (6.79) as first $T \to \infty$ and then $L \to \infty$, the second term goes to zero by (6.78) for every $L > 0$, and the third term goes to zero as $L \to \infty$ by the dominated convergence theorem. Therefore, it remains to prove (6.79). Note that $\tilde{Q}_T^L$ has the same structure as Q_T. Hence, by (6.72), we obtain

$$\lim_{T\to\infty} Th\mathrm{Var}\left(\tilde{Q}_T^L\right) = \mathrm{Var}(YI[|Y| > L]|X = x)f(x)\int_{-\infty}^{+\infty} C^2(u)du.$$

By dominated convergence, the right-hand side converges to 0 as $L \to \infty$. This establishes (6.79) and completes the proof of Theorem 6.3. ∎

6.6.3 Proof of Lemma 6.1

The idea of proving this lemma is a combination of the techniques used in Mack and Silverman (1982) and Theorem 5.3. Recall that C denotes a generic constant, which can vary from one place to another. The proof consists of the following three steps.

(a) (Discretization). Let $Q_h(x) = T^{-1}\sum_{t=1}^T K_h(x - X_t)Y_t$. Partition the interval $[a, b]$ into $N = [(T/h)^{1/2}]$ subintervals $\{I_j\}$ of equal length. Let $\{x_j\}$ be the centers of I_j. Then

$$\sup_{x\in[a,b]} |Q_h(x) - EQ_h(x)| \le \max_{1\le j\le N} |Q_h(x_j) - EQ_h(x_j)| + C(Th)^{-1/2}. \tag{6.80}$$

(b) (Truncation). Let $Q_h^B(x) = T^{-1}\sum_{t=1}^T K_h(x - X_t)Y_t I(|Y_t| \le B_t)$ for an increasing sequences B_t satisfying $\sum_t B_t^{-s} < \infty$. Then, with probability 1,

$$\sup_{x\in[a,b]} |Q_h(x) - Q_h^B(x) - E\{Q_h(x) - Q^B(x)\}| = O(B_T^{1-s}). \tag{6.81}$$

(c) (Maximum deviation for discretized and truncated series). For $\varepsilon_T = (a\log T/Th)^{1/2}$ with sufficiently large a,

$$\begin{aligned} & P\left(\max_{1\le j\le N} |Q_h^B(x_j) - EQ_h^B(x_j)| > \varepsilon_T\right) \\ = \; & O\{B_T^{\beta+1.5}T^{-\beta/2+0.75}h^{-\beta/2-0.75}(\log T)^{\beta/2+0.25}\}, \end{aligned} \tag{6.82}$$

provided that $B_T\varepsilon_T \to 0$.

Suppose that the results in (a)–(c) are correct. Then, by taking $B_T = T^{(s^{-1}+\delta)}$ for some $\delta > 0$, we deduce from (6.81) that

$$\sup_{x\in[a,b]} |Q_h(x) - Q_h^B(x) - E\{Q_h(x) - Q^B(x)\}| = o(T^{-1/2}),$$

which is negligible. This and (6.80) entail

$$\sup_{x\in[a,b]} |Q_h(x)-EQ_h(x)| \le \max_{1\le j\le N} |Q_h^B(x_j)-EQ_h^B(x_j)|+C(Th)^{-1/2}. \quad (6.83)$$

By the condition of this lemma, $B_T\varepsilon_T \to 0$ and the probability given (6.82) tends to zero. Hence

$$\max_{1\le j\le N} |Q_h^B(x_j) - EQ_h^B(x_j)| = O_P\{(\log T/Th)^{1/2}\}.$$

The result follows from (6.83). It remains to prove the results in parts (a)—(c).

The proof of part (a) is very similar to that given in the proof of Theorem 5.3. By using the Lipschitz condition of K, we have

$$\begin{aligned} |Q_h(x) - Q_h(x')| &\le Ch^{-1}|x-x'|T^{-1}\sum_{t=1}^{T}|Y_t| \\ &\le Ch^{-1}|x-x'|E|Y|. \end{aligned} \quad (6.84)$$

Similarly, using the first equality in (6.84), we have

$$|E\{Q_h(x) - Q_h(x')\}| \le Ch^{-1}|x-x'|E|Y|.$$

This and (6.84) prove part (a).

The proof of part (b) is quite similar to that in Mack and Silverman (1982). Note that

$$\sum_t P\{|Y_t| > B_t\} < \sum_t B_t^{-s}E|Y|^s < \infty.$$

By the Borel–Cantelli lemma, with probability 1, $|Y_t| \le B_t$ for sufficiently large t. Hence, for all sufficiently large T,

$$|Y_t| \le B_T \quad \text{for all } t \le T.$$

This implies that $\sup_{x\in[a,b]} |Q_h(x) - Q_h^B(x)|$ is eventually zero with probability 1. It remains to bound the expectation term. By using the fact that

$$\sup_{x\in[a,b]} \int_{|y|\ge B_T} |y|f(x,y)dy \le CB_T^{1-s},$$

we have

$$\begin{aligned} E|Q_h(x) - Q_h^B(x)| &\le \int_{|y|\ge B_T} |K_h(x-u)||y|f(u,y)dydu \\ &\le \sup_{x\in[a,b]} \int_{|y|\ge B_T} |y|f(x,y)dy \int |K(u)|du \\ &\le CB_T^{1-s}. \end{aligned}$$

Combining the two results above, we prove part (b).

We now prove part (c). Let

$$Z_t = K_h(x - X_t)Y_t I(|Y_t| \le B_T) - EK_h(x - X_t)Y_t I(|Y_t| \le B_T).$$

Then $\|Z_t\|_\infty < CB_T/h$. By using the exponential inequality (Theorem 2.18), we have for any $\varepsilon > 0$ and each integer $q \in [1, T/2]$,

$$\begin{aligned} & P\{|Q_h^B(x) - EQ_h^B(x)| > \varepsilon\} \\ \le \; & 4\exp\left(-\frac{\varepsilon^2 q}{8v^2(q)}\right) + 22\left\{1 + \frac{4B_T}{h\varepsilon}\right\}^{1/2} q\alpha(p), \end{aligned} \tag{6.85}$$

where $p = [T/(2q)]$,

$$v^2(q) = 2\sigma^2(q)/p^2 + CB_T\varepsilon/h$$

and

$$\sigma^2(q) = \max_{0 \le j \le 2q-1} \mathrm{Var}\{Y_{jp+1} + \cdots + Y_{(j+1)p+1}\}.$$

By the proof of Theorem 6.3, when T is sufficiently large, $\sigma^2(q) \le Cp/h$. Hence, by taking $p = [(B_T\varepsilon_T)^{-1}]$, by (6.85) and some simple algebra, we have

$$\begin{aligned} & P\{|Q_h^B(x) - EQ_h^B(x)| > \varepsilon_T\} \\ \le \; & 4\exp(-C\varepsilon_T^2 Th) + CB_T^{\beta+1.5} Th^{-1/2}\varepsilon_T^{\beta+0.5}. \end{aligned} \tag{6.86}$$

Rewrite $\varepsilon^2 = a\log T/(CTh)$ for the sufficiently large a. Expression (6.86) is bounded by

$$4T^{-a} + CB_T^{\beta+1.5} T^{-\beta/2+0.75} h^{-\beta/2-0.75} (\log T)^{\beta/2+0.25}.$$

Consequently,

$$\begin{aligned} & P(\max_{1 \le j \le N} |Q_h^B(x_j) - EQ_h^B(x_j)| > \varepsilon) \\ \le \; & N\{4T^{-a} + B_T^{\beta+0.5} T^{-\beta/2+0.75} h^{-\beta/2-0.75} (\log T)^{\beta/2+0.25}\}. \end{aligned}$$

This proves part (c) and hence the lemma. ∎

6.6.4 Proof of Theorem 6.5

We use the notation for the proof of Theorem 6.3. By using Lemma 6.1 with $Y_j \equiv 1$, each element $S_{T,j}$ converges uniformly to its asymptotic counterpart with stochastic error of order $\{Th/\log(T)\}^{-1/2}$ and bias $o(1)$. By (6.65), we have

$$\mathbf{b} = \beta_{p+1}(\mathbf{HSH})^{-1}\mathbf{H}\mathbf{c}_p h^{p+1}\{1 + o(1) + O_P(\{Th/\log(T)\}^{-1/2})\}$$

uniformly for $x \in [a, b]$. Note that each element of $\mathbf{u}$ in (6.67) is of the form given in Lemma 6.1. By Lemma 6.1, it is of order $\{Th/\log(1/h)\}^{-1/2}$ uniformly in $x \in [a, b]$. The result of Theorem 6.5 follows directly from (6.64) and the results above. ∎

6.6.5 Proof for Theorems 6.6 and 6.7

Condition 2:

(i) The kernel functions W and K are symmetric and bounded with bounded supports.

(ii) The process $\{X_j, Y_j\}$ is ρ-mixing with $\sum_\ell \rho(\ell) < \infty$. Furthermore, assume that there exists a sequence of positive integers $s_n \to \infty$ such that $s_n = o\{(nh_1h_2)^{1/2}\}$ and $\{n/(h_1h_2)\}^{1/2}\rho(s_n) \to 0$.

(iii) The function $f(y|x)$ has bounded and continuous third order derivatives at point (x, y), and $f(\cdot)$ is continuous at point x.

(iv) The joint density of the distinct elements of $(X_0, Y_0, X_\ell, Y_\ell)$ $(\ell > 0)$ is bounded by a constant that is independent of ℓ.

Note that the ρ-mixing conditions above can easily be modified for the α-mixing process.

Proof. The proof of Theorem 6.6 is similar to that of Theorem 6.7. The latter uses the same techniques as the proof of Theorem 6.3, so we only outline the proof of Theorem 6.7.

Let $m(x, y) = E\{K_{h_2}(Y_t - y)|X_t = x\}$, $H = \text{diag}(1, h_1, h_1^2)$, and

$$\beta = (m_0(x, y), m_1(x, y), m_2(x, y))^T,$$

where $m_0(x, y) = m(x, y)$ and for $j \geq 1$,

$$m_j(x, y) = \frac{1}{j!}\frac{\partial^j}{\partial x^j}m(x, y).$$

Using matrix notation and simple algebra, we obtain from (6.23) and (6.54) that

$$H(\widehat{\beta} - \beta) = \mathbf{S}_T^{*-1}\{(U_{T,0}, U_{T,1}, U_{T,2})^T + (\gamma_{T,0}, \gamma_{T,1}, \gamma_{T,2})^T\}, \qquad (6.87)$$

where $\mathbf{S}_T^*$ is a 3×3 matrix with the (i, j)-element $S^*_{T,i+j-2}$,

$$\begin{aligned}
U_{T,j} &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{X_t - x}{h_1}\right)^j W_{h_1}(X_t - x)\left\{K_{h_2}(Y_t - y) - m(X_t, y)\right\}, \\
\gamma_{T,j} &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{X_t - x}{h_1}\right)^j W_{h_1}(X_t - x) \times \\
&\quad \left\{m(X_t, y) - m(x, y) - m_1(x, y)(X_t - x) - m_2(x, y)\frac{(X_t - x)^2}{2}\right\}, \\
S^*_{T,j} &= \frac{1}{T}\sum_{t=1}^{T}\left(\frac{X_t - x}{h_1}\right)^j W_{h_1}(X_t - x).
\end{aligned}$$

Let $\mathbf{S}$ and Σ be 3×3 matrices with (i,j)-elements μ_{i+j-2} and ν_{i+j-2}, respectively, and $\gamma = (\mu_3, \mu_4, \mu_5)^T$. We will establish that

(a) $\mathbf{S}_T^*$ converges to $f(x)\mathbf{S}$ in probability.

(b) $h_1^{-3}(\gamma_{T,0}, \gamma_{T,1}, \gamma_{T,2})^T$ converges to $6^{-1}f(x)\gamma\partial^3 f(y|x)/\partial x^3$ in probability.

(c) $(Th_1h_2)^{1/2}(U_{T,0}, U_{T,1}, U_{T,2})$ is asymptotically normal with mean 0 and variance $f(y|x)f(x)\nu_0\nu_K\Sigma$.

Combining these with (6.87), we have

$$\begin{aligned} & (Th_1h_2)^{1/2}\left\{H(\widehat{\beta}-\beta) - \frac{1}{6}h_1^3\frac{\partial^3 f(y|x)}{\partial x^3}\mathbf{S}^{-1}\gamma\right\} \\ \xrightarrow{D} \quad & N\left(0, f(y|x)\nu_0\nu_K\mathbf{S}^{-1}\Sigma\mathbf{S}^{-1}/f(x)\right). \end{aligned} \tag{6.88}$$

It follows from the Taylor expansion that

$$m_j(x,y) = \frac{\partial^j f(y|x)}{\partial x^j} + \frac{1}{2}h_2^2\mu_K\frac{\partial^{j+2}f(y|x)}{\partial x^j\partial y^2} + o(h_2^2).$$

Using this expansion and considering the marginal distribution of (6.88), we obtain the result.

Conclusion (a) has already been shown in (6.28). For (b), by Taylor's expansion, we have

$$\gamma_{T,j} = \frac{m_3(x,y)h_1^3}{T}\sum_{t=1}^{T}\left(\frac{X_t - x}{h_1}\right)^{j+3} W_{h_1}(X_t - x)\{1 + o(1)\}.$$

Using (6.67) again, we have

$$h_1^{-3}\gamma_{T,j} \to m_3(x,y)f(x)\mu_{j+3}$$

This establishes (b).

To prove (c), we consider arbitrary linear combinations of $U_{T,j}$ with constant coefficients η_j $(j = 0, 1, 2)$. Let

$$\begin{aligned} Q_T &= (Th_1h_2)^{1/2}(\eta_0 U_{T,0} + \eta_1 U_{T,1} + \eta_2 U_{T,2}) \\ &= T^{-1/2}\sum_{t=1}^{T}(h_1h_2)^{1/2}D_{h_1}(X_t - x)\{K_{h_2}(Y_t - y) - m(X_i, y)\}, \end{aligned}$$

where $D(u) = (\eta_0 + \eta_1 u + \eta_2 u^2)W(u)$. Write $Q_T = T^{-1/2}(Z_{T,0} + \ldots + Z_{T,T-1})$. Note that Q_T is the sum of a stationary mixing sequence. Its asymptotic normality follows from the small-block and large-block arguments, as shown in the proof of Theorem 6.3. ∎

6.7 Bibliographical Notes

Smoothing in time series is closely related to density estimation and other related problems such as spectral density estimation. It is an extension of the smoothing techniques for independent data and their nonparametric counterparts. See Sections 5.8 and 7.6 for related references. In this section, we mainly focus on some important developments for dependent data.

Nonparametric regression for an independent sample

There are many interactions between the developments of nonparametric density estimation and nonparametric regression. The kernel regression was independently proposed by Nadaraya (1964) and Watson (1964). Other variants include the ones in Priestley and Chao (1972) and Gasser and Müller (1979). Chu and Marron (1991b) compared the merits of various versions of the kernel regression estimator. The optimal rates of convergence were established by Stone (1980, 1982). Mack and Silverman (1982) established uniform consistency for kernel regression. The asymptotic distribution of the maximum deviation between the estimated regression function and the true one was derived in Gruet (1996). The result was independently extended to varying coefficient models by Xia and Li (1999b) and Fan and Zhang (2000) using different estimators. Extensive treatments of kernel regression estimators can be found in the books by Müller (1988), Härdle (1990), and Eubank (1999).

Local polynomial fitting

Local polynomial regression is very useful for estimating regression functions and their derivatives. It has been thoroughly treated in the book by Fan and Gijbels (1996) and the references therein. The idea of local approximation appeared in Woolhouse (1870) and Macaulay (1931), but it dates back at least as early as the time when π was computed. It was first used as a tool for nonparametric regression by Stone (1977) and Cleveland (1979). Tsybakov (1986) demonstrated the asymptotic properties of robust local polynomial estimators. The equivalence between the local polynomial fitting and kernel regression was demonstrated by Müller (1987) for a fixed design setting.

Fan (1992, 1993a) clearly demonstrated the advantages of using local polynomial fitting for nonparametric regression and revived interest in the local polynomial techniques. Subsequently, Fan and Gijbels (1992) and Hastie and Loader (1993) demonstrated that the local linear fitting automatically corrects boundary biases. Ruppert and Wand (1994) extended the results to general local polynomial fitting. Fan, Farmen, and Gijbels (1998) laid out a blueprint for local maximum likelihood estimation, and Carroll, Ruppert, and Welsh (1998) generalized the method further to in-

clude local estimation equations. Data-driven bandwidth selection methods can be found in Fan and Gijbels (1995), Ruppert, Sheather, and Wand (1995), and Ruppert (1997).

Nonparametric regression for a dependent sample

There are various nonparametric regression problems for time series: time-domain smoothing, state-domain smoothing and estimation of conditional density and conditional variance, among others. Yakowitz (1985) considered estimating the conditional mean and transition density for Markov sequences. Roussas (1990) obtained a strong consistent rate for the kernel regression estimator. Nonparametric regression with errors-in-variables was studied by Fan and Masry(1992). Truong and Stone (1992) established rates of convergence under the L_2 and L_∞ norms for local average and local median estimators. The optimal rate of convergence was established in Tran (1993) under some weaker conditions than in previous work. Yao and Tong (1996) estimated conditional expectiles for dependent processes. A semiparametric problem was investigated by Truong and Stone (1994), where the root-n rate was constructed. Vieu (1991) showed that all MISE, ISE, and ASE measures are asymptotically equivalent. This is an extension of a result by Härdle and Marron (1985) from independent to dependent cases; see also the study by Kim and Cox (1995). Tran, Roussas, Yakowitz, and Truong (1996) established asymptotic normality for nonparametric regression estimators under fairly general conditions. Nonparametric multivariate autoregression problems were studied by Härdle, Tsybakov, and Yang (1998). Opsomer, Wang, and Yang (2001) give an overview of nonparametric regression with correlated errors. Jiang and Mack (2001) studied robust local polynomial regression for dependent data, where the one-step properties have also been studied. Nonparametric regression with heavy-tailed dependent errors was studied in Peng and Yao (2001).

Hall and Hart (1990) derived the means-square properties for both long-memory and short-memory errors. Altman (1990) studied time-domain smoothing and bandwidth selection for data with short-memory errors. The rates of convergence of time domain smoothing and semiparametric estimation were investigated by Truong (1991). Brillinger (1996), Wang (1996), and Johnstone and Silverman (1997) studied nonparametric regression using wavelet thresholding estimators. The asymptotic normality for kernel regression under both short- and long-range dependences was studied by Csörgö and Mielniczuk (1995) and Robinson (1997). The asymptotic distribution for the maximum deviation was derived in Csörgö and Mielniczuk (1995).

Nonparametric estimation of drift and diffusion functions were nonparametrically estimated by Pham (1981), Prakasa Rao (1985), Stanton (1997), and Fan and Yao (1998), among others. Wang (2002) investigated the problems of the asymptotic equivalence of ARCH models and diffusions.

Durham and Gallant (2002) studied simulated maximum likelihood estimation based on a discrete sample from diffusion processes. Aït-Sahalia (1999, 2002) derived asymptotic expansions of the transition densities for stochastic diffusion models and investigated the properties of maximum likelihood estimators.

Spline smoothing

The idea of the smoothing spline appeared in Wittaker (1923), Schoenberg (1964), and Reinsch (1967). It was introduced to statistics by Kimeldorf and Wahba (1970) and Wahba (1975). Multivariate spline approximations were discussed by Wong (1984) and Gu (1990). The choice of smoothing parameters was discussed by Utreras (1980) and Li (1985, 1986) in addition to Wahba (1977). Confidence intervals can be constructed by using the Bayesian method described in Nychka (1988).

The knot deletion idea was proposed in Smith (1982) and the book by Breiman, Friedman, Olshen, and Stone (1984); see also its revision in 1993. The current state-of-art of regression splines based on the knot deletion method can be found in Stone, Hansen, Kooperberg and Truong (1997). The method of the sieve for stationary β-mixing observations was studied by Chen and Shen (1998). The asymptotic normality and rate of convergence for nonparametric neural network estimators were established by Chen and White (1999).

Bandwidth selection

The problem of choosing a smoothing parameter exists in virtually all nonparametric estimation. The basic idea is to choose the parameter to minimize either integrated squared errors or the mean integrated squares errors. The methods can basically be classified into two categories: cross-validation methods and plug-in methods. For a survey and development of cross-validation for independent data, see Hall and Johnstone (1992). See also Hall, Marron, and Park (1992) for a smoothed cross-validation method. Most developments and studies in this area are in the i.i.d. density estimation setting. For a survey, see Jones, Marron, and Sheather (1996).

The bandwidth selection for state-domain smoothing problems is very similar to that for independent data when mixing conditions are strong enough. For time-domain smoothing, because of local dependence, the bandwidth selectors for independent samples do not work well. This was observed and studied by Altman (1990), Chu and Marron (1991a), and Hart (1991). Hart and Vieu (1990) and Härdle and Vieu (1992) showed asymptotic optimality for multifold cross-validation. Hall, Lahiri, and Truong (1995) studied properties of cross-validation and plug-in bandwidth selection with dependent data. The asymptotic normality for a cross-validation bandwidth selector was established in Chu (1995). Kim and Cox (1997)

studied the asymptotic convergence rate for a cross-validation bandwidth estimator in a density estimation setting. A generalized cross-validation was considered by Yao and Tong (1998a) for rho-mixing processes.

Robinson (1994) considered data-driven nonparametric estimation for spectral density with singularity at point zero. Ray and Tsay (1997) proposed a plug-in bandwidth selection for kernel regression with long-range dependence, which is an extension of the method by Brockmann, Gasser, and Herrmann (1993).

7

Spectral Density Estimation and Its Applications

7.1 Introduction

Spectral density reveals the power spectrum of a stationary time series. It characterizes the second-moment properties of a stationary time series. By inspecting an estimated spectral density, we may identify the frequency ranges that contribute the most variation of data. It also helps to identify an appropriate family of models that possess the key correlation features of the underlying process. In particular, when an estimated spectral density is nearly a constant, one may infer that an underlying process is a white noise process. This is useful for model diagnostics; after fitting a certain family of models, one wishes to verify if the family of models adequately fits a given time series by checking whether or not the residual series is a white noise process. The latter can be done by inspecting whether the estimated spectral density based on residuals is nearly a constant.

The raw material for estimating spectral density is the periodogram $I_T(\omega_k)$ defined in §2.4.2, where $\omega_k = 2\pi k/T$. Let $V_k = (\xi_{2k-1}^2 + \xi_{2k}^2)/2$, which is a sequence of independent random variables with the standard exponential distribution. Let $n = [(T-1)/2]$. As shown in Theorem 2.14, the periodogram can be written as

$$I_T^*(\omega_k) = g(\omega_k)V_k + R_T(\omega_k), \quad k = 1, \cdots, n, \tag{7.1}$$

where $R_T(\omega_k)$ is asymptotically negligible and $I_T^*(\omega_k) = I_T(\omega_k)/(2\pi)$, which is used as a definition of periodograms by some authors. Let $m(x) =$

$\log g(x)$ and $Y_k = \log I_T^*(\omega_k)$. Then, (7.1) can be written as

$$Y_k = m(\omega_k) + z_k + r_k, \quad k = 1, \cdots, n, \tag{7.2}$$

where $r_k = \log[1 + R_T(\omega_k)/\{g(\omega_k)V_k\}]$, which represents an asymptotically negligible term, and

$$z_k = \log(V_k) \text{ has a density } \exp\{-\exp(x) + x\}. \tag{7.3}$$

Thus, the spectral density estimation is basically a nonparametric regression problem with nearly independent data. Furthermore, the periodogram $I_T^*(\omega_k)$ is not a consistent estimator of $g(\omega_k)$ (see the discussion below Theorem 2.14 and also Figure 7.5). Smoothing is needed in order to obtain a good estimator of the spectral density.

The essence of the approximations above utilizes the Fourier transform. It is a powerful tool for analyzing stationary time series. As shown in Theorem 2.14, it transforms correlated stationary data into nearly independent data. Thus, after the Fourier transform, techniques for independent data can be employed. For example, one can estimate unknown parameters in a stationary process by forming the likelihood based on the periodogram ordinates, relying on an assumption that the spectral density $f(\omega)$ is of a parametric form $f(\omega; \boldsymbol{\theta})$ with unknown parameters $\boldsymbol{\theta}$. This results in the parametric function $m(\omega; \boldsymbol{\theta})$. The parameters $\boldsymbol{\theta}$ can be estimated by forming a parametric likelihood function from (7.2) after ignoring the term r_k, which is the Whittle likelihood; see, for example, Dahlhaus (1984, 1990a) and Dzhaparidze (1986). This idea will be further explored in §9.3.

We begin this chapter with a brief review of traditional approaches on the estimation of spectral density and its relation to kernel smoothing. In §7.3, we apply the techniques introduced in Chapter 6 to estimate spectral densities. An important question in fitting time series data is whether the residuals of a fitted model behave like a white noise series. Nonparametric function estimation provides useful tools for a nonparametric goodness-of-fit test. Several useful tests are introduced in §7.4 and are illustrated by numerical examples.

7.2 Tapering, Kernel Estimation, and Prewhitening

The genesis of smoothing techniques comes from the need for consistent estimates of spectral density. Traditional approaches of estimating spectral density are to smooth periodograms directly using a kernel weight function. It is helpful to take a quick overview of the traditional techniques before we apply the modern smoothing techniques to the problem. Chapter 5 of Brillinger (1981) gives comprehensive coverage of the traditional techniques.

7.2.1 Tapering

The spectral density $g(\omega)$ can be obtained from the autocovariance function (2.36):

$$g(\omega) = \sum_{k=-\infty}^{\infty} \gamma(k) e^{-ik\omega}.$$

The autocorrelation function can be expressed as

$$\gamma(k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ik\omega} g(\omega) d\omega. \tag{7.4}$$

There is much literature concerning the behavior of the partial sums

$$\sum_{k=-p}^{p} \gamma(k) e^{-ik\omega}, \quad p = 1, 2, \cdots; \tag{7.5}$$

see, for example, Zygmund (1968). Fejér (1900, 1904) recognized that the partial sums above might not be good approximations of the spectral density. He therefore introduced a convergence factor into the series above:

$$\sum_{k=-p}^{p} \left(1 - \frac{|k|}{p}\right) \gamma(k) e^{-ik\omega}.$$

This improves the rate of convergence as $p \to \infty$. This idea was then generalized to a general form

$$g_p(\omega) = \sum_{k=-p}^{p} w(|k|/p) \gamma(k) e^{-ik\omega}, \tag{7.6}$$

where the function $w(\cdot)$ is given, and is called a *convergence factor*, *data windows,* or *tapers;* see Tukey (1967). The function $w(\cdot)$ usually satisfies

$$w(0) = 1, \quad |w(x)| \le 1, \quad w(x) = 0, \quad \text{for } |x| > 1.$$

Substituting (7.4) into (7.6), we obtain

$$g_p(\omega) = \int_{-\pi}^{\pi} W_p(\omega - \tau) g(\tau) d\tau, \tag{7.7}$$

where

$$W_p(\tau) = \frac{1}{2\pi} \sum_{k=-p}^{p} w(|k|/p) e^{-ik\tau}. \tag{7.8}$$

As p gets large, the function W_p will get more and more concentrated on the point 0 (see Figures 7.1–7.4). Hence, the function W_p plays the same role as the function $K_h(\cdot)$ in the kernel smoothing. In fact, it holds that

$$\int_{-\pi}^{\pi} W_p(\tau) d\tau = 1.$$

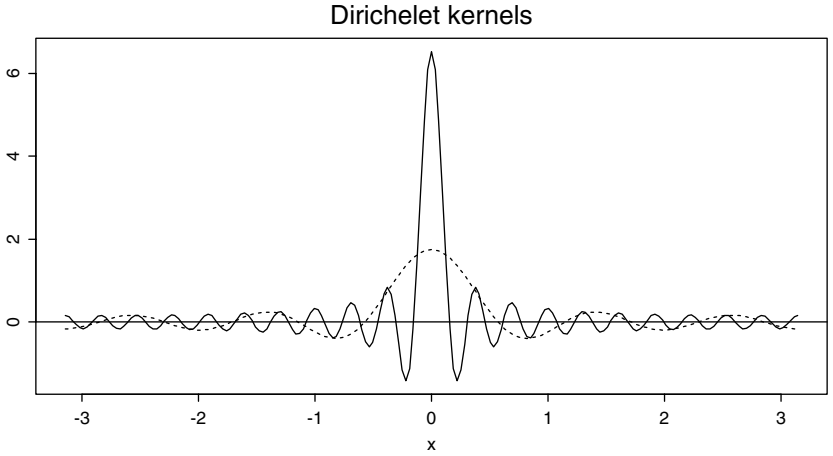


FIGURE 7.1. The Dirichlet kernels with $p = 20$ (solid curve) and $p = 5$ (dashed curve).

The function W_p is called a *frequency window* or a kernel. Here are a few useful examples of the taper functions.

Example 7.1 (*Rectangular or truncated window*). This taper function has the form $w(x) = 1$ if $x \leq 1$ and 0 otherwise. This corresponds to the partial sum series (7.5). The frequency window is

$$W_p(\tau) = \frac{\sin((p + 1/2)\tau)}{2\pi \sin(\tau/2)}, \tag{7.9}$$

which is called the *Dirichlet kernel* (see Figure 7.1). Observe that $W_p(\tau)$ can be negative for certain values of τ. This may lead to negative estimates of the spectral density at certain frequencies. ∎

Example 7.2 (*The Bartlett or triangular window*). This convergence factor is given by $w(x) = (1 - |x|)_+$, and the corresponding frequency window is given by the Fejer kernel (see Figure 7.2),

$$W_p(\tau) = \frac{\sin^2(p\tau/2)}{2\pi p \sin^2(\tau/2)}.$$

This kernel is nonnegative and is indeed a second-order kernel. ∎

Example 7.3 (*The Blackman–Tukey window*). This taper function admits the general form

$$w(x) = \begin{cases} 1 - 2a + 2a\cos x, & |x| \leq 1 \\ 0, & \text{otherwise} \end{cases}.$$

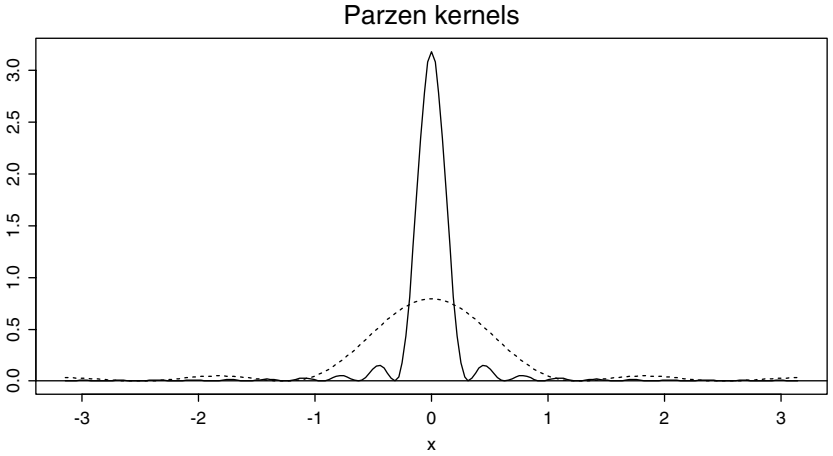


FIGURE 7.2. The Bartlett kernels with $p = 20$ (solid curve) and $p = 5$ (dashed curve).

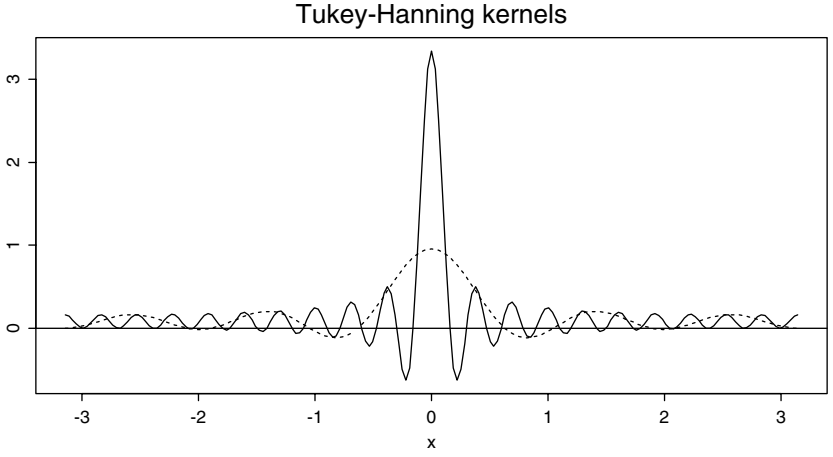


FIGURE 7.3. The Tukey–Hanning kernels with $p = 20$ (solid curve) and $p = 5$ (dashed curve).

The corresponding kernel function is given by

$$W_p(\tau) = aD_p(\tau - \pi/r) + (1 - 2a)D_p(\tau) + aD_r(\tau + \pi/r),$$

where D_p is the Dirichlet kernel given by (7.9). The cases with $a = 0.23$ and $a = 0.25$ are frequently referred to as the Tukey–Hamming and Tukey–Hanning windows (see Figure 7.3). ∎

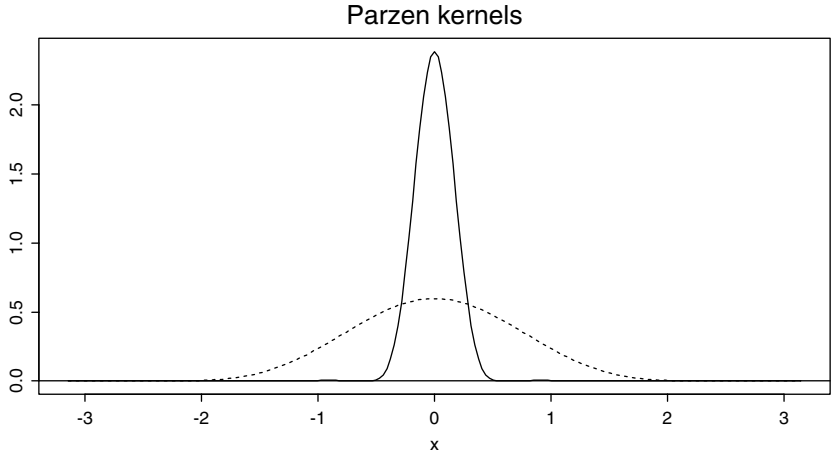


FIGURE 7.4. The Parzen kernels with $p = 20$ (solid curve) and $p = 5$ (dashed curve).

Example 7.4 (*The Parzen window*). In this case,

$$w(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3, & |x| < 1/2 \\ 2(1 - |x|)^3, & 1/2 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}.$$

The corresponding kernel is given by

$$W_p(\tau) = \frac{6 \sin^4(p\tau/4)}{\pi p^3 \sin^4(\tau/2)}.$$

Figure 7.4 depicts the kernel function. ∎

The choice of convergence factors is very much like the choice of a kernel function. The bandwidth parameter in (7.8) is implicitly defined. It is related to the parameter p. Several proposals have been suggested to define the measure of bandwidth explicitly so that one gets an idea of how large a window size has been used for different converging factors. For example, Grenander (1951) suggested the measure

$$\left\{ \int_{-\pi}^{\pi} \tau^2 W_p(\tau) d\tau \right\}^{1/2},$$

Parzen (1961) used the measure

$$\frac{1}{W_p(0)} = \frac{2\pi}{\sum w(|k|/p)},$$

and Brillinger (1981, p. 56) was in favor of the measure

$$\left\{ \int_{-\pi}^{\pi} (1 - \cos\tau) W_p(\tau) d\tau \right\}^{1/2} = \{1 - w(1/n)\}^{1/2}.$$

Although different authors have different measures of the bandwidth, this does not affect the practical usage of the method. In practice, one would tune the parameter p to get a good estimate of the spectral density.

Expression (7.6) suggests the following substitution estimator

$$\widehat{g}_p(\omega) = \sum_{k=-p}^{p} w(|k|/p)\widehat{\gamma}(k)e^{-ik\omega}, \tag{7.10}$$

where $\widehat{\gamma}(k)$ is the sample autocovariance function given in §2.2.2. Note that this is a real function and is called the lag window estimator. It is obvious that (7.10) admits an expression similar to (7.7). To see this, we define

$$\tilde{I}_T(\omega) = \sum_{|k|<T} \widehat{\gamma}(k)e^{-ik\omega}.$$

This is an extension of the periodogram to all frequencies. Then

$$\widehat{\gamma}(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ik\tau}\tilde{I}_T(\tau)d\tau.$$

Substituting this into (7.10), we obtain easily that

$$\widehat{g}_p(\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W_p(\omega-\tau)\tilde{I}_p(\tau)d\tau. \tag{7.11}$$

The smoothing parameter p can be chosen either subjectively by time series analysts or objectively by data. Bühlmann (1996) proposed a local data-driven choice of the parameter p based on an idea of Brockmann, Gasser, and Herrmann (1993).

7.2.2 Smoothing the Periodogram

Partitioning the interval $[-\pi, \pi]$ at the Fourier frequencies $\{\omega_k\}$ and replacing the integral in (7.11) by the corresponding Riemann sum, we obtain

$$\widehat{g}_p(\omega) \approx \frac{1}{2\pi}\sum_{|k|\le n} W_p(\omega-\omega_k)\tilde{I}_T(\omega_k)\frac{2\pi}{T}. \tag{7.12}$$

When p is sufficiently large, W_p is concentrated around the origin. Hence, the summation above is a local average of the periodogram $\tilde{I}_T(\omega)$ around the frequency $\omega_k \approx \omega$. It holds that

$$\frac{1}{2\pi}\sum_{|k|\le n} W_p(\omega-\omega_k)\frac{2\pi}{T} \approx \int_{-\pi}^{\pi} W_p(\omega-\tau)d\tau = 1.$$

This makes the connection to the kernel smoothing in the frequency domain.

Starting directly from (5.1), one can obtain an estimate of the spectral density as

$$\widehat{g}_h(\omega) = \frac{\sum_{k=1}^n K_h(\omega - \omega_k) I_T^*(\omega_k)}{\sum_{k=1}^n K_h(\omega - \omega_k)}, \tag{7.13}$$

where $K_h(u) = K(u/h)/h$, with K being a kernel function and h being a bandwidth. This is the kernel smoothing technique applied to the bivariate data $\{(\omega_k, I_T^*(\omega_k)), k = 1, \cdots, n\}$, resulting in a smoothed periodogram. The kernel estimator can be regarded as an extension of the class of estimators in (7.12).

The smoothing parameters p and h are usually chosen to balance the bias and variance trade-off. For example, a large bandwidth results in a larger window of averaging. This reduces sampling variance but at the same time increases biases. One can achieve the bias and variance trade-off either subjectively via visualization or objectively by data-driven techniques. The latter will be discussed in §7.3 in the context of the local linear estimator.

7.2.3 Prewhitening and Bias Reduction

When the underlying spectral density contains sharp peaks, direct applications of smoothing techniques will widen the peaks and reduce the magnitudes of the peaks unless the bandwidth is very small. However, a small bandwidth will not be able to reduce enough variances of the estimate, resulting in wiggly estimates, particularly at flat regions. A common technique to resolve this problem is *prewhitening*. The basic idea is to apply a linear filter to the series $\{X_t\}$, resulting in

$$Y_t = \sum_{k=-\infty}^{\infty} \phi_k X_{t-k},$$

and estimate the spectral density of the filtered series $\{Y_t\}$. This filtered series has the spectral density (see Theorem 2.12)

$$g_Y(\omega) = g_X(\omega)|\Gamma(\omega)|^2,$$

where

$$\Gamma(\omega) = \sum_{k=-\infty}^{\infty} \phi_k e^{-ik\omega}$$

is a transfer function. Suppose that the filter can be chosen so that the spectral density g_Y of the series $\{Y_t\}$ is nearly constant, namely, the series $\{Y_t\}$ is nearly a white noise series. Applying the smoothing techniques to $\{Y_t\}$ will not create large biases. Hence, the estimate

$$\widehat{g}_X(\omega) = |\Gamma(\omega)|^{-2} \widehat{g}_Y(\omega) \tag{7.14}$$

will have more acceptable biases. This idea is referred to as the spectral density estimate by *prefiltering* or *prewhitening*. It was proposed by Press and Tukey (1956).

The choice of filter is very critical for achieving the goal of the bias reduction. Inspecting (7.14), if $g_Y(\cdot)$ is nearly a constant, then the function $|\Gamma(\omega)|^{-2}$ should be nearly proportional to $g_X(\omega)$, which is unknown to us. Typically, the filter has been determined by ad hoc methods, which aim at reducing sharp peaks of the spectral density g_X. A data-driven procedure is to find coefficients $a_1, \cdots, a_p$ that minimize

$$\sum_{t=p}^{T}(X_t - a_1 X_{t-1} - \cdots - a_p X_{t-p})^2$$

and then to form the filtered series

$$Y_t = X_t - a_1 X_{t-1} - \cdots - a_p X_{t-p} \quad \text{for } t = p+1, \cdots, T.$$

When the series $\{X_t\}$ can be approximated by an AR(p) model, the series $\{Y_t\}$ is basically noise. Hence, the filter achieves the stated objective.

A procedure of similar character, but not requiring any filtering of the data, is as follows. Observe that (7.13) can be approximated as

$$\widehat{g}_h(\omega) \approx |\Gamma(\omega)|^{-2} \frac{\sum_{k=1}^{n} K_h(\omega - \omega_k) I_T^*(\omega_k) \Gamma(\omega_k)^2}{\sum_{k=1}^{n} K_h(\omega - \omega_k)},$$

noting that when $\omega_k \approx \omega$, $\Gamma(\omega_k) \approx \Gamma(\omega)$. This in essence estimates the spectral density of Y directly by

$$\widehat{g}_Y(\omega) = \frac{\sum_{k=1}^{n} K_h(\omega - \omega_k) I_T^*(\omega_k) \Gamma(\omega_k)^2}{\sum_{k=1}^{n} K_h(\omega - \omega_k)}.$$

The spectral density g_X is estimated via (7.14).

The idea of prewhitening was extended by Hjort and Glad (1995), Efron and Tibshirani (1996), Glad (1998), and Mays, Birch and Starnes (2000) to reduce biases in estimating density and regression functions.

7.3 Automatic Estimation of Spectral Density

There are three commonly-used ways to estimate the spectral density. To motivate the procedures, we ignore the negligible terms $R_T(\omega_k)$ and r_k in (7.1) and (7.2). For (7.1), this leads to

$$EI_T^*(\omega_k) = g(\omega_k), \quad \text{Var}\{I_T^*(\omega_k)\} = g(\omega_k)^2.$$

Thus, the model (7.1) can be regarded as a heteroscedastic nonparametric regression problem based on the data $\{(\omega_k, I_T^*(\omega_k))\}$. This results in

smoothing on the periodogram $\{I_T^*(\omega_k)\}$. The procedure will be called a (least-square) *smoothed periodogram.* The second approach is to regard model (7.2) as a homoscedastic nonparametric regression model. This results in an estimate of the log-spectral density $m(\cdot)$ by smoothing the log-periodogram $\{Y_k\}$. We will refer to this procedure as a (least-square) *smoothed log-periodogram.* Note that the distribution of z_k in (7.2) is given in (7.3) and is skewed. The least-squares-based estimator is in fact inefficient. To gain efficiency, we will employ the maximum likelihood approach, resulting in a *local likelihood* estimator. The likelihood function is constructed from model (7.2) by regarding $r_k = 0$ and is often called the *Whittle likelihood* (Whittle 1962). Note that the Whittle likelihood based on model (7.2) is the same as that based on model (7.1).

Most of the traditional approaches are based on the smoothed periodogram. See Brillinger (1981) and Priestley (1981) and references therein. Because of high heteroscedasticity when $g(\cdot)$ varies significantly, smoothing on a periodogram using only a constant bandwidth is not effective. On the other hand, the least squares smoothing for a log-periodogram, as pointed out above, is not efficient because of the nonnormal distribution. We recommend using the local likelihood estimator as a spectral density estimator.

7.3.1 Least-Squares Estimators and Bandwidth Selection

The *smoothed periodogram* applies a smoothing method directly to the data $\{(\omega_k, I_T^*(\omega_k))\}$, ignoring the heteroscedasticity of the data. Asymptotically, the heteroscedasticity does not play any important role since smoothing is conducted locally and hence the data in a small window are nearly homoscedastic. However, this asymptotic theory does not necessarily kick in because the local smoothing window can be reasonably large. In other words, the heteroscedasticity influences somewhat the efficiency of the smoothing for finite sample sizes.

We apply the local linear smoother to the data $\{(\omega_k, I_T^*(\omega_k))\}$ for simplicity. Since the design points $\{\omega_k\}$ are equally spaced over $[0, \pi]$, the key advantage of the local linear fit is its boundary behavior, compared with the traditional kernel approach (7.13). As in (6.26), for a given point ω, let K_T be the effective kernel of the local linear fit ($p = 1$ and $\nu = 0$); namely,

$$K_T(t, \omega) = \frac{1}{h} \cdot \frac{S_{T,2}(\omega) - htS_{T,1}(\omega)}{S_{T,0}(\omega)S_{T,2}(\omega) - S_{T,1}^2(\omega)} K(t) \tag{7.15}$$

with

$$S_{T,j}(\omega) = \sum_{k=1}^{n} K_h(\omega_k - \omega)(\omega_k - \omega)^j, \quad (j = 0, 1, 2).$$

The smoothed periodogram is then given by

$$\widehat{g}_{\mathrm{DLS}}(\omega) = \sum_{j=1}^{n} K_T\left(\frac{\omega - \omega_j}{h}, \omega\right) I_T^*(\omega_j), \tag{7.16}$$

which is the local linear fit to the data $\{(\omega_k, I_T^*(\omega_k))\}$.

To implement the smoothed periodogram estimator (7.16), one needs to choose the bandwidth h. Since many interesting periodograms admit different degrees of smoothness around different frequencies, variable bandwidth methods are more effective. An automatic scheme for selecting a variable bandwidth is given in Fan and Gijbels (1995) and is implemented by Fan and Kreutzberger (1998) for the spectral density estimation.

For each given ω, ignoring the term $R_T(\omega_k)$, by (6.30) and noting that the design density of $\{\omega_k\}$ is $f(\omega) = \pi^{-1}$, one can easily obtain the asymptotic normality of $\widehat{g}_{\mathrm{DLS}}(\omega)$. The next theorem formally shows that the term $R_T(\omega_k)$ is indeed negligible.

Theorem 7.1 *Under Conditions (i), (iii), and (iv) in §7.5.1, if* $g(\omega) > 0$, *then*

$$\sqrt{nh}\{\widehat{g}_{DLS}(\omega) - g(\omega) - h^2 g''(\omega)\mu_2(K)/2 + o(h^2)\} \xrightarrow{D} N\{0, \nu_0(K) g^2(\omega)\pi\}$$

for $\omega \in (0, \pi)$, *where* $\mu_2(K) = \int_{-\infty}^{+\infty} u^2 K(u)du$ *and* $\nu_0(K) = \int_{-\infty}^{+\infty} K^2(u)du$.

To obtain the smoothed log-periodogram, we first note that

$$E(z_k) = C_0 = -0.57721 \quad \text{and} \quad \mathrm{Var}(z_k) = \pi^2/6, \tag{7.17}$$

where C_0 is Euler's constant; see Davis and Jones (1968). Thus, the log-periodogram is a biased estimator for the log-spectral density, and the bias does not vanish even when $T \to \infty$. They differ by an amount $-C_0$. Ignoring the term r_k in (7.1) and correcting the bias $-C_0$ leads to

$$Y_k - C_0 = m(\omega) + (z_k - C_0). \tag{7.18}$$

Model (7.18) is a canonical nonparametric regression model with a uniform design density on $[0, \pi]$ and a homogeneous variance $\pi^2/6$. Thus, one can apply the local linear estimator to the data $\{(\omega_k, Y_k - C_0)\}$ to obtain an estimate of $m(\cdot)$. This leads to the estimator

$$\widehat{m}_{\mathrm{LS}}(\omega) = \sum_{j=1}^{n} K_T\left(\frac{\omega - \omega_j}{h}, \omega\right)(Y_j - C_0). \tag{7.19}$$

Applying (6.30) to the model (7.18) and using (7.17), we have the following theorem.

Theorem 7.2 *Under the conditions in §7.5.1, we have, for each* $0 < \omega < \pi$,

$$\sqrt{nh}\{\widehat{m}_{LS}(\omega) - m(\omega) - h^2 m''(\omega)\mu_2(K)/2 + o(h^2)\} \xrightarrow{D} N\{0, (\pi^3/6)\nu_0(K)\}.$$

The result above is rigorously proved by Fan and Kreutzberger (1998) and will be reproduced in §7.5.4. Both Theorems 7.1 and 7.2 are applicable to a boundary point $\omega_T^* = ch$. To this end, let $\mu_{j,c} = \int_{-\infty}^{c} t^j K(t)dt$ and

$$\mu_2(K,c) = \frac{\mu_{2,c}^2 - \mu_{1,c}\mu_{3,c}}{\mu_{0,c}\mu_{2,c} - \mu_{1,c}^2},\ \nu_0(K,c) = \frac{\int_{-\infty}^{c}(\mu_{2,c} - \mu_{c,1}t)^2 K^2(t)\,dt}{(\mu_{0,c}\mu_{2,c} - \mu_{1,c}^2)^2}.$$

We then have

$$\sqrt{nh}\{\widehat{m}_{\mathrm{LS}}(\omega_T^*) - m(\omega_T^*) - h^2 m''(0+)\mu_2(K,c)/2 + o(h^2)\} \xrightarrow{D} N\{0, (\pi^2/6)\nu_0(K,c)\pi\}.$$

A similar extension for the estimator $\widehat{g}_{\mathrm{DLS}}(\omega^*)$ can easily be made.

The asymptotically optimal bandwidth for $\widehat{m}_{\mathrm{LS}}$, which minimizes the integrated asymptotic squared bias and variance, is given by (see also (6.33))

$$h_{\mathrm{LS,\ OPT}} = \left[\frac{\nu_0(K)(\pi^3/6)}{\mu_2^2(K)\int_0^\pi \{m''(\omega)\}^2 d\omega}\right]^{1/5} n^{-1/5}. \tag{7.20}$$

This bandwidth can be estimated by using the preasymptotic substitution method in §6.3.5, yielding an automatic procedure for estimating spectral densities. Here, a constant bandwidth is used. We recommend using a constant bandwidth because its data-driven version can be estimated more reliably and, furthermore, the log-spectral densities usually do not vary as dramatically as the spectral densities themselves.

7.3.2 Local Maximum Likelihood Estimator

The smoothed periodogram estimator $\widehat{m}_{\mathrm{LS}}$ is not efficient because the distribution z_k is not normal. In fact, the Fisher information for the location model (7.18) is 1, while the variance is $\pi^2/6 = 1.645$; see (7.17). Thus, the efficiency of the least-squares method can be improved by a factor of $\pi^2/6$ by using the likelihood method.

By (7.3), model (7.18) gives the log-likelihood

$$\sum_{k=1}^{n}[-\exp\{Y_k - m(\omega_k)\} + Y_k - m(\omega_k)].$$

This likelihood is equivalent to the Whittle likelihood based on the exponential distribution model

$$I_T^*(\omega_k) \sim \text{Exponential}\{g(\omega_k)\}$$

(see (7.1)). Using the local data around a given point ω and the local linear model $m(\omega_k) \approx \alpha + \beta(\omega_k - \omega)$, we form the local log-likelihood

$$\mathcal{L}(\alpha, \beta) = \sum_{k=1}^{n} [-\exp\{Y_k - \alpha - \beta(\omega_k - \omega)\} + Y_k - \alpha - \beta(\omega_k - \omega)] K_h(\omega_k - \omega), \tag{7.21}$$

where $K_h(\cdot) = K(\cdot/h)/h$. Let $\widehat{\alpha}$ and $\widehat{\beta}$ be the maximizers of (7.21). The proposed local likelihood estimator for $m(\omega)$ is $\widehat{m}_{\text{LK}}(\omega) = \widehat{\alpha}$.

The local likelihood (7.21) is strictly concave in α and β, so the maximizer exists and is unique. The maximizer can be found by the *Newton–Raphson* algorithm or the *Fisher scoring* method. Let $\boldsymbol{\beta} = (\alpha, \beta)^T$ and $\mathcal{L}'(\boldsymbol{\beta})$ and $\mathcal{L}''(\boldsymbol{\beta})$ be the gradient vector and the Hessian matrix of the function $\mathcal{L}(\boldsymbol{\beta})$. Then, the local likelihood estimator solves the likelihood equation $\mathcal{L}'(\widehat{\boldsymbol{\beta}}) = 0$. For a given initial value $\widehat{\boldsymbol{\beta}}_0$, by Taylor's expansion,

$$\mathcal{L}'(\widehat{\boldsymbol{\beta}}) \approx \mathcal{L}'(\widehat{\boldsymbol{\beta}}_0) + \mathcal{L}''(\widehat{\boldsymbol{\beta}}_0)(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_0).$$

Hence, after ignoring the approximation error,

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_0 - \mathcal{L}''(\widehat{\boldsymbol{\beta}}_0)^{-1} \mathcal{L}'(\widehat{\boldsymbol{\beta}}_0).$$

The Newton–Raphson algorithm simply iterates the equation above, while the Fisher scoring method replaces the Hessian matrix $L''(\beta_0)$ by its expectation (namely, the negative Fisher information matrix) in the iteration. The estimator $\widehat{m}_{\text{LS}}(\omega)$ and its associated derivative estimator from the local linear fit can serve as good initial values for the algorithm. Indeed, according to Fan and Chen (1999), with the initial value $\widehat{m}_{\text{LS}}(\omega)$ and its associated derivative estimator, the estimator obtained by only one-step iteration from the Newton–Raphson algorithm is efficient. This is an extension of a result by Bickel (1975).

The following result, due to Fan and Kreutzberger (1998), will be proved in §7.5.5.

Theorem 7.3 *Under the conditions in §7.5.1, for each* $0 < \omega < \pi$,

$$\sqrt{nh}\{\widehat{m}_{LK}(\omega) - m(\omega) - h^2 m''(\omega)\mu_2(K)/2 + o(h^2)\} \xrightarrow{D} N\{0, \nu_0(K)\pi\}$$

and, for a boundary point $\omega_T^* = ch$, *we have*

$$\sqrt{nh}\{\widehat{m}_{LK}(\omega_T^*) - m(\omega_T^*) - h^2 m''(0+)\mu_2(K, c)/2 + o(h^2)\} \xrightarrow{D} N\{0, \nu_0(K, c)\pi\}.$$

The maximum likelihood estimator for the spectral density is given by

$$\widehat{g}_{\mathrm{LK}}(\omega) = \exp\{\widehat{m}_{\mathrm{LK}}(\omega)\}.$$

By a Taylor expansion, one can easily see that

$$\widehat{g}_{\mathrm{LK}}(\omega) - g(\omega) \approx g(\omega)\{\widehat{m}_{\mathrm{LK}}(\omega) - m(\omega)\}.$$

Hence

$$\begin{aligned}\sqrt{nh}\{\widehat{g}_{\mathrm{LK}}(\omega) - g(\omega) - h^2 m''(\omega)g(\omega)\mu_2(K)/2 + o(h^2)\} \\ \xrightarrow{D} N\{0, \nu_0(K)g^2(\omega)\pi\}. \end{aligned} \tag{7.22}$$

Compared with $\widehat{m}_{\mathrm{LS}}$, the asymptotic variance of $\widehat{m}_{\mathrm{LK}}$ is a factor of $\pi^2/6$ smaller, while both estimators share the same asymptotic bias. In other words, $\widehat{m}_{\mathrm{LS}}$ is asymptotically inadmissible. On the other hand, the maximum likelihood estimator for the spectral density has the same asymptotic variance as that of the smoothed periodogram $\widehat{g}_{\mathrm{LS}}(\omega)$. However, their biases are different. Using

$$g''(\omega) = g(\omega)m''(\omega) + g(\omega)\{m'(\omega)\}^2,$$

$\widehat{g}_{\mathrm{DLS}}$ has larger biases than $\widehat{g}_{\mathrm{LK}}$ at convex regions of m, namely where $m''(\omega) > 0$. Furthermore, as pointed out before, the inhomogeneous degree of smoothness of the spectral densities and heteroscedasticity of the periodograms make it hard for smoothed periodograms to estimate the underlying spectral densities efficiently. In addition, the estimate around the peak regions is very unstable. Indeed, the tail of the exponential distribution is not that light, and hence the variability of the periodograms at the peak region is also large. Thus, outliers can often be observed around peaks, which impact significantly on the local least-squares estimate.

The comparisons above provide stark evidence for using the local likelihood estimator $\widehat{m}_{\mathrm{LK}}(\omega)$. Its bandwidth can be selected via the least-squares method. From Theorem 7.3, the asymptotically optimal bandwidth for $\widehat{m}_{\mathrm{LK}}$ is given by

$$h_{\mathrm{LK,\ OPT}} = (6/\pi^2)^{1/5}\, h_{\mathrm{LS,\ OPT}} = 0.9053\, h_{\mathrm{LS,\ OPT}}, \tag{7.23}$$

where $h_{\mathrm{LS,\ OPT}}$ is given by (7.20). Thus, an obvious estimator for $h_{\mathrm{LK,\ OPT}}$ is

$$\widehat{h}_{\mathrm{LK,\ OPT}} = 0.9053\, \widehat{h}_{\mathrm{LS,\ OPT}},$$

where $\widehat{h}_{\mathrm{LS,\ OPT}}$ is the preasymptotic substitution bandwidth estimator.

In summary, the local maximum likelihood estimators for spectral and log-spectral densities are recommended. To implement them, first treat the data $\{(\omega_k, Y_k - C_0)\}$ as an independent sample and apply the local linear techniques to obtain the estimate $\widehat{m}_{\mathrm{LS}}(\omega)$, its associated derivative

estimator, and the optimal bandwidth $\widehat{h}_{\mathrm{LS,\,OPT}}$. Now, use the bandwidth $\widehat{h}_{\mathrm{LK,\,OPT}} = 0.9053\,\widehat{h}_{\mathrm{LS,\,OPT}}$ to obtain an estimate $\widehat{m}_{\mathrm{LK}}(\omega)$ using $\widehat{m}_{\mathrm{LS}}(\omega)$ and its associated derivative estimator as an initial values. Indeed, theoretically, one-step iteration starting from $\widehat{m}_{\mathrm{LS}}(\omega)$ suffices.

We now use the data on yields of the three-month Treasury bill in Example 1.3 to illustrate the proposed procedure. The computation was done with the C-code "spectrum.c". The logarithm of the periodogram is depicted in Figure 7.5 (a). The line there is the pointwise 95% confidence upper limit for the log-spectral density $m(\omega_k)$ above the level of the spectral density $\log(\widehat{\sigma}^2/(2\pi))$, which is computed as

$$\log\left(\frac{\widehat{\sigma}^2}{2\pi}\right) + \log(-\log(0.05)),$$

where $\widehat{\sigma}$ is the sample standard deviation of the data. The bandwidth $\widehat{h}_{\mathrm{LS,\,OPT}} = 0.034$ was selected. The pointwise confidence interval was constructed according to (7.26) below. Clearly, most of the energy is concentrated at the very low frequencies. To examine the rate of decay, in Figure 7.5 (c) we plot the log-periodogram against the log-frequency for small frequency values. The pattern is reasonably linear. From the least-squares fit, it suggests that the estimated spectral density behaves as $g(\omega) = \exp(-3.93)\omega^{-1.87}$ around $\omega \approx 0$. This behavior can be understood as follows. The weekly change in interest rates is very small. Thus, the interest rate is a relatively smooth process that contains high energy at low frequencies. We will analyze the difference series in the next section.

We now examine the spectral densities for the acceleration readings during the crashes of vehicles. The first 50 readings (corresponding to the first 76 ms since the crash) of Figures 1.6 (a) and (c) are used to estimate spectral densities. The remaining readings are unlikely to be related to the severity of crashes. The estimated spectral densities are depicted in Figure 7.6. It appears that both spectral densities are similar. However, the spectral density for the crash that does not require deployment of an airbag has higher energy at low frequencies. This in turn suggests that the acceleration readings oscillate less. Thus, oscillation may be one of the features that differentiate between deployment and nondeployment crashes of vehicles.

7.3.3 Confidence Intervals

Confidence intervals are useful for assessing sampling variability and for testing whether a given series is white noise. By ignoring the bias (see §6.3.4 for more discussion), it follows from Theorem 7.3 that an approximate level $1-\alpha$ confidence interval for $m(\omega)$ is

$$\widehat{m}_{\mathrm{LK}}(\omega) \pm z_{1-\alpha/2}\sqrt{\frac{\|K\|_2^2\pi}{nh}}. \tag{7.24}$$

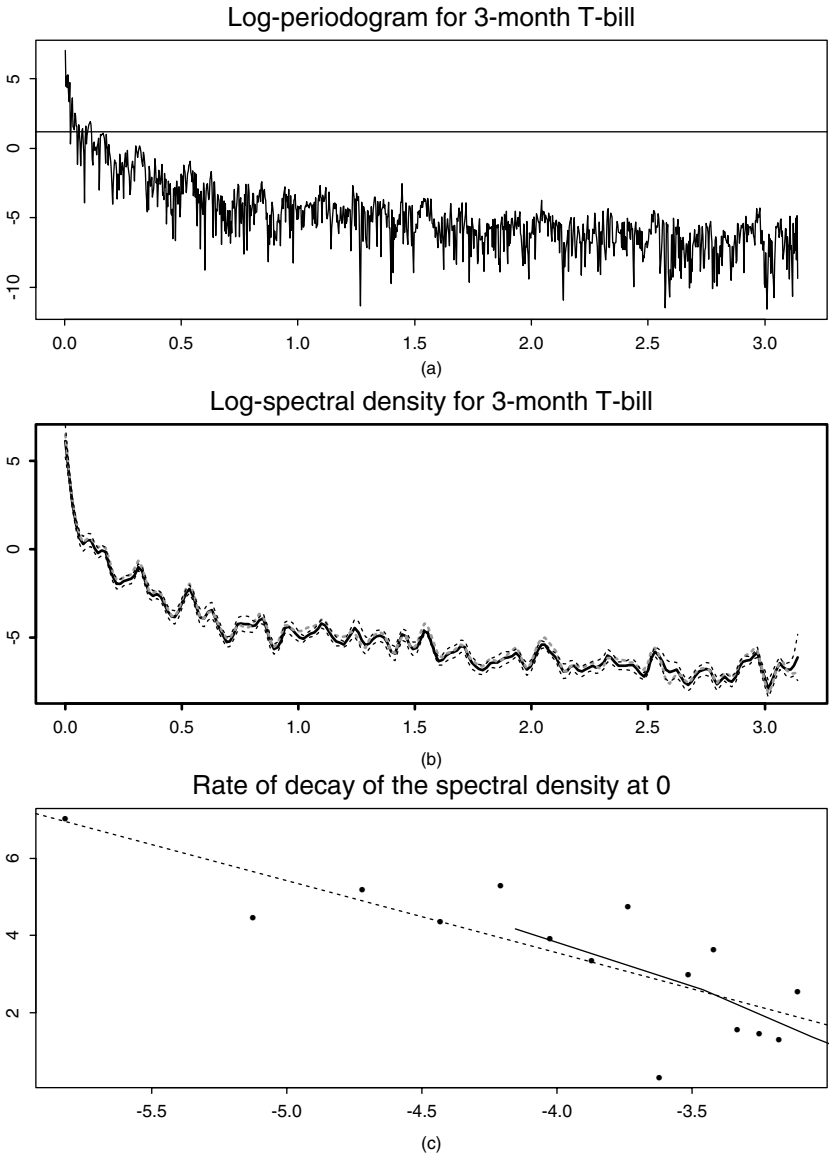


FIGURE 7.5. (a) Log-periodogram of $\{Y_k - C_0\}$ against its Fourier frequency $\{\omega_k\}$. The bar indicates a 95% confidence upper limit above $\log(\frac{\widehat{\sigma}^2}{2\pi})$. (b) Estimated log-spectral density by the local likelihood method $\widehat{m}_{\mathrm{LK}}$ (solid curve) and the least-squares method $\widehat{m}_{\mathrm{LS}}$ (thick dashed-curve) along with the 95% pointwise confidence intervals (7.26) (long dashed-curve). (c) The scatterplot of the log-periodogram of $\{Y_k - C_0\}$ against $\{\log(\omega_k)\}$ at low frequencies $\omega_k \leq 30\pi/2112$ along with the least-squares fit (dashed line). The solid curve is the plot of the estimated log-spectral density against its log-frequency.

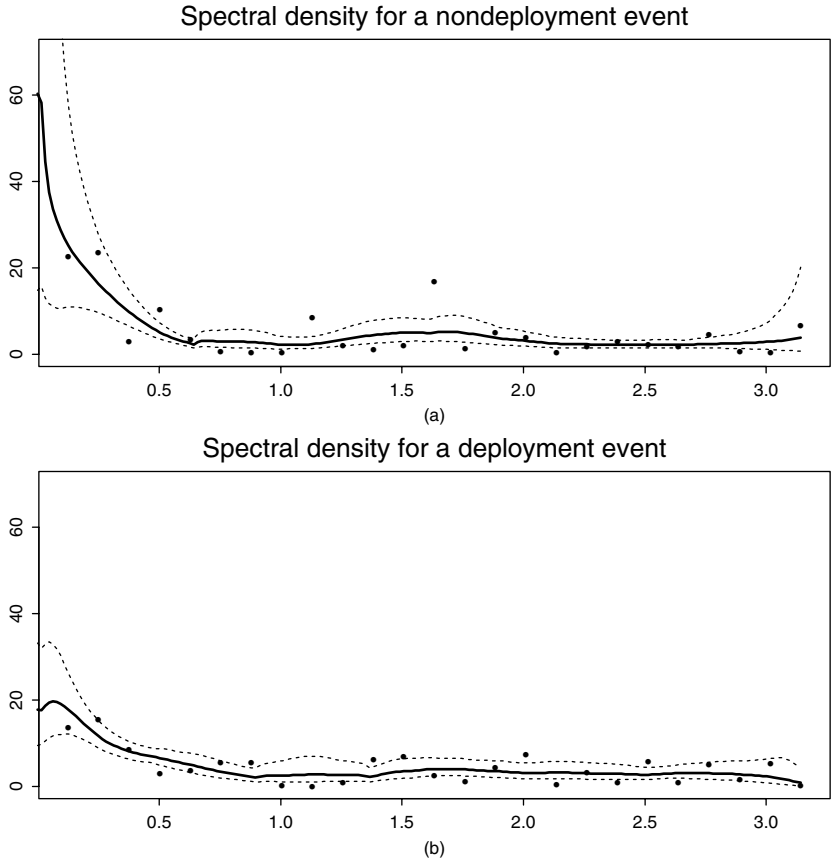


FIGURE 7.6. Estimated spectral densities (thick curves) for car crashes data presented in Figures 1.6 (a) and (c). The dots are the periodogram $\{Y_k\}$ against its Fourier frequency $\{\omega_k\}$. The thin curves are 95% confidence intervals.

Note that this width is known and independent of data. The value of $\|K\|^2$ can be found in Table 5.1. By (7.22), after ignoring the bias term, an approximate level $1-\alpha$ confidence interval is

$$\widehat{g}_{\mathrm{LK}}(\omega)\left\{1 \pm z_{1-\alpha/2}\sqrt{\frac{\|K\|_2^2\pi}{nh}}\right\}. \tag{7.25}$$

The two formulas (7.24) and (7.25) above apply only to interior points: $\omega \pm h \in [0,\pi]$. For boundary points, one needs to replace $\|K\|^2$ by $\nu_0(K,c)$.

The formulas (7.24) and (7.25) are based on the asymptotic variances. As discussed in §6.3.4, the asymptotic variance can also be obtained via (6.34). Translating this formula into the current setting leads to an estimate for

the variance of $\widehat{m}_{\mathrm{LS}}(\omega)$:

$$\frac{\pi^2}{6}\sum_{j=1}^{n} K_T\left(\frac{\omega-\omega_j}{h},\omega\right)^2.$$

Indeed, this formula can be obtained directly from (7.19) by regarding $\{Y_j\}$ as an independent sample. Thus, an alternative asymptotic level $1-\alpha$ confidence interval for $m(\omega)$ is

$$\widehat{m}_{\mathrm{LK}}(\omega) \pm z_{1-\alpha/2}\left\{\sum_{j=1}^{n} K_T\left(\frac{\omega-\omega_j}{h},\omega\right)^2\right\}^{1/2} \tag{7.26}$$

since the asymptotic variance of $\widehat{m}_{\mathrm{LK}}(\omega)$ is a factor of $\pi^2/6$ smaller than $\widehat{m}_{\mathrm{LS}}(\omega)$. Similarly, an approximate level $1-\alpha$ confidence interval for $g(\omega)$ is

$$\widehat{g}_{\mathrm{LK}}(\omega)\left[1 \pm z_{1-\alpha/2}\left\{\sum_{j=1}^{n} K_T\left(\frac{\omega-\omega_j}{h},\omega\right)^2\right\}^{1/2}\right]. \tag{7.27}$$

Confidence intervals (7.26) and (7.27) are both applicable for interior points and boundary points. Note that the confidence interval for $g(\omega)$ can also be obtained directly by exponentiation of the confidence interval (7.26), leading to

$$\exp\left[\widehat{m}_{\mathrm{LK}}(\omega) \pm z_{1-\alpha/2}\left\{\sum_{j=1}^{n} K_T\left(\frac{\omega-\omega_j}{h},\omega\right)^2\right\}^{1/2}\right].$$

By Taylor's expansion, one can easily see that this interval is approximately the same as that given by (7.27). However, when the width of the interval is wide, they may not be equivalent. The latter interval has the advantage that the confidence lower limit is always nonnegative. For this reason, it is implemented in this book.

An application of estimating the spectral density is to examine whether a given time series is a white noise process. Suppose that a series is white noise so that its spectral density is $g(\omega) = \sigma^2/(2\pi)$. Under this null hypothesis, $\widehat{m}_0(\omega) = \log\widehat{\sigma}^2/(2\pi)$ should fall in the confidence interval (7.24) or (7.26) with probability approximately $1-\alpha$, where $\widehat{\sigma}$ is the sample standard deviation of the series. This is equivalent to checking whether $\widehat{m}_{\mathrm{LK}}(\omega)$ falls in the interval

$$\log\widehat{\sigma}^2/(2\pi) \pm z_{1-\alpha/2}\sqrt{\frac{\|K\|_2^2\pi}{nh}} \tag{7.28}$$

or

$$\log\widehat{\sigma}^2/(2\pi) \pm z_{1-\alpha/2}\left\{\sum_{j=1}^{n} K_T\left(\frac{\omega-\omega_j}{h},\omega\right)^2\right\}^{1/2}. \tag{7.29}$$

If one needs to apply the test above to frequencies $\{\omega_j^*, j = 1, \cdots, q\}$ simultaneously, simultaneous confidence intervals for $\{m(\omega_j^*), j = 1, \cdots, q\}$ are needed. This is usually done by the *Bonferroni adjustments.* The idea is very simple. Suppose that I_j $(j = 1, \cdots, q)$ is a level $1 - \alpha_j$ confidence interval for a parameter θ_j; namely,

$$P(\theta_j \in I_j) \geq 1 - \alpha_j.$$

Then, we have the following probability for simultaneous confidence intervals:

$$\begin{aligned} P(\theta_1 \in I_1, \cdots, \theta_q \in I_q) &= 1 - P(\cup_j \{\theta_j \notin I_j\}) \\ &\geq 1 - \sum_{j=1}^{q} P\{\theta_j \notin I_j\} \\ &\geq 1 - \sum_{j=1}^{q} \alpha_j. \end{aligned}$$

By taking $\alpha_j = \alpha/q$, we have that θ_j falls in I_j simultaneously with probability at least $1 - \alpha$. By using this and (7.24), we obtain the following $1 - \alpha$ simultaneous confidence intervals for $\{m(\omega_j^*), j = 1, \cdots, q\}$

$$\widehat{m}_{\mathrm{LK}}(\omega_j^*) \pm z_{1-\alpha/(2q)} \sqrt{\frac{\|K\|_2^2 \pi}{nh}}, \quad j = 1, \cdots, q.$$

In turn, this leads to checking whether all $\widehat{m}_{\mathrm{LK}}(\omega_q^*)$ fall simultaneously in the intervals

$$\log \frac{\widehat{\sigma}^2}{2\pi} \pm z_{1-\alpha/(2q)} \sqrt{\frac{\|K\|_2^2 \pi}{nh}}, \quad j = 1, \cdots, q. \tag{7.30}$$

One can also extend (7.29) to the situation with multiple comparisons, leading to

$$\log \frac{\widehat{\sigma}^2}{2\pi} \pm z_{1-\alpha/(2q)} \left\{ \sum_{k=1}^{n} K_T \left(\frac{\omega_j^* - \omega_k}{h}, \omega_j^* \right)^2 \right\}^{1/2}, \quad j = 1, \cdots, q. \tag{7.31}$$

We now revisit the spectral aspect of the interest rate data. As discussed in the last section, large spectrum values at low frequencies are mainly due to the relatively small weekly changes of interest rates. This leads to considering the difference series and to examining whether the difference series is white noise. Figure 7.7(a) shows the estimated spectral density as well as 95% associated confidence intervals for testing whether the series is a white noise series. The bandwidth 0.034 was selected by our software. On a large portion of regions, the estimated spectral density lies outside the

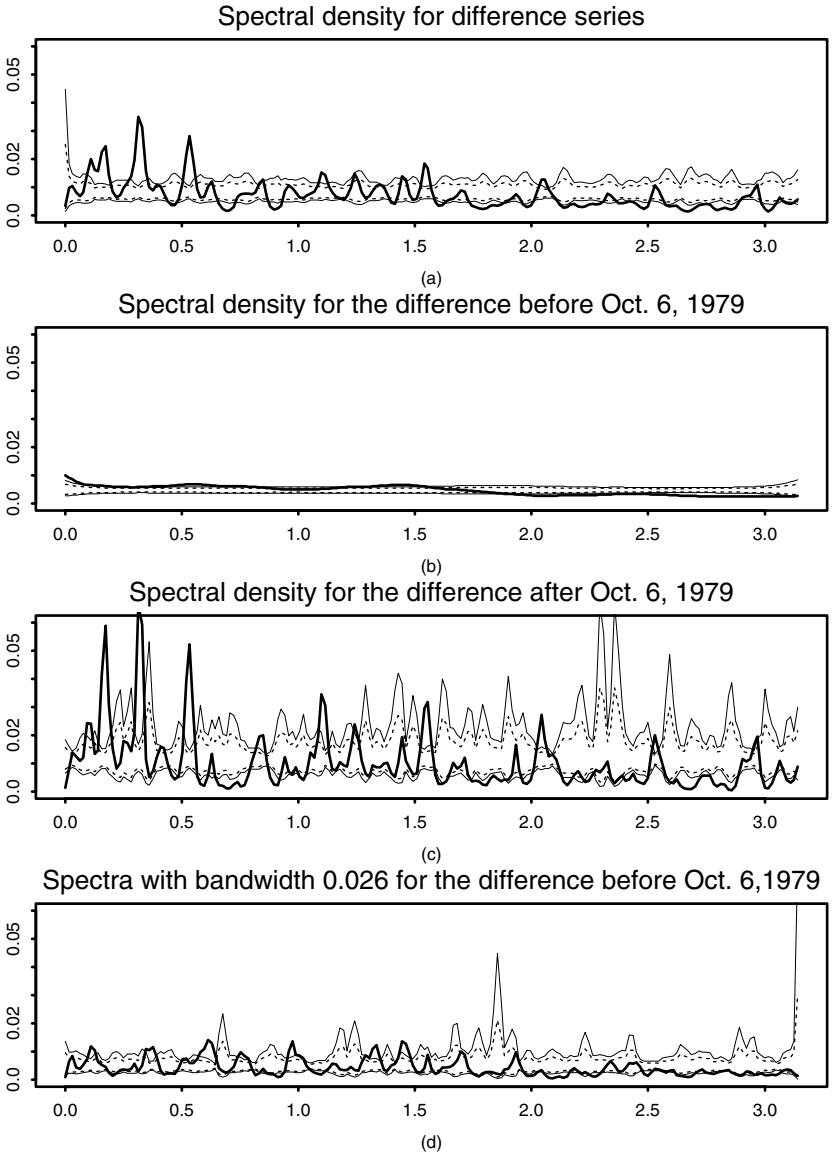


FIGURE 7.7. Estimated spectral densities (thick curves) for the differenced yields of the 3-month Treasury bill and their associated 95% pointwise confidence intervals (7.29) (dashed curves) and 95% simultaneous confidence intervals (7.31) (thin curves) at $q = 15$ different locations for testing whether the difference series is white noise; (a) for the difference series from July 17, 1959 to December 31, 1999; (b) for the difference series from July 17, 1959 to October 5, 1979; (c) for the difference series from October 12, 1979 to December 31, 1999; (d) for the difference series from July 17, 1959 to October 5, 1979 using the same bandwidth as in (c).

simultaneous confidence bands. This provides strong evidence to reject the null hypothesis that the difference series is a white noise process. In fact, the literature on the interest modeling is abundant, see, for example, Cox, Ingersoll and Ross (1985), Chan, Karolyi, Longstaff, and Sanders (1992), and Stanton (1997). None of these models suggest that the difference series is a white noise process. In fact, the variance of the difference series depends on the level of the current interest rate.

The Federal Reserve changed its monetary policy in October 6, 1979 when its newly appointed chairman, Paul Volcker, initiated money supply targeting and abandoned interest rate targeting. The interest rate in the two years following October, 1979 was five times greater than that in the prior two years. To examine whether the interest rate dynamics have changed, we divide the series into two halves. The first half consists of the 20-year data from July 17, 1959 to October 5, 1979, and the second half consists of the 20-year data from October 12, 1979 to December 31, 1999. The volatility over the last twenty years has increased 54%, from 0.172% (the standard deviation of the first twenty years) to 0.265% (the SD of the second twenty years). The estimated spectral densities are presented in Figure 7.7. The spectral density for the second period is higher, which provides further evidence that the interest rate volatility is higher since October, 1979. For the second period of data, there are more significant spectral densities at high frequencies than in the first period. This means that the data in the second period oscillates more than in the first period. The bandwidths 0.342 and 0.026 were selected by the automatic smoothing software for the first and the second periods of data, respectively. This makes the estimated spectral density much smoother for the first period of data. To compare the two estimated densities using the same amount of smoothing, Figure 7.7(d) shows the estimated spectral density for the first period of data using the bandwidth 0.026. The qualitative comments above continue to apply. From the estimated spectral densities and their associated confidence bands, we may conclude that both subseries are not white noise.

In estimating the spectral densities above and their associated confidence intervals, one implicitly assumes that the underlying time series is stationary, which, however, may not be true. If the series is not stationary, the confidence bands are not meaningful. Nevertheless, an estimated spectral density, as a descriptive statistic and a device of spectral decomposition, still provides useful information on the energy distribution over different frequencies for a time series.

7.4 Tests for White Noise

A statistical model is at best only an approximation to the true underlying process that generates the observed data. After fitting a model, one needs to verify various aspects of the assumption. This is usually accomplished by both graphical tools and formal statistical tests. It is a generally accepted principle that a good statistical model is one such that at least the residuals from the fitting behave like a white noise process. In other words, a time series, after extracting the structural part, becomes an unpredictable white noise. For example, after fitting the AR(p) model (1.1), one would expect that the residual series $\{\widehat{\varepsilon}_t\}$ is a white noise process. Systematic departure from this assumption implies the inadequacy of the assumed form of the model. Thus, it is important to develop formal procedures for testing whether a series is white noise.

Different tests explore different aspects of departure from the null hypothesis. Hence, they have different powers against different alternatives. This section aims at introducing some simple and powerful nonparametric procedures. Other related ideas will be further explored in Chapter 9.

We assume throughout this section that the time series $\{X_t\}$ is stationary. Let $g(\omega)$ be its spectral density. Note that $\{X_t\}$ is white noise if and only if its spectral density is constant. Therefore, we only need to test the hypotheses

$$H_0 : g(\omega) = \frac{\sigma^2}{2\pi} \quad \longleftrightarrow \quad H_1 : g(\omega) \neq \frac{\sigma^2}{2\pi}, \tag{7.32}$$

where σ^2 is the variance of $\{X_t\}$. This is a parametric versus nonparametric testing problem. The important raw material is the rescaled periodogram $\{I_T^*(\omega_k)\}$ in (7.1).

In this section, we outline some techniques for testing the problem (7.32). Testing the spectral density of other parametric forms can be found in §9.3. The methods in that section are also applicable to the problem (7.32).

A word of caution: The observed significance level, or the p-value, depends on the sample size. When the sample size is large, a small departure from the null hypothesis may result in a very small p-value. Thus, a small p-value with a large sample size does not necessarily mean that the departure of the model from the null hypothesis is serious.

7.4.1 Fisher's Test

Fisher's test is based on the fact that under H_0 the maximum of the spectral density and its average should be the same. Thus, the large values of the test statistic

$$T_{n,F} = \frac{\max_{1\le k\le n} I(\omega_k)}{n^{-1}\sum_{k=1}^{n} I(\omega_k)} \tag{7.33}$$

indicate the departure from H_0. Hence, we would reject H_0 when $T_{n,F}$ is too large. To obtain the critical value, we need to derive the *null distribution* (i.e., the distribution of the test statistic $T_{n,F}$ under the null hypothesis H_0). The result is summarized as follows.

Theorem 7.4 *Suppose that the conditions of Theorem 2.14 hold. Then, under H_0,*

$$P\{T_{n,F} - \log n \le x\} \to \exp(-\exp(-x)), \quad -\infty < x < \infty.$$

Proof. Let $g_0 = \frac{\sigma^2}{2\pi}$ be the spectral density under the null hypothesis. By (7.1) and Theorem 2.14, we have

$$\max_{1\le k\le n} I_T(\omega_k)/(2\pi) = g_0 \max_{1\le k\le n} V_k + o_P(1)$$

and, by the law of large numbers,

$$n^{-1}\sum_{k=1}^{n} I_T(\omega_k)/(2\pi) = g_0 n^{-1}\sum_{k=1}^{n} V_k + o_P(1) = g_0 + o_P(1).$$

Hence

$$T_{n,F} = \max_{1\le k\le n} V_k + o_P(1). \tag{7.34}$$

For any $x \ge -\log n$, we have

$$\begin{aligned} P\left\{\max_{1\le k\le n} V_k - \log n \le x\right\} &= P[V_k \le \log\{n\exp(x)\}]^n \\ &= (1-\exp(-x)/n)^n \\ &\to \exp(-\exp(-x)). \end{aligned}$$

The conclusion follows from (7.34). ∎

The exact null distribution of $T_{n,F}$ can be obtained when $\{X_t\}$ is a Gaussian white noise process; see page 339 of Brockwell and Davis (1991). For simplicity and brevity, we use the asymptotic distribution, which admits a more explicit formula and applies to more general stationary processes. From Theorem 7.4, we have

$$P\Big\{T_{n,F} \le \log n - \log(-\log(1-\alpha))\Big\} \approx 1-\alpha.$$

Thus, an approximate level α test based on $T_{n,F}$ is given by

$$T_{n,F} > \log n - \log(-\log(1-\alpha)). \tag{7.35}$$

Suppose that, based on the available data $\{x_t, t = 1, \cdots, T\}$, the Fisher statistic is $t_{n,F,\text{obs}}$. Then, the *observed significance level* or *p-value* based on the Fisher test for the problem (7.32) is

$$P\{T_{n,F} \ge t_{n,F,\text{obs}}\} \approx 1 - \exp\left(-n\exp(-t_{n,F,\text{obs}})\right). \tag{7.36}$$

The Fisher test is expected to be powerful against the alternatives with energy concentrated around one frequency, namely, the underlying spectral density has a very sharp peak. It is not expected to be powerful for detecting alternatives with more spread energy.

7.4.2 Generalized Likelihood Ratio Test

After ignoring smaller-order terms in (7.2), the problem is basically a nonparametric testing problem with a smoothed alternative. Thus, one can apply the *generalized likelihood ratio test*, recently developed by Fan, Zhang, and Zhang (2001), to our setting. A comprehensive overview of this subject is given in §9.2.

The basic idea of the generalized likelihood ratio statistic is to find a suitable estimate for $m(\omega)$ in (7.2) under H_0 and H_1, respectively, and then to form a likelihood ratio statistic. A reasonable nonparametric estimator of $m(\omega)$ is the local likelihood estimator $\widehat{m}_{\rm LK}(\omega)$. Then, the log-likelihood with given $\widehat{m}_{\rm LK}(\omega)$ is

$$\log L(H_1) = \sum_{k=1}^{n} \{-\exp(Y_k - \widehat{m}_{\rm LK}(\omega_k)) + Y_k - \widehat{m}_{\rm LK}(\omega_k)\}$$

after ignoring the term r_k in (7.2). Using a similar expression for the log-likelihood under H_0, we obtain the *generalized likelihood ratio statistic*

$$\begin{aligned} \lambda_n &= \log L(H_1) - \log L(H_0) \\ &= \sum_{k=1}^{n} \Big\{ \exp(Y_k - \widehat{m}_0) - \exp(Y_k - \widehat{m}_{\rm LK}(\omega_k)) + \widehat{m}_0 - \widehat{m}_{\rm LK}(\omega_k) \Big\}, \end{aligned} \tag{7.37}$$

where $m_0(\omega) = \log \frac{\widehat{\sigma}^2}{2\pi}$, with $\widehat{\sigma}^2$ being the sample variance.

The generalized likelihood ratio statistic above is a natural extension of the maximum likelihood ratio tests for parametric models. However, there are also several fundamental differences. First, the nonparametric estimate $\widehat{m}_{\rm LK}(\cdot)$ is not the (nonparametric) maximum likelihood estimate. Because of this, there is some chance that λ_n can be negative. Indeed, the parameter space under the full model is an infinite-dimensional function space. The (nonparametric) maximum likelihood estimator for $m(\cdot)$ usually does not exist. Even if it exists, it is hard to compute. Furthermore, it is shown by Fan, Zhang, and Zhang (2001) that the maximum likelihood ratio tests for infinite-dimensional problems are not efficient. This is another remarkable difference from the parametric setting. The generalized likelihood ratio statistic, on the other hand, is shown to be asymptotically optimal, with a proper choice of bandwidth, in the sense that it achieves the asymptotic optimal rate of convergence for testing problems formulated by Ingster (1993) and Spokoiny (1996).

TABLE 7.1. Values of r_K and c_K in (7.38). Adapted from Fan, Zhang, and Zhang (2001).

Kernel	Uniform	Epanechnikov	Biweight	Triweight	Gaussian
r_K	1.2000	2.1153	2.3061	2.3797	2.5375
c_K	0.2500	0.4500	0.5804	0.6858	0.7737

Let h be the bandwidth used in constructing $\widehat{m}_{\mathrm{LK}}$ and K be the kernel function. Denote by $K * K$ the convolution function of K. Assuming that $r_k \equiv 0$, by Theorem 10 of Fan, Zhang, and Zhang (2001), if $h \to 0$ and $nh^{3/2} \to \infty$, we have

$$r_K \lambda_n \overset{a}{\sim} \chi^2_{r_K \mu_n}, \tag{7.38}$$

where

$$\mu_n = \frac{\pi}{h}\{K(0) - \|K\|^2/2\}, \quad r_K = \frac{K(0) - \|K\|^2/2}{\|K - K * K/2\|^2}.$$

Here "$\overset{a}{\sim}$" means "distributed approximately." Note that the normalizing constant is r_K, rather than 2 in the classical Wilks theorem. Note further that the degree of freedom tends to be infinite since $h \to 0$. Formally, (6.27) means that

$$\frac{r_K \lambda_n - r_K \mu_n}{\sqrt{2 r_K \mu_n}} \xrightarrow{D} N(0, 1).$$

Table 7.1 shows the value of constants r_K and $c_K = K(0) - \|K\|^2/2$.

The result above has two practical uses for hypothesis testing. First, an approximate level α for the testing problem (7.32) is to reject the null hypothesis when

$$r_K \lambda_n \geq \chi^2_{r_K \mu_n}(1 - \alpha), \tag{7.39}$$

where $\chi^2_{r_K \mu_n}(1 - \alpha)$ is the $(1 - \alpha)$ quantile of the χ^2 distribution with degrees of freedom $[r_K \mu_n]$, the rounding of $r_K \mu_n$ to its nearest integer. Secondly, it permits one to use the *bootstrap* to obtain the null distribution. In this parametric setting, the bootstrap method is indeed the same as the simulation method. Since the asymptotic distribution does not depend on σ, we can take it to be 1. Generate a random sample of size n from the standard exponential distribution. Create a synthetic periodogram by using (7.1); namely, regarding the random sample as a periodogram under the null hypothesis (7.32). Compute the generalized test statistic λ_n. Repeat the simulation above 1,000 times (say) to obtain 1,000 realizations of the generalized likelihood ratio test statistic λ_n. The 95th sample percentile of the realizations can be used as the critical value. Furthermore, the p-value can be estimated as the upper quantile of the observed test statistic in the empirical distribution of these 1,000 realizations. In other words, if there are m realizations of λ_n that are larger than the observed test statistic, then the p-value is simply estimated by $m/1,000$.

7.4.3 χ^2-Test and the Adaptive Neyman Test

A stationary process is white noise if and only if its autocorrelation function is zero for lag 1 and above. This naturally leads to the test statistic $T\sum_{k=1}^{m}\widehat{\rho}(k)^2$. A better approximation can be achieved by using

$$T_m = T(T+2)\sum_{k=1}^{m}\frac{\widehat{\rho}(k)^2}{T-k} \tag{7.40}$$

for a given parameter m; see Box and Pierce (1970) and Ljung and Box (1978). By Theorem 2.8 (see also Box and Pierce 1970; Li, W.K. 1992), when $\{X_t\}$ is an i.i.d. sequence, $\{\widehat{\rho}(k)^2\}$ are asymptotically independent with mean zero and variance T^{-1}. Thus, for a given m, under the null hypothesis that $\{X_t\}$ is an i.i.d. sequence, we have

$$T_m \overset{a}{\sim} \chi^2_m. \tag{7.41}$$

The test statistic T_m examines only the first m autocorrelation coefficients. To make the procedure consistent among a large class of alternatives, we have to make m depend on T and, furthermore, $m \to \infty$ as $T \to \infty$. Furthermore, $\widehat{\rho}(k)$ is not a good estimate of $\rho(k)$ when k is near T.

The parameter m can be regarded as a smoothing parameter. Its choice would affect the power of the test. For stationary time series, we have prior knowledge that the autocorrelation function is small when the lag is large. Thus, testing on all autocorrelation coefficients (namely, taking $m = T-1$) will accumulate stochastic error in T_m and deteriorate its power. Hence $m = T-1$ is not a good choice.

The test statistic T_m is equivalent to its normalized form:

$$\frac{T_m - m}{\sqrt{2m}}.$$

Different values of m result in different test statistics. A natural way to combine these test statistics is to use the multiscale test

$$T^*_{AN} = \max_{1\le m\le a_T}\frac{T_m - m}{\sqrt{2m}},$$

for some upper limit a_T. This test statistic was introduced by Fan (1996) in a somewhat different context based on power considerations. Fan called it the *adaptive Neyman test* and showed that under the null hypothesis

$$P(T_{AN} < x) \to \exp(-\exp(-x)) \quad \text{as } n \to \infty, \tag{7.42}$$

where

$$T_{AN} = \sqrt{2\log\log a_T}\,T^*_{AN} - \{2\log\log a_T + 0.5\log\log\log a_T - 0.5\log(4\pi)\}. \tag{7.43}$$

TABLE 7.2. The α upper quantiles of the distribution $J_n^\dagger$. Taken from Fan and Lin (1998).

$n\backslash\alpha$	0.001	0.0025	0.005	0.01	0.025	0.05	0.10	0.25	0.50
5	7.80	6.74	5.97	5.21	4.23	3.50	2.77	1.77	0.96
10	9.13	7.73	6.77	5.78	4.57	3.67	2.74	1.49	0.40
20	9.83	8.26	7.16	6.07	4.75	3.77	2.78	1.41	0.18
30	10.11	8.47	7.29	6.18	4.82	3.83	2.81	1.39	0.11
40	10.34	8.65	7.41	6.22	4.87	3.85	2.82	1.39	0.08
50	10.32	8.67	7.43	6.28	4.89	3.86	2.84	1.39	0.07
60	10.56	8.80	7.51	6.32	4.91	3.88	2.85	1.39	0.07
70	10.59	8.81	7.55	6.34	4.92	3.88	2.85	1.40	0.06
80	10.54	8.81	7.57	6.37	4.93	3.89	2.85	1.40	0.06
90	10.79	8.95	7.65	6.40	4.94	3.90	2.86	1.40	0.06
100	10.80	8.95	7.65	6.40	4.94	3.90	2.86	1.40	0.06
120	10.87	8.96	7.65	6.41	4.95	3.90	2.87	1.41	0.05
140	10.80	9.00	7.66	6.42	4.95	3.90	2.86	1.41	0.05
160	10.88	8.95	7.69	6.42	4.95	3.91	2.87	1.41	0.06
180	11.02	9.10	7.77	6.47	4.95	3.90	2.87	1.41	0.06
200	11.10	9.08	7.72	6.43	4.95	3.89	2.86	1.42	0.06

The results are based on 1,000,000 simulations. The relative errors are expected to be around 0.3%–3%.

Hence, an approximate level α test based on T_{AN} is to reject the independent noise assumption when $T_{AN} > -\log(-\log(1-\alpha))$.

Fan (1996) noted that the asymptotic distribution in (7.42) is not a good approximation to T_{AN}. Let us denote the exact distribution of T_{AN} under the null hypothesis by J_{a_T}, depending on the parameter a_T. This distribution does not depend on any unknown parameters and can easily be computed by statistical simulation. We have a C-code "aneyman.table.c" available for computing p-values. Table 7.2 is an excerpt from Fan and Lin (1998).

When the adaptive Neyman test above is applied to residuals based on a parametric model (e.g., an ARMA model), some slight modifications are needed. In the ARMA (p, q), Box and Pierce (1970) show that $T_m \sim \chi^2_{m-p-q}$. Thus, one can modify T^*_{AN} accordingly as

$$T^*_{AN} = \max_{p+q+1 \le m \le a_T} \frac{T_m - (m-p-q)}{\sqrt{2(m-p-q)}}.$$

This modified version would have a better approximation of the null distribution.

In a somewhat different setup, Fan, Zhang, and Zhang (2001) show that the adaptive Neyman test is an adaptively optimal test in the sense that

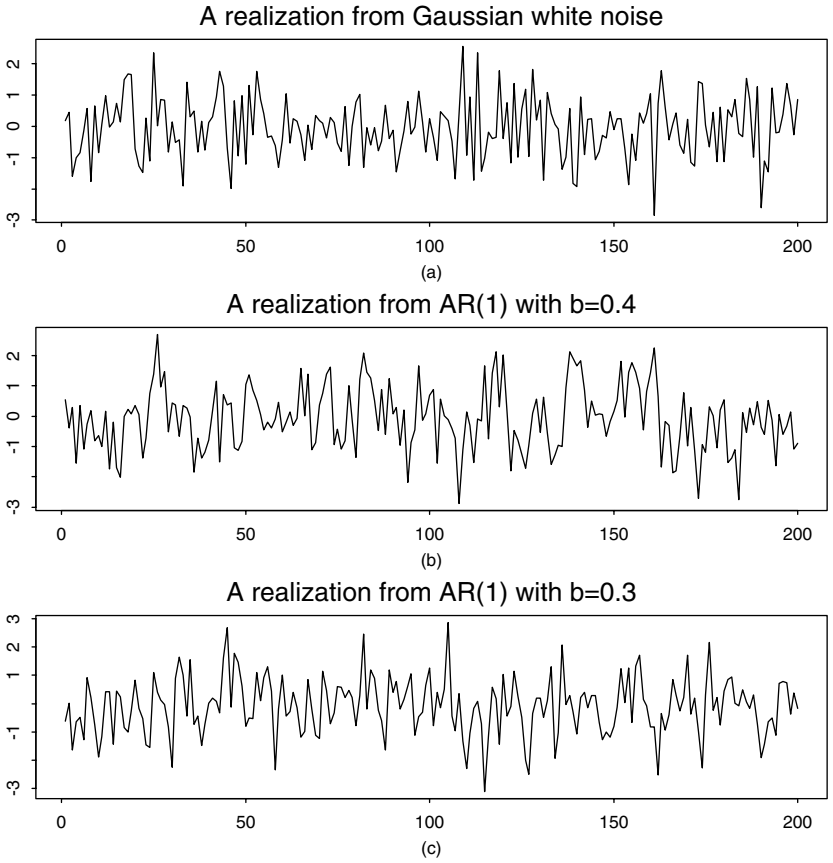


FIGURE 7.8. A realization of length $T = 200$ from (a) a Gaussian white noise model, (b) an AR(1) model with $b = 0.4$, and (c) an AR(1) model with $b = 0.3$.

it achieves adaptively the optimal rate of convergence for nonparametric hypothesis testing with unknown degree of smoothness.

7.4.4 Other Smoothing-Based Tests

After ignoring the smaller order term in (7.2), the problem (7.32) becomes testing whether the regression function in (7.2) is a constant. This problem has been extensively studied; see the books by Bowman and Azzalini (1997) and Hart (1997) and the references therein. The techniques there can also be applied to the current setting. Note that the noise term in (7.2) is not Gaussian, orthogonal transformation methods such as the *Neyman test* and its various adaptive versions are not very convenient to apply.

TABLE 7.3. Results of testing for a Gaussian white noise model based on three simulated data sets.

		Fisher	G-likelihood	A-Neyman	χ^2-test
White noise	statistic	7.177	13.12	1.4728	53.80
	p-value	0.074	0.069	0.239	0.331
	d.f.		7		50
AR(1), $b = 0.4$	statistic	7.847	34.96	27.29	99.48
	p-value	0.038	0.000	0.000	0.001
	d.f.		8		50
AR(1), $b = 0.3$	statistic	5.778	16.71	8.314	59.54
	p-value	0.266	0.010	0.003	0.167
	d.f.		6	50	

7.4.5 Numerical Examples

To gain insights on the various tests, we first simulate three time series of length $T = 200$ from the three models

$$\begin{aligned} \text{White noise:} \quad & X_t = \varepsilon_t, \\ \text{AR(1) with } b = 0.4\text{:} \quad & X_t = 0.4X_{t-1} + \varepsilon_t, \\ \text{AR(1) with } b = 0.3\text{:} \quad & X_t = 0.3X_{t-1} + \varepsilon_t, \end{aligned}$$

where $\{\varepsilon_t\}$ is a sequence of i.i.d. random variables having the standard normal distribution. Figure 7.8 presents a realization from the three models above. Consider the testing problem (7.32) and the following testing procedures: the Fisher test (7.33), the generalized likelihood ratio test (7.37), the adaptive Neyman test (7.43), and the χ^2-test (7.40) with $m = 50$. In the C-code "spectrum.c," the results of the Fisher test and the generalized likelihood ratio test are reported. We have the S-Plus codes "aneyman.s" and "fishertest.s" for computing the adaptive Neyman test and the Fisher test. The p-value of the adaptive Neyman test can be found by using the C-code "aneyman.table.c" or Table 7.2.

The results of the four testing procedures above are reported in Table 7.3. The generalized-likelihood test and the adaptive Neyman test have higher discriminant power, making right decisions at the significance level 5%. Furthermore, the small p-values for the two AR(1) models provide further evidence to support the claim above. On the other hand, as discussed before, the Fisher test and the χ^2-test are less powerful. For the AR(1) model with $b = 0.3$, both made wrong decisions at the significance level 5%.

We now apply the four procedures above to test whether the three difference series in Figure 7.7 are white noise. The results are summarized in Table 7.4. These tests provide further evidence against the hypothesis of white noise. Further, the p-value for the Fisher test is not nearly as small

TABLE 7.4. Testing the Gaussian white noise model for difference series of the yields of the three-month Treasury bill.

		Fisher	G-likelihood	A-Neyman	χ^2-test
Whole series	statistic	10.81	586.4 (d.f.=99)	173.9	478.7
	p-value	0.0211	0.000	0.0000	0.0000
Before 1979	statistic	10.03	284.3 (d.f.=126)	58.75	177.9
	p-value	0.0230	0.000	0.0000	0.0000
After 1979	statistic	10.97	491.7 (d.f.=128)	94.65	404.9
	p-value	0.0090	0.000	0.0000	0.0000

as the three other tests. This is again due to the fact that the Fisher test is less capable of discriminating the alternatives of a nonuniform but more spread-out spectrum (see Figure 7.7).

7.5 Complements

7.5.1 Conditions for Theorems 7.1—-7.3

We first state the technical conditions for Theorems 7.1—-7.3. They are imposed to facilitate technical proofs and are not the minimum possible.

Conditions

(i) The process $\{X_t\}$ is a linear Gaussian process given by

$$X_t = \sum_{j=-\infty}^{\infty} a_j \varepsilon_{t-j}$$

with $\sum_j |a_j| j^2 < \infty$, where $\varepsilon_j \sim$ i.i.d. $N(0, \sigma^2)$.

(ii) The spectral density function $g(\cdot) > 0$ on $[0, \pi]$.

(iii) The kernel function K is a symmetric probability density function and has a compact support.

(iv) $(\log T)^4 h_T \to 0$ in such a way that $T h_T \to \infty$.

It follows from Condition 1(i) and Theorem 2.12 that the spectral density function of $\{X_t\}$ is given by

$$g_X(\omega) = |A(\omega)|^2 f_\varepsilon(\omega) = \frac{\sigma^2}{2\pi} |A(\omega)|^2,$$

where

$$A(\omega) = \sum_{j=-\infty}^{\infty} a_j \exp(-ij\omega).$$

It has a bounded second derivative. We first give two lemmas showing that $R_T(\omega_k)$ in (7.1) and r_k in (7.2) are indeed negligible.

7.5.2 Lemmas

Lemma 7.1 *Under Condition (i), we have*

$$\max_{1\le k\le n} |R_T(\omega_k)| = O\left(\frac{\log T}{\sqrt{T}}\right)$$

almost surely.

Proof. We follow the notation and the proof of Theorem 2.14. By (2.60), the remainder term can be expressed as

$$R_T(\omega_k) = |Y_T(\omega_k)|^2 + A(\omega_k)\alpha_{k,\varepsilon}Y_T(-\omega_k) + A(\omega_k)\bar{\alpha}_{k,\varepsilon}Y_T(\omega_k). \quad (7.44)$$

As shown in Theorem 2.14, $\{\alpha_{k,\varepsilon}\}$ and $\{Y_T(\omega_k)\}$ are independently normally distributed with mean 0 and variance $O(1)$ and $O(T^{-1})$, respectively. Recall that the maximum of n i.i.d. Gaussian white noise is asymptotically equal to $\sqrt{2\log n}$ almost surely. It follows that

$$\max_k |\alpha_{k,\varepsilon}| = O(\sqrt{\log T}) \quad \text{and} \quad \max_k |Y_T(\omega_k)| = O\left(\sqrt{\frac{\log T}{T}}\right)$$

almost surely. Substituting these into (7.44) and using the fact that

$$\max_\omega |A(\omega)| < \infty,$$

we obtain Lemma 7.1. ∎

Lemma 7.2 *Under Conditions (i) and (ii), we have for any sequence c_T,*

$$r_k \le O_P\left(\frac{\log T}{\sqrt{T}}\right)\frac{I(V_k > c_T)}{V_k} + O_P(\log T)I(V_k \le c_T)$$

uniformly for $1 \le k \le n$, where V_k are i.i.d. random variables having the standard exponential distribution.

Proof. Recall that

$$r_k = \log\left\{1 + \frac{R_T(\omega_k)}{g(\omega_k)V_k}\right\}.$$

Using the inequality $\log(1+x) \le x$ for $x > 0$ and dividing the state-space into $V_k > c_T$ and $V_k \le c_T$ for a given sequence c_T, we get that

$$r_k \le \frac{|R_T(\omega_k)|}{g(\omega_k)V_k}I(V_k > c_T) + \log\left\{1 + \frac{\max_k |R_T(\omega_k)|}{\min_\omega g(\omega)\min_k V_k}\right\}I(V_k \le c_T). \quad (7.45)$$

Obviously,

$$P\left(\min_{1\le k\le n} V_k > T^{-2}\right) = \exp(-n/T^2) \to 1.$$

Thus $(\min_k V_k)^{-1} = O_P(T^2)$. Substituting this term into (7.45) and using Lemma 7.1, we obtain

$$r_k \le O_P\left(\frac{\log T}{\sqrt{T}}\right) I(V_k > c_T) V_k^{-1} + O_P(\log T) I(V_k \le c_T).$$

This completes the proof of Lemma 7.2. ∎

7.5.3 Proof of Theorem 7.1

First, by (7.1) and (7.16),

$$\begin{aligned}\widehat{g}_{\mathrm{DLS}}(\omega) &= \sum_{j=1}^{n} K_T\left(\frac{\omega-\omega_j}{h}, \omega\right) g(\omega_j) V_j \\ &\quad + \sum_{j=1}^{n} K_T\left(\frac{\omega-\omega_j}{h}, \omega\right) R_T(\omega_j). \qquad (7.46)\end{aligned}$$

Regarding $g(\omega_j)V_j$ as a response variable Y_j, the first term is the local linear fit based on the data $\{(\omega_j, Y_j)\}$. Applying (6.30), we obtain Theorem 7.2 if we can show that the second term in (7.46) is of order $\frac{\log T}{\sqrt{T}}$. By Lemma 7.1, this in turn requires us to show

$$\sum_{j=1}^{n} \left| K_T\left(\frac{\omega-\omega_j}{h}, \omega\right)\right| = O_P(1). \qquad (7.47)$$

By (7.15), the left-hand side of (7.47) is bounded by

$$\frac{S_{T,2}(\omega)S_{T,0}(\omega) + (\sum_{k=1}^{n} K_h(\omega_k-\omega)|\omega_k-\omega|)^2}{S_{T,2}(\omega)S_{T,0}(\omega) - S_{T,1}(\omega)^2}.$$

By the Cauchy–Schwartz inequality, this is bounded by

$$\frac{2S_{T,2}(\omega)S_{T,0}(\omega)}{S_{T,2}(\omega)S_{T,0}(\omega) - S_{T,1}(\omega)^2}.$$

By (6.28), the quantity above tends to 2. This proves (7.47) and the theorem.

∎

7.5.4 Proof of Theorem 7.2

We will show that

$$\widehat{m}_{\mathrm{LS}}(\omega) = \sum_{j=1}^{n} K_T\left(\frac{\omega-\omega_j}{h}, \omega\right) Y_j' + O_P\left(\frac{\log^2 T}{\sqrt{T}}\right), \tag{7.48}$$

where $Y_j' = m(\omega_j) + \varepsilon_j'$ with $\varepsilon_j' = \varepsilon_j - C_0$. By applying (6.30) to the first term in (7.48), we obtain the result.

We now establish (7.48). Using Lemma 7.2, the remainder term in (7.48) is bounded by

$$\sum_{j=1}^{n}\left|K_T\left(\frac{\omega-\omega_j}{h}, \omega\right) r_j\right| = O_P\left(\frac{\log T}{\sqrt{T}}\right) B_{T,1} + O_P(\log T) B_{T,2}, \tag{7.49}$$

where

$$\begin{aligned} B_{T,1} &= \sum_{j=1}^{n}\left|K_T\left(\frac{\omega-\omega_j}{h}, \omega\right)\right| I(V_j > c_T) V_j^{-1} \\ B_{T,2} &= \sum_{j=1}^{n}\left|K_T\left(\frac{\omega-\omega_j}{h}, \omega\right)\right| I(V_j \le c_T). \end{aligned}$$

Note that when $c_T \to 0$,

$$E\{I(V_j > c_T) V_j^{-1}\} \le \int_{c_T}^{1} t^{-1} dt + \int_{1}^{\infty} \exp(-t) dt = O(\log c_T^{-1})$$

and

$$EI(V_j \le c_T) = 1 - \exp(-c_T) = O(c_T).$$

Using the last two expressions and (7.47), we find

$$E|B_{T,1}| = O(\log c_T^{-1}) \quad \text{and} \quad E|B_{T,2}| = O(c_T).$$

Substituting these into (7.49), we conclude that (7.49) is of order

$$O_P\left(\frac{\log c_T^{-1} \log T}{\sqrt{T}}\right) + O_P\left(c_T \log T\right) = O_P\left(\frac{\log^2 T}{\sqrt{T}}\right)$$

by taking $c_T = T^{-1}$. ∎

7.5.5 Proof of Theorem 7.3

We need the following *quadratic approximation lemma*, due to Fan, Heckman, and Wand (1995), to prove Theorem 7.3. The lemma takes advantage of the fact that the target function $\mathcal{L}(\alpha, \beta)$ in (7.21) is concave. The pointwise convergence of $\mathcal{L}(\alpha, \beta)$ implies automatically the uniform convergence over a compact set. The essence of the following lemma is that it requires only pointwise convergence for the concave target functions. This is much easier to establish than the uniform convergence.

Lemma 7.3 *(Quadratic approximation lemma) Let* $\{\lambda_n(\theta)\colon \theta \in \Theta\}$ *be a sequence of random concave functions defined on a convex open subset* Θ *of* $\mathbb{R}^d$. *Let* $\mathbf{F}$ *and* $\mathbf{G}$ *be nonrandom matrices, with* $\mathbf{F}$ *positive-definite, and let* $\mathbf{U}_n$ *be a stochastically bounded sequence of random vectors. Lastly, let* α_n *be a sequence of constants tending to zero. Write*

$$\lambda_n(\theta) = \mathbf{U}_n^T\theta - \frac{1}{2}\theta^T(\mathbf{F} + \alpha_n\mathbf{G})\theta + f_n(\theta).$$

If, for each $\theta \in \Theta$, $f_n(\theta) = o_P(1)$, *then*

$$\widehat{\theta}_n = \mathbf{F}^{-1}\mathbf{U}_n + o_P(1),$$

where $\widehat{\theta}_n$ *(assumed to exist) maximizes* $\lambda_n(\cdot)$. *If, in addition,* $f'(\theta) = o_P(\alpha_n)$ *and* $f_n''(\theta) = o_P(\alpha_n)$ *uniformly in* θ *in a neighborhood of* $\widehat{\theta}_n$, *then*

$$\widehat{\theta}_n = \mathbf{F}^{-1}\mathbf{U}_n - \alpha_n\mathbf{F}^{-1}\mathbf{G}\mathbf{F}^{-1}\mathbf{U}_n + o_P(\alpha_n).$$

Proof of Theorem 7.3. The idea of the proof is to reduce the problem for dependent data to that for i.i.d. exponentially distributed random variables. The latter can be proved using the first part of the quadratic approximation lemma.

Let $\widehat{\boldsymbol{\beta}} = a_T^{-1}[\widehat{\alpha} - m(\omega), h\{\widehat{\beta} - m'(\omega)\}]^T$, where $a_T = (nh)^{-1/2}$. Define

$$\begin{aligned} L_k(Y_k, \boldsymbol{\beta}) &= -\exp\{Y_k - \bar{m}(\omega, \omega_k) - a_T\boldsymbol{\beta}^T\boldsymbol{\Omega}_k\} \\ &\quad + Y_k - \bar{m}(\omega, \omega_k) - a_T\boldsymbol{\beta}^T\boldsymbol{\Omega}_k, \end{aligned}$$

where $\bar{m}(\omega, \omega_k) = m(\omega) + m'(\omega)(\omega_k - \omega)$ and $\boldsymbol{\Omega}_k = \{1, (\omega_k - \omega)/h\}^T$. Then, it can easily be seen via a linear transform that $\widehat{\boldsymbol{\beta}}$ maximizes

$$\sum_{k=1}^{n} L_k(Y_k, \boldsymbol{\beta})K_h(\omega_k - \omega),$$

or equivalently $\widehat{\boldsymbol{\beta}}$ maximizes

$$\ell_T(\boldsymbol{\beta}) = h\sum_{k=1}^{n}\{L_k(Y_k, \boldsymbol{\beta}) - L_k(Y_k, 0)\}K_h(\omega_k - \omega).$$

Let $Y_k' = m(\omega_k) + z_k$, the main term of (7.2). Then, we can write

$$\ell_T(\boldsymbol{\beta}) = \ell_{1,T}(\boldsymbol{\beta}) + U_T$$

where $\ell_{1,T}(\boldsymbol{\beta})$ is defined in the same way as $\ell_T(\boldsymbol{\beta})$ with Y_k replaced by Y_k', and

$$\begin{aligned} U_T &= -h\sum_{k=1}^{n} R_T(\omega_k)\Big[\exp\{-\bar{m}(\omega, \omega_k) - a_T\boldsymbol{\beta}^T\boldsymbol{\Omega}_k\} \\ &\quad - \exp\{-\bar{m}(\omega, \omega_k)\}\Big]K_h(\omega_k - \omega). \end{aligned}$$

By using Taylor's expansion and Lemma 7.2, for each fixed $\boldsymbol{\beta}$,

$$U_T = O_P(h \cdot a_T \cdot T \cdot \log T/\sqrt{T}) = o_P(1).$$

Thus, we have

$$\ell_T(\boldsymbol{\beta}) = \ell_{1,T}(\boldsymbol{\beta}) + o_P(1). \tag{7.50}$$

We now deal with the term $\ell_{1,T}(\boldsymbol{\beta})$, which is the logarithm of the likelihood based on $\exp(Y_k')$, an independent sample from exponential distributions. By Taylor's expansion around the point 0,

$$\begin{aligned} L_k(Y_k', \boldsymbol{\beta}) - L_k(Y_k', 0) &= a_T[\exp\{Y_k' - \bar{m}(\omega, \omega_k)\} - 1]\boldsymbol{\beta}^T \boldsymbol{\Omega}_k \\ &\quad - \frac{a_T^2}{2} \exp\{Y_k' - \bar{m}(\omega, \omega_k)\}(\boldsymbol{\beta}^T \boldsymbol{\Omega}_k)^2\{1 + o(1)\}. \end{aligned}$$

Thus

$$\ell_{1,T}(\boldsymbol{\beta}) = \mathbf{W}_T \boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}^T \mathbf{A}_T \boldsymbol{\beta}\{1 + o(1)\}, \tag{7.51}$$

where

$$\mathbf{W}_T = a_T h \sum_{k=1}^{n} [\exp\{Y_k' - \bar{m}(\omega, \omega_k)\} - 1]\boldsymbol{\Omega}_k K_h(\omega_k - \omega)$$

and

$$\mathbf{A}_T = \frac{1}{n} \sum_{k=1}^{n} \exp\{Y_k' - \bar{m}(\omega, \omega_k)\}\boldsymbol{\Omega}_k \boldsymbol{\Omega}_k^T K_h(\omega_k - \omega).$$

Note that each element in $\mathbf{W}_T$ and $\mathbf{A}_T$ is a sum of independent random variables of kernel form. Their asymptotic behaviors are relatively easy to characterize. Basically, we will show that $\mathbf{W}_T$ is asymptotically normal and $\mathbf{A}_T$ converges in probability to a matrix. We now outline the key ideas of the proof. Since K has a bounded support, all effective ω_k's are in a local neighborhood of ω. By Taylor's expansion,

$$\bar{m}(\omega, \omega_k) \approx \frac{1}{2} m''(\omega)(\omega_k - \omega)^2.$$

Note that $\exp(Y_k')$ has an exponential distribution with the mean $\exp\{m(\omega_k)\}$. Using these terms, one can easily show that

$$E[\exp\{Y_k' - \bar{m}(\omega, \omega_k)\} - 1] = \frac{1}{2} m''(\omega)(\omega_k - \omega)^2 + o(1), \tag{7.52}$$

$$\mathrm{Var}[\exp\{Y_k' - \bar{m}(\omega, \omega_k)\} - 1] = 1 + o(1). \tag{7.53}$$

Approximating the discrete sum below by its integration, it follows from (7.52) that

$$\begin{aligned} E\mathbf{A}_T &= \frac{1}{n} \sum_{k=1}^{n} \boldsymbol{\Omega}_k \boldsymbol{\Omega}_k K_h(\omega_k - \omega) \\ &= \mathbf{A} + o(1), \end{aligned}$$

where $\mathbf{A} = -\pi^{-1}\text{diag}\{1, \mu_2(K)\}$. Similarly, by using (7.53), one can show that each element of $\mathbf{A}_T$ has variance of order $O(a_T)$. Thus

$$\mathbf{A}_T = \mathbf{A} + o_P(1). \tag{7.54}$$

Combining (7.50), (7.51) and (7.54) leads to

$$\ell_T(\boldsymbol{\beta}) = \mathbf{W}_T^T\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta} + o_P(1). \tag{7.55}$$

To apply the quadratic approximation lemma, we need to establish the asymptotic normality of $\mathbf{W}_T$. First, by (7.52),

$$\begin{aligned} E\mathbf{W}_T &= a_T h \sum_{k=1}^{n} \frac{1}{2} m''(\omega)(\omega_k - \omega)^2 \boldsymbol{\Omega}_k K_h(\omega_k - \omega)\{1 + o(1)\} \\ &= a_T^{-1}\frac{h^2}{2} m''(\omega)\pi^{-1} \begin{pmatrix} \mu_2(K) \\ 0 \end{pmatrix} + o(a_T^{-1}h^2). \end{aligned}$$

Similarly, it follows from (7.53) that we have

$$\begin{aligned} \text{Var}(\mathbf{W}_T) &= a_T^2 h^2 \sum_{k=1}^{n} \boldsymbol{\Omega}_k \boldsymbol{\Omega}_k^T K_h^2(\omega_k - \omega) \\ &= \mathbf{B} + o(1), \end{aligned}$$

where $\mathbf{B} = \pi^{-1}\text{diag}\{\nu_0(K), \int t^2 K^2(t)dt\}$. Since $\mathbf{W}_T$ is the sum of independent random variables, one can easily verify that it satisfies the Lindeberg condition. Hence,

$$\mathbf{W}_T - a_T^{-1}\left\{\frac{m''(x)}{2}h^2\pi^{-1}\{\mu_2(K), 0\}^T + o_P(h^2)\right\} \xrightarrow{D} N(0, \mathbf{B}). \tag{7.56}$$

The quadratic approximation lemma and (7.55) lead to

$$\widehat{\boldsymbol{\beta}} = -\mathbf{A}^{-1}\mathbf{W}_T + o_P(1).$$

By (7.56), $\widehat{\boldsymbol{\beta}}$ is asymptotically normal. Hence, its first element is asymptotically normal. This proves the theorem. ∎

7.6 Bibliographical Notes

Spectral density estimation is closely related to state domain smoothing. Some related references can also be found in §6.7.

Spectral density estimation

The suggestion that an improved spectral estimate might be obtained by smoothing the periodogram was made by Daniels (1946); see also Bartlett (1948, 1950). Brillinger and Rosenblatt (1967) considered estimation of higher-order spectra. The asymptotic normality of higher-order cumulant spectral density estimates was established by Lii and Rosenblatt (1990). Asymptotic properties of spectral estimates were studied by Brillinger (1969). Lii (1978) established the asymptotic normality for the L_2-norm between estimated density and true density. The problem of estimating the spectral density for stationary symmetric α-stable processes was considered by Masry and Cambanis (1984). Nonparametric high-resolution estimation of spectral density was studied by Dahlhaus (1990b). Lii and Masry (1995) studied selection of sampling schemes for spectral density estimation. Lii and Rosenblatt (1998) showed that it is generally impossible to have consistent estimates of spectral mass for a harmonizable process and hence line spectrum was estimated.

Recent advances in smoothing techniques enrich the techniques of spectral density estimation. Wahba (1980) used smoothing splines to smooth a log-periodogram. Extensive efforts have been made in selecting appropriate smoothing parameters for spectral density estimators; see, for example, Swanepoel and van Wyk (1986), Beltrão and Bloomfield (1987), Hurvich and Beltrão (1990), and Franke and Härdle (1992). For multivariate spectral density estimation, Robinson (1991a) considered nonparametric and semiparametric estimation of spectral density with data-dependent bandwidth. A plug-in method for selecting bandwidths for spectral density estimation was proposed in Park, Cho, and Kang (1994). Based on the penalized Whittle likelihood, Pawitan and O'Sullivan (1994) used smoothing splines to estimate the spectral density. Kooperberg, Stone, and Truong (1995a, b) developed log-spline spectral density estimates. Kato and Masry (1999) applied wavelet techniques to spectral density estimation.

Test of independence

Test for independence is usually based on the autocorrelation function and periodogram of a time series. Different tests detect different aspects of deviation from a null hypothesis and hence are powerful for certain given alternatives. Brillinger (1974) derived the asymptotic distribution for testing periodicities. Skaug and Tjøstheim (1993) generalized the idea of Blum, Kiefer and Rosenblatt (1961) for testing serial independence. Robinson (1991b) proposed tests for strong serial correlation and conditional heteroscedasticity. Some recent work on the subject can be found in Deo (2000) and Deo and Chen (2000a,b). A survey and development on various measures of dependence can be found in Tjøstheim (1996) and the references therein.

As mentioned in §7.4 for a stationary time series, testing for white noise in the time domain is equivalent to testing whether spectral density is constant based on nearly independent periodograms. This is basically a

parametric null hypothesis versus nonparametric smooth alternative hypothesis. There is much literature studying this kind of problem. An early paper is Bickel and Rosenblatt (1973), where the asymptotic null distributions were derived. A few new tests were proposed in Bickel and Ritov (1992). Azzalini and Bowman (1993) introduced the F-type test statistic for testing parametric models. Härdle and Mammen (1993) studied nonparametric tests based on an L_2-distance. Various recent testing procedures are motivated by Neyman (1937). They basically focus on selecting the smoothing parameters of the Neyman test and studying the properties of the resulting procedures; see, for example, Eubank and Hart (1992), Eubank and LaRiccia (1992), Inglot, Kallenberg and Ledwina (1994), Kallenberg and Ledwina (1994), and Kuchibhatla and Hart (1996), among others. Fan (1996) proposed simple and powerful methods for constructing tests based on Neyman's truncation and wavelet thresholding. It was shown in Spokoiny (1996) that wavelet thresholding tests are nearly adaptively minimax and in Fan, Zhang, and Zhang (2001) that Fan's version of the adaptive Neyman test is asymptotically adaptively minimax. The asymptotic optimality of data-driven Neyman's tests was also studied by Inglot and Ledwina (1996).

There are various extensions of nonparametric tests to multivariate settings. The largest challenge is how to handle the so-called "curse of dimensionality" in multivariate nonparametric regression. This refers to the fact that a local neighborhood in multidimensional space contains very few data points. Aerts, Claeskens, and Hart (2000) constructed tests based on orthogonal series for a bivariate nonparametric regression problem. Fan and Huang (2001) proposed various testing techniques based on the adaptive Neyman test for various alternative models in a multiple regression setting. A generally applicable method, generalized likelihood ratio tests, was proposed and studied by Fan, Zhang, and Zhang (2001). Horowitz and Spokoiny (2001) studied the problem of adaptive minimax rates for multivariate nonparametric testing problems.

8
Nonparametric Models

8.1 Introduction

Parametric time series models provide powerful tools for analyzing time series data when the models are correctly specified. However, any parametric models are at best only an approximation to the true stochastic dynamics that generates a given data set. The issue of modeling biases always arises in parametric modeling. One conventional technique is to expand the parametric models from a smaller family to a larger family. This eases the concerns on modeling biases but is not necessarily the most effective way to deal with them. As mentioned in §1.3.3, a good fitting for a simple MA series by an AR model may require a high order. Similarly, a simple nonlinear series might require a high order of ARMA model to reasonably approximate it. Thus, the choice for the form of a parametric model is very critical.

Many data in applications exhibit nonlinear features such as nonnormality, asymmetric cycles, bimodality, nonlinearity between lagged variables, and heteroscedasticity. They require nonlinear models to describe the law that generates the data. However, beyond the linear time series models, there are infinitely many nonlinear forms that can be explored. This would be an undue task for any time series analysts to try one model after another. A natural alternative is to use nonparametric methods. The most flexible nonparametric model is the saturated (full) nonparametric model, which does not impose any particular form on autoregression functions. This saturated nonparametric model is certainly flexible in reducing modeling biases.

Yet, in the multivariate setting with more than two lagged variables, its underlying autoregressive function cannot be estimated with reasonable accuracy due to the so-called "curse of dimensionality" of Bellman (1961). The *curse of dimensionality* problem has been clearly illustrated in many books, including Silverman (1986), Härdle (1990), Hastie and Tibshirani (1990), Scott (1992), and Fan and Gijbels (1996).

There are many possibilities between parametric models and saturated nonparametric models. Certain forms are typically imposed on the autoregressive functions. The resulting models are usually generalizations of certain parametric families; see, for example, the functional-coefficient autoregressive (FAR) model (1.11) and the additive autoregressive (AAR) model (1.12). They are better able to reduce possible modeling biases than their parametric counterparts. On the other hand, they are much smaller than the saturated nonparametric model. As a result, the unknown parameters and functions can be estimated with reasonable accuracy.

In this chapter, we will introduce a few nonsaturated nonparametric models. These include functional-coefficient autoregressive (FAR) models, adaptive FAR models, additive autoregressive models, and models for conditional variance. Different models impose different nonparametric forms on the autoregressive regression function and explore different aspects of the data. They together form powerful tools for time-series data analysis.

8.2 Multivariate Local Polynomial Regression

Local polynomial fitting can readily be extended to the multivariate setting. Due to the curse of dimensionality, direct use of the multivariate nonparametric regression is not viable. However, its functionals can be useful for other related problems. For completeness, we briefly outline the idea of the extension of the local polynomial fitting to a multivariate setting.

8.2.1 Multivariate Kernel Functions

To localize data in the p-dimension, we need a *multivariate kernel*. Generally speaking, a multivariate kernel function refers to a p-variate function satisfying

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} K(\mathbf{x}) d\mathbf{x} = 1.$$

Moment conditions similar to (5.13) can be imposed to ameliorate biases. For example, a second-order kernel requires

$$\int x_i K(\mathbf{x}) d\mathbf{x} = 0, \quad i = 1, \cdots, p.$$

and the finite second-moment condition. Here and hereafter, we use "$\int$" to indicate multivariate integration over the p-dimensional Euclidean space.

There are two common approaches for constructing multivariate kernels. For a univariate kernel κ, the *product kernel* is given by

$$K(\mathbf{x}) = \prod_{i=1}^{p} \kappa(x_i),$$

and the spherically symmetric kernel is defined as

$$K(\mathbf{x}) = c_{\kappa,p} K(\|\mathbf{x}\|),$$

where $c_{\kappa,p} = \{\int K(\|\mathbf{x}\|)d\mathbf{x}\}^{-1}$ is a normalization constant and $\|\mathbf{x}\| = (x_1^2 + \cdots + x_p^2)^{1/2}$. Popular choices of K include the standard p-variate normal density

$$K(x) = (2\pi)^{-p/2} \exp(-\|\mathbf{x}\|^2/2)$$

and the spherical Epanechnikov kernel

$$K(\mathbf{x}) = \{p(p+2)\Gamma(p/2)/(4\pi^{p/2})\}(1 - \|\mathbf{x}\|^2)_+$$

The latter is the optimal kernel, according to Fan et al. (1996).

The localization in multivariate nonparametric regression is frequently carried out by the kernel weighting. Let $\mathbf{H}$ be a symmetric positive-definite matrix called a *bandwidth matrix.* The localization scheme at a point x assigns the weight

$$K_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}), \quad \text{with} \quad K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{x}),$$

where $|\mathbf{H}|$ is the determinant of the matrix $\mathbf{H}$. The bandwidth matrix is introduced to accommodate the dependent structure in the independent variables. For practical implementations, one frequently takes the bandwidth matrix $\mathbf{H}$ to be a diagonal matrix. This will accommodate different scales in different independent variables. A further simplification is to take the bandwidth matrix $\mathbf{H} = h\mathbf{I}_p$ with $\mathbf{I}_p$ being the identity matrix of order p, assuming that the independent variables have the same scale (e.g., through some normalizations).

For the spherical Epanechnikov kernel with the bandwidth matrix $\mathbf{H} = h\mathbf{I}_p$, the nonvanishing weights are only those $\mathbf{X}_i$'s that fall in the ball centered at $\mathbf{x}$ with radius h. Such a ball has a size of order $O(h^p)$, which gets smaller and smaller as p increases. For such a small ball, there are not many local data points there. This is the essence of the curse of dimensionality. In order to get a sufficient amount of local data points, the neighborhood has to increase, which introduces an unacceptable level of bias.

8.2.2 Multivariate Local Linear Regression

The best predictor for X_t based on its lag variables $X_{t-1}, \cdots, X_{t-p}$ is

$$G(X_{t-1}, \cdots, X_{t-p}) = E(X_t | X_{t-1}, \cdots, X_{t-p}).$$

Such an *autoregressive function* minimizes the prediction error

$$E\{X_t - g(X_{t-1}, \cdots, X_{t-p})\}^2$$

among the class of measurable functions g. In fact, one can easily see that

$$\begin{aligned} &E\{X_t - g(X_{t-1}, \cdots, X_{t-p})\}^2 = E\{X_t - G(X_{t-1}, \cdots, X_{t-p})\}^2 \\ &+E\{G(X_{t-1}, \cdots, X_{t-p}) - g(X_{t-1}, \cdots, X_{t-p})\}^2. \end{aligned}$$

To estimate the autoregressive function, let $\mathbf{X}_{t-1} = (X_{t-1}, \cdots, X_{t-p})^T$. Then, the *multivariate kernel estimator* is basically the locally weighted average:

$$\widehat{G}(\mathbf{x}) = \frac{\sum_{t=p+1}^{T} X_t K_{\mathbf{H}}(\mathbf{X}_{t-1} - \mathbf{x})}{\sum_{t=p+1}^{T} K_{\mathbf{H}}(\mathbf{X}_{t-1} - \mathbf{x})}.$$

The kernel estimator is based on the local constant approximation. It can be improved by using the local linear approximation

$$G(\mathbf{X}) \approx G(\mathbf{x}) + G'(\mathbf{x})^T(\mathbf{X} - \mathbf{x}),$$

for $\mathbf{X}$ in a local neighborhood of $\mathbf{x}$. This leads to the following least-squares problem:

$$\sum_{t=p+1}^{T} \left\{ X_t - a - \mathbf{b}^T(\mathbf{X}_{t-1} - \mathbf{x}) \right\}^2 K_{\mathbf{H}}(\mathbf{X}_{t-1} - \mathbf{x}).$$

Let $\widehat{a}(\mathbf{x})$ and $\widehat{\mathbf{b}}(\mathbf{x})$ be the minimizers. Then, the local linear estimate of G is simply $\widehat{G}(\mathbf{x}) = \widehat{a}(\mathbf{x})$ and the local linear estimate of the gradient vector $G'(\mathbf{x})$ is $\widehat{G'}(\mathbf{x}) = \widehat{\mathbf{b}}(\mathbf{x})$.

The asymptotic biases and variances can be established along the same lines of argument as those in §6.3 under some regularity conditions. In particular, the asymptotic bias of $\widehat{G}(\mathbf{x})$ is

$$2^{-1}\mathrm{tr}\left\{ G''(\mathbf{x})\mathbf{H}\mathbf{H}^T \int K(\mathbf{u})\mathbf{u}\mathbf{u}^T d\mathbf{u} \right\},$$

with $G''(\cdot)$ the Hessian matrix of the function G, and the asymptotic variance is given by

$$\frac{\sigma^2(\mathbf{x})}{T|H|f(\mathbf{x})} \int K^2(\mathbf{u}) d\mathbf{u},$$

TABLE 8.1. Sample sizes required for p-dimensional nonparametric regression to have a performance comparable with that of one-dimensional nonparametric regression using size 100.

dimension	2	3	4	5	6	7	8	9
sample size	250	630	1580	3,980	10,000	25,000	63,000	158,000

where $\sigma^2(\mathbf{x}) = \text{Var}(X_t|\mathbf{X}_{t-1} = \mathbf{x})$, and $f(\mathbf{x})$ is the joint density of $\mathbf{X}_{t-1}$.

To see the rate of convergence, let us take $\mathbf{H} = h\mathbf{I}_p$. Then, the bias is of order $O(h^2)$ and the variance is of order $O(1/Th^p)$. This leads to the optimal rate of convergence $O(T^{-2/(4+p)})$ by trading off the rates between the bias and variance.

The curse of dimensionality can be quantitatively understood as follows. To have a performance comparable with one-dimensional nonparametric regression with T_1 data points, for p-dimensional nonparametric regression, we need

$$T^{-2/(4+p)} = O(T_1^{-2/5}) \text{ or } T = T_1^{(p+4)/5}.$$

Table 8.1 shows the result with $T_1 = 100$. The increase of required sample sizes is exponentially fast.

8.2.3 Multivariate Local Quadratic Regression

Due to the sparsity of local data in multi-dimensional space, a higher-order polynomial is rarely used. Further, the notation becomes more cumbersome. We use the local quadratic regression to indicate the flavor of the multivariate local polynomial fitting. The technique can be useful for estimating the gradient vector $G'(\cdot)$ where the local linear fit does not give a good enough estimate.

The local quadratic approximation is as follows. By Taylor's expansion to the second order, we have

$$G(\mathbf{X}) \approx G(\mathbf{x}) + G'(\mathbf{x})(\mathbf{X} - x) + \frac{1}{2}(\mathbf{X} - \mathbf{x})^T G''(\mathbf{x})(\mathbf{X} - \mathbf{x}).$$

This leads to minimizing

$$\sum_{t=p+1}^{T} \left\{ X_t - a - \mathbf{b}^T(\mathbf{X}_{t-1} - \mathbf{x}) - \frac{1}{2}(\mathbf{X} - \mathbf{x})^T \mathbf{C}(\mathbf{X} - \mathbf{x}) \right\}^2 K_{\mathbf{H}}(\mathbf{X}_{t-1} - \mathbf{x}),$$

with respect to a, $\mathbf{b}$ and $\mathbf{C}$, where $\mathbf{C}$ is a symmetric matrix. They are an estimate of $G(\mathbf{x})$, $G'(\mathbf{x})$, and $G''(\mathbf{x})$, respectively.

8.3 Functional-Coefficient Autoregressive Model

8.3.1 The Model

The functional coefficient model, introduced by Chen and Tsay (1993), admits the form

$$X_t = a_1(X_{t-d})X_1 + \cdots + a_p(X_{t-d})X_{t-p} + \sigma(X_{t-d})\varepsilon_t, \tag{8.1}$$

where $\{\varepsilon_t\}$ is a sequence of independent random variables with zero mean and unity variance, and ε_t is independent of $X_{t-1}, X_{t-2}, \cdots$. The coefficient functions $a_1(\cdot), \cdots, a_p(\cdot)$ are unknown. The model is a special case of the state-dependent model of Priestley (1981). For simplicity, we will call the variable X_{t-d} the *model-dependent variable* and denote the model by FAR(p, d).

The state-dependent model is a natural extension of the TAR model discussed in §5.2. It allows the coefficient functions to change gradually, rather than abruptly as in the TAR model, as the value of X_{t-d} varies continuously. This can be appealing in many applications such as in understanding the population dynamics in ecological studies. As the population density X_{t-d} changes continuously, it is reasonable to expect that its effects on the current population size X_t will be continuous as well.

The FAR model also includes the generalized exponential autoregressive (EXPAR) model

$$X_t = \sum_{i=1}^{p} \left\{\alpha_i + (\beta_i + \gamma_i X_{t-d}) \exp(-\theta_i X_{t-d}^2)\right\} X_{t-i} + \varepsilon_t, \tag{8.2}$$

where $\theta_i \geq 0$ for $i = 1, \ldots, p$. The model was introduced and studied by Haggan and Ozaki (1981) and Ozaki (1982). The FAR model allows other forms for the coefficient functions.

8.3.2 Relation to Stochastic Regression

All parametric and nonparametric autoregressive models can be regarded as stochastic regression models, so the techniques developed in regression models can be applied to time series. The major difference here is that the data are dependent in the context of time series. This usually has limited impact on estimation procedures but might affect probability statements such as confidence intervals and p-values. However, as illustrated in §5.3, the adverse effects on probability statements for state-domain smoothing are not severe due to the property of "whitening by windowing."

Introduce the independent variable Y as the current observation X_t, the ith independent variable "X_i" as the lag i variable X_{t-i}, and U as the lag

d variable X_{t-d}. The $(t-p)$th observation of these induced variables is

$$Y_t = X_t, X_{t1} = X_{t-1}, \cdots, X_{tp} = X_{t-p}, U_t = X_{t-d}, \quad t = p+1, \cdots, T. \tag{8.3}$$

Here, for simplicity, we assume that $d \leq p$. With the induced variables above, the FAR model (8.1) can be written as

$$Y = a_1(U)X_1 + \cdots + a_p(U)X_{t-p} + \sigma(U)\varepsilon_t \tag{8.4}$$

based on the data $\{(Y_t, X_{t1}, \cdots, X_{tp}, U_t), t = p+1, \cdots, T\}$. To facilitate the notation, with slight abuse of notation, we relabel the data as

$$\{(Y_i, X_{i1}, \cdots, X_{ip}, U_i), i = 1, \cdots, n\},$$

where $n = T - p$. This is indeed a stochastic regression model and has been popularly studied in the setting of the independent observations; see, for example, Hastie and Tibshirani (1993), Fan and Zhang (1999), and Cai, Fan, and Li (2000), among others.

*8.3.3 Ergodicity**

One of the fundamental questions is whether the model (8.1) yields a stationary and ergodic solution. According to Theorem 2.2, we need to establish the ergodicity of the series. The following theorem was established by Chen and Tsay (1993).

Theorem 8.1 *Assume that the functions $a_j(\cdot)$ are bounded by c_j and the density function of ε_t is positive everywhere on the real line. If all roots of the characteristic function*

$$\lambda^p - c_1\lambda^{p-1} - \cdots - c_p = 0$$

are inside the unit circle, then the FAR(p, d) process is geometrically ergodic.

Before we outline the key idea of the proof, let us illustrate the theorem above by a few examples.

Example 8.1 (*AR(p) model*) The AR(p) model corresponds to an FAR(p, d) model with

$$a_1(\cdot) \equiv a_1, \cdots, a_p(\cdot) \equiv a_p \text{ and } \sigma(\cdot) = \sigma.$$

The condition in Theorem 8.1 is the same as that in Theorem 2.1 for the stationarity of an AR(p) process. ∎

Example 8.2 (*EXPAR model*) Consider the EXPAR model (8.2) with $\gamma_i = 0$. The state-dependent coefficient function is given by

$$a_i(u) = \alpha_i + \beta_i \exp(-\theta_i u^2).$$

Since $\theta_i \geq 0$, the coefficient function $a_i(\cdot)$ is bounded by

$$|a_i(u)| \leq |\alpha_i| + |\beta_i|.$$

Let $c_i = |\alpha_i| + |\beta_i|$. By Theorem 8.1, the process is geometrically ergodic as long as the condition in Theorem 8.1 is fulfilled. ∎

Example 8.3 (*TAR(p) model*) Consider the TAR model (1.8). It corresponds to the FAR(p, d) model with

$$a_j(u) = b_j^{(i)} \quad \text{for } u \in \Omega_i, \quad i = 1, \cdots, k, \quad j = 1, \cdots, p.$$

The coefficient function $a_j(\cdot)$ is bounded by $c_j = \max\{|b_j^{(1)}|, \cdots, |b_j^{(k)}|\}$, and Theorem 8.1 is applicable. ∎

We now outline the key idea for the proof of Theorem 8.1. The approach is useful for other similar problems. Following the idea in §2.1.4, we express the series as a Markov chain in the p-dimensional Euclidean space. Let

$$\mathbf{X}_t = (X_t, \cdots, X_{t-p+1})^T, \quad \boldsymbol{\varepsilon}_t = (\varepsilon_t, \cdots, \varepsilon_{t-p+1})^T,$$

and set

$$\mathbf{A}(\mathbf{X}) = \begin{pmatrix} a_1(\mathbf{X}) & a_2(\mathbf{X}) & \cdots & a_{p-1}(\mathbf{X}) & a_p(\mathbf{X}) \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

Then, we can rewrite the FAR model as

$$\mathbf{X}_t = \mathbf{A}(\mathbf{X}_{t-1})\mathbf{X}_{t-1} + \boldsymbol{\varepsilon}_t. \tag{8.5}$$

This is clearly a Markov chain in the p-dimensional Euclidean space.

We need the following concept of *ϕ-irreducibility* and *aperiodicity* of a Markov chain in a topological space. Let $\mathcal{X}$ be a topological space equipped with a nontrivial measure ϕ.

Definition 8.1 *A Markov chain $\{\mathbf{X}_t\}$ is said to be ϕ-irreducible if for any measurable set A with $\phi(A) > 0$, there exists an $n \geq 0$ such that*

$$P\{\mathbf{X}_n \in A | \mathbf{X}_0 = \mathbf{x}\} > 0 \text{ for all } \mathbf{x} \in \mathcal{X}.$$

The ϕ-irreducibility basically says, that starting from any initial value $\mathbf{X}_0 = \mathbf{x}$ with positive probability, the chain will visit the set A in finite steps. The concept of the aperiodicity is that a Markov chain cannot be divided into cyclic subchains. There are a few equivalent definitions, and we take the one that is simplest for description.

Definition 8.2 *A Markov chain* $\{\mathbf{X}_t\}$ *is aperiodic if there exists a measurable set* A *with* $\phi(A) > 0$ *such that, for any subset* B *of* A *with* $\phi(B) > 0$, *there exists a positive integer* n *such that*

$$P\{\mathbf{X}_n \in B|\mathbf{X}_0 = x\} > 0 \quad \textit{and} \quad P\{\mathbf{X}_{n+1} \in B|\mathbf{X}_0 = x\} > 0.$$

The following two lemmas are useful for establishing the ergodicity. The first one is due to Tweedie (1975), and the second one is due to Tjøstheim (1990).

Lemma 8.1 *Let* $\{\mathbf{X}_t\}$ *be a* ϕ*-irreducible Markov chain on a normed topological space. If the transition probability* $P(\mathbf{x}, \cdot)$ *is strongly continuous-namely, the transition probability* $P(\mathbf{x}, A)$ *from* $\mathbf{x}$ *to any measurable set* A *is continuous in* $\mathbf{x}$*-then a sufficient condition for the geometric ergodicity is that there exists a compact set* $\mathbf{K}$ *and a positive constant* $\rho < 1$ *such that*

$$E\Big(\|\mathbf{X}_{t+1}\|\Big|\mathbf{X}_t = \mathbf{x}\Big) < \begin{cases} \infty, & \textit{for } \mathbf{x} \in \mathbf{K} \\ \rho\|\mathbf{x}\|, & \textit{for } \mathbf{x} \notin \mathbf{K}. \end{cases}$$

Lemma 8.2 *Let* $\{X_t\}$ *be an aperiodic Markov chain, and let* m *be a positive integer. Then, the geometric ergodicity of the subsequence* $\{X_{mt}\}$ *entails the geometric ergodicity of the original series* $\{X_t\}$.

The key idea for proving Theorem 8.1 is to show that, for a subsequence $\{X_{mt}\}$, the conditions in Lemma 8.1 are fulfilled. Hence, it is geometric ergodicity. By Lemma 8.2, the whole series must be geometrically ergodic. The details of the proof are given in §8.8.1.

8.3.4 Estimation of Coefficient Functions

The unknown coefficient functions in (8.4) can be estimated by using a local linear regression technique. For any given u_0 and u in a neighborhood of u_0, it follows from a Taylor expansion that

$$a_j(u) \approx a_j(u_0) + a_j'(u_0)(u - u_0) \equiv a_j + b_j\,(u - u_0), \tag{8.6}$$

where a_j and b_j are the local intercept and slope corresponding to $a_j(u_0)$ and $a_j'(u_0)$. Using the data with U_i around u_0 and the local model (8.6), we run the following local linear regression. Minimize with respect to $\{a_j\}$ and $\{b_j\}$

$$\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{p}\{a_j + b_j\,(U_i - u_0)\}\,X_{ij}\right]^2 K_h(U_i - u_0), \tag{8.7}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel function, and h is a bandwidth. Let $\{(\widehat{a}_j, \widehat{b}_j)\}$ be the local least squares estimator. Then, the local linear regression estimator is simply

$$\widehat{a}_j(u_0) = \widehat{a}_j, \quad j = 1, \cdots, n. \tag{8.8}$$

The local linear regression estimator can be easily obtained. Let $\mathbf{e}_{j,2p}$ be the $2p \times 1$ unit vector with 1 at the jth position, $\widetilde{\mathbf{X}}$ denote an $n \times 2p$ matrix with $(\mathbf{X}_i^T, \mathbf{X}_i^T(U_i - u_0))$ as its ith row, and $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$. Set $\mathbf{W} = \text{diag}\{K_h(U_1 - u_0), \ldots, K_h(U_n - u_0)\}$. Then, the local regression problem (8.7) can be written as

$$(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = (a_1, \cdots, a_p, b_1, \cdots, b_p)^T$. The local least squares estimator is simply

$$\widehat{\boldsymbol{\beta}} = \left(\widetilde{\mathbf{X}}^T \mathbf{W} \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^T \mathbf{W} \mathbf{Y},$$

which entails

$$\widehat{a}_j(u_0) = \widehat{a}_j = \mathbf{e}_{j,2p}^T \widehat{\boldsymbol{\beta}}.$$

By simple algebra, it can be expressed in an equivalent kernel form as

$$\widehat{a}_j(u_0) = \sum_{k=1}^{n} K_{n,j}(U_k - u_0, \mathbf{X}_k) Y_k, \tag{8.9}$$

where

$$K_{n,j}(u, \mathbf{x}) = \mathbf{e}_{j,2p}^T \left(\widetilde{\mathbf{X}}^T \mathbf{W} \widetilde{\mathbf{X}}\right)^{-1} \begin{pmatrix} \mathbf{x} \\ u\mathbf{x} \end{pmatrix} K_h(u). \tag{8.10}$$

See §3.2.2 of Fan and Gijbels (1996) for similar derivations.

8.3.5 Selection of Bandwidth and Model-Dependent Variable

Various bandwidth selection techniques (see, e.g., §6.3.5) for nonparametric regression can be extended to the FAR model. Here we introduce a simple and quick method proposed in Cai, Fan, and Yao (2000). It can be regarded as a modified multifold cross-validation criterion that is attentive to the structure of stationary time series data. Let m and Q be two given positive integers such that $n > mQ$. The basic idea is first to use Q subseries of lengths $n - qm$ ($q = 1, \cdots, Q$) to estimate the unknown coefficient functions and then to compute the one-step forecasting errors of the next section of the time series of length m based on the estimated models. This idea is schematically illustrated in Figure 8.1.

Let $\{\widehat{a}_{j,q}(\cdot)\}$ be the estimated coefficients using the qth ($q = 1, \cdots, Q$) subseries $\{(U_i, \mathbf{X}_i, Y_i), 1 \le i \le n - qm\}$ with bandwidth equal to $h\{n/(n - qm)\}^{1/5}$. The bandwidth h is rescaled slightly to accommodate different sample sizes according to its optimal rate (i.e., $h \propto n^{-1/5}$). The average prediction error using the qth subseries is given by

$$\text{APE}_q(h) = \frac{1}{m} \sum_{i=n-qm+1}^{n-qm+m} \left\{ Y_i - \sum_{j=1}^{p} \widehat{a}_{j,q}(U_i) X_{i,j} \right\}^2 .$$

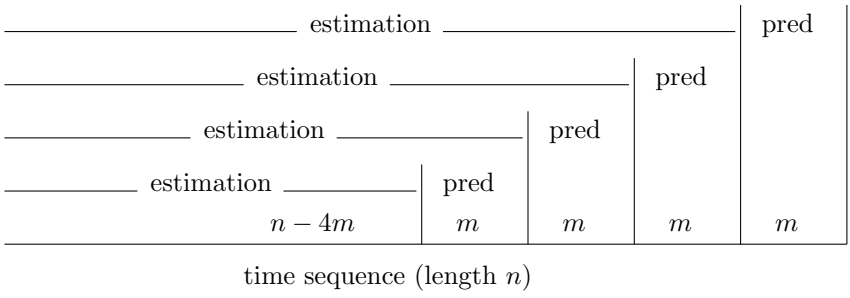


FIGURE 8.1. Illustration of data used for estimation and prediction. The data of length $n - 4m$ are used to construct estimated coefficients, and the prediction errors for the next m data are computed. Then, the data of length $n - 3m$ are used to construct estimated coefficients, and the prediction errors for the next m data are computed, and so on.

The overall average prediction error is given by

$$\text{APE}(h) = Q^{-1} \sum_{q=1}^{Q} \text{APE}_q(h). \tag{8.11}$$

The proposed data-driven bandwidth is the one that minimizes APE(h). In practical implementations, we may use $m = [0.1n]$ and $Q = 4$. The selected bandwidth does not depend critically on the choice of m and Q as long as mQ is reasonably large so that the evaluation of prediction errors is stable. The function APE(h) is minimized by comparing its value at a grid of points $h_j = a^j h_0 (j = 1, \cdots, J)$. For example, one may choose $a = 1.2$, $J = 15$ or 20, and $h_0 = 1.2^{-J}$(range of U). A weighted version of APE(h) can also be used if one wishes to weight down the prediction errors at an earlier time. The choice $m = [0.1n]$ rather than $m = 1$ is taken simply to facilitate computational expediency.

Choosing an appropriate model-dependent variable U is also very important. Knowledge of the physical background of the data may be very helpful, as we have witnessed in modeling lynx data. Without any prior information, it is pertinent to choose U in terms of some data-driven methods such as AIC, cross-validation, and other criteria. Let APE(h, d) be the average prediction error defined by (8.11) using the lagged variable $U = X_{t-d}$. Here, we stress the dependence of the prediction error on the lag variable X_{t-d}. A simple and practical approach is to minimize APE(h, d) simultaneously for h in a certain range and d over the set $\{1, 2, \cdots, p\}$. The order p can also be chosen to minimize the APE.

8.3.6 Prediction

Based on model (8.1), the one-step-ahead predictor is given by

$$\widehat{X}_{t+1} = \widehat{a}_1(X_{t-d})X_t + \cdots + \widehat{a}_p(X_{t-d})X_{t-p+1}. \tag{8.12}$$

This is a predictor whether model (8.1) holds or not. For two-step-ahead forecasting, there are two possible approaches. The iterative two-step-ahead predictor is to use (8.12) iteratively, leading to

$$\widehat{X}_{t+2} = \widehat{a}_1(X_{t+1-d})\widehat{X}_{t+1} + \widehat{a}_2(X_{t+1-d})X_t + \cdots + \widehat{a}_p(X_{t+1-d})X_{t-p+2}. \tag{8.13}$$

The direct two-step ahead predictor is based on fitting the model

$$X_{t+2} = b_1(X_{t-d})X_t + \cdots + b_p(X_{t-d})\, X_{t-p} + \varepsilon_t', \tag{8.14}$$

resulting in the estimated coefficient functions $\widehat{b}_1(\cdot), \cdots, \widehat{b}_p(\cdot)$ and the predictor

$$\widehat{X}_{t+2} = \widehat{b}_1(X_{t-d})X_t + \cdots + \widehat{b}_p(X_{t-d})\, X_{t-p}.$$

Note that model (8.1) does not imply (8.14). In this sense, the direct two-step ahead predictor explores the prediction power of the proposed modeling techniques when the model is misspecified. Since model (8.14) is usually not a correct model, the model-dependent variable X_{t-d} had better be chosen to minimize the estimated prediction error using the techniques in the previous section. For multistep-ahead forecasting, the two approaches above continue to apply.

8.3.7 Examples

We now illustrate the sampling properties of the proposed methods through two simulated and two real data examples. The performance of estimators $\{\widehat{a}_j(\cdot)\}$ can be assessed via the square-root of *average squared errors* (RASE):

$$\begin{aligned} \text{RASE}_j &= \left[n_{\text{grid}}^{-1} \sum_{k=1}^{n_{\text{grid}}} \{\widehat{a}_j(u_k) - a_j(u_k)\}^2 \right]^{1/2}, \\ \text{RASE}^2 &= \sum_{j=1}^{p} \text{RASE}_j^2, \end{aligned}$$

where $\{u_k,\ k = 1, \ldots, n_{\text{grid}}\}$ are regular grid points on an interval over which the functions $a_j(\cdot)$ are evaluated. We also compare the postsample forecasting performance of the new methods with existing methods such as the linear AR model, TAR model, and FAR model that are implemented in Chen and Tsay (1993).

Throughout this section, the Epanechnikov kernel $K(u) = 0.75\ \left(1 - u^2\right)_+$ is employed.

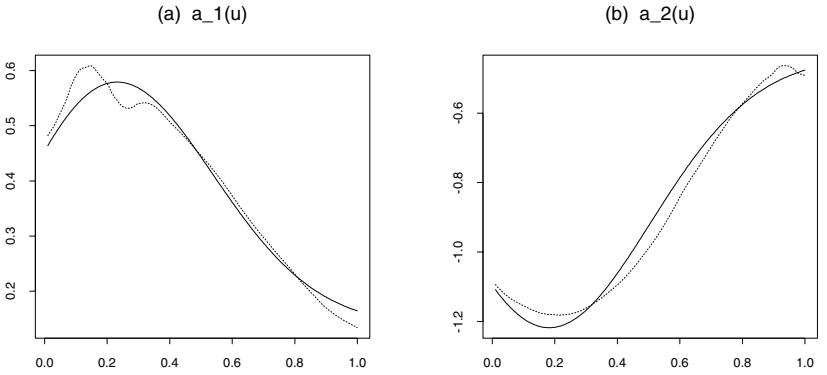


FIGURE 8.2. Simulation results for Example 8.4. The local linear estimators (dotted curves) for the coefficient functions $a_1(\cdot)$ (a) and $a_2(\cdot)$ (b) (solid curves). Adapted from Cai, Fan, and Yao (2000).

Example 8.4 (*Simulation from an EXPAR model*) We drew 400 time series of length 400 from the EXPAR model

$$X_t = a_1(X_{t-1})\, X_{t-1} + a_2(X_{t-1})\, X_{t-2} + \varepsilon_t, \tag{8.15}$$

where $\{\varepsilon_t\}$ are i.i.d. from $N\left(0,\, 0.2^2\right)$ and

$$\begin{aligned} a_1(u) &= 0.138 + (0.316 + 0.982\, u)\, e^{-3.89\, u^2}, \\ a_2(u) &= -0.437 - (0.659 + 1.260\, u)\, e^{-3.89\, u^2}. \end{aligned}$$

Figure 8.2 presents the estimated $a_1(\cdot)$ and $a_2(\cdot)$ from a typical sample. The typical sample is selected in such a way that its RASE-value is equal to the median in the 400 simulations. The optimal bandwidth $h = 0.41$ was chosen. The proposed estimators nicely capture the underlying feature of the true coefficient functions. ■

Example 8.5 (*Simulation from a TAR model*) Instead of using continuous coefficient functions, we now use discontinuous step functions

$$\begin{aligned} a_1(u) &= 0.4\, I(u \le 1) - 0.8\, I(u > 1), \\ a_2(u) &= -0.6\, I(u \le 1) + 0.2\, I(u > 1). \end{aligned}$$

Four hundred series of length 400 were simulated from the TAR model

$$X_t = a_1(X_{t-2})X_{t-1} + a_2(X_{t-2})X_{t-2} + \varepsilon_t. \tag{8.16}$$

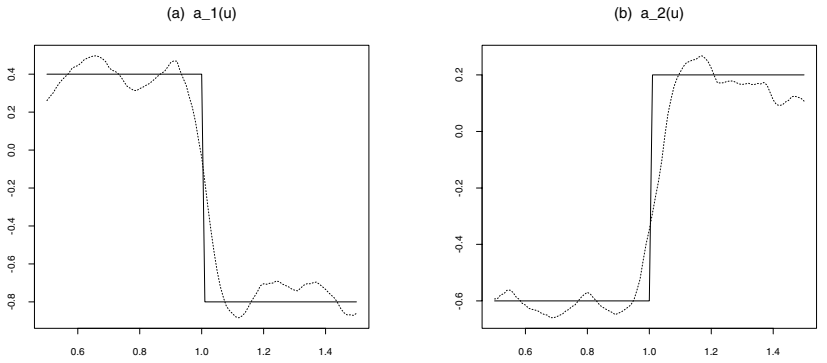


FIGURE 8.3. Simulation results for Example 8.5. The local linear estimators (dotted curves) for the coefficient functions $a_1(\cdot)$ and $a_2(\cdot)$ (solid curves). From Cai, Fan, and Yao (2000).

TABLE 8.2. The mean and SD of AAPE based on 400 replications. Reproduced from Cai, Fan, and Yao (2000).

	One-step	Iterative two-step	Direct two-step
Model (8.16)	0.784(0.203)	0.904(0.273)	0.918(0.281)
Linear AR(2)	1.131(0.485)	1.117(0.496)	

The resulting typical estimates from the 400 simulations are depicted in Figure 8.3. The optimal bandwidth $h_n = 0.325$ was used. The procedure captures the change-point feature quite nicely. A further improvement can be obtained by using nonparametric change-point techniques (see, e.g., Müller 1992, Gijbels, Hall, and Kneip 1995) or the parametric TAR model.

To compare the prediction performance of the predictors from functional-coefficient modeling with the best-fitted linear AR(2) model

$$\widehat{X}_t = \widehat{\beta}_0 + \widehat{\beta}_1 X_{t-1} + \widehat{\beta}_2 X_{t-2},$$

we predict 10 postsample points in each of the 400 replicated simulations. The mean and standard deviation (SD, in parentheses in Table 8.2) of average absolute prediction errors (AAPE) are recorded in Table 8.2. Note that $E|\varepsilon_t| = 0.7979$ and $\mathrm{SD}(|\varepsilon_t|) = 0.6028$ so that the average of 10 absolute deviation errors has an SD of 0.1897. These are indeed very close to the one-step AAPE and its associated SD using model (8.16) and imply that the errors in estimating functions $a_1(\cdot)$ and $a_2(\cdot)$ are negligible in the prediction. The FAR(2,2) model, while somewhat overparametrized in the coefficient functions, provides relevant predictors for the given model

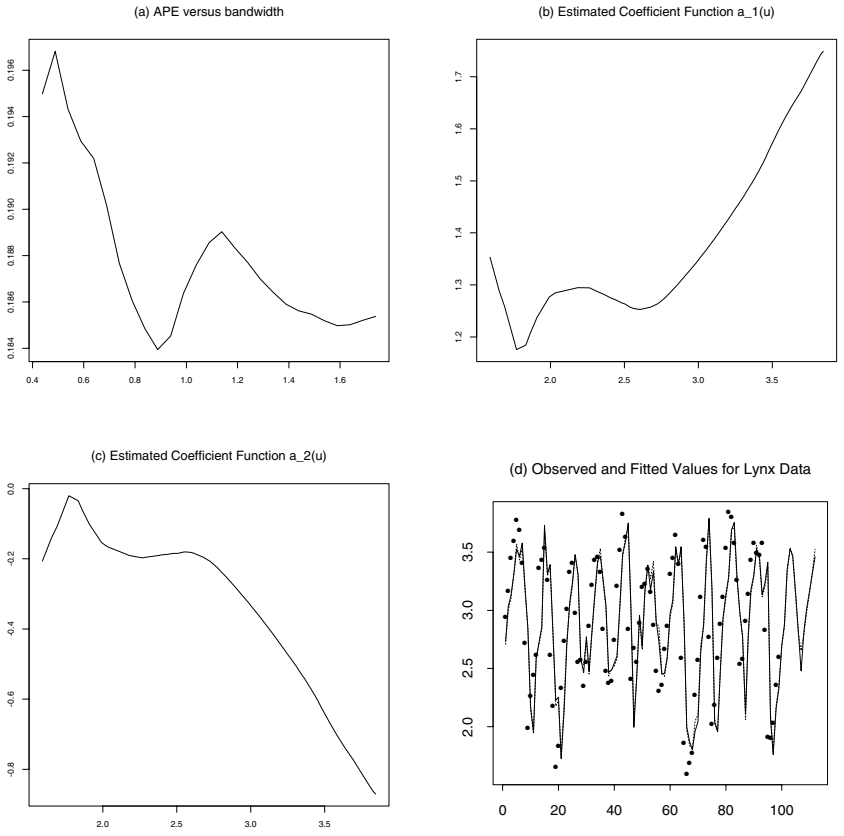


FIGURE 8.4. Canadian lynx data. (a) APE against bandwidth. (b) Local linear estimate $\widehat{a}_1(\cdot)$. (c) Local linear estimate $\widehat{a}_2(\cdot)$. (d) Original series and fitted series by using TAR model (solid) and FAR model (dashed). Adapted from Cai, Fan, and Yao (2000).

(8.16). The direct two-step predictor based on the FAR model (8.14) performs reasonably well. This in turn illustrates the flexibility of this family of models as approximations to true stochastic dynamics. ∎

Example 8.6 (*Canadian lynx data*) A natural alternative model to the TAR model (1.8) for the Canadian lynx data is the FAR(2, 2) model. We apply the APE criterion with $Q = 4$ and $m = 11$ to choose a bandwidth. The function APE against the bandwidth h over a grid of points $h_j = 0.6 + 0.05j (j = 0, \cdots, 12)$ is plotted in Figure 8.4(a). The selected bandwidth is $h = 0.90$. Using this bandwidth and the local linear regression

(8.7), we obtain the estimated coefficients $\widehat{a}_1$ and $\widehat{a}_2$, which are depicted in Figures 8.4 (b) and (c). The fitted values are presented in Figure 8.4(d). For comparison purposes, we also plot the fitted value using the TAR model (1.8). The fitted values are almost undifferentiable.

The resulting FAR(2, 2) model resembles some important features of the TAR(2) model (1.8). Both models admit nice biological interpretation on the predator (lynx) and prey (hare) interaction in ecology (Stenseth et al., 1999). The lower regime of X_{t-2} corresponds roughly to the population increase phase, whereas the upper regime corresponds to the population decrease phase. In the population increase phase, the coefficient functions are nearly constant and are similar to those in the TAR(2) model. The coefficient of X_{t-1} in the model is positive, and more so during the decrease phase, whereas the coefficient of X_{t-2} is negative, and more so during the decrease phase. The signs of those coefficients reveal that lynx and hares relate with each other in a specified prey–predator interactive manner. The dependence of the coefficients on the phases of increase and decrease reflects the so-called phase-dependence and density-dependence in ecology (Stenseth et al. 1999). The phase-dependence refers to the different behavior of preys and predators in hunting and escaping at the increasing or decreasing phase of the population. The density-dependence implies the dependence of reproduction rates of animals as well as their behavior on the abundance of the population. The key difference between the FAR(2, 2) and TAR(2) models is whether the coefficient functions should be smooth or radical in population density. This is an issue of interpretation and belief. In fact, as will be shown in Chapter 9, there is no statistically significant difference between the two models. In other words, given the available data, these two models are statistically indistinguishable.

To compare the prediction performance among various models and several prediction procedures, we fit model (8.16), a TAR model, and a linear AR(2) model using the first 102 data points only, leaving out the last 12 points for assessing the prediction performance. The fitted TAR(2) model is

$$\widehat{X}_t = \begin{cases} 0.424 + 1.255X_{t-1} - 0.348X_{t-2}, & X_{t-2} \le 2.981, \\ 1.882 + 1.516X_{t-1} - 1.126X_{t-2}, & X_{t-2} > 2.981, \end{cases} \tag{8.17}$$

and the fitted linear AR(2) model is

$$\widehat{X}_t = 1.048 + 1.376X_{t-1} - 0.740X_{t-2}.$$

The absolute prediction errors are reported in Table 8.3. The FAR(2, 2) model outperforms both the TAR(2) and linear AR(2) models. ∎

Example 8.7 (*Sunspot data*) We use the sunspot data to illustrate how to use the APE criterion to select the order p and the model-dependent variable X_{t-d} in FAR(p, d). Following the convention in the literature,

TABLE 8.3. The postsample prediction errors for Canadian lynx data. From Cai, Fan, and Yao (2000).

		FAR(2,2) model			TAR model (8.17)		Linear AR(2)	
Year	X_t	OS	Iter	Direct	OS	Iter	OS	Iter
1923	3.054	0.157	0.156	0.209	0.187	0.090	0.173	0.087
1924	3.386	0.012	0.227	0.383	0.035	0.269	0.061	0.299
1925	3.553	0.021	0.035	0.195	0.014	0.038	0.106	0.189
1926	3.468	0.008	0.037	0.034	0.022	0.000	0.036	0.182
1927	3.187	0.085	0.101	0.295	0.059	0.092	0.003	0.046
1928	2.723	0.055	0.086	0.339	0.075	0.015	0.143	0.148
1929	2.686	0.135	0.061	0.055	0.273	0.160	0.248	0.051
1930	2.821	0.016	0.150	0.318	0.026	0.316	0.093	0.434
1931	3.000	0.017	0.037	0.111	0.030	0.062	0.058	0.185
1932	3.201	0.007	0.014	0.151	0.060	0.043	0.113	0.193
1933	3.424	0.089	0.098	0.209	0.076	0.067	0.191	0.347
1934	3.531	0.053	0.175	0.178	0.072	0.187	0.140	0.403
AAPE		0.055	0.095	0.206	0.073	0.112	0.114	0.214

"OS" stands for one-step prediction; "Iter" for iterative two-step estimator; "Direct" for direct two-step estimator.

TABLE 8.4. Selected FAR models for the Sunspot Data. Adapted from Cai, Fan, and Yao (2000).

p	2	3	4	5	6
d	1	3	3	2	2
APE	18.69	13.46	13.90	12.26	13.93
p	7	8	9	10	11
d	3	3	5	3	5
APE	11.68	11.95	14.06	14.26	13.91

the transform $X_t = 2(\sqrt{1+Y_t} - 1)$ was applied to the original series. In order to compare our analysis with the previous ones by Chen and Tsay (1993) and Ghaddar and Tong (1981), only the annual sunspot numbers in 1700–1987 were considered. The parameters $m = 28$ and $Q = 4$ were used to select parameters p, d, and h. For each given $2 \le p \le 11$, the APE-criterion (8.11) is applied to choose the optimal parameter d. The results are recorded in Table 8.4. The overall optimal model is $p = 7$ or $p = 8$; the model-dependent variable is at lag $d = 3$.

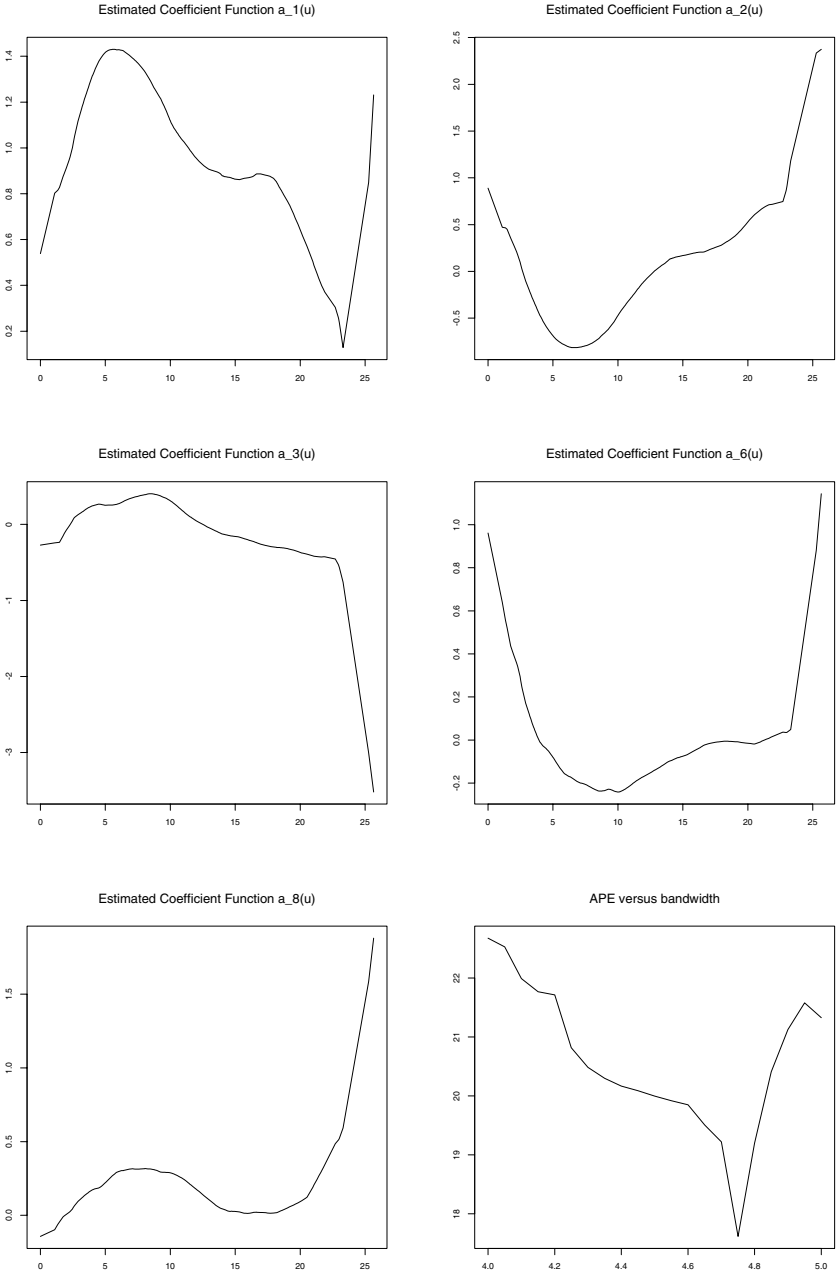


FIGURE 8.5. Wolf's sunspot data. (a)–(e) The estimated functional coefficients in model (8.18). (f) The plot of the APE against bandwidth for model (8.18). Taken from Cai, Fan, and Yao (2000).

TABLE 8.5. The postsample prediction errors for sunspot data from Cai, Fan, and Yao (2000).

		FAR model (8.19)			FAR model (8.18)		FAR model (8.20)	
Year	x_t	OS	Iter	Direct	Error	Iter	Error	Iter
1980	154.7	1.4	1.4	1.4	13.8	13.8	5.5	5.5
1981	140.5	11.4	10.4	3.7	0.0	3.8	1.3	0.0
1982	115.9	15.7	20.7	12.9	10.0	16.4	19.5	22.1
1983	66.6	10.3	0.7	11.0	3.3	0.8	4.8	6.5
1984	45.9	1.0	1.5	4.3	3.8	5.6	14.8	15.9
1985	17.9	2.6	3.4	7.8	4.6	1.7	0.2	2.7
1986	13.4	3.1	0.7	1.9	1.3	2.5	5.5	5.4
1987	29.2	12.3	13.1	18.9	21.7	23.6	0.7	17.5
AAPE		7.2	6.5	7.7	7.3	8.3	6.6	9.5

"OS" stands for one-step prediction; "Iter" for iterative two-step estimator; "Direct" for direct two-step estimator.

Following Table 8.4, we fit an FAR(8, 3) model. Insignificant variables were deleted in Chen and Tsay (1993), leading to the fitted FAR model

$$X_t = \begin{cases} 1.23 + (1.75 - 0.17\,|X_{t-3} - 6.6|)\,X_{t-1} + (-1.28 + \\ \qquad 0.27|X_{t-3} - 6.6|)\,X_{t-2} + 0.20\,X_{t-8} + \varepsilon_t^{(1)}, & \text{if } x_{t-3} < 10.3, \\ 0.92 - 0.24\,x_{t-3} + 0.87\,x_{t-1} + 0.17\,x_{t-2} - 0.06\,x_{t-6} \\ \qquad +0.04\,x_{t-8} + \varepsilon_t^{(2)}, & \text{if } x_{t-3} \geq 10.3. \end{cases} \tag{8.18}$$

This and the model selection result above suggest that we fit the following FAR model

$$\begin{aligned} X_t \;=\; & a_1(X_{t-3})\,X_{t-1} + a_2(X_{t-3})\,X_{t-2} + a_3(X_{t-3})\,X_{t-3} \\ & + a_6(X_{t-3})\,X_{t-6} + a_8(X_{t-3})\,X_{t-8} + \varepsilon_t. \end{aligned} \tag{8.19}$$

The local linear estimator is employed with the bandwidth $h = 4.75$ selected by the APE (see Figure 8.5(f)). The estimated coefficients are reported in Figures 8.5 (a)–(e).

Model (8.18) was fitted by using the first 280 data points (in 1700–1979). To make fair comparisons on the prediction performance, we only use these data to estimate the coefficient functions in (8.19). The following TAR model (Tong 1990, p. 420)

$$X_t = \begin{cases} 1.92 + 0.84\,X_{t-1} + 0.07\,X_{t-2} - 0.32\,X_{t-3} + 0.15\,X_{t-4} \\ -0.20\,X_{t-5} - 0.00\,X_{t-6} + 0.19\,X_{t-7} - 0.27\,X_{t-8} \\ +0.21\,X_{t-9} + 0.01\,X_{t-10} + 0.09\,X_{t-11} + \varepsilon_t^{(1)}, & \text{if } X_{t-8} \leq 11.93, \\ 4.27 + 1.44\,X_{t-1} - 0.84\,X_{t-2} + 0.06\,X_{t-3} + \varepsilon_t^{(2)}, & \text{if } X_{t-8} > 11.93, \end{cases} \tag{8.20}$$

resulting from the fit using the same length of data, was included for comparison. The results are recorded in Table 8.5. According to the average

absolute prediction errors, the nonparametric model performs as well as both the TAR and FAR models in the one-step-ahead prediction. Furthermore, it outperforms in two-step prediction with both iterative and direct methods. ■

*8.3.8 Sampling Properties**

We first present a result on mean square convergence that serves as a building block to our main result. It is also of independent interest. The idea of the proof here is similar to that used in proving Theorem 6.3. We first introduce some notation. Let

$$\mathbf{S}_n = \mathbf{S}_n(u_0) = \begin{pmatrix} \mathbf{S}_{n,0} & \mathbf{S}_{n,1} \\ \mathbf{S}_{n,1} & \mathbf{S}_{n,2} \end{pmatrix} \quad \text{and} \quad \mathbf{T}_n = \mathbf{T}_n(u_0) = \begin{pmatrix} \mathbf{T}_{n,0}(u_0) \\ \mathbf{T}_{n,1}(u_0) \end{pmatrix},$$

where

$$\mathbf{S}_{n,j} = \mathbf{S}_{n,j}(u_0) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \, \mathbf{X}_i^T \left(\frac{U_i - u_0}{h} \right)^j K_h(U_i - u_0)$$

and

$$\mathbf{T}_{n,j}(u_0) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \left(\frac{U_i - u_0}{h} \right)^j K_h(U_i - u_0) \, Y_i.$$

Then, the solution to (8.7) can be written as

$$\widehat{\boldsymbol{\beta}} = \mathbf{H}^{-1} \, \mathbf{S}_n^{-1} \, \mathbf{T}_n. \tag{8.21}$$

Set $\mathbf{H} = \text{diag}\,(1, \ldots, 1, \; h, \ldots, h)$ with the first p diagonal elements ones and the last p diagonal elements h. Denote

$$\Omega = \Omega(u_0) = (\omega_{l,m})_{p \times p} = E\left(\mathbf{X} \, \mathbf{X}^T \, | \, U = u_0 \right). \tag{8.22}$$

Also, let $f(\mathbf{x}, u)$ denote the joint density of $(\mathbf{X}, U)$, and let $f_U(u)$ be the marginal density of U. The following convention is needed: if $U = X_{j_0}$ for some $1 \le j_0 \le p$, then $f(\mathbf{x}, u)$ becomes $f(\mathbf{x})$ — the joint density of $\mathbf{X}$. Recall that

$$\mu_j = \int_{-\infty}^{\infty} u^j \, K(u) \, du, \qquad \nu_j = \int_{-\infty}^{\infty} u^j \, K^2(u) \, du.$$

The following result is established in Cai, Fan, and Yao (2000).

Theorem 8.2 *Assume that Condition 1 in §8.8.2 holds, and $f(\mathbf{x}, u)$ is continuous at the point u_0. Let $h_n \to 0$ in such a way that $n \, h_n \to \infty$. Then*

$$E\{\mathbf{S}_{n,j}(u_0)\} \to f_U(u_0) \, \Omega(u_0) \, \mu_j,$$

and

$$n \, h_n \, Var\{\mathbf{S}_{n,j}(u_0)_{l,m}\} \to f_U(u_0) \, \nu_{2j} \, \omega_{l,m}$$

for all $0 \le j \le 3$ and $1 \le l, \; m \le p$.

The proof of this theorem is similar but less involved than that of Lemma 8.4 in §8.8.3, and thus its proof is omitted.

As a consequence of Theorem 8.2, the variance of each element in $\mathbf{S}_{n,j}$ converges to zero. Hence, each element in $\mathbf{S}_{n,j}$ converges to its expected value in probability. As a result, we have

$$\mathbf{S}_n \xrightarrow{P} f_U(u_0)\,\mathbf{S} \qquad \text{and} \qquad \mathbf{S}_{n,3} \xrightarrow{P} \mu_3\, f_U(u_0)\,\Omega$$

in the sense that each element converges in probability, where $\mathbf{S} = \text{diag}(1, \mu_2) \otimes \Omega$ is the Kronecker product of the 2×2 diagonal matrix $\text{diag}(1, \mu_2)$ and Ω. Denote

$$\sigma^2(\mathbf{x},\, u) = \text{Var}(Y \,|\, \mathbf{X} = \mathbf{x},\, U = u) \tag{8.23}$$

and

$$\Omega^*(u_0) = E\left[\mathbf{X}\,\mathbf{X}^T\,\sigma^2(\mathbf{X},\, U)\,\big|\, U = u_0\right]. \tag{8.24}$$

The following result has been proved in Cai, Fan, and Yao (2000).

Theorem 8.3 *Let $\sigma^2(\mathbf{x},\, u)$ and $f(\mathbf{x},\, u)$ be continuous at the point u_0. Then, under Conditions 1 and 2 in §8.8.2,*

$$\sqrt{n\,h_n}\left[\widehat{\mathbf{a}}(u_0) - \mathbf{a}(u_0) - \frac{h^2}{2}\mu_2\mathbf{a}''(u_0)\right] \xrightarrow{D} N\left(0,\ \Theta^2(u_0)\right), \tag{8.25}$$

provided that $f_U(u_0) \neq 0$, where

$$\Theta^2(u_0) = \frac{\nu_0}{f_U(u_0)}\,\Omega^{-1}(u_0)\,\Omega^*(u_0)\,\Omega^{-1}(u_0). \tag{8.26}$$

Theorem 8.3 reveals that the asymptotic bias of $\widehat{a}_j(u_0)$ is $\frac{h^2}{2}\mu_2 a_j''(u_0)$ and the asymptotic variance is $(n\,h_n)^{-1}\,\theta_j^2(u_0)$, where $\theta_j^2(u_0)$ is the j-th diagonal element of $\Theta^2(u_0)$.

8.4 Adaptive Functional-Coefficient Autoregressive Models

The FAR model (8.1) depends critically on the choice of the model dependent variable X_{t-d}. The model-dependent variable is one of the lagged variables. This limits the scope of its applications. A generalization of this class of models is to allow a linear combination of past values as a model-dependent variable. This is also a generalization of thresholding models with unknown thresholding directions.

8.4.1 The Models

Let $G(x_1, \cdots, x_p) = E(X_t | X_{t-1} = x_1, \cdots, X_{t-p} = x_p)$ be the autoregressive function. Then, one can write

$$X_t = G(X_{t-1}, \cdots, X_{t-p}) + \varepsilon_t, \tag{8.27}$$

with $E\{\varepsilon_t | X_{t-1}, \cdots, X_{t-p}\} = 0$. The autoregressive function G is the best prediction function in the sense that G minimizes the expected prediction error:

$$\min_g E\Big(X_t - g(X_{t-1}, \cdots, X_{t-p})\Big)^2.$$

As mentioned in §8.1, the saturated nonparametric function $G(x_1, \cdots, x_p)$ cannot be estimated with reasonable accuracy due to the curse of dimensionality. Thus, some forms on $G(\cdot)$ are frequently imposed. They are often approximations to the function $G(\cdot)$. For example, the model (8.1) can be viewed as searching for the best FAR model to approximate the function G. The larger the class of the models, the smaller the approximation errors (modeling biases) but the larger the variance of the estimated unknown parameters/functions. Therefore, there is always a trade-off between these two competing demands.

A generalization of the FAR model is to allow its coefficient functions to depend on the linear combinations of past values, called *indices.* Let $\mathbf{X}_{t-1} = (X_{t-1}, \cdots, X_{t-p})$ and $\boldsymbol{\beta}$ be an unknown direction in the p-dimensional space $\Re^p$, namely $\|\boldsymbol{\beta}\| = 1$. The adaptive FAR (AFAR) model approximates the autoregressive function G by the family of functions of form

$$g(\mathbf{x}) = g_0(\boldsymbol{\beta}^T \mathbf{x}) + \sum_{j=1}^{p} g_j(\boldsymbol{\beta}^T \mathbf{x})\, x_j. \tag{8.28}$$

In particular, when the function G admits really the form (8.28), namely $G(\mathbf{x}) = g(\mathbf{x})$, the *adaptive FAR model* (AFAR) is given by (see (8.27))

$$X_t = g_0(\boldsymbol{\beta}^T \mathbf{X}_{t-1}) + \sum_{j=1}^{p} g_j(\boldsymbol{\beta}^T \mathbf{X}_{t-1}) X_{t-j} + \varepsilon_t. \tag{8.29}$$

In addition, it is typically assumed that ε_t is independent of $\mathbf{X}_{t-1}$.

The class of AFAR models is clearly larger than the class of FAR models. This allows one to choose the important model-dependent direction $\boldsymbol{\beta}$ to reduce modeling biases. On the other hand, the extra parameters in $\boldsymbol{\beta}$ do not introduce much extra difficulty in statistical estimation. Indeed, the parameter $\boldsymbol{\beta}$ can usually be estimated at the root-n rate, and g can be estimated as well as if $\boldsymbol{\beta}$ were known. Model (8.29) includes many useful statistical models. Here are some examples.

Example 8.8 (*FAR model*) If we let $\boldsymbol{\beta}$ be the unit vector with the dth position 1 and the rest elements 0, then $\boldsymbol{\beta}^T\mathbf{X}_{t-1} = X_{t-d}$. Thus, the model (8.29) includes the FAR model (8.1) as a specific case. By searching for the best direction $\boldsymbol{\beta}$, we allow the model-dependent variable not only the lagged variables but also their linear combinations. ∎

Example 8.9 (*Expanded variables*) As in multiple linear regression, the FAR model and its related techniques can be applied to the situations with expanded variables such as the transformations of the lagged variables and their interactions. For example, the techniques would allow one to handle the model

$$\begin{aligned} X_t &= g_0(\boldsymbol{\beta}^T\mathbf{X}_{t-1}) + \sum_{j=1}^{p} g_j(\boldsymbol{\beta}^T\mathbf{X}_{t-1})X_{t-j} \\ &\quad + \sum_{i=1}^{p}\sum_{j=1}^{p} g_{ij}(\boldsymbol{\beta}^T\mathbf{X}_{t-1})X_{t-i}X_{t-j}. \end{aligned} \tag{8.30}$$

By regarding this model as a stochastic regression model as in §8.3.2, we introduce the dependent variable $Y_t = X_t$ and the predictors

$$\begin{aligned} &X_{t0} = 1, \quad X_{t1} = X_{t-1}, \quad \cdots, \quad X_{tp} = X_{t-p}, \\ &X_{t,p+1} = X_{t-1}^2, \quad X_{t,p+2} = X_{t-1}X_{t-2}, \quad \cdots, \quad X_{t,q} = X_{t-p}^2, \end{aligned}$$

where $q = p + p(p+1)/2$. Then, the model (8.29) can be written as

$$Y_t = \mathbf{g}(\beta_1 X_{t1} + \cdots + \beta_p X_{tp})^T\mathbf{X}_t^* + \varepsilon_t,$$

where $\mathbf{X}_t^*$ is a vector of length $1+p+p(p+1)/2$ containing all of the aforementioned predictors, and $\mathbf{g}$ is a vector of their corresponding coefficient functions. The techniques in §8.4.3–§8.4.6 continue to apply. ∎

Example 8.10 (*Single index model*) By taking the rest of the coefficients in model (8.30) as zero, the AFAR model can be used to handle the following single-index model:

$$X_t = g_0(\boldsymbol{\beta}^T\mathbf{X}_{t-1}) + \varepsilon_t.$$

This model has been popularly studied in the literature. See, for example, Härdle, Hall, and Ichimura (1993), Ichimura (1993), Newey and Stoker (1993), Samarov (1993), Carroll, Fan, Gijbels, and Wand (1997), and Heckman, Ichimura, Smith, and Todd (1998), among others. ∎

8.4.2 Existence and Identifiability

A fundamental question arises whether there exists a unique function of the form (8.28) such that it minimizes the prediction error

$$E\{X_t - g(\mathbf{X}_{t-1})\}^2.$$

By using the bias and variance decomposition,

$$E\{X_t - g(\mathbf{X}_{t-1})\}^2 = E\{G(\mathbf{X}_{t-1}) - g(\mathbf{X}_{t-1})\}^2 + \mathrm{Var}(X_t),$$

so the problem becomes whether there exist functions $\{g_j\}$ such that the resulting AFAR model best approximates the autoregressive function G.

Another important question is whether the model (8.28) is *identifiable.* First, the model as presented in (8.28) is not identifiable. If the coefficient $\beta_p \neq 0$, then

$$X_{t-p} = (\boldsymbol{\beta}^T \mathbf{X}_{t-1} - \beta_1 X_{t-1} - \cdots - \beta_{p-1} X_{t-p-1})/\beta_p.$$

Substituting this into (8.28), $g(\mathbf{X}_{t-1})$ can be written as

$$g(\mathbf{X}_{t-1}) = g_0^*(\boldsymbol{\beta}^T \mathbf{X}_{t-1}) + \sum_{j=1}^{p-1} g_j^*(\boldsymbol{\beta}^T \mathbf{X}_{t-1}) X_{t-j},$$

where

$$g_0^*(u) = g_0(u) + u g_p(u)/\beta_p, \quad g_j^*(u) = g_j(u) - \beta_j g_p(u)/\beta_p.$$

Thus, there is a redundant term in the model (8.28) and this term should be eliminated. For this reason, we rewrite (8.28) as

$$g(\mathbf{x}) = g_0(\boldsymbol{\beta}^T \mathbf{x}) + \sum_{j=1}^{p-1} g_j(\boldsymbol{\beta}^T \mathbf{x})\, x_j. \tag{8.31}$$

Even after eliminating the redundant term as in (8.31), the model still may not be identifiable. For example, if

$$g(\mathbf{x}) = (\boldsymbol{\beta}_1^T \mathbf{x})(\boldsymbol{\beta}_2^T \mathbf{x}),$$

it is of form (8.28) or (8.31). However, $\boldsymbol{\beta}$ is not identifiable. It can be either $\boldsymbol{\beta}_1$ or $\boldsymbol{\beta}_2$.

To present the results on the existence and identifiability, we consider a more general stochastic regression model in which Y is any response variable and $\mathbf{X} = (X_1, \cdots, X_p)^T$ is a vector of predictors. The following result, due to Fan, Yao, and Cai (2002), shows that the solution to the minimization problem

$$\inf_{\boldsymbol{\alpha}} \inf_{f_0, \ldots, f_{p-1}} E \left\{ Y - f_0(\boldsymbol{\alpha}^T \mathbf{X}) - \sum_{j=1}^{p-1} f_j(\boldsymbol{\alpha}^T \mathbf{X}) X_j \right\}^2 \tag{8.32}$$

exists and the model is identifiable as long as g is not in a class of the specific quadratic functions.

Theorem 8.4 *(i) Assume that* $(\mathbf{X}, Y)$ *has a continuous density and* $Var(Y) + Var(\|\mathbf{X}\|) < \infty$. *Then, there exists a* $g(\cdot)$ *of form (8.31) that minimizes (8.32), provided that* $Var(\mathbf{X}^*|\boldsymbol{\beta}^T\mathbf{X})$ *is nondegenerate, where* $\mathbf{X}^* = (1, X_1, \cdots, X_{p-1})$.
(ii) If $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ *is given and* $\beta_p \neq 0$, *then the functions* $g_j(\cdot)$ $(j = 0, \cdots, p-1)$ *are uniquely determined from* g, *namely, they are identifiable.*
(iii) For any given twice-differentiable $g(\cdot)$ *of the form (8.31), if the first non-zero component of* $\boldsymbol{\beta}$ *is chosen to be positive, such a* $\boldsymbol{\beta}$ *with* $\|\boldsymbol{\beta}\| = 1$ *is unique unless* $g(\cdot)$ *is of the form that*

$$g(\mathbf{x}) = \boldsymbol{\alpha}^T\mathbf{x}\,\boldsymbol{\beta}^T\mathbf{x} + \boldsymbol{\gamma}^T\mathbf{x} + c \tag{8.33}$$

for some constant vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, *and* $\boldsymbol{\gamma}$ *and a constant* c *with* $\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$ *nonparallel.*

The proof of this theorem is given in §8.8.5.

8.4.3 Profile Least-Squares Estimation

The class of model (8.28) was introduced by Ichimura (1993), Xia and Li (1999a), and Fan, Yao, and Cai (2002). For brevity, we only present the methods used in the last paper.

Due to identifiability considerations, from now on we assume that $\beta_p \neq 0$ and take the model (8.31). Furthermore, we assume that g is not of form (8.33) so that the coefficient $\boldsymbol{\beta}$ is identifiable. Let $\{(\mathbf{X}_t, Y_t), t = 1, \cdots, n\}$ be the observed data from a stationary sequence.

The basic idea for fitting such a semiparametric method is the *profile likelihood* method (more precisely, it is a *profile least-squares* method in the current context). The idea has been used before by Severini and Wong (1992), Carroll, Fan, Gijbels, and Wand (1997), and Murphy and van der Vaart (2000), among others. In various contexts (see the aforementioned papers), the profile likelihood estimators for parametric components are semiparametrically efficient (see, e.g., Bickel, Klaassen, Ritov, and Wellner 1993 for a definition), and the nonparametric components are estimated as well as if the parametric components were known. The basic idea of the profile likelihood method is to first estimate the nonparametric functions with a given $\boldsymbol{\beta}$, resulting in estimates $\widehat{g}_j(\cdot; \boldsymbol{\beta})$, and to then estimate the unknown parameter $\boldsymbol{\beta}$ using the estimated nonparametric functions $\widehat{g}_j(\cdot; \boldsymbol{\beta})$, which themselves depend on $\boldsymbol{\beta}$. Substituting the nonparametric estimate into (8.31) results in a synthetic parametric model:

$$g(\mathbf{x}) = \widehat{g}_0(\boldsymbol{\beta}^T\mathbf{x}; \boldsymbol{\beta}) + \sum_{j=1}^{p-1} g_j(\boldsymbol{\beta}^T\mathbf{x}; \boldsymbol{\beta})\, x_j.$$

Using the least-squares method to the "synthetic parametric model," we can obtain an estimate $\widehat{\boldsymbol{\beta}}$. The profile least-squares estimate is to use $\widehat{\boldsymbol{\beta}}$ to

estimate $\boldsymbol{\beta}$ and $\widehat{g}_j(\cdot;\widehat{\boldsymbol{\beta}})$ to estimate the coefficient function $g_j(\cdot)$. To ease the computational burden on the nonlinear least-squares estimate in the "synthetic parametric model," an iterative scheme is frequently employed.

The idea is as follows. Given an initial estimate $\widehat{\boldsymbol{\beta}}_0$ of $\boldsymbol{\beta}$, one obtains estimated coefficient functions $\widehat{g}_j(\cdot;\widehat{\boldsymbol{\beta}}_0)$ and the "synthetic parametric model"

$$g(\mathbf{x}) = \widehat{g}_0(\boldsymbol{\beta}^T\mathbf{x};\widehat{\boldsymbol{\beta}}_0) + \sum_{j=1}^{p-1} g_j(\boldsymbol{\beta}^T\mathbf{x};\widehat{\boldsymbol{\beta}}_0)\, x_j.$$

Applying the least-squares method, we obtain a new estimate $\widehat{\boldsymbol{\beta}}_1$. This updates the estimate of $\boldsymbol{\beta}$. With updated $\widehat{\boldsymbol{\beta}}_1$, we update the nonparametric components and obtain the estimates $\widehat{g}_j(\cdot;\widehat{\boldsymbol{\beta}}_1)$. With the new nonparametric estimates, we can further update the estimate of parametric component $\boldsymbol{\beta}$. Keep iterating until a convergence criterion is met. We now outline the key idea further.

A. Local linear estimators given $\boldsymbol{\beta}$

From (8.32), with given $\boldsymbol{\alpha} = \boldsymbol{\beta}$, the coefficient functions $\{g_j(z)\}$, $j = 0, \cdots, p-1$ are obtained by minimizing

$$E\left\{\left[Y - \sum_{j=0}^{p-1} f_j(z)X_j\right]^2 |\boldsymbol{\beta}^T\mathbf{X} = z\right\}^2$$

with respect to f_j. This in turn suggests that the functions g_j can be obtained by the locally linear regression around the neighborhood $\boldsymbol{\beta}^T\mathbf{X} \approx z$. This leads to minimizing the sum

$$\sum_{t=1}^{n}\left[Y_t - \sum_{j=0}^{p-1}\left\{b_j + c_j\,(\boldsymbol{\beta}^T\mathbf{X}_t - z)\right\} X_{tj}\right]^2 K_h(\boldsymbol{\beta}^T\mathbf{X}_t - z)\, w(\boldsymbol{\beta}^T\mathbf{X}_t) \tag{8.34}$$

with respect to $\{b_j\}$ and $\{c_j\}$. Here, the weight function $w(\cdot)$ is introduced to attenuate the boundary effect. Let

$$\widehat{\boldsymbol{\theta}} \equiv \left(\widehat{b}_0,\ \ldots,\ \widehat{b}_{p-1},\ \widehat{c}_0,\ \ldots,\ \widehat{c}_{p-1}\right)^T$$

be the solution to the local regression problem (8.34). Define the estimators $\widehat{g}_j(z) = \widehat{b}_j$ and $\widehat{\dot{g}}_j(z) = \widehat{c}_j$, where $\widehat{\dot{g}}_j(z)$ is an estimator of the derivative function of g. In fact, for given $\boldsymbol{\beta}$, model (8.31) is an FAR model with the model-dependent variable $U = \boldsymbol{\beta}^T\mathbf{X}$. This step is essentially the same as that in the FAR model; see (8.7).

It follows from the least-squares theory that

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\Sigma}(z)\mathcal{X}^T(z)\mathcal{W}(z)\mathcal{Y}, \quad \text{with} \quad \boldsymbol{\Sigma}(z) = \{\mathcal{X}^T(z)\mathcal{W}(z)\mathcal{X}(z)\}^{-1}, \tag{8.35}$$

where $\mathcal{Y} = (Y_1, \ldots, Y_n)^T$, $\mathcal{W}(z)$ is an $n \times n$ diagonal matrix with $K_h(\boldsymbol{\beta}^T \mathbf{X}_i - z) w(\boldsymbol{\beta}^T \mathbf{X}_i)$ as its ith diagonal element, $\mathcal{X}(z)$ is an $n \times 2p$ matrix with $(\mathbf{U}_i^T, (\boldsymbol{\beta}^T \mathbf{X}_i - z)\mathbf{U}_i^T)$ as its i-th row, and $\mathbf{U}_t = (1, X_{t1}, \ldots, X_{t,p-1})^T$.

B. Estimate $\boldsymbol{\beta}$ for given g's

The property (8.32) suggests estimating $\boldsymbol{\beta}$ by minimizing

$$R(\boldsymbol{\beta}) = \frac{1}{n} \sum_{t=1}^{n} \left\{ Y_t - \sum_{j=0}^{d-1} \widehat{g}_j(\boldsymbol{\beta}^T \mathbf{X}_t) X_{tj} \right\}^2 w(\boldsymbol{\beta}^T \mathbf{X}_t), \tag{8.36}$$

where $X_{t0} = 1$. The weight $w(\cdot)$ is used to mitigate the influence of the nonparametric estimate at tails of $\boldsymbol{\beta}^T \mathbf{X}$. Note that estimates $\{\widehat{g}_j\}$ themselves depend on $\boldsymbol{\beta}$ (i.e., $\widehat{g}_j(\boldsymbol{\beta}^T \mathbf{X}_t)$ is really $\widehat{g}_j(\boldsymbol{\beta}^T \mathbf{X}_t; \boldsymbol{\beta})$). Directly minimizing the profile least-squares function $R(\boldsymbol{\beta})$ is an undue task. Instead, one often uses an iterative scheme to minimize (8.36). Regarding $\boldsymbol{\beta}$ in the nonparametric estimates $\{\widehat{g}_j(\cdot; \boldsymbol{\beta})\}$ and weights $w(\boldsymbol{\beta}^T \mathbf{X}_t)$ as known-namely, given the functions $\{\widehat{g}_j\}$ and weights $w(\boldsymbol{\beta}^T \mathbf{X}_t)$, one minimizes (8.36) with respect to $\boldsymbol{\beta}$-with estimated new $\boldsymbol{\beta}$, one uses (8.34) to update the estimates of $\{g_j\}$ and iterates between these two steps until a certain convergence criterion is met.

Even with given functions $\{g_j\}$, minimizing $R(\boldsymbol{\beta})$ could still be very expensive. Since the minimizer in the iterative scheme is only an intermediate estimate, we do not really need to find the actual minimizer. A simple method is to iterate the Newton–Raphson step once or a few times to update $\boldsymbol{\beta}$ rather than to actually minimize (8.36). The derived estimator is expected to perform well if the initial value is reasonably good (see Bickel, 1975; Robinson 1988; Fan and Chen 1999). We outline the procedure below.

Suppose that $\widehat{\boldsymbol{\beta}}$ is the minimizer of (8.36). Then $\dot{R}\left(\widehat{\boldsymbol{\beta}}\right) = 0$, where $\dot{R}(\cdot)$ denotes the derivative of $R(\cdot)$. For any given $\boldsymbol{\beta}^{(0)}$ close to $\widehat{\boldsymbol{\beta}}$, we have the approximation

$$0 = \dot{R}(\widehat{\boldsymbol{\beta}}) \approx \dot{R}\left(\boldsymbol{\beta}^{(0)}\right) + \ddot{R}\left(\boldsymbol{\beta}^{(0)}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}\right),$$

where $\ddot{R}(\cdot)$ is the Hessian matrix of $R(\cdot)$. The observation above leads to the one-step iterative estimator

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} - \left\{ \ddot{R}\left(\boldsymbol{\beta}^{(0)}\right) \right\}^{-1} \dot{R}\left(\boldsymbol{\beta}^{(0)}\right), \tag{8.37}$$

where $\boldsymbol{\beta}^{(0)}$ is an initial estimator. We rescale $\boldsymbol{\beta}^{(1)}$ such that it has a unit norm whose first nonvanishing element is positive.

In practice, the matrix $\ddot{R}(\cdot)$ could be singular or nearly so. A common technique to deal with this problem is the ridge regression. For this purpose, we propose using the estimator (8.37) with $\ddot{R}$ replaced by $\ddot{R}_r$, which is defined by replacing the matrix $\mathbf{X}_t \mathbf{X}_t^T$ in $\ddot{R}$ by $\mathbf{X}_t \mathbf{X}_t^T + q_n \mathbf{I}_d$ for some positive ridge parameter q_n.

8.4.4 Bandwidth Selection

The *generalized cross-validation* (GCV) method, proposed by Wahba (1977) and Craven and Wahba (1979), is modified to choose the bandwidth h for the estimation of $\{g_j(\cdot)\}$. The criterion is described as follows. For a given $\boldsymbol{\beta}$, let $\widehat{Y}_t = \sum_{j=0}^{d-1} \widehat{g}_j(\boldsymbol{\beta}^T \mathbf{X}_t) X_{tj}$. It is easy to see that all of those predicted values are in fact the linear combinations of $\mathcal{Y} = (Y_1, \ldots, Y_n)^T$ with coefficients depending on $\{\mathbf{X}_t\}$ only; namely,

$$(\widehat{Y}_1, \ldots, \widehat{Y}_n)^T = \mathbf{H}(h)\mathcal{Y},$$

where $\mathbf{H}(h)$ is an $n \times n$ hat matrix, independent of $\mathcal{Y}$. The GCV method selects h minimizing

$$\mathrm{GCV}(h) \equiv \frac{1}{n\{1 - n^{-1}\mathrm{tr}(\mathbf{H}(h))\}^2} \sum_{t=1}^{n} \{Y_t - \widehat{Y}_t\}^2 w(\boldsymbol{\beta}^T \mathbf{X}_t),$$

which indeed is an estimate of the weighted mean integrated square errors. Under some regularity conditions, it holds that

$$\mathrm{GCV}(h) = a_0 + a_1 h^4 + \frac{a_2}{nh} + o_p(h^4 + n^{-1}h^{-1}).$$

Thus, up to the first order asymptote, the optimal bandwidth is $h_{\mathrm{opt}} = (a_2/(4\, n\, a_1))^{1/5}$. The coefficients of a_0, a_1, and a_2 will be estimated from $\{\mathrm{GCV}(h_k)\}$ via least-squares regression. This rule is inspired by the empirical bias method of Ruppert (1997); see §6.3.5.

8.4.5 Variable Selection

As discussed in Example 8.9, the number of predictors in the AFAR model can be large. Hence, it is important to select significant variables. These variables can be chosen by using either local variable selection or global variable selection. The global variable selection refers to testing whether certain sets of coefficient functions are zero. A natural test statistic is to compare the RSS of two competing nonparametric models. This kind of idea can be found in Chapter 9. We now outline an ad hoc local variable selection method.

The basic idea of the local variable selection is to use a stepwise deletion technique for each given z together with a modified AIC and t-statistics. More precisely, the least significant variable is deleted, one variable every time, according to its t-value. This yields a sequence of new and reduced models. The best model is selected according to the modified AIC. This rule is simple and computationally efficient.

For fixed $\boldsymbol{\beta}^T\mathbf{X} = z$, (8.34) could be viewed as a (local) linear regression with $2p$ variables. The residual sum of squares is given by

$$\begin{aligned}\mathrm{RSS}_p(z) &= \sum_{t=1}^{n}\left[Y_t - \sum_{j=0}^{p-1}\{\widehat{b}_j + \widehat{c}_j(\boldsymbol{\beta}^T\mathbf{X}_t - z)\}X_{tj}\right]^2 \\ &\quad \times K_h(\boldsymbol{\beta}^T\mathbf{X}_t - z)w(\boldsymbol{\beta}^T\mathbf{X}_t).\end{aligned}$$

Let $n_z = \mathrm{tr}\{\mathcal{W}(z)\}$ and $d(p,z) = \mathrm{tr}\{\boldsymbol{\Sigma}(z)\mathcal{X}^T(z)\mathcal{W}^2(z)\mathcal{X}(z)\}$. The former may be regarded as the number of observations used in the local estimation and the latter as the number of local parameters. The "degrees of freedom" of $\mathrm{RSS}_p(z)$ is $f(p,z) = n_z - d(p,z)$. Now, we define the AIC for this model as

$$\mathrm{AIC}_p(z) = \log\{\mathrm{RSS}_p(z)/f(p,z)\} + 2\,d(p,z)/n_z.$$

To delete the least significant variable among $X_0, X_1, \ldots, X_{p-1}$, we search for X_k such that both $g_k(z)$ and $\dot{g}_k(z)$ are close to 0. The t-statistics for those two variables in the (local) linear regression are

$$t_k(z) = \frac{\widehat{g}_k(z)}{\sqrt{c_k(z)\mathrm{RSS}(z)/f(p,z)}} \quad \text{and} \quad t_{p+k} = \frac{\widehat{\dot{g}}_k(z)}{\sqrt{c_{p+k}(z)\mathrm{RSS}(z)/f(p,z)}},$$

respectively, where $c_k(z)$ is the $(k+1, k+1)$ element of matrix

$$\boldsymbol{\Sigma}(z)\mathcal{X}^T(z)\mathcal{W}^2(z)\mathcal{X}(z)\boldsymbol{\Sigma}(z).$$

Discarding a common factor, we define

$$T_k^2(z) = \{\widehat{g}_k(z)\}^2/c_k(z) + \{\widehat{\dot{g}}_k(z)\}^2/c_{p+k}(z).$$

Let j be the minimizer of $T_k^2(z)$ over $0 \le k < p$. Then, the variable X_j is deleted from the full model (8.31) at the point z. This leads to a model with $(p-1)$ "linear terms." Repeating the process above, one obtains a sequence of models and their corresponding $\mathrm{AIC}_l(z)$ for all $1 \le l \le p$. The selected model at the point z should be the one that attains the minimum AIC. This local variable selection method is carried out at each grid point where the functions $\{g_j(\cdot)\}$ are computed.

8.4.6 Implementation

The outline of the algorithm is as follows.

Step 1: Standardize the data set $\{\mathbf{X}_t\}$ such that it has sample mean 0 and the sample variance and covariance matrix $\mathbf{I}_p$. Specify an initial value of $\boldsymbol{\beta}$, say, the coefficient of the (global) linear fitting.

Step 2: For each prescribed bandwidth value h_k, $k = 1, \ldots, q$, repeat (a) and (b) below until two successive values of $R(\boldsymbol{\beta})$ differ insignificantly.

(a) For a given direction $\boldsymbol{\beta}$, we estimate the functions $g_j(\cdot)$ by (8.34).

(b) For given $g_j(\cdot)$'s, we update direction $\boldsymbol{\beta}$ using (8.37).

Step 3: For $k = 1, \ldots, q$, calculate $\mathrm{GCV}(h_k)$ with $\boldsymbol{\beta}$ equal to its estimated value. Let $\widehat{a}_1$ and $\widehat{a}_2$ be the solution to the least-squares problem:

$$\sum_{k=1}^{q} \{\mathrm{GCV}(h_k) - a_0 - a_1 h_k^4 - a_2/(nh_k)\}^2.$$

Define the bandwidth $\widehat{h} = \{\widehat{a}_2/(4\, n\, \widehat{a}_1)\}^{1/5}$ if $\widehat{a}_1$ and $\widehat{a}_2$ are positive and $\widehat{h} = \mathrm{argmin}_{h_k} \mathrm{GCV}(h_k)$ otherwise.

Step 4: For $h = \widehat{h}$ selected in Step 3, repeat (a) and (b) in Step 2 until two successive values of $R(\boldsymbol{\beta})$ differ insignificantly.

Step 5: For $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ obtained from Step 4, apply the stepwise deletion technique in the previous section to choose significant variables at each given $\widehat{\boldsymbol{\beta}}^T \mathbf{X}_t = z$ for each fixed point z.

Here are some additional comments on the details of the implementation.

Remark 8.1 The standardization in Step 1 effectively rewrites the model (8.31) as

$$\sum_{j=0}^{p-1} g_j \left(\boldsymbol{\beta}^T \widehat{\Sigma}^{-1/2} (\mathbf{x} - \widehat{\mu}) \right) x_j,$$

where $\widehat{\mu}$ and $\widehat{\Sigma}$ are the sample mean and sample variance, respectively. In the numerical examples in §8.4.7, $\widehat{\Sigma}^{-1/2}\widehat{\boldsymbol{\beta}}/||\widehat{\Sigma}^{-1/2}\widehat{\boldsymbol{\beta}}||$ is reported as the estimated value of $\boldsymbol{\beta}$ defined in (8.31).

Remark 8.2 The weight function $w(z) = I(|z| \leq 2 + \delta)$ is chosen for some small $\delta \geq 0$. The functions $g_j(\cdot)$ in Steps 2 and 3 are estimated on 101 regular grids in the interval $[-1.5,\ 1.5]$ first, and then the values of the functions on this interval are estimated by the linear interpolation. This significantly reduces the computational time. In Step 4, however, we estimate $g_j(\cdot)$'s on the interval $[-2,\ 2]$.

Remark 8.3 The Epanechnikov kernel is employed in the numerical examples. To select the bandwidth $\widehat{h}$, one uses $q = 15$ and $h_k = 0.2 \times 1.2^{k-1}$ in Step 3. Recall that the data have been standardized in Step 1. The values of bandwidth practically cover the range of 0.2 to 2.57 times the standard deviation of the data.

Remark 8.4 In Step 2(b), the required derivatives are estimated based on their estimated function values at the grid points,

$$\widehat{\dot{g}}_j(z) = \{\widehat{g}_j(z_1) - \widehat{g}_j(z_2)\}/(z_1 - z_2), \quad j = 0, \ldots, p-1$$

and

$$\widehat{\ddot{g}}_j(z) = \{\widehat{g}_j(z_1) - 2\widehat{g}_j(z_2) + \widehat{g}_j(z_3)\}/(z_1 - z_2)^2, \quad j = 0, \ldots, p-1,$$

where $z_1 > z_2 > z_3$ are three nearest neighbors of z among the 101 regular grid points. To further stabilize the estimate of $\boldsymbol{\beta}$ in Step 2(b), the estimates of $g_j(\cdot)$ are smoothed further, using a simple moving-average technique: replace an estimate on a grid point by a weighted average on its five nearest neighbors with weights $\{1/2, 1/6, 1/6, 1/12, 1/12\}$. The edge points should be adjusted accordingly. To speed up the convergence, (8.37) is iterated a few times instead of just once.

8.4.7 Examples

In this section, the effectiveness of the proposed method is illustrated by two examples. For more examples, see Fan, Yao, and Cai (2002). There, the effectiveness of the local variable selection techniques is also demonstrated. In the algorithm, the ridge version of (8.37) is iterated two to four times to speed up the convergence. The search in Step 2 is stopped when either the two successive values of $R(\boldsymbol{\beta})$ differ less than 0.001 or the number of replications of (a) and (b) in Step 2 exceeds 30. The ridge parameter $q_n = 0.001\, n^{-1/2}$ is initially set and is kept doubling until the $\ddot{R}_r(\cdot)$ is no longer ill-conditioned with respect to the precision of computers.

Example 8.11 (*Simulations*) Consider a time series model

$$Y_t = -Y_{t-2}\exp(-Y_{t-2}^2/2) + \frac{1}{1+Y_{t-2}^2}\cos(1.5Y_{t-2})Y_{t-1} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ is a sequence of independent normal random variables with mean 0 and variance 0.25. The model is of form (8.29) with $p = 2$, $\boldsymbol{\beta} = (0, 1)$, and

$$g_0(z) = -z\exp(-z^2/2), \qquad g_1(z) = \cos(1.5z)/(1+z^2).$$

Two simulations were conducted with sample sizes 200 and 400, respectively, with 200 replications. The performance for nonparametric functions is assessed by the mean absolute deviation error at the 101 grid points:

$$\mathcal{E}_{\text{MAD}} = \frac{1}{101\, p}\sum_{j=0}^{p-1}\sum_{k=1}^{101}|\widehat{g}_j(z_k) - g_j(z_k)|.$$

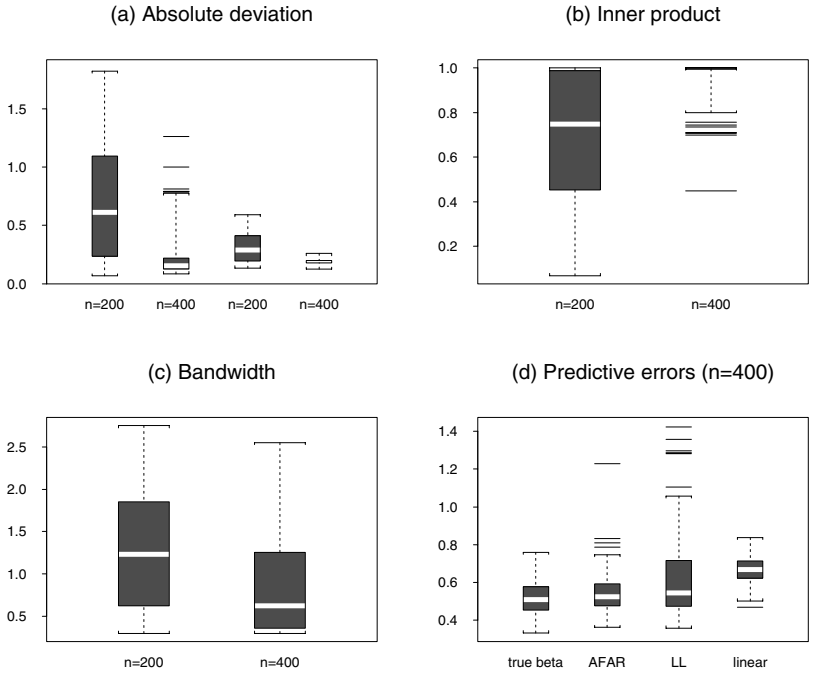


FIGURE 8.6. The boxplots of (a) the mean absolute deviation error $\mathcal{E}_{\text{MAD}}$ (the two panels on the left are based on $\widehat{\boldsymbol{\beta}}$, and the two panels on the right are based on the true $\boldsymbol{\beta}$), (b) the absolute inner product $|\boldsymbol{\beta}^T\widehat{\boldsymbol{\beta}}|$, (c) the selected bandwidths, and (d) the average absolute prediction errors of the varying-coefficient models with $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$, nonparametric model based on local linear regression, and linear AR-model determined by AIC (from left to right).

The performance of the estimated direction is measured by $|\boldsymbol{\beta}^T\widehat{\boldsymbol{\beta}}|$, which is the cosine of the angles between $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$. For each replication, 50 post-sample points are predicted and compared with the true observed values.

The results of the simulation are summarized in Figure 8.6. Figure 8.6(a) displays the boxplots of the mean absolute deviation errors. For sample size $n = 400$, the medians of $\mathcal{E}_{\text{MAD}}$ with estimated and true $\boldsymbol{\beta}$ are about the same, although the distribution of $\mathcal{E}_{\text{MAD}}$ with $\widehat{\boldsymbol{\beta}}$ has a long tail on the right. Figure 8.6(b) shows that the estimator $\widehat{\boldsymbol{\beta}}$ derived from the one-step iterative algorithm is close to the true $\boldsymbol{\beta}$ with high frequencies in the simulation replications. The average number of iterations in searching for $\boldsymbol{\beta}$ is 7.80 for $n = 400$ and 17.62 for $n = 200$. In fact, the search did not converge within 30 iterations for 21 out of 200 replications with $n = 200$ and for one out of 200 replications with $n = 400$. These replications produce outliers in Figure 8.6. Figure 8.6(c) shows the boxplots of the selected bandwidths.

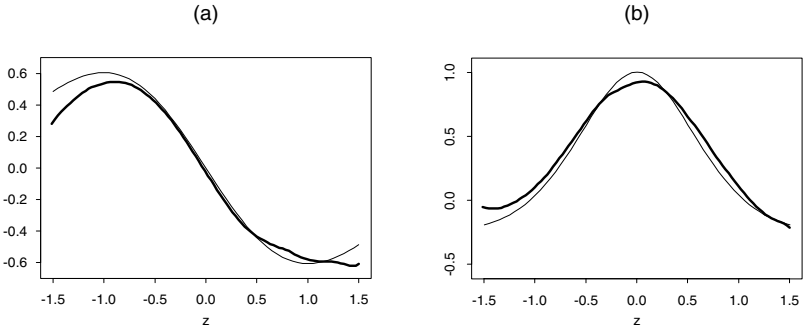


FIGURE 8.7. The plot of estimated coefficient functions (thick line) and true functions (thin line). (a) $g_0(z) = -ze^{-z^2/2}$; (b) $g_1(z) = \cos(1.5z)/(1+z^2)$. The sample size $n = 400$.

The prediction performance of various models is also compared in the simulation with the sample size $n = 400$. For each of 200 realizations, 50 postsample points are predicted from four different models, namely the fitted varying-coefficient models with true and estimated $\boldsymbol{\beta}$, a purely nonparametric model based on local linear regression of Y_t on (Y_{t-1}, Y_{t-2}) with the bandwidth selected by the GCV-criterion, and a linear autoregressive model with the order (≥ 2) determined by AIC. In our simulation, AIC always selected order 2 in the 200 replications. Figure 8.6(d) presents the distributions of the average absolute prediction errors across 200 replications. The AFAR models with true and estimated $\boldsymbol{\beta}$ are the two best predictors since they specify correctly the form of the true model. The median of the prediction errors from the two-dimensional nonparametric model based on local linear regression is about the same as that from the AFAR model, but the variation is much larger. This is largely due to the fact that the full nonparametric model overfits the problem. The linear autoregressive model performs poorly in this example due to the modeling biases.

A typical example of the estimated coefficient functions is depicted in Figure 8.7 with the sample size $n = 400$. The typical example was selected in such a way that the corresponding $\mathcal{E}_{\text{MAD}}$ is equal to its median among the 200 replications in simulation. The curves are plotted on the range from -1.5 to 1.5 times the standard deviation of $\boldsymbol{\beta}^T\mathbf{X}$. For the case with $n = 400$, the selected bandwidth is 0.781, and $\boldsymbol{\beta}^T\widehat{\boldsymbol{\beta}} = 0.999$. (The median of $\boldsymbol{\beta}^T\widehat{\boldsymbol{\beta}}$ in the simulation of 200 replications with $n = 400$ is 0.999.) ∎

Example 8.12 (*Exchange rate data*) This example deals with the daily closing bid prices of the British pound sterling in terms of the U.S. dollar from January 2,1974 to December 30, 1983, consisting of a time series of length 2510. The previous analysis of this "particularly difficult" data set can be found in Gallant, Hsieh and Tauchen (1991) and the ref-

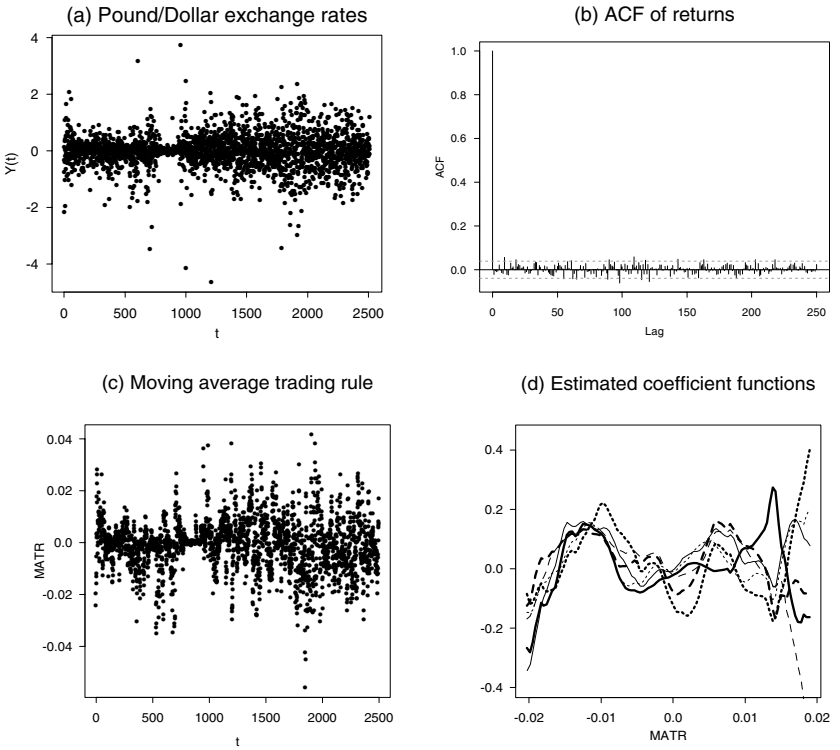


FIGURE 8.8. (a) The plot of pound/dollar exchange rate return series $\{Y_t\}$. (b) The autoregressive function of $\{Y_t\}$. (c) The plot of $\{U_t = 10Y_t/\sum_{i=0}^{9} Y_{t-i} - 1\}$. (d) The estimated coefficient functions of model (8.38) with $Z_t = U_{t-1}$ and $m = 5$. Thick solid lines are $g_0(\cdot)$, thick dotted lines $g_1(\cdot)$, thick dashed lines $g_2(\cdot)$, solid lines $g_3(\cdot)$, dotted lines $g_4(\cdot)$, and dashed lines $g_5(\cdot)$. From Fan, Yao, and Cai (2002).

erences therein. Let X_t be the exchange rate on the tth day and $\{Y_t = 100\log(X_t/X_{t-1})\}$ be the return series. It is well-known that the classical financial theory assumes that time series $\{Y_t\}$ is typically a martingale difference process and that Y_t is unpredictable. Figures 8.8 (a) and (b) show that there exists almost no significant autocorrelation in the series $\{Y_t\}$.

Let

$$U_{t-1} = X_{t-1}\left\{L^{-1}\sum_{j=1}^{L} X_{t-j}\right\}^{-1} - 1$$

be the *moving average technical trading rule* (MATR). Then, $U_{t-1}+1$ is the ratio of the exchange rate at the time $t-1$ to the average rate over a past period of length L. $U_{t-1} > 0$ signals the upward momentum (the position to buy sterling) and $U_{t-1} < 0$ indicates the downward pressure (the position

to sell sterling). For a detailed discussion of the MATR, see the papers by LeBaron (1997, 1999) and Hong and Lee (2002). The performance of this technical trading rule will be evaluated by the *mean trading return* for the postsample forecast (the first 2,400 data points will be used to estimate coefficients in the AFAR model), which is defined as

$$\mathrm{MTR}_{\mathrm{MA}} = \frac{1}{100}\sum_{t=1}^{100}\{I(U_{2410+t-1} > 0) - I(U_{2410+t-1} < 0)\}Y_{2410+t}.$$

The MTR measures the real profits in a financial market, ignoring interest differentials and transaction costs (for the sake of simplicity). According to Hong and Lee (2002), it is more relevant than the conventional mean squared prediction errors or average absolute prediction errors for evaluating the performance of forecasting market movements. Ideally, we would buy sterling at time $t-1$ when $Y_t > 0$ and sell when $Y_t < 0$. (This is not really a trading rule.) The mean trading return for this "ideal" strategy is

$$\mathrm{MTR}_{\mathrm{ideal}} = \frac{1}{100}\sum_{t=1}^{100}|Y_{2410+t}|,$$

which serves as a benchmark for assessing other forecasting procedures. For this particular data set, $\mathrm{MTR}_{\mathrm{MA}}/\mathrm{MTR}_{\mathrm{ideal}} = 12.58\%$ if we let $L = 10$.

To utilize the AFAR technique, we approximate the conditional expectation of Y_t (given its past) by

$$g_0(Z_t) + \sum_{i=1}^{m} g_i(Z_t)Y_{t-i}, \tag{8.38}$$

where

$$Z_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 X_{t-1} + \beta_4 U_{t-1}.$$

Let $\widehat{Y}_t$ be defined as the predicted value by using (8.38) when g's are estimated. The mean trading return for the forecasting based on the AFAR model is defined as

$$\mathrm{MTR}_{\mathrm{AFAR}} = \frac{1}{100}\sum_{t=1}^{100}\{I(\widehat{Y}_{2410+t} > 0) - I(\widehat{Y}_{2410+t} < 0)\}Y_{2410+t}.$$

As the first attempt, we let $m = 5$ and $L = 10$ in (8.38) (i.e., we use the past week's data as the "regressors" in the model and define the MATR as the average rate in the last two weeks). The selected $\boldsymbol{\beta}$ is

$$(0.0068, 0.0077, 0.0198, 0.9998)^T,$$

which suggests that U_t plays an important role in the underlying nonlinear dynamics. The ratio of the RSS of the fitted model to the sample variance of

$\{Y_t\}$ is 93.67%, reflecting the presence of high-level "noise" in the financial data. The selected bandwidth is 0.24. The ratio $\text{MTR}_{\text{AFAR}}/\text{MTR}_{\text{ideal}} = 5.53\%$. The predictability is much lower than that of the MATR. If we include rates in the past two weeks as regressors in the model (i.e., $m = 10$ in (8.38)), the ratio $\text{MTR}_{\text{AFAR}}/\text{MTR}_{\text{ideal}}$ increases to 7.26%, which is still a distance away from $\text{MTR}_{\text{MA}}/\text{MTR}_{\text{ideal}}$, while the ratio of the RSS of the fitted model to the sample variance of $\{Y_t\}$ is 87.96%. The selected bandwidth is still 0.24, and $\widehat{\boldsymbol{\beta}} = (0.0020, 0.0052, 0.0129, 0.9999)^T$.

The calculations above (also others not reported here) seem to suggest that U_t could be a dominated component in the selected index. This leads to use of the model (8.38) with prescribed $Z_t = U_{t-1}$, which is actually the model adopted by Hong and Lee (2002). For $m = 5$, the fitting to the data used in estimation becomes worse; the ratio of the RSS of the fitted model to the sample variance of $\{Y_t\}$ is 97.39%. But it provides a better postsample forecast in terms of MTR: $\text{MTR}_{\text{AFAR}}/\text{MTR}_{\text{ideal}}$ is 23.76%. The selected bandwidth is 0.24. The plots of estimated coefficient functions indicate a possible undersmoothing. By increasing the bandwidth to 0.40, $\text{MTR}_{\text{AFAR}}/\text{MTR}_{\text{ideal}}$ is 31.35%. The estimated coefficient functions are plotted in Figure 8.8(d). The rate of correct predictions for the direction of market movement (sign of Y_t) is 50% for the MATR and 53% and 58% for the AFAR model with bandwidths 0.24 and 0.40, respectively.

One should not take for granted that U_t is always a good index for forecasting. Hong and Lee (2002) conducted empirical studies with several financial data sets with only partial success from using the FAR modeling techniques with U_t as the prescribed index. In fact, for this particular data set, model (8.38) with $Z_t = U_t$ and $m = 10$ gives a negative value of MTR_{AFAR}. Note that the "superdominating" position of U_t in the selected smoothing variable $\widehat{\boldsymbol{\beta}}^T \mathbf{X}_t$ is partially due to the scaling difference between U_t and (Y_t, X_t); see also Figures 8.8(a) and (c). In fact, if we standardize U_t, Y_t, and X_t separately beforehand, the resulting $\widehat{\boldsymbol{\beta}}$ is $(0.59, -0.52, 0.07, 0.62)^T$ when $m = 5$, which is dominated by U_{t-1} and the contrast between Y_{t-1} and Y_{t-2}. ($\text{MTR}_{\text{AFAR}}/\text{MTR}_{\text{ideal}} = 1.42\%$. The ratio of the *residual sum of squares* (RSS) of the fitted model to the sample variance of Y_t is 96.90%.) By doing this, we effectively use a different class of models to approximate the unknown conditional expectation of Y_t; see Remark 8.1. Finally, we remark that a different modeling approach should be adopted if our primary target is to maximize the mean trading return rather than minimize the prediction errors. ∎

8.4.8 Extensions

A further extension of the AFAR model is to allow a *multipleindex* in the model (8.29), leading to

$$X_t = g_0(\boldsymbol{\beta}_1^T \mathbf{X}_{t-1}, \cdots, \boldsymbol{\beta}_d^T \mathbf{X}_{t-1}) + \sum_{j=1}^{p-1} g_j(\boldsymbol{\beta}_1^T \mathbf{X}_{t-1}, \cdots, \boldsymbol{\beta}_d^T \mathbf{X}_{t-1}) X_{t-j} + \varepsilon_t. \tag{8.39}$$

where $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d$ are unknown parameters. This is a specific case of the *multi-index model* frequently studied in the literature for independent samples. The index parameters $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d$ are not identifiable, but the linear space spanned by these index parameters is. For example, the function $g(x_1 - x_4, x_1 + x_4)$ can be written as $g^*(x_1, x_4)$. But the linear space spanned by the vectors $\boldsymbol{\beta}_1 = (1, 0, 0, -1)^T$ and $\boldsymbol{\beta}_2 = (1, 0, 0, 1)^T$ is the same as that spanned by $\boldsymbol{\beta}_1 = (1, 0, 0, 0)^T$ and $\boldsymbol{\beta}_2 = (0, 0, 0, 1)^T$. Popular methods for estimating the linear span include the *sliced inverse regression* (SIR) (Duan and Li, 1991; Li, 1991), principal Hessian directions (Li, K.C. 1992; Cook 1998), the average derivative method (Härdle and Stoker 1989; Samarov 1993, Hristache, Juditsky, Polzehl, and Spokoiny 2002), and other techniques (Cook 1996; Chiaromonte, Cook and Li, 2002; Cook and Li 2002).

The general model (8.39) is unlikely to be useful when $d > 2$. Owing to the curse of dimensionality, the model-coefficient functions $\{g_j\}$ cannot be estimated well when $d > 2$. Thus, for practical purposes, we consider only the model (8.39) with two indices. The parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ can be estimated by one of the methods described in the last paragraph. However, the resulting estimators are not necessarily efficient. The profile likelihood method can be used to improve the efficiency for estimating these parameters. In fact, the profile least-squares method can be applied readily to the current setting. Fan, Yao, and Cai (2002) give an implementation for the two-index AFAR model.

8.5 Additive Models

Additive models (Ezekiel 1924) are very useful for approximating the high-dimensional autoregressive function $G(\cdot)$ given in (8.27). They and their extensions have become one of the widely used nonparametric techniques because of the exemplary monograph by Hastie and Tibshirani (1990) and companion software as described in Chambers and Hastie (1991).

8.5.1 The Models

Direct estimation of the autoregressive function G without imposing restrictions faces the challenge of the curse of dimensionality, as mentioned

in §8.1. A useful class of models are the additive models

$$g(x_{t-1}, \cdots, x_{t-p}) = f_1(x_{t-1}) + \cdots + f_p(x_{t-p}). \tag{8.40}$$

The functions $f_1, \cdots, f_p$ are univariate and can be estimated as well as the one-dimensional nonparametric regression problem (Stone 1985, 1986; Fan, Härdle, and Mammen 1998). Hence, the curse of dimensionality is avoided.

Restricting ourselves to the class of additive models (8.40), the prediction error can be written as

$$\begin{aligned} E\{X_t - g(X_{t-1}, \cdots, X_{t-p})\}^2 &= E\{X_t - G(X_{t-1}, \cdots, X_{t-p})\}^2 \\ &+ E\{G(X_{t-1}, \cdots, X_{t-p}) - g(X_{t-1}, \cdots, X_{t-p})\}^2, \end{aligned} \tag{8.41}$$

where $G(X_{t-1}, \cdots, X_{t-p}) = E(X_t | X_{t-1}, \cdots, X_{t-p})$. Thus, finding the best additive model to minimize the prediction error is equivalent to finding the one that best approximates the autoregressive function G in the sense that g minimizes the third term of (8.41).

When the autoregressive function G admits the additive structure, namely, $G = g$, we have

$$X_t = f_1(X_{t-1}) + \cdots + f_p(X_{t-p}) + \varepsilon_t. \tag{8.42}$$

Denote it by $\{X_t\} \sim \text{AAR}(p)$ (see also (1.12)). This model allows us to examine the extent of the nonlinear contribution of each lagged variable to the variable X_t. In particular, it includes the AR(p) model as its specific case. This allows us to test whether an AR(p) model holds reasonably for a given time series. Details for this will be given in Chapter 9. When the model (8.42) fails, the functions $f_1, \cdots, f_p$ define a best additive approximation to the true autoregression function G in the sense that they minimize the third term of (8.41).

8.5.2 The Backfitting Algorithm

The estimator $f_1, \cdots, f_p$ can easily be estimated by using the *backfitting algorithm*. First, note that we can add a constant to a component and subtract the constant from another component. Thus, the functions $f_1, \cdots, f_p$ are not identifiable. To prevent ambiguity, the following identifiability conditions are frequently imposed:

$$Ef_j(X) = 0, \quad j = 1, \cdots, p. \tag{8.43}$$

With these constraints, $EX_t = 0$ by (8.42). Therefore, as in many other settings, we assume that the mean has already been removed from the series $\{X_t\}$ or add an intercept term μ to the model, resulting in

$$X_t = \mu + f_1(X_{t-1}) + \cdots + f_p(X_{t-p}) + \varepsilon_t,$$

with $\mu = EX_t$.

To highlight the key idea of backfitting, we first consider a spline approximation as in (1.17): $f_j(x) \approx f_j(x, \mathbf{b}_j)$. Our task is then to find the parameters $\mathbf{b}_1, \cdots, \mathbf{b}_p$ to minimize the prediction errors:

$$\sum_{t=p+1}^{T} \{X_t - f_1(X_{t-1}, \mathbf{b}_1) - \cdots - f_p(X_{t-p}, \mathbf{b}_p)\}^2. \tag{8.44}$$

This least-squares problem can be solved directly, resulting in a large parametric problem with an inversion of matrix of high order. Alternatively, the optimization problem can be solved using the following iterative scheme. Given the initial values of $\mathbf{b}_2, \cdots, \mathbf{b}_p$, minimize (8.44) with respect to $\mathbf{b}_1$. This is a much smaller "parametric" problem and can be solved relatively easily. With estimated $\mathbf{b}_1$ and values $\mathbf{b}_3, \cdots, \mathbf{b}_p$, we now minimize (8.44) with respect to $\mathbf{b}_2$. This results in an updated estimate of $\mathbf{b}_2$, and so on. When $\mathbf{b}_p$ is updated, we can now update the parameter $\mathbf{b}_1$ again and then $\mathbf{b}_2$, and so on. This algorithm can be run until some convergence criterion is met. This is the basic idea of the backfitting algorithm (Ezekiel 1924; Breiman and Friedman 1985; Buja, Hastie, and Tibshirani 1989).

Let $\widehat{\varepsilon}_{t,k} = X_t - \sum_{j \neq k} f_j(X_{t-j}, \widehat{\mathbf{b}}_j)$ be the *partial residuals* without using the lag variable X_{t-k}. Then, the backfitting algorithm finds $\mathbf{b}_k$ by minimizing

$$\sum_{t=p+1}^{T} \{\widehat{\varepsilon}_{t,k} - f_k(X_{t-k}, \mathbf{b}_k)\}^2.$$

This is a nonparametric regression problem of $\{\widehat{\varepsilon}_{t,k}\}$ on the variable $\{X_{t-k}\}$ using a polynomial spline method. The resulting estimate is linear in the partial residuals $\{\widehat{\varepsilon}_{t,k}\}$. It can be written as

$$\begin{pmatrix} f_k(X_{p-k+1}, \widehat{b}_k) \\ f_k(X_{p-k+2}, \widehat{b}_k) \\ \vdots \\ f_k(X_{T-k}, \widehat{b}_k) \end{pmatrix} = \mathbf{S}_k \begin{pmatrix} \widehat{\varepsilon}_{p+1,k} \\ \widehat{\varepsilon}_{p+2,k} \\ \vdots \\ \widehat{\varepsilon}_{T,k} \end{pmatrix}. \tag{8.45}$$

The matrix $\mathbf{S}_k$ is called a *smoothing matrix*. To facilitate the notation, we denote the left-hand side of (8.45) as $\widehat{\mathbf{f}}_k$ and write $\mathbf{X} = (X_{p+1}, \cdots, X_T)^T$. Then, (8.45) can be written schematically as

$$\widehat{\mathbf{f}}_k = \mathbf{S}_k \left(\mathbf{X} - \sum_{j \neq k} \widehat{\mathbf{f}}_j \right).$$

The example above utilizes polynomial splines as a nonparametric smoother. The idea can be applied to any nonparametric smoother mentioned in

Chapter 6. Let $\mathbf{S}_k$ be a smoothing matrix that regresses nonparametrically the partial residuals $\{\widehat{\varepsilon}_{t,k}\}$ on the lagged variable $\{X_{t-k}\}$; namely, (8.45) is obtained through the smoother $\mathbf{S}_k$. We now outline the backfitting algorithm with a generic nonparametric smoother $\mathbf{S}_k$.

The backfitting algorithm estimates the functions $f_1, \cdots, f_p$ through the following iterations:

(i) Initialize the functions $\widehat{f}_1, \cdots, \widehat{f}_p$.

(ii) Cycle $k = 1, \cdots, p$, and compute $\widehat{\mathbf{f}}_k^* = \mathbf{S}_k(\mathbf{X} - \sum_{j \neq k} \widehat{\mathbf{f}}_j)$ and center the estimator to obtain

$$\widehat{\mathbf{f}}_k(\cdot) = \widehat{\mathbf{f}}_k^*(\cdot) - (T-p)^{-1} \sum_{t=p+1}^{T} \widehat{\mathbf{f}}_k^*(X_{t-k}).$$

(iii) Repeat (ii) until convergence.

(See, for example, Hastie and Tibshirani 1990, p. 91.) The recentering in step (ii) is to comply with the constraint (8.43). The convergence issue of the algorithm is delicate and has been addressed via the concept of *concurvity* by Buja, Hastie, and Tibshirani (1989). Assuming that concurvity is not present, it is shown there that the backfitting algorithm converges and solves the following equation:

$$\begin{pmatrix} \widehat{\mathbf{f}}_1 \\ \widehat{\mathbf{f}}_2 \\ \vdots \\ \widehat{\mathbf{f}}_p \end{pmatrix} = \begin{pmatrix} I & \mathbf{S}_1 & \cdots & \mathbf{S}_1 \\ \mathbf{S}_2 & I & \cdots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p & \mathbf{S}_p & \cdots & I \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_p \end{pmatrix} \begin{pmatrix} X_{p+1} \\ X_{p+2} \\ \vdots \\ X_T \end{pmatrix}. \tag{8.46}$$

Direct calculation of the right-hand side of (8.46) involves inverting a square matrix of order $O(pT)$ and can hardly be implemented on an average computer for time series with moderate length. In contrast, the backfitting does not share this drawback and is frequently used in practical implementations.

The software for multivariate additive models for cross-sectional data can be used directly here by creating the dependent and independent variables as in (8.3). The function " gam()" in S can easily be used for the computation here. Systematic treatments on additive models and their useful extensions can be found in the monograph by Hastie and Tibshirani (1990).

8.5.3 Projections and Average Surface Estimators

The idea of *projections,* introduced by Tjøstheim and Auestad (1994a, b), provides an explicit estimator for the additive components in model (8.40). Assume that the additive model holds; namely $G(\cdot) = g(\cdot)$. Then

$$EG(\mathbf{X}_{t-1,k})W(\mathbf{X}_{t-1,k}) = f_k(x) + c, \tag{8.47}$$

where c is a constant, and

$$\mathbf{X}_{t-1,k} = (X_{t-1}, \cdots, X_{t-k+1}, x, X_{t-k-1}, \cdots, X_{t-p})^T.$$

This observation provides a simple and direct estimator: estimate the high-dimensional regression surface $\widehat{G}$ directly by using, for example, the kernel or the local polynomial estimator in §8.2, and then average over the variable $\mathbf{X}_{t-1,k}$. Let

$$\widehat{f}_k^*(x) = (T-p)^{-1} \sum_{t=p+1}^{T} \widehat{G}(\mathbf{X}_{t-1,k}) W(\mathbf{X}_{t-1,k}) \tag{8.48}$$

be the *average regression surface*. Then, it estimates $f_k(x) + c$, according to (8.47). The function $\widehat{f}_k$ can be obtained by centering $\widehat{f}_k^*(\cdot)$ as in the backfitting algorithm. This provides an estimator for f_k.

Note that the average operation in (8.48) significantly reduces the variance inherited from $\widehat{G}$. However, it usually does not attenuate the biases in $\widehat{G}$. Thus, a general strategy is to use a small bandwidth to obtain an undersmoothed G first, which has small biases, and then to apply the average surface estimator (8.48). However, the bandwidth h cannot be too small since, for p-dimensional cubes of size $O(h^p)$, they are likely to contain no data points. Therefore, the procedure has implementation difficulties when the number of lagged variables p is large. In their simulated examples, Tjøstheim and Auestad (1994) tested the procedure for $p = 8, 12$ with time series of length $T = 500$. The problem of the sparsity of local data in high-dimensional space is attenuated somewhat by using the Gaussian kernel, which has unbounded support.

The implementation problem above can be significantly attenuated when one uses the following projection idea. Let us for a moment focus on the population version: minimize

$$E\{G(X_{t-1}, \cdots, X_{t-p}) - f_1(X_{t-1}) - \cdots - f_p(X_{t-p})\}^2$$

subject to the constraints (8.43). Let $\tilde{f}_1, \cdots, \tilde{f}_p$ be the solution. Then, by the method of variation, for any $g_1(X_{t-1})$ with $Eg_1(X_{t-1}) = 0$ and any parameter θ, the minimum of

$$E\{G(X_{t-1}, \cdots, X_{t-p}) - \tilde{f}_1(X_{t-1}) - \cdots - \tilde{f}_p(X_{t-p}) - \theta g_1(X_{t-1})\}^2$$

is attained at $\theta = 0$. Taking the derivative with respect to θ and setting it to zero at $\theta = 0$, we have

$$E\{G(X_{t-1}, \cdots, X_{t-p}) - \tilde{f}_1(X_{t-1}) - \cdots - \tilde{f}_p(X_{t-p})\} g(X_{t-1}) = 0.$$

Since it holds for all univariate functions g, the equation above implies necessarily that

$$\begin{aligned} E[\{G(x_1, X_{t-2}, \cdots, X_{t-p}) - \tilde{f}_1(x_1) - \tilde{f}_2(X_{t-2}) - \cdots \\ -\tilde{f}(X_{t-p})\} | X_{t-1} = x_1] = 0. \end{aligned} \tag{8.49}$$

Note that

$$E\{G(x_1, X_{t-2}, \cdots, X_{t-p})|X_{t-1} = x_1\} = E(X_t|X_{t-1} = x_1).$$

Using this and (8.49), we have

$$\begin{aligned}\tilde{f}_1(x_1) &= E(X_t|X_{t-1} = x_1) - \sum_{k=2}^{p} E\Big\{\tilde{f}_2(X_{t-k})|X_{t-1} = x_1\Big\} \\ &= E(X_t|X_{t-1} = x_1) - \sum_{k=2}^{p} \int \tilde{f}_k(x_k) p_{|1-k|}(x_1, x_k)/p(x_1) dx_k\end{aligned}$$

where $p(x)$ is the density of X_{t-1} and $p_{|j-k|}(x_j, x_k)$ is the joint density of (X_{t-j}, X_{t-k}). By applying the method of variation to the other variables, we obtain a system of equations ($j = 1, \cdots, p$):

$$\tilde{f}_j(x_j) = E(X_t|X_{t-j} = x_j) - \sum_{k\neq j} \int \tilde{f}_k(x_k) p_{|j-k|}(x_j, x_k)/p(x_j) dx_k. \quad (8.50)$$

For given functions $E(X_t|X_{t-j} = \cdot)$, $p_{|j-k|}(\cdot, \cdot)$, and $p(\cdot)$, the backfitting algorithm can be used to solve the system of equations above.

The essence of the method of Mammen, Linton, and Nielsen (1999) is to use the backfitting algorithm to solve the empirical version of (8.50), although their motivations are somewhat different from ours. All unknown functions in (8.50) are at most two-dimensional, and hence the curse-of-dimensionality in the implementation is avoided; see also Kim, Linton, and Hengartner (1999).

8.5.4 Estimability of Coefficient Functions

The asymptotic theory on additive models is relatively less-well-developed. Most theory is established for an i.i.d. setting. As explained in §5.3, due to whitening by localization, the results are expected to hold for stationary time series under certain mixing conditions.

The additive components can be estimated as well as the one-dimensional nonparametric problem in terms of the convergence rate (Stone 1985, 1986). In fact, Fan, Härdle, and Mammen (1998) showed further that each additive component can be estimated as well as if other components were known in terms of asymptotic biases and asymptotic variances. Their procedures are based on the projection estimator (8.48) with a suitable choice of weight function W. In other words, in estimating the function f_j, not knowing the other components $\{f_i, i \neq j\}$ doesnot add appreciable extra difficulty. Such a property is frequently called an *oracle property* in the literature. Suppose that there is an oracle who knows the functions $\{f_i, i \neq j\}$. He or she would use the knowledge to estimate f_j. Statisticians who have assistance from the

oracle can construct an estimator for f_j that performs as well as the oracle. Similar oracle properties were obtained by Linton (1997) and Kim, Linton, and Hengartner (1999). Their idea was to use the projection estimator as the initial value in the backfitting algorithm and then to apply univariate smoothers on the partial residuals to improve the efficiency. Mammen, Linton, and Nielsen (1999) modified the backfitting algorithm to obtain efficient estimators for additive components and established the asymptotic normality of estimators. The intuition behind this surprising oracle phenomenon is that the local parameters $\{f_j(x), j = 1, \cdots, p\}$ are asymptotically orthogonal. This can be intuitively understood as follows. For any given point $\mathbf{x}$ in the p-dimensional space, owing to the continuity of the density of $X_{t-1}, \cdots, X_{t-p}$, the joint density is nearly flat, (i.e., nearly uniform) in a small hypercube around $\mathbf{x}$; that is, the variables $X_{t-1}, \cdots, X_{t-p}$ are locally independent, and hence the local parameters $\{f_j(x), j = 1, \cdots, p\}$ are orthogonal. The situation is very much like that in the linear model with orthogonal design matrices. In such a case, knowing the part of parameters does not provide any extra information for the other part of parameters due to the orthogonality. In summary, estimating components in the additive models is not appreciably more difficult, in terms of statistical efficiency, than for the one-dimensional problem.

8.5.5 Bandwidth Selection

Like all nonparametric problems, estimating additive components involves the choice of smoothing parameters. For time series applications, since all lagged variables are in the same order of magnitude, it is reasonable and simple to just use one bandwidth h. For the independent data, Opsomer and Ruppert (1997) proposed a plug-in method and obtained some nice convergence results for the selected bandwidth. The procedure is also applicable to the current problem. However, it is quite delicate.

Here, we outline a simpler but not necessarily more effective method than in Opsomer and Ruppert (1997). The idea is related to the multifold cross-validation criterion for stationary time series in §8.3.5. We adopt the notation from that section. To fix the idea, in the following discussion, we employ the local linear estimator with bandwidth h to construct the smoothing matrix $\mathbf{S}_k$ for the lag variable X_{t-k}.

As in §8.3.5, let $\{\widehat{f}_{j,q}(\cdot)\}$ be the estimated additive functions using the qth $(q = 1, \cdots, Q)$ subseries $\{X_t,\ 1 \le t \le n - qm\}$ with bandwidth equal to $h\{n/(n - qm)\}^{1/5}$. The bandwidth h is rescaled slightly to accommodate different sample sizes according to its optimal rate (i.e., $h \propto T^{-1/5}$), but this can also be omitted without appreciably affecting the result. The

average prediction error using the qth subseries is given by

$$\mathrm{APE}_q(h) = \frac{1}{m} \sum_{t=n-qm+1}^{n-qm+m} \left\{ X_t - \sum_{j=1}^{p} \widehat{f}_{j,q}(X_{t-j}) \right\}^2 .$$

The overall average prediction error is given by

$$\mathrm{APE}(h) = Q^{-1} \sum_{q=1}^{Q} \mathrm{APE}_q(h). \tag{8.51}$$

The proposed data-driven bandwidth is the one that minimizes $\mathrm{APE}(h)$. In practical implementations, as in §8.3.5, we may use $m = [0.1n]$ and $Q = 4$. The selected bandwidth is not expected to depend appreciably on the choice of m and Q as long as mQ is reasonably large so that the evaluation of prediction errors is stable. The function $\mathrm{APE}(h)$ is minimized by comparing its value at a grid of points $h_j = a^j h_0 (j = 1, \cdots, J)$. For example, one may choose $a = 1.2$, $J = 15$ or 20, and $h_0 = 1.2^{-J}$(range of X).

8.5.6 Examples

In this section, we illustrate the procedure using one simulated data set and one real data set from the Standard and Poor's 500 Index described in Example 1.4. The former allows us to examine the performance of the backfitting algorithm and the proposed bandwidth selection rule (8.51), while the latter permits us to see how well the multiperiod forward volatility can be predicted from the observed one-day and multiple-day volatilities.

Example 8.13 (*Simulated data*) A series of length 400 was simulated from the AAR(2) model with

$$\begin{aligned} f_1(X_{t-1}) &= 4X_{t-1}/(1 + 0.8X_{t-1}^2) \\ f_2(X_{t-2}) &= \exp\{3(X_{t-2} - 2)\}/[1 + \exp\{3(X_{t-2} - 2)\}] \end{aligned}$$

and $\varepsilon_t \sim \mathrm{Uniform}(-1, 1)$. The functions f_1 and f_2 are depicted in Figure 8.10 (solid curves). The resulting series is shown in Figure 8.9(a). In Figures 8.9 (b) and (c), the scatterplots for the lag 1 and lag 2 series are displayed. These two figures would easily mislead us to an AR(2) model since the trends are linear. In fact, visualization of the scatterplot of the variable X_{t-1} against the variable X_t amounts to smoothing these two variables visually; namely, estimating visually

$$E(X_t|X_{t-1}) = f_1(X_{t-1}) + E(f_2(X_{t-2})|X_{t-1}). \tag{8.52}$$

It is clear that the right hand side of (8.52) is very different from $f_1(X_{t-1})$ unless X_{t-1} and X_{t-2} are nearly independent. This indicates that the scatterplots are not very informative for visualizing the additive component functions $f_1(\cdot)$ and $f_2(\cdot)$.

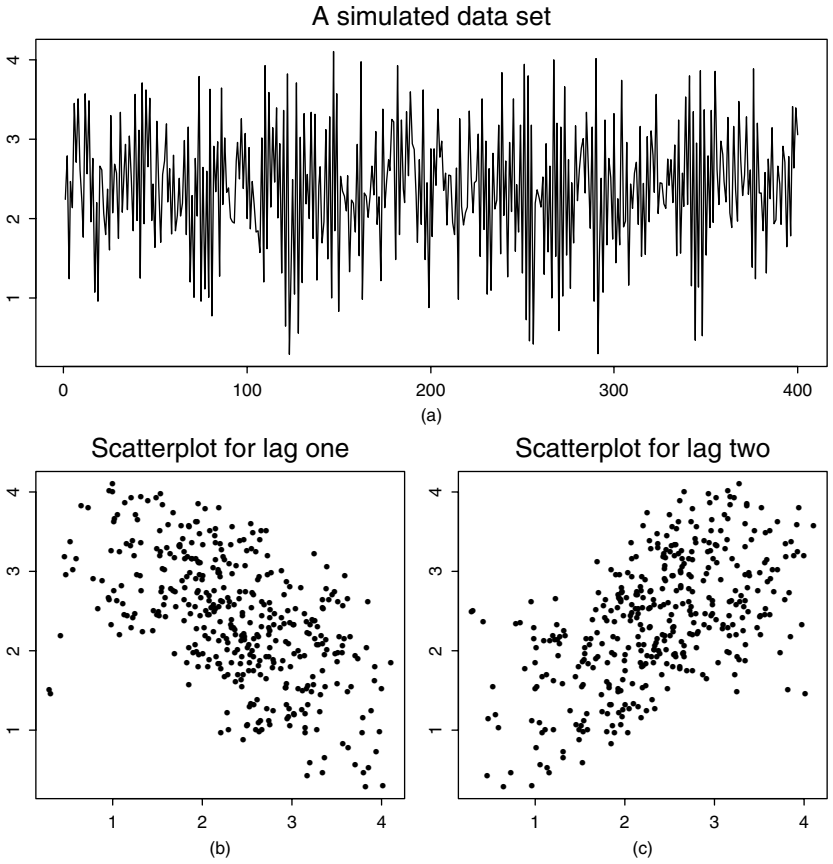


FIGURE 8.9. (a) Time series plot of simulated data from an AAR(2) model. (b) Plot of X_{t-1} against X_t. (c) Plot of X_{t-2} against X_t.

We now apply the backfitting algorithm, with the local linear smoother, to find the estimates for the additive components. The smoothing parameter is selected by (8.51). As noted before, the additive components can only be identifiable within a constant. Thus, the decentered version of function f_1 (i.e., $f_1(X_{t-1}) - \bar{f}_1$) is plotted against X_{t-1} in Figure 8.10(a), where

$$\bar{f}_1 = \frac{1}{398} \sum_{t=3}^{400} f_1(X_{t-1}).$$

A similar remark can be made for the function f_2. The resulting plots are displayed in Figure 8.10. The relatively poor performance for $\widehat{f}_1$ at the left boundary is mainly due to the fact that there are not many data points in that region. Overall, the performance is quite satisfactory. ∎

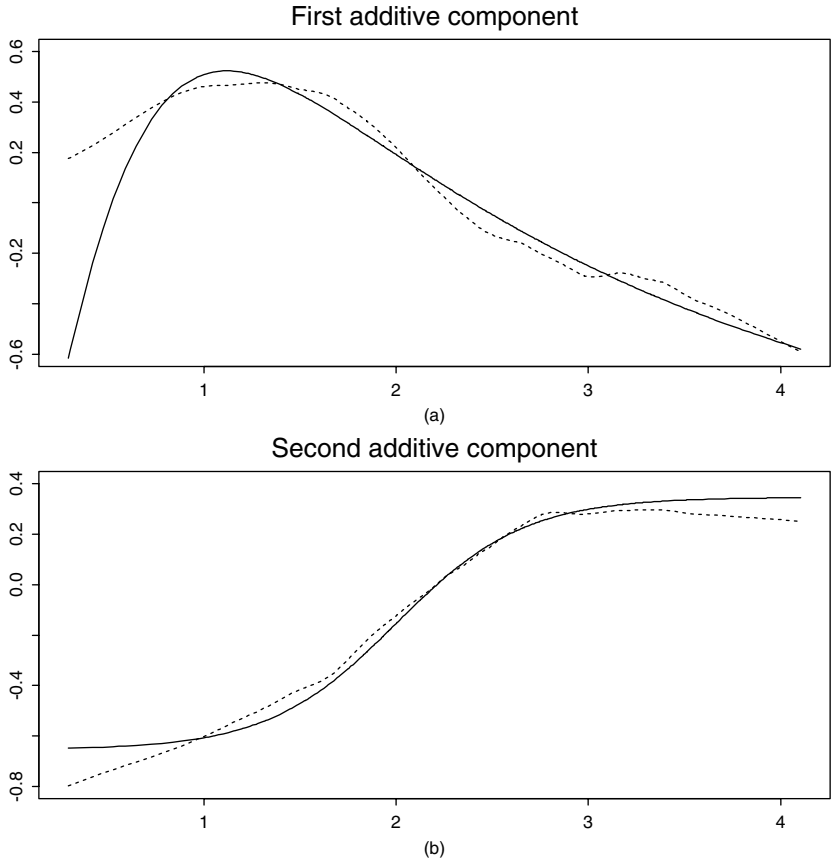


FIGURE 8.10. Estimated additive functions for (a) $\widehat{f}_1$ and (b) $\widehat{f}_2$. The bandwidth was selected by using (8.51).

The additive model can apply not only to AAR(p) with its lag variables but also to other stochastic regression models. We illustrate this point in the next example.

Example 8.14 (*Value at risk*) In response to the infamous financial catastrophes of the 1990s that engulfed companies such as Barings and Daiwa as well as Orange County, CA, and Asian countries, risk management has become important for financial institutions, regulators, nonfinancial corporations, and asset managers. *Value at risk* (VaR) is a fundamental tool for measuring market risks. It measures the worst loss to be expected of a portfolio over a given time horizon under normal market conditions at a given confidence level. VaR has been popularly used to control and manage various risks, including credit risk, market risk, and operational risk. Jorion (2000) provides an informative introduction to the subject.

Let S_t be the price of a portfolio at time t. Let

$$r_t = \log(S_t/S_{t-1}) \approx \frac{S_t - S_{t-1}}{S_{t-1}}$$

be the observed return at time t. The aggregate return at time t for a predetermined holding period τ is

$$R_{t,\tau} = \log(S_{t+\tau-1}/S_{t-1}) = r_t + \cdots + r_{t+\tau-1}. \tag{8.53}$$

The VaR measures the extreme loss $V_{t+1,\tau}$, in terms of percentage, of the portfolio over a predetermined holding period τ with a prescribed confidence level $1-\alpha$, namely

$$P(R_{t+1,\tau} > V_{t+1,\tau}|\Omega_t) = 1-\alpha,$$

where Ω_t is the historical information-namely, the σ-field generated by $S_t, S_{t-1}, \cdots$.

An important contribution to the forecast of VaR is the *RiskMetrics* of J.P. Morgan (1996). The RiskMetrics method consists of the following three steps. First, it estimates the one-period volatility $\widehat{\sigma}_t$ by the exponential smoothing (see §6.2.4)

$$\widehat{\sigma}_t^2 = (1-\lambda)r_{t-1}^2 + \lambda\widehat{\sigma}_{t-1}^2. \tag{8.54}$$

Second, for a τ-period return, the square-root rule is used for computing the volatilities of τ-period returns $R_{t,\tau}$:

$$\widehat{\theta}_{t,\tau} = \sqrt{\tau}\widehat{\sigma}_t. \tag{8.55}$$

J.P. Morgan recommends using (8.55) with $\lambda = 0.97$ for forecasting the monthly ($\tau = 25$ trading days) volatilities of aggregate returns. The final step is to forecast the VaR through the normality assumption on the standardized return process $\{R_{t,\tau}/\widehat{\theta}_{t,\tau}\}$; that is, the τ-period VaR is forecasted as

$$\widehat{V}_{t+1,\tau} = \Phi^{-1}(\alpha)\widehat{\theta}_{t,\tau}. \tag{8.56}$$

RiskMetrics has been scrutinized by many practitioners and regulators. For example, the square-root rule (8.55) was criticized by Diebold, Hickman, Inoue, and Schuermann (1998). Fan and Gu (2001) showed that the normal quantile in (8.56) can be improved by using a symmetric nonparametric method, and the volatility estimate in Step 1 can also be ameliorated. In this example, we show how the additive model can be used to improve the prediction of multiperiod volatility. To be more specific, we will compute the monthly ($\tau = 25$ trading days) VaR using the Standard and Poor's 500 Index. The in-sample period is set from January 2, 1990 to April 30, 1996, which consists of a series of length 1,500. The confidence level is taken to be $1-\alpha = 0.95$.

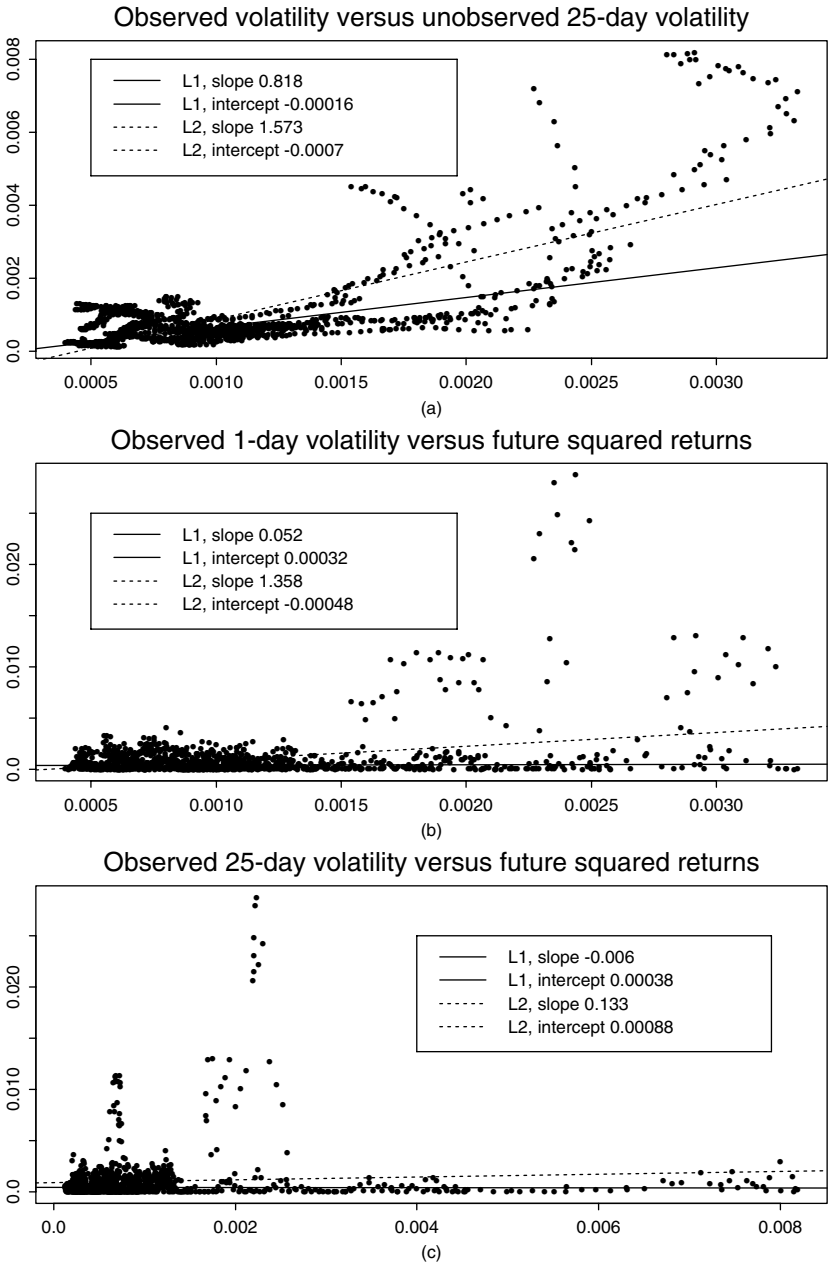


FIGURE 8.11. (a) Scatterplot of $\widehat{\theta}_{t,\tau}$ versus unobserved $\widehat{\sigma}^2_{t,\tau}$ along with regression lines using the least-squares fit (dashed) and the least-absolute deviation fit (solid). (b) Scatterplot of $\widehat{\theta}_{t,\tau}$ versus aggregate squared return $R^2_{t,\tau}$ along with their regression lines. (c) Scatterplot of $\sigma^2_{t-\tau,\tau}$ against $R^2_{t,\tau}$ along with their regression lines.

Let $\widehat{\sigma}^2_{t,\tau}$ be the estimated volatility via the exponential smoothing of the aggregate return $\{R_{t,\tau}\}$:

$$\widehat{\sigma}^2_{t,\tau} = (1-\lambda)R^2_{t-1,\tau} + \lambda\widehat{\sigma}^2_{t-1,\tau}.$$

In our illustration below, we use $\lambda = 0.97$ for computing the monthly volatility. Note that $\widehat{\sigma}^2_{t,\tau}$ is unobservable at the time t since it involves future observations. However, the J.P. Morgan estimate $\widehat{\theta}_{t,\tau}$ in (8.55) is observable. Figure 8.11(a) shows the relationship between these two quantities using the data from January 2, 1991 (the initial year's estimates were discarded to avoid boundary effects) to April 30, 1996. Shown in the figure are the regression lines using the least-squares fit and the least absolute deviation fit. If RiskMetrics gives good estimates, then the intercept and slope of the regression lines should be near 0 and 1, respectively. This is nearly the case for the least-absolute deviation (L_1) fit but not so for the least-squares (L_2) fit. The outliers at the upper right-hand corner are also influential points. This pushes the least-squares estimate upward significantly. In Figure 8.11(b), we show directly the scatterplot between the aggregated squared return $R^2_{t,\tau}$, which depends on future value, and the estimated volatility $\widehat{\theta}_{t,\tau}$.

The volatility for the aggregated return $R_{t,\tau}$ is also related to the observed τ-period volatility $\widehat{\sigma}^2_{t-\tau,\tau}$ at time t. The relation between this observed τ-period volatility and the future squared return $R^2_{t,\tau}$ is depicted in Figure 8.11(c). Clearly, the observed values $\widehat{\theta}^2_{t,\tau}$ and $\widehat{\sigma}^2_{t-\tau,\tau}$ are very relevant for predicting the volatility of the unobserved squared aggregate return $R^2_{t,\tau}$. By collecting this information $\{(\widehat{\theta}^2_{t,\tau}, \widehat{\sigma}^2_{t-\tau,\tau}, R^2_{t,\tau})\}$ in the in-sample period, from January 2, 1991 to April 30, 1996, we aim to build a stochastic regression model for predicting multiple-period volatility.

As an illustration, we fit the additive regression model

$$R^2_{t,\tau} = \mu + f_1(\widehat{\theta}^2_{t,\tau}) + f_2(\widehat{\sigma}^2_{t-\tau,\tau}) + \varepsilon_t. \tag{8.57}$$

By using the backfitting algorithm along with the local linear fit, we obtain the fitted functions $\widehat{f}_1$ and $\widehat{f}_2$ displayed in Figure 8.12. To reduce the influence of outliers, 5% of data points, whose aggregate returns are at both tails, were discarded. This gives 1,161 data points for fitting the additive model. The intercept $\widehat{\mu} = 0.0007$.

The additive model (8.57) was applied to forecast the volatility in the in-sample period and out-sample period (May 1, 1996 to December 31, 1999, with length 1,003). The results are depicted in Figure 8.13. Shown in Figure 8.13(b) is $R_{t,\tau}/1.645$ when returns are negative. They indicate the extent to which the negative returns exceed the forecasted VaR.

Following the recommendation by J.P. Morgan, the multiple-period VaR is simply calculated by (8.56). This can be ameliorated by using the symmetric nonparametric method given by Fan and Gu (2001), but this is

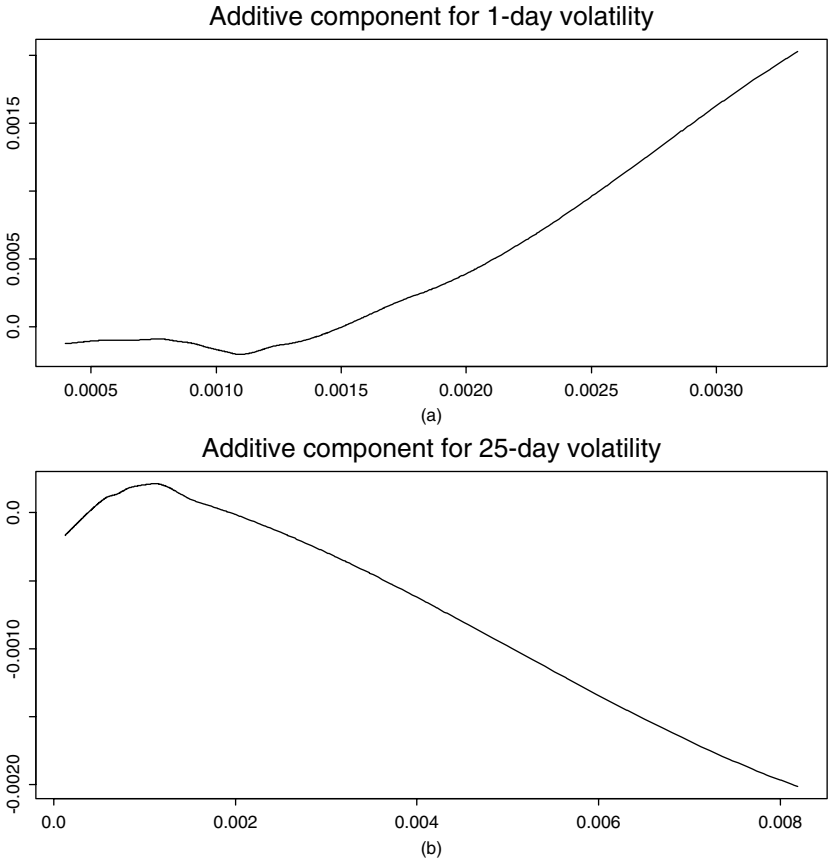


FIGURE 8.12. Estimated additive functions for model (8.57). (a) $\widehat{f}_1$; (b) $\widehat{f}_2$.

beyond the scope of this example. The common measure of volatility estimation includes the *exceedence ratio* (ER), which is defined as

$$\text{ER} = n^{-1} \sum_{t=T+1}^{T+n} I(R_{t,\tau} < \Phi^{-1}(\alpha)\widehat{\sigma}_{t,\tau}),$$

where $T+1$ and $T+n$ are the first and last days of the out-sample period. This is to be compared with the confidence level $1-\alpha$. The measure gives us an idea how well the volatility forecast can be used for the calculation of VaR. An alternative criterion is the mean square error defined by

$$\text{MSE} = n^{-1} \sum_{t=T+1}^{T+n} (R_{t,\tau}^2 - \widehat{\sigma}_{t,\tau}^2)^2.$$

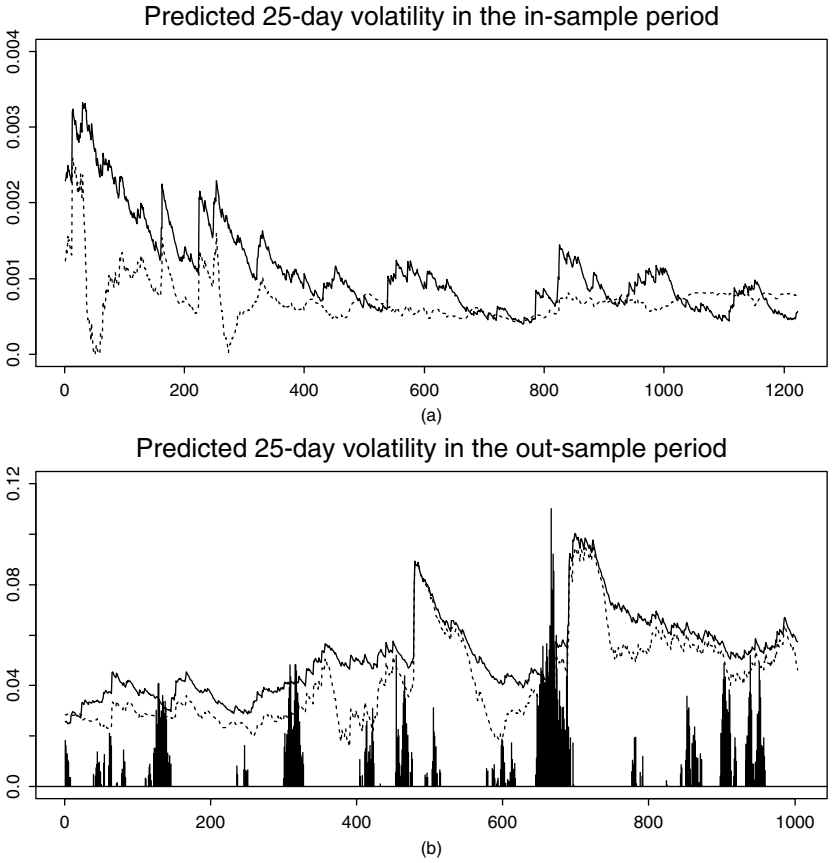


FIGURE 8.13. Predicted 25-day volatility. (a) In-sample period from January 2, 1991 to April 30, 1996. (b) Out-sample period from May 1, 1996 to December 31, 1999. Solid curves: J.P. Morgan's forecast; dashed curves: volatility forecast by the additive model (8.57). The bars in (b) are $-R_{t,\tau}/1.645$ when returns are negative.

Letting $\sigma^2_{t,\tau} = E(R^2_{t,\tau}|\Omega_t)$ be the true multiple-period volatility, the expected value can be decomposed as

$$E(\text{MSE}) = n^{-1} \sum_{t=T+1}^{T+n} E(\sigma^2_{t,\tau} - \widehat{\sigma}^2_{t,\tau})^2 + n^{-1} \sum_{t=T+1}^{T+n} E(R^2_{t,\tau} - \sigma^2_{t,\tau})^2.$$

The first term reflects the effectiveness of the estimated volatility, while the second term is the size of the stochastic error, independent of estimators. The second term is usually much larger than the first term. Thus, a small improvement in MSE could mean substantial improvement over the estimated volatility. However, due to the well-known fact that financial time series contain outliers due to large market movements, the mean-square

TABLE 8.6. Performance comparisons for multiple-period volatility estimates.

	Exceedence ratio	Mean-square error	Mean absolute error
RiskMetrics	2.37%	1.161×10^{-5}	2.41×10^{-3}
Additive model	5.08%	1.165×10^{-5}	2.25×10^{-3}

error is not a robust measure. Therefore, we also used the mean-absolute deviation error:

$$\text{MADE} = n^{-1} \sum_{t=T+1}^{T+n} \mid R_{t,\tau}^2 - \widehat{\sigma}_{t,\tau}^2 \mid .$$

The performance of the multiple-period volatility estimates is summarized in Table 8.6. In terms of the exceedence ratio, the method based on the additive model performs much better. In fact, RiskMetrics performs even worse for the period from January 2, 1991 to April 30, 1996, which gives an exceedence ratio of 1%. (Note that the J.P. Morgan estimate of multiple-period volatility $\widehat{\theta}_{t,\tau}$ does not depend on the in-sample training.) For the same period, the multiple-period volatility produced by the additive model gives the exceedence ratio 2.37%. These results indicate that the square-root rule (8.55) tends to overforecast the multiple-period volatility VaR. This can easily be seen from Figure 8.13. In terms of the mean-square errors, both methods perform approximately the same. However, the approach based on the additive model outperforms *RiskMetrics* in terms of the mean absolute deviation error, which is more robust to the outliers caused by large market movements. ∎

8.6 Other Nonparametric Models

There are many nonparametric models for multivariate regression data. They can be extended to the time series context to model autoregressive functions. We highlight some of these models and techniques in the context of analyzing nonlinear time series. For an overview of multivariate nonparametric models, see Chapter 7 of Fan and Gijbels (1996).

Different models exploit different aspects of data structure. Together they form useful tool kits for processing time series data and for checking the adequacy of commonly-used parametric models. Some models are more general than others. The choice of model depends critically on practical needs. However, some general guidelines from statistical considerations are also helpful. A larger family of nonparametric models implies, in principle, smaller modeling biases, yet the unknown parameters and functions in such a model may not be estimated accurately. On the other hand, a smaller family of models may create large modeling biases, but unknown parameters and functions can be estimated with reasonable accuracy. Therefore,

a compromise between estimability and modeling biases should be reached after careful consideration and investigation. Other considerations being equal, parsimonious models are preferable due to their interpretability. The discussion above indicates that the choice of models should also take into account sample sizes, which are directly related to the estimability issue.

8.6.1 Two-Term Interaction Models

The autoregressive regression surface G can be better approximated when the additive model (8.40) is replaced by the *two-term interaction model*:

$$X_t = \sum_{j=1}^{p} f_j(X_{t-j}) + \sum_{1 \le j < k \le p} f_{jk}(X_{t-j}, X_{t-k}) + \varepsilon_t. \tag{8.58}$$

This is a more flexible family of models than the additive models. The univariate components are regarded as main-effect functions, whereas the bivariate components are interaction terms. As in §8.5, when the model (8.58) does not hold, the effort of model (8.58) is to search a two-term interaction model that best approximates the true autoregression surface G.

The issue of identifiability arises naturally. In addition to the identifiability of the main effect terms, the conditions on the bivariate functions f_{jk} should also be imposed. Indeed, we can add an arbitrary function $h(X_{t-j})$ to the component $f_j(X_{t-j})$ and then subtract it from the function $f_{jk}(X_{t-j}, X_{t-k})$. Thus, in addition to the constraints (8.40), the following requirements should also be imposed on the bivariate interactions:

$$E\{f_{jk}(X_{t-j}, X_{t-k})|X_{t-j}\} = E\{f_{jk}(X_{t-j}, X_{t-k})|X_{t-k}\} = 0, 1 \le j < k \le d. \tag{8.59}$$

These are simple and convenient conditions for the identifiability. The backfitting algorithm can be extended to estimate the functions f_j and f_{jk}; see, for example, Hastie and Tibshirani (1990) and Fan and Gijbels (1996).

By using the backfitting algorithm, only one- and two-dimensional nonparametric smoothers are used. Hence, the curse of dimensionality is not very severe with moderate sample sizes. In fact, according to Stone (1994), the interaction terms can be estimated at rate $O(n^{-s/(2s+2)})$ and the main-effect functions at rate $O(n^{-s/(2s+1)})$ when they have the sth derivative. In other words, the problem (8.58) is as hard as a two-dimensional nonparametric smoothing problem. The projection method in §8.5.3 can be employed here as well.

A specific implementation of estimating functions in (8.58) is the regression spline method. By selecting a sequence of knots, one forms the spline basis $\{B_j(\cdot), j = 1, \cdots, L\}$; see §6.4. Approximate the univariate functions

by

$$f_j(x) \approx \sum_{k=1}^{L} \theta_{jk} B_k(x)$$

and the bivariate functions by using the multivariate tensor-product spline basis:

$$f_{ij}(x_i, x_j) \approx \sum_{k=1}^{L} \sum_{l=1}^{L} \theta_{ijkl} B_k(x_i) B_l(x_j).$$

Then, the model (8.58) is approximated as

$$\begin{aligned} X_t &\approx \alpha + \sum_{j=1}^{p} \left\{ \sum_{k=1}^{L} \theta_{jk} B_k(X_{t-j}) \right\} \\ &\quad + \sum_{1 \le i < j \le p} \left\{ \sum_{k=1}^{L} \sum_{l=1}^{L} \theta_{ijkl} B_k(X_{t-i}) B_l(X_{t-j}) \right\}. \end{aligned}$$

With the approximations above, the problem becomes estimating the coefficients θ. Friedman (1991) gives a specific implementation to the spline method above, resulting in the *multivariate adaptive regression splines method* (MARS). The techniques above have been applied to the time series by Lewis and Stevens (1991).

8.6.2 Partially Linear Models

Partially linear models are a specific member of the FAR models and additive models. They allow us to examine whether a specific lagged variable has nonlinear contributions to the autoregressive function G. Since the models are more parsimonious than the FAR and additive models, the parameters and functions can be estimated more accurately.

To facilitate the presentation, we assume, without loss of generality, that the first component is possibly nonlinear. The *partially linear models* postulate the following structure on the autoregressive function:

$$X_t = f(X_{t-1}) + \beta_2 X_{t-2} + \cdots + \beta_p X_{t-p} + \varepsilon_t. \tag{8.60}$$

The method of fitting is typically the *profile least-squares* method or, more generally, the profile likelihood method; see, for example, §8.4.3, Speckman (1988), Carroll, Fan, Gijbels, and Wand (1997), and Murphy and van der Vaart (2000). The idea is as follows. Pretend, for a moment, that $\beta_2, \cdots, \beta_p$ are known. Then, the function f can be estimated by the nonparametric method in Chapter 6. Let us denote the resulting estimate by $\widehat{f}(\cdot; \boldsymbol{\beta})$, where we stress the dependence of the estimate on

$$\boldsymbol{\beta} = (\beta_2, \cdots, \beta_p)^T.$$

Substituting the estimate into (8.60), we have

$$X_t = \widehat{f}(X_{t-1}, \boldsymbol{\beta}) + \beta_2 X_{t-2} + \cdots + \beta_p X_{t-p} + \varepsilon_t. \tag{8.61}$$

Now, the parameters $\boldsymbol{\beta}$ can be estimated by using the least-squares method that minimizes

$$\sum_{t=p+1}^{T} \{X_t - \widehat{f}(X_{t-1}, \boldsymbol{\beta}) - \beta_2 X_{t-2} - \cdots - \beta_p X_{t-p}\}^2. \tag{8.62}$$

Let $\widehat{\boldsymbol{\beta}}$ be the resulting profile least-squares estimator. Then, the nonparametric component is simply estimated by $\widehat{f}(\cdot; \widehat{\boldsymbol{\beta}})$.

As an illustration, consider the kernel estimator

$$\widehat{f}(x, \boldsymbol{\beta}) = \frac{\sum_{t=p+1}^{T} K_h(X_{t-1} - x)(X_t - \beta_2 X_{t-2} - \cdots - \beta_p X_{t-p})}{\sum_{t=p+1}^{T} K_h(X_{t-1} - x)}.$$

Then, model (8.61) is a linear model in the sense that it depends linearly on $\boldsymbol{\beta}$. Hence, the least-squares problem (8.62) can easily be solved. Let

$$g_j(x) = \frac{\sum_{t=p+1}^{T} K_h(X_{t-1} - x) X_{t-j}}{\sum_{t=p+1}^{T} K_h(X_{t-1} - x)}.$$

Then (8.62) is simply

$$\sum_{t=p+1}^{T} [X_t - \widehat{g}_0(X_{t-1}) - \beta_2\{X_{t-2} - \widehat{g}_2(X_{t-1})\} - \cdots - \beta_p\{X_{t-p} - \widehat{g}_p(X_{t-1})\}]^2$$

and hence can easily be minimized. The approach applies to all linear smoothers, including the local polynomial estimators and spline estimators.

For the multivariate regression model, it has been shown by Speckman (1988) that the parametric components $\boldsymbol{\beta}$ can be estimated at rate $T^{-1/2}$, and the nonparametric function f can be estimated as well as if $\boldsymbol{\beta}_2$ were known. For a more detailed account of the partially linear model, see the monograph by Härdle, Liang, and Gao (2000).

8.6.3 Single-Index Models

The single-index model, translated into the time series context, becomes

$$X_t = g(\beta_1 X_{t-1} + \cdots + \beta_p X_{t-p}) + \varepsilon_t. \tag{8.63}$$

By allowing an unknown *link function* g, the linear autoregressive predictor is used to predict X_t.

This single-index model (8.63) is a specific case of the AFAR model (8.29). The parameter β is identifiable up to a sign change if we impose $\|\boldsymbol{\beta}\|^2 = 1$. The profile least-squares in §8.4.3 can be applied. Other techniques mentioned in §8.4.8 can also be used to estimate the index parameters. The model has been widely used in econometrics and statistics; see, for example, Härdle, Hall and Ichimura (1993), Ichimura (1993), Newey and Stoker (1993), Samarov (1993), Carroll, Fan, Gijbels, and Wand (1997), and Heckman, Ichimura, Smith, and Todd (1998), among others.

8.6.4 Multiple-Index Models

A further extension of the single-indices model is to allow multiple-index in the model (8.63), leading to

$$X_t = g(\beta_{11}X_{t-1}+\cdots+\beta_{1p}X_{t-p}, \cdots, \beta_{d1}X_{t-1}+\cdots+\beta_{dp}X_{t-p})+\varepsilon_t. \quad (8.64)$$

The number of indices d is usually small in order to avoid the curse of dimensionality. The model is frequently used in an exploratory stage of study, when time series analysts look for the possible low-dimensional structure to approximate the autoregression surface G. As mentioned in §8.4.8, the index parameters $\boldsymbol{\beta}$ are not identifiable, but the linear space spanned by these index parameters is. Popular methods for estimating the linear span include the sliced inverse regression, principal Hessian directions, the average derivative method, and other forward regression methods. To our knowledge, some of these methods have not yet been applied to the time series context.

The models in §8.6.2–§8.6.4 involve both parametric and nonparametric parts. These kinds of models are frequently referred to as *semiparametric models*. A comprehensive account of this subject can be found in Bickel, Klaassen, Ritov, and Wellner (1993), where efficient estimation of parametric components is emphasized.

To illustrate the usefulness of this class of models, we apply the model (8.64) to the environmental data set in Example 1.5, analyzed recently by Xia, Tong, Li, and Zhu (2002) by using the *minimum average variance estimation* (MAVE) technique. To this end, we briefly introduce their MAVE technique. To facilitate the notation, let

$$\mathbf{B} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d), \quad \boldsymbol{\beta}_j = (\beta_{j1}, \cdots, \beta_{jp})^T.$$

Model (8.64) is a specific case of the stochastic regression model

$$Y_t = g(\mathbf{B}^T\mathbf{X}_t) + \varepsilon_t, \quad (8.65)$$

where $\mathbf{X}_t$ is a vector of p-covariates observed at time t. The true parameter $\mathbf{B}_0$ is a solution to the problem

$$\min_{\mathbf{B}} E\{Y - E(Y|\mathbf{B}^T\mathbf{X})\}^2 = E\sigma^2_{\mathbf{B}}(\mathbf{B}^T\mathbf{X}), \tag{8.66}$$

where

$$\sigma^2_{\mathbf{B}}(\mathbf{B}^T\mathbf{X}) = E\left[\{Y - E(Y|\mathbf{B}^T\mathbf{X})\}^2|\mathbf{B}^T\mathbf{X}\right]$$

is the conditional variance. The MAVE is to find a $\mathbf{B}$ to minimize the empirical version of (8.66).

To implement the idea of MAVE, we first need to estimate the conditional variance. A more detailed study of this will be given in the next section. Let $g_B(\mathbf{v}) = E(Y|\mathbf{B}^T\mathbf{X} = \mathbf{v})$, which is a d-variate nonparametric function. The local linear fit can be readily applied to estimate $g_B(\cdot)$. For a given $\mathbf{v}_0$, by Taylor expansion,

$$g_B(\mathbf{v}) \approx g_B(\mathbf{v}_0) + g'_B(\mathbf{v}_0)(\mathbf{v} - \mathbf{v}_0) \equiv a + \mathbf{b}^T(\mathbf{v} - \mathbf{v}_0).$$

For a given sample $\{(\mathbf{X}_t, Y_t) : t = 1, \cdots, T\}$, by (8.65), we have the approximate local linear model

$$Y_t = a + \mathbf{b}^T(\mathbf{B}^T\mathbf{X}_t - \mathbf{v}_0) + \varepsilon_t \tag{8.67}$$

for $\mathbf{B}^T\mathbf{X}_t \approx \mathbf{v}_0$. For a given $\mathbf{B}$, the local parameters a and $\mathbf{b}^T$ can be estimated via the local least-squares method, which minimizes

$$\sum_{t=1}^{T}\{Y_t - a - \mathbf{b}^T(\mathbf{B}^T\mathbf{X}_t - \mathbf{v}_0)\}^2 w(\mathbf{B}^T\mathbf{X}_t, \mathbf{v}_0),$$

where $w(\mathbf{B}^T\mathbf{X}_t, \mathbf{v}_0) \geq 0$ is a weight function and typically decreases with the distance between $\mathbf{B}^T\mathbf{X}_t$ and $\mathbf{v}_0$, satisfying

$$\sum_{t=1}^{T} w(\mathbf{B}^T\mathbf{X}_t, \mathbf{v}_0) = 1.$$

As an example, one can take the kernel weight:

$$w(\mathbf{B}^T\mathbf{X}_t, \mathbf{v}_0) = K_h(\mathbf{B}^T\mathbf{X}_t - \mathbf{v}_0)/\sum_{t'=1}^{T} K_h(\mathbf{B}^T\mathbf{X}_{t'} - \mathbf{v}_0). \tag{8.68}$$

Just as in the ordinary linear model, the variance $\widehat{\sigma}^2_{\mathbf{B}}(\mathbf{v}_0)$ can be estimated by the residual sum of squares

$$\widehat{\sigma}^2_{\mathbf{B}}(\mathbf{v}_0) = \min_{a,\mathbf{b}} \sum_{t'=1}^{T}\{Y_{t'} - a - \mathbf{b}^T(\mathbf{B}^T\mathbf{X}_{t'} - \mathbf{v}_0)\}^2 w(\mathbf{B}^T\mathbf{X}_t, \mathbf{v}_0). \tag{8.69}$$

Based on (8.66), one would find a $\mathbf{B}$ to minimize

$$\sum_{t=1}^{T} \widehat{\sigma}^2_{\mathbf{B}}(\mathbf{B}^T\mathbf{X}_t).$$

By (8.69), this is the same as minimizing with respect to $\mathbf{B}$ the following function:

$$\min_{\{a_t,b_t\}} \sum_{t=1}^{T}\sum_{t'=1}^{T} \{Y_{t'} - a_t - \mathbf{b}_t^T(\mathbf{B}^T\mathbf{X}_{t'} - \mathbf{B}^T\mathbf{X}_t)\}^2 w(\mathbf{B}^T\mathbf{X}_{t'}, \mathbf{B}^T\mathbf{X}_t). \quad (8.70)$$

As explained above, there are multiple solutions of $\mathbf{B}$. Only the space spanned by the columns of the matrix $\mathbf{B}$ is identifiable.

The MAVE estimator utilizes an idea similar to the profile least-squares method in §8.4.3. To see this, let $\widehat{a}_t$ and $\widehat{b}_t$ be the solutions in (8.70). Then (8.70) can be written as

$$\sum_{t=1}^{T}\sum_{t'=1}^{T} \{Y_{t'} - \widehat{a}_t - \widehat{\mathbf{b}}_t^T(\mathbf{B}^T\mathbf{X}_{t'} - \mathbf{B}^T\mathbf{X}_t)\}^2 w(\mathbf{B}^T\mathbf{X}_{t'}, \mathbf{B}^T\mathbf{X}_t). \quad (8.71)$$

The iterative algorithm in §8.4.3 can be employed to compute an estimate of $\mathbf{B}$ with kernel weights (8.68): Given $\mathbf{B}$, compute the weights (8.68) and find the local linear estimate $\widehat{a}_t$ and $\widehat{\mathbf{b}}_t$ from (8.70). For a given least-squares estimate with kernel weights computed from the most recent estimate of $\mathbf{B}$, find a $\mathbf{B}$ to minimize (8.71).

Due to the local weighting scheme, the double sum in (8.71) is basically concentrated on terms that are near the diagonal. Thus, since the effective terms in (8.71) satisfy $\mathbf{B}^T\mathbf{X}_t \approx \mathbf{B}^T\mathbf{X}_{t'}$, it follows that

$$\widehat{a}_t \approx \widehat{a}_{t'} \quad \text{and} \quad \widehat{\mathbf{b}}_t^T(\mathbf{B}^T\mathbf{X}_{t'} - \mathbf{B}^T\mathbf{X}_t) \approx 0.$$

Using these approximations, (8.71) becomes

$$\sum_{t=1}^{T}\sum_{t'=1}^{T} \{Y_{t'} - \widehat{a}_{t'}\}^2 w(\mathbf{B}^T\mathbf{X}_{t'}, \mathbf{B}^T\mathbf{X}_t) = \sum_{t'=1}^{T} \{Y_{t'} - \widehat{a}_{t'}\}^2, \quad (8.72)$$

when $\sum_{t=1}^{T} w(\mathbf{B}^T\mathbf{X}_{t'}, \mathbf{B}^T\mathbf{X}_t) = 1$, which is satisfied for the kernel weights (8.68). The minimization problem (8.72) is the profile least-squares method. In other words, the MAVE and the profile least-squares estimate would give approximately the same solution when h is small enough. They may differ for a given bandwidth h.

When the dimensional d is large, both the MAVE and the profile least-squares estimates face the challenge of the curse of dimensionality. The inverse regression method of Li (1991) averts this kind of problem, but other assumptions are needed that are hardly valid for nonlinear time series.

In practical implementation, the issue of the choice of the bandwidth and number of indices d arises. Xia, Tong, Li, and Zhu (2002) suggested a cross-validation technique to choose the parameters d and h and established nice sampling properties of the proposed methods.

8.6.5 An Analysis of Environmental Data

Consider the environmental data discussed in Example 1.5. The analysis is taken from Xia, Tong, Li, and Zhu (2002). The daily number of admissions (Y_t) is taken as the response variable, whereas the pollutants and the weather variables are taken as the covariates. They are the average levels of sulfur dioxide X_{1t}, nitrogen dioxide X_{2t}, respirable suspended particulates X_{3t}, ozone X_{4t}, temperature X_{5t} (in ^{o}C), and relative humidity X_{6t} (in percent). The variables Y_t, X_{1t}, X_{2t}, and X_{3t} are presented in Figure 1.5. The variables X_{4t}, X_{5t}, and X_{6t} are presented in Figure 8.14.

Xia, Tong, Li, and Zhu (2002) preprocessed the response data Y_t. Due to the release of additional hospital beds to accommodate circulatory and respiratory patients in the course of this study, the time effect is expected. The time trend was removed by using a kernel smoother. Let the resulting (residual) series be Y_t'. It is also expected to have the day-of-the-week effect, presumably due to the hospital booking system. This effect was estimated by a simple regression method using dummy variables,

$$Y_t' = \beta_0 + \sum_{j=1}^{6} \beta_j I(d(t) = j) + \varepsilon_t,$$

where $d(t)$ is the day of the week for time t. Only six indicators are used because

$$\sum_{j=1}^{7} I(d(t) = j) = 1,$$

which is the same as the intercept term. Let $\widehat{\varepsilon}_t$ be the residual series. To simplify the notation, we still use Y_t to denote $\widehat{\varepsilon}_t$, which is a series with the time effect and the day-of-the-week effect removed. Figure 8.10(d) shows the resulting filtered series.

Since the pollutant and weather may affect the circulatory and respiratory systems with a time delay, the covariates in the last 7-days were considered. This results in 42 covariates:

$$\mathbf{X}_t = (X_{1,t-1}, X_{1,t-2}, \cdots, X_{1,t-7}, \cdots, X_{6,t-1}, X_{6,t-2}, \cdots, X_{6,t-7})'.$$

All variables are standardized to have a mean 0 and variance 1 before the fitting using the multiple-index model (8.64). We use the same notation to denote the standardized variables. Applying the MAVE method with cross-validation, Xia, Tong, Li, and Zhu (2002) obtained three indices $d = 3$. Their corresponding coefficients are shown in Table 8.7.

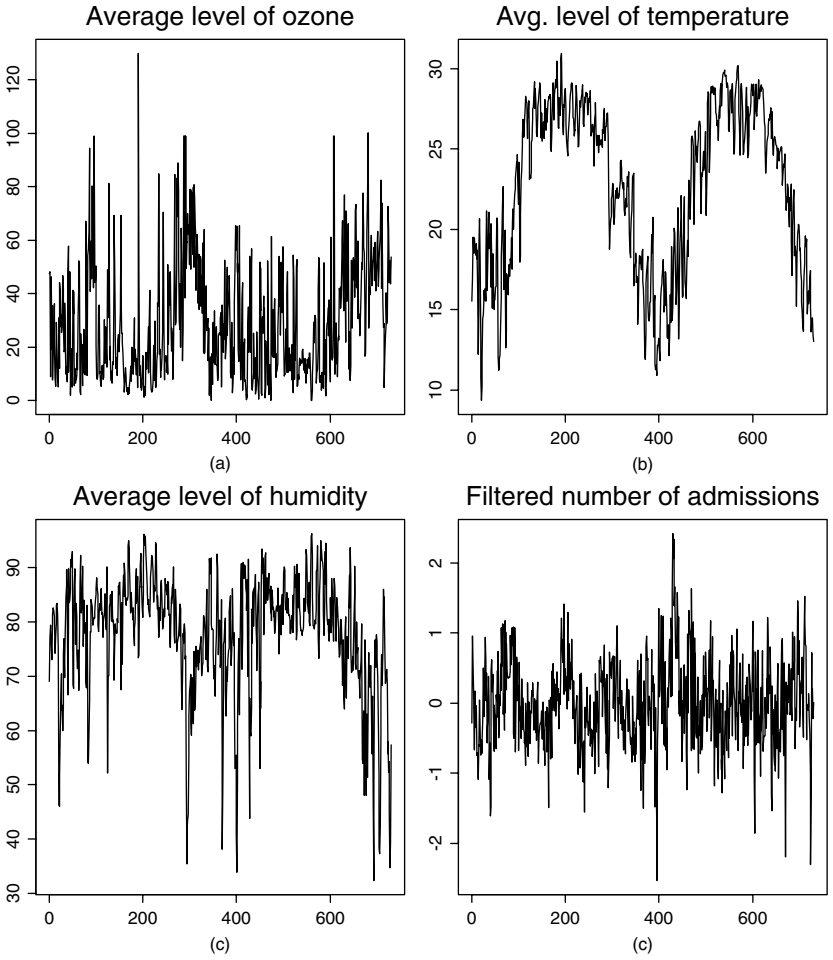


FIGURE 8.14. The daily average of (a) ozone level, (b) temperature, (c) humidity, and (d) filtered number of daily admissions of circulatory and respiratory patients by removing the time trend and the day-of-the-week effect.

Figures 8.15 (a)–(c) showed Y_t plotted against the index variable $\widehat{\boldsymbol{\beta}}_j^T \mathbf{X}_t$ $(j = 1, 2, 3)$. Based on these plots together with Table 8.7, Xia, Tong, Li, and Zhu (2002) suggested the following features.

- Rapid temperature changes play an important role. Notice that the dominant coefficients in $\widehat{\boldsymbol{\beta}}_1$ are associated with variables $X_{5,t}$ and $X_{5,t-1}$.
- Among the pollutants, the most influential one seems to be the particulates. Note that the large coefficients are mainly associated with the lag variables of $X_{3,t}$.

TABLE 8.7. Estimated coefficients $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2$, and $\widehat{\boldsymbol{\beta}}_3$ (boldfaced entries have relatively large absolute values). Adapted from Xia, Tong, Li, and Zhu (2002) with permission from the Royal Statistical Society.

lags	1	2	3	4	5	6	7
x_1	0.0586	−0.0854	0.0472	−0.0152	**0.1083**	−0.0942	0.0734
x_2	0.0876	0.0313	**−0.1964**	0.0893	−0.0867	0.0951	**−0.1068**
x_3	**−0.2038**	**0.1103**	0.0153	0.0740	−0.0756	**0.1283**	−0.0520
x_4	0.0155	0.0692	**0.1622**	**−0.2624**	**0.1312**	**0.1342**	0.0976
x_5	**0.5065**	**−0.4079**	0.0743	0.0859	**−0.3024**	−0.1734	−0.0302
x_6	−0.0294	−0.0610	0.0129	−0.0392	−0.0075	**0.2850**	0.0513
x_1	**−0.1525**	0.0962	**−0.1112**	**0.1170**	−0.0388	−0.0605	−0.0326
x_2	−0.0029	**0.1614**	−0.0955	**−0.1160**	**−0.2185**	0.0826	**0.1696**
x_3	−0.0096	**−0.1874**	**0.2422**	−0.0047	**0.3272**	**−0.2646**	−0.0041
x_4	−0.0013	**−0.1162**	0.0673	**0.2113**	**−0.2193**	**0.1235**	**−0.1282**
x_5	**0.1410**	**0.1193**	**−0.1425**	**0.1819**	**−0.2793**	−0.0880	−0.0325
x_6	−0.0345	**−0.1479**	−0.0400	**0.4033**	0.0474	0.0899	**0.1336**
x_1	0.0701	0.0065	−0.0535	**−0.1570**	−0.0553	−0.0091	−0.0363
x_2	−0.0529	**0.1360**	0.0723	**0.1045**	−0.0045	−0.0200	0.0221
x_3	−0.0121	**−0.1189**	0.0715	−0.0814	0.0112	0.0155	**0.1214**
x_4	**0.2215**	0.0103	**−0.3304**	**0.1028**	0.0160	**−0.1805**	**0.1341**
x_5	**0.2909**	**−0.2372**	0.0621	−0.0211	0.0950	−0.0954	**0.2507**
x_6	**0.2797**	**−0.1094**	**−0.3038**	0.0452	**0.1754**	**−0.3937**	**0.2597**

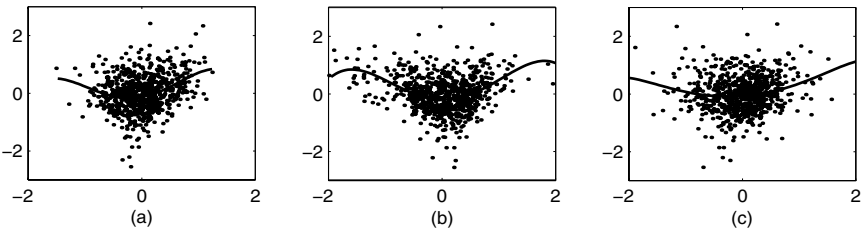


FIGURE 8.15. The observed series y_t is plotted against (a) index $\boldsymbol{\beta}_1^T \mathbf{X}_t$; (b) index $\boldsymbol{\beta}_2^T \mathbf{X}_t$; and (c) index $\boldsymbol{\beta}_3^T \mathbf{X}_t$. The lines are drawn simply by polynomial regression to make the trends more visualizable. Adapted from Xia, Tong, Li, and Zhu (2002) with permission from the Royal Statistical Society.

- The weather covariates are influential. Note the many large coefficients for the weather covariates $X_{6,t-j}, j = 1, \cdots, 7$ in all of the three $\widehat{\boldsymbol{\beta}}$'s.

To obtain a more parsimonious description of the model, Xia, Tong, Li, and Zhu (2002) carried out the analysis focusing on the suspended particulates X_3, the ozone X_4, the temperature X_5, and their associated lagged variables. Further simplification is obtained by selecting only one lag for each covariate. By applying the method of Yao and Tong (1994b), the lagged variables $X_{3,t-2}$, $X_{4,t-6}$, $X_{5,t-4}$, and $X_{6,t-2}$ were selected. As indicated in the analysis above, the temperature variation plays an important role. Thus, an additional variable

$$V_t = \mathrm{SD}(X_{5,t-j}, j = 1, 2, 3, 4, 5)$$

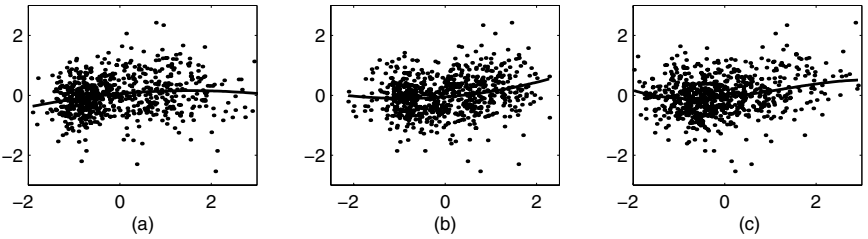


FIGURE 8.16. The observed series y_t is plotted against (a) index $\boldsymbol{\beta}_1^T \mathbf{Z}_t$, (b) index $\boldsymbol{\beta}_2^T \mathbf{Z}_t$, and (c) index $\boldsymbol{\beta}_3^T \mathbf{Z}_t$. The lines are drawn simply by polynomial regression to make the trends more visualizable. Adapted from Xia, Tong, Li, and Zhu (2002) with the permission from the Royal Statistical Society.

is introduced to measure the temperature variation in the past 5 days prior to the time t. Let $\mathbf{Z}_t = (X_{3,t-2}, X_{4,t-6}, X_{5,t-4}, X_{6,t-2}, V_t)^T$. Applying the multiple-index model to the data $(\mathbf{Z}_t, Y_t)$, three indices were chosen with the following estimated coefficients:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_1 &= (-0.1316, \ -0.0772, \ -0.8366, \ -0.0235, \ \ 0.5256)^T, \\
\widehat{\boldsymbol{\beta}}_2 &= \ \ (0.4809, \ -0.3154, \ -0.5078, \ \ 0.0018, \ -0.6414)^T, \\
\widehat{\boldsymbol{\beta}}_3 &= \ \ (0.0101, \ \ 0.3815, \ -0.0734, \ -0.9115, \ \ 0.1345)^T.
\end{aligned}$$

Figures 8.16 (a)–(c) show Y_t plotted against these three indices $\widehat{\boldsymbol{\beta}}_j^T \mathbf{Z}$ $(j = 1, 2, 3)$. Compared with the fit using three indices of 42 covariates, the fit using the reduced set of five covariates decreases the variance explained by covariates (multiple R^2) from about 73% to about 66%. In return, further insights are gained. Specifically, Xia, Tong, Li, and Zhu (2002) made the following observations.

- The temperature and temperature variation are dominant components in the index $\boldsymbol{\beta}_1^T \mathbf{Z}_t$, yet they donot seem to cause a large variation in hospital admissions; see Figure 8.16(a).

- The dominant components in the second index $\widehat{\boldsymbol{\beta}}_2^T \mathbf{Z}_t$ and Figure 8.16 (b) suggest that high levels of suspended particulates and/or high levels of ozone during cold weather tend to cause high admissions for patients with circulatory and respiratory problems.

- The coefficients of $\widehat{\boldsymbol{\beta}}_3$ together with Figure 8.16(c) suggest that high ozone levels on dry days tend to be associated with high admissions.

8.7 Modeling Conditional Variance

Conditional variance is pivotal for statistical data analysis. It can be used to construct confidence intervals and conduct other statistical inferences.

It can be applied to assess the size of prediction errors. In financial applications, the conditional variance is also refered to as volatility. It is directly related to the price of financial derivatives such as options and the measures of risks of traded assets such as value at risk.

8.7.1 Methods of Estimating Conditional Variance

A general time series model is to assume that the series is generated from

$$X_t = f(X_{t-1}, \cdots, X_{t-p}) + \sigma(X_{t-1}, \cdots, X_{t-p})\varepsilon_t, \tag{8.73}$$

where ε_t is a random variable, having mean zero and variance unity, independent of $\{X_{t-1}, X_{t-2}, \cdots\}$. Hence, the function $f(\cdot)$ is the autoregressive function and $\sigma^2(\cdot)$ is the conditional variance function.

Model (8.73) allows the autoregression and the conditional variance functions to depend on different lag variables. For example, the function $\sigma(\cdot)$ can depend only on the first q lag variables, while the function $\sigma(\cdot)$ depends on the last r lag variables. As mentioned in §8.1, the saturated nonparametric model is not very useful due to the curse of dimensionality. Thus, some structures on f and σ should be imposed.

Suppose that we have a method to estimate f, such as those mentioned in §8.3–§8.6. Let $\widehat{f}$ be an estimate and

$$r_t = X_t - \widehat{f}(X_{t-1}, \cdots, X_{t-p}). \tag{8.74}$$

Denote

$$\widehat{f}_t = \widehat{f}(X_{t-1}, \cdots, X_{t-p}), \quad \text{and} \quad f_t = f(X_{t-1}, \cdots, X_{t-p}).$$

Then,

$$\begin{aligned} r_t^2 &= (\widehat{f}_t - f_t)^2 + \sigma^2(X_{t-1}, \cdots, X_{t-p})\varepsilon_t^2 \\ &\quad -2(\widehat{f}_t - f_t)\sigma(X_{t-1}, \cdots, X_{t-p})\varepsilon_t. \end{aligned} \tag{8.75}$$

The cross-product term has mean approximately zero. The local average of this cross-product term is usually of the same order as that of the first term. Suppose that $\widehat{f}_t - f_t = O_p(b_T)$ for some $b_T \to 0$. Then, the errors in estimation of f affect the estimation of $\sigma^2(\cdot)$ only in $O_P(b_T^2)$. This is negligible in many cases. Consequently, without knowing the function f, we can estimate $\sigma^2(\cdot)$ as well as if f were known-that is, as well as that based on ideal data $\{X_t - f_t\}$. The conditional variance estimator based on these ideal data is referred to as the oracle estimator.

To elucidate the heuristic argument above, let us consider the case where $p = 1$. In this case, according to Theorem 6.4, $b_T = h^2 + 1/(Th)^{1/2}$, for the local linear fit with bandwidth h. For a large range of choices of h, $b_T^2 = o(T^{-2/5})$ or even $b_T^2 = o(T^{-1/2})$. Yet, the function $\sigma^2(\cdot)$ can only be

estimated at rate $O(T^{-2/5})$ or $O(T^{-1/2})$ if $\sigma(\cdot)$ is a constant. Thus, the error $O(b_T^2)$ is negligible for estimating $\sigma^2(\cdot)$.

The heuristic argument above also shows that the choice of smoothing parameter for the regression f is not very critical for estimating $\sigma(\cdot)$ since, for a large range of bandwidths, the error in estimating $\widehat{f}$ is negligible. The heuristic argument applies to other models such as the FAR-model and AAR-model.

By (8.75), we have the following model:

$$r_t^2 \approx \sigma^2(X_{t-1}, \cdots, X_{t-p})\varepsilon_t^2. \tag{8.76}$$

It is easy to see from (8.76) that

$$E(r_t^2|X_{t-1}, \cdots, X_{t-p}) \approx \sigma^2(X_{t-1}, \cdots, X_{t-p}). \tag{8.77}$$

Thus $\sigma^2(\cdot)$ can be regarded as the nonparametric regression of r_t^2 on X_{t-1}, $\cdots$, X_{t-p}. The techniques introduced in §8.3–§8.6 continue to apply to model the conditional variance function. In particular, the local least-squares approach can be used to estimate $\sigma^2(\cdot)$.

Another technique is the local *pseudolikelihood* method. Regarding $\varepsilon_t \sim N(0,1)$ and after ignoring the errors in the estimation of $\widehat{f}$, the conditional density of r_t given $X_{t-1}, \cdots, X_{t-p}$ is

$$\frac{1}{\sqrt{2\pi}\sigma(X_{t-1}, \cdots, X_{t-p})} \exp\{-r_t^2/2\sigma^2(X_{t-1}, \cdots, X_{t-p})\}.$$

Taking the logarithm and adding the likelihood, we have the following *pseudolikelihood*

$$\sum_{t=p+1}^{T} \{-\log \sigma^2(X_{t-1}, \cdots, X_{t-p}) - r_t^2/\sigma^2(X_{t-1}, \cdots, X_{t-p})\}. \tag{8.78}$$

Here, we have dropped a constant factor and multiplied the logarithm of the likelihood by a factor of 2. This does not affect our estimate of $\sigma(\cdot)$.

The local least-squares method and the local pseudolikelihood approach [a local version of (8.78), see (8.83)] are two popular procedures for estimating conditional variance. They can be applied to nonparametric estimation of $\sigma^2(\cdot)$ by either local modeling or global modeling of the conditional variance $\sigma^2(\cdot)$.

8.7.2 Univariate Setting

We now use the univariate setting to illustrate the aforementioned idea. The idea is also applicable to a more general stochastic regression setting.

Let $\{(X_t, Y_t) : t = 1, \cdots, n\}$ be a strictly stationary process from the model

$$Y_t = f(X_t) + \sigma(X_t)\varepsilon_t. \tag{8.79}$$

The autoregressive model (8.73) with $p = 1$ corresponds to the setting with $Y_t = X_{t+1}$, $n = T-1$, and a slight rearrangement of the index. Let $\widehat{f}$ be the local linear estimator, using the data $\{(X_t, Y_t) : t = 1, \cdots, n\}$, with kernel K and bandwidth h_2 (see §6.3.2). By applying the nonparametric regression approach (8.77) using the local linear estimator with the kernel W and bandwidth h_1, we obtain an estimate of $\widehat{\sigma}^2(x)$ for each given x; namely, $\sigma^2(x)$ is obtained by the local linear fit of the data $\{(X_t, r_t^2) : t = 1, \cdots, n\}$. The following result was proved by Fan and Yao (1998).

Theorem 8.5 *Suppose that Condition 3 in §8.8.6 is met. Then,*

$$(nh_1)^{\frac{1}{2}}\{\widehat{\sigma}^2(x) - \sigma^2(x) - \theta_n\}$$

is asymptotically normal with mean 0 and variance

$$p^{-1}(x)\sigma^4(x)\lambda^2(x)\int W^2(t)dt,$$

where $p(\cdot)$ denotes the marginal density function of X, $\lambda^2(x) = E\{(\varepsilon^2 - 1)^2 | X = x\}$, and

$$\theta_n = \frac{h_1^2}{2}\sigma_W^2\ddot{\sigma}^2(x)\int t^2 W(t)dt + o(h_1^2 + h_2^2)$$

with $\ddot{\sigma}^2(x)$ being the second derivative of the function $\sigma^2(x)$.

The result above reveals that, as long as $h_2 = O(h_1)$, the estimator $\widehat{\sigma}^2(\cdot)$ performs as well as the *oracle estimator* where the function f is known. The condition that f is twice differentiable is not minimal. The function $\sigma(\cdot)$ can be estimated with optimal rates under weaker smoothness conditions on $m(\cdot)$; see Hall and Carroll (1989) and Müller and Stadtmüller (1993).

Theorem 8.5 permits us to take advantage of existing bandwidth selection methods for the local linear fit (see §6.3.5). Let $\widehat{h}(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$ be a data-driven bandwidth selection rule for local linear regression based on the data $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$. The selected bandwidth

$$\widehat{h}(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$$

is usually a consistent estimate of the asymptotic optimal bandwidth, which is of order $O(n^{-1/5})$. Our bandwidth selection rule for estimating conditional variance consists of the following steps:

1. Use bandwidth $h_2 = \widehat{h}(X_1, \cdots, X_n; Y_1, \cdots, Y_n)$ in local linear regression to obtain the estimate $\widehat{f}(X_i)$ for $i = 1, \cdots, n$.

2. Compute squared residuals $\widehat{r}_i^2 = \{Y_i - \widehat{f}(X_i)\}^2$, $i = 1, \cdots, n$.

3. Apply bandwidth $h_1 = \widehat{h}(X_1, \cdots, X_n; \widehat{r}_1^2, \cdots, \widehat{r}_n^2)$ in local linear regression to obtain $\widehat{\sigma}^2(\cdot)$.

In the algorithm above, we keep the bandwidth selection method flexible. In our numerical implementation below, we use the preasymptotic substitution method of Fan and Gijbels (1995) (see also §6.3.5). We have written the C-code "autovar.c" for automatic estimation of the variance function.

We now illustrate the technique above with two numerical examples.

Example 8.15 (*Interest rate data*) This example concerns the yields of the three-month Treasury bill presented in Example 1.3. Let R_t be the yield of the three-month Treasury bill at time t. There is much literature on modeling the dynamics of the short-term rates. Famous models include the Vasicek (1977) model

$$dR_t = (\alpha_0 + \alpha_1 R_t)dt + \sigma dW_t,$$

the Cox, Ingersoll, and Ross (1985) model

$$dR_t = (\alpha_0 + \alpha_1 R_t)\, dt + \sigma R_t^{1/2}\, dW_t,$$

and the parametric model by Chan, Karolyi, Longstaff, and Sanders (1992),

$$dR_t = (\alpha_0 + \alpha_1 R_t)\, dt + \sigma R_t^{\gamma}\, dW_t,$$

where $\{W_t\}$ is a standard one-dimensional *Brownian motion*. By a Brownian motion, we mean a zero-mean Gaussian process starting at zero with covariance function $EW(t)W(\tau) = \min(t, \tau)$. These time-homogeneous models are specific cases of the following nonparametric model:

$$dR_t = \mu(R_t)dt + \sigma(R_t)dW_t. \tag{8.80}$$

The function $\mu(\cdot)$ is called the *instantaneous return,* and $\sigma(\cdot)$ is frequently referred to as a volatility function.

The nonparametric model (8.80) has been studied by Stanton (1997), Fan and Yao (1998), Chapman and Pearson (2000), and Fan and Zhang (2003), among others. There are also many time-inhomogeneous models; see, for example, Black, Derman, and Toy (1990), Hull and White (1990), Black, and Karasinski (1991), and Fan, Jiang, Zhang, and Zhou (2003).

Let $\{X_i,\ i = 1, \ldots, n+1\}$ be the observed yields of the three-month Treasury Bill at discrete time points: $t_1 < \cdots < t_n < t_{n+1}$; namely, $X_i = R_{t_i}$. For the weekly data, the time unit is years, with $t_i = t_0 + \frac{i}{52}$. Denote

$$Y_i = X_{i+1} - X_i,\ Z_i = W_{t_{i+1}} - W_{t_i} \text{ and } \triangle_i = t_{i+1} - t_i.$$

By the independent increments of the Brownian motion, $\{Z_i\}$ are independently distributed as $N(0, \triangle_i)$. The discretized version of (8.72) can be expressed as

$$\begin{aligned} Y_i &\approx \mu(X_i)\Delta_i + \sigma(X_i)Z_i \\ &= \mu(X_i)\Delta_i + \sigma(X_i)\sqrt{\Delta_i}\varepsilon_i, \quad i = 1, \ldots, n, \end{aligned} \tag{8.81}$$

where $\{\varepsilon_i\}$ are independently distributed as $N(0,1)$. As pointed out in Chan, Karolyi, Longstaff, and Sanders (1992) and demonstrated by Stanton (1997), the discretized approximation error to the continuous-time model is of second order when the data are observed within a short time gap. Indeed, according to Stanton (1997), as long as data are sampled monthly or more frequently, the errors introduced by using approximations rather than true drift and diffusion are extremely small when compared with the likely size of estimation errors.

The difference data Y_i are plotted against X_i in Figure 8.17(a), where the estimated drift function, namely the regression function, is also displayed. Figure 8.17(b) shows the estimated drift function with its pointwise 95% confidence intervals. The bandwidth $\widehat{h}_2 = 2.61$ was chosen by the program "autovar.c." The nonlinear appearance led Stanton (1997) to speculate that the drift function is nonlinear. Based on empirical simulation studies, Chapman and Pearson (2000) suggested that the nonlinearity might be spurious due to the boundary effect of kernel estimators and the "mean-reversion" of the interest rate dynamic. A formal statistical test of whether the drift function is nonlinear is presented in Fan and Zhang (2002) (see also Chapter 9).

Displayed in Figure 8.17(c) are $\widehat{\sigma}(\cdot)$ and $\widehat{\sigma}^2(\cdot)$ and their associated pointwise confidence intervals. The bandwidth $\widehat{h}_1 = 3.16$ was chosen by the data. The correlation between $\log(x)$ and $\log\{\widehat{\sigma}(x)\}$ at 101 grid points is 0.93. This suggests that we fit the parametric model

$$\sigma(x) = \alpha x^{\beta}.$$

Substituting the parametric form into (8.78), we obtain the pseudolikelihood

$$\ell(\alpha, \beta) = \sum_{i=1}^{n} \{-\log(\alpha^2 X_i^{2\beta}) - r_i^2/(\alpha^2 X_i^{2\beta})\}, \tag{8.82}$$

where $r_i = Y_i - \widehat{\mu}(X_i)\Delta_i$. For each given β, the maximum is obtained at

$$\widehat{\alpha}^2(\beta) = n^{-1} \sum_{i=1}^{n} \frac{r_i^2}{X_i^{2\beta}}.$$

Thus, we need only to maximize the univariate function $\ell(\widehat{\alpha}(\beta), \beta)$. The results are $\widehat{\beta} = 0.90$ and $\widehat{\alpha} = 0.046$. The corresponding parametric function is displayed in Figure 8.17(c). The figure suggests that the parametric fit does not accurately capture the curvature of the nonparametric fit. ∎

Example 8.16 (*Motorcycle data*) Presented in Figure 8.18(a) are 133 observations of motorcycle data from Schmidt, Mattern, and Schüler (1981). The time (in milliseconds) after a simulated impact on motorcycles was recorded and serves as the covariate X_t. The response variable Y_t is the

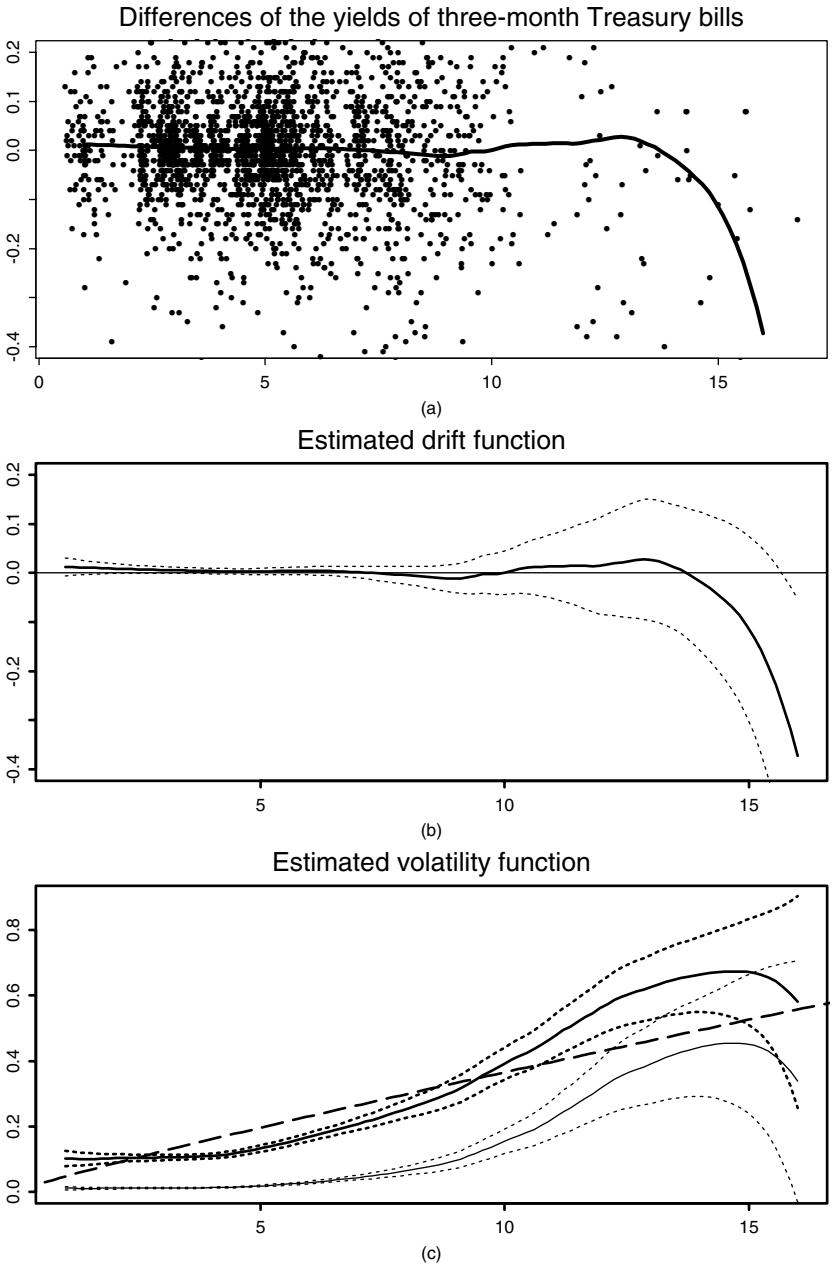


FIGURE 8.17. Three-month Treasury bill data. (a) The difference Y_i is plotted against X_i, the solid curve being the function $\mu(\cdot)/52$. (b) The estimated instantaneous return function in (a). (c) The estimated volatility function $\sigma(\cdot)/\sqrt{52}$ (thick curve), the conditional variance function $\sigma^2(\cdot)/52$ (thin curve), and the parametric fit of $\sigma(\cdot)/\sqrt{52}$ (long-dashed curve).

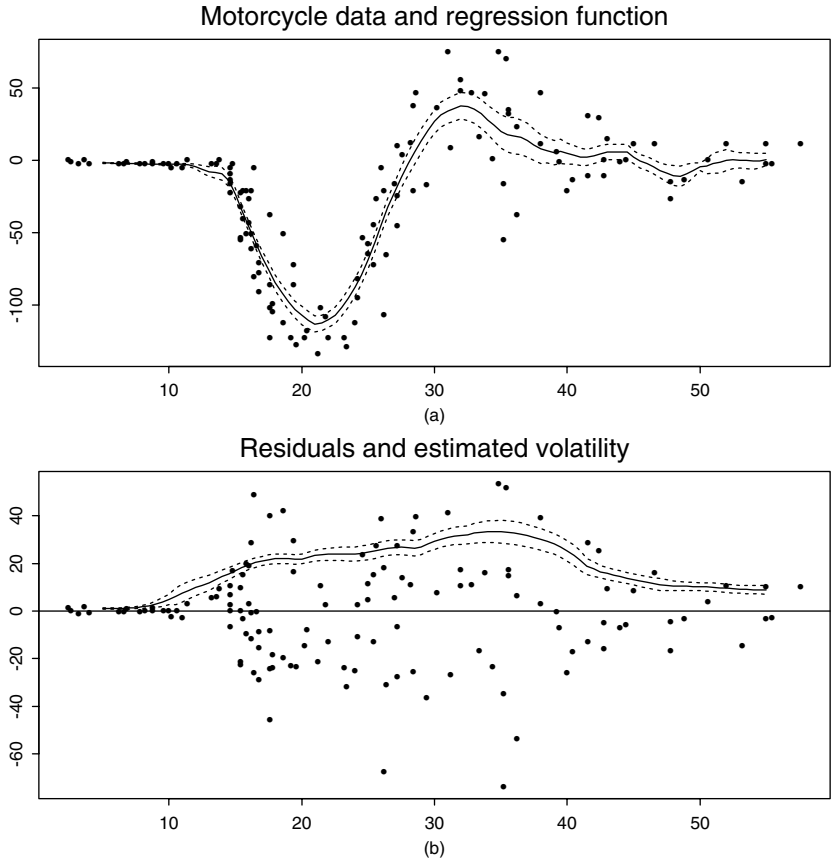


FIGURE 8.18. Motorcycle data. (a) Raw data and their estimated regression function. (b) The residuals and the estimated volatility function.

head acceleration (in a) of a test object. We fit the data with model (8.79). The estimated regression function $\widehat{f}(\cdot)$ is depicted in Figure 8.18(a) along with 95% pointwise confidence intervals. Figure 8.18(b) describes the residuals and the estimated conditional standard deviation $\widehat{\sigma}(\cdot)$. The bandwidths selected by data for estimating the regression function and the conditional variance function are 3.230 and 6.293, respectively. Figure 8.18(b) shows that $\widehat{\sigma}^2(\cdot)$ captures the changes of variability in the data. ∎

Finally, we now briefly discuss how to employ the pseudolikelihood method in estimating the volatility. For each given x and for every $X_t \approx x$, approximate the conditional variance function locally by a linear function

$$\sigma^2(X_t) \approx \alpha + \beta(X_t - x).$$

Using the pseudolikelihood (8.78) locally, we obtain the local pseudolikelihood

$$\sum_{i=1}^{n}\{-\log(\alpha+\boldsymbol{\beta}(X_t-x))-r_t^2/(\alpha+\boldsymbol{\beta}(X_t-x))\}W_{h_1}(X_t-x). \quad (8.83)$$

Maximizing (8.83) with respect to α and β yields the estimate $\widehat{\sigma}^2(x)=\widehat{\alpha}$. This is the maximum local pseudo-likelihood estimator.

8.7.3 Functional-Coefficient Models

As mentioned in §8.7.1, for the multivariate case, one needs to impose the structure on the volatility function. For example, one can impose the form

$$\sigma^2(X_{t-1},\cdots,X_{t-p})=a_1(X_{t-d})X_1^2+\cdots+a_p(X_{t-d})X_{t-p}^2. \quad (8.84)$$

This is analogous to the FAR(p) model in (8.1). By regarding $\sigma^2(\cdot)$ as the regression function of r_t^2 as in (8.77), one can apply the techniques in §8.3 to estimate the coefficient functions $a_1(\cdot),\cdots,a_p(\cdot)$, using the data

$$\{(X_{t-d},X_{t-p}^2,\cdots,X_{t-1}^2,r_t^2):t=p+1,\cdots,T\}.$$

Note that the model (8.84) is a generalization of the ARCH(p) model with the coefficient allowed to vary.

8.7.4 Additive Models

A useful application of the additive model in the regression setting leads to the additive model

$$\sigma^2(X_{t-1},\cdots,X_{t-p})=\sigma_1(X_{t-1}^2)+\cdots+\sigma_p(X_{t-p}^2). \quad (8.85)$$

By applying the backfitting algorithm or other techniques in §8.5 to the data

$$\{(X_{t-p}^2,\cdots,X_{t-1}^2,r_t^2):t=p+1,\cdots,T\},$$

one can easily obtain an estimate of the functions $\sigma_1(\cdot),\cdots,\sigma_p(\cdot)$.

Note that model (8.85) is a generalization of the ARCH(p) model. Thus, it allows one to examine whether an ARCH model adequately fits a given data set.

Example 8.17 (*Standard and Poor's 500 Index*). We revisit the data analyzed in Example 8.14. As in that example, let r_t be the observed return at time t. Similarly to that in Example 8.15, the conditional mean of r_t is negligible. As an illustration, we fit the additive model

$$r_t^2=\{\mu+\sigma_1(r_{t-1}^2)+\cdots+\sigma_{t-6}(r_{t-6}^2)\}\varepsilon_t^2$$

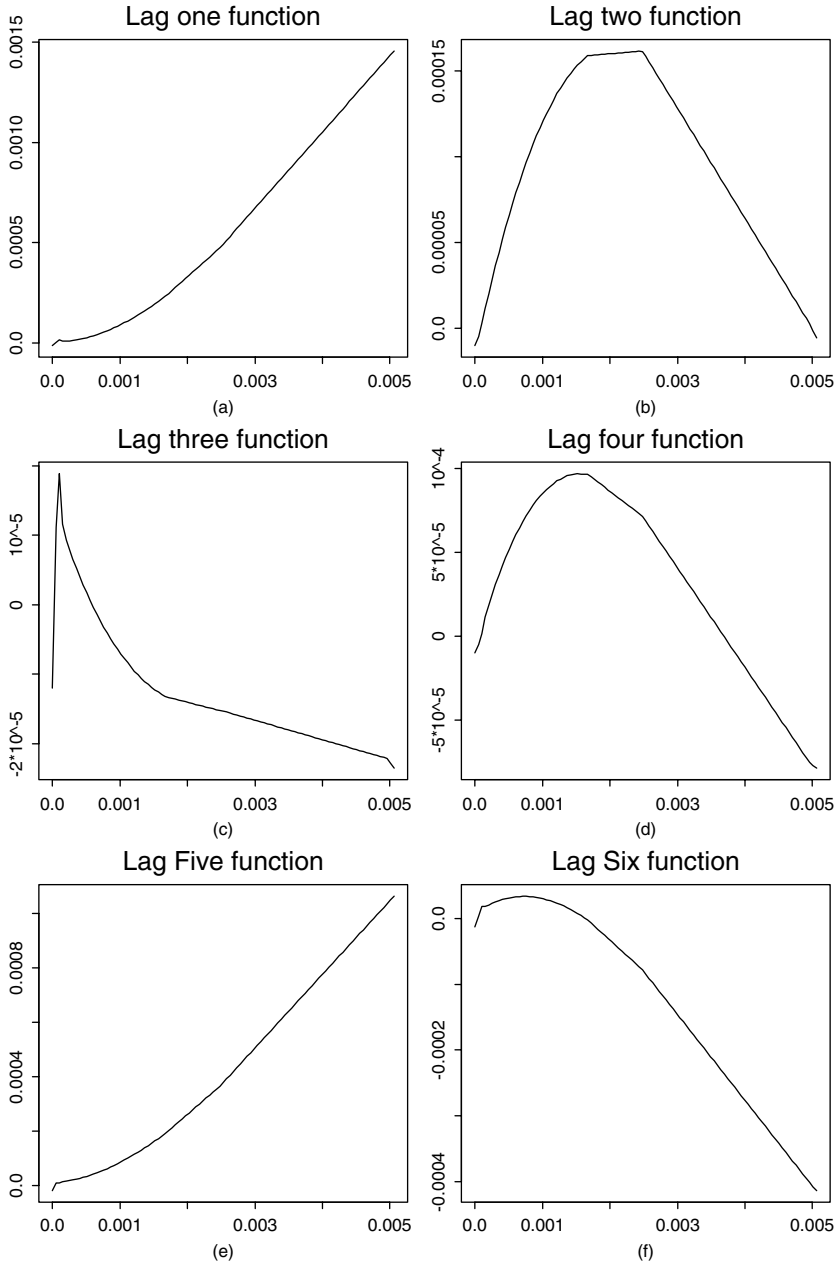


FIGURE 8.19. Fitted additive functions for the conditional variance based on the Standard and Poor's 500 Index.

using the data from January 3, 1990 to December 31, 1999, a series of length 2529. The fitted functions are depicted in Figure 8.19. The nonlinearity of fitted functions is very apparent. This suggests that an ARCH(6) model is inadequate for fitting the volatility function. ∎

8.7.5 Product Models

The models (8.84) and (8.85) can possibly result in a negative fit of the variance function. One way to avoid this is to impose the models on $\log \sigma^2(\cdot)$. For example, one can impose the model

$$\log\{\sigma^2(X_{t-1}, \cdots, X_{t-p})\} = a_1(X_{t-d})X_1 + \cdots + a_p(X_{t-d})X_{t-p} \quad (8.86)$$

or the model

$$\log\{\sigma^2(X_{t-1}, \cdots, X_{t-p})\} = \sigma_1(X_{t-1}^2) + \cdots + \sigma_p(X_{t-p}^2). \quad (8.87)$$

From the scale model (8.76), we have

$$\log r_t^2 \approx \log \sigma^2(X_{t-1}, \cdots, X_{t-p}) + E \log \varepsilon_t^2 + \varepsilon_t',$$

where $\varepsilon_t' = \log \varepsilon_t^2 - \log \varepsilon_t^2$ is a random noise with mean zero. If $\varepsilon_t \sim N(0, 1)$, then $E \log \varepsilon_t^2 \approx -1.27$. Thus, models (8.86) and (8.87) correspond to the functional-coefficient model and the additive model for the dependent variable $\log r_t^2$, respectively.

The localization idea in §8.3 and the backfitting algorithm in §8.5 can be directly employed to estimate the one-dimensional functions in (8.86) and (8.87), respectively. Use of the pseudolikelihood can be more effective than the least-squares approach.

8.7.6 Other Nonparametric Models

As shown in (8.77), the conditional variance function is nothing but a mean regression function. Thus, nonparametric techniques in §8.6 for modeling the mean function can be applied to model the variance function. We do not pursue this general idea further here.

8.8 Complements

8.8.1 Proof of Theorem 8.1

For a more general autoregressive model

$$X_t = h(X_{t-1}, X_{t-2}, \cdots, X_{t-p}) + \varepsilon_t,$$

where ε_t is an i.i.d. sequence, it can also be expressed in a form of (8.5). The following lemma is due to Chan and Tong (1985).

Lemma 8.3 *The Markov chain $\{\mathbf{X}_t\}$ is aperiodic and ϕ-irreducible with ϕ being the Lebesgue measure if ε_t has an absolutely continuous component with a positive density everywhere and $h(\cdot)$ is bounded over a bounded set.*

Let $\|\mathbf{x}\|$ be the Euclidean norm

$$\|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_p^2}$$

and $\mathbf{C}$ be the matrix that is similar to the matrix $\mathbf{A}(\mathbf{X})$, except that the element $a_i(\cdot)$ is replaced by c_i.

Proof of Theorem 8.1. First, by Lemma 8.3, the chain $\{\mathbf{X}_t\}$ is aperiodic and ϕ-irreducible. Thus, we can apply Lemmas 8.1 and 8.2. To check the moment condition in Lemma 8.1, we first give a bound on $\mathbf{C}^n$.

Let $\mathbf{I}_p$ be the identity matrix of order p. Then, the determinant of $\lambda\mathbf{I} - \mathbf{C}$ is given by (see, for example, Anderson 1971, p. 180)

$$|\lambda\mathbf{I} - \mathbf{C}| = \lambda^p - c_1\lambda^{p-1} - \cdots - c_p.$$

Hence, all of the roots of the characteristic function are eigenvalues of the matrix $\mathbf{C}$. Let $\lambda_{\max}$ be the maximum eigenvalue of $\mathbf{C}$. Then $\|\mathbf{C}^n\|^{1/n} \to |\lambda_{\max}| < 1$. Consequently, there exists a positive constant $\delta < 1$ and an integer m such that $\|\mathbf{C}^m\| < \delta$.

We now verify the moment condition of Lemma 8.1 for the subchain $\{\mathbf{X}_{mt}, t = 1, 2, \cdots\}$. Using the iterative formula (8.5), we have

$$\mathbf{X}_{m(t+1)} = \prod_{i=0}^{m-1} \mathbf{A}(\mathbf{X}_{mt+i})\mathbf{X}_{mt} + \sum_{i=1}^{m} \left[\prod_{j=i}^{m-1} \mathbf{A}(\mathbf{X}_{mt+j})\right] \boldsymbol{\varepsilon}_{mt+i}. \tag{8.88}$$

For any vector $\mathbf{b} = (b_1, \cdots, b_p)^T$, let $(d_1, \cdots, d_p)^T = \mathbf{A}(\mathbf{X})\mathbf{b}$. Then

$$\begin{aligned} |d_1| &= |a_1(\mathbf{X})b_1 + \cdots + a_p(\mathbf{X})b_p| \\ &\leq c_1|b_1| + \cdots + c_p|b_p| \end{aligned}$$

and $|d_i| = |b_i|$ for $i = 2, \cdots, p$. Consequently, with the vector $|\mathbf{b}| = (|b_1|, \cdots, |b_p|)^T$, we have

$$\|A(\mathbf{X})\mathbf{b}\| \leq \|\mathbf{C}|\mathbf{b}|\|.$$

Repeatedly applying this to (8.88), we obtain

$$\|\mathbf{X}_{(m+1)t}\| \leq \left\|\mathbf{C}^m|\mathbf{X}_{mt}|\right\| + \left\|\sum_{i=1}^{m} \mathbf{C}^{m-i}|\boldsymbol{\varepsilon}_{mt+i}|\right\|.$$

The first term is bounded by

$$\|\mathbf{C}^m\|\ \|\mathbf{X}_{mt}\| \leq \delta\|\mathbf{X}_{mt}\|.$$

Hence

$$E\left\{\|\mathbf{X}_{m(t+1)}\|\Big|\mathbf{X}_{mt}=\mathbf{x}\right\}\le\delta\|\mathbf{x}\|+E\left\|\sum_{i=1}^{m}\mathbf{C}^{m-i}|\boldsymbol{\varepsilon}_{mt+i}|\right\|.$$

Note that the second term is bounded and is independent of $\mathbf{x}$. Let D denote the bound. Then

$$E\left\{\|\mathbf{X}_{(m+1)t}\|\Big|\mathbf{X}_{mt}=\mathbf{x}\right\}\le\delta\|\mathbf{x}\|+D.$$

Let $\rho\in(\delta,1)$ and set $M=D(\rho-\delta)^{-1}$. Then, for all $\|\mathbf{x}\|>M$,

$$E\left\{\|\mathbf{X}_{m(t+1)}\|\Big|\mathbf{X}_{mt}\right\}\le\rho\|\mathbf{x}\|.$$

Hence, by Lemma 8.1, with $K=\{\mathbf{x}:\|\mathbf{x}\|\le M\}$, the sequence $\{\mathbf{X}_{mt}\}$ is geometrically ergodic. As a result, the original sequence $\{\mathbf{X}_t\}$ is geometrically ergodic by Lemma 8.2. ∎

8.8.2 Technical Conditions for Theorems 8.2 and 8.3

Conditions 1:

(i) The kernel function $K(\cdot)$ is a bounded density with the bounded support $[-1,\ 1]$. Furthermore, $\int_{-\infty}^{+\infty}uK(u)du=0$.

(ii) $|f(u,\,v\,|\,\mathbf{x}_0,\,\mathbf{x}_1;\,l)|\le M<\infty$, for all $l\ge1$, where $f(u,\,v,\,|\,\mathbf{x}_0,\,\mathbf{x}_1;\,l)$ is the conditional density of $(U_0,\,U_l)$ given $(\mathbf{X}_0,\,\mathbf{X}_l)$, and $f(u\,|\,\mathbf{x})\le M<\infty$, where $f(u\,|\,\mathbf{x})$ is the conditional density of U given $\mathbf{X}=\mathbf{x}$.

(iii) The process $\{(U_i,\,\mathbf{X}_i,\,Y_i)\}$ is α-mixing with $\sum\ell^c[\alpha(\ell)]^{1-2/\delta}<\infty$ for some $\delta>2$ and $c>1-2/\delta$.

(iv) $E|\mathbf{X}|^{2\,\delta}<\infty$, where δ is given in Condition 1(iii).

Note that the conditions imposed here are just for the convenience of technical derivations. They are not the minimum possible. For example, the requirement on the bounded support of the kernel K can be relaxed at the expense of lengthier proofs. In particular, the Gaussian kernel is allowed.

Conditions 2:

(i) For all $l\ge1,\ \mathbf{x}_0,\ \mathbf{x}_1\in\Re^p,\ u$ and v in a neighborhood of u_0,

$$E\left\{Y_0^2+Y_l^2\,|\,\mathbf{X}_0=\mathbf{x}_0,\,U_0=u;\,\mathbf{X}_l=\mathbf{x}_1,\,U_l=v\right\}\le M<\infty.$$

(ii) There exists a sequence of positive integers s_n such that $s_n \to \infty$, $s_n = o\left((n\,h_n)^{1/2}\right)$, and $(n/h_n)^{1/2}\,\alpha(s_n) \to 0$, as $n \to \infty$.

(iii) There exists a positive constant $\delta^* > \delta$, where δ is given in Condition 1(iii), such that

$$E\left\{|Y|^{\delta^*}\,|\mathbf{X} = \mathbf{x},\, U = u\right\} \leq M < \infty$$

for all $\mathbf{x} \in \Re^p$ and u in a neighborhood of u_0, and $\alpha(n) = O\left(n^{-\theta^*}\right)$ for some $\theta^* \geq \delta\,\delta^*/\{2(\delta^* - \delta)\}$.

(iv) $E|\mathbf{X}|^{2\,\delta^*} < \infty$, and $n^{1/2-\delta/4}\,h^{\delta/\delta^*-1} = O(1)$.

We now provide a sufficient condition for the mixing coefficient $\alpha(n)$ to satisfy Conditions 1(iii) and 2(ii). Suppose that $h_n = A\,n^{-a}$ for some $a \in (0,1)$ and $A > 0$, $s_n = (n\,h_n/\log n)^{1/2}$, and $\alpha(n) = O\left(n^{-d}\right)$ for some $d > 0$. Then Condition 1(iii) is satisfied for $d > 2(\delta-1)/(\delta-2)$, and Condition 2(ii) is satisfied if $d > (1+a)/(1-a)$. Hence, both conditions are satisfied if

$$\alpha(n) = O\left(n^{-d}\right), \qquad d > \max\left\{\frac{1+a}{1-a},\ \frac{2(\delta-1)}{\delta-2}\right\}.$$

Note that the larger the order δ, the weaker the decay rate of $\alpha(n)$. This is a trade-off between the order δ of the moment of Y and the rate of decay of the mixing coefficient.

8.8.3 Preliminaries to the Proof of Theorem 8.3

To study the joint asymptotic normality of $\widehat{\mathbf{a}}(u_0)$, we need to center the vector $\mathbf{T}_n(u_0)$ by replacing Y_i with $Y_i - m(\mathbf{X}_i, U_i)$ in the expression of $\mathbf{T}_{n,j}(u_0)$, where $m(\mathbf{X}_i, U_i) = \mathbf{X}_i^T\mathbf{a}(U_i)$. Let

$$\mathbf{T}^*_{n,j}(u_0) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\left(\frac{U_i - u_0}{h}\right)^j K_h(U_i - u_0)\,[Y_i - m(\mathbf{X}_i, U_i)]$$

and

$$\mathbf{T}^*_n = \begin{pmatrix} \mathbf{T}^*_{n,0} \\ \mathbf{T}^*_{n,1} \end{pmatrix}.$$

For $|U_i - u_0| < h$, by Taylor's expansion,

$$\begin{aligned} m(\mathbf{X}_i, U_i) \;&=\; \mathbf{X}_i^T\,\mathbf{a}(u_0) + (U_i - u_0)\,\mathbf{X}_i^T\,\mathbf{a}'(u_0) \\ &\quad + \frac{h^2}{2}\left(\frac{U_i - u_0}{h}\right)^2 \mathbf{X}_i^T\,\mathbf{a}''(u_0) + o_p(h^2), \end{aligned}$$

where $\mathbf{a}'(u_0)$ and $\mathbf{a}''(u_0)$ are the vectors consisting of the first and the second derivatives of the functions $a_j(\cdot)$. Thus

$$\mathbf{T}_{n,0} - \mathbf{T}^*_{n,0} = \mathbf{S}_{n,0}\,\mathbf{a}(u_0) + h\,\mathbf{S}_{n,1}\,\mathbf{a}'(u_0) + \frac{h^2}{2}\,\mathbf{S}_{n,2}\,\mathbf{a}''(u_0) + o_p(h^2)$$

and

$$\mathbf{T}_{n,1} - \mathbf{T}^*_{n,1} = \mathbf{S}_{n,1}\,\mathbf{a}(u_0) + h\,\mathbf{S}_{n,2}\,\mathbf{a}'(u_0) + \frac{h^2}{2}\,\mathbf{S}_{n,3}\,\mathbf{a}''(u_0) + o_p(h^2).$$

As a result, we have

$$\mathbf{T}_n - \mathbf{T}^*_n = \mathbf{S}_n\,\mathbf{H}\,\boldsymbol{\beta} + \frac{h^2}{2}\begin{pmatrix}\mathbf{S}_{n,2}\\ \mathbf{S}_{n,3}\end{pmatrix}\mathbf{a}''(u_0) + o_p(h^2),$$

where $\boldsymbol{\beta} = (\mathbf{a}(u_0)^T,\, \mathbf{a}'(u_0)^T)^T$. Therefore, it follows from (8.21) and Theorem 8.2 that

$$\begin{aligned}\mathbf{H}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) &= \mathbf{S}_n(\mathbf{T}^*_n + \mathbf{T}_n - \mathbf{T}^*_n) - \mathbf{H}\boldsymbol{\beta}\\ &= f_U^{-1}(u_0)\,\mathbf{S}^{-1}\,\mathbf{T}^*_n + \frac{h^2}{2}\,\mathbf{S}^{-1}\begin{pmatrix}\mu_2\,\Omega\\ \mu_3\Omega\end{pmatrix}\mathbf{a}''(u_0) + o_p(h^2).\end{aligned}$$

Using $\mathbf{S}^{-1} = \mathrm{diag}(1, \mu_2^{-1}) \otimes \Omega^{-1}$, we have

$$\widehat{\mathbf{a}}(u_0) - \mathbf{a}(u_0) = \frac{\Omega^{-1}}{f_U(u_0)}\mathbf{T}^*_{n,0} + \frac{h^2}{2}\mu_2\,\mathbf{a}''(u_0) + o_p(h^2). \tag{8.89}$$

Expression (8.89) indicates that the asymptotic bias of $\widehat{\mathbf{a}}(u_0)$ is $\frac{h^2}{2}\,\mu_2\,\mathbf{a}''(u_0)$. Let

$$\mathbf{Q}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i,$$

where

$$\mathbf{Z}_i = \mathbf{X}_i K_h(U_i - u_0)\,[Y_i - m(\mathbf{X}_i,\, U_i)]\,.$$

It follows from this and (8.89) that

$$\begin{aligned}&\sqrt{n\,h_n}\left[\widehat{\mathbf{a}}(u_0) - \mathbf{a}(u_0) - \frac{h^2}{2}\mu_2\mathbf{a}''(u_0) + o(h^2)\right]\\ =\;& \frac{\Omega^{-1}}{f_U(u_0)}\,\sqrt{n\,h_n}\,\mathbf{Q}_n.\end{aligned}$$

Thus, the main task becomes establishing the asymptotic normality of $\mathbf{Q}_n$. To this end, we need the following lemma.

Lemma 8.4 *Under Conditions 1 and 2 and the assumption that $h_n \to 0$ and $n\,h_n \to \infty$, as $n \to \infty$, if $\sigma^2(\mathbf{x},\, u)$ and $f(\mathbf{x},\, u)$ are continuous at the point u_0, then we have*

(a) $h_n \operatorname{Var}(\mathbf{Z}_1) \to \nu_0 f_U(u_0)\,\Omega^*(u_0)$;

(b) $h_n \sum_{l=1}^{n-1} |\operatorname{Cov}(\mathbf{Z}_1, \mathbf{Z}_{l+1})| = o(1)$;

(c) $n\,h_n \operatorname{Var}(\mathbf{Q}_n) \to \nu_0 f_U(u_0)\,\Omega^*(u_0)$.

Proof: Let C be a generic constant, which may take different values at different places. First, by conditioning on $(\mathbf{X}_1, U_1)$, we have

$$\begin{aligned}\operatorname{Var}(\mathbf{Z}_1) &= E\left[\mathbf{X}_1\,\mathbf{X}_1^T\,\sigma^2(\mathbf{X}_1, U_1)K_h^2(U_1 - u_0)\right] \\ &= \frac{1}{h}\left[f_U(u_0)\,\Omega^*(u_0) + o(1)\right].\end{aligned}$$

The result (c) follows trivially from (a) and (b) along with

$$\operatorname{Var}(\mathbf{Q}_n) = \frac{1}{n}\operatorname{Var}(\mathbf{Z}_1) + \frac{2}{n}\sum_{l=1}^{n-1}\left(1 - \frac{l}{n}\right)\operatorname{Cov}(\mathbf{Z}_1, \mathbf{Z}_{l+1}).$$

Therefore, it remains to prove part (b). To this end, let $d_n \to \infty$ be a sequence of positive integers such that $d_n\,h_n \to 0$. Define

$$J_1 = \sum_{l=1}^{d_n - 1} |\operatorname{Cov}(\mathbf{Z}_1, \mathbf{Z}_{l+1})| \qquad \text{and} \qquad J_2 = \sum_{l=d_n}^{n-1} |\operatorname{Cov}(\mathbf{Z}_1, \mathbf{Z}_{l+1})|.$$

It remains to show that $J_1 = o\left(h^{-1}\right)$ and $J_2 = o\left(h^{-1}\right)$.

We remark that since $K(\cdot)$ has a bounded support $[-1,\ 1]$, $a_j(u)$ is bounded in the neighborhood of $u \in [u_0 - h,\ u_0 + h]$. Let

$$B = \max_{1 \le j \le p} \sup_{|u - u_0| < h} |a_j(u)| \quad \text{and} \quad \|\mathbf{x}\|_1 = \sum_{j=1}^{p} |x_j|.$$

Then $\sup_{|u-u_0|<h} |m(\mathbf{x}, u)| \le B\|\mathbf{x}\|_1$. By conditioning on $(\mathbf{X}_1, U_1)$ and $(\mathbf{X}_{l+1}, U_{l+1})$, and using Conditions 1(ii) and 2(i), we have, for all $l \ge 1$,

$$\begin{aligned}& |\operatorname{Cov}(\mathbf{Z}_1, \mathbf{Z}_{l+1})| \\ \le\ & C\,E\Big[\left|\mathbf{X}_1\,\mathbf{X}_{l+1}^T\right| \{|Y_1| + B\|\mathbf{X}_1\|_1\}\{|Y_{l+1}| + B\|\mathbf{X}_{l+1}\|_1\} \\ & \qquad K_h(U_1 - u_0)\,K_h(U_{l+1} - u_0)\Big] \\ \le\ & C\,E\Big[\left|\mathbf{X}_1\mathbf{X}_{l+1}^T\right| \left\{C + B^2\|\mathbf{X}_1\|_1^2\right\}^{1/2}\left\{C + B^2\|\mathbf{X}_{l+1}\|_1^2\right\}^{1/2} \\ & \qquad K_h(U_1 - u_0)K_h(U_{l+1} - u_0)\Big] \\ \le\ & C\,E\left[\left|\mathbf{X}_1\,\mathbf{X}_{l+1}^T\right| \{1 + \|\mathbf{X}_1\|_1\}\{1 + \|\mathbf{X}_{l+1}\|_1\}\right] \\ \le\ & C.\end{aligned}$$

Here, the finite fourth moment of $\mathbf{X}$ follows from Condition 2(iv). It follows that

$$J_1 \leq C\, d_n = o\left(h^{-1}\right)$$

by the choice of d_n.

Next, we consider the upper bound of J_2. To this end, by using Davydov's inequality (Proposition 2.5 with $p = q = \delta$), we obtain, for all $1 \leq j,\ m \leq p$ and $l \geq 1$,

$$|\mathrm{Cov}(Z_{1j},\, Z_{l+1,m})| \leq C\,[\alpha(l)]^{1-2/\delta}\ \left[E|Z_j|^{\delta}\right]^{1/\delta}\ \left[E|Z_m|^{\delta}\right]^{1/\delta}. \tag{8.90}$$

By conditioning on $(\mathbf{X},\, U)$ and using Conditions 1(ii) and 2(iii), we obtain

$$\begin{aligned} E\left[|Z_j|^{\delta}\right] &\leq C\,E\left[|X_j|^{\delta}\, K_h^{\delta}(U-u_0)\left\{|Y|^{\delta} + B^{\delta}\,\|\mathbf{X}\|_1^{\delta}\right\}\right] \\ &\leq C\,E\left[|X_j|^{\delta}\, K_h^{\delta}(U-u_0)\left\{C + B^{\delta}\,\|\mathbf{X}\|_1^{\delta}\right\}\right] \\ &\leq C\,h^{1-\delta}\,E\left[|X_j|^{\delta}\ \left\{C + B^{\delta}\,\|\mathbf{X}\|_1^{\delta}\right\}\right] \\ &\leq C\,h^{1-\delta}. \end{aligned} \tag{8.91}$$

A combination of (8.90) and (8.91) leads to

$$\begin{aligned} J_2 &\leq C\,h^{2/\delta-2}\ \sum_{l=d_n}^{\infty}[\alpha(l)]^{1-2/\delta} \\ &\leq C\,h^{2/\delta-2}\,d_n^{-c}\ \sum_{l=d_n}^{\infty} l^c\,[\alpha(l)]^{1-2/\delta} \\ &= o\left(h^{-1}\right) \end{aligned}$$

by choosing d_n such that $h^{1-2/\delta}\,d_n^c = C$, so that the requirement that $d_n\,h_n \to 0$ is satisfied. ∎

8.8.4 Proof of Theorem 8.3

We employ the small-block and large-block techniques as in the proof of Theorem 6.3. The basic ideas and the techniques are almost identical to that theorem so that the notation introduced there is also used here with the understanding that $T = n$ and $x = u_0$. We apply the Cramér–Wold device to derive the asymptotic normality of $\mathbf{Q}_n$. For any unit vector $\mathbf{c} \in \Re^p$, let $Z_{n,i} = \sqrt{h}\,\mathbf{c}^T\,\mathbf{Z}_{i+1},\ \ i = 0,\, \ldots,\, n-1$. Then

$$\sqrt{n\,h}\ \mathbf{c}^T\,\mathbf{Q}_n = \frac{1}{\sqrt{n}}\ \sum_{i=0}^{n-1} Z_{n,i}.$$

By Lemma 8.4, we have

$$\mathrm{Var}(Z_{n,0}) = \theta^2(u_0)\{1 + o(1)\},$$

where $\theta^2(u_0) = \nu_0 f_U(u_0)\, \mathbf{c}^T\, \Omega^*(u_0)\mathbf{c}$ and

$$\sum_{l=0}^{n-1} |\mathrm{Cov}(Z_{n,0},\, Z_{n,l})| = o(1).$$

Recall in the proof of Theorem 6.3 that

$$\sqrt{n\,h}\; \mathbf{c}^T\, \mathbf{Q}_n = \frac{1}{\sqrt{n}} \left\{ Q'_n + Q''_n + Q'''_n \right\}.$$

Following the same idea as in the proof of Theorem 6.3, we need to verify conditions (6.73)—(6.76).

First, choose the same block size as in the proof of Theorem 6.3. Observe that

$$E\left[Q''_n\right]^2 = \sum_{j=0}^{q-1} \mathrm{Var}(\xi_j) + 2 \sum_{0 \le i < j \le q-1} \mathrm{Cov}(\xi_i,\, \xi_j) \equiv I_1 + I_2.$$

It follows from stationarity and Lemma 8.4 that

$$I_1 = q_n\, \mathrm{Var}(\xi_1) = q_n\, \mathrm{Var} \left(\sum_{j=1}^{s_n} Z_{n,j} \right) = q_n\, s_n\, [\theta^2(u_0) + o(1)].$$

Let $r_j^* = j(r_n + s_n)$, then $r_j^* - r_i^* \ge r_n$ for all $j > i$, and we therefore have

$$\begin{aligned} |I_2| &\le 2 \sum_{0 \le i < j \le q-1} \sum_{j_1=1}^{s_n} \sum_{j_2=1}^{s_n} \left| \mathrm{Cov}(Z_{n, r_i^* + r_n + j_1},\, Z_{n, r_j^* + r_n + j_2}) \right| \\ &\le 2 \sum_{j_1=1}^{n-r_n} \sum_{j_2=j_1+r_n}^{n} \left| \mathrm{Cov}(Z_{n,j_1},\, Z_{n,j_2}) \right|. \end{aligned}$$

By stationarity and Lemma 8.4, we obtain

$$|I_2| \le 2n \sum_{j=r_n+1}^{n} \left| \mathrm{Cov}(Z_{n,1},\, Z_{n,j}) \right| = o(n).$$

Hence

$$\frac{1}{n}\, E[Q''_n]^2 = O\left(q_n\; s_n\; n^{-1}\right) + o(1) = o(1).$$

It follows from stationarity and Lemma 8.4 that

$$\mathrm{Var}\left[Q'''_n\right] = \mathrm{Var} \left(\sum_{j=1}^{n - q_n(r_n + s_n)} Z_{n,j} \right) = O(n - q_n(r_n + s_n)) = o(n).$$

The last two expressions entail (6.73). Similarly, (6.74) can be established using the identical argument as in the proof of Theorem 6.3. As for (6.75), by stationarity, (6.77), and Lemma 8.4, it is easily seen that

$$\frac{1}{n}\sum_{j=0}^{q_n-1} E\left(\eta_j^2\right) = \frac{q_n}{n} E\left(\eta_1^2\right) = \frac{q_n\, r_n}{n} \cdot \frac{1}{r_n} \mathrm{Var}\left(\sum_{j=1}^{r_n} Z_{n,j}\right) \to \theta^2(u_0).$$

This proves (6.75).

It remains to establish (6.76). For this purpose, we employ Theorem 4.1 in Shao and Yu (1996) and Condition 2 to obtain

$$\begin{aligned} & E\left[\eta_1^2\, I\left\{|\eta_1| \geq \varepsilon\, \theta(u_0)\, \sqrt{n}\right\}\right] \\ \leq\ & C\, n^{1-\delta/2}\, E\left(|\eta_1|^{\delta}\right) \\ \leq\ & C\, n^{1-\delta/2}\, r_n^{\delta/2} \left\{E\left(|Z_{n,0}|^{\delta^*}\right)\right\}^{\delta/\delta^*}. \end{aligned} \tag{8.92}$$

As in (8.91),

$$E\left(|Z_{n,0}|^{\delta^*}\right) \leq C\, h^{1-\delta^*/2}.$$

Therefore, by (8.92)

$$E\left[\eta_1^2\, I\left\{|\eta_1| \geq \varepsilon\, \theta(u_0)\, \sqrt{n}\right\}\right] \leq C\, n^{1-\delta/2}\, r_n^{\delta/2}\, h^{(2-\delta^*)\delta/(2\,\delta^*)}.$$

Thus, by using Conditions 2(iii) and (iv), we obtain

$$\frac{1}{n}\sum_{j=0}^{q-1} E\left[\eta_j^2 I\left\{|\eta_j| \geq \varepsilon\, \theta(u_0)\, \sqrt{n}\right\}\right] \leq C\, \gamma_n^{1-\delta/2}\, n^{1/2-\delta/4}\, h_n^{\delta/\delta^*-1} \to 0$$

since $\gamma_n \to \infty$. This completes the proof of the theorem. ∎

8.8.5 Proof of Theorem 8.4

(i) It follows from the ordinary least-squares theory that there exists a minimum value of

$$E\left[\left\{Y - f(X)\right\}^2 \,\middle|\, \boldsymbol{\alpha}^T\mathbf{X} = z\right]$$

over the class of functions of the form $f(\mathbf{x}) = \sum_{i=0}^{p-1} f_i(\boldsymbol{\alpha}^\tau \mathbf{x}) x_i$ with $x_0 = 1$. Let $f_0^*(z), \ldots, f_{p-1}^*(z)$ be the minimizer. Then

$$\left(f_0^*(z), \ldots, f_{p-1}^*(z)\right)^T = \left\{\mathrm{Var}\left(\mathbf{X}^* \,\middle|\, \boldsymbol{\alpha}^T\mathbf{X} = z\right)\right\}^{-1} \mathrm{Cov}\left(\mathbf{X}^*, Y \,\middle|\, \boldsymbol{\alpha}^T\mathbf{X} = z\right).$$

By the continuity assumption, the functions $f_0^*(z), \ldots, f_{p-1}^*(z)$ are continuous in z. It follows immediately from the least-squares theory that

$$E\left\{\left[Y - f_0^*(z) - \sum_{j=1}^{p-1} f_j^*(z) X_j\right]^2 \,\middle|\, \boldsymbol{\alpha}^T\mathbf{X} = z\right\} \leq \mathrm{Var}(Y|\boldsymbol{\alpha}^T\mathbf{X} = z).$$

Consequently,

$$R(\boldsymbol{\alpha}) \equiv E\left\{ Y - f_0^*(\boldsymbol{\alpha}^T\mathbf{X}) - \sum_{j=1}^{p-1} f_j^*(\boldsymbol{\alpha}^T\mathbf{X})X_j \right\}^2$$

is bounded by $\mathrm{Var}(Y)$ and continuous on the compact set $\{\boldsymbol{\alpha} \in R^d : ||\boldsymbol{\alpha}|| = 1\}$. Hence, there exists a $\boldsymbol{\beta}$ in the set above such that $R(\boldsymbol{\alpha})$ obtains its minimum. This establishes (i).

(ii) By letting $x_1 = \cdots = x_{p-1} = 0$ in (8.31), we have $g_0(u) = g(0, \cdots, 0, z/\beta_p)$. Now, letting $\mathbf{u}_j$ be the vector with $x_j = 1$, $x_p = (u - \beta_j)/\beta_p$, and the rest $x_k = 0$ for $k \neq j, p$, we deduce from (8.31) that

$$g_j(u) = g(\mathbf{u}_j) - g_0(u).$$

Thus, the functions g_j are uniquely determined from g.

(iii) Suppose that there exist two nonzero and nonparallel vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in R^p such that

$$g(\mathbf{x}) = g_0(\boldsymbol{\beta}^T\mathbf{x}) + \sum_{j=1}^{p-1} g_j(\boldsymbol{\beta}^T\mathbf{x})x_j \tag{8.93}$$

$$= f_0(\boldsymbol{\alpha}^T\mathbf{x}) + \sum_{j=1}^{p-1} f_j(\boldsymbol{\alpha}^T\mathbf{x})x_j. \tag{8.94}$$

By a rotation transform if necessary, without loss of generality, assume $\boldsymbol{\beta} = (c, 0, \ldots, 0)^T$. Then, it follows from (8.93) that $\partial^2 g(\mathbf{x})/\partial x_i^2 = 0$ for $i = 2, \ldots, p$. Write $\boldsymbol{\alpha}^T\mathbf{x} = z$. Choose $2 \leq i \leq p$ fixed for which $\alpha_i \neq 0$. Then, from (8.94), we have that

$$\frac{\partial^2 g(\mathbf{x})}{\partial x_i^2} = \alpha_i^2 \ddot{f}_0(z) + \alpha_i^2 \sum_{j=1}^{p-1} \ddot{f}_j(z)x_j + 2\alpha_i \dot{f}_i(z) = 0,$$

namely,

$$\{\alpha_i \ddot{f}_0(z) + z\ddot{f}_i(z) + 2\dot{f}_i(z)\} + \alpha_i \sum_{j \neq i} \left\{ \ddot{f}_j(z) - \frac{\alpha_j}{\alpha_i}\ddot{f}_i(z) \right\} x_j = 0. \tag{8.95}$$

Letting $x_j = 0$ for $j \neq i$ and $x_i = x/\alpha_i$ in the equation above, we have

$$\alpha_i \ddot{f}_0(x) + x\ddot{f}_i(x) + 2\dot{f}_i(x) = 0. \tag{8.96}$$

Hence (8.95) reduces to

$$\sum_{j \neq i} \left\{ \ddot{f}_j(z) - \frac{\alpha_j}{\alpha_i}\ddot{f}_i(z) \right\} x_j = 0.$$

This entails that

$$\ddot{f}_k(x) = \ddot{f}_i(x)\frac{\alpha_k}{\alpha_i}, \quad 1 \le k \le p-1,$$

by letting $x_k = x/\alpha_k$ and all other $x_j = 0$ for $k \neq i$ and $\alpha_k \neq 0$, or $x_k \neq 0$, $x_i = x/\alpha_i$, and all other $x_j = 0$ for $k \neq i$ and $\alpha_k = 0$. This implies that $f_k(z) = f_i(z)\alpha_k\alpha_i^{-1} + a_k z + b_k$ with $a_i = b_i = 0$. Substituting this into (8.94), we have

$$\begin{aligned} g(\mathbf{x}) &= f_0(\boldsymbol{\alpha}^T\mathbf{x}) + \alpha_i^{-1} f_i(\boldsymbol{\alpha}^T\mathbf{x})\boldsymbol{\alpha}^T\mathbf{x} + \sum_{j\neq i}(a_j\boldsymbol{\alpha}^T\mathbf{x} + b_j)x_j \\ &\equiv f_0^*(\boldsymbol{\alpha}^T\mathbf{x}) + \sum_{j\neq i}(a_j\boldsymbol{\alpha}^T\mathbf{x} + b_j)x_j. \end{aligned}$$

Now, an application of the argument (8.96) to the last expression above shows that $f_0^*(z) = a_0 z + b_0$. Thus

$$g(\mathbf{x}) = a_0\boldsymbol{\alpha}^T\mathbf{x} + b_0 + \sum_{j\neq i}(a_j\boldsymbol{\alpha}^T\mathbf{x} + b_j)x_j.$$

Now, $\partial^2 g(\mathbf{x})/\partial x_i \partial x_j = a_j\alpha_i$ for any $j \ge 2$, which should be 0 according to (8.93) since $\boldsymbol{\beta} = (c, 0, \ldots, 0)^T$. Hence, all a_j $(j \ge 2)$ in the expression above are zero. This implies that

$$g(\mathbf{x}) = \boldsymbol{\gamma}^T\mathbf{x} + b_0 + a_1 x_1 \boldsymbol{\alpha}^T\mathbf{x} = \boldsymbol{\gamma}^T\mathbf{x} + b_0 + c^{-1}a_1\boldsymbol{\beta}^T\mathbf{x}\,\boldsymbol{\alpha}^T\mathbf{x},$$

where $\boldsymbol{\gamma} = a_0\boldsymbol{\alpha} + \mathbf{b}$, and $\mathbf{b} = (b_1, \cdots, b_p)^T$. ∎

8.8.6 Conditions of Theorem 8.5

Conditions 3:

(i) For a given point x, $p(x) > 0$, $\sigma^2(x) > 0$, and the functions $E\{Y^k|X = z\}$ are continuous at x for $k = 3, 4$. Furthermore, $\ddot{f}(z) \equiv d^2 f(z)/dz^2$ and $\ddot{\sigma}^2(z) \equiv d^2\{\sigma^2(z)\}/dz^2$ are uniformly continuous on an open set containing the point x.

(ii) $E\{Y^{4(1+\delta)}\} < \infty$, where $\delta \in [0, 1)$ is a constant.

(iii) The kernel functions W and K are symmetric density functions each with a bounded support. Furthermore, $|W(x_1) - W(x_2)| \le c|x_1 - x_2|$, $|K(x_1) - K(x_2)| \le c|x_1 - x_2|$, and also $|p(x_1) - p(x_2)| \le c|x_1 - x_2|$ for real value x_1, x_2.

(iv) The strictly stationary process $\{(X_i, Y_i)\}$ is absolutely regular; that is,

$$\beta(j) \equiv \sup_{i \geq 1} E \left\{ \sup_{A \in \mathcal{F}_{i+j}^{\infty}} |\mathrm{P}(A|\mathcal{F}_1^i) - \mathrm{P}(A)| \right\} \to 0, \quad \text{as } j \to \infty,$$

where $\mathcal{F}_i^j$ is the σ-field generated by $\{(X_k, Y_k) : k = i, \ldots, j\}$, $(j \geq i)$. Furthermore, for the same δ as in (ii),

$$\sum_{j=1}^{\infty} j^2 \beta^{\frac{\delta}{1+\delta}}(j) < \infty.$$

(We use the convention $0^0 = 0$.)

(v) As $n \to \infty$, $h_i \to 0$, and $\liminf_{n\to\infty} nh_i^4 > 0$ for $i = 1, 2$.

We impose the boundedness on the supports of $K(\cdot)$ and $W(\cdot)$ for brevity of proofs; it may be removed at the cost of lengthier proofs. In particular, the Gaussian kernel is allowed. The assumption of the convergence rate of $\beta(j)$ is also for technical convenience. For other types of mixing coefficients, the result can also be established. The assumption on the convergence rates of h_1 and h_2 is not the weakest possible.

When $\{(X_t, Y_t)\}$ are independent, condition (iv) holds with $\delta = 0$ and condition (ii) reduces to $E(Y^4) < \infty$. On the other hand, if (iv) holds with $\delta = 0$, there are at most finitely many nonzero $\beta(j)$'s. This means that there exists an integer $0 < j_0 < \infty$ for which (X_i, Y_i) is independent of $\{(X_j, Y_j),\ j \geq i + j_0\}$, for all $i \geq 1$.

8.8.7 Proof of Theorem 8.5

We outline only the key steps of the proofs. We always assume that conditions (i)—(v) hold. We say that $B_n(x) = B(x) + o_p(b_n)$, or $O_p(b_n)$, uniformly for $x \in G$ if

$$\sup_{x \in G} |B_n(x) - B(x)| = o_p(b_n), \text{ or } O_p(b_n).$$

We only present the proof for the cases with $\delta > 0$. The case with $\delta = 0$ can be dealt with in a more direct and simpler way.

The proof is based on the following lemma, which follows directly from Lemma 2 of Yao and Tong (2000).

Lemma 8.5 *Let $G \subset \{p(x) > 0\}$ be a compact subset for which Condition 3(i) holds. As $n \to \infty$, uniformly for $x \in G$,*

$$\begin{aligned} & \widehat{\sigma}^2(x) - \sigma^2(x) \\ = \; & \frac{1}{nh_1 p(x)} \sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right) \{\widehat{r}_i^2 - \sigma^2(x) - \dot{\sigma}^2(x)(X_i - x)\} \\ & + O_p\{R_{n,1}(x)\} \end{aligned} \tag{8.97}$$

and

$$\begin{aligned} \widehat{f}(x) - f(x) &= \frac{1}{nh_2 p(x)} \sum_{i=1}^{n} \sigma(X_i)\varepsilon_i K\left(\frac{X_i - x}{h_2}\right) \\ &\quad + \frac{h_2^2 \sigma_K^2}{2} \ddot{f}(x) + O_p\{R_{n,2}(x)\}, \end{aligned} \tag{8.98}$$

where $\sigma_K^2 = \int x^2 K(x)dx$,

$$\begin{aligned} R_{n,1}(x) &= \frac{1}{np(x)} \left[\left| \sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right) \{\widehat{r}_i^2 - \sigma^2(x) - \dot{\sigma}^2(x)(X_i - x)\} \right| \right. \\ &+ \left. \left| \sum_{i=1}^{n} \frac{X_i - x}{h_1} W\left(\frac{X_i - x}{h_1}\right) \{\widehat{r}_i^2 - \sigma^2(x) - \dot{\sigma}^2(x)(X_i - x)\} \right| \right], \\ R_{n,2}(x) &= \frac{1}{np(x)} \left\{ \left| \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_2}\right) \sigma(X_i)\varepsilon_i \right| \right. \\ &\quad \left. + \left| \sum_{i=1}^{n} \frac{X_i - x}{h_2} K\left(\frac{X_i - x}{h_2}\right) \sigma(X_i)\varepsilon_i \right| \right\} + O(h_2^3). \end{aligned}$$

■

We are now ready to prove the results. It follows from (8.75) and (8.97) that

$$\begin{aligned} \widehat{\sigma}^2(x) - \sigma^2(x) &= I_1 + I_2 - I_3 + I_4 \\ &\quad + O_p(h_1)(|I_1 + I_2 - I_3 + I_4| + |I_1' + I_2' - I_3' + I_4'|), \end{aligned}$$

where

$$
\begin{aligned}
I_1 &= \frac{1}{nh_1p(x)}\sum_{i=1}^{n} W\left(\frac{X_i-x}{h_1}\right)\{\sigma^2(X_i)-\sigma^2(x)-\dot{\sigma}^2(x)(X_i-x)\},\\
I_2 &= \frac{1}{nh_1p(x)}\sum_{i=1}^{n} W\left(\frac{X_i-x}{h_1}\right)\sigma^2(X_i)(\varepsilon_i^2-1),\\
I_3 &= \frac{2}{nh_1p(x)}\sum_{i=1}^{n} W\left(\frac{X_i-x}{h_1}\right)\sigma(X_i)\varepsilon_i\{\widehat{f}(X_i)-f(X_i)\},\\
I_4 &= \frac{1}{nh_1p(x)}\sum_{i=1}^{n} W\left(\frac{X_i-x}{h_1}\right)\{\widehat{f}(X_i)-f(X_i)\}^2,
\end{aligned}
\tag{8.99}
$$

and I_j' $(1 \le j \le 4)$ is defined in the same way as I_j with one more factor $h_1^{-1}(X_i - x)$ in the ith summand. It is easy to see that the theorem follows directly from statements (a)—(d) below.

(a) $I_1 = \frac{1}{2}h_1^2\ddot{\sigma}^2(x)\sigma_W^2 + o_p(h_1^2)$ and $I_1' = o_p(h_1^2)$.

(b) $(nh_1)^{\frac{1}{2}}I_2 \xrightarrow{D} N(0,\ \sigma^4(x)\lambda^2(x)\int W^2(t)dt/p(x))$ and

$$(nh_1)^{\frac{1}{2}}I_2' \xrightarrow{D} N(0,\ \sigma^4(x)\lambda^2(x)\int t^2W^2(t)dt/p(x)).$$

(c) $I_3 = o_p(h_1^2 + h_2^2)$ and $I_3' = o_p(h_1^2 + h_2^2)$.

(d) $I_4 = o_p(h_1^2 + h_2^2)$ and $I_4' = o_p(h_1^2 + h_2^2)$.

In the following, we establish the statements on I_j in (a)—(d) only. The cases with I_j' can be proved in the same manner.

It is easy to see that (a) follows from a Taylor expansion and a direct application of the ergodic theorem. Conditions 3(ii) and 3(iii) imply that

$$E\left\{W\left(\frac{X_i-x}{h_1}\right)\sigma^2(X_i)(\varepsilon_i^2-1)\right\}^{2+\delta/2} < \infty.$$

Note that the condition of absolutely regularity implies α-mixing with $\alpha(j) \le \beta(j)$. By Condition 3(iv) and Theorem 2.21, I_2 is asymptotically normal with mean 0 and variance σ_*^2/nh_1, where

$$
\begin{aligned}
\sigma_*^2 &= \frac{1}{h_1}E\left\{W\left(\frac{X-x}{h_1}\right)\frac{\sigma^2(X)}{p(X)}(\varepsilon^2-1)\right\}^2\\
&\quad + \frac{1}{h_1}\sum_{i=2}^{n} E\left\{W\left(\frac{X_1-x}{h_1}\right)\frac{\sigma^2(X_1)}{p(X_1)}(\varepsilon_1^2-1)\right.\\
&\quad \left.\times W\left(\frac{X_i-x}{h_1}\right)\frac{\sigma^2(X_i)}{p(X_i)}(\varepsilon_i^2-1)\right\}.
\end{aligned}
\tag{8.100}
$$

It is easy to see that the first term in the expression above converges to

$$\sigma^4(x)\lambda^2(x)\int W^2(t)dt/p(x).$$

Note that, for any $i \geq 2$,

$$E\left\{W\left(\frac{X_1-x}{h_1}\right)\frac{\sigma^2(X_1)}{p(X_1)}(\varepsilon_1^2-1)W\left(\frac{X_i-x}{h_1}\right)\frac{\sigma^2(X_i)}{p(X_i)}(\varepsilon_i^2-1)\right\}^{1+\delta} = O(h_1^2),$$

$$E\left\{W\left(\frac{X-x}{h_1}\right)\frac{\sigma^2(X)}{p(X)}(\varepsilon^2-1)\right\} = 0,$$

and

$$E\left|W\left(\frac{X-x}{h_1}\right)\frac{\sigma^2(X)}{p(X)}(\varepsilon^2-1)\right|^{1+\delta} = O(h_1).$$

It follows from Condition (iv) and Lemma 1 of Yoshihara (1976) that the absolute value of the second term on the right-hand side of (8.100) is bounded above by

$$ch_1^{(1-\delta)/(1+\delta)}\{\beta^{\frac{\delta}{1+\delta}}(1)+\cdots+\beta^{\frac{\delta}{1+\delta}}(n-1)\} = o(1).$$

Hence (b) holds.

Note that $W(\cdot)$ has a bounded support contained in the interval $[-s_w, s_w]$, say. Therefore, in the summation on the right-hand side of (8.99), only those terms with $X_i \in [x-h_2s_w, x+h_2s_w]$ might not be 0. Let

$$\varphi_{ij} = K\left(\frac{X_i-X_j}{h_2}\right)\sigma(X_i)\sigma(X_j)\varepsilon_i\varepsilon_j\left\{p^{-1}(X_i)W\left(\frac{X_i-x}{h_1}\right) + p^{-1}(X_j)W\left(\frac{X_j-x}{h_1}\right)\right\}.$$

It follows from (8.98) that we may write $I_3 = I_{31} + I_{32} + I_{33}$, where

$$\begin{aligned} I_{31} &= \frac{1}{n^2 h_1 h_2 p(x)} \sum_{i,j=1}^{n} \varphi_{ij} \\ &= \frac{2}{n^2 h_1 h_2 p(x)} \sum_{1 \le i < j \le n} \varphi_{ij} + O_p\left(\frac{1}{n h_2}\right), \qquad (8.101) \\ I_{32} &= \frac{h_2^2 \sigma_K^2}{n h_1 p(x)} \sum_{i=1}^{n} W\left(\frac{X_i - x}{h_1}\right) \sigma(X_i) \varepsilon_i \ddot{f}(X_i) \;=\; o_p(h_2^2), \\ |I_{33}| &\le \frac{O_p(1)}{n^2 h_1} \left| \sum_{i,j=1}^{n} W\left(\frac{X_i - x}{h_1}\right) K\left(\frac{X_i - X_j}{h_2}\right) \frac{\sigma(X_i)\sigma(X_j)|\varepsilon_i|\varepsilon_j}{p(X_i)} \right| \\ &\quad + \frac{O_p(1)}{n^2 h_1} \left| \sum_{i,j=1}^{n} \frac{X_j - X_i}{h_2} W\left(\frac{X_i - x}{h_1}\right) K\left(\frac{X_i - X_j}{h_2}\right) \sigma(X_i)\sigma(X_j) \right. \\ &\quad \left. \times |\varepsilon_i|\varepsilon_j / p(X_i)| + o_p(h_2^2). \right. \end{aligned}$$

It follows from Lemma A(ii) of Hjellvik, Yao, and Tjøstheim (1998) that, for any $\varepsilon_0 > 0$ and $\varepsilon > 0$,

$$\begin{aligned} & \mathrm{P}\left\{ n^{-1} (h_1 h_2)^{-(\frac{1}{1+\delta} - \varepsilon_0)/2} \Big| \sum_{I<j} \varphi_{ij} \Big| > \varepsilon \right\} \\ \le\; & \frac{c (h_1 h_2)^{\varepsilon_0}}{n^2} E\left\{ (h_1 h_2)^{-\frac{1}{2(1+\delta)}} \sum_{i<j} \varphi_{ij} \right\}^2 \\ =\; & o\left((h_1 h_2)^{\varepsilon_0}\right). \end{aligned}$$

Therefore, the first term on the right-hand side of (8.101) is

$$o_p\{n^{-1}(h_1 h_2)^{-(\frac{1+2\delta}{1+\delta} + \varepsilon_0)/2}\}.$$

Thus

$$I_{31} = o_p(n^{-1}(h_1 h_2)^{-(\frac{1+2\delta}{1+\delta} + \varepsilon_0)/2}) + O_p(n^{-1} h_2^{-1}).$$

Condition 3(v) implies that the terms on the right-hand side of the expression above are of order $o_p(h_1^2 + h_2^2)$ if we choose $\varepsilon_0 < (1+\delta)^{-1}$. Performing Hoeffding's projection decomposition of U-statistics, we can prove that $I_{33} = o_p(h_1^2 + h_2^2)$ in the same way.

The proof of (d) is similar to that of (c) and therefore is omitted here.

8.9 Bibliographical Notes

Functional-coefficient models

The functional-coefficient models are also referred to as the varying coefficient models . Varying-coefficient models arise from many contexts and have been successfully applied to multidimensional nonparametric regression, generalized linear models, nonlinear time series models, longitudinal and functional data analysis, and interest rate modeling in finance.

Early applications of varying-coefficient models appear in Haggan and Ozaki (1981), Ozaki (1982), and Shumway (1988, p. 245). However, the nonparametric techniques of the varying-coefficient model were not popularized until the work of Cleveland, Grosse, and Shyu (1991), Chen and Tsay (1993), and Hastie and Tibshirani (1993).

For the independent data, the conditional bias and variance of the estimators were derived in Carroll, Ruppert, and Welsh (1998) and Fan and Zhang (1999), where a two-step procedure is also proposed. The asymptotic normality and bandwidth selection were presented in Zhang and Lee (2000). The distribution of the maximum discrepancy between the estimated coefficients and true ones was discussed by Xia and Li (1999b) and Fan and Zhang (2000).

The varying-coefficient models have been popularly used to analyze the longitudinal data. They are a specific case of the functional linear model discussed in Ramsay and Silverman (1997) in the context of functional data analysis. They allow one to examine the extent to which the association between independent and dependent variables varies over time. The kernel and spline methods have been proposed in Brumback and Rice (1998) and Hoover, Rice, Wu, and Yang (1998). Fan and Zhang (2000) proposed a two-step approach to improve the efficiency of estimated coefficient functions, while Wu and Chiang (2000) used a different approach. Approaches for constructing confidence regions based on the kernel method were introduced in Wu, Chiang, and Hoover (1998). Fan, Jiang, Zhang, and Zhou (2003) use varying-coefficient models to model term structure dynamics.

Additive models

The use of additive models can be dated back at least to Ezekiel (1924). The idea of extending additive models includes the projection pursuit model due to Friedman and Stuetzle (1981), the transformed additive model by Breiman and Friedman (1985), and the generalized additive model by Hastie and Tibshirani (1986). The convergence of the backfitting algorithm has been studied by a number of authors, including Breiman and Friedman (1985), Buja, Hastie, and Tibshirani (1989), Härdle and Hall (1993), Ansley and Kohn (1994), and Opsomer and Ruppert (1997). The asymptotic bias and variance of the backfitting estimator using the local polynomial fitting were investigated by Opsomer and Ruppert (1997) and Opsomer (2000). The asymptotic normality of the backfitting estimator using the local polynomial estimator was established by Wand (2000). Yee and Wild (1996) extended additive models for multivariate responses. Smith, Wong,

and Kohn (1998) considered additive nonparametric regression with autocorrelated errors. Model selection for semiparametric and additive models is discussed in Härdle and Tsybakov (1995) and Simonoff and Tsai (1999).

The optimal rates of convergence for additive models have been established by Stone (1985, 1986). Further extensions of the results can be found in Stone (1994). The efficiency issues of the additive models were studied by Linton (1997, 2000) and Fan, Härdle, and Mammen (1998). Fan, Härdle, and Mammen (1998) were among the first to discover the oracle property for the additive model. Mammen, Linton, and Nielsen (1999) modified the backfitting algorithm to construct an efficient estimator for additive models. Kim, Linton, and Hengartner (1999) provided an efficient algorithm for obtaining efficient estimators for additive components. Diagnostic tools for additive models were given in Breiman (1993). Linton, Chen, Wang, and Härdle (1997) proposed additive regression models with a parametrically transformed response variable.

The projection estimator was proposed by Tjøstheim and Auestad (1994a) for identification of nonlinear time series models. It was applied to selecting significant lags in Tjøstheim and Auestad (1994b). The procedure was further extended and studied by Masry and Tjøstheim (1995) and Cai and Fan (2000). In the multiple regression setting, the idea was independently proposed by Linton and Nielsen (1995) under the name of "marginal integration estimator." The procedure was then modified and extended by several authors, including Linton and Härdle (1996), Linton (1997, 2000), Fan, Härdle, and Mammen (1998), Kim, Linton, Hengartner (1999) and Mammen, Linton, and Nielsen (1999), , among others. Nielsen and Linton (1998) gave an optimization interpretation of integration and backfitting estimators.

Several papers have addressed the issue of whether an additive model reasonably fits a given data set. Additive tests for nonlinear autoregression were proposed and studied by Chen, Liu, and Tsay (1995). For multiple nonparametric regression with factorial type of designs, Eubank, Hart, Simpson, and Stefanski (1995) studied the Tukey type of test for additivity.

Other nonparametric models

Partially linear models were introduced by Wahba (1984) and studied further by Heckman (1986), Chen (1988), Cuzick (1992), Severini and Staniswalis (1994) and Hunsberger (1994), among others. The idea of the profile least-squares method for the partially linear models was proposed by Speckman (1988). Severini and Wong (1992) gave an insightful study of the partially linear model from the profile likelihood perspective. Carroll, Fan, Gijbels, and Wand (1997) extended the partially linear models to generalized partially single-index models and proposed semiparametric efficient methods. The techniques have been successfully applied and extended to the assessment of selection biases of job training programs by Heckman,

Ichimura, Smith, and Todd (1998). The partially linear models have also been applied to other contexts such as the errors-in-variables regression and survival analysis (Huang 1999, Liang, Härdle, and Carroll 1999). More references and techniques can be found in the recent monograph by Härdle, Liang, and Gao (2000).

Several methods have been proposed for estimating the indices in the multiple-index models. Sliced inverse regression was proposed in Duan and Li (1991), Li (1991) and . The idea was extended to handle the predictors that contain binary regressors by Carroll and Li (1995). Principal Hessian directions, introduced by Li, K.C. (1992) and revisited by Cook (1998), are alternative methods of estimating multiple indices. Hsing and Carroll (1992) studied the asymptotic normality of the two-slice estimate of covariance matrix used in the sliced inverse regression. Hall and Li (1993) studied the shapes of low-dimensional projections from high-dimensional data. The results broaden the scope of the applicability of the sliced inverse regression. Schott (1994) addressed the problem of determining the number of indices. Zhu and Fang (1996) established the asymptotic normality of the estimated covariance matrix used in sliced inverse regression.

The average derivative method is a forward regression approach for estimating the indices in the multiple-index models; see, for example, Härdle and Stoker (1989). The technique has been applied to econometric models by Hildenbrand and Jerison (1991), Stoker (1992), and Newey and Stoker (1993). Samarov (1993) used the average derivative method for model selection and diagnostics. Chaudhuri, Doksum, and Samarov (1997) applied the technique to estimate quantile functions. The approach has recently been improved by Hristache, Juditsky, Polzehl, and Spokoiny (2002). Estimation and optimal smoothing in single-index models have been studied by Härdle, Hall, and Ichimura (1993). Other innovative ideas appear in Cook (1996, 1998), Cook and Lee(1999), Cook, Chiaromonte, and Li (2002), and Cook and Li (2002).

For a detailed treatment of classifications and regression trees, see Breiman, Friedman, Olshen, and Stone (1993). Breiman (1993) introduced the concept of hinging hyperplanes for regression, classification, and function approximation. Hastie, Tibshirani, and Buja (1994) studied nonparametric versions of discriminant analysis with multiple linear regression replaced by any nonparametric regression technique. Applications of tree-based regression models to the health sciences can be found in the book by Zhang and Singer (1999). Recently, Li, Lue, and Chen (2000) applied principal Hessian directions to construct an interactive tree-structured regression.

Estimation of conditional variance

Several authors have studied the problem of estimating conditional variance in the nonparametric regression model. When the variance function is constant, it can be estimated at a root-n rate. An early paper on the

estimation of constant variance in a nonparametric setting is Rice (1984b). Gasser, Sroka, and Jennen-Steinmetz (1986) proposed an estimator based on residuals. A class of difference-based estimators of the variance was studied by Hall, Kay, and Titterington (1990), where optimal weights are obtained.

The problem of estimating the conditional variance function has been extensively studied in the literature. Müller and Stadtmüller (1987) established the uniform rate of convergence and the asymptotic normality of a kernel estimator. Hall and Carroll (1989) examined the oracle property of conditional variance estimation under mild smoothness conditions on the regression function. A general class of nonparametric estimators was studied by Müller and Stadtmüller (1993). Neumann (1994) studied a bandwidth selection problem for the estimation of conditional variance. The mean-square error of the local polynomial estimation of the variance function was studied by Ruppert, Wand, Holst, and Hössjer (1997). Härdle and Tsybakov (1997) studied nonparametric estimation of a volatility function in nonparametric autoregression using a local polynomial. Fan and Yao (1998) studied the local linear estimation of conditional variance when data are dependent.

Estimation of drift function and volatility function in diffusion models has also received considerable attention in the literature. Pham (1981) and Prakasa Rao (1985) proposed nonparametric drift estimators. Florens-Zmirou (1993) studied the problem of estimating the diffusion coefficient from discrete observations. A similar problem for multidimensional diffusion processes was studied by Genon-Catalot and Jacod (1993). Uniformly strong consistency for the Nadaraya–Watson kernel estimator of the drift function was established by Arfi (1995, 1998) under ergodic conditions. A semiparametric procedure for estimating the diffusion function was proposed by Aït-Sahalia (1996). Jiang and Knight (1997) developed a nonparametric kernel estimator for the diffusion function. Stanton (1997) used a kernel method to estimate the volatility function. Inferences on the volatility functions have been studied by Fan and Zhang (2003). Time-inhomogeneous nonparametric estimation of the volatility function was studied by Fan, Jiang, Zhang, and Zhou (2003).

This page intentionally blank

9
Model Validation

9.1 Introduction

Parametric time series models provide explanatory power and a parsimonious description of stochastic dynamical systems. Yet, there is a risk that misspecification of an underlying stochastic model can lead to misunderstanding of the systems, wrong conclusions, and erroneous forecasting. It is common statistical practice to check whether a parametric model fits a given data set reasonably well. To achieve this, in the Neyman–Pearson framework, we need to specify a class of alternative models. The traditional approach is to use a large family of parametric models under an alternative hypothesis. This is basically a parametric approach for model diagnostics. The implicit assumption is that the large family of parametric models specifies the form of the true underlying dynamics correctly. However, this is not always warranted and leads naturally to a nonparametric alternative hypothesis. It is clear that nonparametric models will reduce the danger of model misspecification.

As seen in Chapter 8, there are many nonparametric models that contain parametric models as their specific examples. All can serve as alternative models. For example, the AR(p) model can be embedded into the FAR(p, d) model (8.1), the AAR(p) model (8.42), and the saturated nonparametric autoregressive model (8.27). Depending on prior knowledge, we choose a class of alternative models. As in statistical estimation, the larger the family of models, the poorer the parameters are estimated but the less the modeling biases. Similarly, in hypothesis testing, the larger the alternative

models, the lower the power of tests but the smaller the danger of model misspecification.

Despite extensive developments on nonparametric estimation techniques, there are few generally applicable methods for nonparametric inferences. For hypothesis testing, there are extensive developments based on the univariate nonparametric model (see Bowman and Azzalini, 1997; Hart, 1997). However, there are only a few papers on multivariate nonparametric models. Most methods are designed for some specific problems. There are virtually no developments on model validation using nonparametric techniques for dependent data.

In this chapter, we will introduce the idea of the generalized likelihood ratio test. This is a generally applicable method for testing against nonparametric models. It has been developed for independent data. Nevertheless, the idea is applicable to time series data, even though the theory for dependent data needs to be developed. In particular, we will introduce the technique for validating ARMA models in the spectral domain and AR models and threshold models in the time-domain.

9.2 Generalized Likelihood Ratio Tests

We first introduce the generalized likelihood ratio statistic. The technique will be used repeatedly. This section follows mainly the development of Fan, Zhang, and Zhang (2001). Although their development was based on independent data, the idea and techniques can be extended readily to the time series setting. It is expected that under some mixing conditions, the results should also hold for the dependent data. Extension of the results in this section to the dependent data remains as an interesting problem for further investigation.

9.2.1 Introduction

The maximum likelihood ratio test is a useful method that is generally applicable to most parametric hypothesis-testing problems. The most fundamental property that contributes tremendously to the success of the maximum likelihood ratio tests is that their asymptotic distributions under null hypotheses are independent of nuisance parameters. This property was referred to as the Wilks phenomenon by Fan, Zhang, and Zhang (2001). With such a property, one can determine the null distribution of the likelihood ratio statistic by using either the asymptotic distribution or the Monte Carlo simulation by setting nuisance parameters at some fitted value. The latter is also referred to as the parametric *bootstrap*.

The question arises naturally whether the maximum likelihood ratio test is still applicable to the problems with nonparametric models as alternative.

First, nonparametric maximum likelihood estimators usually do not exist. Even when they exist, they are hard to compute. Furthermore, the resulting maximum likelihood ratio tests are not optimal. We use the simplest example to illustrate the points above.

Example 9.1 (*Univariate nonparametric model*) Suppose that we have n data $\{(X_i, Y_i)\}$ sampled from the nonparametric regression model

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \cdots, n, \tag{9.1}$$

where $\{\varepsilon_i\}$ are a sequence of i.i.d. random variables from $N(0, \sigma^2)$ and X_i has a density f with support $[0, 1]$. Suppose that the parameter space is

$$\mathcal{F} = \{m : \sup_{x \in [0,1]} |m''(x)| \le 1\}.$$

Consider testing the simple linear regression model

$$H_0 : m(x) = \beta_0 + \beta_1 x \quad \longleftrightarrow \quad H_1 : m(x) \neq \beta_0 + \beta_1 x \tag{9.2}$$

with nonparametric alternative model (9.1). Then, the conditional log-likelihood function given $X_1, \cdots, X_n$ is

$$\ell(m, \sigma) = -n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - m(X_i))^2. \tag{9.3}$$

Let $(\widehat{\beta}_0, \widehat{\beta}_1)$ be the maximum likelihood estimator (MLE) under H_0 and $\widehat{m}_{\text{MLE}}(\cdot)$ be the MLE under the nonparametric model. The latter is used to solve the following optimization problem:

$$\min \sum_{i=1}^{n} (Y_i - m(X_i))^2, \text{ subject to } \sup_{x \in [0,1]} |m''(x)| \le 1.$$

The optimization of such a problem, even if it exists, is hard to find.

Now, let us consider the parameter space

$$\mathcal{M} = \left\{ m : \int_0^1 m''(x)^2 dx \le 1 \right\}.$$

Then, the nonparametric maximum likelihood estimator is used to find m to minimize

$$\sum_{i=1}^{n} (Y_i - m(X_i))^2 \text{ subject to } \int_0^1 m''(x)^2 dx \le 1.$$

As discussed in §6.4.3, the resulting estimator $\widehat{m}_{\text{MLE}}$ is a smoothing spline (Wahba 1990; Eubank 1999), with the smoothing parameter chosen to satisfy $\|\widehat{m}_{\text{MLE}}\|_2^2 = 1$. Define the residual sum of squares RSS_0 and RSS_1 as follows:

$$\text{RSS}_0 = \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2, \qquad \text{RSS}_1 = \sum_{i=1}^{n} \{Y_i - \widehat{m}_{\text{MLE}}(X_i)\}^2. \tag{9.4}$$

Then, it is easy to see that the logarithm of the conditional maximum likelihood ratio statistic for the problem (9.2) is given by

$$T = \ell_n(\widehat{m}_{\text{MLE}}, \widehat{\sigma}) - \ell(\widehat{m}_0, \widehat{\sigma}_0) = \frac{n}{2} \log \frac{\text{RSS}_0}{\text{RSS}_1}, \tag{9.5}$$

where $\widehat{\sigma}^2 = n^{-1}\text{RSS}_1$, $\widehat{m}_0(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$, and $\widehat{\sigma}_0^2 = n^{-1}\text{RSS}_0$. Even in this simplest situation, where the nonparametric MLE exists, Fan, Zhang, and Zhang (2001) have shown that the nonparametric maximum likelihood ratio test is not optimal. This is due to the limited choice of the smoothing parameter in $\widehat{m}_{\text{MLE}}$, which makes it satisfy $\|\widehat{m}_{\text{MLE}}\| = 1$. This choice is optimal for estimating the function m but not for estimating the functional $\|m\|^2$. ∎

The example above reveals that the nonparametric MLE may not exist and hence cannot serve as a generally applicable method. It illustrates further that, even when it exists, the nonparametric MLE chooses smoothing parameters automatically. This is too restrictive for the procedure to possess the optimality of testing problems. Further, we need to know the nonparametric space exactly. For example, the constant "one" in $\mathcal{F}$ and $\mathcal{M}$ needs to be specified. This is an unrealistic assumption in practice. To attenuate these difficulties, we replace the maximum likelihood estimator under the alternative nonparametric model by any reasonable nonparametric estimator. This is the essence of the *generalized likelihood ratio* (GLR) statistics. In Example 9.1, the GLR statistic is used to replace $\widehat{m}_{\text{MLE}}$ by, for example, the local linear estimator $\widehat{m}$. This significantly enhances the flexibility of the test statistic by varying the smoothing parameter. By proper choices of the smoothing parameter, the GLR tests achieve the optimal rates of convergence in the sense of Ingster (1993) and Lepski and Spokoiny (1999). Further, they are applicable to both nonparametric spaces $\mathcal{F}$ and $\mathcal{M}$, even without knowing the exact constant (e.g., the constant "one" in Example 9.1) in these spaces.

9.2.2 Generalized Likelihood Ratio Test

The basic idea of the generalized likelihood ratio test is as follows. Let $\mathbf{f}$ be the vector of functions of main interest and $\boldsymbol{\eta}$ be the nuisance parameters. Suppose that the logarithm of the likelihood of a given set of data is $\ell(\mathbf{f}, \boldsymbol{\eta})$. Given $\boldsymbol{\eta}$, we have a good nonparametric estimator $\widehat{\mathbf{f}}_{\boldsymbol{\eta}}$. The nuisance parameters $\boldsymbol{\eta}$ can be estimated by the *profile likelihood* by maximizing $\ell(\widehat{\mathbf{f}}_{\boldsymbol{\eta}}, \boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$, resulting in the profile likelihood estimator $\widehat{\boldsymbol{\eta}}$. This gives the profile likelihood $\ell(\widehat{\mathbf{f}}_{\widehat{\boldsymbol{\eta}}}, \widehat{\boldsymbol{\eta}})$, which is not the maximum likelihood since $\widehat{\mathbf{f}}_{\boldsymbol{\eta}}$ is not an MLE.

Now, suppose that we are interested in testing whether a parametric family $\mathbf{f}_{\boldsymbol{\theta}}$ fits a given set of data. Formally, the null hypothesis is

$$H_0 : \mathbf{f} = \mathbf{f}_{\boldsymbol{\theta}}, \quad \theta \in \Theta, \tag{9.6}$$

and we use the nonparametric model $\mathbf{f}$ as the alternative model. Let $\widehat{\boldsymbol{\theta}}_0$ and $\widehat{\boldsymbol{\eta}}_0$ be the maximum likelihood estimator under the null model (9.6) maximizing the function $\ell(\mathbf{f}_{\boldsymbol{\theta}}, \boldsymbol{\eta})$. Then $\ell(\mathbf{f}_{\widehat{\boldsymbol{\theta}}_0}, \widehat{\boldsymbol{\eta}}_0)$ is the maximum likelihood under the null hypothesis. The GLR test statistic simply compares the log-likelihood under the two competing classes of models:

$$T = \ell(\widehat{\mathbf{f}}_{\widehat{\boldsymbol{\eta}}}, \widehat{\boldsymbol{\eta}}) - \ell(\mathbf{f}_{\widehat{\boldsymbol{\theta}}_0}, \widehat{\boldsymbol{\eta}}_0). \tag{9.7}$$

Before we proceed further, let us revisit Example 9.1.

Example 9.2 (*Example 9.1 revisited*). In that example, $\mathbf{f} = m$ and $\boldsymbol{\eta} = \sigma$. The log-likelihood function $\ell(m, \sigma)$ is given by (9.3). For a given σ, let $\widehat{m}(\cdot)$ be the local linear estimator based on the data $\{(X_i, Y_i), i = 1, \cdots, n\}$, which is independent of σ. Substituting it into the likelihood, we obtain the profile likelihood

$$-n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \mathrm{RSS}_1,$$

where $\mathrm{RSS}_1 = \sum_{i=1}^n (Y_i - \widehat{m}(X_i))^2$. Maximizing the profile likelihood above with respect to σ, we obtain $\widehat{\sigma}^2 = n^{-1}\mathrm{RSS}_1$ and the profile likelihood

$$-\frac{n}{2} \log(\sqrt{2\pi}\mathrm{RSS}_1/n) - \frac{n}{2}.$$

Under the null hypothesis, $\mathbf{f}_{\boldsymbol{\theta}}(x) = \beta_0 + \beta_1 x$. One can easily obtain the maximum likelihood estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and hence RSS_0 as in (9.4) and $\widehat{\sigma}_0^2 = n^{-1}\mathrm{RSS}_0$. Substituting these into the definition of the GLR, we obtain

$$T = \frac{n}{2} \log(\mathrm{RSS}_0/\mathrm{RSS}_1).$$

This is a GLR test statistic. ∎

As with parametric inferences, the GLR test does not have to use the true likelihood. For example, the test statistic T in Example 9.2 applies to problem (9.2) whether ε_i is normally distributed or not. The normality assumption is simply used to motivate the procedure. Similarly, the GLR statistic does not have to require the MLE under the null model (9.6). In fact, any reasonable parametric methods are applicable since they typically have a faster rate of convergence than nonparametric methods. In addition, as will be discussed in §9.2.6, the approach is also applicable to the case where nuisance parameters contain nonparametric functions.

9.2.3 Null Distributions and the Bootstrap

To utilize the GLR test statistic (9.7), we need to derive the distribution under the null hypothesis (9.6). The question arises naturally whether the

asymptotic null distribution depends on the nuisance parameter under the null hypothesis, namely, whether the Wilks phenomenon continues to hold for the GLR tests with nonparametric alternatives. Furthermore, we ask whether the resulting tests are powerful enough.

For a number of models and a number of hypotheses, studied by Fan, Zhang, and Zhang (2001), it has been shown that the Wilks type of results continue to hold. Like Wilks (1938), Fan, Zhang, and Zhang (2001) are not able to show that the Wilks type of results hold for all problems, but their results indicate that such a phenomenon holds with generality. We use the *varying-coefficient model*, which is closely related to the FAR(p, d) model, to illustrate the results.

Suppose that we have a random sample $\{(U_i, X_{i1}, \cdots, X_{ip}, Y_i)\}_{i=1}^n$ from

$$Y = a_1(U)X_1 + \cdots + a_p(U)X_p + \varepsilon, \tag{9.8}$$

where ε is independent of covariates $(U, X_1, \cdots, X_p)$, having mean zero and variance σ^2. To facilitate the notation, we denote

$$\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^T, \quad \mathbf{A}(U) = (a_1(\mathbf{U}), \cdots, a_p(\mathbf{U}))^T$$

and rewrite the model (9.8) as

$$Y = \mathbf{A}(\mathbf{U})^T\mathbf{X} + \varepsilon.$$

Consider first the simple null hypothesis testing problem

$$H_0 : \mathbf{A} = \mathbf{A}_0, \qquad \longleftrightarrow \qquad H_1 : \mathbf{A} \neq \mathbf{A}_0 \tag{9.9}$$

for a given vector of functions $\mathbf{A}_0$. The GLR statistic can be constructed by using the local linear fit.

Let $\widehat{\mathbf{A}}$ be the vector of the local linear estimator constructed by using (8.7). Define $\text{RSS}_1 = \sum_{i=1}^n (Y_i - \widehat{\mathbf{A}}(\mathbf{U}_i)^T\mathbf{X}_i)^2$, the residual sum of squares under the nonparametric model. Using the same derivation as in Example 9.2, when $\varepsilon \sim N(0, \sigma^2)$, the GLR test statistic is given by

$$T_{n,1} = \frac{n}{2}\log(\text{RSS}_0/\text{RSS}_1), \tag{9.10}$$

where $\text{RSS}_0 = \sum_{i=1}^n (Y_i - \mathbf{A}_0(\mathbf{U}_i)^T\mathbf{X}_i)^2$. By Taylor's expansion, we have the following approximation:

$$\begin{aligned} T_{n,1} &= \frac{n}{2}\log\left(1 + \frac{\text{RSS}_0 - \text{RSS}_1}{\text{RSS}_1}\right) \\ &\approx \frac{n}{2}\frac{\text{RSS}_0 - \text{RSS}_1}{\text{RSS}_1}. \end{aligned}$$

This approximation is valid as long as RSS_0 and RSS_1 are close. This is often the case when we need the test statistics to differentiate between the

two classes of models. For this reason, the F-type of test statistic provides a useful variant of the GLR test statistic; see for example, Azzalini, Bowman, and Härdle (1989).

As will be demonstrated in the next theorem, the normality assumption on ε is merely used to motivate the testing procedure. We use the notation $cT_n \overset{a}{\sim} \chi^2_{a_n}$ (with $a_n \to \infty$) to denote a sequence of random T_n having

$$\{2a_n\}^{-1/2}(cT_n - a_n) \xrightarrow{D} N(0,1).$$

The technical arguments for establishing the asymptotic null distribution of $T_{n,1}$ are complicated, and we refer readers to the paper by Fan, Zhang, and Zhang (2001) for details. Here, we only give technical conditions of the theorem.

Condition (A)

(A1) The marginal density of $\mathbf{U}$ is Lipschitz continuous and bounded away from 0. $\mathbf{U}$ has a bounded support Ω.

(A2) $\mathbf{A}(u)$ has the continuous second derivative.

(A3) The function $K(t)$ is symmetric and bounded. Furthermore, the functions $t^3K(t)$ and $t^3K'(t)$ are bounded and $\int t^4K(t)dt < \infty$.

(A4) $E|\varepsilon|^4 < \infty$.

(A5) $\mathbf{X}$ is bounded. The $p \times p$ matrix $E(\mathbf{X}\mathbf{X}^T|\mathbf{U} = u)$ is invertible for each $u \in \Omega$ and Lipschitz continuous.

Theorem 9.1 *Under Condition* **(A)**, *if* $\mathbf{A}_0$ *is linear or* $nh^{9/2} \to 0$, *then as* $nh^{3/2} \to \infty$,

$$r_K T_{n,1} \overset{a}{\sim} \chi^2_{\mu_n},$$

where $\mu_n = r_K c_K p|\Omega|/h$, $|\Omega|$ *is the length of the support of* U,

$$r_K = \frac{K(0) - \frac{1}{2}\int K^2(t)dt}{\int (K(t) - \frac{1}{2}K * K(t))^2 dt}, \quad \textit{and} \quad c_K = K(0) - \frac{1}{2}\int K^2(t)dt.$$

The theorem above reveals that the asymptotic null distribution is independent of any nuisance parameters, such as σ^2 and the density function of the covariate vector $(U, X_1, \cdots, X_p)$. The normalization factor is r_K rather than 2 in the parametric maximum likelihood ratio test. The degree of freedom depends on $p|\Omega|/h$. This can be understood as follows. Imagine that we partition the support of U into equispaced intervals, each with length h. Model $a_j(\cdot)$ by a constant in each subinterval, resulting in $p|\Omega|/h + 1$ parameters for model (9.8). Yet the number of parameters under the null hypothesis is 1. Hence, the degree of freedom is $p|\Omega|/h$. Since the local

linear estimator uses overlapping intervals, the effective number of parameters is slightly different from $p|\Omega|/h$. The constant factor $r_K c_K$ reflects this difference.

The asymptotic null distribution is known and can be used to compute p-values. Table 7.1 shows the values r_K and c_K. However, the asymptotic distribution does not necessarily give a good approximation for finite samples. For example, from the asymptotic point of view, the $\chi^2_{\mu_n+20}$-distribution and the $\chi^2_{\mu_n}$-distribution are approximately the same, but for moderate μ_n, they can be quite different. What this means is that we need a second order term. Assume that the appropriate degree of freedom is $\mu_n + c$ for a constant c. The constant c can be determined as follows. When the bandwidth is large ($h \to \infty$), the local linear fit becomes a global linear fit. The GLR test becomes the maximum likelihood ratio test. Hence, $T_{n,1} \xrightarrow{D} \chi^2_{2p}$ according to the classical Wilks type of result. It is reasonable to expect that the degree of freedom $\mu_n + c \to 2p$ as $h \to \infty$; namely $c = 2p$. This kind of calibration idea appeared in Zhang (2002b). In conclusion, the $\chi^2_{\mu_n+2p}$-distribution might be a closer approximation to the null distribution of $T_{n,1}$ than that of the $\chi^2_{\mu_n}$-distribution.

A better alternative is to use the parametric conditional bootstrap. The only nuisance parameter under the null hypothesis is σ^2. For a given data set, we are not certain whether the null model holds. Thus, we use the fits from the alternative model, which is consistent under both classes of models. This yields an estimate $\widehat{\sigma}_1^2 = n^{-1}\text{RSS}_1$. The conditional bootstrap method reads as follows:

1. Simulate ε_i^* from $N(0, \widehat{\sigma}_1^2)$ and construct the conditional bootstrap sample:
$$Y_i^* = \mathbf{A}_0(U_i)^T \mathbf{X}_i + \varepsilon_i^*, \quad i = 1, \cdots, n.$$

2. Use the bootstrap sample $\{(U_i, \mathbf{X}_i, Y_i^*)\}_{i=1}^n$ to construct the GLR statistic T^*.

3. Repeat Steps 1 and 2 B times (say, 1,000 times) and obtain B GLR statistics $T_1^*, \cdots, T_B^*$.

4. Use the empirical distribution
$$\widehat{F}_B(x) = B^{-1} \sum_{i=1}^{B} I(T_i^* \le x)$$
as an approximation to the distribution of $T_{n,1}$ under the null hypothesis.

In particular, the p-value is simply the percentage of $\{T_i^*\}_{i=1}^B$ greater than $T_{n,1}$.

When the normality assumption on ε_i is removed, one can replace Step 1 by drawing ε_i^* from the centered residuals $\{\widehat{\varepsilon}_i - \bar{\widehat{\varepsilon}}\}_{i=1}^n$, where $\{\widehat{\varepsilon}_i\}$ are the

residuals from the alternative model and $\bar{\widehat{\varepsilon}}$ is their average. This is basically the conditional nonparametric bootstrap.

Theorem 9.1 is also applicable to testing composite null hypotheses. To see this, let us consider the composite null hypothesis

$$H_0 : \mathbf{A}(u) = \mathbf{A}(u, \boldsymbol{\beta}). \tag{9.11}$$

Let $\widehat{\boldsymbol{\beta}}$ be the least-squares estimator minimizing

$$\sum_{i=1}^{n} \{Y_i - \mathbf{A}(u, \boldsymbol{\beta})^T \mathbf{X}_i\}^2.$$

Let RSS_0^* be the residual sum of squares under model (9.11). Then, the GLR statistic is given by

$$T_{n,2} = \frac{n}{2} \log(\mathrm{RSS}_0^*/\mathrm{RSS}_1).$$

To derive the asymptotic distribution of $T_{n,2}$, consider two fabricated testing problems,

$$H_0^{(1)} : \mathbf{A}(u) = \mathbf{A}(u, \boldsymbol{\beta}_0) \quad \longleftrightarrow \quad H_1^{(1)} : \mathbf{A}(u) = \mathbf{A}(u, \boldsymbol{\beta}) \tag{9.12}$$

and

$$H_0^{(2)} : \mathbf{A}(u) = \mathbf{A}(u, \boldsymbol{\beta}_0) \quad \longleftrightarrow \quad H_1^{(2)} : \mathbf{A}(u) \neq \mathbf{A}(u, \boldsymbol{\beta}_0), \tag{9.13}$$

where $\boldsymbol{\beta}_0$ is the true parameter under H_0. Both of them have the same simple null hypothesis. Then, decompose

$$T_{n,2} = \frac{n}{2} \log(\mathrm{RSS}_0/\mathrm{RSS}_1) - \frac{n}{2} \log(\mathrm{RSS}_0/\mathrm{RSS}_0^*). \tag{9.14}$$

Note that the first term is the GLR statistic for the problem (9.9), or more precisely (9.13), and the second term is the traditional likelihood ratio test for the problem (9.12). Applying Theorem 9.1, $r_K T_{n,1} \overset{a}{\sim} \chi^2_{\mu_n}$. When $\widehat{\beta}$ is the MLE, by the traditional Wilks theorem, it has an asymptotic χ^2_q-distribution, which is stochastically bounded in the sense that it does not diverge, where q is the number of parameters in $\boldsymbol{\beta}$. Hence, the first term in (9.14) dominates the second term. Therefore,

$$r_K T_{n,2} \overset{a}{\sim} \chi^2_{\mu_n}. \tag{9.15}$$

The result (9.15) reveals that the asymptotic null distribution is independent of the nuisance parameters $\boldsymbol{\beta}$ and σ^2. Theoretically speaking, we can set $\boldsymbol{\beta}$ and σ^2 at any predetermined value in the conditional bootstrap above. In practice, we use $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}_1^2$ constructed from the alternative model (i.e., pretending $\mathbf{A}_0(\cdot) = \mathbf{A}(\cdot, \widehat{\boldsymbol{\beta}})$ in the conditional bootstrap above).

9.2.4 Power of the GLR Test

We now consider the power of the GLR test for the problem (9.9). Consider the contiguous alternatives of form

$$H_{1n} : \mathbf{A}(u) = \mathbf{A}_0(u) + \mathbf{G}_n(u), \tag{9.16}$$

where $\mathbf{G}_n(u) = (g_{1n}(u), \cdots, g_{pn}(u))^T$ is the deviation from the null hypothesis. This class of alternative hypothesis is wider than

$$H_{1n} : \mathbf{A}(u) = \mathbf{A}_0(u) + a_n\mathbf{G}(u) \tag{9.17}$$

for some given $a_n \to 0$ and $\mathbf{G}(\cdot)$ and

$$H_{1n} : \mathbf{A}(u) = \mathbf{A}_0(u) + a_n\mathbf{G}_0((u-u_0)/b_n) + a_n'\mathbf{G}_1((u-u_1)/b_n') \tag{9.18}$$

for some given $a_n, a_n', b_n, b_n' \to 0$ and $u_0, u_1, \mathbf{G}_0$, and $\mathbf{G}_1$. In many different contexts, the power of tests has been popularly studied under the alternative of form (9.17) (see, e.g., Hart 1997). Despite its popularity and its simplicity, the class of alternatives (9.17) is limited for nonparametric applications. For example, it implies that not only does the function deviate from the null hypothesis at rate a_n but also so do its derivatives. By the proper choice of rates in (9.18), this problem can be avoided. Yet, this kind of alternative is not included in (9.17) but in (9.16).

The power under the alternative (9.16) has been calculated by Fan, Zhang, and Zhang (2001). In particular, they show that when the bandwidth is of order $n^{-2/9}$, the GLR test can detect alternatives with rate

$$\{E\mathbf{G}_n^T(\mathbf{U})\mathbf{X}\mathbf{X}^T\mathbf{G}_n(\mathbf{U})\}^{1/2} \geq n^{-4/9}$$

uniformly over a large class of functions $\mathbf{G}_n$ that satisfy some regularity conditions. This rate $n^{-4/9}$ is optimal, according to Ingster (1993). In other words, the GLR test is optimal in uniformly detecting a large class of alternative models that deviate from the null hypothesis with rate $O(n^{-4/9})$. The optimality here is not the same as the most powerful test in the classical sense. In fact, for problems as complex as ours, the uniformly most powerful test does not exist.

9.2.5 Bias Reduction

When the null hypothesis $\mathbf{A}(\cdot;\boldsymbol{\beta})$ is not linear, a local linear fit will have biases under the null hypothesis. This affects the null distribution of the generalized likelihood statistic. This problem was handled in Theorem 9.1 by letting bandwidth go to zero sufficiently fast. Although this solves the problem theoretically, it does not solve the bias problem in practice. In practice, it is unknown how small the bandwidth is in order to have negligible biases. Furthermore, too small a bandwidth can have an adverse impact on the power of the GLR.

Our approach is quite simple. We use the framework in §9.2.2 to illustrate the idea. Reparameterize the unknown functions as $\mathbf{f}^* = \mathbf{f} - \mathbf{f}_{\widehat{\theta}_0}$. Then, the problem (9.6) becomes testing

$$H_0 : \mathbf{f}^* = 0 \quad \longleftrightarrow \quad \mathbf{f}^* \neq 0 \tag{9.19}$$

with the likelihood function $\ell^*(\mathbf{f}^*, \eta) = \ell(\mathbf{f}^* + \mathbf{f}_{\widehat{\theta}_0}, \eta)$. Now, apply the GLR test to this reparameterized problem with the new likelihood function $\ell^*(\mathbf{f}^*, \eta)$. The bias problem in the null distribution has disappeared for this reparame-
terized problem, since any reasonable nonparametric estimator will not have biases when the true function is zero. This approach is related to an idea of Härdle and Mammen (1993), Hjort and Glad (1995), and Glad (1998).

Let $\widehat{\mathbf{f}}^*$ be a profile likelihood estimate as in §9.2.2 based on $\ell^*(\mathbf{f}^*, \eta)$. Then, the bias-corrected version of the GLR test is

$$T^* = \ell^*(\widehat{\mathbf{f}}^*, \widehat{\eta}) - \ell^*(0, \widehat{\eta}_0).$$

Compare this with (9.17).

9.2.6 Nonparametric versus Nonparametric Models

For multivariate nonparametric models such as (9.8), we naturally ask whether a few covariates are statistically significant. This is equivalent to testing problems such as

$$H_0 : a_1(U) = a_2(U) = 0,$$

corresponding to whether X_1 and X_2 are statistically significant. This problem is different from (9.11) in that the null model is still a nonparametric model since the functions $a_3, \cdots, a_p$ are not restricted to a parametric family.

We use the varying-coefficient model (9.8) to illustrate the basic idea. Consider more generally the testing problem

$$H_0' : \mathbf{A}_1 = \mathbf{A}_{1,0} \qquad \longleftrightarrow \qquad H_1' : \mathbf{A}_1 \neq \mathbf{A}_{1,0}, \tag{9.20}$$

where we partition the parameter functions as $\mathbf{A}(u) = (\mathbf{A}_1(u)^T, \mathbf{A}_2(u)^T)^T$. Under the null hypothesis H_0', the problem is still a varying-coefficient model. Let $\widehat{\mathbf{A}}_{2,0}$ be the local linear estimator using the same bandwidth h. Define the residual sum of squares as

$$\mathrm{RSS}_0^{**} = \sum_{i=1}^{n} \{Y_i - \mathbf{A}_{1,0}(U_i)^T \mathbf{X}_i^{(1)} - \widehat{\mathbf{A}}_{2,0}(U_i)^T \mathbf{X}_i^{(2)}\}^2,$$

where $\mathbf{X}_i$ is partitioned as $(\mathbf{X}_i^{(1)T}, \mathbf{X}_i^{(2)T})^T$. Following the same derivation as in Example 9.2, we obtain the GLR statistic

$$T_{n,3} = \frac{n}{2} \log(\mathrm{RSS}_0^{**}/\mathrm{RSS}_1). \tag{9.21}$$

Theorem 9.1 is still relevant to the derivation of the distribution of the test statistic $T_{n,3}$ under the null hypothesis H_0'. Decompose

$$T_{n,3} = \frac{n}{2} \log(\mathrm{RSS}_0/\mathrm{RSS}_1) - \frac{n}{2} \log(\mathrm{RSS}_0/\mathrm{RSS}_0^{**}).$$

The first term is just the GLR statistic for the problem (9.9), and the second term is the GLR statistic for the simple null hypothesis

$$H_0^{(2)} : \mathbf{A}_2 = \mathbf{A}_{2,0} \quad \longleftrightarrow \quad H_1^{(2)} \mathbf{A}_2 \neq \mathbf{A}_{2,0}$$

with the function $\mathbf{A}_1 = \mathbf{A}_{10}$ being given, where $\mathbf{A}_{20}$ is the true parameter function of $\mathbf{A}_2$. Using this observation, Fan, Zhang, and Zhang (2001) have derived the following result.

Theorem 9.2 *Under Condition* **(A)**, *when* $\mathbf{A}_{1,0}$ *is linear or* $nh^{9/2} \to 0$, *if* $nh^{3/2} \to \infty$, *then*

$$r_K T_{n,3} \overset{a}{\sim} \chi^2_{r_K c_K p_1 |\Omega|/h},$$

where p_1 *is the dimension of the vector* $\mathbf{A}_1$.

Once again, we unveil the Wilks phenomenon for problem (9.20). Although the asymptotic distribution provides a method to compute the approximate p-value, the conditional bootstrap seems to provide better approximations. Because of Theorem 9.2, we can in theory set the vector of nuisance functions $\mathbf{A}_2$ at any reasonable value in the conditional bootstrap since the null distribution depends sensitively on $\mathbf{A}_2$. In practice, we use the estimate from the alternative model, which is a consistent estimate of $\mathbf{A}_2$ under both null and alternative models. The conditional bootstrap is identical to the one outlined above, except in Step 1, where we construct the conditional bootstrap sample as

$$Y_i^* = \mathbf{A}_{10}(U_i)^T \mathbf{X}_i^{(1)} + \widehat{\mathbf{A}}_{20}(U_i)^T \mathbf{X}_i^{(2)} + \varepsilon_i^*,$$

where $\widehat{\mathbf{A}}_{20}$ is the nonparametric estimate from the alternative model. See also §9.4.1 for a similar algorithm. The bias reduction technique in §9.2.5 should be employed if $A_{1,0}$ is nonlinear.

9.2.7 Choice of Bandwidth

For each given smoothing parameter h, the GLR statistic $T_n(h)$ is a test statistic. This forms a family of test statistics indexed by h. In general, a

larger choice of bandwidth is more powerful for testing smoother alternatives, and a smaller choice of bandwidth is more powerful for testing less smooth alternatives. Depending on the background of applications, one can choose an appropriate size of bandwidth. An adaptive choice of bandwidth, inspired by the adaptive Neyman test of Fan (1996) (see §7.4), is to choose h to maximize the normalized test statistic

$$\widehat{h} = \text{argmax}_h \frac{T_n(h) - \mu_n(h)}{\sqrt{2\mu_n(h)}}$$

over a certain range of h, where $\mu_n(h)$ is the degree of freedom of the test statistic $T_n(h)$ (see, e.g., Theorem 9.1). This results in a *multiscale GLR* test statistic

$$\max_h \frac{T_n(h) - \mu_n(h)}{\sqrt{2\mu_n(h)}}.$$

The idea above appeared in Fan, Zhang, and Zhang (2001). Zhang (2002a) further studied the properties of the multiscale test statistic. In particular, she proposed methods to estimate the null distribution of the multiscale GLR test.

In practical implementations, one would find the maximum in the multiscale GLR test over a grid of bandwidths. Zhang (2002a) calculated the correlation between $T_n(h)$ and $T_n(ch)$ for some inflation factor c. The correlation is quite large when $c = 1.3$. Thus, a simple implementation is to choose a grid of points $h = h_0 1.5^j$ for $j = -1, 0, 1$, representing "small," "right," and "large" bandwidths. A natural choice of h_0 is the optimal bandwidth in the function estimation.

9.2.8 A Numerical Example

The Wilks phenomenon for the GLR statistic has been demonstrated for distributions of an *exponential family* (see McCullagh and Nelder 1989) by Fan, Zhang, and Zhang (2001) using asymptotic theory and by Cai, Fan, and Li (2000) using various simulation models. We use one of their examples to illustrate finite sample properties.

Example 9.3 (*Logistic regression*) We drew random samples of size $n = 400$ from the following nonparametric *logistic regression* model. The covariates X_1 and X_2 are standard normal random variables with correlation coefficient $2^{-1/2}$, and U is uniformly distributed over $[0, 1]$, independent of X_1 and X_2. Given X_1 and X_2, the conditional probability is

$$\begin{aligned} P(Y = 1|X_1, X_2, U) &= 1 - P(Y = 0|X_1, X_2, U) \\ &= \frac{\exp\{a_0(U) + a_1(U)X_1 + a_2(U)X_2\}}{1 + \exp\{a_0(U) + a_1(U)X_1 + a_2(U)X_2\}}. \end{aligned}$$

The conditional likelihood function is

$$P(Y = 1|X_1, X_2, U)^Y P(Y = 0|X_1, X_2, U)^{1-Y}.$$

Consequently, by substituting the conditional formula into the conditional likelihood function above, we obtain the log-likelihood as

$$\begin{aligned} \ell(a_0, a_1, a_2) &= \{a_0(U) + a_1(U)X_1 + a_2(U)X_2\}Y \\ &\quad - \log(1 + \exp\{a_0(U) + a_1(U)X_1 + a_2(U)X_2\}). \end{aligned}$$

Consider testing the hypothesis

$$H_0 : a_j(\cdot) = a_j, \quad j = 0, 1, 2.$$

The local likelihood estimator with the Epanechnikov kernel was used to construct the nonparametric estimation of the functions $a_0(\cdot), a_1(\cdot)$, and $a_2(\cdot)$. According to Table 7.1, the normalization constant $r_K = 2.1153$. Thus, the normalization was simply taken as $T_n = 2T_{n,1}$. To verify whether the distributions of T_n depend sensitively on the parameters under the null hypothesis, five different sets of values of $\{a_j\}$ were taken. They are quite far apart. The distributions of T_n were estimated by using 1,000 Monte Carlo simulations. The resulting estimates are depicted in Figure 9.1 and do not depend sensitively on the choices of parameters under the null hypothesis. This is consistent with the asymptotic theory.

To validate our conditional bootstrap method, five typical data sets were actually simulated from an alternative model

$$a_{0,0}(u) = \exp(2u - 1), \quad a_{1,0}(u) = 8u(1 - u), \quad a_{2,0}(u) = 2\sin^2(2\pi u).$$

The conditional bootstrap estimates of null distributions, based on 1,000 bootstrap samples, are plotted as thin curves in Figure 9.1. They are almost undifferentiable from the true distributions. This in turn demonstrates that our conditional bootstrap method works very reasonably, even when the null hypothesis is wrong.

Next, we examine the power of the GLR test. We evaluate the power of the test under the alternative model

$$H_1 : a_j(u) = \bar{a}_{j,0} + \beta(a_{j,0}(u) - \bar{a}_{j,0}), \quad j = 0, 1, 2,$$

where $\bar{a}_{j,0} = Ea_{j,0}(U)$ and β is a given parameter, measuring the distance between the null hypothesis and the alternative model. The GLR test was performed at five different significance levels: 0.01, 0.05, 0.10, 0.25, and 0.5. Figure 9.2 shows the power, based on 1,000 simulations, of the GLR test for various choices of β. In particular, when $\beta = 0$, the alternative hypothesis becomes the null hypothesis. The power should be approximately the same as the significance level. This was indeed the case. The powers for five different levels of tests when $\beta = 0$ are, respectively, 0.012, 0.047, 0.101,

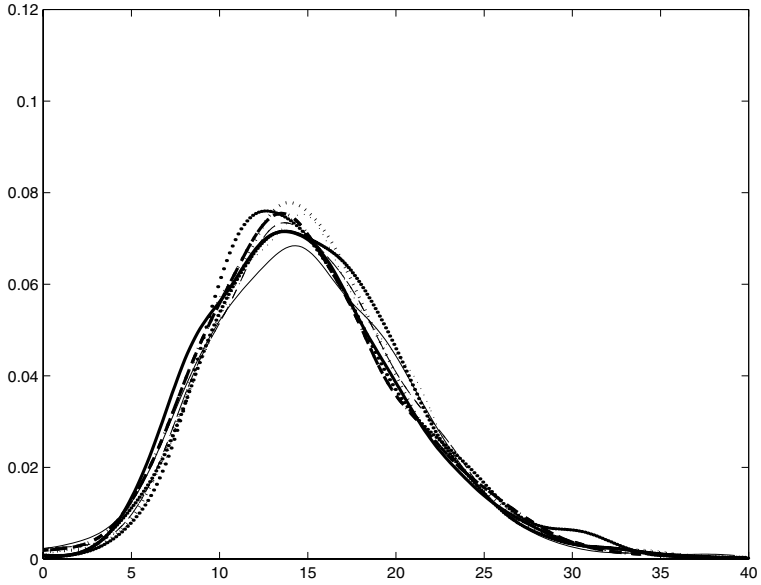


FIGURE 9.1. Null distributions of the GLR statistic T_n for five different sets of parameters (thick curves). The estimated distribution of T_n using the conditional bootstrap for five different sets of data were also depicted (thin curves). Adapted from Cai, Fan, and Li (2000).

0.281, and 0.532. This is another verification that the estimates using the conditional bootstrap approximate the null distribution very well. When the significance level is 5%, the power is approximately 0.8 when $\beta = 0.6$ and approximately 1 when $\beta = 0.8$. ∎

9.3 Tests on Spectral Densities

ARMA models are one of the most popularly used families of models in linear time series analysis. They provide powerful prediction tools and useful understanding on the probability law that governs the data-generation process. Yet, models need to be verified before meaningful conclusions can be drawn. One method for such a validation is to check whether the residuals from an ARMA model form a white noise process using the nonparametric technique in §7.4. It checks one important aspect of the model. An alternative method is to check whether the spectral density of the ARMA model is significantly different from the nonparametric estimates in §7.2 and §7.3. This gives a different aspect of model diagnostics and forms the

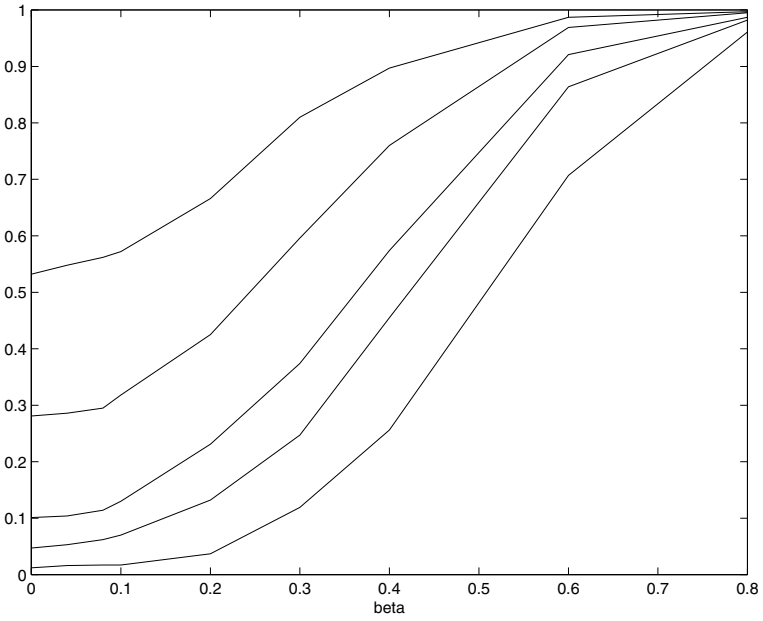


FIGURE 9.2. Power of the GLR statistic T_n at significance levels 0.01, 0.05, 0.10, 0.25, and 0.5 based on 1,000 simulations for different choices of β. Reproduced from Cai, Fan, and Li (2000).

subject of this section. It checks whether the autocovariance structure of an underlying process is consistent with that of an ARMA model.

As a generalization, we may test whether the spectral density admits a certain specific form,

$$H_0 : g = g_\theta \quad \longleftrightarrow \quad H_1 : g \neq g_\theta, \tag{9.22}$$

where g_θ is a given parametric family of spectral densities. Recall that the spectral density characterizes the autocovariance structure of a stationary time series. It governs entirely the stochastic dynamic of a stationary time series when the stochastic process is Gaussian. The testing problem (9.22) is really on the issue of whether the autocovariance structure of a given time series is of a certain specific parametric form. Note that nonlinear time series such as GARCH and bilinear processes admit the same forms of spectral densities as (linear) ARMA processes. Note also that if a time series $\{X_t\}$ is stationary, then its transformed series such as $\{X_t^2\}$ is also stationary. Testing (9.22) based on the transformed series checks a different model on $\{X_t\}$. Hence, the scope of the applicability is wider than what we present here.

9.3.1 Relation with Nonparametric Regression

Let $\{X_1, \cdots, X_T\}$ be a sequence of observed time series with spectral density g. As in §7.1, let $Y_k = \log I_T(\omega_k)/(2\pi)$ be the logarithm of the periodogram at frequency $\omega_k = 2\pi k/T$. Then, from (7.2), we have

$$Y_k = m(\omega_k) + z_k + r_k, \quad k = 1, \cdots, n, \tag{9.23}$$

where $m(\cdot) = \log g(\cdot)$, $n = [(T-1)/2]$, r_k is an asymptotically negligible term, and $\{z_k\}$ is a sequence of i.i.d. random variables having density

$$f_z(x) = \exp(-\exp(x) + x). \tag{9.24}$$

Ignoring the small order r_k, the model (9.23) is a nonparametric model, with the error distribution given by (9.24). The testing problem (9.22) becomes testing whether the "regression function" $m(\cdot)$ is of form $m_\theta = \log(g_\theta)$.

Testing problem (9.22) has a number of applications. It can be applied to testing whether an underlying stochastic process follows an ARMA model. It can also be employed to verify other models by checking whether a residual series follows a white noise process.

Suppose that we wish to test whether $\{X_1, \cdots, X_T\}$ follows an ARMA(p, q) model (2.46):

$$X_t - b_1 X_{t-1} - \cdots - b_p X_{t-p} = \varepsilon_t + a_1 \varepsilon_{t-1} + \cdots + a_q \varepsilon_{t-q}. \tag{9.25}$$

Then, it has spectral density [see (2.47)]

$$g_\theta(\omega) = \frac{\sigma^2}{2\pi} \frac{|1 + a_1 \exp(-i\omega) + \cdots + a_q \exp(-iq\omega)|^2}{|1 - b_1 \exp(-i\omega) - \cdots - b_p \exp(-ip\omega)|^2},$$

where $\theta = (a_1, \cdots, a_q, b_1, \cdots, b_p, \sigma)$ is a vector of parameters. In the spectral domain, validating an ARMA(p, q) model becomes the hypothesis-testing problem (9.22). When the deviation is evidenced, it is clear that the underlying series cannot be well-modeled by an ARMA process. If the deviation from the null hypothesis in (9.22) is not substantiated, one can only conclude that the series has the same autocovariance structure as the ARMA model. This is a drawback of the procedure. However, for Gaussian processes, the validation of H_0 implies that the process is ARMA.

For a nonlinear time series, it is frequently assumed that stochastic errors are normally distributed. Thus, the idea above can be applied to a residual series to check whether it is a white noise series. This significantly expands the scope of the application of this approach.

9.3.2 Generalized Likelihood Ratio Tests

We now employ the GLR test on the problem (9.22). As discussed in §9.2.5, to reduce the biases under the null hypothesis, we need to reparameterize

the problem. Let θ be an estimated value under the null hypothesis. For example, it can be estimated by maximizing the Whittle likelihood

$$\sum_{k=1}^{n}[-\exp\{Y_k - m_\theta(\omega)\} + Y_k - m_\theta(\omega_k)]. \tag{9.26}$$

This is an approximate likelihood in the frequency domain. In the case of testing an ARMA model, we can also use the maximum likelihood method to estimate the parameters in the ARMA(p, q) model (9.25). In general, we require only that θ be estimated at $O(n^{-1/2})$. Let $g_{\widehat{\theta}}$ be the estimated spectral density $g_{\widehat{\theta}}$. Set

$$Y_k^* = Y_k - \log(g_{\widehat{\theta}}(\omega_k)).$$

Then, the problem is reduced to testing

$$H_0 : m^*(x) = 0 \quad \longleftrightarrow \quad H_1 : m^*(x) \neq 0 \tag{9.27}$$

based on the data

$$Y_k^* \approx m^*(\omega_k) + z_k, \quad k = 1, \cdots, n, \tag{9.28}$$

where $m^*(\omega) = m(\omega) - \log(g_{\widehat{\theta}}(\omega))$. Ignoring the smaller order term in (9.28), the log-likelihood of $Y_1^*, \cdots, Y_n^*$ is

$$\ell(m^*) = \sum_{k=1}^{n}[Y_k^* - m^*(\omega_k) - \exp\{Y_k^* - m^*(\omega_k)\}]. \tag{9.29}$$

This is in fact the same as the Whittle likelihood in §7.3.2.

As in (7.21), the function m^* can be estimated by using the local-likelihood fit, resulting in $\widehat{m}^*$. Now, the GLR statistic is simply

$$\begin{aligned} T_{n,4} &= \ell(\widehat{m}^*) - \ell(0) \\ &= \sum_{k=1}^{n}[\exp(Y_k^*) - \exp\{Y_k^* - \widehat{m}^*(\omega_k)\} - \widehat{m}^*(\omega_k)]. \end{aligned} \tag{9.30}$$

The distribution of the GLR statistic can be estimated by using the conditional bootstrap method outlined in §9.2. To expedite the computation, we can simulate the sample directly from (9.28), and hence the p-value of the test can be obtained. More precisely, the schematic algorithm goes as follows:

1. Obtain the parametric estimate $\widehat{\theta}$ and compute $\{Y_k^*\}$.
2. Apply the local likelihood method in §7.3.2 to obtain an estimate $\widehat{m}^*$ and the selected bandwidth $\widehat{h}$.

3. Compute the observed test statistic $T_{n,4}$ as in (9.30).

4. Generate a random sample $\{z_k\}$ with the density (9.24), and set the bootstrap sample $Y_k^{**} = z_k$ (see (9.28)).

5. Use the local likelihood method in §7.3.2 with the bandwidth $\widehat{h}$, which is independent of the bootstrap sample, to obtain the local likelihood estimate $\widehat{m}^{**}$ and the bootstrap test statistic $T_{n,4}^*$.

6. Repeat steps 4 and 5 B times (say, 1,000 times) and obtain B bootstrap test statistics.

7. Estimate the p-value as the percentage of the bootstrap statistics that exceed the observed statistic $T_{n,4}$.

Note that z_k in step 4 can be generated from $z_k = \log(-\log(u_k))$, where $\{u_k\}$ are a sequence of random variables uniformly distributed on $(0, 1)$. This expedites the computational burden. If a more precise null distribution is needed, we can modify step 4 as follows. Generate a sequence of Gaussian white noise with $\sigma = 1$ and length T, and compute the logarithm of the periodogram of this series to obtain $\{Y_k^{**}\}$. For the bootstrap samples in step 5, we do not use the data-driven bandwidth. This avoids the variability and possible instability of the data-driven bandwidths.

A simple alternative method is to regard (9.28) as the least-squares problem:

$$Y_k^* - C_0 \approx m^*(\omega_k) + (z_k - C_0), \quad k = 1, \cdots, n,$$

where $C_0 = Ez_k = -.57721$ from (7.6). By using the local linear fit, we obtain the least-squares estimate

$$\widehat{m}_{\mathrm{LS}}^*(\omega) = \sum_{j=1}^{n} K_T\left(\frac{\omega - \omega_j}{h}, \omega\right)(Y_j^* - C_0),$$

where as in (7.16) K_T is the weight induced by the local linear fit. Let

$$\mathrm{RSS}_1 = \sum_{k=1}^{n}\{Y_k^* - C_0 - \widehat{m}_{\mathrm{LS}}^*(\omega_k)\}^2$$

be the residual sum of squares under the nonparametric model. Similarly, let

$$\mathrm{RSS}_0 = \sum_{k=1}^{n}(Y_k^* - C_0)^2.$$

Pretending that the distribution of z_k is normal, we obtain the GLR

$$T_{n,5} = \log(\mathrm{RSS}_0/\mathrm{RSS}_1).$$

FIGURE 9.3. Powers of the GLR statistics at significance levels 0.01, 0.05, and 0.1 based on 1,000 simulations. The left panel corresponds to the test statistics without bias correction, and the right panel is for the tests with bias correction. The solid lines are for the likelihood-based approach, the dotted lines are for the least-squares-based approach, and the dashed lines are for the residual-based approach.

According to Theorem 9.1,

$$r_K T_{n,5} \overset{a}{\sim} \chi^2_{r_K c_K/h}$$

if model (9.28) holds exactly. It is expected that the result continues to hold even with the approximate model (9.28). A similar result is expected to hold for the GLR statistic $T_{n,4}$. This is indeed shown in Fan and Zhang (2002).

It is interesting to compare the power of three testing procedures: the residual method discussed in §7.4 and the test statistics $T_{n,4}$ and $T_{n,5}$. It will also be of interest to study the accuracy of the conditional bootstrap method as an approximation to the null distribution. Furthermore, we may ask whether the bias reduction technique outlined in §9.2.5 provides any reasonable gains. These issues were studied in Fan and Zhang (2002). Their studies show that the Whittle likelihood-based methods are more powerful than the least-squares approaches and that the bias correction improves the power of the tests further. The residual-based GLR test and the likelihood-based GLR test with bias correction are the most powerful procedures among the methods they studied.

We use a simulated example to illustrate the accuracy and the power of the test statistics $T_{n,4}$ and $T_{n,5}$ and their bias-corrected versions. The example is adapted from Fan and Zhang (2002).

Example 9.4 (*autoregressive model*) We simulated 1,000 series of length 500 from the model

$$X_t = \Big\{0.8(1-\beta) + \beta g(X_{t-1})\Big\} X_{t-1} - 0.56 X_{t-2} + 0.6 X_{t-3} + \varepsilon_t$$

for several values of $\beta \in [0,1]$, where ε_t is a sequence of independent random variables from $N(0,1)$ and

$$g(x) = 0.95I(x \in [-5,0)) - 0.18xI(x \in [0,5]).$$

The model can be regarded as either the FAR(3,1) or AAR(3). Consider the null hypothesis

$$H_0 : X_t = b_1 X_{t-1} + b_2 X_{t-2} + b_3 X_{t-3} + \varepsilon_t,$$

namely, testing whether the generated series was from an AR(3) model. The test statistics $T_{n,4}$ and $T_{n,5}$ are employed. Their bias-corrected versions and residual-based method using the likelihood approach are also employed. In computing nonparametric estimates of spectral density, we applied the techniques in §7.3 using the Epanechnikov kernel with bandwidth $h = 0.23$. The tests were performed at three different significance levels: 0.01, 0.05, and 0.1. The null distributions were estimated based on the conditional bootstrap with number of repetitions B=10,000. The power of the test statistics is depicted in Figure 9.3 based on 1,000 simulations. Note that when $\beta = 0$, the generated time series is from an AR(3) model, and hence its power should be the same as the significance level. This is indeed the case, as shown in the figure. When β is farther away from zero, the model gets more deviated from an AR(3) model and hence its power increases. As anticipated, the least-squares-based method is less powerful than the likelihood-based method, while the bias-corrected version dominates uncorrected counterparts. The residual-based approach has about the same performance as the likelihood method with bias correction. ∎

9.3.3 Other Nonparametric Methods

Regarding model (9.23) as a nonparametric regression model, problem (9.22) is testing a family of parametric models against the nonparametric alternative. Hence, a wealth of nonparametric testing techniques can be employed; see, for example, Bowman and Azzalini (1997) and Hart (1997) for a good collection of procedures.

One can construct a test statistic based on a distance between a nonparametric estimator and a parametric estimator, resulting in a family of test statistics of form

$$\widehat{T}_1 = \|\widehat{m} - \log(g_{\widehat{\theta}})\|,$$

where $\widehat{m}$ is a nonparametric estimate of the log-spectral density and $\|\cdot\|$ is a certain norm. This family of test statistics includes

$$\widehat{T}_2 = \sup_{\tau \in (0,\pi)} |\widehat{m}(\tau) - \log(g_{\widehat{\theta}}(\tau))| w(\tau)$$

and

$$\widehat{T}_3 = \int_0^\pi |\widehat{m}(\tau) - \log(g_{\widehat{\theta}}(\tau))|^2 w(\tau) d\tau,$$

where $w(\cdot)$ is a given weight function. This idea appears in Bickel and Rosenblatt (1973).

As explained in §9.2.5, the estimator $\widehat{m}$ can be biased under the null hypothesis. The bias correction idea can be employed. Let $\widehat{m}^*(\cdot)$ be the estimator given in §9.3.2. Then, the bias-corrected version of the test statistics takes the form

$$\widehat{T}_4 = \|\widehat{m}^*\|.$$

In particular, when a weighted L_2 distance is used, the resulting test is basically the same as that in Härdle and Mammen (1993).

The Neyman test can also be applied to the testing problem (9.27). We outline the basic idea of the Neyman test below. First, expand the regression function in (9.28) into the Fourier series

$$m^*(\omega) = \sum_{j=0}^{k} \beta_j \phi_j(\omega),$$

where

$$\phi_0(\omega) = \frac{1}{\sqrt{\pi}}, \phi_{2j-1}(\omega) = \sqrt{\frac{2}{\pi}} \sin(2j\omega), \phi_{2j}(\omega) = \sqrt{\frac{2}{\pi}} \cos(2j\omega), j = 1, 2, \cdots$$

is a family of the orthonormal basis in $L^2(0, \pi)$. Apply the least-squares method

$$\min_{\beta} \sum_{l=1}^{n} \left\{ Y_l^* - \sum_{j=0}^{k} \beta_j \phi_j(\omega_l) \right\}^2$$

to obtain the estimated coefficients $\widehat{\beta}_1, \cdots, \widehat{\beta}_k$. Let $\mathbf{X}$ be the design matrix with (l, j) element $\phi_j(\omega_l)$. Then, it is an orthogonal matrix with

$$\sum_{l=1}^{n} \phi_j^2(\omega_l) = \frac{2}{\pi} \frac{n}{2} = \frac{n}{\pi}.$$

Note that $\mathrm{Var}(Y_k^*) \approx \pi^2/6$ by (7.17). Pretending that model (9.28) holds exactly, from the least-squares theory, we have

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \frac{\pi^2}{6} (\mathbf{X}^T \mathbf{X})^{-1} = \frac{\pi^3}{6n} I_k,$$

where $\widehat{\boldsymbol{\beta}}$ is the vector of the least-squares coefficients and I_k is the identity matrix of order k. Under the null hypothesis, using the asymptotic theory of the classical linear model,

$$\widehat{\beta}_j \overset{a}{\sim} N\left(0, \frac{\pi^3}{6n}\right)$$

for $0 \le j \le a_n$, with the given sequence a_n diverging to infinity at a certain rate. Further, these coefficients are asymptotically independent. The Neyman test is to construct the χ^2 statistic

$$\widehat{T}_5(k+1) = \frac{6n}{\pi^3} \sum_{j=0}^{k} \widehat{\beta}_j^2,$$

which aims at testing the null hypothesis

$$H_0 : \beta_1 = \cdots = \beta_k = 0.$$

The test statistic $\widehat{T}_5(k+1)$ is called the Neyman statistic.

There is much literature on the choice of the parameter k in the Neyman test. For example, Fan (1996) proposed the following choice of k, which maximizes the normalized partial sum process

$$\frac{\widehat{T}_5(k+1) - (k+1)}{\sqrt{2(k+1)}},$$

leading to the adaptive Neyman test

$$\max_{0 \le k \le a_n} \frac{T_5(k+1) - (k+1)}{\sqrt{2(k+1)}}$$

for some given sequence a_n; see also §7.4.3. Kuchibhatla and Hart (1996) proposed the following adaptive version of the Neyman test:

$$\max_{0 \le k \le a_n} \frac{\widehat{T}_5(k+1)}{k+1}.$$

It is shown by Fan and Huang (2001) that Fan's version tends to test more dimensions (selecting a larger k) than that of Kuchibhatla and Hart and that Fan's version is adaptively optimal in terms of the rate of convergence. Proposals to use cross-validation and other model selection techniques such as the AIC and BIC have also been suggested in the literature; see, for example, Eubank and Hart (1992), Inglot and Ledwina (1996), and Kallenberg and Ledwina (1997).

9.3.4 Tests Based on Rescaled Periodogram

The techniques in §9.3.1–§9.3.3 are based on the log-periodogram. One can also construct tests directly based on a periodogram. To reduce the biases for nonparametric estimates, we first employ the bias reduction technique. In the spectral density (rather than the logarithm of the spectral density in §9.3.3) domain, this is equivalent to considering the function

$$g^*(\omega) = g(\omega)/g_{\widehat{\theta}}(\omega)$$

and testing

$$H_0 : g^*(\cdot) = 1 \quad \longleftrightarrow \quad g^*(\cdot) \neq 1.$$

The function $g^*(\omega)$ can be estimated by smoothing on the rescaled periodogram $\frac{I_T^*(\omega_k)}{g_\theta(\omega_k)}$. Let $\widehat{g}^*$ be the resulting estimate using a nonparametric technique. Then, one can construct a test statistic of form $\|\widehat{g}^* - 1\|$.

As an example, Paparoditis (2000) considered the estimate of $g^* - 1$ based on the Priestley and Chao estimator:

$$\widehat{g}^*(\tau) - 1 = \frac{1}{2n+1} \sum_{j=-n}^{n} K_h(\tau - \omega_j) \left(\frac{I_T^*(\omega_k)}{g_\theta(\omega_k)} - 1 \right).$$

He constructed a test statistic based on an L_2-distance, resulting in the following test statistic

$$\widehat{T}_6 = (2n+1) \int_{-\pi}^{\pi} \{\widehat{g}^*(\tau) - 1\}^2 d\tau.$$

Under certain regularity conditions, it is shown by Paparoditis (2000) that under the null hypothesis

$$\sigma(K,h)^{-1}\{\widehat{T}_6 - \mu(K,h)\} \xrightarrow{D} N(0,1),$$

where $\mu(K,h) = 2h^{-1}\pi\|K\|^2$ and

$$\sigma(K,h) = \pi^{-1}h^{-1/2} \int_{-2\pi}^{2\pi} \left\{ \int_{-\pi}^{\pi} K(u)K(u+x)du \right\}^2 dx.$$

Hence, an asymptotic level α test is of rejection region

$$\widehat{T}_6 \geq \mu(K,h) + z_{1-\alpha}\sigma(K,h).$$

A way to avoid smoothing is to construct tests based on cumulated rescaled spectral density:

$$G(x) = \int_0^x g(\tau)/g_{\widehat{\theta}}(\tau)d\tau.$$

Under the null hypothesis, $G(x) = x$. Replacing the spectral density by the periodogram, we obtain a family of test statistics of form

$$\|\widehat{G}(\tau) - \tau\widehat{G}(\pi)/\pi\|, \quad x \in [0,\pi],$$

where

$$\widehat{G}(\tau) = \int_0^\tau I^*(x)/g_{\widehat{\theta}}(x)dx.$$

Here, the constant factor $\widehat{G}(\pi)/\pi \approx 1$ makes $\widehat{G}(\tau) - \tau\widehat{G}(\pi)/\pi = 0$ at both end points $\tau = 0$ and $\tau = \pi$. In particular, Dzhaparidze (1986) proposed the test statistic

$$\widehat{T}_7 = \frac{1}{2\pi^3} T \int_0^\pi \left\{\widehat{G}(\tau) - \tau\widehat{G}(\pi)/\pi\right\}^2 d\tau.$$

The asymptotic distribution of $\widehat{T}_7$ is the distribution of $\int_0^1 \{W(\tau) - \tau W(1)\}^2 d\tau$, where $\{W(\tau), 0 \le \tau \le 1\}$ is the Brownian motion (see Example 8.15). This asymptotic distribution is identical to that of the Cramér–von Mises test based on the integrated squared difference between the standardized sample spectral distribution function and the standardized spectral distribution of the model under H_0; see, for example, Anderson (1993).

We now briefly describe procedures of Anderson (1993). Let

$$F(\omega) = \frac{1}{\pi}\left\{\omega + 2\sum_{k=1}^{\infty} \rho(k)\sin(k\omega)/k\right\}, \quad \omega \in [0, \pi]$$

be the normalized spectral distribution (see also (2.35)), where $\rho(k)$ is the autocorrelation function. Let

$$\widehat{F}(\omega) = \frac{1}{\pi}\left\{\omega + 2\sum_{k=1}^{T-1} \widehat{\rho}(k)\sin(k\omega)/k\right\}$$

be an estimator of $F(\omega)$. Suppose that we wish to test

$$H_0 : F(\omega) = F_0(\omega)$$

for some given F_0. Anderson (1993) considered the Cramér–von Mises test of form

$$\frac{T}{2\pi c(f_0)^2} \int_0^\pi \{\widehat{F}(\tau) - F(\tau)\}^2 f_0^2(\tau) d\tau,$$

where $f_0(\tau) = F_0'(\tau)$ is a spectral density and $c(f_0) = 2\int_0^\pi f_0(\tau)^2 d\tau$, and the Kolmogorov–Smirnov test

$$\sup_{0 \le \omega \le \tau} \frac{\sqrt{T}}{2\sqrt{\pi c(f_0)}} |\widehat{F}(\tau) - F(\tau)|.$$

He derived the asymptotic null distribution of the test statistic.

In addition to the asymptotic distributions, as approximations to the null distribution of test statistics, the null distributions of the test statistics can also be approximated via the frequency domain bootstrap; see, for example, Paparoditis (2002) and references therein.

It is important to have some general understanding of the differences between the class of tests based on the spectral density and those based on cumulative spectral density. In general, the former tests give more weight to high frequency deviations from the null hypothesis, whereas the latter class of tests focuses mainly on the local frequency components. See Fan (1996) for more discussion on this subject.

9.4 Autoregressive versus Nonparametric Models

After fitting the FAR(p, d) model,

$$X_t = a_1(X_{t-d})X_1 + \cdots + a_p(X_{t-d})X_{t-p} + \sigma(X_{t-d})\varepsilon_t, \tag{9.31}$$

we frequently ask whether the coefficient functions are really varying. This is equivalent to testing the hypothesis

$$H_0 : a_1(\cdot) = a_1, \cdots, a_p(\cdot) = a_p, \tag{9.32}$$

where parameters $a_1, \cdots, a_p$ are unspecified. Under the null hypothesis (9.32), the series $\{X_t\}$ is an AR(p) model. Thus, the problem is equivalent to testing an AR(p) model against the FAR(p, d) model. Note that the problem here is different from that outlined in the last section since the alternative model in this problem is more specific. Hence, it should be expected that the resulting test statistics are more powerful than those introduced in the last section.

A similar question arises after fitting the autoregressive additive model:

$$X_t = f_1(X_{t-1}) + \cdots + f_p(X_{t-p}) + \varepsilon_t. \tag{9.33}$$

We are interested in testing whether the AR(p) model is adequate for the given data. This is equivalent to testing

$$H_0 : f_1(x) = a_1 x, \cdots, f_p(x) = a_p x. \tag{9.34}$$

Again, the alternative model is structured as (9.33). Thus, the tests designed for this particular setting should be more powerful than those generic tests introduced in §9.3.

An alternative view of the problems (9.32) and (9.34) is that we wish to validate an AR(p) model, and we embed it in the structured alternative models. One possibility is to use the FAR(p, d) model as a family of alternative models, assuming implicitly that the latter contains a model that fits a given time series reasonably well. Another possibility is to use the additive model as an alternative model. With the structured alternative models, one can have higher power in discriminating the null and alternative models.

9.4.1 Functional-Coefficient Alternatives

By introducing lagged variables as the covariates as in (8.3), the FAR(p, d) can be written as the varying-coefficient form (9.8). Hence, following §8.3.4, one can employ the local linear fit to obtain the estimated coefficient functions, resulting in $\widehat{a}_1(\cdot), \cdots, \widehat{a}_p(\cdot)$. Define

$$\text{RSS}_1 = \sum_{t=p+1}^{T} \{X_t - \widehat{a}_1(X_{t-d})X_{t-1} - \cdots - \widehat{a}_p(X_{t-d})X_{t-p}\}^2. \tag{9.35}$$

Similarly, let $\widehat{a}_1, \cdots, \widehat{a}_p$ be the estimated coefficients under the AR(p) model using the maximum likelihood technique introduced in §3.3. Define the residual sum of squares as

$$\mathrm{RSS}_0 = \sum_{t=p+1}^{T} \{X_t - \widehat{a}_1 X_{t-1} - \cdots - \widehat{a}_p X_{t-p}\}^2. \tag{9.36}$$

Then, the GLR statistic is $(n/2)\log(\mathrm{RSS}_0/\mathrm{RSS}_1)$, which under contiguous alternatives has the approximation

$$\log\left(\frac{\mathrm{RSS}_0}{\mathrm{RSS}_1}\right) = \log\left\{1 + \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1}\right\} \approx (\mathrm{RSS}_0 - \mathrm{RSS}_1)/\mathrm{RSS}_1$$

by Taylor's expansion. We therefore define the test statistic as

$$T_{n,6} = \frac{n}{2}(\mathrm{RSS}_0 - \mathrm{RSS}_1)/\mathrm{RSS}_1, \tag{9.37}$$

where $n = T - p$.

For theoretical and practical considerations, the functions $\widehat{a}_j(\cdot)$ can not be estimated well at the tails of the distribution of $\{X_{t-d}\}$. Thus, we may wish to restrict computing RSS_1 and RSS_0 to those cases where X_{t-d} falls in a prescribed set Ω. Following Theorem 9.1, we would expect that

$$r_K T_{n,6} \overset{a}{\sim} \chi^2_{r_K c_K p|\Omega|/h}.$$

In other words, the asymptotic null distribution is independent of the nuisance parameters. Hence, the conditional bootstrap can be employed to approximate the null distribution of the test statistic $T_{n,6}$.

Since we do not impose a restriction on the distribution of $\{\varepsilon_t\}$, we apply the conditional nonparametric bootstrap. In the time series context, the algorithm goes as follows.

1. Generate the bootstrap residuals $\{\varepsilon_i^*\}_{i=1}^n$ of the empirical distribution of the centered residuals $\{\widehat{\varepsilon}_i - \bar{\widehat{\varepsilon}}\}_{i=1}^n$ from the FAR(p, d) model, where $\bar{\widehat{\varepsilon}}$ is the average of $\{\widehat{\varepsilon}_i\}$. Construct the bootstrap sample: $X_{t,1}^* = X_{t-1}, \cdots, X_{t,p}^* = X_{t-p}$ and

$$Y_t^* = \widehat{a}_1 X_{t-1} + \cdots + \widehat{a}_p X_{t-p} + \varepsilon_t^*$$

for $t = p+1, \cdots, T$.

2. Calculate the test statistic $T_{n,6}^*$ based on the bootstrap sample

$$\{(X_{t,1}^*, \cdots, X_{t,p}^*, Y_t^*), t = p+1, \cdots, T\}.$$

This step computes the test statistic $T_{n,6}^*$ as if the data came from a regression model. More precisely, by applying the linear regression

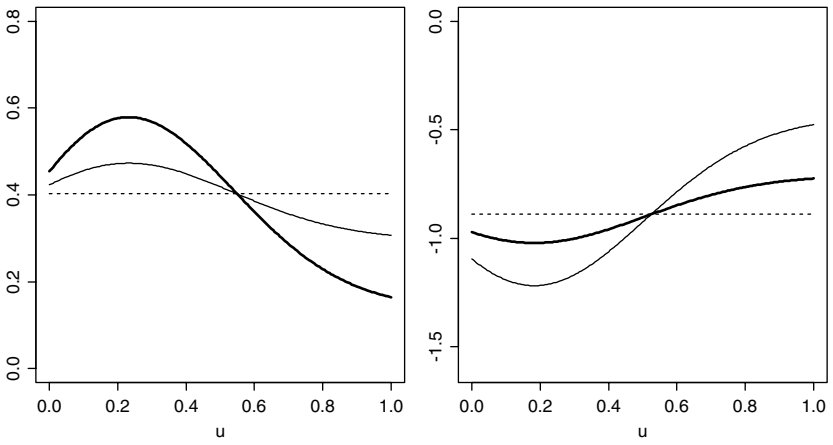


FIGURE 9.4. Functions $a_1(\cdot)$ (left-panel) and $a_2(\cdot)$ (right-panel) for $\beta = 0$ (dashed lines), $\beta = 0.4$ (solid curves), and $\beta = 1$ (thick curves).

technique to the bootstrap sample, we obtain the residual sum of squares RSS_0^* under the null hypothesis; by applying the technique as in (8.7) with $U_t^* = X_{t-d}$, we obtain the residual sum of squares RSS_1^* under the nonparametric model. This gives

$$T_{n,6}^* = \frac{T-p}{2}(\text{RSS}_0^* - \text{RSS}_1^*)/\text{RSS}_1^*.$$

3. Repeat the two steps above B times, and use the empirical distribution of $\{T_n^*\}$ as an approximation to the null distribution of the GLR statistic $T_{n,6}$.

4. Use the percentage of $\{T_{n,6}^*\}$ greater than the test statistic $T_{n,6}$ as an estimate of the p-value of the test.

We now use two examples from Cai, Fan, and Yao (2000) to illustrate the procedure.

Example 9.5 (*EXPAR model*) Consider the FAR$(2,1)$ model with $\varepsilon_t \sim N(0, 0.2^2)$. We wish to examine the power of the test statistic $T_{n,6}$ for the alternative model

$$a_j(u) = \bar{a}_j + \beta(a_{j,0}(u) - \bar{a}_j), \quad j = 1, 2,$$

where $\{a_{j,0}(\cdot), j = 0, 1, 2\}$ are the functions given in Example 8.4 and $\{\bar{a}_j, j = 0, 1, 2\}$ are their average heights (see Figure 9.4). Figure 9.4 displays the functions $a_1(\cdot)$ and $a_2(\cdot)$ for $\beta = 0, 0.4$, and 1. As in Example 9.3, β is related to the distance between the null hypothesis and the alternative

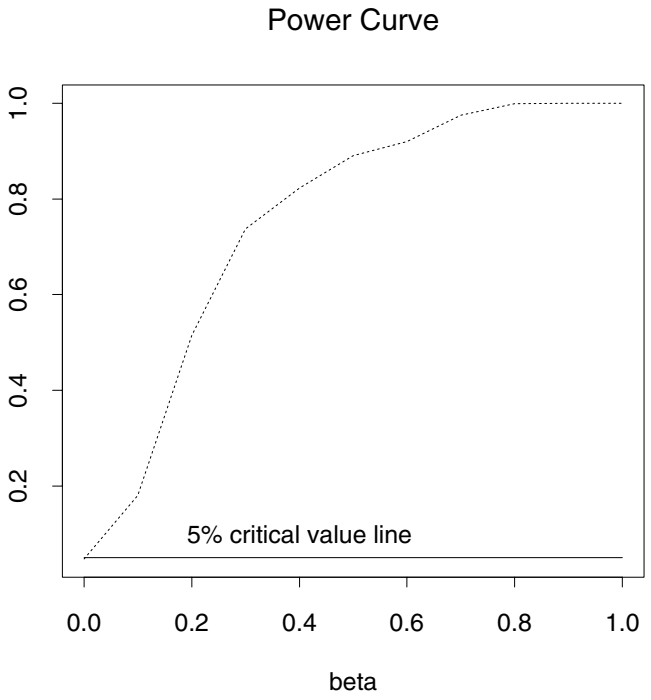


FIGURE 9.5. Power of the GLR statistic T_n at the significance level 5% based on 500 simulations for different choices of β. Adapted from Cai, Fan, and Yao (2000).

hypothesis. In particular, when $\beta = 0$, the alternative hypothesis becomes the null hypothesis. Thus, we can examine whether the power of the GLR test is approximately the same as the significance level 5%. This verifies whether the conditional nonparametric bootstrap gives reasonable approximations to the null distributions.

The power was computed based on 400 simulations with length $T = 400$. For each realization, 500 bootstrap samples were drawn to compute the p-value of the GLR test. The significance level was taken as 5%. When the P-value is less than 5%, the null hypothesis is rejected. The power is computed as the percentage chance of rejecting the null hypothesis among 400 simulations. For the nonparametric estimate, the Epanechnikov kernel with the bandwidth $h = 0.41$ was employed. The results are summarized in Figure 9.5.

When $\beta = 0$, the power is 4.7%, which is very close to the significance level 5%. This demonstrates that the bootstrap estimate of the null distribution gives a very good approximation of the size of the test. When $\beta = 0.4$, the power is already 80%. The power increases rapidly to 1 as β increases. This shows that the GLR test is very powerful. When $\beta = 0.8$, the power is already 100%. ∎

Example 9.6 (*Canadian lynx data*) After fitting the FAR(2, 2) model to the Canadian lynx data as in Example 8.6, we would naturally question whether the AR(2) model is adequate for the data. Applying the GLR test with the nonparametric estimates given by Figure 8.4, $\text{RSS}_1 = 4.5785$. On the other hand, under the null hypothesis, the estimated AR(2) is

$$X_t = 1.0732 + 1.3504X_{t-1} - 0.7200X_{t-2} + \varepsilon_t, \tag{9.38}$$

where $\varepsilon \sim N(0, 0.2275^2)$, resulting in $\text{RSS}_0 = 5.7981$. This gives the test statistic $T_{n,6} = 14.9174$. The estimated coefficients are slightly different from those in Example 8.6 since the whole series, rather than only the first 102 data points, are used here. Based on 1,000 bootstrap samples, the p-value was estimated as 0%. This provides very strong evidence against the AR(2) model. The result reinforces the existence of nonlinearity in the lynx data. We have written an S-Plus code "Ex96test.s" to implement the GLR test. ∎

9.4.2 Additive Alternatives

The aforementioned GLR test can be applied directly to test AR(p) model (9.34) against AAR(p) model (9.33). Let $\widehat{f}_1, \cdots, \widehat{f}_p$ be the estimated functions under the AAR(p) model using the techniques in §8.5 (e.g., the backfitting algorithm). Then, the residual sum of squares under the AAR(p) model is simply

$$\text{RSS}_1 = \sum_{t=p+1}^{T} \{X_t - \widehat{f}_1(X_{t-1}) - \cdots - \widehat{f}_p(X_{t-p})\}^2. \tag{9.39}$$

The GLR test statistic is

$$T_{n,7} = (n/2)\log(\text{RSS}_0/\text{RSS}_1), \tag{9.40}$$

where RSS_0 is given by (9.36). The conditional nonparametric bootstrap in §9.4.1 can similarly be employed here to compute the p-value of the test statistic $T_{n,7}$.

As we mentioned in §9.1 for all hypothesis-testing problems, we implicitly assumed that the alternative hypothesis contains a model that reasonably fits a given data set. When this is violated, the result can be misleading. We use the following example to elucidate the point.

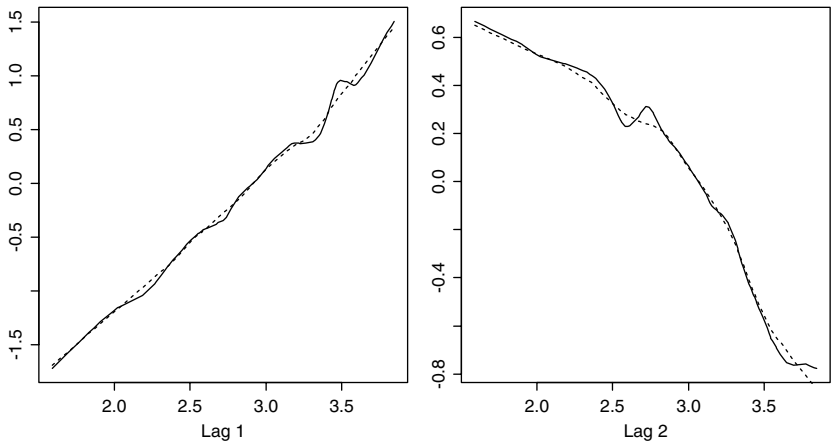


FIGURE 9.6. Estimated functions $\widehat{f}_1(\cdot)$ (left-panel) and $\widehat{f}_2(\cdot)$ (right-panel) for the Canadian lynx data using the backfitting algorithm with bandwidth selected by (8.51) (solid curves) and by the user (dashed curves).

Example 9.7 *(Canadian lynx data, revisited)* Following Tjøstheim and Auestad (1994a), we are curious how well the AAR(2) model fits the Canadian lynx data. We apply the backfitting technique introduced in §8.5 with bandwidth selected by (8.51) to estimate the functions f_1 and f_2. The resulting estimates are plotted in Figure 9.6. The RSS of the AAR(2) model is $\text{RSS}_1 = 3.3947$. Compared with the FAR$(2,1)$ fit, which has $\text{RSS}_1 = 4.5785$, the AAR(2) model fits the series better, although two models could have used different amounts of smoothing parameters. For example, if we increase the smoothing parameter in the AAR(2) model by 75%, the RSS_1 now increases to 4.1762.

Without reference to the TAR(2) model or FAR$(2,1)$ model, Figure 9.6 would reveal that an AR(2) model might be adequate for the lynx data. This leads to testing the problem (9.34), resulting in the estimated AR(2) model (9.38) with $\text{RSS}_0 = 5.7981$. Thus $T_{n,7} = 29.978$. By using 1,000 conditional nonparametric bootstrap samples, we obtained 1,000 test statistics $T^*_{n,7}$. Their average and the variance are 13.2130 and 12.5158, respectively. Normalize $T^*_{n,7}$ as

$$T'_{n,7} = 2\frac{13.2130}{12.5158}T^*_{n,7} = 2.1114T^*_{n,7}$$

so that its average is half its variance, a property possessed by the χ^2-distribution. Figure 9.7 shows the estimated distribution, using the kernel density estimator (5.1) with the Gaussian kernel and bandwidth (5.9), of the 1,000 normalized bootstrapped statistics $T'_{n,7}$. Note that the average of the 1,000 statistics $T'_{n,7}$ is 27.8979. Thus, its distribution can be ap-

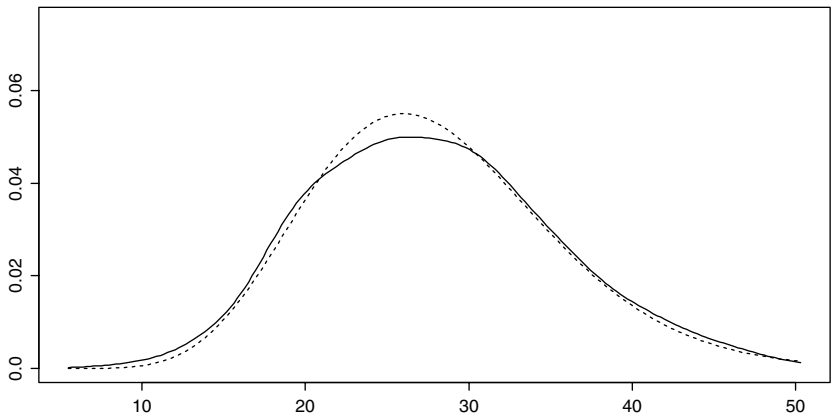


FIGURE 9.7. Estimated density (solid curve) of 1,000 normalized GLR statistics from bootstrap samples and the density of a χ^2-distribution with 28 degrees of freedom.

proximated by the χ^2-distribution with 28 degrees of freedom. As shown in Figure 9.7, the approximation is quite good. This supports the asymptotic theory. The p-value is estimated as 0. Hence, the AR(2) model does not fit the lynx data. The numerical implementation was carried out by the S-Plus code "Ex97test.s."

The complexity of alternative models depends on the choice of the bandwidth. When the bandwidth is excessively large, the family of alternative models becomes small and hence might not contain the true model. For the AAR(2) model, when the bandwidth is very large, it becomes close to an AR(2) model when the local linear fit is used since the local linear fit now becomes basically a global linear model. In this case, the alternative family of models may not necessarily contain the true model; for example, if we do not use (8.51) to select the bandwidth but choose the one that is about 75% larger than the one selected by (8.51). Then, the resulting fitted functions are given in Figure 9.6 (dashed curves). This gives $\text{RSS}_1 = 4.1762$, resulting in the test statistic $T_{n,7} = 18.3741$. By using the conditional nonparametric bootstrap, we obtain the estimated p-value 8.4%. Hence, we would have concluded that the AR(2) model fits the data well. This wrong conclusion is due to the assumption that the AAR(2) model with the given bandwidth reasonably fits the data. A more conservative interpretation is that the AR(2) model reasonably fits the lynx data among the family of AAR(2) models with the bandwidth that was used. ∎

Example 9.7 illustrates a few interesting points. First, the results of tests depend on the choice of bandwidth in the AAR(p) model. However, the dependence is not strong as shown here since the bandwidth was artificially

inflated here. Second, when we embed a parametric model into a family of alternative models (both parametric or nonparametric), we need to have reasonable assurance that the family of alternative models is large enough to fit the data reasonably well. Finally, it seems that the AAR(2) model fits the lynx data reasonably well.

9.5 Threshold Models versus Varying-Coefficient Models

As indicated in the introduction and illustrated in the last section, when the alternative models do not contain a model that fits the data well, the testing results can be misleading. To verify a nonlinear model, we would naturally embed it into a nonparametric family of models. The saturated nonparametric model would have little discriminability power. Nonsaturated models provide an immediate trade-off between the two contradictory demands: modeling biases and power of tests.

To validate threshold models, we choose FAR(p, d) as the alternative family of models. The idea of the GLR test continues to apply. The TAR model (4.1) can more conveniently be written as

$$X_t = a_1(X_{t-d}, \boldsymbol{\theta})X_{t-1} + \cdots + a_p(X_{t-d}, \boldsymbol{\theta})X_{t-p} + \varepsilon_t, \tag{9.41}$$

where $\boldsymbol{\theta}$ is the vector of unknown parameters. Let $\widehat{\boldsymbol{\theta}}$ be the estimated parameters. Define the residual sum of squares as

$$\text{RSS}_0 = \sum_{t=p+1}^{T} \{X_t - a_1(X_{t-d}, \widehat{\boldsymbol{\theta}})X_{t-1} - \cdots - a_p(X_{t-d}, \widehat{\boldsymbol{\theta}})X_{t-p}\}^2.$$

Suppose that we take FAR(p, d) as the family of alternative models. Then, RSS_1 is given by (9.35). Define the GLR test statistic as

$$T_{n,8} = (n/2)\log(\text{RSS}_0/\text{RSS}_1) \tag{9.42}$$

or its approximation

$$T'_{n,8} = (n/2)\frac{\text{RSS}_0 - \text{RSS}_1}{\text{RSS}_1}, \tag{9.43}$$

where $n = T-p$. The distribution of $T_{n,8}$ can be obtained by the conditional nonparametric bootstrap as in §9.4.1. Hence, the p-value of the test statistic can be estimated.

The method above applies not only to threshold models but also to other parametric families such as the EXPAR model introduced in Example 8.2. We now apply the procedure to the Canadian lynx data and the sunspot data to check whether TAR models reasonably fit the data. The numerical results are from the work of Cai, Fan, and Yao (2000).

Example 9.8 (*Canadian lynx data and TAR(2) models*). The residual sum of squares under the TAR(2) model discussed in §4.1.4 is $\text{RSS}_0 = 4.6367$. The residual sum of squares under the FAR(2) model, which is the same as that in Example 9.6, is $\text{RSS}_1 = 4.5785$. This results in the GLR statistic

$$T_{n,8} = 114/2 \log(4.6367/4.5785) = 0.7200 \quad \text{or} \quad T'_{n,8} = 0.7116.$$

By using the conditional nonparametric bootstrap with $B = 500$ on the statistic $T'_{n,8}$, the p-value is estimated as 68.7%. Hence, the evidence against the TAR(2) model is very weak. In fact, the fitted values between the nonparametric fit and the TAR(2) model are undifferentiable (Figures 8.4(d)). The difference lies in the interpretation of whether the population dynamic changes radically or gradually. Based on the available data, these two models are undifferentiable.

As illuminated in Example 9.7, the foregoing analysis does not warrant that the TAR(2) model must be correct. It simply shows that the TAR(2) model fits the lynx data very reasonably among the family of FAR(2, 2) models. Given the fact that the AAR(2) model has a 35% smaller RSS than that of the FAR(2, 2) model, both with the optimally chosen bandwidths, it is conceivably possible that the FAR(2, 2) model does not fit the data. ∎

Example 9.9 (*Sunspot data and TAR models*) The sunspot data have been analyzed in Example 8.7. The FAR(8, 3) model was fitted. After deleting insignificant variables, a specific model of FAR(8, 3) was used, leading to model (8.19). Suppose that we wish to test whether model (8.18) is reasonable using model (8.19) as the alternative. This is equivalent to testing a TAR model against the FAR(8, 3) model with known coefficient functions $a_4(\cdot) = 0$, $a_5(\cdot) = 0$, and $a_7(\cdot) = 0$. The residual sums of squares under the null and alternative hypotheses are, respectively, $\text{RSS}_0 = 3.277$ and $\text{RSS}_1 = 2.932$, resulting in the test statistic

$$T_{n,8} = 6.1740 \quad \text{or} \quad T'_{n,8} = 6.531.$$

By applying the nonparametric bootstrap method to the statistic $T'_{n,8}$, the p-value was estimated as 45.4%. In other words, we have very little evidence against the model (8.18).

The same technique has been applied to testing the TAR model (8.20) against the FAR(11, 8) model, resulting in $\text{RSS}_0 = 3.685$ and $\text{RSS}_1 = 2.077$. The test statistic is

$$T_{n,8} = 31.8207 \quad \text{or} \quad T'_{n,8} = 42.9677.$$

Using the nonparametric bootstrap method on statistic $T'_{n,8}$, the p-value was estimated as 10.1%. Once again, we do not have any strong evidence against the null model (8.20).

Even though we have weak evidence against both TAR models, as discussed in Example 9.7, this does not imply that both TAR models correctly capture the underlying stochastic dynamic that produced the observed series. In particular, we have not verified whether the family of the FAR models contains a model that fits the data well. A more conservative interpretation of the results is that among the family of FAR(8, 3) models, model (8.18) seems to fit the data. Similarly, among the family of FAR(11, 8) models, the TAR model (8.20) seems quite reasonable. ∎

9.6 Bibliographical Notes

There are many collective efforts on hypothesis testing in nonparametric regression problems. Most of them focus on the one-dimensional setting and cannot easily be extended to non-saturated multivariate models; §7.6 gives some bibliographical notes on the development of hypothesis testing with nonparametric alternatives. For an overview and references, see the books by Bowman and Azzalini (1997) and Hart (1997).

Generalized likelihood ratio tests were introduced by Fan, Zhang and Zhang (2001). They can be applied to many nonsaturated multivariate models. They test one aspect of models by comparing the likelihood ratios of two competing classes of models. An alternative approach is to check whether the residuals are of any structure. The challenge is how to use the prior information on the nonsaturated multivariate models to improve the power of the generic nonparametric tests. Fan and Huang (2001) have made a start on this kind of problem.

There are other techniques designed for testing some specific nonsaturated nonparametric models. Various techniques have been proposed for testing additive structure when data points have some specific designs; see, for example, Berry (1993), Chen, Liu, and Tsay (1995), Eubank, Hart, Simpson, and Stefanski (1995), Gozalo and Linton (2000), and Gao, Tong, and Wolff (2002). Fan and Li (1996) considered tests for the significance of a subset of regressors and tests for the specification of the semiparametric functional form of the regression function. Chen, Härdle, and Li (2002) used empirical likelihood to construct a statistic for testing against a parametric family of autoregressive models. Aerts, Claeskens, and Hart (2000) constructed tests based on orthogonal series expansions using model selection criteria. Horowitz and Spokoiny (2001) studied nonparametric tests based on L_2-distances. The preceding two papers use the saturated nonparametric alternatives. Härdle, Sperlich, and Spokoiny (2001) considered problems of testing the form of additive components using a variation of a multiscale test and wavelets.

This page intentionally blank

10
Nonlinear Prediction

Forecasting the future is one of the fundamental tasks of time series analysis. Although linear methods, such as those introduced in §3.7, are useful, a prediction from a nonlinear point of view is one-step closer to reality. Anybody who has first-hand experience of the stock-market knows that we can forecast the future better at the right moment than at another time. Such common sense can be naturally reflected in nonlinear forecasting only! In this chapter, we first discuss the general properties of nonlinear prediction, paying particular attention to those features that distinguish nonlinear prediction from linear prediction. The sensitivity to initial condition, a key concept in deterministic chaos, plays an important role in understanding nonlinearity. Three types of predictors-namely point predictors, predictive intervals, and predictive distributions-constructed based on local regression will be presented.

10.1 Features of Nonlinear Prediction

10.1.1 Decomposition for Mean Square Predictive Errors

Let $X_1, \cdots X_T$ be observations from a time series process. Suppose that we have no information on the underlying process. It still makes sense to consider the problem of predicting future values $X_{T+1}, X_{T+2}, \cdots$ based on the observed data $X_T, X_{T-1}, \cdots$. To highlight what nonlinearity can do for prediction, we temporarily ignore the problem of estimating relevant unknown functions. This is similar to the approach in §3.7 in which we

assumed that the coefficients in linear models were known. Furthermore, we predict X_{T+m} $(m \geq 1)$ based on the last p observed values

$$\mathbf{X}_T \equiv (X_T, X_{T-1}, \cdots, X_{T-p+1})^\tau$$

only. The least squares predictor is defined as

$$f_{T,m}(\mathbf{X}_T) = \arg\inf_f E\{X_{T+m} - f(\mathbf{X}_T)\}^2, \tag{10.1}$$

where the infimum is taken over all (measurable) functions of $\mathbf{X}_T$. It is easy to see that

$$f_{T,m}(\mathbf{x}) = E(X_{T+m}|\mathbf{X}_T = \mathbf{x}); \tag{10.2}$$

see Proposition 3.2. Furthermore, the mean square predictive error of $f_{T,m}$ is

$$\begin{aligned} & E\{X_{T+m} - f_{T,m}(\mathbf{X}_T)\}^2 \\ = \; & E[E\{(X_{T+m} - f_{T,m}(\mathbf{X}_T))^2|\mathbf{X}_T\}] \\ = \; & E\{\mathrm{Var}(X_{T+m}|\mathbf{X}_T)\}, \end{aligned}$$

which is the average of conditional variances of X_{T+m} given $\mathbf{X}_T$. If $\{X_t\}$ is a linear AR(p) process with innovations satisfying conditions (3.34) and (3.37), the conditional variance

$$\sigma^2_{T,m}(\mathbf{x}) \equiv \mathrm{Var}(X_{T+m}|\mathbf{X}_T = \mathbf{x})$$

is a constant; see Proposition 3.4. This, however, is no longer true in general. In practice, we are concerned with how good the prediction is based on an observed and known value of $\mathbf{X}_T$. Therefore, the conditional mean square predictive error

$$E[\{X_{T+m} - f_{T,m}(\mathbf{x})\}^2|\mathbf{X}_T = \mathbf{x}] \; = \; \sigma^2_{T,m}(\mathbf{x})$$

is in fact a practically more relevant measure for the performance of the prediction. It reflects the reality that how well we can predict depends on where we are. We argue that this statement has a further implication that is not always fully appreciated in statistical literature. To this end, let $\mathbf{x}$ be the observed value of $\mathbf{X}_T$. Therefore, we predict X_{T+m} by $f_{T,m}(\mathbf{x})$. However, our observation is subject to an error, and the true and unobserved value of $\mathbf{X}_T$ is $\mathbf{x}+\boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is a small drift, counting for a measurement and/or an experimental error and so on. Now, the following decomposition theorem holds, which was first presented in Yao and Tong (1994a).

Theorem 10.1 *For the least squares m-step-ahead predictor $f_{T,m}(\mathbf{X}_T)$, it holds that*

$$\begin{aligned} & E[\{X_{T+m} - f_{T,m}(\mathbf{x})\}^2|\mathbf{X}_T = \mathbf{x} + \boldsymbol{\delta}] \\ = \; & \sigma^2_{T,m}(\mathbf{x} + \boldsymbol{\delta}) + \{f_{T,m}(\mathbf{x} + \boldsymbol{\delta}) - f_{T,m}(\mathbf{x})\}^2 \\ = \; & \sigma^2_{T,m}(\mathbf{x} + \boldsymbol{\delta}) + \{\boldsymbol{\delta}^\tau \dot{f}_{T,m}(\mathbf{x})\}^2 + o(||\boldsymbol{\delta}||^2), \end{aligned} \tag{10.3}$$

where $\dot{f}_{T,m}$ denotes the gradient vector of $f_{T,m}$. The second equality requires the condition that $\dot{f}_{T,m}$ be continuous in a small neighborhood of $\mathbf{x}$.

The proof of the theorem above is trivial. Note that the right-hand side of (10.3) may be written as

$$E\big([\{X_{T+m} - f_{T,m}(\mathbf{x}+\boldsymbol{\delta})\} + \{f_{T,m}(\mathbf{x}+\boldsymbol{\delta}) - f_{T,m}(\mathbf{x})\}]^2 \big| \mathbf{X}_T = \mathbf{x}+\boldsymbol{\delta}\big).$$

Now, the first equality in (10.3) follows from the above and the fact that

$$E\{X_{T+m} - f_{T,m}(\mathbf{x}+\boldsymbol{\delta})|\mathbf{X}_T = \mathbf{x}+\boldsymbol{\delta}\} = 0.$$

The decomposition (10.3) indicates that the goodness of the prediction is dictated by two factors: (a) the error due to the randomness in the system represented by conditional variance $\sigma^2_{T,m}(\mathbf{x}+\boldsymbol{\delta})$ and (b) the error caused by the drift $\boldsymbol{\delta}$ at the initial condition. Usually, the conditional variance

$$\sigma^2_{T,m}(\mathbf{x}+\boldsymbol{\delta}) = \sigma^2_{T,m}(\mathbf{x}) + O(||\boldsymbol{\delta}||)$$

is the dominant term. However, for some nonlinear process with very small stochastic noise (such as operational deterministic systems, treated in Yao and Tong 1998b), the error due to the drift $\boldsymbol{\delta}$ may no longer be negligible. We will see in §10.1.2 and §10.1.3 below that for nonlinear processes both types of errors may be amplified rapidly at some places in the state-space. Therefore, the prediction for the future depends crucially on where we are at present.

To highlight the essence of nonlinearity, we assume in the rest of this section that $\{X_t\}$ is generated from a simple model

$$X_t = f(X_{t-1}) + \varepsilon_t, \tag{10.4}$$

where $\{\varepsilon_t\} \sim \text{IID}(0, \sigma^2)$, and ε_t is independent of $\{X_{t-k},\ k \geq 1\}$. Note that we do not impose any stationarity condition on $\{X_t\}$ at this stage, which is only required when we need to estimate predictive functions later on (see also §3.7). Now, it is easy to see that under this model, due to the Markovian property,

$$f_{T,m}(\mathbf{x}) = E(X_{T+m}|X_T = x) \equiv f_m(x) \tag{10.5}$$

and

$$\sigma^2_{T,m}(\mathbf{x}) = \text{Var}(X_{T+m}|X_T = x) \equiv \sigma^2_m(x), \tag{10.6}$$

where x is the first component of $\mathbf{x}$. In particular, $f_{T,1}(\mathbf{x}) = f_1(x) = f(x)$ and $\sigma^2_{T,1}(\mathbf{x}) = \sigma^2_1(x) \equiv \sigma^2$. The decomposition (10.3) may be written as

$$E[\{X_m - f_m(x)\}^2|X_0 = x+\delta] = \sigma^2_m(x+\delta) + \{\delta \dot{f}_m(x)\}^2 + o(\delta^2). \tag{10.7}$$

10.1.2 Noise Amplification

For the linear AR(1) process with coefficient b and $|b| < 1$, it follows from Proposition 3.4 that the mean square error of the least squares m-step-ahead predictor is

$$\sigma^2 \sum_{j=0}^{m-1} b^{2j} = \sum_{j=0}^{m-1} b^{2j} \mathrm{Var}(\varepsilon_{T+1+j}). \tag{10.8}$$

Although it increases monotonically as m increases, the noise entering at a fixed time decays exponentially as m increases. As we will see below, the noise contraction is not always observed in nonlinear prediction.

To simplify the discussion, we further assume that in model (10.4) $|\varepsilon_t| \leq \zeta$ almost surely, where $\zeta > 0$ is a small constant. By Taylor expansion, it is easy to see that for $m \geq 1$

$$\begin{aligned} X_m &= f\{f(X_{m-2}) + \varepsilon_{m-1}\} + \varepsilon_m \\ &= f^{(2)}(X_{m-2}) + \dot{f}\{f(X_{m-2})\}\varepsilon_{m-1} + \varepsilon_m. \end{aligned}$$

Note that $X_{m-2} = f^{(m-2)}(X_0) + O(\zeta)$. Hence

$$X_m = f^{(2)}(X_{m-2}) + \dot{f}\{f^{(m-1)}(X_0)\}\varepsilon_{m-1} + \varepsilon_m + O_P(\zeta^2).$$

By applying iteratively the Taylor expansion above, we have

$$\begin{aligned} X_m &= f^{(m)}(X_0) + \varepsilon_m + \dot{f}\{f^{(m-1)}(X_0)\}\varepsilon_{m-1} + \cdots \\ &+ \left\{ \prod_{k=1}^{m-1} \dot{f}[f^{(k)}(X_0)] \right\} \varepsilon_1 + O_p(\zeta^2), \end{aligned}$$

where $\dot{f}$ denotes the derivative of f, and $f^{(k)}$ denotes the k-fold composition of f. Thus

$$\sigma_m^2(x) = \mathrm{Var}(X_m | X_0 = x) = \mu_m(x)\, \sigma^2 + O(\zeta^3), \tag{10.9}$$

where

$$\mu_m(x) = 1 + \sum_{j=1}^{m-1} \left\{ \prod_{k=j}^{m-1} \dot{f}[f^{(k)}(x)] \right\}^2. \tag{10.10}$$

The fact that $\sigma_m^2(x)$ varies with respect to x reflects that our ability to predict depends critically on where we are when the prediction is made. For linear processes, $\dot{f}(\cdot)$ is a constant. Therefore, both $\sigma_m^2(x)$ and $\mu_m(x)$ are constants.

The *noise amplification* is dictated by $\mu_m(x)$. The values μ_m are determined by those of the derivative $\dot{f}$. If $|\dot{f}(\cdot)| > 1$ defines a large subset of the state-space, $\mu_m(\cdot)$ could be large or very large for moderate or even

small m. The rapid increase of $\sigma_m^2(x)$ with respect to m is a manifestation of noise amplification. In such cases, only very-short-range prediction is practically meaningful. Thus, how far in the future we can predict also depends on where we are. Again, this is a well-known fact among many working forecasters, especially in fields such as meteorology.

For multistep linear prediction, the mean square predictive error monotonically increases as m increases; see, for example, (10.8). However, this is not always the case with nonlinear prediction. Note that, by (10.10), it holds that

$$\mu_{m+1}(x) = 1 + \mu_m(x)\dot{f}\{f^{(m)}(x)\}^2.$$

Thus $\mu_{m+1}(x) < \mu_m(x)$ if $\dot{f}\{f^{(m)}(x)\}^2 < 1 - 1/\mu_m(x)$. By (10.9), it is possible that for such x and m, $\sigma_m^2(x) > \sigma_{m+1}^2(x)$. This suggests that at some initial value the error of an $(m+1)$-step-ahead prediction could be smaller than that of the m-step-ahead prediction.

Example 10.1 Consider a simple quadratic model

$$X_t = 0.235X_{t-1}(16 - X_{t-1}) + \varepsilon_t, \tag{10.11}$$

where ε_t are independent and uniformly distributed on the interval $[-0.52, 0.52]$. The scatterplots of X_{t+m}, for $m = 2$ and 3, against X_t from a sample of size 300 are displayed in Figures 10.1 (a) and (b) together with the least squares predictor $f_m(\cdot)$ and the conditional variance $\sigma_m^2(\cdot)$. The amount of variation of the data varies with respect to the initial value, indicating that the predictive error depends on the initial condition. Furthermore, the variation is well-depicted by the conditional variance functions. For example, both $\sigma_2^2(x)$ and $\sigma_3^2(x)$ obtain their maximum values at $x = 8$, where the variation in both scatterplots is largest. Figure 10.1(c) plots the two conditional variance functions $\sigma_2^2(x)$ and $\sigma_3^2(x)$ together. Around $x = 5.4$ and 10.6, it holds that $\sigma_2^2(x) > \sigma_3^2(x)$. Thus, in those areas, three-step-ahead prediction is more accurate than two-step-ahead prediction. ∎

10.1.3 Sensitivity to Initial Values

The sensitivity to initial conditions is the key feature for chaotic behavior of nonlinear deterministic systems. A compact introduction on deterministic chaos is available in Chapter 2 of Chan and Tong (2001). We examine below the sensitivity to initial values in the context of point prediction. More specifically, we elaborate the term $\{\delta\dot{f}_m(x)\}^2$ in (10.7) further and reveal how it evolves in nonlinear stochastic dynamic systems. Like noise amplification, it also boils down to the property of the derivation $\dot{f}(\cdot)$.

Suppose that $\{X_t(x), t \geq 1\}$ is a trajectory of the process $\{X_t\}$ defined by (10.4) starting at $X_0 = x$. How will the two trajectories starting at nearby initial values, say x and $x + \delta$, diverge? Since different trajectories receive different random shocks (i.e., different realizations) from ε_t at each

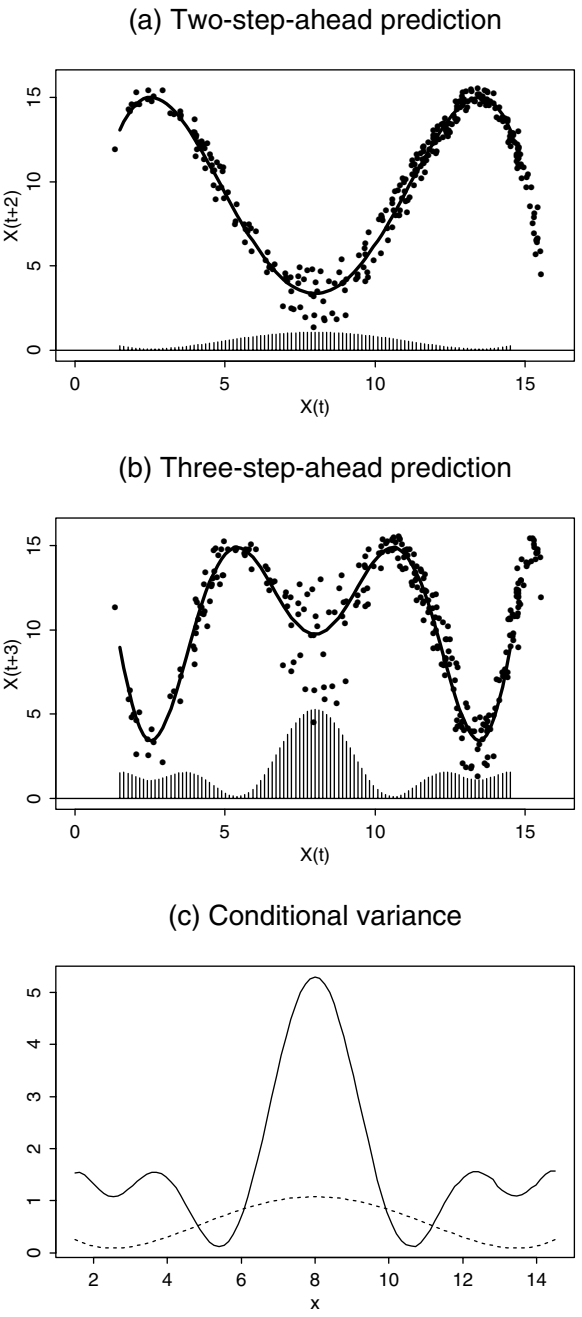


FIGURE 10.1. Scatterplots of X_{t+m} against X_t of a sample draw from model (10.11), together with the m-step-ahead predictor $f_m(\cdot)$ (thick solid curves) and conditional variance function $\sigma_m^2(\cdot)$ (impulses) for (a) $m = 2$ and (b) $m = 3$. (c) Plots of $\sigma_m^2(x)$ against x for $m = 2$ (dotted curve) and $m = 3$ (solid curve).

time t, it is more sensible to consider the divergence of the (conditional) expected values of those two trajectories; namely,

$$\begin{aligned} & E\{X_m(x+\delta)|X_0 = x+\delta\} - E\{X_m(x)|X_0 = x\} \\ &= f_m(x+\delta) - f_m(x) \;=\; \delta \dot{f}_m(x) + o(|\delta|). \end{aligned}$$

Now, we take a close look at how the derivative $\dot{f}_m(x) = \frac{d}{dx} f_m(x)$ evolves when m increases. Note that

$$\begin{aligned} f_m(x) &= E\{f(X_{m-1})|X_0 = x\} = E[f\{f(X_{m-2}) + \varepsilon_{m-1}\}|X_0 = x] \\ &= E\left(f[\cdots\{f(x) + \varepsilon_1\} + \cdots + \varepsilon_{m-1}]|X_0 = x\right). \end{aligned}$$

By assuming that the order of expectation and differentiation is interchangeable, it follows from the chain rule that

$$\dot{f}_m(x) = E\left\{ \prod_{k=1}^{m} \dot{f}(X_{k-1}) \middle| X_0 = x \right\}. \tag{10.12}$$

By assuming that all of the factors on the right-hand side of this expression are of comparable size, it is plausible that $\dot{f}_m(x)$ grows (or decays) exponentially with m. Again, the values of the derivative $\dot{f}$ are crucial. If $|\dot{f}(\cdot)| > 1$ on a large part of the state-space, $\dot{f}_m(x)$ could be substantially large for moderate or even small m.

Combining (10.9), (10.10), and (10.12), the first two terms on the right-hand side of (10.7) depend critically on the behavior of the derivative $\dot{f}(\cdot)$, which is the key factor dictating the quality of nonlinear prediction.

10.1.4 Multiple-Step Prediction versus a One-Step Plug-in Method

In m-step-ahead prediction, a frequently used strategy in practice is to repeat one-step-ahead prediction m times, treating the predicted value from the last round as the true value. We refer to this as a one-step plug-in method. This method is justified for the model-based linear prediction presented in §3.7; see (3.33) and (3.38). However, it is different from the least squares prediction in general. For example, for model (10.4), the one-step plug-in predictor for X_{T+m} based on X_T is $f^{(m)}(X_T)$, which is of course different from the least squares predictor $f_m(X_T) = E(X_{T+m}|X_T)$ unless $f(\cdot)$ is linear. Therefore

$$E[\{X_{T+m} - f^{(m)}(X_T)\}^2|X_T] \geq E[\{X_{T+m} - f_m(X_T)\}^2|X_T].$$

Hence, the one-step plug-in method is not desirable in principle.

The comparison of the two approaches above is almost purely theoretical, as we ought to estimate f or f_m in practice. Suppose that we adopt the

same method to estimate both f and f_m, resulting in estimators $\widehat{f}^{(m)}$ and $\widehat{f}_m$. Under standard regularity conditions, we may prove that both $\widehat{f}^{(m)}(X_T) - X_{T+m}$ and $\widehat{f}_m(X_T) - X_{T+m}$ will be asymptotically normal with the same convergence rate, but the former has a constant bias and the latter has a bias converging to 0. Therefore, again we should use the direct multistep-ahead prediction method.

Under some circumstances, physical laws allow us to formulate $f(\cdot)$ into a given functional form subject to a few unknown parameters. Such a formulation, however, is hardly available for f_m due to complex nonlinear dynamics. Therefore, typically f can be estimated globally through its parameters, and its estimator enjoys the convergence rate $T^{1/2}$. On the other hand, f_m can only be estimated locally, and the estimation has to contend with a slower convergence rate. Now, the picture is less clear. The advantage of using the predictor $f_m(X_T)$ may well be evened out by the larger errors incurred in the estimation of f_m. This partially explains why the model-based one-step plug-in method is still popular among practitioners.

It is also worth mentioning that the definition of the least squares predictor f_m is model-free. Therefore, it is robust against the misspecification of the model (i.e., $f(\cdot)$) in the first place. See a relevant study in Tiao and Tsay (1994).

In summary, the least squares m-step-ahead predictor $f_m(\cdot)$ is the right predictor to use in principle, although its performance may be hampered by the lack of efficient means to identify and to estimate it effectively in practice, especially when prediction is based on several observed lagged values (i.e., $p > 1$).

10.1.5 Nonlinear versus Linear Prediction

There is no reason why real-life generating processes should all be linear. However, time series forecasting is still very much dominated by *linear prediction methods* in the sense that the predicted values are the linear combinations of their observed lagged values. (Note that we require ARMA processes to be invertible in §3.7.4). This is partially due to both mathematical and practical convenience. But empirical studies indicate that linear methods often work well despite their simplicity, and the gain from nonlinear prediction is not always significant and sometimes is not even guaranteed; see §3.4.1 of Chatfield (2001) and the references therein. Although we should not take numerical comparisons on faith (see, §6.6.3 of Chan and Tong 2001), the robust performance of linear forecasting methods is undeniable. Since this issue has rarely been addressed explicitly in the literature, we provide an explanation below.

First, it is worth pointing out that the linear prediction method can in fact be applied to any time series as long as it has finite second moments; see §3.2. To simplify the discussion, let $\{X_t\}$ be a (weakly) stationary time series. We may seek the best linear predictor for X_t based on $\{X_{t-1},\ k \geq 1\}$

(i.e., the predictor that is a linear combination of $\{X_{t-k}, k \geq 1\}$ such that the mean square error attains the minimum). Then, by the Wold decomposition theorem (see, for example, (5.7.1) and (5.7.2) in Brockwell and Davis 1991),

$$X_t = e_t + \sum_{j=1}^{\infty} \psi_j e_{t-j} + V_t, \tag{10.13}$$

where $\{e_t\} \sim \text{WN}(0, \sigma^2)$ and

$$e_t = X_t - \sum_{i=1}^{\infty} \varphi_i X_{t-i}, \tag{10.14}$$

the coefficients $\{\psi_j\}$ and $\{\varphi_i\}$ are square-summable, and V_t is a deterministic component in the sense that it is entirely determined by its lagged values (and hence can be predicted relatively easily). Note that the AR(∞) representation (10.14) holds for any stationary $\{X_t\}$, including those generated by nonlinear AR models of the form (10.4). In general, $\{e_t\}$ is not i.i.d., and further

$$E(X_t | X_{t-k},\, k \geq 1) \neq \sum_{i=1}^{\infty} \varphi_i X_{t-i} \equiv \widehat{X}_t.$$

However, $\widehat{X}_t$ is in fact the *best linear predictor* of X_t from its lagged values in the sense that it minimizes

$$E\left\{ X_t - \sum_{i=1}^{\infty} b_i X_{t-i} \right\}^2$$

over all square-summable coefficients $\{b_i\}$ (see also §3.2). The mean square error of this linear predictor is

$$E(X_t - \widehat{X}_t)^2 = E(e_t^2) = \sigma^2.$$

This illustrates that it is perfectly sensible to seek the best linear predictor for general stationary nonlinear time series. Furthermore, if the innovation has a small variance in its Wold decomposition, the linear prediction is reliable. Nevertheless, we emphasize that the best linear predictor is not the least squares predictor in general and therefore is not the best estimator. In fact, the best linear prediction can be viewed as merely the best prediction based on the first two moment properties since it does not make use of any properties of the process beyond the second moment. The two predictors will be identical if the innovations $\{e_t\}$ are i.i.d. and e_t is independent of $\{X_{t-j},\, j \geq 1\}$. Under these conditions, $\{X_t\}$ is often called a linear process.

We end this section by summarizing some general principles for prediction.

(i) Under the least squares criterion, the least squares prediction is recommended, although often it is not a linear procedure and may be hampered by the difficulties associated with the estimation for predictive functions. Typically, the estimation is only possible in a nonparametric manner, although some semiparametric approximations could be employed (Chapter 8).

(ii) Linear prediction is easy to implement. It is reliable and robust, as the model (10.14) is always valid for a stationary process. Linear prediction may also provide a useful yardstick as a basis for comparison with nonlinear methods. But the best linear predictor is not the best predictor in general.

(iii) The prediction based on a nonlinear parametric model will be much more efficient for one-step-ahead prediction if the model is correctly specified. But such an approach is less robust against model misspecification. Also, the concrete parametric form is rarely useful for identifying multiple-step predictive functions.

10.2 Point Prediction

10.2.1 Local Linear Predictors

Let $X_1, \cdots, X_T$ be observations from a strictly stationary time series. By (10.1) and (10.2), the least squares predictor for X_{T+m} based on p lagged variables $\mathbf{X}_T = \mathbf{x}$ is

$$f_m(\mathbf{x}) = E(X_{T+m}|\mathbf{X}_T = \mathbf{x}),$$

where $\mathbf{X}_t = (X_t, X_{t-1}, \cdots, X_{t-p+1})^\tau$. The problem of estimating the predictive function f_m is a standard nonparametric regression, which may be tackled by using the techniques presented in §8.2. As an illustration, we simply adopt local linear regression to estimate f_m and its derivative $\dot{f}_m$, which is useful in calculating its mean square error; see Theorem 10.1. Let $\widehat{f}_m(\mathbf{x}) = \widehat{a}$ and $\widehat{\dot{f}}_m(\mathbf{x}) = \widehat{\mathbf{b}}$, and $(\widehat{a}, \widehat{\mathbf{b}})$ is the minimizer of the weighted sum

$$\sum_{t=p}^{T-m} \{X_{t+m} - a - \mathbf{b}^\tau(\mathbf{X}_t - \mathbf{x})\}^2 K\left(\frac{\mathbf{X}_t - \mathbf{x}}{h}\right),$$

where $K(\cdot)$ is a kernel function on R^p, and $h = h(T)$ is a bandwidth. Simple calculation yields

$$\widehat{f}_m(\mathbf{x}) = \frac{T_0(\mathbf{x}) - S_1^\tau(\mathbf{x})S_2^{-1}(\mathbf{x})T_1(\mathbf{x})}{S_0(\mathbf{x}) - S_1^\tau(\mathbf{x})S_2^{-1}(\mathbf{x})S_1(\mathbf{x})}, \tag{10.15}$$

$$\widehat{\dot{f}}_m(\mathbf{x}) = \{S_2(\mathbf{x}) - S_1(\mathbf{x})S_1^\tau(\mathbf{x})/S_0(\mathbf{x})\}^{-1}\{S_1(\mathbf{x})T_0(\mathbf{x})/S_0(\mathbf{x}) - T_1(\mathbf{x})\}, \tag{10.16}$$

where

$$S_0(\mathbf{x}) = \sum_{t=p}^{T-m} K\left(\frac{\mathbf{X}_t - \mathbf{x}}{h}\right), \quad S_1(\mathbf{x}) = \sum_{t=p}^{T-m} (\mathbf{x} - \mathbf{X}_t)K\left(\frac{\mathbf{X}_t - \mathbf{x}}{h}\right),$$

$$S_2(\mathbf{x}) = \sum_{t=p}^{T-m} (\mathbf{x} - \mathbf{X}_t)K\left(\frac{\mathbf{X}_t - \mathbf{x}}{h}\right)(\mathbf{x} - \mathbf{X}_t)^\tau,$$

and

$$T_0(\mathbf{x}) = \sum_{t=p}^{T-m} X_{t+m}K\left(\frac{\mathbf{X}_t - \mathbf{x}}{h}\right),$$

$$T_1(\mathbf{x}) = \sum_{t=p}^{T-m} (\mathbf{x} - \mathbf{X}_t)X_{t+m}K\left(\frac{\mathbf{X}_t - \mathbf{x}}{h}\right).$$

The decomposition theorem (i.e., Theorem 10.1) still holds asymptotically if we replace the theoretical predictor f_m by its estimator given in (10.15). In fact, Yao and Tong (1994a) showed that under some regularity conditions the estimator $\widehat{f}_m(\mathbf{x})$ is mean square consistent in the sense that

$$E[\{f_m(\mathbf{x}) - \widehat{f}_m(\mathbf{x})\}^2 | \mathbf{X}_T = \mathbf{x} + \boldsymbol{\delta}] \to 0$$

almost surely as $T \to \infty$. Now, it follows from Theorem 10.1 and the Cauchy–Schwarz inequality that

$$\begin{aligned} & E[\{X_{T+m} - \widehat{f}_m(\mathbf{x})\}^2 | \mathbf{X}_T = \mathbf{x} + \boldsymbol{\delta}] \\ = \; & E[\{X_{T+m} - f_m(\mathbf{x})\}^2 | \mathbf{X}_T = \mathbf{x} + \boldsymbol{\delta}] + R_T \\ = \; & \sigma_m^2(\mathbf{x} + \boldsymbol{\delta}) + \{\boldsymbol{\delta}^\tau \dot{f}_m(\mathbf{x})\}^2 + R_T + o(||\boldsymbol{\delta}||^2), \end{aligned} \tag{10.17}$$

where $R_T \to 0$ almost surely as $T \to \infty$. Note that this result holds for general strictly stationary processes, and we do not impose any model assumptions on $\{X_t\}$; see §3 of Yao and Tong (1994a).

From (10.17), we need to estimate $\sigma_m^2(\cdot)$ and $\dot{f}_m(\cdot)$ in order to gauge how good the predictor $\widehat{f}_m(\cdot)$ is. The estimation for conditional variance $\sigma_m^2(\cdot)$ has been discussed in detail in §8.7. The local linear regression gives a natural estimator (10.16) for the derivative $\dot{f}_m(\cdot)$. Of course, if our primary goal is to estimate the first derivative, we may adopt local quadratic regression instead; see §6.3.

10.2.2 An Example

Example 10.2 below and Figures 10.2 and 10.3 illustrate the predictor constructed above. More examples can be found in Yao and Tong (1994a).

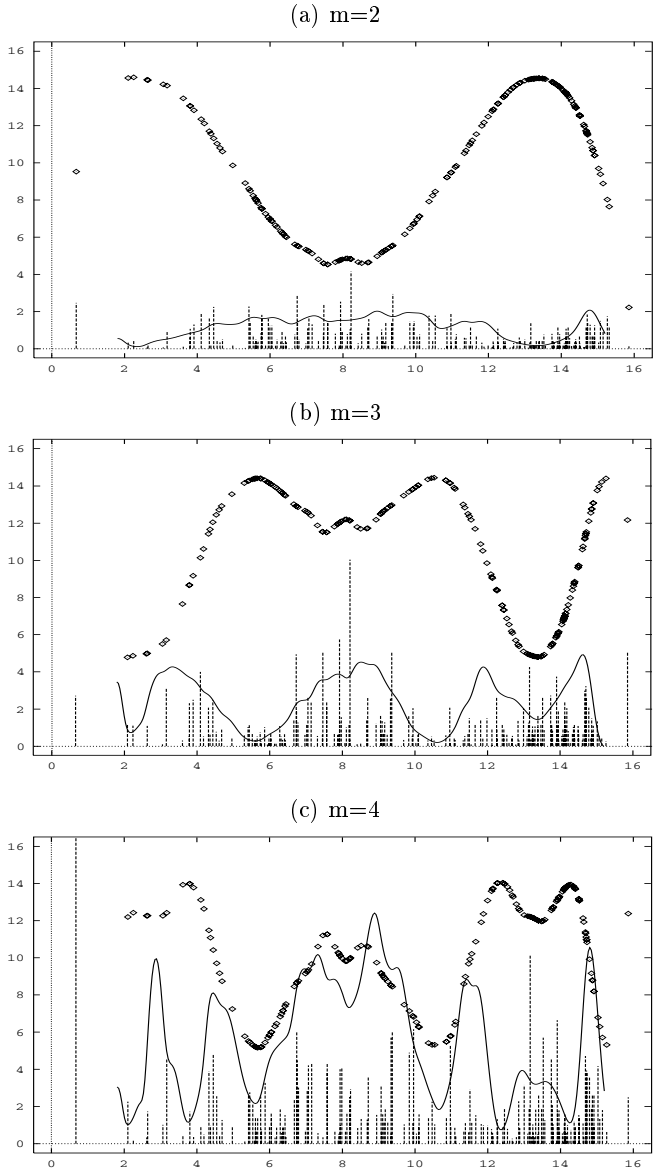


FIGURE 10.2. Example 10.2. Plots of the 200 m-step-ahead predicted values (diamonds) and the corresponding absolute errors (impulses) against their initial values as well as the estimated conditional variance $\widehat{\sigma}^2_m(x)$ (solid curves) for (a) $m = 2$ ($h = 0.25$), (b) $m = 3$ ($h = 0.2$), and (c) $m = 4$ ($h = 0.18$). From Yao and Tong (1995b).

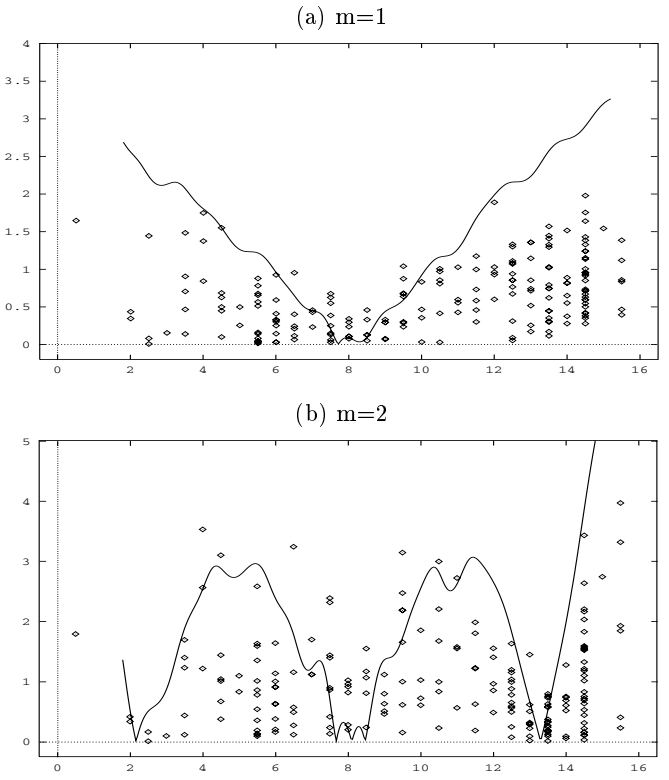


FIGURE 10.3. Example 10.2. Plots of the 200 m-step-ahead predicted values (diamonds) and the corresponding absolute errors (impulses) against their rounded initial values as well as the estimated function $|\widehat{f}_m(x)|$ (solid curves) for (a) $m = 1$ ($h = 0.32$) and (b) $m = 2$. From Yao and Tong (1995b).

Example 10.2 Consider the simple one-dimensional model

$$X_t = 0.23X_{t-1}(16 - X_{t-1}) + 0.4\varepsilon_t,$$

where $\{\varepsilon_t\}$ is a sequence of independent $N(0, 1)$ random variables truncated in the interval $[-12, 12]$. We generate a sample of size 1,200 from this model. Since $\sigma_1^2(x) \equiv 0.16$, the one-step-ahead prediction is uniformly good for different initial values, which we do not report here. We use the first 1,000 data to estimate predictive functions for $m = 2, 3$, and 4 and the last 200 points to check the performance. The predicted values for those 200 post-sample points together with their absolute predictive errors and estimated conditional variance $\widehat{\sigma}_m^2(x)$ are shown in Figure 10.2. Since the rounding error in the calculation is below 10^{-6}, the accuracy is dominated by the conditional variance. For example, Figure 10.2(b) shows that the

three-step-ahead prediction is at its worst when the initial value is around 8 and is at its best when the initial value is near 5.6 or 10.4, which is in agreement with the profile of $\widehat{\sigma}_3(x)$.

To see how a small shift in the initial values affects the prediction, we round the initial value x to the nearest value from among $[x]$, $[x]+0.5$, and $[x]+1$, where $[x]$ denotes the integer part of x. Hence $|\delta| \leq 0.5$. Figure 10.3 shows that, for $m = 1, 2$, the absolute prediction error increases as $|\widehat{\hat{f}}_m(x)|$ increases, which is consistent with the decomposition formula (10.17). ∎

10.3 Estimating Predictive Distributions

In any serious attempt to forecast, a point prediction is only a beginning. A predictive interval or, more generally, a predictive set is more informative. All information on the future is of course contained in a predictive distribution function, which is in fact a conditional distribution of a future variable given the present state. We deal with the estimation for predictive distribution functions in this section. Predictive intervals will be discussed in the next section.

In general, the *predictive distribution* of X_{T+m} based on $\mathbf{X}_T = (X_T, \cdots, X_{T-p+1})$ is the conditional distribution of X_{T+m} given $\mathbf{X}_T$. In the context of linear time series models with normally distributed innovations, the predictive distributions are normal. Therefore, the problem of estimating predictive distributions reduces to the estimation of means and variances. However, for nonlinear time series, the predictive distributions typically are not normal. Furthermore even for a process generated by a parametric nonlinear model, multiple-step-ahead predictive distributions are of unknown form and may only be estimated in a nonparametric manner. Below, we introduce two estimation methods proposed in Hall, Wolff, and Yao (1999). The first, local logistic distribution estimation, produces estimators of arbitrarily high order that always lie strictly between 0 and 1. In spirit, this approach is related to recently-introduced local parametric methods. The second method is an "adjusted" version of the Nadaraya–Watson estimator. It is designed to reproduce the superior bias properties of local linear methods while preserving the property that the Nadaraya–Watson estimator is always a distribution function. It is based on weighted, or biased, bootstrap methods (Hall and Presnell 1999).

In the rest of this section, we assume that data are available in the form of a strictly stationary stochastic process $\{(\mathbf{X}_t, Y_t)\}$, where Y_t is a scalar and $\mathbf{X}_t$ is a p-dimensional vector. In the time series context, $\mathbf{X}_t = (X_t, \cdots, X_{t-p+1})^\tau$ typically denotes a vector of lagged values of $Y_t = X_{t+m}$ for some $m \geq 1$. Naturally, our setting also includes the case where the pairs $(\mathbf{X}_t, Y_t)$ are independent and identically distributed. We wish to estimate

the conditional distribution function

$$F(y|\mathbf{x}) \equiv P(Y_t \le y|\mathbf{X}_t = \mathbf{x}).$$

If we write $Z_t = I(Y_t \le y)$ then,

$$E(Z_t|\mathbf{X}_t = \mathbf{x}) = F(y|\mathbf{x}),$$

so our estimation problem may be viewed as regression of Z_t on $\mathbf{X}_t$; see also Example 6.2. Hence $F(y|\mathbf{x})$ can be estimated by the local linear technique in §8.2. Although such an approach is useful in practice, the estimator $\widehat{F}(y|\mathbf{x})$ is not necessarily a cumulative distribution function. This is a drawback.

To simplify discussion, we introduce our methods and develop theory only in the case where $\mathbf{X}_t = X_t$ is a scalar (i.e., $p = 1$). The multivariate case will be illustrated through a real data example.

10.3.1 Local Logistic Estimator

For fixed y, write $P(x) = F(y|x)$ and assume that P has $r-1$ continuous derivatives. A generalized local logistic model for $P(x)$ has the form $L(x, \boldsymbol{\theta}) \equiv A(x, \boldsymbol{\theta})/\{1 + A(x, \boldsymbol{\theta})\}$, where $A(\cdot, \boldsymbol{\theta})$ denotes a nonnegative function that depends on a vector of parameters $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_r)$ that "represent" the values of $P(x), P^{(1)}(x), \cdots, P^{(r-1)}(x)$. Here, "represent" means that, for each sequence $\omega_1 \in (0, 1), \omega_2, \cdots, \omega_r$ denoting potential values of $P(x), P^{(1)}(x), \cdots, P^{(r-1)}(x)$, respectively, there exist $\theta_1, \cdots, \theta_r$ such that

$$\frac{A(u, \boldsymbol{\theta})}{1 + A(u, \boldsymbol{\theta})} = \omega_1 + \omega_2(u - x) + \cdots + (r!)^{-1}\omega_r(u - x)^{r-1} + o\big(|u - x|^{r-1}\big)$$

as $u \to x$. Arguably, the simplest function A with which to work is $A(u, \boldsymbol{\theta}) \equiv e^{p(u,\boldsymbol{\theta})}$, where $p(u, \boldsymbol{\theta}) = \theta_1 + \theta_2 u + \cdots + \theta_r u^{r-1}$ is a polynomial of degree $r-1$. Fitting this model locally to indicator-function data leads to an estimator $\widehat{F}(y|x) \equiv L(0, \widehat{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\theta}}$ denotes the minimizer of

$$R(\boldsymbol{\theta}; x, y) = \sum_{t=1}^{T} \{I(Y_t \le y) - L(X_t - x, \boldsymbol{\theta})\}^2 K_h(X_t - x)\,, \qquad (10.18)$$

K is a kernel function, $K_h(\cdot) = h^{-1}K(\cdot/h)$, and $h > 0$ is a bandwidth. We call this approach local logistic distribution estimation. Depending on bandwidth choice, it also furnishes consistent estimators of the derivatives $F^{(i)}(y|x) \equiv (\partial/\partial x)^i F(y|x)$ in the form $\widehat{F}^{(i)}(y|x) = L^{(i)}(0, \widehat{\boldsymbol{\theta}}_{xy})$ for $i = 1, \cdots, r-1$, where $L^{(i)}(x, \boldsymbol{\theta}) \equiv (\partial/\partial x)^i L(x, \boldsymbol{\theta})$. In practice, $\widehat{\boldsymbol{\theta}}_{xy}$ may be computed using the "downhill simplex" algorithm (see §10.4 in Press et al. 1992).

We expect the estimator $\widehat{F}(y|x)$ to have a bias of order h^r and variance of order $(Th)^{-1}$ under an asymptotic scheme where $h = h(T) \to 0$ and $Th \to \infty$ as $T \to \infty$. A more detailed account of this property will be given in §10.3.5.

10.3.2 Adjusted Nadaraya–Watson Estimator

Let $p_t = p_t(x)$, for $1 \le t \le T$, denote weights (functions of the data $X_1, \cdots, X_T$, as well as of x) with the property that each $p_t \ge 0$, $\sum_t p_t = 1$, and

$$\sum_{t=1}^{T} p_t(x)\,(X_t - x)\,K_h(X_t - x) = 0\,. \tag{10.19}$$

Of course, p_t's satisfying these conditions are not uniquely defined, and we specify them concisely by asking that $\prod_t p_t$ be as large as possible subject to the constraints. Define

$$\widetilde{F}(y|x) = \frac{\sum_{t=1}^{T} I(Y_t \le y)\,p_t(x)\,K_h(X_t - x)}{\sum_{t=1}^{T} p_t(x)\,K_h(X_t - x)}\,. \tag{10.20}$$

Note particularly that $0 \le \widetilde{F}(y|x) \le 1$ and $\widetilde{F}$ is monotone in y. We will show in §10.3.5 that $\widetilde{F}$ is first-order equivalent to a local linear estimator, which does not enjoy either of these properties of $\widetilde{F}$.

Another way of viewing the biased bootstrap estimator $\widetilde{F}$ is as a local linear estimator in which the weights for the least-squares step are taken to be $p_t(x)\,K_h(X_t - x)$, rather than simply $K_h(X_t - x)$, for $1 \le t \le T$. To appreciate why this is so, we refer to the definition of general local linear estimators given by Fan and Gijbels (1996, p. 20) and note that in view of (10.19), with the suggested change of weights, their estimator $\widehat{m}_0$ reduces to (see (6.10))

$$\widehat{m}_0(x) = \left\{ \sum_{t=1}^{T} w_t(x)\,I(Y_t \le y) \right\} \Big/ \left\{ \sum_{t=1}^{T} w_t(x) \right\},$$

where

$$w_t(x) = p_t(x)\,K_h(X_t - x).$$

Therefore, $\widehat{m}_0 = \widetilde{F}(y|x)$.

Computation of the weights p_t can be carried out via the Lagrange multiplier method. The constrained optimization reduces to maximizing

$$\sum_{t=1}^{T} \log p_t + \lambda_1 \sum_{t=1}^{T} p_t + \lambda_2 \sum_{t=1}^{T} p_t(X_t - x)K_h(X_t - x)$$

with the Lagrange multipliers λ_1 and λ_2. By taking the derivative with respect to p_t and setting it to zero, we have

$$p_t^{-1} + \lambda_1 + \lambda_2(X_t - x)K_h(X_t - x) = 0.$$

This gives the solution

$$p_t = -\lambda_1^{-1}\{1 + \lambda(X_t - x)K_h(X_t - x)\}^{-1},$$

where $\lambda = \lambda_2/\lambda_1$. The constraint on the unit total weight leads to

$$p_t = \frac{\{1 + \lambda(X_t - x)K_h(X_t - x)\}^{-1}}{\sum_{j=1}^{T}\{1 + \lambda(X_j - x)K_h(X_j - x)\}^{-1}}.$$

The parameter λ (a function of the data and of x) is uniquely defined by (10.19). It is easily computed using a Newton–Raphson argument. Now, it follows from (10.19) that

$$0 = \sum_{j=1}^{T} \frac{(X_j - x)K_h(X_j - x)}{1 + \lambda(X_j - x)K_h(X_j - x)} = T - \sum_{j=1}^{T} \frac{1}{1 + \lambda(X_j - x)K_h(X_j - x)}.$$

Hence

$$p_t(x) = p_t = T^{-1}\{1 + \lambda(X_t - x)K_h(X_t - x)\}^{-1}.$$

10.3.3 Bootstrap Bandwidth Selection

Particularly in the time series case, deriving asymptotically optimal bandwidths for either the local logistic or biased bootstrap methods is a tedious matter. Using plug-in methods requires explicit estimation of complex functions with dependent data, and using cross-validation demands selection of the amount of data that is left out. Instead, Hall, Wolff, and Yao (1999) suggested a bootstrap approach, which we introduce below. We first fit a simple parametric model such as

$$Y_i = a_0 + a_1 X_i + \cdots + a_k X_i^k + \sigma\varepsilon_i,$$

where ε_i is standard normal, $a_0, \cdots, a_k, \sigma$ are estimated from the data, and k is determined by AIC. We form a parametric estimator $\check{F}(y|x)$ based on the model. By Monte Carlo simulation from the model, we compute a bootstrap version of $\{Y_1^*, \cdots, Y_T^*\}$ based on given observations $\{X_1, \cdots, X_T\}$, and hence a bootstrap version $\widehat{F}_h^*(y|x) = \widehat{F}^*(y|x)$ of $\widehat{F}(y|x)$, derived from (10.18) with $\{(X_i, Y_i)\}$ replaced by $\{(X_i, Y_i^*)\}$. The bootstrap estimator of the absolute deviation error of $\widehat{F}(y|x)$ is

$$M(h; x, y) = E\left[|\widehat{F}_h^*(y|x) - \check{F}(y|x)| \,\middle|\, \{(X_i, Y_i)\} \right].$$

Choose $h = \widehat{h}(x, y)$ to minimize $M(h; x, y)$. Sometimes we use the x-dependent bandwidth $\widehat{h}(x)$, which minimizes

$$M(h; x) = \int M(h; x, y)\, \check{F}(y|x)\, dy.$$

In practice, $M(h; x, y)$ and $M(h; x)$ are evaluated via repeated bootstrap sampling. The approach above can also be applied to choosing h for estimating $\widetilde{F}(y|x)$.

When we are working with time-series data (e.g., $Y_t = X_{t+m}$ for some $m \geq 1$), we propose an alternative resampling scheme as follows. Assume that the data $\{X_1, \cdots, X_T\}$ represent a segment of a Gaussian autoregression:

$$X_t = b_1 X_{t-1} + \cdots + b_p X_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2).$$

Select the order p via AIC, and estimate its parameters $b_1, \cdots, b_p$ and σ^2. Resample the segment $\{X_1^*, \cdots, X_T^*\}$ from the parametric model. The bootstrap estimator $F_h^*(y|x)$ is calculated using this segment and then substituted into the formula above for $M(h; x, y)$.

10.3.4 Numerical Examples

We illustrate the methodology through one simulated model and the Canadian lynx data for which we also consider the conditional distributions given two lagged values. We always use the Gaussian kernel in our calculation.

Example 10.3 We compared various estimators of the conditional distribution function $F(.|.)$ through the simulated model

$$Y_t = 3.76\, Y_{t-1} - 0.235\, Y_{t-1}^2 + 0.3\, \varepsilon_t, \tag{10.21}$$

where the errors ε_t were independent with common distribution $U[-\sqrt{3}, \sqrt{3}]$. The estimators concerned are the Nadaraya–Watson estimator (NW), the local linear regression estimator (LL), the adjusted Nadaraya–Watson estimator (ANW), and the local logistic estimators with $r = 2$ (LG-2). We treated two- and three-step-ahead prediction by taking $Y_t = X_{t+m}$ for $m = 2$ and 3. The performance of the estimator was evaluated in terms of the Mean Absolute Deviation Error (MADE),

$$\text{MADE} = \frac{\sum_i |F_e(y_i|x_i) - F(y_i|x_i)|\ I\{0.001 \leq F(y_i|x_i) \leq 0.999\}}{\sum_i I\{0.001 \leq F(y_i|x_i) \leq 0.999\}},$$

where $F_e(.|.)$ denotes an estimator of $F(.|.)$, and $\{(x_i, y_i)\}$ are grid points with step 0.4 in the x-direction and steps 0.1 for $m = 2$ and 0.19 for $m = 3$ in the y-direction.

We conducted the simulation in two stages. First, we calculated MADEs for the various estimators over grid points evenly distributed across the whole sample space. For each estimator, we used the optimal bandwidth defined by

$$h_{\text{opt}}(x) = \int h_{\text{opt}}(x, y)\, F(y|x)\, dy\,,$$

where $h_{\text{opt}}(x, y)$ is the minimizer of the asymptotic mean squared error (up to first order) of the estimator. This guarantees a fair comparison among different methods. Note that the conditional distributions concerned no

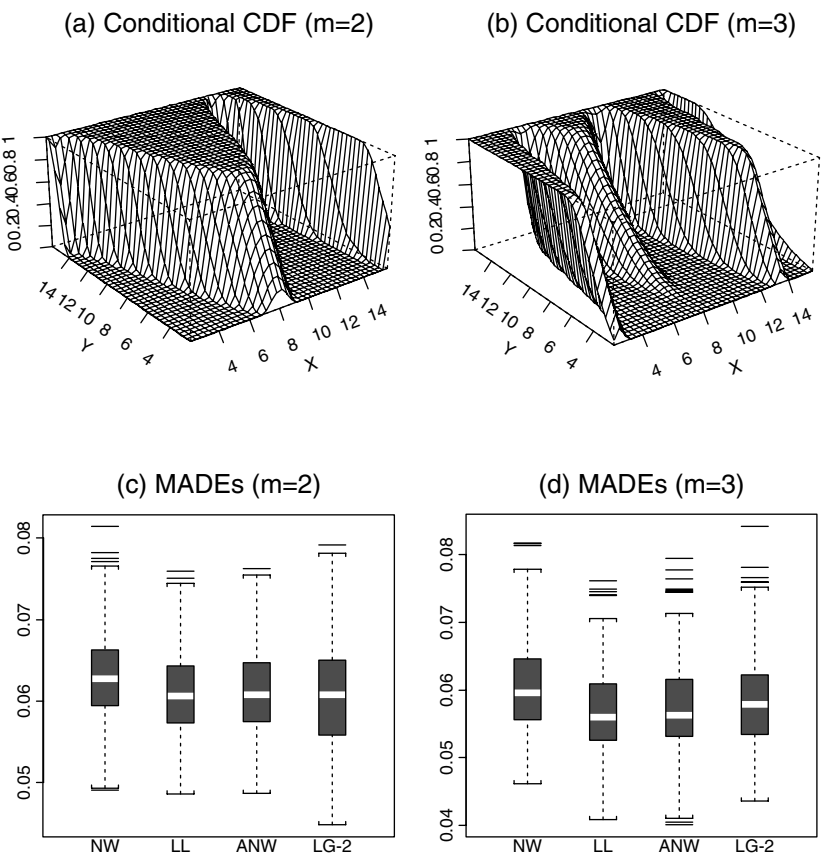


FIGURE 10.4. Example 10.3. Conditional distribution function $z = F(y|x)$ for (a) $m = 2$ and (b) $m = 3$. Boxplots of MADEs for the Nadaraya–Watson estimate (NW), local linear regression estimate (LL), adjusted NW estimate (ANW), and local logistic estimate with $r = 2$ (LG-2) when (c) $m = 2$ and (d) $m = 3$. From Hall, Wolff, and Yao (1999).

longer admit simple explicit forms. In order to calculate $h_{\text{opt}}(x)$, we evaluated the true values of $F(y|x)$ and its derivatives by simulation, as follows. We generated 50,000 random samples by iterating (10.21) two (or three) times starting at a fixed value x. The relative frequency of the sample exceeding y was regarded as the true value of $F(y|x)$. The resulting conditional distribution functions are plotted in Figures 10.4 (a) and (b). We used kernel methods to estimate the marginal density function with a sample of size 100,000. Figures 10.4 (c) and (d) are the boxplots of MADEs for the 400 replications. Both the ANW and LG-2 methods provide competi-

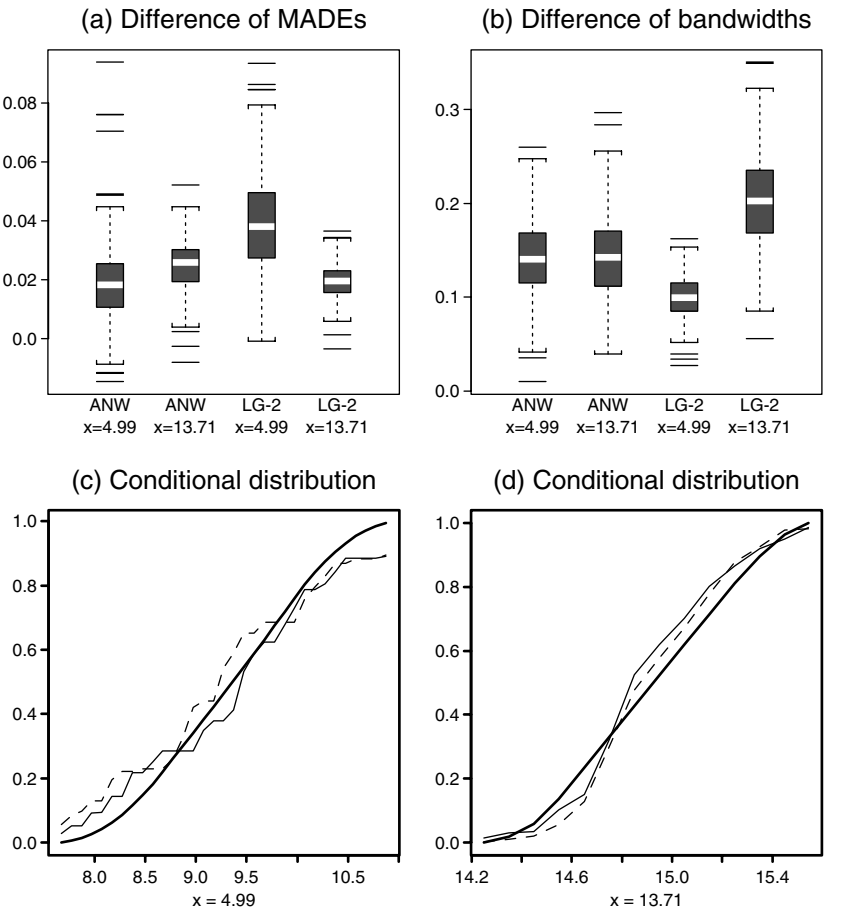


FIGURE 10.5. Example 10.3. (a) Boxplots of the MADEs based on $\widehat{h}(x)$ minus the MADEs based on $h_{\text{opt}}(x)$. (b) Boxplots of $\widehat{h}(x) - h_{\text{opt}}(x)$. (c)—(d) The curves representing the conditional distribution functions $F(.|x)$: thick line, $F(.|x)$; thin line, adjusted Nadaraya–Watson estimate; dashed line, local logistic estimate ($r = 2$). From Hall, Wolff, and Yao (1999).

tive performance relative to the LL method in terms of the absolute error of estimation. The larger MADE values for the NW estimator are due to its larger bias and poor boundary effect.

Second, we demonstrated the usefulness of the bootstrap scheme for choosing bandwidth stated in §10.3.3. For each of 200 random samples of size $n = 600$, we estimated the two-step-ahead predictive distribution $F(.|x)$ using the bandwidth $\widehat{h}(x)$ selected by the bootstrap scheme for $x = 4.99$ and 13.71. The bootstrap resampling was conducted as follows. We fitted a linear AR(1) model to the original data and sampled time series

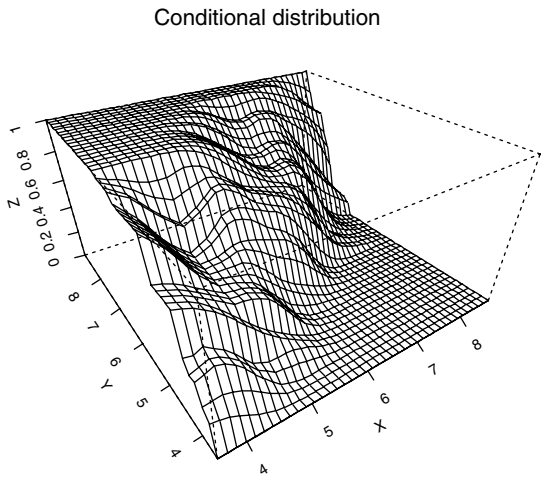


FIGURE 10.6. Example 10.4. Estimated conditional distribution $z = F(y|x)$ of $Y_t = X_{t+1}$ given $X_t = x$.

(with length 600) from the fitted model. We replicated bootstrap sampling 40 times. We considered only the adjusted Nadaraya–Watson estimator and the local logistic estimator with $r = 2$. We compared the estimates with those based on the optimal bandwidth $h_{\text{opt}}(x)$, which is equal to 0.182 for $x = 4.99$ and 0.216 for $x = 13.71$ in the case of the ANW estimate and equal to 0.241 for $x = 4.99$ and 0.168 for $x = 13.71$ in the case of the LG-2 estimate. Figure 10.5(a) presents boxplots of the differences of MADEs based on $\widehat{h}(x)$ over the MADEs based on $h_{\text{opt}}(x)$. Figure 10.5(b) displays boxplots of $\widehat{h}(x) - h_{\text{opt}}(x)$ in the simulation with 200 replications. Since we used a simple linear model to fit the nonlinear structure, it is not surprising that $\widehat{h}(x)$ always overestimates $h_{\text{opt}}(x)$. But the estimates for $F(y|x)$ remain reasonably reliable. Figures 10.5(c) and (d) depict typical examples of the estimated conditional distribution functions $\widehat{F}(.|x)$ and $\widetilde{F}(.|x)$. The typical example was selected in such a way that the corresponding MADE was equal to its median in the simulation with 200 replications. ∎

Example 10.4 We illustrate our method with the Canadian lynx data (on a natural logarithmic scale) for the years 1821–1934; see Figure 1.2. We estimated the conditional distribution of X_{t+1} given X_t by the adjusted Nadaraya–Watson method. The bandwidths were selected by the bootstrap scheme based on resampling the whole time series from the best-fitted linear AR(1) model. We did 40 replications in the bootstrap resampling step. The estimated conditional distribution function is depicted in Figure 10.6.

TABLE 10.1. Predictive intervals for Canadian lynx in 1925–1934 based on the data in 1821–1924. The nominal coverage probability is $\alpha = 0.9$. From Hall, Wolff, and Yao (1999).

Year	True value	Predictor from one lagged value	$\widehat{h}(x)$	Predictor from two lagged values	$\widehat{h}(x_1, x_2)$
1925	8.18	[5.89, 8.69]	0.123	[6.86, 8.60]	0.245
1926	7.98	[5.99, 8.81]	0.340	[6.86, 8.81]	0.570
1927	7.34	[5.94, 8.75]	0.485	[6.40, 8.26]	0.715
1928	6.27	[5.43, 8.35]	0.195	[5.44, 6.86]	0.715
1929	6.18	[4.69, 7.71]	0.268	[4.60, 6.16]	1.095
1930	6.50	[4.65, 7.70]	0.340	[5.43, 7.03]	0.860
1931	6.91	[5.21, 7.72]	0.268	[5.71, 7.50]	0.860
1932	7.37	[5.37, 7.82]	0.268	[6.38, 8.12]	0.860
1933	7.88	[5.44, 8.38]	0.123	[7.17, 8.25]	0.715
1934	8.13	[5.89, 8.74]	0.485	[7.26, 8.81]	1.205

As an alternative application, we constructed the predictive interval

$$[F^{-1}(0.5 - 0.5\alpha|x),\ F^{-1}(0.5 + 0.5\alpha|x)], \qquad \alpha \in (0, 1), \tag{10.22}$$

based on the estimated conditional distribution function. To check on performance, we used the data for 1821–1924 (i.e., $T = 104$) to estimate $F(y|x)$ and the last ten data points to check the predicted values. This time, we used the local logistic method with $r = 2$. The results with $\alpha = 0.9$ are reported in Table 10.1. All of the predictive intervals contain the corresponding true values. The average length of the intervals is 2.80, which is 53.9% of the dynamic range of the data.

We also include in Table 10.1 the predictive intervals based on the estimated conditional distribution of X_t given both X_{t-1} and X_{t-2}. To obtain these results, we used the local (linear) logistic method to estimate $F(y|x_1, x_2)$. To this end, let $L(x_1, x_2, \boldsymbol{\theta}) = A(x_1, x_2, \boldsymbol{\theta})/\{1 + A(x_1, x_2, \boldsymbol{\theta})\}$ with $A(x_1, x_2, \boldsymbol{\theta}) = \exp(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$. The estimator is defined as $\widehat{F}(y|x_1, x_2) \equiv L(0, 0, \widehat{\theta})$, where $\widehat{\theta}$ denotes the minimizer of

$$\sum_{t=3}^{T} \{I(X_t \le y) - L(X_{t-1} - x_1, X_{t-2} - x_2, \boldsymbol{\theta})\}^2 K\left(\frac{X_{t-1} - x_1}{h_1}, \frac{X_{t-2} - x_2}{h_2}\right),$$

K is a symmetric probability density on R^2, and h_1 and h_2 are bandwidths. In our calculation, we simply chose K to be the standard Gaussian kernel and $h_1 = h_2$. The bandwidths were selected by the bootstrap scheme based on resampling time series from the best-fitted linear AR(2) model. Out of ten predictive intervals, only one (for the year 1929) missed the true value, and then only narrowly. The average length of the intervals is now reduced to 1.63, which is 32.8% of the dynamic range of the data. ∎

10.3.5 Asymptotic Properties

We present the asymptotic normality for both the local logistic estimator and adjusted Nadaraya–Watson estimator. For the local logistic estimator $\widehat{F}(y|x)$, we only consider function A of exponential-polynomial type, with $r \geq 2$: $A(x, \boldsymbol{\theta}) = \exp(\theta_1 x^0 + \ldots + \theta_r x^{r-1})$. Let f denote the marginal density of X_i. We impose the following regularity conditions:

(C1) For fixed y and x, $f(x) > 0$, $0 < F(y|x) < 1$, f is continuous at x, and $F(y|\cdot)$ has $2[(r+1)/2]$ continuous derivatives in a neighborhood of x, where $[t]$ denotes the integer part of t. The conditional density function of (X_1, X_j) given (Y_1, Y_j) is bounded by a positive constant independent of j.

(C2) The kernel K is a symmetric, compactly supported probability density satisfying $|K(x_1) - K(x_2)| \leq C\,|x_1 - x_2|$ for x_1, x_2.

(C3) The process $\{(X_i, Y_i)\}$ is α-mixing (see Definition 2.11). Furthermore, its α-mixing coefficients fulfill the condition

$$\sum_{j=1}^{\infty} j^{\lambda} \alpha(j)^{\gamma} < \infty \quad \text{for some } \gamma \in [0, 1) \text{ and } \lambda > \gamma.$$

(We define $a^b = 0$ when $a = b = 0$.)

(C4) As $T \to \infty$, $h \to 0$ and $\liminf_{n\to\infty} Th^{2r} > 0$.

Assumption (C3) with $\gamma = 0$ implies that the process $\{(X_i, Y_i)\}$ is m-dependent for some $m \geq 1$. The requirement in (C2) that K be compactly supported is imposed for the sake of brevity of proofs and can be removed at the cost of lengthier arguments. In particular, the Gaussian kernel is allowed. The last condition in (C4) may be relaxed if we are prepared to strengthen (C3) somewhat. For example, if the process $\{(X_i, Y_i)\}$ is m-dependent, then, for Theorem 10.2 below, we need only $nh \to \infty$, not nh^{2r} bounded away from 0. However, since (C4) is always satisfied by bandwidths of optimal size (i.e. $h \approx CT^{-1/(2r+1)}$), we will not concern ourselves with such refinements.

Define $\kappa_j = \int u^j\, K(u)\, du$ and $\nu_j = \int u^j\, K(u)^2\, du$. Let $\mathbf{S}$ denote the $r \times r$ matrix with (i, j)th element κ_{i+j-2} and $\kappa^{(i,j)}$ be the (i, j)th element of $\mathbf{S}^{-1}$. Let $r_1 = 2[(r+1)/2]$ and put $\tau(y|x)^2 = F(y|x)\,\{1 - F(y|x)\}/f(x)$,

$$\begin{aligned} \mu_r(x) &= (r!)^{-1} \left\{ F^{(r_1)}(y|x) - L^{(r_1)}(0, \boldsymbol{\theta}^0) \right\} \sum_{i=1}^{r} \kappa^{(1,i)}\, \kappa_{r_1+i-1}\,, \\ \tau_r^2 &= \int \left(\sum_{i=1}^{r} \kappa^{(1,i)}\, u^{i-1} \right)^2 K(u)^2\, du, \end{aligned}$$

where $\boldsymbol{\theta}^0$ is determined by

$$F^{(i)}(y|x) = L^{(i)}(0, \boldsymbol{\theta}^0), \qquad i = 0, 1, \cdots, r-1. \tag{10.23}$$

Let N_1, N_2, N_3 denote random variables with the standard normal distribution.

Theorem 10.2 (*i*) *Suppose that* $r \geq 2$ *and conditions* $(C1)$–$(C4)$ *hold. Then, as* $T \to \infty$,

$$\begin{aligned} & \widehat{F}(y|x) - F(y|x) \qquad (10.24) \\ = \; & (Th)^{-1/2}\, \tau(y|x)\, \tau_r\, N_1 + h^{r_1} \mu_r(x) + o_p\left\{h^{r_1} + (Th)^{-1/2}\right\}. \end{aligned}$$

(*ii*) *Assume conditions* $(C1)$–$(C4)$ *with* $r = 2$. *Then, as* $T \to \infty$,

$$\begin{aligned} & \widetilde{F}(y|x) - F(y|x) \qquad (10.25) \\ = \; & (Th)^{-1/2}\, \tau(y|x)\, \nu_0^{1/2}\, N_2 + \tfrac{1}{2}\, h^2\, \kappa_2\, F^{(2)}(y|x) + o_p\left\{h^2 + (Th)^{-1/2}\right\}. \end{aligned}$$

The theorem above was first established for β-mixing processes by Hall, Wolff, and Yao (1999). A proof for α-mixing processes based on Theorem 2.22 will be outlined in §10.5. Some remarks are now in order.

(a) Comparison of $\widehat{F}$ *and the local polynomial estimator.* To first order, and for general x, the asymptotic variance of $\widehat{F}(y|x)$ is exactly the same as in the case of local polynomial regression estimators of order r; for the latter, see, for example, Ruppert and Wand (1994) and Fan and Gijbels (1996). This similarity extends also to the bias term, to the extent that for both $\widehat{F}$ and local polynomial estimators the bias is of order h^r for even r and h^{r+1} for odd r, and (to this order) does not depend on the design density. However, the forms of bias as functionals of the "regression mean" F are quite different. This is a consequence of the fact that, unlike a local polynomial estimator, $\widehat{F}(y|x)$ is constrained to lie within $(0, 1)$.

(b) Comparison of $\widetilde{F}$ *and the local linear estimator.* It can be shown that under conditions (C1)–(C4) for $r = 2$, the asymptotic formula (10.25) for $\widetilde{F}(y|x)$ is shared exactly by the standard local linear estimator $\widehat{F}_{\mathrm{LL}}(y|x)$, derived by minimizing

$$\sum_{t=1}^{T} \{I(Y_t \leq y) - \alpha - \beta\,(X_t - x)\}^2\, K_h(X_t - x)$$

with respect to (α, β) and taking $\widehat{F}_{\mathrm{LL}}(y|x) = \widehat{\alpha}$. Compare (10.25) with the state-domain version of (6.10). Note, however, that, unlike $\widetilde{F}$, $\widehat{F}_{\mathrm{LL}}$ is constrained neither to lie between 0 and 1 nor to be monotone in y. Additionally, $\widetilde{F}$ is somewhat more resistant against data sparseness than $\widehat{F}_{\mathrm{LL}}$.

(c) Comparison of $\widehat{F}$ and $\widetilde{F}$. In the case $r = 2$, (10.24) reduces to

$$\begin{aligned} & \widehat{F}(y|x) - F(y|x) \qquad (10.26) \\ = \; & (Th)^{-1/2}\,\tau(y|x)\,N_1 + \tfrac{1}{2}\,h^2\,\kappa_2\,\mu_2(y|x) + o_p\left\{h^2 + (Th)^{-1/2}\right\}, \end{aligned}$$

where

$$\mu_2(y|x) = F^{(2)}(y|x) - \frac{F^{(1)}(y|x)^2\,\{1 - 2F(y|x)\}}{F(y|x)\{1 - F(y|x)\}}$$

and $F^{(i)} = (\partial/\partial x)^i F$. Comparing (10.25) and (10.26), we see that $\widehat{F}(y|x)$ (with $r = 2$) and $\widetilde{F}(y|x)$ have the same asymptotic variance but that the first-order bias formula of the former contains an additional term. In consequence, if $F(y|x) < \frac{1}{2}$, then $\widehat{F}(y|x)$ is biased downward relative to $\widetilde{F}(y|x)$, while if $F(y|x) > \frac{1}{2}$, then it is biased upward.

(d) Comparison with the Nadaraya–Watson estimator. The analog of (10.25) and (10.26) in the case of the Nadaraya–Watson estimator,

$$\widehat{F}_{\mathrm{NW}}(y|x) = \left\{\sum_{t=1}^{T} I(Y_t \le y)\,K_h(X_t - x)\right\} \Big/ \left\{\sum_{t=1}^{T} K_h(X_t - x)\right\},$$

is

$$\begin{aligned} & \widehat{F}_{\mathrm{NW}}(y|x) - F(y|x) \\ = \; & (Th)^{-1/2}\,\tau(y|x)\,\nu_0^{1/2}\,N_3 + \tfrac{1}{2}\,h^2\,\kappa_2\,\mu(y|x) + o_p\left\{h^2 + (Th)^{-1/2}\right\}, \end{aligned}$$

where $\mu(y|x) = F^{(2)}(y|x) + 2\,f(x)^{-1} f'(x)\,F^{(1)}(y|x)$. Note particularly that, unlike any of $\widehat{F}$, $\widetilde{F}$, and $\widehat{F}_{\mathrm{LL}}$, $\widehat{F}_{\mathrm{NW}}$ has a bias that depends to first order on the density f of X_t. However, the variances of all four estimators ($\widehat{F}$ with $r = 2$) are identical to first order.

(e) Continuity of $F(y|x)$ with respect to y. Conditions (C1)–(C4) require continuity of $F(y|x)$ with respect only to x, not to y. In principle, we could exploit smoothness of $F(y|x)$ in y by taking, for example, the integral average of $\widehat{F}(\cdot|x)$ or $\widetilde{F}(\cdot|x)$ in the neighborhood of y, thereby obtaining an estimator that had potentially lower variance. However, any improvement in performance is available only to second order. To appreciate why, note that if $y_1 \le y_2$, then, to first order, the covariance of $\widehat{F}(y_1|x)$ and $\widehat{F}(y_2|x)$ equals $(nh)^{-1}F(y_1|x)\,\{1 - F(y_2|x)\}\,\tau_r^2$, which, as $y_1, y_2 \to y$, converges to the first-order term in the variance of $\widehat{F}(y|x)$. It follows that no first-order reductions in variance are obtainable by averaging over values of $\widehat{F}(z|x)$ for z in a decreasingly small neighborhood of y. The same argument applies to $\widetilde{F}$.

10.3.6 Sensitivity to Initial Values: A Conditional Distribution Approach

(a) Sensitivity measures K_m and D_m

In §10.1.3, we dealt with the sensitivity of the conditional means to initial values, which is relevant to nonlinear point prediction. A more informative way is to consider the global deviation of the conditional distribution of X_{T+m} given X_T (Yao and Tong 1995b; Fan, Yao, and Tong 1996), which measures the error in the m-step-ahead predictive distribution due to a drift δ in the initial value. For this purpose, it is more convenient to consider the conditional density function $g_m(\cdot|x)$ of X_{T+m} given $X_T = x$ instead of the conditional distribution. Naturally, we may use the mutual information based on the Kullback–Leibler information, which is expressed as

$$K_m(x;\delta) = \int \{g_m(y|x+\delta) - g_m(y|x)\} \log\{g_m(y|x+\delta)/g_m(y|x)\} dy.$$

It may be shown that as $\delta \to 0$, $K_m(x;\delta)$ has the approximation

$$K_m(x;\delta) = I_{1,m}(x)\delta^2 + o(\delta^2), \tag{10.27}$$

where

$$I_{1,m}(x) = \int \{\dot{g}_m(y|x)\}^2 / g_m(y|x) dy, \tag{10.28}$$

where $\dot{g}_m(y|x)$ denotes the partial derivative of $g_m(y|x)$ with respect to x. If we treat the initial value x as a parameter of the distribution, $I_{1,m}(x)$ is the Fisher's information, which represents the information on the initial value $X_T = x$ contained in X_{T+m}. Therefore (10.27) may be interpreted as that the more information X_{T+m} contains about the initial state X_T, the more sensitively the distribution depends on the initial condition. The converse is also true. We will see from Proposition 10.1 below that $I_{1,m}(x)$ also controls the sensitivity to initial values of predictive intervals.

We also consider a simple L_2-distance defined as

$$D_m(x;\delta) = \int \{g_m(y|x+\delta) - g_m(y|x)\}^2 dy.$$

It follows from Taylor's expansion that

$$D_m(x;\delta) = I_{2,m}(x)\delta^2 + o(\delta^2),$$

where

$$I_{2,m}(x) = \int \{\dot{g}_m(y|x)\}^2 dy. \tag{10.29}$$

The measures $I_{1,m}$ and $I_{2,m}$ are more informative than the measure derived from the conditional mean approach in §10.1.3. To illustrate this point, we consider the following one-dimensional model

$$X_{t+1} = \alpha X_t + \sigma(X_t)\varepsilon_{t+1},$$

where α is a real constant, $\sigma(\cdot)$ is a positive and differentiable function, and $\{\varepsilon_t\}$ are i.i.d. standard normal. It is easy to see from (10.12) that $\dot{f}_m(x) = \alpha^m$ is a constant, which indicates that when $|\alpha| < 1$, the system is globally as well as locally stable as far as the conditional mean is concerned. However, both $I_{1,m}(\cdot;\cdot)$ and $I_{2,m}(\cdot;\cdot)$ are no longer constants. For example,

$$I_{1,1}(x) = \frac{1}{\sigma^2(x)}\{\alpha^2 + 2[\dot{\sigma}(x)]^2\}, \qquad I_{2,1}(x) = \frac{1}{4\sqrt{\pi}\sigma^3(x)}\left\{\alpha^2 + \frac{3}{2}[\dot{\sigma}(x)]^2\right\}.$$

Therefore, there is some variation in the sensitivity of the conditional distribution with respect to the initial value x, which is due to the presence of the conditional heteroscedasticity in the model.

The sensitivity of conditional distribution is closely related to the sensitivity of the conditional mean. In fact

$$I_{1,m}(x) \geq \{\dot{f}_m(x)\}^2 / \mathrm{Var}(X_{T+m}|X_T = x).$$

(see Theorem 4.1 of Blyth 1994). This is a conditional version of the famous *Cramer–Rao inequality*. Note that $\dot{f}_m(x)$ measures the sensitivity of the conditional expectation $f_m(x) = E(X_{T+m}|X_T = x)$ (see §10.1.3). The relation above indicates that when the conditional variance is large, $I_{1,m}(x)$ will be small. This reflects the fact that the sensitivity of the conditional distribution will be masked by stochastic noise in the system. For very noisy time series, the predictive errors due to a drift in initial values are relatively negligible.

(b) Monitoring predictive intervals

Let $\Omega_m(X_T)$ be a predictive set for X_{T+m} based on X_T with coverage probability $\alpha \in (0,1)$, namely

$$P\{X_{T+m} \in \Omega_m(x)|X_T = x\} = \alpha. \tag{10.30}$$

When $\Omega_m(x)$ is an interval, it is called an interval predictor . If the observation is subject to an error, the real coverage probability may differ adversely from α. Proposition 10.1 indicates that the deviation in the coverage probability may be monitored by the measure $I_{1,m}$ defined in (10.28).

Proposition 10.1 *Suppose that* $b(z) \equiv |\int (\frac{\partial}{\partial z})^2 g_m(y|z)dy|$ *is bounded in a neighborhood of* x. *For any predictive set* $\Omega(\cdot)$ *satisfying (10.30), it holds that*

$$\left|P\{X_{T+m} \in \Omega_m(x)|X_T = x+\delta\} - \alpha\right| \leq |\delta|\{\alpha\, I_{1,m}(x)\}^{1/2} + O(\delta^2).$$

Proof. It follows from (10.30) that

$$\begin{aligned} P\{Y_m \in \Omega_m(x)|X_0 = x+\delta\} &= \int_{\Omega_m(x)} g_m(y|x+\delta)dy \\ &= \int_{\Omega_m(x)} \{g_m(y|x) + \delta \dot{g}_m(y|x)\}dy + O(\delta^2) \\ &= \alpha + \int_{\Omega_m(x)} \delta \dot{g}_m(y|x)dy + O(\delta^2). \end{aligned}$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned} \left| \int_{\Omega_m(x)} \delta \dot{g}_m(y|x)dy \right| &\le \left\{ \int_{\Omega_m(x)} g_m(y|x)dy \int \frac{\{\delta \dot{g}_m(y|x)\}^2}{g_m(y|x)} dy \right\}^{\frac{1}{2}} \\ &\le |\delta| \{\alpha I_{1,m}(x)\}^{\frac{1}{2}}. \end{aligned}$$

■

(c) Estimation of $I_{1,m}(x)$ and $I_{2,m}(x)$

The estimators for $I_{1,m}(x)$ and $I_{2,m}(x)$ may be easily constructed by plugging in the estimators for $g_m(y|x)$ and $\dot{g}_m(y|x)$ presented in §6.5. Due to the simple form of (10.29), the resulting estimator for $I_{2,m}(x)$ reduces to a relatively simple explicit form (10.32) below. In contrast, the plug-in estimator for $I_{1,m}(x)$ involves an integral of a ratio of two estimators, which is less stable. An alternative will also be suggested. We outline those estimation methods below. A more detailed account of the asymptotic properties, bandwidth selection, and numerical illustration is available from Fan, Yao, and Tong (1996).

Suppose that $X_1, \cdots, X_T$ are observations from a strictly stationary time series. Let $p = 2$ in (6.54). The resulting local quadratic estimators can be written as $\widehat{g}_m(y|x) = \beta_0(x,y)$ and $\widehat{\dot{g}}_m(y|x) = \beta_1(x,y)$, where

$$\widehat{\beta}_j(x,y) = h_1^{-1} \sum_{t=1}^{T-m} W_j^T \left(\frac{X_t - x}{h_1} \right) K_{h_2}(X_{t+m} - y), \quad j = 0, 1. \tag{10.31}$$

In the expression above, K and W are kernel functions, $h_1, h_2 > 0$ are bandwidth,

$$W_j^T(t) = e_j^\tau S_T^{-1}(1, h_1 t, h_1^2 t^2)^\tau \times W(t),$$

with e_j being the unit vector with $(j+1)$ element 1, and

$$S_T = \begin{pmatrix} s_0 & s_1 & s_2 \\ s_1 & s_2 & s_3 \\ s_2 & s_3 & s_4 \end{pmatrix}, \quad s_j = \sum_{t=1}^{T-m} (X_t - x)^j W_{h_1}(X_t - x).$$

With the derivative of the conditional density estimated by (10.31), a natural estimator for $I_{2,m}(x)$ is

$$\widehat{I}_{2,m}(x) = \int \widehat{\beta}_1^2(x,y)dy = \frac{1}{h_1^2}\sum_{i=1}^{T-m}\sum_{j=1}^{T-m} W_1^T\left(\frac{X_i - x}{h_1}\right) \times W_1^T\left(\frac{X_j - x}{h_1}\right)\int K_{h_2}(X_{i+m} - y)K_{h_2}(X_{j+m} - y)dy.$$

Assume that the kernel $K(\cdot)$ is symmetric. Then

$$\int K_{h_2}(X_i - y)K_{h_2}(X_j - y)dy = K^*_{h_2}(X_i - X_j),$$

where $K^* = K * K$ is a convolution of the kernel function K with itself. Thus, the proposed estimator can be expressed as

$$\widehat{I}_{2,m}(x) = \frac{1}{h_1^2}\sum_{i=1}^{T-m}\sum_{j=1}^{T-m} W_1^T\left(\frac{X_i - x}{h_1}\right) W_1^T\left(\frac{X_j - x}{h_1}\right) K^*_{h_2}(X_{i+m} - X_{j+m}). \tag{10.32}$$

The asymptotic normality for the estimator above was established in Fan, Yao, and Tong (1996).

Analogously, an estimator for $I_{1,m}(x)$ can be defined by

$$\widehat{I}_{1,m}(x) = \int \widehat{\beta}_1^2(x,y)/\widehat{\beta}_0(x,y)dy,$$

with the usual convention $0/0 = 0$. The integration above is typically finite under some mild conditions. However, this estimator cannot be simplified easily.

An alternative estimator to $I_{1,m}(x)$ originates from the fact that

$$I_{1,m}(x) = 4\int \left\{\frac{\partial\sqrt{g_m(y|x)}}{\partial x}\right\}^2 dy.$$

For given bandwidths h_1 and h_2, define

$$C(X_i, X_{i+m}) = \#\{(X_t, X_{t+m}), 1 \le t \le T-m \,:\, |X_t - X_i| \le h_1 \text{ and } |X_{t+m} - X_{i+m}| \le h_2\},$$

$$C(X_i) = \#\{X_t, 1 \le t \le T-m, \,:\, |X_t - X_i| \le h_1\},$$

for $1 \le i \le n$. Then

$$Z_t \equiv [C(X_t, X_{t+m})/\{C(X_t)\, h_2\}]^{1/2}$$

is a natural estimate of $q(x,y) \equiv \{g_m(y|x)\}^{1/2}$ at $(x,y) = (X_t, X_{t+m})$. Fitting it into the context of locally quadratic regression, we may estimate

$q(x,y)$ and its first- and second-order partial derivatives with respect to x, which are denoted by $\dot{q}(x,y)$ and $\ddot{q}(x,y)$, respectively, by using $\widehat{q}(x,y)=\widehat{a}$, $\widehat{\dot{q}}(x,y)=\widehat{b}$, and $\widehat{\ddot{q}}(x,y)=\widehat{c}$, where $(\widehat{a},\widehat{b},\widehat{c})$ is the minimizer of the function

$$\sum_{t=1}^{T-m}\{Z_t-a-b(X_t-x)-c(X_t-x)^2/2\}^2 H\left(\frac{X_t-x}{h_1},\frac{X_{t+m}-y}{h_2}\right),$$

H being a probability density function on R^2. Consequently, we estimate $I_{1,m}(x)$ by

$$\widetilde{I}_{1,m}(x)=4\int\{\widehat{q}(x,y)\}^2\,dy.$$

10.4 Interval Predictors and Predictive Sets

For linear time series models with normally distributed errors, the predictive distributions are normal. Therefore, the predictive intervals are easily obtained using the mean plus and minus a multiple of the standard deviation. The width of such an interval, even for multistep-ahead prediction, is constant over the whole state-space. Predictive intervals constructed in this way have also been used for some special nonlinear models, such as threshold autoregressive models (see Davis, Pembertn, and Petruccelli 1988, Tong and Moeanaddin 1988). However, the method above is no longer pertinent when the predictive distribution is not normal, which, unfortunately, is the case for most nonlinear time series, especially with multiple-step-ahead prediction. Yao and Tong (1995b, 1996) proposed to construct predictive intervals using conditional quantiles (percentiles) (see also (10.22)). However, interval predictors so constructed are inappropriate when the predictive distributions are asymmetric and/or multimodal.

Asymmetric distributions have been widely used in modeling economic and financial data. Further, skewed predictive distributions may occur in multistep-ahead prediction even though the errors in the models have symmetric distributions (see Figure 10.7(a) below). Multimodal phenomena often indicate model uncertainty. The uncertainty may be caused by factors beyond the variables specified in the prediction (see Figure 10.7(b)). In order to cope with the possible skewness and multimodality of the underlying predictive distribution, Polonik and Yao (2000) suggested searching for the set with the minimum length (i.e., Lebesgue measure) among all candidate predictive sets. We introduce their approach in this section. For the theoretical properties of the minimum-length predictors, we refer to Polonik and Yao (2000, 2002).

10.4.1 Minimum-Length Predictive Sets

Suppose that $\{(Y_t, \mathbf{X}_t)\}$ is a strictly stationary process. We consider the predictive sets for Y_t based on $\mathbf{X}_t$. In the time series context, $Y_t = X_{t+m}$ for some $m \geq 1$ fixed and $\mathbf{X}_t = (X_t, \cdots, X_{t-p+1})$. Let $F(\cdot|\mathbf{x})$ denote the conditional distribution of Y_t given $\mathbf{X}_t = \mathbf{x}$. Now, we treat $F(\cdot|\mathbf{x})$ as a function defined on all measurable sets in R, and we adopt the convention that

$$F(y|\mathbf{x}) = F((-\infty, y]|\mathbf{x}), \qquad y \in R.$$

A general form of predictive set is defined as in (10.30). However, in practice, we would restrict our attention to some simple types of predictive sets only. Let $\mathcal{C}$ denote a class of measurable subsets of R, which defines candidate predictive sets. Usually, $\mathcal{C}$ consists of all intervals in R, or all unions of two intervals, and so on. For $\alpha \in [0, 1]$ and $\mathbf{x} \in R^p$, define

$$\mathcal{C}_\alpha(\mathbf{x}) = \{C \in \mathcal{C} \,:\, F(C|\mathbf{x}) \geq \alpha\}.$$

The minimum length predictor may be formally defined as follows.

Definition 10.1 *The set in* $\mathcal{C}_\alpha(\mathbf{x})$ *with the minimum Lebesgue measure is called the* minimum length predictor *for* Y_t *based on* $\mathbf{X}_t = \mathbf{x}$ *in* $\mathcal{C}$ *with coverage probability* α*, which is denoted* $M_{\mathcal{C}}(\alpha|\mathbf{x})$.

The minimum-length predictor depends on the current position $\mathbf{X}_t = \mathbf{x}$. It defines a set on which the predictive distribution $F(\cdot|\mathbf{x})$ has the largest (probability) mass concentration in the sense that it has the minimum Lebesgue measure among all sets in a given class with the nominal coverage probability. Suppose that the conditional density $g(y|\mathbf{x})$ of Y_t given $\mathbf{X}_t = \mathbf{x}$ exists. Then, it is clear that the minimum-length predictor is given by

$$\{y : g(y|\mathbf{x}) \geq \lambda_\alpha\}$$

(i.e., the high-density region), where λ_α is the constant such that the prediction set has coverage probability α.

The minimum-length predictor of all of the predictive intervals is the one with the shortest length. For a symmetric and unimodal predictive distribution, a minimum-length predictor reduces to the quantile interval

$$I(\alpha|\mathbf{x}) \equiv [\, F^{-1}(0.5 - 0.5\alpha|\mathbf{x}),\ F^{-1}(0.5 + 0.5\alpha|\mathbf{x})\,]. \tag{10.33}$$

In general, the minimum-length predictor $M_{\mathcal{C}}(\alpha|\mathbf{x})$ may not be unique. Furthermore, it may not exist for some $\mathcal{C}$; see Polonik and Yao (2000). Ideally, we should specify the candidate set $\mathcal{C}$ according to the profile of the predictive distribution $F(\cdot|\mathbf{x})$. For example, when $F(\cdot|\mathbf{x})$ is k modal, $\mathcal{C}$ may consist of all of the unions of at most k intervals. In practice, we often let $\mathcal{C}$ be the set of all (connected) intervals, or all unions of two intervals. We denote them as

$$M_1(\alpha|\mathbf{x}) = M_{\mathcal{C}}(\alpha|\mathbf{x}) \quad \text{for} \quad \mathcal{C} = \{[a, b] \,:\, -\infty \leq a < b \leq \infty\} \tag{10.34}$$

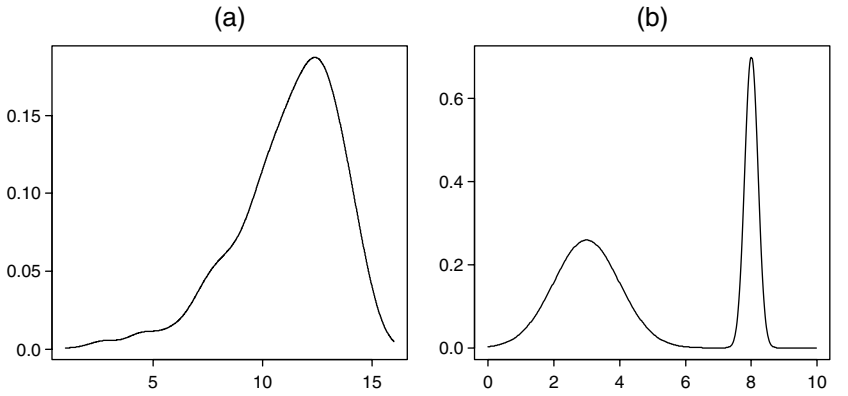


FIGURE 10.7. (a) Conditional density function of X_{t+3} given $X_t = 8$ for model (10.36). (b) Conditional density function of X_{t+1} given $X_t = 0$ for model (10.37). From Polonik and Yao (2000).

and

$$M_2(\alpha|\mathbf{x}) = M_{\mathcal{C}}(\alpha|\mathbf{x}) \quad \text{for} \quad \mathcal{C} = \{[a,b] \cup [c,d] \,:\, -\infty \le a < b \le c < d \le \infty\}. \tag{10.35}$$

The minimum-length predictor $M_1(\alpha|\mathbf{x})$ is the predictive interval with the shortest length. The predictor $M_2(\alpha|\mathbf{x})$ may consist of two disconnected intervals if $F(\cdot|\mathbf{x})$ has more than one mode.

To illustrate the basic ideas of minimum length predictors, we look into two toy models first.

We start with a simple quadratic model

$$X_t = 0.23X_{t-1}(16 - X_{t-1}) + 0.4\varepsilon_t, \tag{10.36}$$

where $\{\varepsilon_t\}$ is a sequence of independent random variables each with the standard normal distribution truncated in the interval $[-12,\ 12]$. The conditional distribution of $Y_t \equiv X_{t+m}$ given X_t is symmetric for $m = 1$ but not necessarily so for $m > 1$. For example, the conditional density function at $X_t = 8$ with $m = 3$ is depicted in Figure 10.7(a), which is obviously skewed to the left. Based on this density function, two types of predictive intervals with three different coverage probabilities are specified in Table 10.2. For example, the quantile interval $I(\alpha|x) = [5.11, 14.96]$ for $\alpha = 0.95$ and $x = 8$. It contains some lower-density points near its left end-point due to the skewness of the distribution; see Figure 10.7(a). The minimum-length interval $M_1(\alpha|x) = [6.50, 15.30]$ could be regarded as a compressed shift to the right of the quantile interval with 10.57% reduction in its length. Obviously, the accuracy of prediction has been substantially improved by using the minimum-length interval.

TABLE 10.2. Two predictive sets for X_{t+3} based on $X_t = 8$ (i.e. $x = 8$) for model (10.36). The percentage decreases in length relative to $I(\alpha|x)$ are recorded in parentheses.

$\alpha = 0.95$	$I(\alpha\|x)$	[5.11, 14.96]	
	$M_1(\alpha\|x)$	[6.50, 15.30]	(10.57%)
$\alpha = 0.70$	$I(\alpha\|x)$	[8.66, 13.95]	
	$M_1(\alpha\|x)$	[9.64, 14.15]	(8.72%)
$\alpha = 0.50$	$I(\alpha\|x)$	[9.83, 12.95]	
	$\mathrm{M}_1(\alpha\|x)$	[10.69, 13.56]	(8.01%)

Now, we consider the model

$$X_t = 3\cos\left(\frac{\pi X_{t-1}}{10}\right) + Z_{t-1} + \frac{1}{0.8Z_{t-1}+1}\varepsilon_t, \tag{10.37}$$

where $\{\varepsilon_t\}$ and $\{Z_t\}$ are two independent i.i.d. sequences with $\varepsilon_t \sim N(0,1)$ and $P(Z_1 = 0) = 0.65 = 1 - P(Z_1 = 5)$. For the sake of illustration, we assume that the "exogenous" variable Z_t is unobservable. We predict $Y_t \equiv X_{t+1}$ from X_t only. Thus, the (theoretical) least squares conditional point predictor is $3\cos(0.1\pi X_t) + 1.75$, which obviously is not satisfactory. It is easy to see that the conditional distribution of X_{t+1} given X_t is a mixture of two normal distributions. Figure 10.7(b) depicts the conditional density function at $X_t = 0$. For the three different values of α, Table 10.3 records the three types of predictive sets: the quantile interval $I(\alpha|x)$, the minimum-length intervals $M_1(\alpha|x)$, and the minimum-length set with at most two intervals $M_2(\alpha|x)$. We can see that the percentile interval fails to do a reasonable job simply because the predictive intervals are too wide. The improvement by using $M_1(\alpha|x)$ is not substantial unless the coverage probability α is small enough that the probability mass around one mode exceeds α. The set $M_2(\alpha|x)$ is much shorter in length (i.e., Lebesgue measure) and therefore offers a much more accurate prediction. All three $M_2(\alpha|x)$ consist of two disconnected intervals, which clearly reveals the uncertainty in X_{t+1} caused by the "hidden" variable Z_t. The coverage probabilities of the two intervals centered at 3 and 8 are 0.61 and 0.34, 0.38 and 0.32, and 0.20 and 0.30 when the global coverage probability is 0.95, 0.70, and 0.50, respectively.

The two simple examples above indicate that we should seek the minimum-length predictive sets when the conditional distribution of Y given X is skewed and/or multimodal. The number of intervals used in the predictor should be equal, or at least close, to the number of modes of the conditional distribution, subject to practical feasibility.

TABLE 10.3. Three predictive sets for X_{t+1} based on $X_t = 0$ (i.e., $x = 0$) for model (10.37). The percentage decreases in length relative to $I(\alpha|x)$ are recorded in parentheses.

	$I(\alpha\|x)$	[1.23, 8.30]	
$\alpha = 0.95$	$\mathrm{M}_1(\alpha\|x)$	[1.50, 8.42]	(2.12%)
	$\mathrm{M}_2(\alpha\|x)$	[1.15, 4.84]∪[7.54, 8.46]	(34.79%)
	$I(\alpha\|x)$	[2.26, 8.04]	
$\alpha = 0.70$	$\mathrm{M}_1(\alpha\|x)$	[2.80, 8.29]	(5.02%)
	$\mathrm{M}_2(\alpha\|x)$	[2.18, 3.82]∪[7.66, 8.34]	(59.86%)
	$I(\alpha\|x)$	[2.71, 7.89]	
$\alpha = 0.50$	$\mathrm{M}_1(\alpha\|x)$	[1.80, 4.20]	(53.67%)
	$\mathrm{M}_2(\alpha\|x)$	[2.60, 3.40]∪[7.71, 8.29]	(73.36%)

10.4.2 Estimation of Minimum-Length Predictors

Constructing a minimum-length predictor involves three steps: (i) estimating the conditional distribution $F(\cdot|\mathbf{x})$, (ii) specifying the set $\mathcal{C}$, and (iii) searching for $M_{\mathcal{C}}(\alpha|\mathbf{x})$ with F replaced by its estimator; see Definition 10.1. As we have discussed in §10.4.1, the specification of $\mathcal{C}$ is dictated by the profile of F as well as practical considerations. In fact, interval predictors such as (10.34) are often preferred in practice. In this case, the search for the minimum-length predictive interval $M_1(\alpha|\mathbf{x})$ reduces to the search for its two end points, which may be carried out, for example, in the manner of exhaustive searching among the observed data points.

Suppose that $(Y_t, \mathbf{X}_t), 1 \le t \le T$, are observations from a strictly stationary process. An estimator for the conditional distribution $F(C|\mathbf{x})$ may be obtained by any local regression of $I(Y_t \in C)$ on $\mathbf{X}_t$. For example, the local logistic estimation or the adjusted Nadaraya–Watson estimator of §10.3 may be adopted for this purpose. As a simple illustration, we use the standard Nadaraya–Watson estimator

$$\widehat{F}(C|\mathbf{x}) = \sum_{t=1}^{T} I(Y_t \in C) K\left(\frac{\mathbf{X}_t - \mathbf{x}}{h}\right) \Bigg/ \sum_{t=1}^{T} K\left(\frac{\mathbf{X}_t - \mathbf{x}}{h}\right),$$

where $K(\cdot)$ is a kernel function defined on R^p, and $h > 0$ is a bandwidth. Replacing $F(\cdot|\mathbf{x})$ by $\widehat{F}(\cdot|\mathbf{x})$ in Definition 10.1, we obtain an estimator for the minimum-length predictor, which may be expressed as

$$\widehat{M_{\mathcal{C}}}(\alpha|\mathbf{x}) = \arg\min_{C \in \mathcal{C}} \{ \mathrm{Leb}(C) \ : \ \widehat{F}(C|\mathbf{x}) \ge \alpha \}, \tag{10.38}$$

where $\mathrm{Leb}(C)$ denotes the Lebesgue measure of C. Its true coverage probability

$$\widehat{\alpha} \equiv F\{\widehat{M_{\mathcal{C}}}(\alpha|\mathbf{x})|\mathbf{x}\}$$

converges to the nominal coverage probability α. In fact, Polonik and Yao (2000) show that under some regularity conditions,

$$\{Th^p f(\mathbf{x})\}^{1/2}(\widehat{\alpha} - \alpha) \xrightarrow{D} N\,(0,\, \alpha(1-\alpha)\nu_2)$$

as $T \to \infty$, where $\nu_2 = \int K(\mathbf{u})^2 d\mathbf{u}$, and $f(\cdot)$ is the density function of $\mathbf{X}_t$. Furthermore,

$$(Th^p)^{1/2}\frac{f(\mathbf{x})}{\mu(\mathbf{x})}[\mathrm{Leb}\{\widehat{M}_{\mathcal{C}}(\alpha|\mathbf{x})\} - \mathrm{Leb}\{M_{\mathcal{C}}(\alpha|\mathbf{x})\}] \xrightarrow{D} N\,(0,\, \alpha(1-\alpha)\nu_2)\,,$$

where $\mu(\mathbf{x})$ is the partial derivative of $\mathrm{Leb}\{M_{\mathcal{C}}(\alpha|\mathbf{x})\}$ with respect to α.

Like all local regression methods, we need to specify the smooth parameter h in the estimation above. The bootstrap bandwidth selection procedure of §10.3.3 can be adapted for this purpose. To simplify the presentation, we outline the scheme for the case where $\mathcal{C}$ is the set of intervals, as defined in (10.34).

We fit a parametric model

$$Y_t = G(\mathbf{X}_t) + \varepsilon_t, \tag{10.39}$$

where $G(\mathbf{x})$ denotes, for example, a polynomial function of $\mathbf{x}$. We assume that $\{\varepsilon_t\}$ are independent $N(0, \sigma^2)$. The parameters in G and σ^2 are estimated from the data. We form a parametric estimator $\check{M}_1(\alpha|\mathbf{x})$ based on the above model. By Monte Carlo simulation from the model, we compute a bootstrap version $\{Y_1^*, \ldots, Y_n^*\}$ from (10.39) based on given observations $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, and with that a bootstrap version $\widehat{M}_1^*(\alpha|\mathbf{x})$ of $\widehat{M}_1(\alpha|\mathbf{x})$ with $\{(\mathbf{X}_i, Y_i)\}$ replaced by $\{(\mathbf{X}_i, Y_i^*)\}$. Define

$$D(h) = E[\mathrm{Leb}\{\widehat{M}_1^*(\alpha|\mathbf{x}) \,\triangle\, \check{M}_1(\alpha|\mathbf{x})\}|\{\mathbf{X}_i, Y_i\}],$$

where $A \triangle B = (A - B) \cup (B - A)$ is the symmetric difference of sets A and B. Choose $h = \widehat{h}$ to minimize $D(h)$. In practice, $D(h)$ is evaluated via repeated bootstrap sampling.

In principle, there are no difficulties in extending the idea above for estimation of $M_2(\alpha|\mathbf{x})$ defined in (10.35). We may, for example, let ε_t have a mixture of two normal distributions, although the bootstrap search for $\widehat{M}_2^*(\alpha|\mathbf{x})$ is computationally more expensive. Our experience suggests that the choice between the two predictors $\widehat{M}_1(\alpha|\mathbf{x})$ and $\widehat{M}_2(\alpha|\mathbf{x})$ does not depend on the bandwidth sensitively unless there occurs a bifurcation to the conditional distribution $F(\cdot|\mathbf{x})$ around $\mathbf{x}$. Note that our problem is different in nature from that in Silverman's test for multimodality (Silverman 1981). If we were interested in determining the number of modes in the curve $P(\mathbf{x}) \equiv F(C|\mathbf{x})$, the bandwidth used in estimation would play a critical role.

An obvious alternative to the minimum-length approach introduced above is to construct a predictive set based on an estimated conditional density

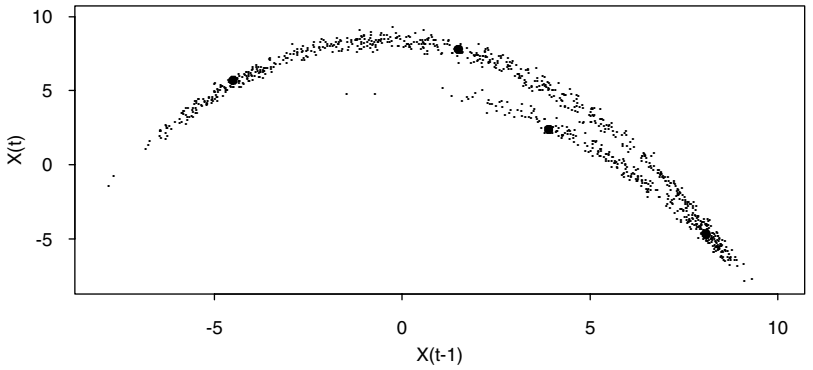


FIGURE 10.8. Scatterplot of X_t against X_{t-1} from a sample of size 1,000 generated from model (10.40). The positions marked with a "•" are (from left to right) $(-4.5, 5.7)$, $(1.5, 7.8)$, $(3.9, 2.4)$, and $(8.1, -4.7)$. From Polonik and Yao (2000).

function (see §6.5). Let $g(\cdot|\mathbf{x})$ be the conditional density function of Y_t given $\mathbf{X}_t = \mathbf{x}$. The minimum-length predictor may be defined as

$$M(\alpha|x) = \{y \,:\, g(y|\mathbf{x}) \geq \lambda_\alpha\},$$

where λ_α is the maximum value for which

$$\int_{\{y:\, g(y|\mathbf{x}) \geq \lambda_\alpha\}} g(y|\mathbf{x})dy \geq \alpha.$$

Note that we do not need to specify the candidate set $\mathcal{C}$ in this approach. However, the estimation of $g(y|\mathbf{x})$ involves smoothing in both y and $\mathbf{x}$ directions; see §6.5. Further, we argue that minimum-length predictors in different sets $\mathcal{C}$ may provide valuable information on the shape of the conditional distribution of Y_t given $\mathbf{X}_t$; see Examples 10.5 and 10.6 in §10.4.3 below.

10.4.3 Numerical Examples

To appreciate the finite sample properties of the estimated minimum-length predictors, we illustrate the methods via one nonlinear AR(2) model and the rainfall and river flow data from a catchment in Wales. (This data set was provided by Professor Peter C. Young.) We always use the Gaussian kernel. Setting $\alpha = 0.9$, we calculate estimators for the minimum-length interval $M_1(\alpha|\mathbf{x})$ of (10.34) and minimum-length predictors with at most two intervals $M_2(\alpha|\mathbf{x})$ of (10.35).

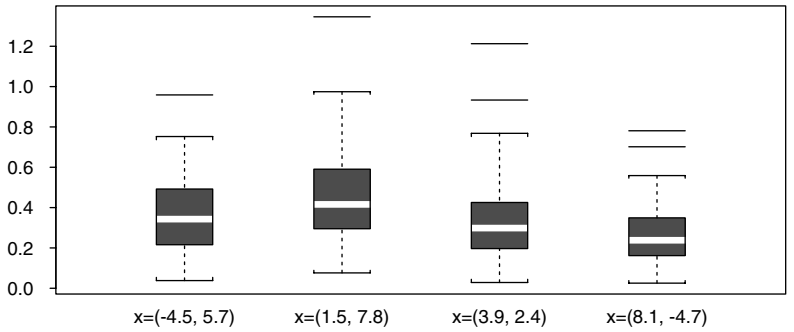


FIGURE 10.9. Boxplots of the ratio of Leb$\{\widehat{M}_1(\alpha|\mathbf{x})\Delta M_1(\alpha|\mathbf{x})\}$ to Leb$\{M_1(\alpha|\mathbf{x})\}$ for model (10.40). From Polonik and Yao (2000).

Example 10.5 First we consider the simulated model

$$X_t = 6.8 - 0.17X_{t-1}^2 + 0.26X_{t-2} + 0.3\varepsilon_t, \tag{10.40}$$

where $\{\varepsilon_t\}$ is a sequence of independent random variables, each with the standard normal distribution truncated in the interval $[-12, 12]$. We conduct the simulation in two stages to estimate the minimum-length predictors for $Y_t \equiv X_{t+1}$ given (i) $\mathbf{X}_t \equiv (X_t, X_{t-1})$ and (ii) X_t, respectively.

(i) For four fixed values of $\mathbf{X}_t = (X_t, X_{t-1})$, we repeat the simulation 100 times with sample size n =1,000. Figure 10.8 is the scatterplot of a sample of size n =1,000. The four positions marked with a "•" are the values of $\mathbf{X}_t = \mathbf{x}$ at which the predictor $M_1(\alpha|\mathbf{x})$ is estimated. Figure 10.9 presents the boxplots of Leb$\{\widehat{M}_1(\alpha|\mathbf{x})\Delta M_1(\alpha|\mathbf{x})\}$/Leb$\{M_1(\alpha|\mathbf{x})\}$. The bandwidths were selected by the bootstrap scheme stated in §10.4.2 based on parametric models determined by AIC. With the given sample sizes, AIC always identified the correct model from the candidate polynomial model of order 3.

(ii) For a sample of size 1,000, we estimate both predictors $M_1(\alpha|x)$ and $M_2(\alpha|x)$ for $Y_t \equiv X_{t+1}$ given its first lagged value $X_t = x$ only. We let x range over 90% of the inner samples. We use a postsample of size 100 to check the performance of the predictors. For estimating bandwidths using the proposed bootstrap scheme, the parametric model selected by the AIC is

$$X_t = 8.088 - 0.316X_{t-1} - 0.179X_{t-1}^2 + 0.003X_{t-1}^3 + 0.825\varepsilon_t.$$

Figure 10.10(a) displays the estimated $M_1(\alpha|x)$ together with the 100 postpoints. Within the range of values of x on which estimation is conducted, $\widehat{M}_1(\alpha|x)$ contains about 90% of the postsample. Note that $\widehat{M}_1(\alpha|x)$ has an abrupt change in the width around $x = 1.5$. In fact, the predictor $M_1(\alpha|x)$

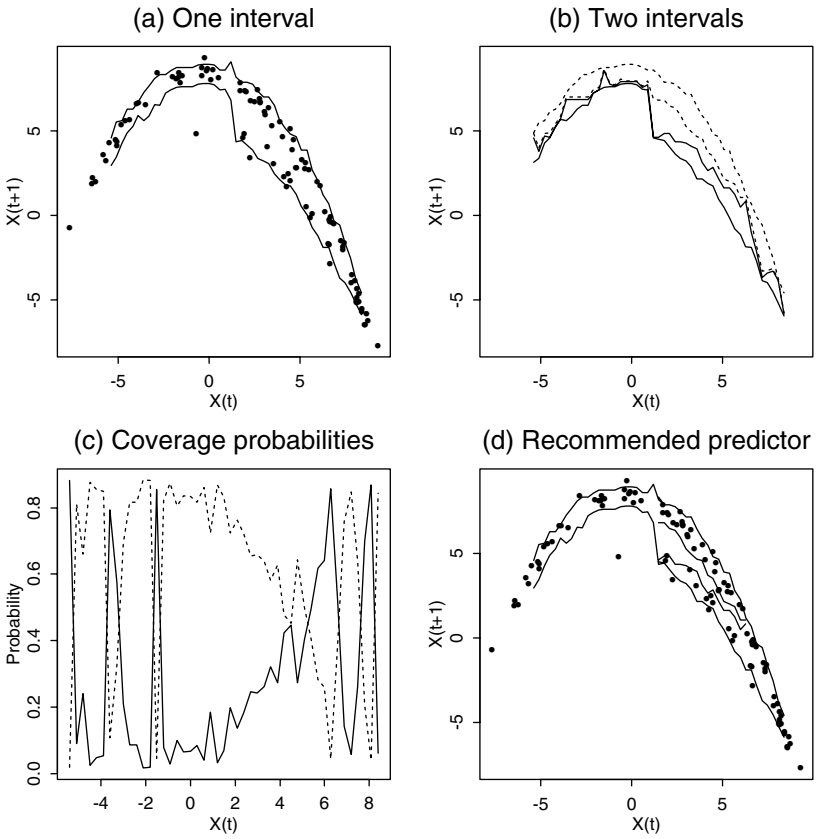


FIGURE 10.10. Simulation results for model (10.40). (a) $\widehat{M}_1(\alpha|x)$ together with 100 postsamples. (b) $\widehat{M}_2(\alpha|x)$. The two disconnected intervals are bounded with solid lines and dashed lines, respectively. (c) Coverage probabilities of the two intervals in (b): solid curve is the coverage probability for the interval with a solid boundary; dashed curve is the coverage probability for the interval with a dashed boundary. (d) The recommended minimum-length predictor for X_{t+1} given X_t, together with 100 postsamples. From Polonik and Yao (2000).

is not satisfactory for x between 1.5 and 6 because the center of the intervals is void for those x-values (see also Figure 10.8). Therefore, it is possible to construct more accurate predictive sets (i.e., with smaller lengths) for those x-values. The estimator $\widehat{M}_2(\alpha|x)$ is plotted in Figure 10.10(b). Due to sampling fluctuation, the estimator always consists of two disconnected intervals over the whole sample space. The coverage probabilities of the two intervals are plotted in Figure 10.10(c). From Figures 10.10 (b) and (c), we note that when $x \notin [1.5, 6.3]$, the two intervals of $\widehat{M}_2(\alpha|x)$ are almost connected, and the two corresponding coverage probabilities are either erratic

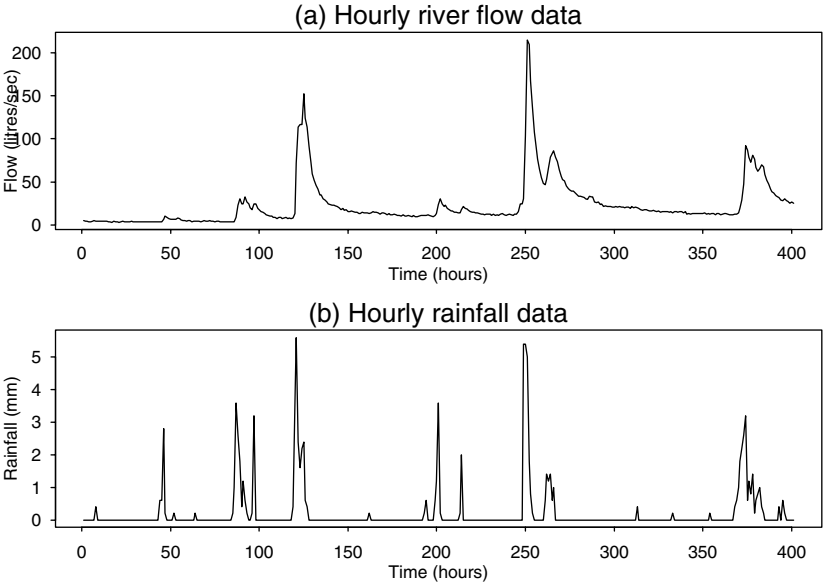


FIGURE 10.11. Hourly rainfall and river flow data from a catchment in Wales. (a) flow (liters/sec); (b) rainfall (mm).

or very close to 0 and 0.9, respectively. Therefore, it seems plausible to use $\widehat{M}_2(\alpha|x)$ for $x \in [1.5, 6.3]$ and $\widehat{M}_1(\alpha|x)$ for $x \notin [1.5, 6.3]$. The combined minimum-length predictor is depicted in Figure 10.10(d) together with the postsample. The combined predictor covers the postsample as well as the $\widehat{M}_1(\alpha|x)$, although its Lebesgue measure has been reduced significantly for $x \in [1.5, 6.3]$. ∎

Example 10.6 Figure 10.11 plots the 401 hourly rainfall and river flow data from a catchment in Wales. We predict the flow from its logged values and the rainfall data. Note that the flow data themselves are strongly autocorrelated (Figure 10.12(a)). Figures 10.12 (b)–(d) indicate that the point-cloud in the scatterplot of flow against rainfall with time lag 2 is slightly thinner than those with time lags 0 and 1, which seems to suggest that the effect of rainfall on the river flow has about a two-hour delay in time. This is further supported by various statistical modeling procedures. In fact, the cross-validation method (Yao and Tong 1994b) specified that the optimal regression subset with two regressors for the flow at the $(t+1)$ hour Y_{t+1} consists of its lagged value Y_t and the rainfall within the $(t-1)$ hour X_{t-1}. This was further echoed by a fitted MARS model (Friedman 1991). We now predict Y_{t+1} from Y_t and X_{t-1} using three predictors given in (10.33)–(10.35). We estimate the predictors using the data with sample size $n = 394$ resulting from leaving out the 373rd, the 375th, and the last

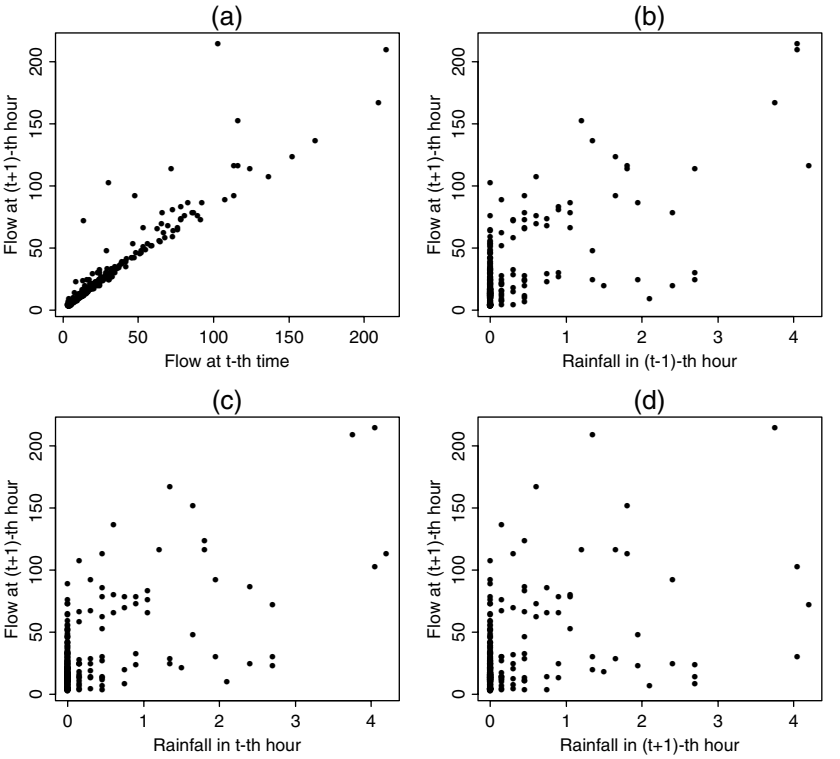


FIGURE 10.12. Hourly rainfall and river flow data from a catchment in Wales. (a) scatterplot of flow data; (b)–(d) scatterplots of rainfall against flow at time lags 2, 1, and 0, respectively.

five flow data (therefore also their corresponding lagged values and the rainfall data) in order to check the reliability of the prediction. We standardize the observations of regressors before the fitting. We adopt the bootstrap scheme to select the bandwidth. The parametric model determined by AIC is

$$\begin{aligned} Y_{t+1} &= -1.509 + 1.191Y_t + 0.924X_{t-1} + 0.102Y_tX_{t-1} \\ &\quad -0.004Y_t^2 + 7.902\varepsilon_{t+1}, \end{aligned}$$

where ε_t is the standard normal. Table 10.4 reports the estimated predictors for the seven data points that are not used in estimation. All of the quantile intervals cover the corresponding true values. For the minimum-length predictor $M_1(\alpha|\mathbf{x})$, six out of seven intervals contain the true value. The only exception occurs when there is a high burst of river flow at the value 86.9. It is easy to see from Figure 10.11 that data are sparse at this level and upward. Due to the quick river flow caused by rainfall, we expect the predictive distributions to be skewed to the right. Therefore, the im-

TABLE 10.4. The three predictive sets for the river flow Y_{t+1} from its lagged value Y_t and the lagged rainfall X_{t-1}. The coverage probabilities of the two intervals in $\widehat{M}_2(\alpha|x)$ are recorded in parentheses. From Polonik and Yao (2000).

Y_{t+1}	(Y_t, X_{t-1})	$\widehat{I}(\alpha\|\mathbf{x})$	$\widehat{M}_1(\alpha\|\mathbf{x})$	$\widehat{M}_2(\alpha\|\mathbf{x})$
47.9	(29.1, 1.8)	[3.8, 71.8]	[3.3, 51.1]	[3.6, 42.3]∪[65.4, 67.5] (0.89, 0.01)
86.9	(92.4, 2.6)	[3.8, 102.9]	[3.3, 76.2]	[3.3, 53.1]∪[62.6, 78.5] (0.83, 0.07)
28.0	(30.7, 0.6)	[5.6, 39.3]	[3.8, 33.5]	[4.3, 5.6]∪[6.6, 34.6] (0.03, 0.87)
27.5	(28.0, 0.2)	[4.7, 35.8]	[3.8, 32.4]	[4.1, 26.9]∪[30.7, 34.6] (0.82, 0.08)
25.4	(27.5, 0.0)	[4.4, 34.6]	[3.8, 32.4]	[3.6, 24.7]∪[30.7, 34.6] (0.02, 0.88)
26.9	(25.4, 0.0)	[7.7, 33.5]	[9.3, 33.5]	[9.7, 26.9]∪[29.1, 34.1] (0.81, 0.09)
25.4	(26.9, 0.0)	[4.7, 34.1]	[3.6, 31.7]	[3.3, 25.9]∪[30.7, 34.1] (0.83, 0.07)

provement in prediction should be observed by using the minimum-length predictor $M_1(\alpha|\mathbf{x})$ instead of the quantile interval $I(\alpha|\mathbf{x})$. In fact, even if we discard the case where the true Y_{t+1} lies outside of $\widehat{M}_1(\alpha|\mathbf{x})$, the relative decrease in length of $\widehat{M}_1(\alpha|\mathbf{x})$ with respect to the quantile predictor $\widehat{I}(\alpha|\mathbf{x})$ is between 4.4% and 27.9% for the six other cases. Actually, $\widehat{M}_1(\alpha|\mathbf{x})$ could be regarded as a compressed shift of $\widehat{I}(\alpha|\mathbf{x})$ to its left in six out of seven cases. For application to data sets like this, it is pertinent to use the state-dependent bandwidths. For example, for estimating $M_1(\alpha|\mathbf{x})$, our bootstrap scheme selected, respectively, the quite large bandwidths 1.57 and 2.99 for the first two cases in Table 10.4 in response to the sparseness of data in the area with positive rainfall and quick river flow. The selected bandwidths for the last five cases are rather stable and are between 0.33 and 0.43. There seems little evidence suggesting multimodality, for the estimated $M_2(\alpha|\mathbf{x})$ always contains one interval with very tiny coverage probability. For this example, we recommend using the interval predictor $M_1(\alpha|\mathbf{x})$.

In the application above, we have included a single rainfall point X_{t-2} in the model for the sake of simplicity. A more pertinent approach should take into account the moisture condition of soil, which depends on prior rainfall. For a more detailed discussion of this topic, see Young (1993) and Young and Bevan (1994). ∎

10.5 Complements

We prove Theorem 10.2 now. We derive only (10.24), noting that a proof of (10.25) is similar but simpler. We introduce a lemma first.

Lemma 10.1 *Under conditions (C1) — (C4),* $\widehat{F}(y|x) \xrightarrow{P} F(y,x)$*, and for* $i = 1, \ldots r-1$*,*

$$\widehat{F}^{(i)}(y|x) \equiv L^{(i)}(0, \widehat{\boldsymbol{\theta}}_{xy}) \xrightarrow{P} F^{(i)}(y|x).$$

In fact, $\widehat{\boldsymbol{\theta}}_{xy} \xrightarrow{P} \boldsymbol{\theta}^{(0)}$*, where* $\boldsymbol{\theta}^{(0)}$ *is uniquely determined by (10.23).*

Proof. We only need to prove that $\widehat{\boldsymbol{\theta}}_{xy} \xrightarrow{P} \boldsymbol{\theta}^0$; see (10.23). Since $\widehat{\boldsymbol{\theta}}_{xy}$ is the minimizer of $R(\boldsymbol{\theta}; x, y)$ defined in (10.18), $D_n(x, \widehat{\boldsymbol{\theta}}_{xy}) = 0$, where

$$\begin{aligned} D_n(x,\boldsymbol{\theta}) \;=\; & \frac{1}{nh^{r-1}} \sum_{i=1}^{n} \left(1, \frac{X_i - x}{h}, \ldots, \left(\frac{X_i - x}{h}\right)^{r-1}\right)^{\tau} \\ & \times \{I(Y_i \le y) - L(X_i - x, \boldsymbol{\theta})\} L(X_i - x, \boldsymbol{\theta}) \{1 - L(X_i - x, \boldsymbol{\theta})\} \\ & \times K_h(X_i - x). \end{aligned}$$

Define

$$\begin{aligned} D(x,\boldsymbol{\theta},h) \;=\; & \frac{f(x)}{h^{r-1}} \int (1, t, \ldots, t^{r-1})^{\tau} L(0,\boldsymbol{\theta}) \{1 - L(0,\boldsymbol{\theta})\} K(t) \\ & \times \sum_{i=0}^{r-1} \frac{(th)^i}{i!} \{F^{(i)}(y|x) - L^{(i)}(0,\boldsymbol{\theta})\} dt. \end{aligned}$$

Obviously, $D(x, \boldsymbol{\theta}^0, h) \equiv 0$. Furthermore, it can be proved that for any compact set G,

$$\sup_{\boldsymbol{\theta} \in G} ||D_n(x,\boldsymbol{\theta}) - D(x,\boldsymbol{\theta},h)|| \xrightarrow{P} 0.$$

Let us assume that $\widehat{\boldsymbol{\theta}}_{xy} \not\xrightarrow{P} \boldsymbol{\theta}^0$. Then, there exists a subsequence of $\{n\}$, still denoted as $\{n\}$ for simplicity of notation, for which $P\{||\widehat{\boldsymbol{\theta}}_{xy} - \boldsymbol{\theta}^0|| > \varepsilon\} > \varepsilon$ for all sufficiently large n, where $\varepsilon > 0$ is a constant. Consequently, $\inf_{||\boldsymbol{\theta} - \boldsymbol{\theta}^0|| \le \varepsilon} ||D_n(x,\boldsymbol{\theta})|| \not\xrightarrow{P} 0$. Hence, we have that

$$\begin{aligned} & \inf_{||\boldsymbol{\theta} - \boldsymbol{\theta}^0|| \le \varepsilon} ||D(x,\boldsymbol{\theta},h)|| \\ \ge\; & \inf_{||\boldsymbol{\theta} - \boldsymbol{\theta}^0|| \le \varepsilon} ||D_n(x,\boldsymbol{\theta})|| - \sup_{||\boldsymbol{\theta} - \boldsymbol{\theta}^0|| \le \varepsilon} ||D_n(x,\boldsymbol{\theta}) - D(x,\boldsymbol{\theta},h)|| \\ =\; & \inf_{||\boldsymbol{\theta} - \boldsymbol{\theta}^0|| \le \varepsilon} ||D_n(x,\boldsymbol{\theta})|| + o_p(1) \not\xrightarrow{P} 0, \end{aligned}$$

which contradicts to the fact that $D(x, \boldsymbol{\theta}^0, h) \equiv 0$. Therefore, it must hold that $\widehat{\boldsymbol{\theta}}_{xy} \xrightarrow{P} \boldsymbol{\theta}^0$. ∎

Proof of Theorem 10.2. For any $\varepsilon \in (0,1)$, it follows from Lemma 10.1 that there exists $\varepsilon_1 \in (0,\infty)$ for which $P\{||\widehat{\boldsymbol{\theta}}_{xy} - \boldsymbol{\theta}^{(0)}|| \le \varepsilon_1\} \ge 1-\varepsilon$ for all sufficiently large n. Let $G \equiv G(\varepsilon_1)$ be the closed ball centered at $\boldsymbol{\theta}^{(0)}$ with radius ε_1. Let $\widehat{\boldsymbol{\theta}}_{xy,G}$ be the minimizer of (10.18) with $\boldsymbol{\theta}$ restricted on G. Define $\widehat{F}_G(y|x) = L(0|\widehat{\boldsymbol{\theta}}_{xy,G})$. Then $P\{\widehat{F}_G(y|x) \ne \widehat{F}(y|x)\} < \varepsilon$ when n is sufficiently large. The argument above indicates that we only need to establish (10.24) for $\widehat{F}_G(y|x)$. Therefore, we proceed with the proof below by assuming that $\widehat{\boldsymbol{\theta}}_{xy}$ is always within a compact set G.

We consider only the case where r is even. Note that $K(.)$ has a bounded support. By simple Taylor expansion on L in (10.18), we have that

$$\begin{aligned} R(\boldsymbol{\theta};x,y) &= \sum_{i=1}^{n}\left(I(Y_i \le y) - \sum_{j=0}^{r-1}\frac{L^{(j)}(0,\boldsymbol{\theta})}{j!}(X_i-x)^j\right. \\ &- \left.\frac{1}{r!}L^{(r)}(c_i(X_i-x),\boldsymbol{\theta})(X_i-x)^r\right)^2 K_h(X_i-x), \end{aligned}$$

where $c_i \in [0,1]$. Define $R^*(\boldsymbol{\theta};x,y)$ as $R(\boldsymbol{\theta};x,y)$ with $\boldsymbol{\theta}$ in $L^{(r)}(c_i(X_i-x),\boldsymbol{\theta})$ replaced by $\widehat{\boldsymbol{\theta}}_{xy}$. Let $\widehat{\boldsymbol{\theta}}^*_{xy}$ be the minimizer of $R^*(\boldsymbol{\theta};x,y)$ and $\widehat{F}^*(y|x) = L(0|\widehat{\boldsymbol{\theta}}^*_{xy})$. In the following, we first prove that (10.24) holds for $\widehat{F}^*(y|x)$. Then, we show that

$$\widehat{F}(y|x) = \widehat{F}^*(y|x) + o_p(h^r). \tag{10.41}$$

It is easy to see that (10.24) follows immediately from the two statements above.

It follows from the least squares theory that

$$\begin{aligned} &\widehat{F}^*(y|x) - F(y|x) \\ =\ & \frac{1}{nh}\sum_{i=1}^{n} W_n\left(\frac{X_i-x}{h},x\right)\left\{I(Y_i \le y) - \sum_{j=0}^{r-1}\frac{F^{(j)}(y|x)}{j!}(X_i-x)^j\right. \\ -\ & \left.\frac{1}{r!}L^{(r)}(c_i(X_i-x),\widehat{\boldsymbol{\theta}}_{xy})(X_i-x)^r\right\} \\ =\ & \frac{1}{nh}\sum_{i=1}^{n} W_n\left(\frac{X_i-x}{h},x\right)\left\{\epsilon_i + \frac{1}{r!}\{F^{(r)}(y|x+c_i'(X_x-x))\right. \\ -\ & \left. L^{(r)}(c_i(X_i-x),\widehat{\boldsymbol{\theta}}_{xy})\}(X_i-x)^r\right\}, \end{aligned} \tag{10.42}$$

where $\epsilon_i = I(Y_i \le y) - F(y|X_i)$, $c_i' \in [0,1]$,

$$W_n(u,x) = (1,0,\ldots,0)S_n(x)^{-1}(1,u,\ldots,u^{r-1})^\tau K(u),$$

and $S_n(x)$ is an $r \times r$ matrix with $s_{i+j-2}(x)$ as its (i,j)th element, and

$$s_j(x) = \frac{1}{nh^j}\sum_{i=1}^{n} K_h(X_i-x)(X_i-x)^j.$$

It follows from the ergodic theorem that $S_n(x) \xrightarrow{P} f(x)S = f(x)(\kappa_{i+j-2})$. We write

$$\xi_i = \sum_{j=1}^{r} \kappa^{(1,j)} \left(\frac{X_i - x}{h} \right)^{j-1},$$

$$R_i = (r!)^{-1}[F^{(r)}\{y|x + c_i'(X_x - x)\} - L^{(r)}\{c_i(X_i - x), \widehat{\boldsymbol{\theta}}_{xy}\}].$$

We have that

$$\begin{aligned} & \widehat{F}^*(y|x) - F(y|x) \\ = & \left\{ \frac{1}{nhf(x)} \sum_{i=1}^{n} \xi_i K \left(\frac{X_i - x}{h} \right) \{\epsilon_i + R_i(X_i - x)^r\} \right\} \{1 + o_p(x)\}. \end{aligned}$$

Note that we have assumed that $\widehat{\boldsymbol{\theta}}_{xy} \in G$. By the ergodic theorem,

$$\frac{1}{nhf(x)} \sum_{i=1}^{n} \xi_i K \left(\frac{X_i - x}{h} \right) R_i(X_i - x)^r = h^r \mu_r(x) + o_p(h^r).$$

In the case where γ given in (C3) is positive, it follows from Theorem 2.22 that $(nh)^{-1/2} \sum_{i=1}^{n} \xi_i K\left(\frac{X_i - x}{h}\right) \epsilon_i$ is asymptotically normal with mean 0 and asymptotic variance

$$f(x)F(y|x)\{1 - F(y|x)\}\tau_r^2(x).$$

The asymptotic normality above can be established in a simpler manner when $\gamma = 0$. We have proved that (10.24) holds for $\widehat{F}^*(y|x)$.

To prove (10.41), note that all of the $L^{(i)}(0, \widehat{\boldsymbol{\theta}}_{xy}^*)$ $(i = 0, 1, \ldots, r-1)$ have explicit expressions such as (10.42). Therefore, it is easy to prove that $L^{(i)}(0, \widehat{\boldsymbol{\theta}}_{xy}^*) \xrightarrow{P} L^{(i)}(0, \boldsymbol{\theta}^0)$, where $\boldsymbol{\theta}^0$ is determined by (10.23). This implies that $\widehat{\boldsymbol{\theta}}_{xy}^* \xrightarrow{P} \boldsymbol{\theta}^0$. Consequently, $|\widehat{\boldsymbol{\theta}}_{xy}^* - \widehat{\boldsymbol{\theta}}_{xy}| \xrightarrow{P} 0$, which implies that $R(\widehat{\boldsymbol{\theta}}_{xy}^*; x, y) = R^*(\widehat{\boldsymbol{\theta}}_{xy}^*; x, y) + o_p(nh^{2r})$ because $\frac{\partial R^*(\boldsymbol{\theta}; x, y)}{\partial \boldsymbol{\theta}} = 0$ at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{xy}^*$. Note that $R(\widehat{\boldsymbol{\theta}}_{xy}; x, y) = R^*(\widehat{\boldsymbol{\theta}}_{xy}; x, y)$ and $\widehat{\boldsymbol{\theta}}_{xy}$ and $\widehat{\boldsymbol{\theta}}_{xy}^*$ are the minimizers of R and R^*. From

$$0 < R(\widehat{\boldsymbol{\theta}}_{xy}^*; x, y) - R(\widehat{\boldsymbol{\theta}}_{xy}; x, y) = R^*(\widehat{\boldsymbol{\theta}}_{xy}^*; x, y) - R^*(\widehat{\boldsymbol{\theta}}_{xy}; x, y) + o_p(nh^{2r}),$$

we have that

$$\frac{1}{n} R(\widehat{\boldsymbol{\theta}}_{xy}; x, y) = \frac{1}{n} R(\widehat{\boldsymbol{\theta}}_{xy}^*; x, y) + o_p(h^{2r}).$$

Since $\frac{\partial R(\boldsymbol{\theta};x,y)}{\partial \boldsymbol{\theta}} = 0$ at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{xy}$, the expression above implies that

$$\begin{aligned} & h^{-2r}(\widehat{\boldsymbol{\theta}}_{xy} - \widehat{\boldsymbol{\theta}}^*_{xy})^\tau \tilde{R}(\widehat{\boldsymbol{\theta}}_{xy})(\widehat{\boldsymbol{\theta}}_{xy} - \widehat{\boldsymbol{\theta}}^*_{xy}) \\ = & \left(\frac{\widehat{\theta}_{xy,1} - \widehat{\theta}^*_{xy,1}}{h^r}, \frac{\widehat{\theta}_{xy,2} - \widehat{\theta}^*_{xy,2}}{h^{r-1}}, \ldots, \frac{\widehat{\theta}_{xy,r} - \widehat{\theta}^*_{xy,r}}{h} \right) \\ \times & R^* \begin{pmatrix} \frac{\widehat{\theta}_{xy,1} - \widehat{\theta}^*_{xy,1}}{h^r} \\ \frac{\widehat{\theta}_{xy,2} - \widehat{\theta}^*_{xy,2}}{h^{r-1}} \\ \vdots \\ \frac{\widehat{\theta}_{xy,r} - \widehat{\theta}^*_{xy,r}}{h} \end{pmatrix} \xrightarrow{P} 0, \end{aligned}$$

where $\tilde{R}(\boldsymbol{\theta}) = \frac{1}{2n} \frac{\partial^2 R(\boldsymbol{\theta};x,y)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\tau}$ and

$$R^* = \operatorname{diag}(1, h^{-1}, \ldots, h^{-(r-1)})\, R(\widehat{\boldsymbol{\theta}}_{xy}) \operatorname{diag}(1, h^{-1}, \ldots, h^{-(r-1)}).$$

It can be proved that $R^* \xrightarrow{P} f(x)F(y|x)\{1 - F(y|x)\}S$. Note that $S = (\kappa_{i+j-2})$ is a positive-definite matrix, and we have that

$$\widehat{\theta}_{xy,i} = \widehat{\theta}^*_{xy,i} + o_p(h^{r-i+1})$$

for $i = 1, \ldots, r$. Now (10.41) follows from the fact that

$$\widehat{F}(y|x) = \exp(\widehat{\theta}_{xy,1})/\{1 + \exp(\widehat{\theta}_{xy,1})\}.$$

We have completed the proof. ∎

10.6 Additional Bibliographical Notes

Chatfield (2001) is a specific monograph on time series forecasting. Clements and Hendry (1998) focuses on forecasting in economics. Yao and Tong (1994a) appeared to be the first to address the features of nonlinear prediction presented in §10.1.1–§10.1.3 in a systematic manner.

The literature on nonparametric (and nonlinear) point prediction includes Chen (1996), Matzner-Løber, Gannoun and De Gooijer (1998), and De Gooijer and Zerom (2000) on kernel methods, Sugihara and May (1990) and Jensen (1993) on nearest-neighbor methods, and Faraway and Chatfield (1998) and Zhang, Patuwo, and Hu (1998) on neural networks methods (see also Weigend and Gershenfeld 1994).

Interval prediction based on conditional quantiles was studied in Yao and Tong (1996), De Gooijer, Gannoun and Zerom (2001), and Cai (2002). Nonparametric estimation for conditional quantiles (for independent data) was treated in Sheather and Marron (1990), Fan, Hu, and Truong (1994),

and Yu and Jones (1998). Hyndman (1995, 1996) and De Gooijer and Gannoun (2000) dealt with minimum-length predictive sets. Constrained estimation for conditional density functions was considered in Hyndman and Yao (2002). Hall and Yao (2002) proposed a nonparametric estimation for an optimum approximation of the conditional distribution function of a scalar Y given a vector $\mathbf{X}$ via dimension reduction. Cai, Yao, and Zhang (2001) dealt with nonparametric and semiparametric estimation for conditional distributions of discrete-valued time series.

References

Abramson, I.S. (1982). On bandwidth variation in kernel estimates–a square root law. *The Annals of Statistics*, **10**, 1217–1223.

Adak, S. (1998). Time-dependent spectral analysis of nonstationary time series. *Journal of the American Statistical Association*, **93**, 1488-1501.

Adams, T.M. and Nobel, A.B. (1998). On density estimation from ergodic processes. *Annals of Probability*, **26**, 794–804.

Aerts, M., Claeskens, G., and Hart, J.D. (2000). Testing lack of fit in multiple regression. *Biometrika*, **87**, 405–424.

Ahmad, I.A. (1979). Strong consistency of density estimation by orthogonal series methods for dependent variables with applications. *Annals of the Institute of Statistical Mathematics*, **31**, 279–288.

Ahmad, I.A. (1982). Integrated mean square properties of density estimation by orthogonal series methods for dependent variables. *Annals of the Institute of Statistical Mathematics*, **34**, 339–350.

Aït-Sahalia, Y. (1996). Nonparametric pricing of interest rate derivative securities. *Econometrica*, **64**, 527–560.

Aït-Sahalia, Y. (1999). Transition densities for interest rate and other nonlinear diffusions. *Journal of Finance*, **54**, 1361–1395.

Aït-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica*, **70**, 223–262.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, **21**, 243–247.

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, **22**, 203–217.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium in Information Theory* (B.N. Petroc and F. Caski, eds.). Akademiai Kiado, Budapest, pp. 276–281.

Akaike, H. (1977). On the entropy maximisation principle. In *Applications of Statistics* (P.R. Krishnaiah, ed.). North-Holland, Amsterdam, pp. 27–41.

Akaike, H. and Kitagawa, G. (1999). *The Practice of Time Series Analysis*. Springer-Verlag, New York.

Allen, D.M. (1974). The relationship between variable and data augmentation and a method of prediction. *Technometrics*, **16**, 125–127.

Altman, N.S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association*, **85**, 749–759.

An, H.Z. and Chen, S.G. (1997). A note on the ergodicity of the non-linear autoregressive model. *Statistics and Probability Letters*, **34**, 365–372.

An, H.Z. and Huang, F.C. (1996). The geometrical ergodicity of nonlinear autoregressive models. *Statistica Sinica*, **6**, 943–956.

Andersen, T.G. and Lund, J. (1997). Estimating continuous time stochastic volatility models of the short term interest rate. *Journal of Econometrics*, **77**, 343–77.

Anderson, T.W. (1971). *The Statistical Analysis of Time Series*. John Wiley, New York.

Anderson, T.W. (1993). Goodness of fit tests for spectral distributions. *The Annals of Statistics*, **21**, 830-847.

Andrews, D. (1984). Nonstrong mixing autoregressive processes. *Journal of Applied Probability*, **21**, 930–934.

Ansley, C.F. and Kohn, R. (1994). Convergence of the backfitting algorithm for additive models. *Journal of the Australian Mathematical Society Series A*, **57**, 316–329.

Antoniadis, A. and Fan, J. (2001). Regularized wavelet approximations (with discussion). *Journal of the American Statistical Association*, **96**, 939–967.

Arfi, M. (1995). Non-parametric drift estimation from ergodic samples. *Journal of Nonparametric Statistics*, **5**, 381–389.

Arfi, M. (1998). Non-parametric variance estimation from ergodic samples. *Scandinavian Journal of Statistics*, **25**, 225–234.

Azzalini, A. and Bowman, A.N. (1993). On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society, Series B*, **55**, 549–557.

Azzalini, A., Bowman, A.N., and Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika*, **76**, 1–11.

Baillie, R.T., Bollerslev, T., and Mikkelsen, H.O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **74**, 3–30.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, **113**, 301–413.

Bartlett, M.S. (1946). On theoretical specification of sampling properties of auto-correlated time series. *Journal of the Royal Statistical Society, Series B*, **8**, 27–41.

Bartlett, M.S. (1948). Smoothing periodograms from time series with continuous spectra. *Nature*, **161**, 686–687.

Bartlett, M.S. (1950). Periodogram analysis and continuous spectra. *Biometrika*, **37**, 1–16.

Bartlett, M.S. (1954). A note on some multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society, Series B*, **16**, 296–298.

Bartlett, M.S. (1963). The spectral analysis of point processes. *Journal of the Royal Statistical Society, Series B*, **25**, 264–296.

Basawa, I.V. and Prakasa Rao, B.L.S. (1980). *Statistical Inference for Stochastic Processes*. Academic Press, New York.

Basrak, B., Davis, R.A. and Mikosch, T. (2002). Regular variation of GARCH processes. *Stochastic Processes and Their Applications*, **99**, 95–115.

Bellman, R.E. (1961). *Adaptive Control Processes*. Princeton University Press, Princeton.

Beltrão, K.I. and Bloomfield, P. (1987). Determining the bandwidth of a kernel spectrum estimate. *Journal of Time Series Analysis*, **8**, 21–36.

Bera, A.K. and Higgins, M.L. (1992). A test for conditional heteroskedasticity in time series models. *Journal of Time Series Analysis*, **13**, 501–519.

Beran, J. (1995). *Statistics for Long-Memory Processes*. Chapman and Hall, New York.

Beran, J. and Feng, Y. (2001). Local polynomial fitting with long-memory, short-memory and antipersistent errors. *Annals of the Institute of Statistical Mathematics*, **54**, 291–311.

Bernstein, S.N. (1926). Sur l'estension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Annals of Mathematics*, **97**, 1–59.

Berry, D. (1993). Testing for additivity of a regression function. *The Annals of Statistics*, **21**, 235–254.

Bhaskara Rao, M., Subba Rao, T., and Walker, A.M. (1983). On the existence of some bilinear time series models. *Journal of Time Series Analysis*, **4**, 95–110.

Bhattacharya, R.N. and Lee, C. (1995). Ergodicity of nonlinear first order autoregressive models. *Journal of Theoretical Probability*, **8**, 207-219.

Bickel, P.J. (1975). One-step Huber estimates in linear models. *Journal of the American Statistical Association*, **70**, 428–433.

Bickel, P.J. and Doksum, K.A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, San Francisco.

Bickel, P., Götze, F., and van Zwet, W.R. (1995). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statistica Sinica*, **7**, 1–31.

Bickel, P.J., Klaassen, A.J., Ritov, Y., and Wellner, J.A. (1993). *Efficient and Adaptive Inference in Semi-parametric Models*. Johns Hopkins University Press, Baltimore.

Bickel, P.J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: Sharp order of convergence estimates. *Sankhyā, Series A*, **50**, 381–393.

Bickel, P.J. and Ritov, Y. (1992). Testing for goodness of fit: A new approach. *In Nonparametric Statistics and Related Topics* (A.K.Md.E. Saleh, ed.). Elsevier Science Publishers B.V., Amsterdam, pp. 51–57.

Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviation of density function estimates. *The Annals of Statistics*, **1**, 1071–1095.

Billingsley, P. (1961). *Statistical Inference for Markov Processes*. Holt, New York.

Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density. *The Annals of Statistics*, **23**, 11–29.

Black, F., Derman, E., and Toy, E. (1990). A one-factor model of interest rates and its application to treasury bond options. *Financial Analysts' Journal*, **46**, 33–39.

Black, F., and Karasinski, P. (1991). Bond and option pricing when short rates are lognormal. *Financial Analysts' Journal*, **47**, 52–59.

Bloomfield, P. (1976). *Fourier Analysis of Time Series: An Introduction*, John Wiley, New York.

Blum, J.R., Kiefer, J., and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *Annals of Mathematical Statistics*, **32**, 485–498.

Blyth, S. (1994). Local divergence and association. *Biometrika*, **81**, 579–584.

Bochner, S. (1936). Summation of multiple Fourier series by spherical mean. *Transactions of the American Mathematical Society*, **40**, 175–207.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*, **31**, 307–327.

Bollerslev, T., Engle, R.F., and Nelson, D.B. (1994). ARCH models in finance. In *Handbook of Econometrics* (R.F. Engle and D.L. McFadden, ed.), Vol. IV, Chapter 49. Elsevier Sciences B.V., Amsterdam.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction* (2nd ed.). Lecture Notes in Statistics **110**. Springer-Verlag, Berlin.

Bougerol, P. and Picard, N. (1992a). Strict stationarity of generalized autoregressive processes. *Annals of Probability*, **4**, 1714–1730.

Bougerol, P. and Picard, N. (1992b). Stationarity of GARCH processes and some nonnegative time series. *Journal of Econometrics*, **52**, 115–127.

Bowman, A.W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, Oxford.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26**, 211–252.

Box, G.E.P. and Jenkins, G.M. (1970). *Time Series Analysis, Forecasting, and Control*. Holden-Day, San Francisco.

Box, G.E.P. and Pierce, D.A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, **65**, 1509–1526.

Bradley, R.C. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics* (E. Eberlein and M.S. Taqqu, ed.). Birkhauser, Boston, pp. 165–192.

Bradley, R.C. and Tran, L.T. (1999). Density estimation for nonisotropic random fields. *Journal of Statistical Planning and Inference*, **81**, 51–70.

Brandt, A. (1986). The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Advances in Applied Probability*, **18**, 211–220.

Breidt, F.J. and Davis, R.A. (1992). Time reversibility, identifiability and independence of innovations for stationary time series. *Journal of Time Series Analysis*, **13**, 377–390.

Breiman, L. (1993). Fitting additive models to regression data: Diagnostics and alternative views. *Computational Statistics & Data Analysis*, **15**, 13–46.

Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, **80**, 580–619.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1993). *CART: Classification and Regression Trees* (1st ed., 1984). Wadsworth, Belmont, CA.

Breiman, L., Meisel, W., and Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics*, **19**, 135–144.

Brillinger, D.R. (1969). Asymptotic properties of spectral estimates of second order. *Biometrika*, **56**, 375–390.

Brillinger, D.R. (1974). The asymptotic distribution of the Whittaker periodogram and a related chi-squared statistic for stationary processes. *Biometrika*, **61**, 419–422.

Brillinger, D.R. (1981). *Time Series Analysis: Data Analysis and Theory* (2nd ed.). Holt, Rinehart & Winston, New York.

Brillinger, D.R. (1991). Some history of the study of higher-order moments and spectra. *Statistica Sinica*, **1**, 465–476.

Brillinger, D.R. (1996). Some uses of cumulants in wavelet analysis. *Journal of Nonparametric Statistics*, **6**, 93–114.

Brillinger, D.R. and Rosenblatt, M. (1967). Asymptotic theory of estimates of k-th order spectra. In *Spectral Analysis Time Series* (Proc. Advanced Sem., Madison, WI., 1966). Wiley, New York, pp. 153–188.

Brillinger, D.R. and Segundo, P. (1979). Empirical examination of the threshold model of neuron firing. *Biological Cybernetics*, **35**, 213–220.

Brockmann, M., Gasser, T., and Herrmann, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. *Journal of the American Statistical Association*, **88**, 1302–1309.

Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*. (2nd ed.). Springer-Verlag, New York.

Brockwell, P.J. and Davis, R.A. (1996). *Introduction to Time Series and Forecasting*. Springer-Verlag, New York.

Brown, L.D. and Low, M. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *The Annals of Statistics*, **24**, 2524–2535.

Brumback, B. and Rice, J.A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association*, **93**, 961–994.

Bühlmann, P. (1976). Locally adaptive lag-window spectral estimation. *Journal of Time Series Analysis*, **17**, 247–270.

Buja, A., Hastie, T.J., and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *The Annals of Statistics*, **17**, 453–555.

Cai, Z. (2002). Regression quantiles for time series. *Econometric Theory*, **18**, 169–192.

Cai, Z. and Fan, J. (2000). Average regression surface for dependent data. *Journal of Multivariate Analysis*, **75**, 112–142.

Cai, Z., Fan, J., and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, **95**, 888–902.

Cai, Z., Fan, J., and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, **95**, 941–956.

Cai, Z. and Roussas, G.G. (1992). Uniform strong estimation under α-mixing, with rates. *Statistics and Probability Letters*, **15**, 47–55.

Cai, Z., Yao, Q., and Zhang, W. (2001). Smoothing for discrete-valued time series. *Journal of the Royal Statistical Society, Series B*, **63**, 357–375.

Carbon, M., Hallin, M., and Tran, L.T. (1996). Kernel density estimation for random fields: The L_1 theory. *Journal of Nonparametric Statistics*, **6**, 157–170.

Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, **92**, 477–489.

Carroll, R.J. and Li, K.C. (1995). Binary regressors in dimension reduction models: A new look at treatment comparisons. *Statistica Sinica*, **5**, 667–688.

Carroll, R.J., Ruppert, D., and Welsh, A.H. (1998). Nonparametric estimation via local estimating equations. *Journal of the American Statistical Association*, **93**, 214–227.

Chambers, J.M. and Hastie, T.J. (1991). *Statistical Models*. Wadsworth/ Brooks Cole, Pacific Grove, CA.

Chan, K.C., Karolyi, A.G., Longstaff, F.A., and Sanders, A.B. (1992). An empirical comparison of alternative models of the short-term interest rate. Journal of Finance, **47**, 1209–1227.

Chan, K.S. (1990a). Deterministic stability, stochastic stability and ergodicity. An appendix in *Nonlinear Time Series* by H. Tong. Oxford University Press, Oxford, pp. 448–466.

Chan, K.S. (1990b). Testing for threshold autoregression. *The Annals of Statistics*, **18**, 1886–1894.

Chan, K.S. (1991). Percentage point of likelihood ratio tests for threshold autoregression. *Journal of the Royal Statistical Society, Series B*, **53**, 691–696.

Chan, K.S. (1993a). Consistency and limiting distribution of a least squares estimator of a threshold autoregressive model. *The Annals of Statistics*, **21**, 520–533.

Chan, K.S. (1993b). A review of some limit theorems of Markov chains and their applications. In *Dimension Estimation and Models* (H. Tong, ed.). World Scientific, Singapore, pp. 108–135.

Chan, K.S., Petruccelli, J.D., Tong, H., and Woolford, S.W. (1985). A multiple-threshold AR(1) model. *Journal of Applied Probability*, **22**, 267–279.

Chan, K.S. and Tong, H. (1985). On the use of the deterministic Lyapounov function for the ergodicity of stochastic difference equations. *Advances in Applied Probability*, **17**, 666–678.

Chan, K.S. and Tong, H. (1990). On likelihood ratio tests for threshold autoregression. *Journal of the Royal Statistical Society, Series B*, **52**, 469–476.

Chan, K.S. and Tong, H. (1994). A note on noisy chaos. *Journal of the Royal Statistical Society, Series B*, **56**, 301–311.

Chan, K.S. and Tong, H. (2001). *Chaos: A Statistical Perspective*. Springer, New York.

Chan, K.S. and Tsay, R.S. (1998). Limiting properties of the least squares estimator of a continuous threshold autoregressive model. *Biometrika*, **85**, 413–426.

Chapman, D.A. and Pearson, N.D. (2000). Is the short rate drift actually nonlinear? Journal of Finance, **55**, 355–388.

Chaudhuri, P., Doksum, K., and Samarov, A. (1997). On average derivative quantile regression. *The Annals of Statistics*, **25**, 715–744.

Chatfield, C. (1996). *The Analysis of Time Series: An Introduction* (5th ed.). Chapman and Hall, London.

Chatfield, C. (2001). *Time-Series Forecasting*. Chapman and Hall/CRC, Boca Raton.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *The Annals of Statistics*, **16**, 136–146.

Chen, M. and An, H.Z. (1997). A Kolmogorov–Smirnov type test for conditional heteroskedasticity in time series. *Statistics and Probability Letters*, **33**, 321–331.

Chen, R. (1996). A nonparametric multi-step prediction estimator in Markovian structure. *Statistica Sinica*, **6**, 603–615.

Chen, R., Liu, J.S., and Tsay, R.S. (1995). Additivity tests for nonlinear autoregressions. *Biometrika*, **82**, 369–383.

Chen, R. and Tsay, R.S. (1991). On the ergodicity of TAR(1) processes. *Annals of Applied Probability*, **1**, 613–634.

Chen, R. and Tsay, R.S. (1993). Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, **88**, 298–308.

Chen, S.X., Härdle, W., and Li, M. (2002). An empirical likelihood goodness-of-fit test for time series. *Journal of the Royal Statistical Society, Series B*, to appear.

Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, **66**, 289–314.

Chen, X. and White, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, **45**, 682–691.

Chen, Z.G., Dahlhaus, R. and Wu, K.H. (2000). Hidden frequency estimation with data taper. *Journal of Time Series Analysis*, **21**, 113–142.

Cheng, B. and Robinson, P.M. (1991). Density estimation in strongly dependent nonlinear time series. *Statistica Sinica*, **1**, 335–359.

Chiaromonte, F., Cook, R.D., and Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *The Annals of Statistics*, **30**, 475–497.

Choi, B.S. (1992). *ARMA Model Identification*. Springer-Verlag, New York.

Choi, E., Hall, P., and Rousson, V. (2000). Data sharpening methods for bias reduction in nonparametric regression. *The Annals of Statistics*, **28**, 1339–1355.

Chow, Y.S. and Teicher, H. (1997). *Probability Theory* (3rd ed.). Springer-Verlag, New York.

Chu, C.K.(1995). Bandwidth selection in nonparametric regression with general errors. *Journal of Statistical Planning and Inference*, **44**, 265–275.

Chu, C.K. and Marron, J.S. (1991a). Comparison of two bandwidth selectors with dependent errors. *The Annals of Statistics*, **19**, 1906–1918.

Chu, C.K. and Marron, J.S. (1991b). Choosing a kernel regression estimator (with discussions). *Statistical Science*, **6**, 404–436.

Claeskens, G. and Hall, P. (2002). Effect of dependence on stochastic measures of accuracy of density estimators. *The Annals of Statistics*, **30**, 431–454.

Clements, M.P. and Hendry, D.F. (1998). *Forecasting Economic Time Series*. Cambridge University Press, Cambridge.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Cleveland, W.S., Grosse, E., and Shyu, W.M. (1991). Local regression models. In *Statistical Models in S* (J.M. Chambers and T.J. Hastie, ed.). Wadsworth & Brooks, Pacific Grove. CA. pp. 309–376.

Cohen, J.E., Kesten, H., and Newman, C.H. (1986). Random matrices and their applications. *Contemporary Mathematics*, **50**.

Cook, R.D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983–992.

Cook, R.D. (1998). Principal Hessian directions revisited (with discussion). *Journal of the American Statistical Association*, **93**, 84–100.

Cook, R.D. and Lee, H. (1999). Dimension reduction in regressions with a binary response. *Journal of the American Statistical Association*, **94**, 1187–1200.

Cook, R.D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *The Annals of Statistics*, **30**, 455–474.

Cox, D.D. (1984). Multivariate smoothing spline functions. *SIAM Journal on Numerical Analysis*, **21**, 789–813.

Cox, D.D. and Kim, T.Y. (1995). Moment bounds for mixing random variables useful in nonparametric function estimation. *Stochastic Processes and Their Applications*, **56**, 151–158.

Cox, D.R. (1981). Statistical analysis of time series: Some recent developments (with discussion). *Scandinavian Journal of Statistics*, **8**, 93–115.

Cox, J.C., Ingersoll, J.E., and Ross, S.A. (1985). A theory of the term structure of interest rates. *Econometrica*, **53**, 385–467.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, **31**, 377–403.

Cryer, J.D. (1986). *Time Series Analysis*. PWS-Kent Publishing Co., Boston.

Csörgö, S. and Mielniczuk, J. (1995). Nonparametric regression under long-range dependent normal errors. *The Annals of Statistics*, **23**, 1000–1014.

Cuzick, J. (1992). Semiparametric additive regression. *Journal of the Royal Statistical Society, Series B*, **54**, 831–843.

Dahlhaus, R. (1984). Parametric estimation of stationary process with spectra containing strong peaks. *Robust and Nonlinear Time Series Analysis*. Lecture Notes in Statistics, **26**. Springer, New York, pp. 50-86.

Dahlhaus, R. (1990a). Small sample effects in time series analysis: a new asymptotic theory and a new estimate. *The Annals of Statistics*, **16**, 808–841.

Dahlhaus, R. (1990b). Nonparametric high resolution spectral estimation. *Probability Theory and Related Fields*, **85**, 147–180.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics*, **25**, 1–37.

Daniels, H.E. (1946). Discussion of paper by M.S. Bartlett. *Journal of the Royal Statistical Society, Series B*, **24**, 185–198.

Davis, H.T. and Jones, R.H. (1968). Estimation of the innovation variance of a stationary time series. *Journal of the American Statistical Association*, **63**, 141–149.

Davis, K.B. (1975). Mean square error properties of density estimates. *The Annals of Statistics*, **3**, 1025–1030.

Davis, N., Pembertin., J., and Petruccelli, J.D. (1988). An automatic procedure for identification, estimation and forecasting univariate self-exciting threshold autoregressive models. *The Statistician*, **37**, 119–204.

Davis, R.A., Knight, K., and Liu, J. (1992). M-estimation for autoregression with infinite variances. *Stochastic Processes and Their Applications*, **40**, 145–180.

Davis, R.A. and Mikosch, T. (1998). The sample autocorrelations of heavy-tailed processes with applications to ARCH. *Applications of Statistics*, **26**, 2049–2080.

Davydov, Y.A. (1973). Mixing conditions for Markov chains. *Theory of Probability and Its Applications*, **18**, 313–328.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag, New York.

De Gooijer, J.G. and Gannoun, A. (2000). Nonparametric conditional predictive regions for time series. *Computational Statistics & Data Analysis*, **33**, 259–275.

De Gooijer, J.G., Gannoun, A. and Zerom, D. (2001). Multi-stage kernel-based conditional quantile prediction in time series. Preprint.

De Gooijer, J.G. and Zerom, D. (2000). Kernel-based multiple-ahead prediction of the U.S. short-term interest rate. *Journal of Forecasting*, **19**, 335–353.

Denker, M. and Keller, G. (1983). On U-statistics and v. Mises' statistics for weakly dependent processes. *Zeitschriftfuer Wahrscheinlichkeitstheorie verw. Gebiete*, **64**, 505–522.

Deo, R. (2000). Spectral tests of the martingale hypothesis under conditional heteroscedasticity. *Journal of Econometrics*, **99**, 291–315.

Deo, R. and Chen, W.(2000a). A generalized portmanteau goodness-of-fit test for time series models. Unpublished Manuscript.

Deo, R. and Chen, W. (2000b). Power transformations to induce normality and their applications. Unpublished Manuscript.

Devroye, L.P. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. Wiley, New York.

Diebold, F.X., Hickman, A., Inoue, A. and Schuermann, T. (1998). Converting 1-day volatility to h-day volatility: Scaling by $\sqrt{h}$ is worse than you think. *Risk*, **11**, 104–107.

Diggle, P. J. (1990). *Time Series. A Biostatistical Introduction*. Oxford University Press, Oxford.

Ding, Z. and Granger, C.W.J. (1996). Modeling volatility persistence of speculative returns: A new approach. *Journal of Econometrics*, **73**, 185–215.

Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, **90**, 1200–1224.

Donoho, D.L. and Johnstone, I.M. (1996). Neo-classical minimax problems, thresholding, and adaptive function estimation. *Bernoulli*, **2**, 39–62.

Donoho, D.L. and Johnstone, I.M. (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, **26**, 879–921.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B*, **57**, 301–369.

Donoho, D.L. and Liu, R.C. (1991a). Geometrizing rate of convergence II. *The Annals of Statistics*, **19**, 633–667.

Donoho, D.L. and Liu, R.C. (1991b). Geometrizing rate of convergence III. *The Annals of Statistics*, **19**, 668–701.

Doob, J.L. (1953). *Stochastic Processes*. Wiley, New York.

Doukhan, P. (1994). *Mixing*. Springer-Verlag, New York.

Doukhan, P. and Louhichi, S. (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and Their Applications*, **84**, 313–342.

Duan, N. and Li, K.-C. (1991). Slicing regression: A link-free regression method. *The Annals of Statistics*, **19**, 505–530.

Duffie, D. (1996). *Dynamic Asset Pricing Theory* (2nd ed.). Princeton University Press, Princeton, N.J..

Durbin, J. and Koopman, S.J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

Durham, G.B. and Gallant, A.R. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion Processes. *Journal of Business and Economic Statistics*, **20**, 297–316.

Dzhaparidze, K. (1986). *Parameter Estimation and Hypothesis Testing on Spectral Analysis of Stationary Time Series*. Springer-Verlag, New York.

Efromovich, S. (1985). Nonparametric estimation of a density with unknown smoothness. *Theory of Probability and Its Applications*, **30**, 557–568.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. Springer-Verlag, New York.

Efromovich, S.Y. and Pinsker, M.S. (1982). Estimation of square-integrable probability density of a random variable. *Problems of Information Transmission*, **18**, 175–189.

Efron, B. and Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *The Annals of Statistics*, **24**, 2431–2461.

Elton, C. and Nicholson, M. (1942). The ten-year cycle in numbers of the lynx in Canada. *Journal of Animal Ecology*, **11**, 215–244.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events*. Springer-Verlag, New York.

Engle, R.F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica*, **50**, 987–1008.

Engle, R.F. and Granger, C.W.J. (1987). Cointegration and error correction: Representation, estimation and testing. *Econometrica*, **55**, 251–276.

Engle, R.F. and Bollerslev, T. (1986). Modelling the persistence of conditional variances. *Econometric Reviews*, **5**, 1-50.

Epanechnikov, V.A. (1969). Nonparametric estimation of a multidimensional probability density. *Theory of Probability and Its Applications*, **13**, 153–158.

Eubank, R.L. (1999). *Spline Smoothing and Nonparametric Regression* (2nd ed.). Marcel Dekker, New York.

Eubank, R.L. and Hart, J.D. (1992). Testing goodness-of-fit in regression via order selection criteria. *The Annals of Statistics*, **20**, 1412–1425.

Eubank, R.L., Hart, J.D., Simpson, D.G., and Stefanski, L.A. (1995). Testing for additivity in nonparametric regression. *The Annals of Statistics*, **23**, 1896–1920.

Eubank, R.L. and LaRiccia, V.N. (1992). Asymptotic comparison of Cramér–von Mises and nonparametric function estimation techniques for testing goodness-of-fit. *The Annals of Statistics*, **20**, 2071–2086.

Eubank, R.L. and Speckman, P.L. (1991). A bias reduction theorem with applications in nonparametric regression. *Scandinavian Journal of Statistics*, **18**, 211–222.

Ezekiel, A. (1924). A method for handling curvilinear correlation for any number of variables. *Journal of the American Statistical Association*, **19**, 431–453.

Fama, E. (1965). The behaviour of stock market prices. *Journal of Business*, **38**, 34–105.

Fan, J. (1991). On the estimation of quadratic functionals. *The Annals of Statistics*, **19**, 1273–1294.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998–1004.

Fan, J. (1993a). Local linear regression smoothers and their minimax efficiency. *The Annals of Statistics*, **21**, 196–216.

Fan, J. (1993b). Adaptively local one-dimensional subproblems with application to a deconvolution problem. *The Annals of Statistics*, **21**, 600–610.

Fan, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *Journal of the American Statistical Association*, **91**, 674–688.

Fan, J. and Chen, J. (1999). One-step local quasi-likelihood estimation. *Journal of the Royal Statistical Society, Series B*, **61**, 303–322.

Fan, J., Farmen, M. and Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society, Series B*, **60**, 591–608.

Fan, J., Gasser, T., Gijbels, I., Brockmann, M., and Engel, J. (1996). Local polynomial fitting: Optimal kernel and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics*, **49**, 79–99.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics*, **20**, 2008–2036.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B*, **57**, 371–394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

Fan, J. and Gijbels, I. (2000). Local polynomial fitting. In Smoothing and Regression: Approaches, Computation and Application (M.G. Schimek, ed.). John Wiley & Sons, New York, pp. 229-276.

Fan, J. and Gu, J. (2001). Semiparametric estimation of value-at-risk. Manuscript.

Fan, J., Härdle, W., and Mammen, E. (1998). Direct estimation of additive and linear components for high-dimensional data. *The Annals of Statistics*, **26**, 943–971.

Fan, J., Heckman, N.E. and Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, **90**, 141–150.

Fan, J., Hu, T.-C., and Truong, Y.K. (1994). Robust nonparametric function estimation. *Scandinavian Journal of Statistics*, **21**, 433–446.

Fan, J. and Huang, L. (2001). Goodness-of-fit test for parametric regression models. *Journal of the American Statistical Association*, **96**, 640–652.

Fan, J., Jiang, J., Zhang, C., and Zhou, Z. (2003). Time-dependent diffusion models for term structure dynamics and the stock price volatility. *Statistica Sinica*, to appear.

Fan, J. and Kreutzberger, E. (1998). Automatic local smoothing for spectral density estimation. *Scandinavian Journal of Statistics*, **25**, 359–369.

Fan, J. and Li, R. (2001). Variable selection via penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

Fan, J. and Lin, S.K. (1996). Test of Significance when data are curves. *Journal of the American Statistical Association*, **93**, 1007-1021.

Fan, J. and Masry, E. (1992). Multivariate regression estimation with errors-in-variables: Asymptotic normality for mixing processes. *Journal of Multivariate Analysis*, **43** 1992–?.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**, 645–660.

Fan, J., Yao, Q., and Cai, Z. (2002). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society, Series B*, to appear.

Fan, J., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, **83**, 189–206.

Fan, J. and Zhang, C. (2003). A re-examination of diffusion estimations with applications to financial model validation. *Journal of the American Statistical Association*, to appear.

Fan, J., Zhang, C., and Zhang, J. (2001). Generalized likelihood test statistic and Wilks phenomenon. *The Annals of Statistics*, **29**, 153–193.

Fan, J. and Zhang, J.T. (2000). Functional linear models for longitudinal data. *Journal of the Royal Statistical Society, Series B*, **62**, 303–322.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, **27**, 1491–1518.

Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, **27**, 715–731.

Fan, J. and Zhang, W. (2002). Generalized likelihood ratio tests for spectral density. Manuscript.

Fan, Y. and Li, Q. (1996). Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica*, **64**, 865–890.

Faraway, J. and Chatfield, C. (1998). Time series forecasting with neural networks: A comparative study using the airline data. *Applications of Statistics*, **47**, 231–250.

Farrell, R.H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Annals of Mathematical Statistics*, **43**, 170–180.

Fejèr, L. (1900). Sur les foctions bornèes et integrables. *Comptes Rendus de l' Academie des Sciences Paris*, **131**, 984–987.

Fejèr, L. (1904). Untersuchungen über Fouriersche Reihen. *Annals of Mathematics*, **58**, 501–569.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Vol. 2* (2nd ed.). John Wiley & Sons, New York.

Finkenstädt, B. (1995). *Nonlinear Dynamics in Economics*. Springer, Berlin.

Fix, E. and Hodges, J.L. (1951). Discriminatory analysis—nonparametric discrimination: Consistency properties. Report No. 4, Project no. 21-29-004. USAF School of Aviation Medicine, Randolph Field, TX.

Florens-Zmirou, D. (1993). On estimating the diffusion coefficient from discrete observations. *Journal of Applied Probability*, **30**, 790–804.

Franke, J. and Härdle, W. (1992). On bootstrapping kernel spectral estimates. *The Annals of Statistics*, **20**, 121–145.

Franke, J., Kreiss, J.P., and Mammen, E. (2002). Bootstrap of kernel smoothing in nonlinear time series. *Bernoulli*, **8**, 1–37.

Friedman, J.H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, **19**, 1–142.

Friedman, J.H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, **76**, 817–823.

Fu, W.J. (1998). Penalized regression: The bridge versus the LASSO, *Journal of Computational and Graphical Statistics*, **7**, 397–416.

Fuller, W.A. (1976). *Introduction to Statistical Time Series*, John Wiley, New York.

Gallant, A.R., Hsieh, D.A., and Tauchen, G.E. (1991) On fitting a recalcitrant series: The pound/dollar exchange rate, 1974-1983. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics* (W.A. Barnett, J. Powell, and G.E. Tauchen, ed.). Cambridge University Press, Cambridge, pp. 199–240.

Gallant, A.R. and Tauchen, G. (1997). Estimation of continuous time models for stock returns and interest rates. *Macroeconomic Dynamics*, **1**, 135–168.

Gao, J., Tong, H., and Wolff, R. (2002). Model specification tests in nonparametric stochastic regression models. *Journal of Multivariate Analysis*, **83**, 324-359.

Gardner, E.S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting* **4**, 1–28.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, Lecture Notes in Mathematics **757**. Springer-Verlag, New York, pp. 23–68.

Gasser, T., Müller, H.-G., and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society, Series B*, **47**, 238–252.

Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625–633.

Genon-Catalot, V. and Jacod, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Annales de l' Institut Henri Poincare Probabilite'et Statistique*, **29**, 119–151.

Ghaddar, D.K. and Tong, H. (1981). Data transformation and self-exciting threshold autoregression. *Journal of the Royal Statistical Society, Series C*, **30**, 238–248.

Gijbels, I., Hall, P. and Kneip, A. (1999). On the estimation of jump points in smooth curves. *Annals of the Institute of Statistical Mathematics*, **51**, 231–251.

Gijbels, I., Pope, A., and Wand, M.P. (1999). Understanding exponential smoothing via kernel regression. *Journal of the Royal Statistical Society, Series B*, **61**, 39–50.

Giraitis, L., Kokoszka, P., and Leipus, R. (2000). Stationary ARCH models: Dependence structure and central limit theorem. *Econometric Theory*, **16**, 3–22.

Giraitis, L. and Robinson, P.M. (2001). Whittle estimation of ARCH models. *Econometric Theory*, **17**, 608–623.

Glad, I.K. (1998). Parametrically guided non-parametric regression. *Scandinavian Journal of Statistics*, **25**, 649–668.

Goldie, C.M. (1991). Implicit renewal theory and tails of solutions of random equations. *Annals of Applied Probability*, **1**, 126–166.

Golyandina, N., Nekrutkin, V., and Zhigljavsky, A. (2001). *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman and Hall/CRC, Boca Raton.

Good, I.J. and Gaskins, R.A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, **58**, 255–277.

Gouriéroux, C. (1997). *ARCH Models and Financial Applications*. Springer-Verlag, New York.

Gozalo, P. and Linton, O. (2000). Local nonlinear least squares: using parametric information in nonparametric regression. *Journal of Econometrics*, **99**, 63–106.

Granger, C.W.J. and Anderson, A.P. (1978a). *An Introduction to Bilinear Models*. Van derhoeck & Ruprecht, Gottingen.

Granger, C.W.J. and Anderson, A.P. (1978b). On the invertibility of time series models. *Stochastic Processes and Their Applications*, **8**, 87–92.

Granger, C.W.J. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, **1**, 15–29.

Granger, C.W.J. and Terasvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford University Press, New York.

Granovsky, B.L. and Müller, H.-G. (1991). Optimizing kernel methods: A unifying variational principle. *International Statistical Review*, **59**, 373–388.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, London.

Grenander, U. (1951). On empirical spectral analysis of stochastic processes. *Arkiv Matematik.*, **1**, 503–531.

Grenander, U. and Rosenblatt, M. (1952). On spectral analysis of stationary time series. *Proceedings of the National Academy of Sciences USA*, **38**, 519–521.

Gruet, M.-A. (1996). A nonparametric calibration analysis. *The Annals of Statistics*, **24**, 1474–1492.

Gu, C. (1990). Adaptive spline smoothing in non-Gaussian regression models. *Journal of the American Statistical Association.*, **85**, 801–807.

Guegan, D. and Pham, D.T. (1989). A note on the strong consistency of the least squares estimates for the diagonal bilinear time series model. *Scandinavian Journal of Statistics*, **6**, 129–136.

Guegan, D. and Pham, D.T. (1992). Power of the score test against bilinear time series models. *Statistica Sinica*, **2**, 157–169.

Guo, M. and Petruccelli, J.D. (1991). On the null recurrence and transience of a first-order SETAR model. *Journal of Applied Probability*, **28**, 584–592.

Györfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series*. Lecture Notes in Statistics **60**. Springer-Verlag, Berlin.

Györfi, L. and Lugosi, G. (1992). Kernel density estimation from an ergodic sample is not universally consistent. *Computational Statistics & Data Analysis*, **14**, 437–442.

Györfi, L., Lugosi, G., and Morvai, G. (1999). A simple randomized algorithm for sequential prediction of ergodic time series. *IEEE Transactions on Information Theory*, **45**, 2642–2650.

Györfi, L. and Masry, E. (1990). The L_1 and L_2 strong consistency of recursive kernel density estimation from dependent samples. *IEEE Transactions on Information Theory*, **36**, 531–539.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Boston.

Härdle, W. and Hall, P. (1993). On the backfitting algorithm for additive regression models. *Statistica Neerlandica*, **47**, 43–57.

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *The Annals of Statistics*, **21**, 157–178.

Härdle, W., Hildenbrand, W., and Jerison, M. (1991). Empirical evidence on the law of demand. *Econometrica*, **59**, 1525–1549.

Härdle, W., Liang, H., and Gao, J. (2000). *Partially Linear Models*. Physica-Verlag, Heidelberg.

Härdle, W., Lütkepohl, H. and Chen, R. (1997). A review of nonparametric time series analysis. *International Statistical Review*, **65**, 49–72.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, **21**, 1926–1947.

Härdle, W., Sperlich, S., and Spokoiny, V. (2001). Structural tests in additive regression. *Journal of the American Statistical Association*, **96**, 1333–1347.

Härdle, W. and Stoker, T.M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, **84**, 986–995.

Härdle, W. and Tsybakov, A.B. (1995). Additive nonparametric regression on principal components. *Journal of Nonparametric Statistics*, **5**, 157–184.

Härdle, W. and Tsybakov, A.B. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics*, **81**, 223–242.

Härdle, W., Tsybakov, A., and Yang, L. (1998). Nonparametric vector autoregression. *Journal of Statistical Planning and Inference*, **68**, 221–245.

Härdle, W. and Vieu, P. (1992). Kernel regression smoothing of time series. *Journal of Time Series Analysis*, **13**, 209–232.

Haan, L. de, Resnick, S.I., Rootzen, H., and Vries, C.G. de (1989). Extremal behaviour of solutions to a stochastic difference equation with applications to ARCH processes. *Stochastic Processes and Their Applications*, **32**, 213–224.

Haggan, V. and Ozaki, T. (1981). Modeling nonlinear vibrations using an amplitude-dependent autoregressive time series model. *Biometrika*, **68**, 189–196.

Hall, P. (1978). Representations and limit theorems for extreme value distributions. *Journal of Applied Probability*, **15**, 639–644.

Hall, P. (1990). On the bias of variable bandwidth curve estimators. *Biometrika*, **77**, 529–535.

Hall, P. and Carroll, R.J. (1989). Variance function estimation in regression: The effect of estimation of the mean. *Journal of the Royal Statistical Society, Series B*, **51**, 3–14.

Hall, P. and Hart, J.D. (1990). Nonparametric regression with long-range dependence. *Stochastic Processes and Their Applications*, **36**, 339–351.

Hall, P. and Heyde, C.C. (1980). *Martingale Limit Theory and its Applications*. Academic Press, New York.

Hall, P. and Johnstone, I. (1992). Empirical functional and efficient smoothing parameter selection (with discussion). *Journal of the Royal Statistical Society, Series B*, **54**, 475–530.

Hall, P., Kay, J.W., and Titterington, D.M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, **77**, 521–528.

Hall, P., Lahiri, S.N., and Truong, Y.K. (1995). On bandwidth choice for density estimation with dependent data. *The Annals of Statistics*, **23**, 2241–2263.

Hall, P. and Li, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics*, **21**, 867–889.

Hall, P. and Marron, J.S. (1988). Variable window width kernel estimates of probability densities. *Probability Theory and Related Fields*, **80**, 37–49.

Hall, P., Marron, J.S., and Park, B.U. (1992). Smoothed cross-validation. *Probability Theory and Related Fields*, **92**, 1–20.

Hall, P., Peng, L., and Yao, Q. (2002). Prediction and nonparametric estimation for time series with heavy tails. *Journal of Time Series Analysis*, **23**, 313–331.

Hall, P. and Presnell, B. (1999). Intentionally biased bootstrap methods. *Journal of the Royal Statistical Society, Series B*, **61**, 143–158.

Hall, P. and Wehrly, T.E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Journal of the American Statistical Association*, **86**, 665–672.

Hall, P., Wolff, R.C.L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, **94**, 154-163.

Hall, P. and Yao, Q. (2002). Estimating conditional distribution functions using dimension reduction. A preprint.

Hall, P. and Yao, Q. (2003). Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica*, to appear.

Hamilton, J.D. (1989). A new approach to economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357–384.

Hamilton, J.D. (1994). *Time Series Analysis*. Princeton University Press, Princeton, NJ.

Hannan, E.J. (1970). *Multiple Time Series*. John Wiley, New York.

Hannan, E.J. (1973). The asymptotic theory of linear time-series models. *Journal of Applied Probability*, **10**, 130–145.

Hannan, E.J. (1980). The estimation of the order of an ARMA process. *The Annals of Statistics*, **8**, 1071–1081.

Hannan, E.J. (1982). A note on bilinear time series models. *Stochastic Processes and Their Applications*, **12**, 21–24.

Hannan, E.J. (1986). Remembrance of things past. In *The Craft of Probability Modeling* (J. Gani, ed.). Springer-Verlag, New York.

Hannan, E.J. and Deistler, M. (1988). *The Statistical Theory of Linear Systems*. Wiley, New York.

Hansen, B.E. (1999). Threshold effects in non-dynamic panels: estimation, testing, and inference. *Journal of Econometrics*, **93**, 345–368.

Hart, J.D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society, Series B*, **53**, 173–187.

Hart, J.D. (1996). Some automated methods of smoothing time-dependent data. *Journal of Nonparametric Statistics*, **6**, 115–142.

Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer, New York.

Hart, J.D. and Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *The Annals of Statistics*, **18**, 873–890.

Hart, J.D. and Wehrly, T.E. (1986). Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association*, **81**, 1080–1088.

Harvey, A. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

Harvey, A.C., Ruiz, E., and Shephard, N. (1994). Multivariate stochastic variance models. *Review of Economic Studies*, **61**, 247–264.

Hasminskii, R.Z. (1978). A lower bound on the risks of nonparametric estimates densities in the uniform metric. *Theory of Probability and Its Applications*, **23**, 794–798.

Hastie, T.J. and Loader, C. (1993). Local regression: Automatic kernel carpentry (with discussion). *Statistical Science*, **8**, 120–143.

Hastie, T.J. and Tibshirani, R. (1986). Generalized additive models (with discussion). *Statistical Science*, **1**, 297–318.

Hastie, T.J. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hastie, T.J. and Tibshirani, R.J. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, **55**, 757–796.

Hastie, T.J., Tibshirani, R. and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, **89**, 1255–1270.

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica*, **66**, 1017–1098.

Heckman, N. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society, Series A*, **48**, 244–248.

Hinich, M.J. (1982). Testing for Gaussianity and linearity of a stationary time series. *Journal of Time Series Analysis*, **3**, 169–176.

Hjellvik, V., Yao, Q., and Tjøstheim, D. (1998). Linearity testing using local polynomial approximation. *Journal of Statistical Planning and Inference*, **68**, 295–321.

Hjort, N.L. and Glad, I.K. (1995). Nonparametric density estimation with a parametric start. *The Annals of Statistics*, **23**, 882–904.

Hjort, N.L. and Jones, M.C. (1996a). Locally parametric nonparametric density estimation. *The Annals of Statistics*, **24**, 1619–1647.

Hjort, N.L. and Jones, M.C. (1996b). Better rules of thumb for choosing bandwidth in density estimation. Statistical Research Report, Department of Mathematics, University of Oslo, Oslo, Norway.

Hong, P.Y. (1991). The autocorrelation structure for the GARCH-M process. *Economic Letters*, **37**, 129–132.

Hong, Y. (1997). One-sided testing for conditional heteroskedasticity in time series models. *Journal of Time Series Analysis*, **18**, 253–277.

Hong, Y. and Lee, T.-H. (2002). Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics*, to appear.

Hoover, D.R., Rice, J.A., Wu, C.O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809–822.

Horowitz, J.L. and Spokoiny, G.G. (2001). An adaptive, rate-optimal test of a parametric model against a nonparametric alternative. *Econometrica*, **69**, 599–631.

Hosking, J.R.M. (1981). Fractional differencing. *Biometrika*, **68**, 165–176.

Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny, V. (2002). Structure adaptive approach for dimension reduction. *The Annals of Statistics*, to appear.

Hsing, T. and Carroll, R.J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, **20**, 1040–1061.

Huang, J. (1999). Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics*, Vol 26, 242–272.

Hull, J. (1997). *Options, Futures, and Other Derivatives* (3rd ed.). Prentice-Hall, Upper Saddle River, NJ.

Hull, J. and White, A. (1990). Pricing interest rate derivative securities. *Review of Financial Studies*, **3**, 573–592.

Hunsberger, S. (1994). Semiparametric regression in likelihood-based models. *Journal of the American Statistical Association*, **89**, 1354–1365.

Hurvich, C.M. and Beltrão, K.I. (1990). Cross-validatory choice of a spectral estimate and its connections with AIC. *Journal of Time Series Analysis*, **11**, 121–137.

Hurvich, C.M. and Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Hyndman, R.J. (1995). Highest density forecast regions for non-linear and non-normal time series models. *Journal of Forecasting*, **14**, 431–441.

Hyndman, R.J. (1996). Computing and graphing highest density regions. *The American Statistician*, **50**, 120–126.

Hyndman, R.J. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *Journal of Nonparametric Statistics*, **14**, 259–278.

Ibragimov, I.A. and Hasminskii, R.Z. (1984). On nonparametric estimation of a linear functional in a Gaussian white noise. *Theory of Probability and Its Applications*, **29**, 19–32.

Ibragimov, I.A. and Linnik, Y.V. (1971). *Independent and Stationary Sequences of Random Variables*. Walters-Noordhoff, Gröningen.

Ichimura, H. (1993). Semiparametric least-squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, **58**, 71–120.

Inglot, T., Kallenberg, W.C.M., and Ledwina, T. (1994). Power approximations to and power comparison of smooth goodness-of-fit tests. *Scandinavian Journal of Statistics*, **21**, 131–145.

Inglot, T. and Ledwina, T. (1996). Asymptotic optimality of data-driven Neyman's tests for uniformity. *The Annals of Statistics*, **24**, 1982–2019.

Ingster, Yu.I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives I–III. *Math. Methods Statist.*, **2**, 85–114; **3**, 171–189; **4**, 249–268.

Izenman, A.J. (1983). J.R. Wolf and H.A. Wolfer: An historical note on Zurich sunspot relative numbers. *Journal of the Royal Statistical Society, Series A*, **146**, 311–318.

Jensen, J.L. (1993). Comments on non-parametric predictions of sunspot numbers. *The Astronomical Journal*, **105**, 350–352.

Jiang, G.J. and Knight, J.L. (1997). A nonparametric approach to the estimation of diffusion processes, with an application to a short-term interest rate model. *Econometric Theory*, **13**, 615–645.

Jiang, J. and Mack, M.P. (2001). Robust local polynomial regression for dependent data. *Statistica Sinica*, **11**, 705–722.

Johnstone, I.M. and Silverman, B.W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B*, **59**, 319–351.

Jones, M.C. (1997). A variation on local linear regression. *Statistica Sinica*, **7**, 1171–1180.

Jones, M.C., Marron, J.S., and Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, **91**, 401–407.

Jones, R.H. (1975). Fitting autoregressions. *Journal of the American Statistical Association*, **70**, 590–592.

Jorion, P. (2000). Value at Risk: The New Benchmark for Managing Financial Risk (2nd ed.). McGraw–Hill, New York.

Kallenberg, W.C.M. and Ledwina, T. (1997). Data-driven smooth tests when the hypothesis is composite. *Journal of the American Statistical Association*, **92**, 1094–1104.

Kashyap, R.L. (1977). A Bayesian comparison of different classes of dynamic models using empirical data. *IEEE Transactions on Automatic Control*, **22**, 715–727.

Kato, T. and Masry, E. (1999). On the spectral density of the wavelet transform of fractional Brownian motion. *Journal of Time Series Analysis*, **20**, 559–563.

Kesten, H. (1973). Random difference equations and renewal theory for products of random matrices. *Acta Mathematica*, **131**, 207–248.

Kim, J.H. and Hart, J.D. (1998). Tests for change in a mean function when the data are dependent. *Journal of Time Series Analysis*, **19**, 399–424.

Kim, T.Y. and Cox, D.D. (1995). Asymptotic behaviors of some measures of accuracy in nonparametric curve estimation with dependent observations. *Journal of Multivariate Analysis*, **53**, 67–93.

Kim, T.Y. and Cox, D.D. (1997). A study on bandwidth selection in density estimation under dependence. *Journal of Multivariate Analysis*, **62**, 190–203.

Kim, W., Linton, O.B., and Hengartner, N.W. (1999). A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, **8**, 278–297.

Kim, W.K., Billard, L., and Basawa, I.V. (1990). Estimation of first order diagonal bilinear time series model. *Journal of Time Series Analysis*, **11**, 215–230.

Kimeldorf, G.S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, **41**, 495–502.

Kitagawa, G. and Akaike, H. (1978). A procedure for the modeling of non-stationary time series. *Annals of the Institute of Statistical Mathematics*, **30**, 351-363.

Kitagawa, G. and Gersch, W. (1996). *Smoothness Priors Analysis of Time Series*. Springer-Verlag, New York.

Kohn, R., Ansley, C.F., and Wong, C.M. (1992). Nonparametric spline regression with autoregressive moving average errors. *Biometrika*, **79**, 335–346.

Kokoszka, P. and Leipus, R. (2000). Change-point estimation in ARCH models. *Bernoulli*, **6**, 513–539.

Kooperberg, C. and Stone, C.J. (1991). A study of logspline density estimation. *Computational Statistics & Data Analysis*, **12**, 327–347.

Kooperberg, C., Stone, C.J., and Truong, Y.K. (1995a). Logspline estimation of a possibly mixed spectral distribution. *Journal of Time Series Analysis*, **16**, 359–388.

Kooperberg, C., Stone, C.J., and Truong, Y.K. (1995b). Rate of convergence for logspline spectral density estimation. *Journal of Time Series Analysis*, **16**, 389–401.

Koopmans, L.H. (1974). *The Spectral Analysis of Time Series*. Academic Press, New York.

Koul, H. and Stute, W. (1999). Nonparametric model checks for time series. *Applications of Statistics*, **27**, 204–236.

Kuchibhatla, M. and Hart, J.D. (1996). Smoothing-based lack-of-fit tests: Variations on a theme. *Journal of Nonparametric Statistics*, **7**, 1–22.

Laïb, N. (2002). Nonparametric test for conditional variance functions in time series. Preprint.

Lawrance, A.J. and Lewis, P.A.W. (1980). The exponential autoregressive moving average EARAM(p, q) process. *Journal of the Royal Statistical Society, Series B*, **42**, 150–161.

Lawrance, A.J. and Lewis, P.A.W. (1985). Modelling and residual analysis of non-linear autoregressive time series in exponential variables (with discussion). *Journal of the Royal Statistical Society, Series B*, **47**, 165–202.

LeBaron, B. (1997). Technical trading rule and regime shifts in foreign exchange. In *Advances in Trading Rules* (E. Acar and S. Satchell, eds.). Butterworths-Heinemann, London.

LeBaron, B. (1999). Technical trading rule profitability and foreign exchange intervention. *Journal of International Economics*, **49**, 125–143.

Lee, A.W. and Hansen, B.E. (1994). Asymptotic theory for a GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory*, **10**, 29–52.

Lee, J.H.H. (1991). A Lagrange multiplier test for GARCH models. *Economic Letters*, **37**, 265–271.

Lee, J.H.H. and King, M.L. (1993). A locally most mean powerful based score test for ARCH and GARCH regression disturbances. *Journal of Business and Economic Statistics*, **11**, 17–27.

Lehmann, E.L. (1983). *Theory of Point Estimation*. John Wiley & Sons, New York.

Lepski, O.V. (1991) Asymptotically minimax adaptive estimation I. *Theory of Probability and Its Applications*, **36**, 4, 682–697.

Lepski, O.V. (1992) Asymptotically minimax adaptive estimation II. *Theory of Probability and Its Applications*, **37**, 3, 433–448.

Lepski, O.V. and Spokoiny, V.G. (1999). Minimax nonparametric hypothesis testing: The case of an inhomogeneous alternative. *Bernoulli*, **5**, 333–358.

Lewis, P.A.W. and Stevens, J.G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). *Journal of the American Statistical Association*, **87**, 864–877.

Li, C.W. and Li, W.K. (1996). On a double threshold autoregressive heteroscedastic time series model. *Journal of Applied Econometrics*, **11**, 253–274.

Li, K.-C. (1985). From Stein's unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics*, **13**, 1352–1377.

Li, K.-C. (1986). Asymptotic optimality for C_L and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, **14**, 1101–1112.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316–342.

Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, **87**, 1025–1039.

Li, K.C., Lue, H.H. and Chen, C. H. (2000). Interactive tree-structured regression via principal Hessian directions. *Journal of the American Statistical Association*, **95**, 547–560.

Li, W.K. (1992). On the asymptotic standard errors of residual autocorrelations in nonlinear time series modelling. *Biometrika*, **79**, 435–437.

Li, W.K. and Lam, K. (1995). Modelling asymmetry in stock returns by a threshold ARCH model. *The Statistician*, **44**, 333–341.

Liang, H., Hrdle, W. and Carroll, R. (1999). Estimation in a semiparametric partially linear errors-in-variables model. *The Annals of Statistics*, **27**, 1519–1535.

Lii, K.S. (1978). A global measure of a spline density estimate. *The Annals of Statistics*, **6**, 1138–1148.

Lii, K.S. and Masry, E. (1995). On the selection of random sampling schemes for the spectral estimation of continuous time processes. *Journal of Time Series Analysis*, **16**, 291–311.

Lii, K.S. and Rosenblatt, M. (1990). Asymptotic normality of cumulant spectral estimates. *Journal of Theoretical Probability*, **3**, 367–385.

Lii, K.S. and Rosenblatt, M. (1998). Line spectral analysis for harmonizable processes. *Proceedings of the National Academy of Sciences USA*, **95**, 4800–4803.

Linton, O. (1993). Adaptive estimation in ARCH models. *Econometric Theory*, **9**, 539–569.

Linton, O. and Nielsen, J.P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, **82**, 93–100.

Linton, O.B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika*, **84**, 469–473.

Linton, O.B. (2000). Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory*, **16**, 502–523.

Linton, O.B., Chen, R., Wang, N., and Härdle, W. (1997). An analysis of transformations for additive nonparametric regression. *Journal of the American Statistical Association*, **92**, 1512–1521.

Linton, O.B. and Härdle, W. (1996). Estimation of additive regression models with known links. *Biometrika*, **83**, 529–540.

Liu, J. (1990). A note on causality and invertibility of a general bilinear time series model. *Advances in Applied Probability* **22**, 247–250.

Liu, J. (1992). On the stationary and asymptotic inference of bilinear time series models. *Statistica Sinica*, **2**, 479–494.

Liu, J. and Brockwell, P.J. (1982). On the general bilinear time series. *Journal of Time Series Analysis*, **10**, 33–40.

Liu, J. and Chen, Z.G. (1991). Bilinear$(p, 0, 1, 1)$ model and its consistent identification. Preprint.

Ljung, G.M. and Box, G.E.P. (1978). On a measure of lack of fit in time series models. *Biometrika*, **65**, 297–303.

Loader, C.R. (1996). Local likelihood density estimation. *The Annals of Statistics*, **24**, 1602–1618.

Low, M.G. (1993). Lower bounds for the integrated risk in nonparametric density and regression estimation. *The Annals of Statistics*, **21**, 577–589.

Lumsdaine, R. (1996). Consistency and asymptotic normality of the quasi-maximum likelihood estimator for IGARCH(1,1) and covariance stationary GARCH(1,1) models. *Econometrica*, **16**, 575–596.

Lütkepohl, H. (1993). *Introduction to Multiple Time Series*. Springer-Verlag, New York.

Macaulay, F.R. (1931). *The Smoothing of Time Series*. National Bureau of Economic Research, New York.

Mack, Y.P. and Silverman, B.W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Zeitschriftfuer Wahrscheinlichkeitstheorie verw. Gebiete*, **61**, 405–415.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

Mammen, E. (1992). *When Does Bootstrap Work? Asymptotic Results and Simulations*. Springer, New York.

Mammen, E., Linton, O., Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics*, **27**, 1443–1490.

Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business*, **36**, 394–419.

Marron, J.S. and Nolan, D. (1988). Canonical kernels for density estimation. *Statistics and Probability Letters*, **7**, 195–199.

Masry, E. (1983). Probability density estimation from sampled data. *IEEE Transactions on Information Theory*, **29**, 696–709.

Masry, E. (1993). Asymptotic normality for deconvolution estimators of multivariate densities of stationary processes. *Journal of Multivariate Analysis*, **44**, 47–68.

Masry, E. and Cambanis, S. (1984). Spectral density estimation for stationary stable processes. *Stochastic Processes and Their Applications*, **18**, 1–31.

Masry, E. and Fan, J. (1997). Local polynomial estimation of regression functions for mixing processes. *Scandinavian Journal of Statistics*, **24**, 165–179.

Masry, E. and Tjøstheim, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series. *Econometric Theory*, **11**, 258–289.

Matzner-Løber, E., Gannoun, A. and De Gooijer, J.G. (1998). Nonparametric forecasting: a comparison of three kernel-based methods. *Communications in Statistics – Theory and Methods*, **27**, 1593–1617.

Mays, J.E., Birch, J.B., and Starnes, B.A. (2000). Model robust regression: combining parametric, nonparametric and semiparametric methods. *Journal of Nonparametric Statistics*, **13**, 245–277.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

McLeod, A.L. and Li, W.K. (1983). Diagnostic checking of ARMA time series models using squared residual autoregressions. *Journal of Time Series Analysis*, **4**, 269–273.

Mélard, G. and Roy, R. (1988). Modèles de series chronologiques avec seuils. *Revues de Statistiques Appliqués*, **36**, 5–24.

Melino, A. and Turnbull, M.S. (1990). Pricing foreign currency options with stochastic volatility. *Journal of Econometrics*, **54**, 239–265.

Messer, K. (1991). A comparison of a spline estimate to an equivalent kernel estimate. *The Annals of Statistics*, **19**, 817–829.

Meyn, S.P. and Tweedie, R.L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, London.

Meyn, S.P. and Tweedie, R.L. (1994). State-dependent criteria for convergence Markov chains. *Annals of Applied Probability*, **4**, 149–168.

Mikosch, T. and Stărică, C. (1999). Long range dependence effects and ARCH modelling. *Technical report*.

Mikosch, T. and Straumann, D. (2000). Whittle estimation in a heavy-tailed GARCH(1,1) model. *Technical report*.

Milhoj, A. (1985). The moment structure of ARCH processes. *Scandinavian Journal of Statistics*, **12**, 281–292.

Mittnik, S. and Rachev, S.T. (2000). *Stable Paretian Models in Finance*. Wiley, New York.

Mittnik, S., Rachev, S.T., and Paolella, M.S. (1998). Stable Paretian modeling in finance: Some empirical and theoretical aspects. In *A Practical Guide to Heavy Tails* (R.J. Adler, R.E. Feldman, and M.S. Taqqu, ed.). Birkhäuser, Boston, pp. 79–110.

Moran, P.A.P. (1953). The statistical analysis of the Canadian lynx cycle, I: Structure and prediction. *Australian Journal of Zoology*, **1**, 163–173.

Morgan, J.P. (1996). RiskMetrics Technical Document (4th ed.). J.P. Morgan, New York.

Müller, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association*, **82**, 231–238.

Müller, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Lecture Notes in Statistics **46**. Springer-Verlag, Berlin.

Müller, H.-G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika*, **78**, 521–530.

Müller, H.-G. (1992). Change-points in nonparametric regression analysis. *The Annals of Statistics*, **20**, 737–761.

Müller, H.-G. (1993). On the boundary kernel method for non-parametric curve estimation near endpoints. *Scandinavian Journal of Statistics*, **20**, 313–328.

Müller, H.G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *The Annals of Statistics* **15**, 610–625.

Müller, H.G. and Stadtmüller U. (1993). On variance function estimation with quadratic forms. *Journal of Statistical Planning and Inference*, **35**, 213–231.

Murphy, S.A. and van der Vaart, A.W. (2000). On profile likelihood (with discussion). *Journal of the American Statistical Association*, **95**, 449–485.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and Its Applications*, **9**, 141–142.

Nadaraya, E.A. (1989). *Nonparametric Estimation of Probability Densities and Regression Curves*. (S. Kotz, trans.). Kluwer, Dordrecht.

Needham, J. (1959). *Science and Civilisation in China*, Vol. III. Cambridge University Press, Cambridge.

Nelson, D.B. (1990). Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory*, **6**, 318–334.

Nelson, D.B. (1991). Conditional heteroscedasticity in asset pricing: A new approach. *Econometrica*, **59**, 347–370.

Nelson, D.B. and Cao, C.Q. (1992). Inequality constraints in the univariate GARCH models. *Journal of Business and Economic Statistics*, **10**, 229–235.

Neumann, M.H. (1994). Fully data-driven nonparametric variance estimators. *Statistics*, **25**, 189–212.

Neumann, M.H. and von Sachs, R. (1997). Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *The Annals of Statistics*, **25**, 38–76.

Newey, W.K. and Stoker, T.M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica*, **61**, 1199–1223.

Neyman, J. (1937). Smooth test for goodness of fit. *Skandinavisk Aktuarietidskrift*, **20**, 149–199.

Nicholls, D.F. and Quinn, B.G. (1982). *Random Coefficient Autoregressive Models: An Introduction*. Springer-Verlag, New York.

Nielsen, J.P. and Linton, O.B. (1998). An optimization interpretation of integration and back-fitting estimators for separable nonparametric models. *Journal of the Royal Statistical Society, Series B*, **60**, 217–222.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press, Cambridge.

Nummelin, E. and Tuominen, P. (1982). Geometric ergodicity of Harris recurrent Markov chains. *Stochastic Processes and Their Applications*, **3**, 187–202.

Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *The Annals of Statistics*, **13**, 984–997.

Nychka, D. (1988). Bayesian 'confidence' intervals for smoothing splines. *Journal of the American Statistical Association*, **83**, 1134–1143.

Ogden, T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser, Boston.

Opsomer, J.D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, **73**, 166–179.

Opsomer, J.D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics*, **25**, 186–211.

Opsomer, J.D., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, **16**, 134–153.

Ozaki, T. (1982). The statistical analysis of perturbed limit cycle processes using nonlinear time series models. *Journal of Time Series Analysis*, **3**, 29–41.

Ozaki, T. (1985). Non-linear time series models and dynamical systems. *Handbook of Statistics*. Vol. 5 (E.J. Hannan, P.R. Krishnaiah, and M.M. Rao, eds.). North-Holland, Amsterdam.

Ozaki, T. and Tong, H. (1975). On the fitting of non-stationary autoregressive model analysis. *Proceedings of the 8th Hawaii International Conference on System Sciences*, pp. 224–226.

Paparoditis, E. (2000). Spectral density based goodness-of-fit tests in time series models. *Scandinavian Journal of Statistics*, **27**, 143–176.

Paparoditis, E. (2002). Frequency domain bootstrap for time series. In *Empirical Process Techniques for Dependent Data* (H. Dehling, M. Sorensen, and T. Mikosch, eds.). Birkhauser: Boston, p. 365-381.

Park, B.U., Cho, S., and Kang, K.H. (1994). An automatic spectral density estimate. *Journal of the Korean Statistical Society*, **23**, 79–88.

Parzen, E. (1961). Mathematical considerations in the estimation of spectra. *Technometrics*, **3**, 167–190.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, **33**, 1065–1076.

Pawitan, Y. and O'Sullivan, F. (1994). Nonparametric spectral density estimation using penalized Whittle likelihood. *Journal of the American Statistical Association*, **89**, 600–610.

Peligrad, M. (1986). Recent advances in the central limit theorems and its weak invariance principle for mixing sequences of random variables (a survey). In *Dependence in Probability and Statistics* (E. Eberlein and M.S. Taqqu, eds.). Birkhauser, Boston, pp. 193-223.

Pemberton, J. (1985). Contributions to the theory of non-linear time series models. Ph.D. Thesis, University of Manchester, Manchester, U.K.

Peng, L. and Yao, Q. (2000). Nonparametric regression under infinite variance dependent errors. Preprint.

Peng, L. and Yao, Q. (2002). Least absolute deviations estimation for ARCH and GARCH models. Preprint.

Petruccelli, J.D. and Woolford, S.W. (1984). A threshold AR(1) model. *Journal of Applied Probability*, **21**, 207–286.

Pham, D.T. (1981). Nonparametric estimation of the drift coefficient in the diffusion equation. *Math. Operationsforsch. Statist. Ser. Statist.*, **12**, 61–73.

Pham, D.T. (1985). Bilinear Markovian representation and bilinear model. *Stochastic Processes and Their Applications*, **20**, 295–306.

Pham, D.T. (1986). The mixing properties of bilinear and generalized random coefficients autoregressive models. *Stochastic Processes and Their Applications*, **23**, 291–300.

Pham, D.T. (1993). Bilinear time series models. In *Dimension Estimation and Models* (H.Tong, ed.). World Scientific, Singapore.

Pham, T.D. and Tran, L.T. (1985). Some mixing properties of time series models. *Stochastic Processes and Their Applications*, **19**, 279–303.

Pham, T.D. and Tran, L.T. (1991). Kernel density estimation under a locally mixing condition. *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed.). pp. 419–430

Pinsker, M.S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems of Information Transmission*, **16**, 52–68.

Polonik, W. and Yao, Q. (2000). Conditional minimum volume predictive regions for stochastic processes. *Journal of the American Statistical Association*, **95**, 509–519.

Polonik, W. and Yao, Q. (2002). Set-indexed conditional empirical and quantile processes based on dependent data. *Journal of Multivariate Analysis*, **80**, 234–255.

Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation*. Academic Press, New York.

Prakasa Rao, B.L.S. (1985). Estimation of the drift for a diffusion process. *Statistics*, **16**, 263–275.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1992). *Numerical Recipes in C* (2nd ed.). Cambridge University Press, Cambridge.

Press, W.H. and Tukey, J.W. (1956). Power Spectral Methods of Analysis and Their Application to Problems in Airplane Dynamics. Bell Telephone System Monograph 2606. Bell telephone System.

Priestley, M.B. (1981). *Spectral Analysis and Time Series*, Vols. 1 and 2. Academic Press, New York.

Priestley, M.B. (1988). *Nonlinear and Non-stationary Time Series Analysis*. Academic Press, London.

Priestley, M.B. and Chao, M.T. (1972). Nonparametric function fitting. *Journal of the Royal Statistical Society, Series B*, **34**, 384–392.

Quinn, B.G. (1982). A note on the existence of a strictly stationary solution to bilinear equations. *Journal of Time Series Analysis*, **3**, 249–252.

Ramsay, J.O. and Silverman, B.W. (1997). *The Analysis of Functional Data*. Springer-Verlag, Berlin.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. (2nd ed.). John Wiley & Sons, New York.

Ray, B.K. and Tsay, R.S. (1997). Bandwidth selection for kernel regression with long-range dependence. *Biometrika*, **84**, 791–802.

Reinsch, C. (1967). Smoothing by spline functions. *Numerische Mathematik*, **10**, 177–183.

Reinsel, G.C. (1997). *Elements of Multivariate Time Series Analysis*. Springer-Verlag, New York.

Rice, J. (1984a). Bandwidth choice for nonparametric kernel regression. *The Annals of Statistics*, **12**, 1215–1230.

Rice, J.A. (1984b). Boundary modification for nonparametric regression. *Communications in Statistics – Theory and Methods*, **13**, 893–900.

Rice, J.A. and Rosenblatt, M. (1981). Integrated mean squared error of a smoothing spline. *Journal of Approximation Theory*, **33**, 353–365.

Rissanen, J. (1980). Consistent order estimates of autoregressive processes by shortest description of data. In *Analysis and Optimization of Stochastic Systems* (O.L.R. Jacobs et.al.,eds.). Academic Press, New York, pp. 451–461.

Robinson, P.M. (1983). Nonparametric estimators for time series. *Journal of Time Series Analysis*, **4**, 185–207.

Robinson, P.M. (1987). Time series residuals with application to probability density estimation. *Journal of Time Series Analysis*, **8**, 329–344.

Robinson, P.M. (1988). The stochastic difference between econometrics and statistics. *Econometrica*, **56**, 531–547.

Robinson, P.M. (1991a). Automatic frequency domain inference on semiparametric and nonparametric models. *Econometrica*, **59**, 1329–1363.

Robinson, P.M. (1991b). Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *Journal of Econometrics*, **47**, 67–84.

Robinson, P.M. (1994). Rates of convergence and optimal spectral bandwidth for long range dependence. *Probability Theory and Related Fields*, **99**, 443–473.

Robinson, P.M. (1997). Large-sample inference for nonparametric regression with dependent errors. *The Annals of Statistics*, **25**, 2054–2083.

Robinson, P.M. (1998). Inference-without-smoothing in the presence of nonparametric autocorrelation. *Econometrica*, **66**, 1163–1182.

Robinson, P.M. and Zaffaroni, P. (1998). Nonlinear time series with long memory: a model for stochastic volatility. *Journal of Statistical Planning and Inference*, **68**, 359–371.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832–837.

Rosenblatt, M. (1970). Density estimates and Markov sequences. In *Nonparametric Techniques in Statistical Inference* (M.L. Puri, ed.). Cambridge University Press, London, pp. 199–213.

Rosenblatt, M. (1991). *Stochastic Curve Estimation*. NSF-CBMS Regional Conference Series in Probability and Statistics, **3**. Institute of Mathematical Statistics, California.

Roussas, G. (1967). Nonparametric estimation in Markov processes. *Annals of the Institute of Statistical Mathematics*, **21**, 73–87.

Roussas, G.G. (1969). Nonparametric estimation of the transition distribution function of a Markov process. *Annals of Mathematical Statistics*, **40**, 1386–1400.

Roussas, G.G. (1990). Nonparametric regression estimation under mixing conditions. *Stochastic Processes and Their Applications*, **36**, 107–116.

Roussas, G.G. (1995). Asymptotic normality of a smooth estimate of a random field distribution function under association. *Statistics and Probability Letters*, **24**, 77–90.

Roussas, G.G. and Tran, L.T. (1992). Asymptotic normality of the recursive kernel regression estimate under dependence conditions. *The Annals of Statistics*, **20**, 98–120.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, **92**, 1049–1062.

Ruppert, D., Sheather, S.J., and Wand, M.P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257–1270.

Ruppert, D. and Wand, M.P. (1994). Multivariate weighted least squares regression. *The Annals of Statistics*, **22**, 1346–1370.

Ruppert, D., Wand, M.P., Holst, U., and Hössjer, O. (1997). Local polynomial variance function estimation. *Technometrics*, **39**, 262–273.

Rydberg, T. (2000). Realistic statistical modelling of financial data. *International Statistical Review*, **68**, 233–258.

Saikkonen, P. and Luukkonen, R. (1991). Power properties of a time series linear test against some simple bilinear alternatives. *Statistica Sinica*, **1**, 453–464.

Samarov, A.M. (1993). Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association*, **88**, 836–847.

Schmidt, G., Mattern, R., and Schüler, F. (1981). Biomechanical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column with and without protective helmet under effects of impact. EEC Research Program on Biomechanics of Impacts. Final Report Phase III, Project 65, Institut für Rechtsmedizin, Universität Heidelberg, Heidelberg, Germany.

Schoenberg, I.J. (1964). Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences USA*, **52**, 947–950.

Schott, J.R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, **89**, 141–148.

Schucany, W.R. and Sommers, J.P. (1977). Improvement of kernel type density estimators. *Journal of the American Statistical Association*, **72**, 420–423.

Schuster, E.F. (1985). Incorporating support constraints into nonparametric estimates of densities. *Communications in Statistics – Theory and Methods*, **14**, 1123–1126.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.

Severini, T.A. and Staniswalis, J.G. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association*, **89**, 501–511.

Severini, T.A. and Wong, W.H. (1992). Generalized profile likelihood and conditional parametric models. *The Annals of Statistics*, **20**, 1768–1802.

Sesay, S.A.O. and Subba Rao, T. (1992). Frequency domain estimation of a bilinear time series model. *Journal of Time Series Analysis*, **13**, 521–545.

Shao, Q. and Yu, H. (1996). Weak convergence for weighted empirical processes of dependent sequences. *Annals of Probability*, **24**, 2098–2127.

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, **53**, 683–690.

Sheather, S.J. and Marron, J.S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, **85**, 410–416.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In *Time Series Models in Econometrics, Finance and Other Fields* (D.R. Cox, D.V. Hinkley, and O.E. Barndorff-Nielsen, eds.). Chapman and Hall, London. pp. 1–67.

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *The Annals of Statistics*, **8**, 147–164.

Shiryayev, A.N. (1984). *Probability*. Springer-Verlag, New York.

Shumway, R.H. (1988). *Applied Statistical Time Series Analysis*. Prentice–Hall, Englewood Cliffs, NJ.

Shumway, R.H. and Stoffer, D.S. (2000). *Time Series Analysis and Its Applications*. Springer-Verlag, New York.

Silverman, B.W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, Series B*, **43**, 97–99.

Silverman, B.W. (1984). Spline smoothing: The equivalent variable kernel method. *The Annals of Statistics*, **12**, 898–916.

Silverman, B.W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society, Series B*, **47**, 1–52.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, **24**, 1–24.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer, New York.

Simonoff, J.S. and Tsai, C.L. (1999). Semiparametric and additive model selection using an improved Akaike information criterion. *Journal of Computational and Graphical Statistics*, **8**, 22–40.

Skaug, H.J. and Tjøstheim, D. (1993). A nonparametric test of serial independence based on the empirical distribution function. *Biometrika*, **80**, 591–602.

Smith, M., Wong, C.M., and Kohn, R. (1998). Additive nonparametric regression with autocorrelated errors. *Journal of the Royal Statistical Society, Series B*, **60**, 311–331.

Smith, P.L. (1982). Curve fitting and modeling with splines using statistical variable selection methods, NASA Report 166034, NASA Langley Research Center, Hampton, VA.

Speckman, P. (1981). The asymptotic integrated mean square error for smoothing noisy data by splines. Unpublished manuscript.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Series B*, **50**, 413–436.

Spokoiny, V.G. (1996). Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, **24**, 2477–2498.

Stanton, R. (1997). A nonparametric model of term structure dynamics and the market price of interest rate risk. Journal of Finance, **5**, 1973–2002.

Stenseth, N.C., Falck, W., Chan, K.S., Bjørnstad, O.N., O'Donoghue, M., Tong, H., Boonstra, R., Boutin, S., Krebs, C.J., and Yoccoz, N.G. (1999). From ecological patterns to ecological processes: Phase- and density-dependencies in Canadian lynx cycle. *Proceedings of the National Academy of Sciences USA*, **95**, 15430–15435.

Stoker, T.M. (1992). *Lectures on Semiparametric Econometrics*. Center for Operational Research and Econometrics, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Stone, C.J. (1977). Consistent nonparametric regression. *The Annals of Statistics*, **5**, 595–645.

Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, **8**, 1348–1360.

Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, **10**, 1040–1053.

Stone, C.J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13**, 689–705.

Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, **14**, 590–606.

Stone, C.J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *The Annals of Statistics*, **22**, 118–184.

Stone, C.J., Hansen, M., Kooperberg, C., and Truong, Y.K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *The Annals of Statistics*, **25**, 1371–1470.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B*, **36**, 111–147.

Stramer, O., Tweedie, R.L., and Brockwell, P.J. (1996). Existence and stability of continuous time threshold ARMA processes. *Statistica Sinica*, **6**, 715–723.

Subba Rao, T. (1981). On the theory of bilinear models. *Journal of the Royal Statistical Society, Series B*, **43**, 224–255.

Subba Rao, T. (1983). The bispectral analysis of nonlinear time series with reference to bilinear time series models. In *Handbook of Statistics* , Vol. 3 (D.R. Brillinger and P.R. Krishnaiah, eds.). North-Holland, Amsterdam, pp. 293–391,

Subba Rao, T. and Gabr, M.M. (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models*. Springer-Verlag, New York.

Sugihara, G. and May, R.M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement errors in time series. *Nature*, **344**, 734-741.

Swanepoel, J.W. and van Wyk, J.W.J. (1986). The bootstrap applied to spectral density function estimation. *Biometrika*, **73**, 135–142.

Taniguchi, M. and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer, New York.

Taylor, S.J. (1986). *Modelling Financial Time Series*. Weily, New York.

Terdik, G. (1999). *Bilinear Stochastic Models and Related Problems of Nonlinear Time Series Analysis: A Frequency Domain Approach*. Springer-Verlag, New York.

Tiao, G.C. and Tsay, R.S. (1994). Some advances in nonlinear and adaptive modeling in time series. *Journal of Forecasting*, **13**, 109–131.

Tibshirani, R. (1996). Regression shrinkage and selection via lasso. *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Tjøstheim, D. (1990). Non-linear time series and Markov chains. *Advances in Applied Probability*, **22**, 587–611.

Tjøstheim, D. (1994). Non-linear time series: A selective review. *Scandinavian Journal of Statistics*, **21**, 97–130.

Tjøstheim, D. (1996). Measures of dependence and tests of independence. *Statistics*, **28**, 249–284.

Tjøstheim, D. and Auestad, B.H. (1994a). Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association*, **89**, 1398–1409.

Tjøstheim, D. and Auestad, B.H. (1994b). Nonparametric identification of nonlinear time series: Selecting significant lags. *Journal of the American Statistical Association*, **89**, 1410–1419.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis*. Springer-Verlag, New York.

Tong, H. (1990). *Non-linear Time Series: A Dynamical Systems Approach*, Oxford University Press, Oxford.

Tong, H. (1995). A personal overview of non-linear time series analysis from a chaos perspective (with discussion). *Scandinavian Journal of Statistics*, **22**, 399–445.

Tong, H. and Lim, K.S. (1980). Threshold autoregression, limit cycles and cyclical data (with discussion). *Journal of the Royal Statistical Society, Series B*, **42**, 245–292.

Tong, H. and Moeanaddin, R. (1988). On multi-step non-linear least squares prediction. *The Statistician*, **37**, 101–110.

Tran, L.T. (1989). Recursive density estimation under dependence. *IEEE Transactions on Information Theory*, **35**, 1103–1108.

Tran, L.T. (1993). Nonparametric function estimation for time series by local average estimators. *The Annals of Statistics*, **21**, 1040–1057.

Tran, L.T., Roussas, G., Yakowitz, S.J., and Truong, V.B. (1996). Fixed-design regression for linear time series. *The Annals of Statistics*, **24**, 975–991.

Truong, Y.K. (1991). Nonparametric curve estimation with time series errors. *Journal of Statistical Planning and Inference*, **28**, 167–183.

Truong, Y.K. and Stone, C.J. (1992). Nonparametric function estimation involving time series. *The Annals of Statistics*, **20**, 77–97.

Truong, Y.K. and Stone, C.J. (1994). Semiparametric time series regression. *Journal of Time Series Analysis*, **15**, 405–428.

Tsay, R.S. (1989). Testing and modelling threshold autoregressive processes. *Journal of the American Statistical Association*, **84**, 231–240.

Tsay, R.S. (2002). *Analysis of Financial Time Series*. Wiley, New York.

Tsybakov, A.B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmission*, **22**, 133–146.

Tsybakov, A. B. (1998). Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *The Annals of Statistics*, **26**, 2420–2469.

Tukey, J.W. (1967). An introduction to the calculations of numerical spectrum analysis. In *Advanced Seminar on Spectral Analysis of Time Series* (B. Harris, ed.). Wiley, New York, pp. 25–46.

Tweedie, R.L. (1975). Sufficient conditions for ergodicity and recurrence of Markov chain on a general state space. *Stochastic Processes and Their Applications*, **3**, 385–402.

Tweedie, R.L. (1976). Criteria for classifying general Markov chains. *Advances in Applied Probability*, **8**, 737–771.

Tweedie, R.L. (1983). Criteria for rate of convergence of Markov chain with application to queuing and storage theory. In *Probability, Statistics and Analysis* (J.F.K. Kingman and G.E.H. Reuter, eds.). Springer-Verlag, New York, pp. 260–276,

Tyssedal, J.S. and Tjøstheim, D. (1988). An autoregressive model with suddenly changing parameters and an application to stock market prices. *Applications of Statistics*, **37**, 353–369.

Utreras, F.D. (1980). Sur le choix du parametre d'adjustement dans le lissage par fonctions spline. *Numerische Mathematik*, **34**, 15–28.

Vasicek, O.A. (1977). An equilibrium characterization of the term structure. **5**, 177–188.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. Wiley, New York.

Vieu, P. (1991). Quadratic errors for nonparametric estimates under dependence. *Journal of Multivariate Analysis*, **39**, 324–347.

Volkonskii, V.A. and Rozanov, Yu.A. (1959). Some limit theorems for random functions. *Theory Prob. Appl.*, **4**, 178–197.

Wahba, G. (1975). Smoothing noisy data with spline functions. *Numerische Mathematik*, **24**, 383–393.

Wahba, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In *Applications of Statistics* (P.R. Krishnaiah, ed.). North-Holland, Amsterdam, pp. 507–523.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society, Series B*, **40**, 364–372.

Wahba, G. (1980). Automatic smoothing of the log periodogram. *Journal of the American Statistical Association*, **75**, 122–132.

Wahba, G. (1984). Partial spline models for semiparametric estimation of functions of several variables. In *Statistical Analysis of Time Series*, Proceedings of the Japan U.S. Joint Seminar, Tokyo, 319–329. Institute of Statistical Mathematics, Tokyo.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wand, M.P. (2000). A central limit theorem for local polynomial backfitting estimators. *Journal of Multivariate Analysis*, **70**, 57–65.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Wand, M.P., Marron, J.S., and Ruppert, D. (1991). Transformations in density estimation (with discussion). *Journal of the American Statistical Association*, **86**, 343–361.

Wang, Y. (1996). Function estimation via wavelet shrinkage for long-memory data. *The Annals of Statistics*, **24**, 466–484.

Wang, Y. (2002). Asymptotic nonequivalence of GARCH models and diffusions. *The Annals of Statistics*, **30**, 754–783.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā, Series A*, **26**, 359–372.

Watson, G.S. and Leadbetter, M.R. (1963). On the estimation of the probability density, I. *Annals of Mathematical Statistics*, **34**, 480–491.

Weigend, A.S. and Gershenfeld, N.A. (1994). *Time Series Prediction*. Proceedings Volume 15, Santa Fe Institute Studies in the Sciences of Complexity. Addison-Wesley, New York.

Weiss, A.A. (1984). ARMA models with ARCH errors. *Journal of Time Series Analysis*, **5**, 129–143.

Weiss, A.A. (1986). Asymptotic theory for ARCH models: estimation and testing. *Econometric Theory*, **2**, 107–131.

West, W. and Harrison, P.J. (1989). Bayesian Forecasting and Dynamic Models (2nd ed.). Springer-Verlag, New York.

Whittle, P. (1962). Gaussian estimation in stationary time series. *Bulletin of the International Statistical Institute*, **39**, 105–129.

Wilks, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, **9**, 60–62.

Withers, C.S. (1981). Central limit theorems for dependent variables I. *Zeitschriftfuer Wahrscheinlichkeitstheorie verw. Gebiete*, **57**, 509–534.

Wittaker, E.T. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, **41**, 63–75.

Wong, W.H. (1984). On constrained multivariate splines and their approximations. *Numerische Mathematik*, **43**, 141–152.

Woodroofe, M. (1982). On model selection and the arc sine laws. *The Annals of Statistics*, **10**, 1182–1194.

Woolhouse, W.S.B. (1870). Explanation of a new method of adjusting mortality tables, with some observations upon Mr. Makeham's modification of Gompertz's theory. *J. Inst. Act.*, **15**, 389–410.

Wu, C.O. and Chiang, C.T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variables. *Statistica Sinica*, **10**, 433–456.

Wu, C.O., Chiang, C.T., and Hoover, D.R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association*, **93**, 1388–1401.

Xia, Y. and Li, W.K. (1999a). On single-index coefficient regression models. *Journal of the American Statistical Association*, **94**, 1275–1285.

Xia, Y. and Li, W.K. (1999b). On the estimation and testing of functional-coefficient linear models. *Statistica Sinica*, **9**, 735–757.

Xia, Y., Tong, H., Li, W.K., and Zhu, L.X. (2002). An adaptive estimation of dimension reduction space (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**, 363–410.

Xie, Z. (1993). *Case Studies in Time Series Analysis*. World Scientific, Singapore.

Yakowitz, S.J. (1979). Nonparametric estimation of Markov transition functions. *The Annals of Statistics*, **7**, 671–679.

Yakowitz, S.J. (1985). Nonparametric density estimation, prediction, and regression for Markov sequences. *Journal of the American Statistical Association*, **80**, 215–221.

Yang, L. and Marron, J.S. (1999). Iterated transformation-kernel density estimation. *Journal of the American Statistical Association*, **94**, 580–589.

Yao, Q. and Brockwell, P.J. (2001). Gaussian maximum likelihood estimation for ARMA models I: time series. (Submitted.)

Yao, Q. and Tong, H. (1994a). Quantifying the inference of initial values on nonlinear prediction. *Journal of the Royal Statistical Society, Series B*, **56**, 701–725.

Yao, Q. and Tong, H. (1994b). On subset selection in non-parametric stochastic regression. *Statistica Sinica*, **4**, 51–70.

Yao, Q. and Tong, H. (1995a). On initial-condition sensitivity and prediction in nonlinear stochastic systems. *Bull. Int. Statist. Inst.*, **IP 10.3**, 395–412.

Yao, Q. and Tong, H. (1995b). On prediction and chaos in stochastic systems. In *Chaos and Forecasting* ed. by H Tong, World Scientific, Singapore, pp. 57–86. (A short version was published in *Philosophical Transactions of the Royal Society (London)*, **A**, **348**, pp. 357–369 (1994).)

Yao, Q. and Tong, H. (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Journal of Nonparametric Statistics*, **6**, 273–292.

Yao, Q. and Tong, H. (1998a). Cross-validatory bandwidth selections for regression estimation based on dependent data. *Journal of Statistical Planning and Inference*, **68**, 387–415.

Yao, Q. and Tong, H. (1998b). A bootstrap detection for operational determinism. *Physica*, D, **115**, 49–55.

Yao, Q. and Tong, H. (2000). Nonparametric estimation of ratios of noise to signal in stochastic regression. *Statistica Sinica*, **10**, 751–770.

Yee, T.W. and Wild, C.J. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society, Series B*, **58**, 481–493.

Yoshihara, K. (1976). Limiting behaviour of U-statistics for a stationary absolutely regular process. *Zeitschriftfuer Wahrscheinlichkeitstheorie verw. Gebiete*, **35**, 237–252.

Young, P.C. (1993). Time variable and state dependent modelling of non-stationary and nonlinear time series. *Developments in Time Series Analysis* (T.Subba Rao, ed.). Chapman and Hall, London, pp. 374–413.

Young, P.C. and Beven, K.J. (1994). Data-based mechanistic modelling and the rainfall-flow nonlinearity. *Environmetrics*, **5**, 335–363.

Yu, K. and Jones, M.C. (1998). Local linear quantile regression. *Journal of the American Statistical Association*, **93**, 228–237.

Yule, G.U. (1927). On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society (London)*, **A**, **226**, 267-298.

Zhang, C. (2002a). Adaptive tests of regression functions via multi-scale generalized likelihood ratios. *Canadian Journal of Statistics*, to appear.

Zhang, C. (2002b). Calibrating the degrees of freedom for automatic data-smoothing and effective curve-checking. Unpublished Manuscript.

Zhang, G., Patuwo, B.E., and Hu, M.Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, **14**, 35–62.

Zhang, H.P. and Singer, B. (1999) *Recursive Partitioning in the Health Sciences*. Springer-Verlag: New York.

Zhang, W. and Lee, S.Y. (2000). Variable bandwidth selection in varying coefficient models. *Journal of Multivariate Analysis*, **74**, 116–134.

Zhang, Z. and Tong, H. (2001). On some distributional properties of first order non-negative bilinear time series models. *Journal of Applied Probability*, **38**, 659–671.

Zhu, L.X. and Fang, K.T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, **24**, 1053–1068.

Zygmund, D. (1968). *Trigonometric Series*, Vols. I, II. Cambridge University Press, Cambridge.

Author index

Subject index