Optimal Statistical Inference in Financial Engineering

Optimal Statistical Inference in Financial Engineering

Masanobu Taniguchi
Junichi Hirukawa
Kenichiro Tamaki

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an **informa** business

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-58488-591-7 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Taniguchi, Masanobu.
Optimal statistical inference in financial engineering / Masanobu Taniguchi, Junichi Hirukawa, and Kenichiro Tamaki.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-58488-591-7 (alk. paper)
1. Mathematical statistics. 2. Financial engineering. 3. Finance--Statistical methods. I. Hirukawa, Junichi. II. Tamaki, Kenichiro. III. Title.

QA276.T27 2008
332.01'51923--dc22 2007027986

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

To Our Families

Contents

Preface

The field of financial engineering has developed as a huge integration of economics, mathematics, probability theory, statistics, time series analysis, operation research, etc. over the last decade. First, we describe financial assets as stochastic processes. Using stochastic differential equations, probabilists developed highly sophisticated mathematical theory in this field. On the other hand empirical people in financial econometrics studied various numerical aspects of financial data by means of statistical methods. However, systematic studies based on optimal statistical inference for stochastic processes have been barren although they are very important and fundamental in financial engineering. Black and Scholes provided the modern option pricing theory assuming that the price process of an underlying asset follows a geometric Brownian motion. But, a lot of empirical studies for the price processes of assets show that they do not follow the geometric Brownian motion. Therefore it is important to investigate which stochastic models can describe the actual financial data sufficiently, and how to estimate the proposed models optimally. We think that financial engineering should be constructed on this ground. The purpose of this book is motivated by the above view.

First, we explain the elements of probability and statistical inference for independent observations in Chapters 2 and 3. Chapter 4 discusses the problem of testing hypothesis and discriminant analysis for independent observations. Chapter 5 provides an introduction to stochastic processes, which includes the spectral theory for stationary processes, and martingale and central limit theorems for stochastic processes. In Chapter 6, we deal with many famous time series models, and discuss their asymptotically optimal inference. The problem of prediction and discriminant analysis is also presented. In Chapter 7, we give a bridge for financial engineering based on the statistical inference for stochastic processes. Chapter 7 concretely addresses the problems of option pricing theory, statistical estimation for portfolio coefficients, and the VaR problem using the residual empirical return processes. In Chapter 8 we introduce some models for interest rates and discount bonds, and discuss their no-arbitrage pricing theory. Some empirical studies will be given. In Chapter 9 we investigate problems of credit rating, based on a methodology of time series analysis discussed in Chapter 6. We will execute the clustering of stock returns in both the New York and Tokyo Stock Exchanges.

This book can be used as a textbook of mathematical statistics for under-

graduate students, and as one of time series analysis for graduate students. Also, this may be a research book for people in the fields of statistics, financial engineering, econometrics and mathematics. We hope that readers recognize the importance of financial engineering constructed on statistical optimal inference theory.

We are indebted to Professor Howell Tong, London School of Economics, who recommended the original version of this book to the financial mathematics series at Chapman & Hall/CRC. We are indebted also to Professor Tadashi Uratani, Hosei University, for his comments on the original version. Finally we thank all the members of the statistical group of Waseda University, especially Professor Takeru Suzuki, and the editors of Chapman & Hall/CRC for their cooperation.

CHAPTER 1

Introduction

Statistics for independent samples has been developing based on the fundamental theory = mathematical statistics, and the applications have been expanding to enormous fields, e.g., engineering, medical science, economics, finance, psychology, etc. The entirety may be called statistical science. Further, the statistics for independent samples has been extended to that for stochastic processes which are probability models describing that past, present and future phenomena are interacting (dependent). The statistics for stochastic processes is called time series analysis, and is now developing based on mathematical statistics.

Recently the field of finance, showing very complicated aspects, has constituted a huge domain involving economics, mathematics, probability theory, statistics, operations research, etc., which is called financial engineering. Let us think about an option security of a specified asset e.g., stock. If the price of asset S on a specified future date is higher than a specified price K, the option security bears the profit $S - K$. Assuming that the price of asset follows a geometric Brownian motion process, Black and Scholes (1973) valuated the reasonable price, called the Black-Scholes formula, which is the origin of financial engineering. Therefore, as the foundation of financial engineering it is most important how the stochastic process models for assets describe the real financial data sufficiently. This is exactly the theme of time series analysis. In this book, first we review the mathematical statistics for independent samples. Next we develop the statistical analysis for stochastic processes, i.e., dependent data, and describe the optimal inference theory for time series. Based on this we discuss option pricing, optimal portfolio estimation, VaR problems, term structure for spot rates, and cluster analysis for financial time series etc. Hence, the purpose of this book is to build a bridge between the optimal statistical inference for time series and financial engineering.

Now, let us look at actual stock price data. Figure 1.1 plots the daily return of Hewlett-Packard company from February 2, 1984 to December 31, 1991.

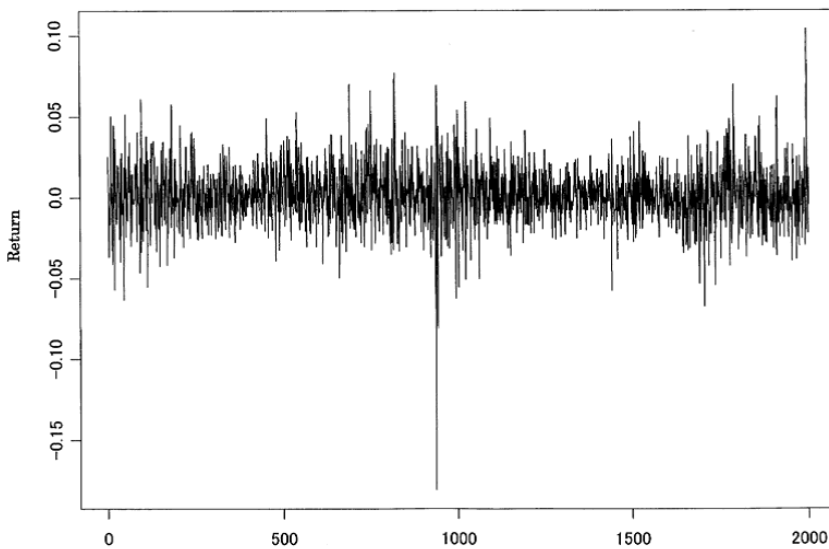


Figure 1.1 *The daily return of Hewlett-Packard Company from February 2, 1984 to December 31, 1991.*

In what follows we write the observed stretch as $X_1, X_2, ..., X_n$. As an elementary time series analysis, we often check the behavior of the following sample autocorrelation function

$$S_{X_t}(l) = \frac{\sum_{t=1}^{n-l}(X_{t+l} - \bar{X}_n)(X_t - \bar{X}_n)}{\sum_{t=1}^{n}(X_t - \bar{X}_n)^2}, \tag{1.1}$$

where $\bar{X}_n = n^{-1}\sum_{t=1}^{n} X_t$. Figure 1.2 plots the values of $S_{X_t}(l)$, and Figure 1.3 plots the values of the sample autocorrelation function of the square transformed process X_t^2, i.e., $S_{X_t^2}(l)$.

The sample autocorrelation function $S_{X_t}(l)$ shows a strength of interaction between X_{t+l} and X_t. If X_t's are mutually independent or uncorrelated, $S_{X_t}(l)$ would be near zero except for $l \neq 0$ (we will explain the fundamental terminology of probability and statistics in Chapters 2-6).

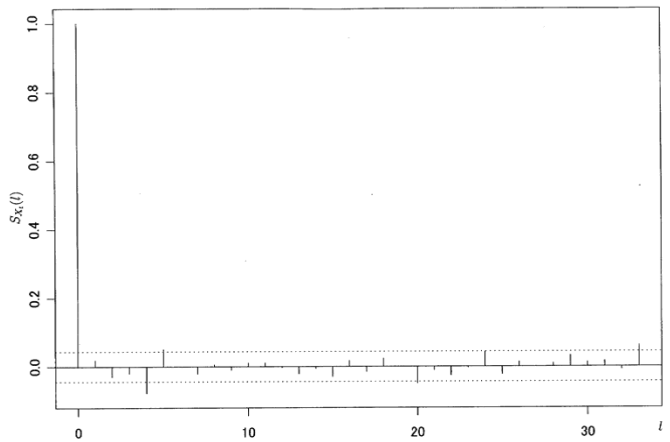


Figure 1.2 *The sample autocorrelation function* $S_{X_t}(l)$ *of* $\{X_t\}$.

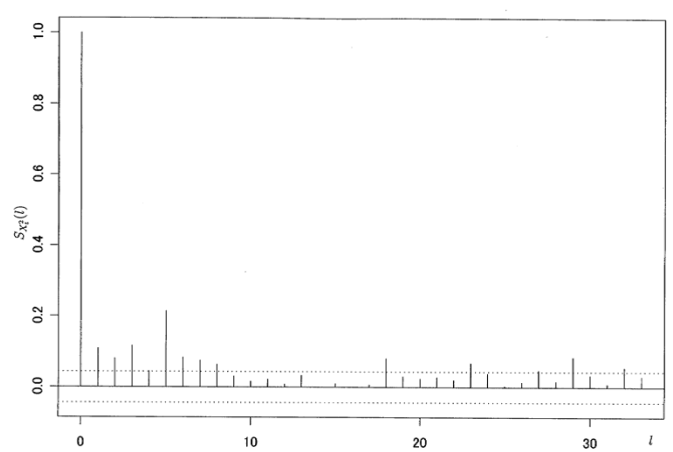


Figure 1.3 *The sample autocorrelation function* $S_{X_t^2}(l)$ *of the square transformed process* X_t^2.

From Figures 1.2 and 1.3 we observe

(1) X_t's are almost uncorrelated,

(2) X_t^2's are not uncorrelated.

which imply that X_t's are not mutually independent and that the distribution is not normal (Gaussian), i.e.,

(3) X_t's are not mutually independent,

(4) the distribution of $\{X_t\}$ is non-Gaussian.

In view of substantial analysis for economic data, we usually observe the above findings for general financial returns. As a model for the price of asset, Black and Scholes proposed a geometric Brownian motion process whose logarithm has independent Gaussian increments. However, from the observation of (3) and (4) we have to say that their financial world does not describe the actual financial world sufficiently, which motivates the subject of this book. That is, we will develop the statistical financial engineering based on the optimal statistical inference for non-Gaussian dependent processes.

This book is organized as follows. Chapter 2 reviews the elements of mathematical statistics, which include probability space, random variable and probability distribution. In actual problems, we often argue about the distribution of multidimensional variables and the existence of effects between variables. In relation to this, we consider the distribution of an n-dimensional random vector, and the independence between their components. Furthermore, we explain the concepts of expectation, characteristic function and conditional distribution and introduce various types of convergence of sequences of random variables and the central limit theorems.

Chapter 3 gives a brief survey of the statistical inference. We introduce the idea of sufficient statistics and construct minimum variance unbiased estimators by using them. We also derive the Cramér-Rao bound which gives a lower bound for the variance of unbiased estimators, and discuss the efficient estimator which attains it. Furthermore, in the case that the sample sizes are "large", we define a "goodness" of asymptotic inference theory and describe the asymptotic optimality of estimators.

In Chapter 4 we present various statistical methods including interval estimations, testing problems and discriminant analyses. First, we discuss a method for seeking interval, in which unknown parameter lies at certain level of probability accuracy (interval estimation). The process of determining whether the observation indicates that the hypothesis is true or not is called the testing

hypothesis. We use a statistic to make a decision whether the hypothesis is true or not and call this the test statistic. We describe the fundamental theory concerned with the optimality of test statistics. Furthermore, we explain various types of testing hypotheses. At the end of this chapter we introduce the discriminant analysis which is fundamental and of importance in various fields of applied statistics.

Chapter 5 explains elements of stochastic processes, e.g., stationarity, spectral structure, ergodicity, mixing properties, martingale, etc. Because the statistical analysis for stochastic processes largely relies on the asymptotic theory with respect to the length of observations, we provide some useful limit theorems and central limit theorems.

Chapter 6 explains typical linear and nonlinear parametric models of time series, and states the asymptotically optimal estimation for their unknown parameters. We also discuss the problem of model selection by use of some information criteria.

Since it is often difficult for parametric models to describe the real world sufficiently, we address nonparametric and semiparametric estimation problems for spectra of stationary processes and trend functions of time series regression models. Usually we assume stationarity of the concerned processes. However, this assumption is often severe for actual time series data. We also study the statistical inference for a class of important nonstationary processes, called locally stationary processes.

Next the problem of prediction and discriminant analysis is addressed. We derive the best predictor in terms of spectral density and conditional expectation, and discuss its statistical estimation. Regarding discriminant analysis we give an asymptotically optimal discriminator, and evaluate the asymptotic misclassification probabilities. Discriminant analysis for time series is applied to clustering financial time series data.

Chapter 7 provides an introduction to statistical financial engineering, which includes option pricing by use of the CHARN model, higher order option pricing via Edgeworth expansion for non-Gaussian dependent return processes, and asymptotically efficient estimation theory for optimal portfolio coefficients in the case of non-Gaussian dependent returns. Also the problem of VaR by use of asymptotics of the residual empirical return processes is addressed. Then we propose a feasible VaR which keeps assets away from a specified risk with high confidence level.

In Chapter 8 we introduce some models for interest rates and discount bonds, and discuss their no-arbitrage pricing theory. Some empirical studies for the term structure of discount bond, yield-to-maturity and forward rate are given.

One of the most interesting topics in the field of financial engineering is problems of credit rating. Usually credit rating has been done by use of i.i.d. settings. In Chapter 9 we investigate problems of credit rating, based on a

methodology of time series analysis. We consider the case that concerned time series are locally stationary processes. We discuss a clustering problem of stock data, using both parametric and nonparametric approaches for estimation of time varying spectral densities. Furthermore, we suggest a credit rating based on taking into account not only covariance structures but mean structures.

CHAPTER 2

Elements of Probability

In modern probability theory the mathematical model of random phenomena is represented by a probability space $(\Omega, \mathcal{A}, P)$, where Ω is the set which denotes the entirety of possible outcomes of the random phenomena, $P(B)$ is the probability assigned to the outcome belonging to a subset B of Ω and $\mathcal{A}$ is the collection of subsets which consists of all B on which $P(B)$ are defined. All the concepts related to probability are mathematically described based on this probability space. The probability space is mathematically a measure space. Although we provide a concise description of the measure theory and the integral calculus in the Appendix, readers who are interested in detail may refer to, e.g., Ash and Doléans-Dade (2000).

2.1 Probability and Probability Distribution

In a toss of one die, let the outcome be the number of spots on the die. Then, the collection of possible outcomes becomes $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let A and B be the subsets of elements in Ω for which even and odd numbers come out, respectively. Thus A and B are the sets $\{2, 4, 6\}$ and $\{1, 3, 5\}$. Let us start from defining a set of sets whose probabilities are defined, such as A, B, Ω. In general, the collection of all possible outcomes is denoted by Ω and is called the *sample space*. We may take any abstract set as a sample space. The probabilities are not necessarily defined on all subsets of Ω. The following definition provides the collection $\mathcal{A}$ of subsets whose probabilities are defined and we call the element of $\mathcal{A}$ *event*.

Definition 2.1 *A collection $\mathcal{A}$ of subsets of Ω satisfying the following (A1)-(A3) is called the* σ-field.

(A1) $\Omega \in \mathcal{A}$,

(A2) *If* $A \in \mathcal{A}$, *then* $A^c \in \mathcal{A}$,

(A3) *If* $A_1, A_2, \ldots \in \mathcal{A}$, *then* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

Next we define the probability. For example, think of having made n repeated tossing of one die. Let C denote the event for which $\{1\}$ comes out in the toss of one die. Then we can count the number k of times (the frequency) that

the event C actually occurred in the n tossings. The ratio $f_C = k/n$ is called the *relative frequency* of the event C in the n experiments. Therefore, it seems reasonable that the probability of the event C is defined by

$$\text{“}\lim_{n\to\infty}\text{”}\frac{k}{n}. \tag{2.1}$$

However, since the above is the limit of outcome in the n experiments, this definition is inadequate for the mathematical definition. In modern probability theory the definition of probability consists of three axioms motivated by the following three intuitive properties of relative frequency. Namely, $f_C \geq 0$, $f_C \leq 1$ and $f_{C_1 \cup C_2} = f_{C_1} + f_{C_2}$ for disjoint events C_1 and C_2. From the axioms of probability, the readers can understand that Theorem 2.1 below agrees with intuition of relative frequency.

Definition 2.2 (Probability) *Let Ω be a sample space and let $\mathcal{A}$ be a σ-field on Ω. Let P be a real valued function defined on $\mathcal{A}$. Then P is called a* probability (probability measure) *if P satisfies the following three conditions:*

(P1) *For all $A \in \mathcal{A}$, $P(A) \geq 0$.*

(P2) $P(\Omega) = 1$.

(P3) *If $A_1, A_2, \ldots \in \mathcal{A}$, and for all i, j $(i \neq j)$, $A_i \cap A_j = \emptyset$, then*

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \tag{2.2}$$

Henceforth, we call the triple $(\Omega, \mathcal{A}, P)$ a *probability space*. All the arguments of modern probability and mathematical statistics start from Definition 2.2 and agree to the so-called measure theory.

The following theorem gives us some other properties of probability. In the statements of this theorem, $P(A)$ is taken to be the probability defined on a σ-field $\mathcal{A}$ of a sample space Ω.

Theorem 2.1 **(i)** $P(\emptyset) = 0$.

(ii) *If $A_1, \ldots, A_n \in \mathcal{A}$, and for any i, j $(i \neq j)$, $A_i \cap A_j = \emptyset$, then*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i). \tag{2.3}$$

(iii) *For each event $A \in \mathcal{A}$,*

$$P(A^c) = 1 - P(A). \tag{2.4}$$

(iv) *If $A, B \in \mathcal{A}$, and $A \subset B$, then $P(A) \leq P(B)$.*

(v) *For each $A \in \mathcal{A}$, $0 \leq P(A) \leq 1$.*

(vi) *For any* $A_1, A_2, \ldots \in \mathcal{A}$,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i). \tag{2.5}$$

(vii) *If* $A_1 \subset A_2 \subset \cdots \subset A_n \subset \cdots$, $A_n \in \mathcal{A}$ $(n = 1, 2, \ldots)$, *then*

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n). \tag{2.6}$$

(viii) *If* $A_1 \supset A_2 \supset \cdots \supset A_n \supset \cdots$, $A_n \in \mathcal{A}$ $(n = 1, 2, \ldots)$, *then*

$$P\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n). \tag{2.7}$$

PROOF

(i) Let $A_1 = \Omega$ and $A_i = \emptyset$, $i \geq 2$ in (P3). Then we have $\bigcup_{i=1}^{\infty} A_i = \Omega$. Hence $P(\Omega) = P(\Omega) + \sum_{k=2}^{\infty} P(\emptyset)$. From (P1) and (P2), it follows that $P(\emptyset) = 0$.

(ii) Let $A_i = \emptyset$, $i \geq n+1$ in (P3). Then we have $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{n} A_i$ and $A_i \cap A_j = \emptyset$ for all $i, j \in \mathbf{N}$ $(i \neq j)$. Hence

$$P\left(\bigcup_{i=1}^{n} A_i\right) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{n} P(A_i) + \sum_{i=n+1}^{\infty} P(\emptyset). \tag{2.8}$$

Thus, the assertion follows from (i).

(iii) We have $\Omega = A \cup A^c$ and $A \cap A^c = \emptyset$. Thus, from (P2) and (ii), it follows that

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c). \tag{2.9}$$

(iv) Now $B = A \cup (A^c \cap B)$ and $A \cap (A^c \cap B) = \emptyset$. Hence, from (ii),

$$P(B) = P(A) + P(A^c \cap B). \tag{2.10}$$

From (P1), $P(A^c \cap B) \geq 0$. Hence, $P(B) \geq P(A)$.

(v) Since $A \subset \Omega$, from (P1) and (iv), we have

$$0 \leq P(A) \leq P(\Omega) = 1. \tag{2.11}$$

(vi) Let $\tilde{A}_1 = A_1$ and $\tilde{A}_n = A_n \backslash \left(\bigcup_{i=1}^{n-1} A_i\right)$, $n \geq 2$, where $B \backslash A$ denotes $B \cap A^c$. Then, the events $\tilde{A}_n$, $n \in \mathbf{N}$ become mutually disjoint, $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} \tilde{A}_n$ and $\tilde{A}_n \subset A_n$. Therefore, we have

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = P\left(\bigcup_{n=1}^{\infty} \tilde{A}_n\right) = \sum_{n=1}^{\infty} P\left(\tilde{A}_n\right) \leq \sum_{n=1}^{\infty} P\left(A_n\right). \tag{2.12}$$

(vii) Let $A_0 \equiv \emptyset$, $\tilde{A}_n \equiv A_n \setminus A_{n-1}$, $n \in \mathbf{N}$. Then the events $\tilde{A}_n$, $n \in \mathbf{N}$ become mutually disjoint and $\bigcup_{n=1}^{\infty} \tilde{A}_n = \bigcup_{n=1}^{\infty} A_n$. Since $A_n = A_{n-1} \cup \tilde{A}_n$, $A_{n-1} \cap \tilde{A}_n = \emptyset$, $n \in \mathbf{N}$ and $P(A_0) = 0$, it follows that

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = P\left(\bigcup_{n=1}^{\infty} \tilde{A}_n\right) = \sum_{n=1}^{\infty} P\left(\tilde{A}_n\right) = \lim_{N\to\infty} \sum_{n=1}^{N} P\left(\tilde{A}_n\right)$$
$$= \lim_{N\to\infty} \sum_{n=1}^{N} \{P(A_n) - P(A_{n-1})\} = \lim_{N\to\infty} P(A_N). \qquad (2.13)$$

(viii) Applying (vii) above to the complement A_n^c, $n \in \mathbf{N}$ leads to

$$P\left(\bigcup_{n=1}^{\infty} A_n^c\right) = \lim_{n\to\infty} P(A_n^c). \qquad (2.14)$$

From (iii) the assertion follows. □

If the elements of a sample space Ω are not numbers, it may be inconvenient to deal with mathematically. We shall formulate a rule that the elements ω of Ω correspond to numbers. We begin with the toss of a coin. The sample space is $\Omega = \{\omega : \omega \text{ is } T \text{ or } H\}$ where T and H represent tails and heads, respectively. Let X be a function such that $X(\omega) = 0$ if ω is T and $X(\omega) = 1$ if ω is H. Thus X is a real-valued function defined on the sample space Ω which leads us from the sample space Ω to a space of real numbers $\mathcal{D} = \{0, 1\}$. In general X is a function from the sample space Ω to the reals (or the extended reals). We now formulate X mathematically.

Definition 2.3 *In a probability space $(\Omega, \mathcal{A}, P)$, if a real-valued function $X = X(\omega)$ defined on Ω satisfies*

$$\{\omega |\ X(\omega) \le x\} \ \in\ \mathcal{A}, \qquad (2.15)$$

for every $x \in \mathbf{R}$, then X is called a random variable *and*

$$F_X(x) \equiv P(\{\omega |\ X(\omega) \le x\}) \qquad (2.16)$$

is called the distribution function *of X.*

Henceforth, the smallest σ-field of subsets of $\mathbf{R}$ containing all intervals $(a, b]$, $a, b \in \mathbf{R}$ is called the family of *Borel sets* on $\mathbf{R}$ and is denoted by $\mathcal{B}(\mathbf{R})$ or $\mathcal{B}$. Also, if a σ-field contains all open intervals, it must contain intervals of the form $(a, b]$, and conversely, since

$$(a, b] = \bigcap_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right), \quad (a, b) = \bigcup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right]. \qquad (2.17)$$

Similarly, we may replace the intervals $(a, b]$ in the definition of $\mathcal{B}$ by other types of intervals (see Exercise 2.6). Although random variable X is defined as satisfying the relation (2.15), we can show that this implies

$$\{\omega |\ X(\omega) \in B\} \ \in\ \mathcal{A}, \qquad (2.18)$$

for any $B \in \mathcal{B}$. X satisfying the relation (2.18) is called an *$\mathcal{A}$-measurable function*, and if we set

$$P_X(B) \equiv P\{X(\omega) \in B\}, \quad (B \in \mathcal{B}), \tag{2.19}$$

then P_X is the probability on $\mathcal{B}$ and induces the probability space which is described by the mathematical triple $(\mathbf{R}, \mathcal{B}, P_X)$. From now on, we call P_X the *probability distribution* of X.

Now, let us check general properties of the distribution function.

Theorem 2.2 *For the distribution function F_X, we have the following:*

(1) *If $x \leq y$, then $F_X(x) \leq F_X(y)$ (monotonicity).*
(2) *If $x \searrow c$, then $F_X(x) \searrow F_X(c)$ (right-continuity).*
(3) $\lim_{x\to\infty} F_X(x) = 1$, $\lim_{x\to-\infty} F_X(x) = 0$.

PROOF

(1) If $x \leq y$, then $(-\infty, x] \subset (-\infty, y]$. From Theorem 2.1 (iv), we have

$$F_X(x) = P_X[(-\infty, x]] \leq P_X[(-\infty, y]] = F_X(y).$$

(2) It is sufficient to show that for an arbitrary sequence $\{x_n\}$ with $x_n \geq c$, $n \in \mathbf{N}$ and $\lim_{n\to\infty} x_n = c$, we have $F_X(c) = \lim_{n\to\infty} F_X(x_n)$. Indeed, let $A_k = (-\infty, x_k]$ and $B_n = \bigcup_{k=n}^{\infty} A_k$, then we have $\sup_{k\geq n} P_X(A_k) \leq P_X(B_n)$ and B_n is monotone decreasing. Therefore, it is seen that

$$\limsup_{n\to\infty} F_X(x_n) = \limsup_{n\to\infty} P_X(A_n) = \lim_{n\to\infty} \sup_{k\geq n} P_X(A_k) \leq \lim_{n\to\infty} P_X(B_n)$$

and, from Theorem 2.1 (viii),

$$\lim_{n\to\infty} P_X(B_n) = P_X\left(\bigcap_{n=1}^{\infty} B_n\right) = P_X\left(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} A_k\right) = P_X\left(\limsup_{n\to\infty} A_n\right).$$

Note that $\limsup_{n\to\infty} A_n = (-\infty, c]$, then we have $\limsup_{n\to\infty} F_X(x_n) \leq F_X(c)$. On the other hand, by (1) F_X has the monotonicity, hence $F_X(c) \leq \liminf_{n\to\infty} F_X(x_n)$. Therefore, $F_X(c) = \lim_{n\to\infty} F_X(x_n)$.

(3) If $x_n \nearrow \infty$, then $(-\infty, x_n]$ is increasing and converges to $\bigcup_{n=1}^{\infty}(-\infty, x_n] = \mathbf{R}$. From Theorem 2.1 (iv) and (vii), we have $F_X(x_n) \nearrow 1$. The monotonicity of F_X leads to $\lim_{x\to\infty} F_X(x) = 1$. Similarly, $\lim_{x\to-\infty} F_X(x) = 0$ follows from the fact that if $x_n \searrow -\infty$, then $(-\infty, x_n]$ monotonically converges to $\emptyset$. □

Given a probability distribution P_X, the distribution function is determined by (2.16). Conversely, given a distribution function F_X, there is a unique probability distribution P_X whose distribution function is F_X. Moreover, for a real-valued function F which satisfies (1)-(3), we can construct random variable X and the probability distribution P_X on $(\mathbf{R}, \mathcal{B})$ with the underlying F as its distribution function. (see e.g. Chung (2001)).

Definition 2.4 (i) *For the probability distribution P_X of a random variable X, if there exists a countable set $B = \{x_1, x_2, \ldots\}$ on $\mathbf{R}$ such that $P_X(B) = 1$, then X is called a* discrete random variable. *This implies that $\sum_{i=1}^{\infty} p(x_i) = 1$, where $p(x_i) = P(\{x_i\})$, and henceforth we call the function $p(x)$ the* probability function *of X.*

(ii) *Let the distribution function of a random variable X be F_X. If there exists a non-negative $\mathcal{B}$-measurable function f_X, which satisfies*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

for arbitrary $x \in \mathbf{R}$, then X is called a continuous random variable *and we call f_X the* probability density function *of X.*

Let us see some typical examples of probability distribution.

Examples of discrete distributions:

(1) Binomial distribution. Let X denote the number of successes in a sequence of n trials in which the probability of success is the same for all n trials, say, $P(\text{success}) = \theta$. The distribution of X, called the *binomial distribution* $B(n, \theta)$, is given by

$$p(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad (x = 0, 1, \ldots, n,\ 0 \le \theta \le 1) \tag{2.20}$$

In particular, $B(1, \theta)$ is called a *Bernoulli distribution.*

(2) Poisson distribution. The probability distribution whose probability function is given by

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad (x = 0, 1, 2, \ldots, \lambda > 0) \tag{2.21}$$

is called the *Poisson distribution* and denoted by $P_o(\lambda)$. This is the distribution of the number of events occurring in a fixed interval of time or space if the probability of more than one occurrence in a very short interval is a smaller order of magnitude than that of a single occurrence, and if the number of events in nonoverlapping intervals is not mutually affected.

Examples of continuous distribution:

(3) Normal distribution. The central role in what follows is played by the *normal distribution.* Its importance for large sample theory stems from the fact that the distributions of various fundamental statistics under sufficiently well-controlled conditions are often approximately normal (see Chapter 3). The probability density function of the normal distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{2.22}$$

If X is a random variable with this density, it is said to be normally distributed or, more precisely, to have the normal distribution $N(\mu, \sigma^2)$. If X is normally distributed, so is any linear function $Y = aX + b$. Moreover, if the distribution of X is $N(\mu, \sigma^2)$, that of Y is $N(a\mu + b, a^2\sigma^2)$ (Exercise 2.7). In particular, the random variable $Z = (X - \mu)/\sigma$ then has the *standard normal distribution* with probability density function

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{2.23}$$

obtained from (2.22) by setting $\mu = 0$ and $\sigma = 1$. The normal density (2.22) is symmetric about μ and unimodal. It therefore takes on its maximum value at μ and decreases as x tends away from μ on either side. In the standardized form (2.23), for example, the density is inversely proportional to $e^{x^2/2}$, which tends to infinity much more rapidly even than e^x.

Let us now consider briefly some other densities that have the same general shape (symmetric, unimodal) as the normal density but which decrease more slowly. In each case we shall give only one member of the family of densities, which is centered at 0 and in which the scale is chosen so as to give the density a simple form. If X is a random variable with this density, more general associated location-scale family is obtained as the family of densities of $aX + b$ for arbitrary b and $a > 0$.

(4) Cauchy distribution. The probability density function of the *Cauchy distribution* is

$$f(x) = \frac{1}{\pi(1 + x^2)} \tag{2.24}$$

which tends to zero at a rate of $1/x^2$. A curious consequence of this tail behavior affects the average $S_n \equiv n^{-1} \sum_{i=1}^{n} X_i$ of n independent, identically distributed Cauchy random variables $X_1, \ldots, X_n$. Usually, the average S_n of independent observations from the same distribution has a distribution that is much more concentrated than that of a single observation X_i (see the central limit theorem (Theorem 2.10)). However, in the Cauchy case, the distribution of S_n is the same as that of a single observation, regardless of the n observations being averaged (Exercise 2.8). In Figure 2.1, the empirical density function of S_n of a Cauchy distribution with $n = 100$ and that of a single observation X_i for 100 time experiments are plotted, which confirms the above fact. Although we do not believe in the realism of the Cauchy distribution, we are interested in how well the procedure stands under such extreme tail behavior. Hence, it is often of interest to test the performance of a statistical procedure in the Cauchy case.

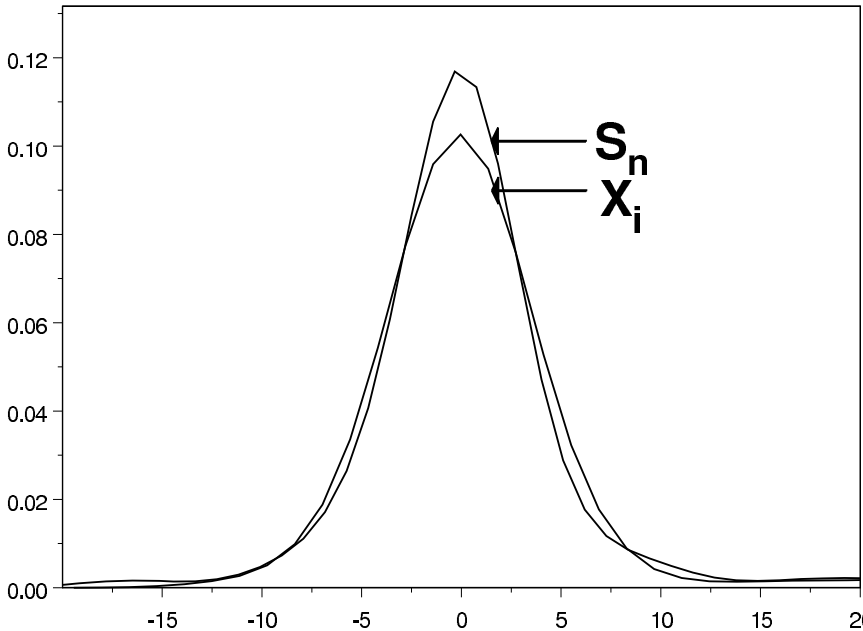


Figure 2.1 *The empirical density function of the average S_n of a Cauchy distribution with $n = 100$ and that of a single observation X_i for 100 time experiments.*

(5) Logistic distribution. A distribution whose tail behavior is intermediate between that of the normal and Cauchy distributions is the *logistic distribution*, whose density is

$$f(x) = \frac{e^{-x}}{(1 + e^{-x})^2}. \tag{2.25}$$

This is again symmetric about zero and unimodal (Exercise 2.9), and tends to zero at the rate of e^{-x} as x tends to infinity, more slowly than $e^{-x^2/2}$, but much faster than $1/x^2$. The probability density functions of the standard normal (2.23), Cauchy (2.24) and logistic (2.25) distributions are plotted in Figure 2.2. From this figure, we see that the tail of the probability density function of the logistic distribution tends to zero, more slowly than that of the standard normal distribution, but much faster than that of the Cauchy distribution.

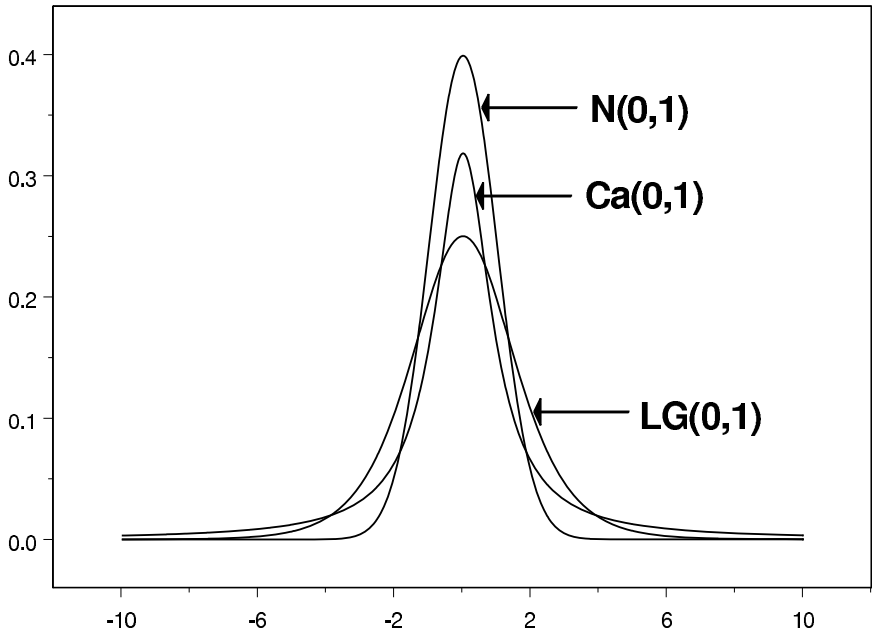


Figure 2.2 *The probability density functions of the standard normal (2.23), Cauchy (2.24) and logistic (2.25) distributions.*

(6) Double exponential distribution. A similar tail behavior to the logistic distribution is observed by the *double exponential distribution* with density

$$f(x) = \frac{1}{2}e^{-|x|} \tag{2.26}$$

which, however, has a quite different shape in the center. The probability density functions of the logistic (2.25) and double exponential (2.26) distributions are plotted in Figure 2.3. From this figure, we see that the tail behavior of both probability density functions are almost the same. On the other hand the probability density function of the double exponential distribution has a spike in its center.

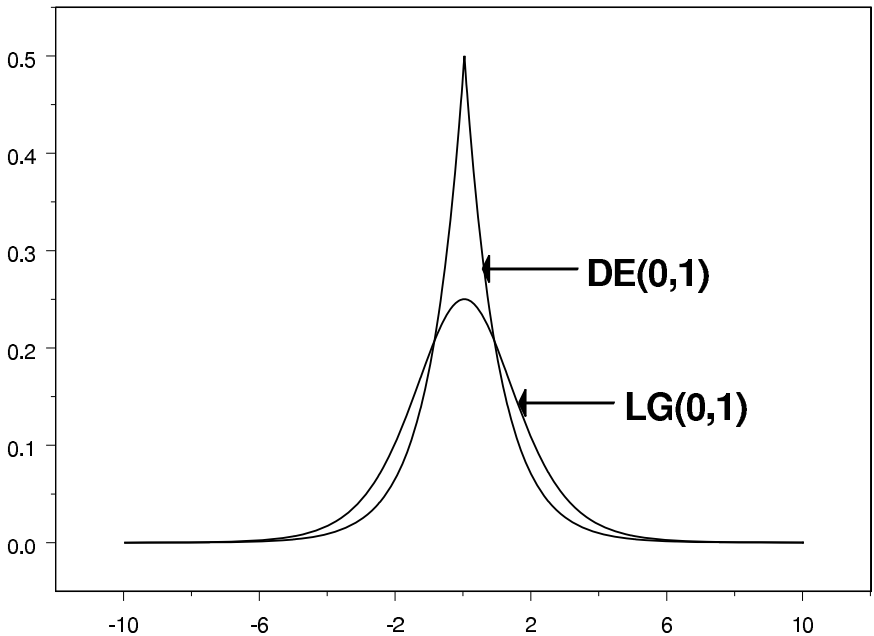


Figure 2.3 *The probability density functions of the logistic (2.25) and double exponential (2.26) distributions.*

For each of the densities (2.24), (2.25) and (2.26), the function

$$\frac{1}{a} f\left(\frac{x-b}{a}\right) \tag{2.27}$$

defines a probability density provided $a > 0$. These more general densities and their associated distributions are again referred to as Cauchy, logistic, and double exponential, respectively. Moreover, for fixed f in (2.27) but varying a and b, constitutes a location-scale family, the parameter a being a measure of scale and b a measure of location. Two further important location-scale families are provided by the *uniform* and *exponential distributions*.

(7) Uniform distribution. The probability density function of a *uniform distribution* is

$$f(x) = \begin{cases} 1/a, & (x \in (b - 1/2a, b + 1/2a)), \\ 0, & \text{otherwise.} \end{cases}$$

A random variable X with this density is said to have the uniform distribution $U(b-1/2a, b+1/2a)$. This is the probability density function of a point selected at random from the interval $(b-1/2a, b+1/2a)$.

(8) Exponential distribution. The probability density function of an *exponential distribution* is

$$f(x) = \begin{cases} \frac{1}{a}e^{-(x-b)/a}, & (x > b), \\ 0, & \text{otherwise} \end{cases} \qquad (0 < a,\ -\infty < b < \infty). \tag{2.28}$$

A random variable X with this density is said to have the exponential distribution $Exp(1/a, b)$. This distribution plays an important role in reliability theory, and survival analysis (see Lehmann (1975)).

2.2 Vector Random Variable and Independence

Very often in practice we are interested in measuring multivariate variables on each individual and their interaction. This section explains the fundamental notions of multivariate distributions.

Definition 2.5 *Let $X_1, \ldots, X_n$ be random variables defined on a probability space $(\Omega, \mathcal{A}, P)$. Then $\mathbf{X} = (X_1, \ldots, X_n)'$ is said to be an* n-dimensional random vector *and the function*

$$F_{X_1 \cdots X_n}(x_1, \ldots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n), \quad (x_i \in \mathbf{R},\ i = 1, \ldots, n)$$

is called the joint distribution function *of* $\mathbf{X}$. *Let $\{i_1, i_2, \ldots, i_k\}$, $(i_1 < i_2 < \cdots < i_k)$ be a subset of $\{1, 2, \ldots, n\}$. Then*

$$F_{X_{i_1} \cdots X_{i_k}}(x_{i_1}, \ldots, x_{i_k}) = F_{X_1 \cdots X_n}(\infty, \ldots, \infty, x_{i_1}, \infty, \ldots, \infty, x_{i_k}, \infty, \ldots, \infty)$$

is called the marginal distribution function *of* $(X_{i_1}, \ldots, X_{i_k})$.

Suppose that $\mathcal{B}_n$ is the smallest σ-algebra which contains every n-dimensional interval $(a_1, b_1] \times \cdots \times (a_n, b_n]$. Then

$$P_{\mathbf{X}}(\mathbf{B}) \equiv P\{\mathbf{X} \in \mathbf{B}\}, \quad (\mathbf{B} \in \mathcal{B}^n) \tag{2.29}$$

is said to be the *probability distribution* of $\mathbf{X} = (X_1, \ldots, X_n)'$. Therefore, $\mathbf{X}$ induces the probability space $(\mathbf{R}^n, \mathcal{B}^n, P_{\mathbf{X}})$.

Similarly as in the one-dimensional case, discrete and continuous probability distributions are defined for n-dimensional random vectors.

Definition 2.6 *(i) For the probability distribution $P_{\mathbf{X}}$ of an n-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_n)'$, if there exists a countable set $\mathbf{A} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ on $\mathbf{R}^n$ such that*

$$P_{\mathbf{X}}(\mathbf{A}) = 1,$$

then $\mathbf{X}$ *is said to be an* n-dimensional discrete random vector, *and* $p(\mathbf{x}_i) = P_{\mathbf{X}}(\{\mathbf{x}_i\})$ *is called the* probability function *of* $\mathbf{X}$.

(ii) Let $F_{X_1 \cdots X_n}$ *be the distribution function of an* n*-dimensional random vector* $\mathbf{X} = (X_1, \ldots, X_n)'$. *For any* $(x_1, \ldots, x_n)' \in \mathbf{R}^n$, *if there exists a nonnegative* $\mathcal{B}^n$*-measurable function* $f_{X_1 \cdots X_n}$ *on* $\mathbf{R}^n$ *such that*

$$F_{X_1 \cdots X_n}(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1 \cdots X_n}(t_1, \ldots, t_n) dt_1 \cdots dt_n,$$

then $\mathbf{X}$ *is said to be an* n-dimensional continuous random vector, *and* $f_{X_1 \cdots X_n}(\cdot)$ *is called the* n-dimensional joint probability density function *of* $\mathbf{X}$. *Here, in view of the Radon-Nikodym theorem (see Theorem A.4),* $f_{X_1 \cdots X_n}(\cdot)$ *is uniquely determined a.e. on* $\mathbf{R}^n$.

Let us see an important example of an n-dimensional continuous distribution.

n-dimensional normal distribution. If $\mathbf{X} = (X_1, \ldots, X_n)'$ has the n-dimensional joint probability density function

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}, \tag{2.30}$$

then we say that $\mathbf{X}$ has an *n-dimensional normal distribution* and denote this as $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, where $\mathbf{x} = (x_1, \ldots, x_n)' \in \mathbf{R}^n$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$, and $\Sigma = \{\sigma_{ij}\}$ $(i, j = 1, \ldots, n)$ is an $n \times n$ positive definite matrix.

Plural events are statistically independent if the probability of any one of them is unaffected by the occurrence of the others. Similarly, independence of random variables is defined as follows:

Definition 2.7 *Random variables* $X_1, \ldots, X_n$ *with the joint distribution function* $F_{X_1 \cdots X_n}$ *are said to be* mutually independent *if*

$$F_{X_1 \cdots X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n) \tag{2.31}$$

for any $x_i \in \mathbf{R}$ $(i = 1, 2, \ldots, n)$, *where* F_{X_i} *is the marginal distribution function of* X_i.

If $X_1, \ldots, X_n$ has the joint probability density function $f_{X_1 \cdots X_n}$, each X_i has the probability density function f_{X_i}, respectively, and if $X_1, \ldots, X_n$ are mutually independent, then we have

$$\begin{aligned} F_{X_1 \cdots X_n}(x_1, \ldots, x_n) &= F_{X_1}(x_1) \cdots F_{X_n}(x_n) \\ &= \int_{-\infty}^{x_1} f_{X_1}(t_1) dt_1 \cdots \int_{-\infty}^{x_n} f_{X_n}(t_n) dt_n \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1}(t_1) \cdots f_{X_n}(t_n) dt_1 \cdots dt_n. \end{aligned}$$

Hence, from Definition 2.6 (ii),

$$f_{X_1\cdots X_n}(t_1,\ldots,t_n)=\prod_{i=1}^{n} f_{X_i}(t_i). \tag{2.32}$$

Conversely, it is easy to see that (2.32) implies (2.31). Thus we have the following.

Theorem 2.3 *Let $\mathbf{X}_n=(X_1,\ldots,X_n)'$ be a continuous random vector. Then $X_1,\ldots,X_n$ are mutually independent if and only if*

$$f_{X_1\cdots X_n}(x_1,\ldots,x_n)=\prod_{i=1}^{n} f_{X_i}(x_i) \tag{2.33}$$

holds for any $(x_1,\ldots,x_n)'\in\mathbf{R}^n$. More precisely, it is sufficient that (2.33) *holds on $\mathbf{R}^n$ a.e.*

2.3 Expectation and Conditional Distribution

Roughly speaking, the subject of probability and statistics consists in the description of some judgments and decisions for the probability distribution P_X of the concerned random variable X. However, since P_X is a function which has uncountable degrees of freedom in general, it is difficult to infer P_X completely from the observation of X. Therefore, we often discuss, not about P_X itself, but about some characteristic quantities of P_X. The most typical characteristic quantities are the expectation (the mean) and the variance.

Let X be a random variable on a probability space $(\Omega,\mathcal{B},P)$. The *expectation* $E(X)$ of X is defined by the integral of X with respect to the probability measure P, that is,

$$E(X)\equiv\int_{\Omega} X dP, \tag{2.34}$$

which is rewritten by the Lebesgue-Stieltjes integral with respect to the distribution function F of X

$$\int_{\mathbf{R}} x dF(x). \tag{2.35}$$

If X is discrete and has the probability function $p(\cdot)$, then

$$E(X)=\sum_i x_i p(x_i), \tag{2.36}$$

or if X is continuous and has the probability density function $f(x)$, then

$$E(X)=\int_{-\infty}^{\infty} x f(x)dx. \tag{2.37}$$

The representations (2.36) and (2.37) are adequate for the substantial understanding of expectation. The expectation (the mean) of X is a characteristic quantity which measures the center of the distribution of X. One of the characteristic quantities which measure the magnitudes of spread or variation around its expectation is the *variance* defined by

$$\begin{aligned} V(X) &\equiv E\left[\{X - E(X)\}^2\right] \\ &= \int_{\mathbf{R}} (x-\mu)^2 dF(x), \quad (\mu = E(X)). \end{aligned} \tag{2.38}$$

In general, the expectation of functions of a vector-valued random variable is defined as follows. Let $\mathbf{X} = (X_1, \ldots, X_n)'$ be an n-dimensional random variable and have the distribution function $F_{\mathbf{X}}(\mathbf{x})$, $\mathbf{x} \in \mathbf{R}^n$, and let $\varphi : \mathbf{R}^n \to \mathbf{R}$ be a $\mathcal{B}^n$-measurable function. Then, the expectation of $\varphi(\mathbf{X})$ is defined by

$$E\{\varphi(\mathbf{X})\} \equiv \int_{\mathbf{R}^n} \varphi(\mathbf{x}) dF_{\mathbf{X}}(\mathbf{x}). \tag{2.39}$$

Moreover, φ may be a vector- or matrix-valued function. If $\varphi : \mathbf{R}^n \to \mathbf{R}^m$ and $\varphi(\mathbf{x}) = (\varphi_1(\mathbf{x}), \ldots, \varphi_m(\mathbf{x}))'$ in the right-hand side of (2.39), then we define

$$E\{\varphi(\mathbf{X})\} \equiv \left[\int_{\mathbf{R}^n} \varphi_1(\mathbf{x}) dF_{\mathbf{X}}(\mathbf{x}), \ldots, \int_{\mathbf{R}^n} \varphi_m(\mathbf{x}) dF_{\mathbf{X}}(\mathbf{x})\right]' \tag{2.40}$$

and in particular, we call $E(\mathbf{X})$ obtained by setting $\varphi(\mathbf{X}) = \mathbf{X}$ in the above equation the *mean vector* of $\mathbf{X}$. Furthermore, $E\{\varphi(\mathbf{X})\}$ with $\varphi(\mathbf{X}) = \{\mathbf{X} - E(\mathbf{X})\}\{\mathbf{X} - E(\mathbf{X})\}'$ is called the *covariance matrix* of $\mathbf{X}$ and is denoted by $V(\mathbf{X})$, and the (i,j)th element of $V(\mathbf{X})$ is called the *covariance* of X_i and X_j and is denoted by $Cov(X_i, X_j)$.

Now, we summarize some fundamental properties of expectation. Since expectation is basically an integral such as (2.35) and (2.39), it satisfies the properties of integrals, so we omit the proofs (Exercise 2.11).

Theorem 2.4 *If random variables X, Y satisfy $P(X = Y) = 1$ and $P(X \leq Y) = 1$, we denote $X = Y$ a.e. and $X \leq Y$ a.e., respectively. Then, we have the following:*

(i) *If $E|X| < \infty$, $E|Y| < \infty$, then for any $a, b \in \mathbf{R}$ we have*

$$E(aX + bY) = aE(X) + bE(Y). \tag{2.41}$$

(ii) *If $X = c$ a.e., then $E(X) = c$, where c is a constant.*

(iii) *If $X = Y$ a.e., then $E(X) = E(Y)$.*

(iv) *If $X \leq Y$ a.e., then $E(X) \leq E(Y)$.*

(v) $|E(X)| \leq E|X|$.

Next, we give some fundamental inequalities related to expectation.

Theorem 2.5 *For random variables X and Y, we have the following:*

(i)

$$E\left(|X+Y|^r\right) \le c_r E\left(|X|^r\right) + c_r E\left(|Y|^r\right), \tag{2.42}$$

where r is a positive number and $c_r = \begin{cases} 1, & (r \le 1), \\ 2^{r-1}, & (r > 1). \end{cases}$

(ii) ***Hölder's inequality.***

$$E\left(|XY|\right) \le \left\{E\left(|X|^r\right)\right\}^{\frac{1}{r}} \left\{E\left(|Y|^s\right)\right\}^{\frac{1}{s}}, \tag{2.43}$$

where r satisfies $r > 1$ and $1/r + 1/s = 1$. In particular, the case that $r = s = 2$ is called Schwarz's inequality.

(iii) ***Jensen's inequality.*** *If g is a convex function and $E(X)$ exists, then*

$$g\left[E(X)\right] \le E\left\{g(X)\right\}. \tag{2.44}$$

(iv) ***Markov's inequality.*** *If g is a nonnegative even function and is monotone nondecreasing on $[0, \infty)$, then for any $a > 0$, it holds that*

$$P\left(|X| \ge a\right) \le \frac{E\left\{g(X)\right\}}{g(a)}. \tag{2.45}$$

PROOF

(i) If $0 < r \le 1$, then $(a+b)^r \le a^r + b^r$ holds for arbitrary $a, b \ge 0$. Therefore, if we take $a = |X|$ and $b = |Y|$, we have $|X+Y|^r \le (|X| + |Y|)^r \le |X|^r + |Y|^r$. On the other hand, if $r > 1$, x^r is convex on $x \ge 0$, so we have $\{(|X| + |Y|)/2\}^r \le (|X|^r + |Y|^r)/2$. Hence, $|X+Y|^r \le (|X| + |Y|)^r \le 2^{r-1}\left(|X|^r + |Y|^r\right)$ and the assertion follows from Theorem 2.4 (iv).

(ii) For any positive numbers $a, b > 0$, we have $\frac{a^r}{r} + \frac{b^s}{s} \ge ab$. Put $A \equiv \left\{E\left(|X|^r\right)\right\}^{\frac{1}{r}}$ and $B \equiv \left\{E\left(|Y|^s\right)\right\}^{\frac{1}{s}}$. In the case that $A = 0$ or $B = 0$, it is seen that $XY = 0$, a.e. Therefore, from Theorem 2.4 (ii), we have $E|XY| = 0$, so the assertion holds. Furthermore, in the case that $A = \infty$ or $B = \infty$, the inequality evidently holds. Hence, it is sufficient to prove the inequality for the case that $0 < A < \infty$ and $0 < B < \infty$. Let $a \equiv A^{-1}|X|$ and $b \equiv B^{-1}|Y|$, then we have

$$\frac{|XY|}{AB} \le \frac{|X|^r}{rA^r} + \frac{|Y|^s}{sB^s}. \tag{2.46}$$

Taking the expectation of both sides leads to

$$\frac{E\left(|XY|\right)}{AB} \le \frac{1}{r} + \frac{1}{s} = 1. \tag{2.47}$$

(iii) For fixed a, put $M = \sup_{s<a} \frac{g(a)-g(s)}{a-s}$. Then, for all t, we have

$$g(t) \ge g(a) + M(t-a). \tag{2.48}$$

Indeed, if $t = a$ the equality holds. If $t < a$, from the definition of M, the inequality evidently holds. If $a < t$, the inequality follows from the fact that for $s < a < t$, the convex function g satisfies $\frac{g(a)-g(s)}{a-s} \leq \frac{g(t)-g(a)}{t-a}$. Put $t = X$ and $a = E(X)$, then we have

$$g(X) \geq g\{E(X)\} + M\{X - E(X)\}. \tag{2.49}$$

Taking the expectations of both sides leads to the assertion.

(iv) For an arbitrary $a > 0$, from the assumption on $g(x)$ it follows that

$$\begin{aligned} g(a)\chi_{\{|X|\geq a\}} &\leq g(|X|)\chi_{\{|X|\geq a\}} \\ &\leq g(|X|) = g(X). \end{aligned} \tag{2.50}$$

Using Theorem 2.4 (iv) and taking the expectation of the above inequality leads to

$$E\left[g(a)\chi_{\{|X|\geq a\}}\right] \leq E\{g(X)\}. \tag{2.51}$$

The assertion follows from the fact $E\left[g(a)\chi_{\{|X|\geq a\}}\right] = g(a)P(|X| \geq a)$. □

When we argue about the probability distribution, we often find that it is more convenient that we consider the Fourier transform of the probability distribution than the probability distribution itself. Let X be a random variable. For each $t \in \mathbf{R}$, the function defined by

$$\phi_X(t) = E\left\{e^{itX}\right\}, \quad (i = \sqrt{-1}) \tag{2.52}$$

is called the *characteristic function* of X. Similarly, the characteristic function of a vector-valued random variable is defined by

$$\phi(\mathbf{t}) = E\left\{e^{i\mathbf{t}'\mathbf{X}}\right\}, \quad (\mathbf{t} \in \mathbf{R}^n). \tag{2.53}$$

We will later introduce techniques by use of characteristic function. Here first, we list the expectations, variances and characteristic functions of typical distributions in Table 2.1 (see also Exercise 2.12).

Table 2.1 *The probability (density) functions, expectations, variances and characteristic functions of typical distributions.*

Name Symbol	The probability (density) function	Expectation Variance	Characteristic function
Binomial $B(n,\theta)$	$\begin{pmatrix} n \\ x \end{pmatrix}\theta^x(1-\theta)^{n-x}$, $x=0,1,\ldots,n$, $0\le\theta\le 1$	$n\theta$ $n\theta(1-\theta)$	$\left\{\theta e^{it}+(1-\theta)\right\}^n$
Poisson $P_o(\lambda)$	$e^{-\lambda}\lambda^x(x!)^{-1}$, $x=0,1,2,\ldots,$ $\lambda>0$	λ λ	$e^{\left\{-\lambda(1-e^{it})\right\}}$
Uniform $U(a,b)$	$\frac{1}{b-a}$, $x\in(a,b)$	$\frac{a+b}{2}$ $\frac{(b-a)^2}{12}$	$\frac{e^{ibt}-e^{iat}}{it(b-a)}$
Exponential $Exp(\theta,\mu)$	$\theta e^{-\theta(x-\mu)}$, $x\ge\mu$, $0<\theta$, $-\infty<\mu<\infty$	$\mu+\frac{1}{\theta}$ $\frac{1}{\theta^2}$	$e^{i\mu t}\left(1-\frac{it}{\theta}\right)^{-1}$
Normal $N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$, $x\in\mathbf{R}$, $\mu\in\mathbf{R}$, $\sigma>0$	μ σ^2	$e^{i\mu t-\frac{1}{2}\sigma^2t^2}$
Cauchy $C_a(\mu,\alpha)$	$\frac{1}{\pi\alpha}\left\{1+\frac{(x-\mu)^2}{\alpha^2}\right\}^{-1}$, $x\in\mathbf{R}$, $\mu\in\mathbf{R}$, $\alpha>0$	Does not exist Does not exist	$e^{i\mu t-\alpha\lvert t\rvert}$
Logistic $LG(a,b)$	$\frac{\exp\left(\frac{x-b}{a}\right)}{a\left\{1+\exp\left(\frac{x-b}{a}\right)\right\}^2}$, $x\in\mathbf{R}$, $b\in\mathbf{R}$, $a>0$	b $\frac{a^2\pi^2}{3}$	omitted
Double exponential $DE(a,b)$	$\frac{1}{2a}e^{-\lvert(x-b)/a\rvert}$, $x\in\mathbf{R}$, $b\in\mathbf{R}$, $a>0$	b $2a^2$	$e^{ibt}\left\{1+(at)^2\right\}^{-1}$
Multivariate normal $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$	$\frac{\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}}{(2\pi)^{\frac{n}{2}}\sqrt{\lvert\boldsymbol{\Sigma}\rvert}}$, $\mathbf{x}\in\mathbf{R}^n$, $\boldsymbol{\mu}\in\mathbf{R}^n$, $\boldsymbol{\Sigma}$ is positive definite	$\boldsymbol{\mu}$ $\boldsymbol{\Sigma}$	$e^{i\mu'\mathbf{t}-\frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}}$

Next, we consider conditional distributions and conditional expectations.

Definition 2.8 **(i)** *Let a random vector (X, Y) be discrete and have the joint probability function $p_{X,Y}(x, y)$, and let X and Y have the probability functions $p_X(x)$ and $p_Y(y)$, respectively. Then we call*

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x, y)}{p_Y(y)} \tag{2.54}$$

the conditional probability function *of X given $Y = y$.*

(ii) *Let a random vector (X, Y) be continuous and have the joint probability density function $f_{X,Y}(x, y)$, and let X and Y have the probability density functions $f_X(x)$ and $f_Y(y)$, respectively. Then we call*

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \tag{2.55}$$

the conditional probability density function *of X given $Y = y$.*

(iii) *Under the assumptions (i) and (ii) above, we define*

$$E(X|Y = y) = \begin{cases} \sum_x x\ p_{X|Y}(x|y), & ((i) \text{ discrete case}), \\ \int_{-\infty}^{\infty} x\ f_{X|Y}(x|y)dx, & ((ii) \text{ continuous case}) \end{cases} \tag{2.56}$$

and call it the conditional expectation *of X given $Y = y$.*

If X and Y are n-dimensional vectors $\mathbf{X}$ and $\mathbf{Y}$, respectively, we can similarly define the conditional probability (density) function as in (i) and (ii). Furthermore, from (iii) we can also define $E(\mathbf{X}|\mathbf{Y} = \mathbf{y})$, $\mathbf{y} \in \mathbf{R}^n$ and $E\{\varphi(\mathbf{X})|\mathbf{Y} = \mathbf{y}\}$ for a measurable function $\varphi(\cdot)$.

Since $E(X|Y = y)$ is a function of y, we can write it as $h(y)$. Then $h(Y)$ becomes a random variable. Henceforth we denote it as $E(X|Y)$. Moreover, we similarly define $E\{\varphi(\mathbf{X})|\mathbf{Y}\}$ and so on.

Theorem 2.6 *Let X, Y, Z be random variables with $E|X| < \infty$, $E|Y| < \infty$, and a, b be arbitrary real constants. Then, we have*

(i) *If $X \geq 0$ a.e., then $E(X|Y) \geq 0$ a.e.,*

(ii) *$E(1|Y) = 1$ a.e.,*

(iii) *$E(aX + bY|Z) = aE(X|Z) + bE(Y|Z)$ a.e.,*

(iv) *$E\{E(X|Y)\} = E(X)$,*

(v) *$E(XY|Y) = YE(X|Y)$ a.e.*

PROOF

We only check the properties (iv) and (v) for the discrete case, which characterize conditional expectation (see Exercises 2.15 and 2.16).

(iv) First, we have

$$E\{E(X|Y)\} = \sum_y \left\{ \sum_x x p_{X|Y}(x|y) \right\} p_Y(y)$$
$$= \sum_y \left\{ \sum_x x p_{X,Y}(x,y) \right\}. \tag{2.57}$$

From the assumption $E|X| < \infty$, we can exchange the order of summations, so (2.57) becomes

$$\sum_x x \sum_y p_{X,Y}(x,y) = \sum_x x p_X(x) = E(X). \tag{2.58}$$

Hence, the assertion follows.

(v) Let $y, \tilde{y}$ be possible arbitrary values of Y. Then, we have

$$\begin{aligned} E\{XY|Y=y\} &= \sum_x \sum_{\tilde{y}} x\tilde{y} P(X=x, Y=\tilde{y}|Y=y) \\ &= \sum_x \sum_{\tilde{y}} x\tilde{y} P(X=x, Y=\tilde{y}, Y=y)/P(Y=y) \\ &= \sum_x xy P(X=x, Y=y)/P(Y=y) \\ &= y \sum_x x p_{X|Y}(x|y) \\ &= yE\{X|Y=y\}, \end{aligned} \tag{2.59}$$

so the assertion follows. □

Although we defined in Definition 2.8 the conditional distribution in discrete and continuous cases, separately, we can make the unified measure theoretic definition in terms of the Radon-Nikodym theorem (Theorem A.4).

Let X, Y be random variables on $(\Omega, \mathcal{A}, P)$. If we let $\mathcal{A}_Y \equiv \left\{Y^{-1}(B) : B \in \mathcal{B}\right\}$, then it becomes a σ-field. We say that this is the *σ-field generated by Y*. If φ is a measurable function, then the conditional expectation $E\{\varphi(X)|Y\}$ is defined as follows. For $A \in \mathcal{B}$, define

$$Q_Y(A) \equiv \int_{Y^{-1}(A)} \varphi(X) dP, \tag{2.60}$$

then for any $A \in \mathcal{B}$ satisfying $P_Y(A)(\equiv P\{Y(\omega) \in A\}) = 0$, we have $Q_Y(A) = 0$, therefore, from the Radon-Nikodym theorem, there exists an $\mathcal{A}_Y$-measurable function g satisfying

$$\int_{Y^{-1}(A)} \varphi(X) dP = \int_A g dP_Y, \tag{2.61}$$

where g is unique a.e. We write this g as $E\{\varphi(X)|Y\}$ or $E\{\varphi(X)|\mathcal{A}_Y\}$ and

call the conditional expectation given Y. In particular, if we take $\varphi = \chi_B$, $B \in \mathcal{A}$ (indicator function of B), then we write $E\{\chi_B|Y\}$ as $P(B|Y)$ or $P(B|\mathcal{A}_Y)$ and call the conditional probability of B given Y. The conditional expectations for discrete and continuous cases defined in Definition 2.8 satisfy the relationship (2.61) and agree with the measure theoretic definition above.

2.4 Convergence and Central Limit Theorems

Let $X_0, X_1, \ldots, X_n$ be a sequence of random variables on a probability space $(\Omega, \mathcal{A}, P)$. There are several different notions of convergence for them. We now define the following four types of convergence of the sequence $X_1, \ldots, X_n$.

Definition 2.9 *(i) A sequence $\{X_n, n = 1, 2, \ldots\}$ of random variables is said to* converge almost surely *to a random variable X_0 if*

$$P\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X_0(\omega)\right\} = 1.$$

We denote this convergence as $X_n \xrightarrow{a.s.} X$.

(ii) A sequence $\{X_n, n = 1, 2, \ldots\} \subset L^p(\Omega) \equiv \{Y : E(|Y|^p) < \infty\}$ $(p \geq 1)$ is said to converge in pth mean *to $X_0 \in L^p(\Omega)$ if*

$$\lim_{n\to\infty} E(|X_n - X_0|^p) = 0.$$

We denote this convergence as $X_n \xrightarrow{L^p} X_0$.

(iii) A sequence $\{X_n, n = 1, 2, \ldots\}$ of random variables is said to converge in probability *to a random variable X_0 if, for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} P(|X_n - X_0| \geq \varepsilon) = 0.$$

We denote this convergence as $X_n \xrightarrow{p} X_0$.

Henceforth, if $X_n \xrightarrow{p} 0$ we write $X_n = o_p(1)$. We say that a sequence $\{X_n, n = 1, 2, \ldots\}$ of random variables is said to be bounded in probability*, denoted by $X_n = O_p(1)$, if for every $\varepsilon > 0$ there exists a $\delta(\varepsilon) > 0$ such that*

$$P\{|X_n| \geq \delta(\varepsilon)\} < \varepsilon \quad \text{for all } n \in \mathbf{N}.$$

(iv) Let $\{X_n, n = 0, 1, 2, \ldots\}$ be a sequence of random variables with the corresponding distribution functions F_n. The sequence $\{X_n, n = 1, 2, \ldots\}$ is said to converge in distribution *to X_0 if*

$$\lim_{n\to\infty} F_n(x) = F_0(x)$$

at all continuity points x of $F_0(\cdot)$. We denote this convergence as $X_n \xrightarrow{d} X_0$ (or $F_n \xrightarrow{d} F_0$, or $X_n \xrightarrow{d} F_0$).

Interrelations of the above four convergence concepts are stated as follows:

Theorem 2.7 *(i) If* $X_n \xrightarrow{L^p} X_0$ *then* $X_n \xrightarrow{p} X_0$.

(ii) If $X_n \xrightarrow{p} X_0$ *then* $X_n \xrightarrow{d} X_0$.

(iii) If $X_n \xrightarrow{a.s.} X_0$ *then* $X_n \xrightarrow{p} X_0$.

PROOF

(i) Setting $g(x) = |x|^p$ in Markov's inequality (Theorem 2.5 (iv)), we have for every $\varepsilon > 0$

$$P(|X_n - X_0| \geq \varepsilon) \leq \frac{E(|X_n - X_0|^p)}{\varepsilon^p}.$$

The result follows from $E(|X_n - X_0|^p) \to 0$ $(n \to \infty)$.

(ii) For every $\varepsilon > 0$, we obtain

$$\begin{aligned} F_n(x) &= P\{X_n \leq x\} \\ &= P\{X_n \leq x, X_0 > x + \varepsilon\} + P\{X_n \leq x, X_0 \leq x + \varepsilon\} \\ &\leq P\{|X_n - X_0| \geq \varepsilon\} + P\{X_0 \leq x + \varepsilon\} \\ &= P\{|X_n - X_0| \geq \varepsilon\} + F_0(x + \varepsilon). \end{aligned} \tag{2.62}$$

Similarly,

$$F_0(x - \varepsilon) - P\{|X_n - X_0| \geq \varepsilon\} \leq F_n(x). \tag{2.63}$$

Hence, since $P\{|X_n - X_0| \geq \varepsilon\} \to 0$ as $n \to \infty$,

$$F_0(x - \varepsilon) \leq \liminf_{n\to\infty} F_n(x) \leq \limsup_{n\to\infty} F_n(x) \leq F_0(x + \varepsilon).$$

If x is a continuity point of $F_0(\cdot)$, then $F_0(x - \varepsilon)$ and $F_0(x + \varepsilon)$ both converge to $F_0(x)$ as $\varepsilon \to 0$, which implies that $\lim_{n\to\infty} F_n(x) = F_0(x)$.

(iii) Let

$$S_{k,l} = \bigcap_{n=l}^{\infty} \left\{\omega : |X_n(\omega) - X_0(\omega)| < \frac{1}{k}\right\}.$$

Then $\{\omega : \lim_{n\to\infty} X_n(\omega) = X_0(\omega)\} = \{\omega : \text{for all } 1/k, \text{ there exists an } l \text{ such that } |X_n(\omega) - X_0(\omega)| < 1/k \text{ for all } n \geq l\} = \bigcap_{k=1}^{\infty} \bigcup_{l=1}^{\infty} S_{k,l}$. Thus $X_n \xrightarrow{a.s.} X_0$ is equivalent to $P(\bigcap_{k=1}^{\infty} \bigcup_{l=1}^{\infty} S_{k,l}) = 1$. Note that the sequence of sets $S_{k,l}$ monotonously increases to the set S_k as $l \to \infty$ and the sequence of sets $S_k = \bigcup_{l=1}^{\infty} S_{k,l}$ monotonously decreases to the set $S = \bigcap_{k=1}^{\infty} \bigcup_{l=1}^{\infty} S_{k,l}$ as $k \to \infty$. From Theorem 2.1 (vii) and (viii),

$$P(S_k) = \lim_{l\to\infty} P(S_{k,l}), \tag{2.64}$$

$$P(S) = \lim_{k\to\infty} P(S_k). \tag{2.65}$$

Because the sequence of sets S_k decreases, from (2.65) it follows that

$$P(S_k) = 1 \quad (k = 1, 2, \dots). \tag{2.66}$$

For any given $\varepsilon > 0$,

$$\begin{aligned}\{\omega : |X_n(\omega) - X_0(\omega)| \geq \varepsilon\} &\subset \{\omega : |X_n(\omega) - X_0(\omega)| \geq 1/k\} \\ &\subset \Omega \cap S_{k,l}^c\end{aligned}$$

for all $1/k < \varepsilon$ and all $n \geq l$. By letting $n \to \infty$,

$$P\{\omega : |X_n(\omega) - X_0(\omega)| \geq \varepsilon\} \to 0 \quad (n \to \infty)$$

from Theorem 2.1 (iii), (2.64) and (2.66). □

Independence of random variables was defined in Section 2.2. In addition, if random variables are mutually independent and have the same distribution, we say that they are independent and identically distributed, which we abbreviate as i.i.d.

The following theorem is known as the *weak law of large numbers* for the average of i.i.d. random variables.

Theorem 2.8 *Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with mean $E(X_j) = \mu$ and variance $V(X_j) = \sigma^2$, and let $\bar{X}_n = n^{-1} \sum_{j=1}^n X_j$. Then $\bar{X}_n \xrightarrow{p} \mu$.*

PROOF Note that $\bar{X}_n - \mu = n^{-1} \sum_{j=1}^n (X_j - \mu)$ and Exercise 2.10. Since $E\{(X_i - \mu)(X_j - \mu)\} = E(X_i - \mu)E(X_j - \mu) = 0$ $(i \neq j)$,

$$\begin{aligned}E\{(\bar{X}_n - \mu)^2\} &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n E\{(X_i - \mu)(X_j - \mu)\} \\ &= \frac{1}{n^2} \sum_{i=1}^n E\{(X_i - \mu)^2\} + \frac{1}{n^2} \sum_{i \neq j} E\{(X_i - \mu)(X_j - \mu)\} \\ &= \frac{1}{n^2} \sum_{i=1}^n E\{(X_i - \mu)^2\} \\ &= \frac{\sigma^2}{n} \to 0, \quad (n \to \infty),\end{aligned}$$

which implies, from Theorem 2.7 (i), $\bar{X}_n \xrightarrow{p} \mu$. □

Remark 2.1 *Under the assumptions in Theorem 2.8, a stronger result*

$$X_n \xrightarrow{a.s.} \mu$$

holds. This result is known as the strong law of large numbers. *The proof is given in, e.g., Ash and Doléans-Dade (2000, p.242).*

Next, the following theorem states the relation between the convergence of distribution functions and that of their characteristic functions. The proof is given in, e.g., Ash and Doléans-Dade (2000, p.304).

Theorem 2.9 *Let $X_0, X_1, X_2, \dots$ be a sequence of random variables, and let ϕ_n be the characteristic function of X_n $(n = 0, 1, 2, \dots)$. If for any $t \in \boldsymbol{R}$ $\phi_n(t) \to \phi_0(t)$, then $X_n \xrightarrow{d} X_0$.*

The following result states that the distribution of sum of i.i.d. random variables is approximately normal even if the distribution of each random variable is not normal.

Theorem 2.10 (Central Limit Theorem) *Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with mean $E(X_j) = \mu$ and variance $V(X_j) = \sigma^2$, and let $\bar{X}_n = n^{-1} \sum_{j=1}^n X_j$. Then*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1) \quad (\textit{standard normal distribution})$$

PROOF Let $S_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma = n^{-1/2} \sum_{j=1}^n Z_j$, where $Z_j = (X_j - \mu)/\sigma$. From Theorem 2.9, it is sufficient to prove that for all $t \in \boldsymbol{R}$ the characteristic function $\phi_{S_n}(t)$ of S_n converges to $\phi(t) = e^{-t^2/2}$ which is the characteristic function of the standard normal distribution. Since $\{Z_1, \dots, Z_n\}$ is a sequence of i.i.d. random variables with mean 0 and variance 1, we have

$$\begin{aligned}
\phi_{S_n}(t) &= E\{e^{itS_n}\} \\
&= E\left\{e^{itn^{-1/2}\sum_{j=1}^n Z_j}\right\} \\
&= \prod_{j=1}^n E\left\{e^{i\frac{t}{\sqrt{n}}Z_j}\right\} \quad \text{(Exercise 2.13)} \\
&= \prod_{j=1}^n \phi_{Z_j}\left(\frac{t}{\sqrt{n}}\right) \\
&= \left\{\phi_{Z_1}\left(\frac{t}{\sqrt{n}}\right)\right\}^n .
\end{aligned} \tag{2.67}$$

We expand $\phi_{Z_1}(t/\sqrt{n})$ in a Taylor series around 0, from Exercise 2.14, leading to

$$\phi_{Z_1}\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + \frac{t^2}{2n} R\left(\frac{t}{\sqrt{n}}\right), \tag{2.68}$$

where

$$\begin{aligned} R\left(\frac{t}{\sqrt{n}}\right) &= \phi''_{Z_1}\left(\frac{ht}{\sqrt{n}}\right) - \phi''_{Z_1}(0) \\ &= -\left\{E\left(Z_1^2 e^{\frac{ihtZ_1}{\sqrt{n}}}\right) - E(Z_1^2)\right\} \\ &= -E\left\{Z_1^2\left(e^{\frac{ihtZ_1}{\sqrt{n}}} - 1\right)\right\} \quad (0 < h < 1). \end{aligned} \tag{2.69}$$

Note that

$$\left|Z_1^2\left(e^{\frac{ihtZ_1}{\sqrt{n}}} - 1\right)\right| \le 2Z_1^2, \quad (E(Z_1^2) = 1).$$

We can see that from Lebesgue's convergence theorem (Theorem A.1), (2.68) is written as

$$\phi_{Z_1}\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right).$$

Recalling (2.67), we have

$$\begin{aligned} \phi_{S_n}(t) &= \left\{1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right\}^n \\ &= \left(1 - \frac{t^2}{2n}\right)^n + n \times o\left(\frac{t^2}{n}\right) \\ &\to e^{-\frac{t^2}{2}}, \quad (n \to \infty) \end{aligned}$$

which completes the proof. □

Remark 2.2 *Similarly, we can extend the results in this section to the case of random vectors. Let* $\mathbf{X}_1, \mathbf{X}_2, \ldots$ *be a sequence of i.i.d. m-dimensional random vectors with mean vector* $E(\mathbf{X}_1) = \boldsymbol{\mu}$ *and variance matrix* $V(\mathbf{X}_1) = \Sigma$*, and let* $\bar{\mathbf{X}}_n = n^{-1}\sum_{j=1}^n \mathbf{X}_j$*. Then*

$$\bar{\boldsymbol{X}}_n \xrightarrow{p} \boldsymbol{\mu} \tag{2.70}$$

$$\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \Sigma). \tag{2.71}$$

Exercises

2.1 Show that

$$\left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c \quad \text{and} \quad \left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c. \tag{2.72}$$

2.2 Let $\mathcal{A}$ be a σ-field, and $A_n \in \mathcal{A}$ for all $n \in \mathbf{N}$. Show that the following statements hold:

(B4) $\emptyset \in \mathcal{A}$.

(B5) $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$.
(B6) $\bigcap_{i=1}^{n} A_i \in \mathcal{A}$.
(B7) $\bigcap_{i=1}^{\infty} A_i \in \mathcal{A}$.
(B8) $\limsup_{k\to\infty} A_k \in \mathcal{A}$.
(B9) $\liminf_{k\to\infty} A_k \in \mathcal{A}$.

Here, we define $\limsup_{k\to\infty} A_k \equiv \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$ and $\liminf_{k\to\infty} A_k \equiv \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$.

2.3 We denote the collection of all subsets of Ω by $\mathcal{S}(\Omega)$. Show that $\mathcal{S}(\Omega)$ is a σ-field of Ω.

2.4 Let $A \subset \Omega$. Show that $\sigma[A] \equiv \{\emptyset, A, A^c, \Omega\}$ is a σ-field of Ω.

2.5 Suppose that $\mathcal{B}$ is a σ-field of Ω and A is a nonempty subset of Ω. Show that the collection of events

$$\mathcal{B} \cap A \equiv \{B \cap A : B \in \mathcal{B}\} \tag{2.73}$$

is a σ-field of A.

2.6 Show that we can replace intervals of the form $(a, b]$ in the definition of the family of Borel sets by other classes of intervals, for instance,

all closed intervals,
all intervals $[a, b)$, $a, b \in \mathbf{R}$,
all intervals (a, ∞), $a \in \mathbf{R}$,
all intervals $[a, \infty)$, $a \in \mathbf{R}$,
all intervals $(-\infty, b)$, $b \in \mathbf{R}$,
all intervals $(-\infty, b]$, $b \in \mathbf{R}$.

2.7 Let the distribution of X be $N(\mu, \sigma^2)$ and $Y = aX + b$. Show that the distribution of Y is $N(a\mu + b, a^2\sigma^2)$.

2.8 (i) Let X_1, X_2 be independent, each distributed according to the Cauchy distribution (2.24). Find the distribution of $a_1X_1 + a_2X_2$. (ii) Use (i) to prove (by induction) that if $X_1, \ldots, X_n$ are independent, each distributed according to (2.24), then S_n is also distributed according to (2.24).

2.9 Show that the density (2.25) is symmetric about zero and unimodal.

2.10 Let X and Y be independent random variables with $E|X| < \infty$ and $E|Y| < \infty$. Show that $E|XY| < \infty$ and

$$E(XY) = E(X)E(Y). \tag{2.74}$$

2.11 Prove the statements (i)-(v) of Theorem 2.4.

2.12 Compute the expectation, variance and characteristic function of each distribution in Table 2.1.

2.13 Let $X_1, \ldots, X_n$ be independent random variables, and let $S_n = X_1 + \cdots + X_n$. Show that the characteristic function of S_n is the product of the characteristic functions of the X_i.

2.14 Let $E|X|^n < \infty$ for some positive integer n. Show that the nth derivative of $\phi_X(t)$ exists and is continuous on $\mathbf{R}$, and

$$\phi_X^{(n)}(t) \equiv \frac{d^n}{dt^n}\phi_X(t) = E\left\{(iX)^n e^{itX}\right\}, \tag{2.75}$$

and in particular,

$$E(X^n) = \frac{1}{i^n}\phi_X^{(n)}(0). \tag{2.76}$$

2.15 Prove the statements of (i)- (iii) of Theorem 2.6 for the discrete case.

2.16 Prove the statements of (i)- (v) of Theorem 2.6 for the continuous case.

2.17 Suppose that $X_0, X_1, \ldots$ is a sequence of random variables defined on the probability space $(\Omega, \mathcal{A}, P)$. If $g : (\mathbf{R}, \mathcal{B}^1) \to (\mathbf{R}, \mathcal{B}^1)$ is a continuous function, then show that

(1) If $X_n \xrightarrow{a.s.} X_0$ then $g(X_n) \xrightarrow{a.s.} g(X_0)$.

(2) If $X_n \xrightarrow{p} X_0$ then $g(X_n) \xrightarrow{p} g(X_0)$.

(3) If $X_n \xrightarrow{d} X_0$ then $g(X_n) \xrightarrow{d} g(X_0)$.

2.18 (Slutsky's lemma) Let $\{X_n, n = 0, 1, 2 \ldots\}$ and $\{Y_n, n = 0, 1, 2 \ldots\}$ be sequences of random variables. If $X_n \xrightarrow{d} X_0$ and $Y_n \xrightarrow{p} c$ for a constant c as $n \to \infty$, then show that

(1) $X_n + Y_n \xrightarrow{d} X_0 + c$.

(2) $X_n Y_n \xrightarrow{d} cX_0$.

(3) $X_n/Y_n \xrightarrow{d} X_0/c$, $(c \neq 0)$.

2.19 Let $\{X_n, n = 0, 1, 2 \ldots\}$ be a sequence of random variables. If $X_n \xrightarrow{d} c$ for some constant c, then show that $X_n \xrightarrow{p} c$.

CHAPTER 3

Statistical Inference

In this chapter we give a compact survey of statistical inference based on a random sample. Concretely speaking, we introduce the idea of sufficient statistics and construct minimum variance unbiased estimators by using them. Furthermore, we derive the Cramér-Rao bound which gives a lower bound for the variance of unbiased estimators, and discuss the efficient estimator which achieves it. Heretofore, we discussed approaches in random samples with fixed size n. In these cases it is often difficult to obtain the exact distribution of estimators from samples of size n. However, if the sample sizes are "large", then the structure of estimation becomes clearer and simpler. Therefore, we define "goodness" in asymptotic inference theory and describe the asymptotic optimality of estimators.

3.1 Sufficient Statistics

Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables on probability space $(\Omega, \mathcal{A}, P)$, which are mutually independent and have an identical probability distribution. Denote $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ and henceforth, we call it a random sample of size n. Write the probability distribution of $\mathbf{X}$ as $P_{\mathbf{X}}(B) \equiv P\left\{\mathbf{X}^{-1}(B)\right\}$, $B \in \mathcal{B}^n$. Then, since X_i's are mutually independent in this case, it is represented by the product of probability distributions P_{X_i} of each X_i, namely represented in the form $P_{\mathbf{X}} = P_{X_1} \times \cdots \times P_{X_n}$. The probability space induced by $\mathbf{X}$ becomes $(\mathbf{R}^n, \mathcal{B}^n, P_{\mathbf{X}})$. We call this the *$n$-dimensional sample space* and henceforth, for simplicity, denote it by $(\mathcal{X}, \mathcal{G}, \mathbb{P})$. If $\mathbb{P}$ is described in the form $\mathbb{P}_\theta$, we call θ a *parameter* and a certain set Θ in which θ lies is called the *parameter space*. Henceforth, if the probability distribution of $\mathbf{X}$ is $\mathbb{P}_\theta$, then we denote $\mathbf{X} \sim \mathbb{P}_\theta$ and $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. We define a *statistic* $T = T(\mathbf{X})$ by a function $T : (\mathcal{X}, \mathcal{G}) \to \left(\mathbf{R}^k, \mathcal{B}^k\right)$ that is $\mathcal{G}$-measurable and independent of θ. One of the most fundamental statistics is sufficient statistic, which is defined as follows:

Definition 3.1 *Let $\mathbf{X} \sim \mathbb{P}_\theta$. If for any $A \in \mathcal{G}$, the conditional probability $\mathbb{P}_\theta(A|T)$ given a statistic $T = T(\mathbf{X})$ is independent of θ $(\in \Theta)$, then T is said to be a* sufficient statistic *for the family $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$.*

In general, checking sufficiency directly is difficult, since we need to compute the conditional distribution. Fortunately, a simple necessary and sufficient criterion for a statistic to be sufficient is available.

Theorem 3.1 (Factorization theorem) *Let* $\mathbf{X}$ *be a random sample of size* n *on a probability space* $(\mathcal{X}, \mathcal{G}, \mathbb{P}_\theta)$, $\theta \in \Theta$ *which has the probability density function (probability function if* $\mathbf{X}$ *is discrete)* $f_\theta(\mathbf{x})$, $x \in \mathbf{R}^n$. *Then,* $T : \mathbf{X} \to \mathbf{R}^k$ *is a sufficient statistic if and only if there exist a nonnegative* $\mathcal{B}^k$*-measurable function* g_θ *on* $\mathbf{R}^k$ *and a nonnegative* $\mathcal{G}$*-measurable function* h *on* $\mathcal{X}$ *such that*

$$f_\theta(\mathbf{x}) = g_\theta\{T(\mathbf{x})\}\, h(\mathbf{x}), \tag{3.1}$$

where in the right hand of (3.1), the first factor may depend on θ *but depends on* $\mathbf{x}$ *only through* $T(\mathbf{x})$, *while the second factor is independent of* θ.

PROOF

We prove this only for the case that $\mathbf{X}$ is discrete (see, for example, Lehmann (1986) for the general case (Exercise 3.1)). Denote the probability function of T as $p_T^\theta(\mathbf{t}) = \mathbb{P}_\theta(T = \mathbf{t})$, $\mathbf{t} \in \mathbf{R}^k$ and the conditional probability function of $\mathbf{X}$ given $T = \mathbf{t}$ as $p_\theta(\mathbf{x}|T = \mathbf{t})$, $\mathbf{x} \in \mathbf{R}^n$. Suppose that (3.1) holds. If $T(\mathbf{x}) \neq \mathbf{t}$, then evidently $p_\theta(\mathbf{x}|T = \mathbf{t}) = 0$, so let $T(\mathbf{x}) = \mathbf{t}$. Then,

$$p_T^\theta(\mathbf{t}) = \mathbb{P}_\theta\{T(\mathbf{x}) = \mathbf{t}\} = \sum_{\mathbf{x}'} f_\theta(\mathbf{x}') = g_\theta(\mathbf{t}) \sum_{\mathbf{x}'} h(\mathbf{x}') \tag{3.2}$$

where $\sum_{\mathbf{x}'}$ is summed over all points $\mathbf{x}'$ with $T(\mathbf{x}') = \mathbf{t}$, and the conditional probability

$$p_\theta(\mathbf{x}|T = \mathbf{t}) = \frac{f_\theta(\mathbf{x})}{p_T^\theta(\mathbf{t})} = \frac{h(\mathbf{x})}{\sum_{\mathbf{x}'} h(\mathbf{x}')} \tag{3.3}$$

is independent of θ. Conversely, if this conditional distribution does not depend on θ and is equal to, say $p(\mathbf{x}|T = \mathbf{t})$, then $f_\theta(\mathbf{x}) = p_T^\theta(\mathbf{t})p(\mathbf{x}|T = \mathbf{t})$, so that (3.1) holds. □

Now, let us see some concrete examples.

Example 3.1 *Let* $X_1, X_2, \ldots, X_n$ *be a sequence of random variables, where* X_j*'s are mutually independent and each of them has probability density* $p_\theta(x) = \theta^x(1-\theta)^{(1-x)}$, $x = 0, 1$ $(0 \leq \theta \leq 1)$, *namely distributed according to* $B(1, \theta)$. *Henceforth let us denote the above as* $\{X_j\} \sim$ *i.i.d.* $B(1, \theta)$ *for simplicity. Write* $\mathbf{X} = (X_1, \ldots, X_n)'$, $\mathbf{x} = (x_1, \ldots, x_n)'$, $x_j \in \{0, 1\}$, *then the joint probability function of* $\mathbf{X}$ *is*

$$f_\theta(\mathbf{x}) = \prod_{j=1}^n \theta^{x_j}(1-\theta)^{1-x_j} = \theta^{\sum_{j=1}^n x_j}(1-\theta)^{n-\sum_{j=1}^n x_j}. \tag{3.4}$$

Therefore, taking $T(\mathbf{x}) = \sum_{j=1}^n x_j$, $g_\theta\{T(\mathbf{x})\} = \theta^{T(\mathbf{x})}(1-\theta)^{n-T(\mathbf{x})}$, $h(\mathbf{x}) =$

$\chi_A(\mathbf{x})$, $A = \{0,1\}^n$ *in Theorem 3.1, we see that* $T(\mathbf{X}) = X_1 + \cdots + X_n$ *is a sufficient statistic.*

Example 3.2 *Let* $\{X_j\} \sim$ *i.i.d.* $N(\theta, 1)$, $-\infty < \theta < \infty$. *In this case,* $\mathbf{X} = (X_1, \ldots, X_n)'$ *has the joint probability density function*

$$f_\theta(\mathbf{x}) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n}(x_j - \theta)^2\right\}, \quad \left(\mathbf{x} = (x_1, \ldots, x_n)' \in \mathbf{R}^n\right). \tag{3.5}$$

Taking $T(\mathbf{x}) = n^{-1}\sum_{j=1}^{n} x_j$, *we have*

$$f_\theta(\mathbf{x}) = \exp\left\{nT(\mathbf{x})\theta - \frac{n\theta^2}{2}\right\} \times \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{j=1}^{n} x_j^2\right\}. \tag{3.6}$$

Hence, if we put $g_\theta\{T(\mathbf{x})\} = \exp\left\{nT(\mathbf{x})\theta - n\theta^2/2\right\}$ *and* $h(\mathbf{x}) = (1/2\pi)^{n/2} \exp\left\{-1/2\sum_{j=1}^{n} x_j^2\right\}$ *in Theorem 3.1, it is seen that* $T(\mathbf{X})$ *is a sufficient statistic.*

Example 3.3 *Let* $\{X_j\} \sim$ *i.i.d.* $N(\mu, \sigma^2)$, $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$ *and* $\theta = (\mu, \sigma)^2$. *In this case, the joint probability density function of* $\mathbf{X} = (X_1, \ldots, X_n)'$ *is*

$$\begin{aligned} f_\theta(\mathbf{x}) &= \left(\sqrt{2\pi}\sigma\right)^{-1} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2\right\} \\ &= \left(\sqrt{2\pi}\sigma\right)^{-1} \exp\left[-\frac{n}{2\sigma^2}\left\{s^2 + (\overline{x} - \mu)^2\right\}\right], \end{aligned} \tag{3.7}$$

where $\overline{x} = n^{-1}\sum_{j=1}^{n} x_j$ *and* $s^2 = n^{-1}\sum_{j=1}^{n}(x_j - \overline{x})^2$. *In Theorem 3.1, let* $T(\mathbf{x}) = \left(\overline{x}, s^2\right)'$ *and* $h(\mathbf{x}) = 1$, *and* $g_\theta\{T(\mathbf{x})\}$ *as the right-hand side of (3.7). Then we can see that*

$$T(\mathbf{X}) = \left\{n^{-1}\sum_{j=1}^{n} X_j, n^{-1}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2\right\}' \tag{3.8}$$

is a sufficient statistic.

Now, let us see an implication of sufficient statistics in Example 3.1 with $n = 2$.

First, let $X_1, X_2 \sim$ *i.i.d.* $B(1, \theta)$. We consider $T = T(\mathbf{X}) = X_1 + X_2$ as a statistic. T has the probability function

$$\binom{2}{x}\theta^x(1-\theta)^{2-x}, \quad (x = 0, 1, 2). \tag{3.9}$$

Next, we define new random variables Y_1 and Y_2 as follows:

(i) When $T = 0$, we define $Y_1 = 0$, $Y_2 = 0$.

(ii) When $T = 2$, we define $Y_1 = 1$, $Y_2 = 1$.

(iii) When $T = 1$, we flip a fair coin. If heads, we define $Y_1 = 1$, $Y_2 = 0$ and if tails, we define $Y_1 = 0$, $Y_2 = 1$.

Then, we can show that

$$\text{“The distribution of } \mathbf{Y} = (\mathrm{Y}_1, \mathrm{Y}_2)' \text{ is equivalent to that of } \mathbf{X} = (\mathrm{X}_1, \mathrm{X}_2)'\text{''} \tag{3.10}$$

(Exercise 3.2). This fact implies that we can reconstruct the distribution of $\mathbf{X}$ from T even if we do not know the value of θ. So, we can conclude T includes all the probability information of θ.

Now, we give N times experiments of the new random variable $\mathbf{Y} = (Y_1, Y_2)'$ for $\theta = 0.1, 0.2, \ldots, 0.5$ and $N = 100, 1000, 10000$. The results are listed in Tables 3.1-3.3. From Table 3.3 we can see that the frequencies of $\mathbf{Y} = \mathbf{y}$ of large N are close to the theoretical values $N\theta^{y_1+y_2}(1-\theta)^{2-y_1+y_2}$.

Table 3.1 *The number of* $T(\mathbf{X}) = X_1 + X_2 = t$, $t = 0, 1, 2$, *of* N *trials.*

(θ, N)	$T = 0$	$T = 1$	$T = 2$
$(0.1, 10)$	80	18	2
$(0.1, 100)$	803	188	9
$(0.1, 1000)$	8095	1800	105
$(0.2, 10)$	55	40	5
$(0.2, 100)$	653	306	41
$(0.2, 1000)$	6302	3310	388
$(0.3, 10)$	49	44	7
$(0.3, 100)$	488	415	97
$(0.3, 1000)$	4885	4179	936
$(0.4, 10)$	39	47	14
$(0.4, 100)$	361	482	157
$(0.4, 1000)$	3638	4744	1618
$(0.5, 10)$	28	43	29
$(0.5, 100)$	241	510	249
$(0.5, 1000)$	2477	5033	2490

Table 3.2 *The number of heads and tails of N trials in step (iii).*

(θ, N)	Heads	Tails
$(0.5, 18)$	13	5
$(0.5, 188)$	94	94
$(0.5, 1800)$	934	866
$(0.5, 40)$	19	21
$(0.5, 306)$	150	156
$(0.5, 3310)$	1617	1693
$(0.5, 44)$	23	21
$(0.5, 415)$	181	234
$(0.5, 4179)$	2159	2020
$(0.5, 47)$	23	24
$(0.5, 482)$	218	264
$(0.5, 4744)$	2393	2351
$(0.5, 43)$	27	16
$(0.5, 510)$	251	259
$(0.5, 5033)$	2540	2493

Table 3.3 *The frequencies of the new random variables $\mathbf{Y} = \mathbf{y}$ of N trials.*

(θ, N)	(Y_1, Y_2) $= (0, 0)$	(Y_1, Y_2) $= (1, 0)$	(Y_1, Y_2) $= (0, 1)$	(Y_1, Y_2) $= (1, 1)$
$(0.1, 10)$	80	13	5	2
$(0.1, 100)$	803	94	94	9
$(0.1, 1000)$	8095	934	866	105
$(0.2, 10)$	55	19	21	5
$(0.2, 100)$	653	150	156	41
$(0.2, 1000)$	6302	1617	1693	388
$(0.3, 10)$	49	23	21	7
$(0.3, 100)$	488	181	234	97
$(0.3, 1000)$	4885	2159	2020	936
$(0.4, 10)$	39	23	24	14
$(0.4, 100)$	361	218	264	157
$(0.4, 1000)$	3638	2393	2351	1618
$(0.5, 10)$	28	27	16	29
$(0.5, 100)$	241	251	259	249
$(0.5, 1000)$	2477	2540	2493	2490

3.2 Unbiased Estimators

Let $\mathbf{X} = (X_1, \ldots, X_n)' \sim \mathbb{P}_\theta$, $\theta \in \Theta$, where the parameter space Θ is a subset of $\mathbf{R}$. Let $E_\theta(\cdot)$ and $V_\theta(\cdot)$ be the expectation and variance with respect to $\mathbb{P}_\theta$, respectively.

Definition 3.2 *Let $\psi(\mathbf{X})$ denote a statistic. We say that $\psi(\mathbf{X})$ is an* unbiased estimator *of θ if $E_\theta\{\psi(\mathbf{X})\} = \theta$ for all $\theta \in \Theta$.*

Unbiasedness is a desirable property of the estimator because it ensures that the estimator reproduces in expectation the true value of the parameter. However, unbiased estimators of θ do not always take very close values to θ. Therefore, it is required to construct an estimator whose variance is the smallest among all the unbiased estimators. The following theorem gives a fundamental result.

Theorem 3.2 (Rao-Blackwell theorem) *Let $\mathbf{X} \sim \mathbb{P}_\theta$, $\theta \in \Theta \subset \mathbf{R}$, and let $T = T(\mathbf{X})$ be a sufficient statistic. Let $\hat{\theta}(\mathbf{X})$ be an arbitrary unbiased estimator of θ whose variance $V_\theta\{\hat{\theta}(\mathbf{X})\}$ exists. Define $\tilde{\theta}(T) \equiv E_\theta\{\hat{\theta}(\mathbf{X})|T\}$. Then the following (i) and (ii) hold.*

(i) $\tilde{\theta}(T)$ is an unbiased estimator of θ.

(ii) $V_\theta\{\tilde{\theta}(\mathbf{X})\} \leq V_\theta\{\hat{\theta}(\mathbf{X})\}$, $\theta \in \Theta$.

PROOF

(i) Since T is a sufficient statistic, $E_\theta\{\hat{\theta}(\mathbf{X})|T\}$ does not depend on θ. From Theorem 2.6 (iv) we have

$$\begin{aligned} E_\theta\{\tilde{\theta}(T)\} &= E_T[E_\theta\{\hat{\theta}(\mathbf{X})|T\}] \\ &= E_\theta\{\hat{\theta}(\mathbf{X})|T\} \\ &= \theta, \quad (\theta \in \Theta), \end{aligned}$$

which implies that $\tilde{\theta}(T)$ is an unbiased estimator of θ.

(ii) For simplicity, write $\hat{\theta} = \hat{\theta}(\mathbf{X})$ and $\tilde{\theta} = \tilde{\theta}(T)$. From Theorem 2.6 (v) it follows that

$$\begin{aligned} E_\theta\{(\hat{\theta} - \tilde{\theta})\tilde{\theta}|T\} &= \tilde{\theta} E_\theta\{(\hat{\theta} - \tilde{\theta})|T\} \quad a.e. \\ &= \tilde{\theta}(\tilde{\theta} - \tilde{\theta}) \quad a.e. \\ &= 0 \quad a.e., \end{aligned}$$

which leads to

$$E_\theta\{(\hat{\theta} - \tilde{\theta})\tilde{\theta}\} = 0. \tag{3.11}$$

Note that

$$\begin{aligned} V_\theta(\hat{\theta}) &= E_\theta\{(\tilde{\theta} - \theta + \hat{\theta} - \tilde{\theta})^2\} \\ &= E_\theta\{(\tilde{\theta} - \theta)^2\} + E_\theta\{(\hat{\theta} - \tilde{\theta})^2\} + 2E_\theta\{(\tilde{\theta} - \theta)(\hat{\theta} - \tilde{\theta})\}. \end{aligned}$$

From (3.11) we obtain

$$\begin{aligned} V_\theta(\hat{\theta}) &= E_\theta\{(\tilde{\theta} - \theta)^2\} + E_\theta\{(\hat{\theta} - \tilde{\theta})^2\} \\ &\geq E_\theta\{(\tilde{\theta} - \theta)^2\} \\ &= V_\theta(\tilde{\theta}), \quad (\theta \in \Theta). \end{aligned}$$

□

Another basic concept is completeness, which is defined as follows:

Definition 3.3 *Let* $\mathbf{X} \sim \mathbb{P}_\theta$, $\theta \in \Theta$. *A statistic* $T = T(\mathbf{X})$ *is* complete *if, for every measurable function* $g(T)$

$$E_\theta\{g(T)\} = 0 \quad \textit{for all } \theta \in \Theta$$

implies that $g(T) = 0$ *a.e.*

Example 3.4 *Let* $\{X_j\} \sim$ *i.i.d.* $B(1, \theta)$, $\theta \in \Theta = (0, 1)$. *Then* $T = \sum_{j=1}^n X_j$ *has the binomial distribution* $B(n, \theta)$, *whose probability function is*

$$p(t) = \binom{n}{t} \theta^t (1 - \theta)^{n-t}, \quad (t = 0, 1, \ldots, n).$$

Hence, if

$$\begin{aligned} \mathrm{E}_\theta\{g(T)\} &= \sum_{t=0}^n g(t) \binom{n}{t} \theta^t (1 - \theta)^{n-t} \\ &= (1 - \theta)^n \sum_{t=0}^n g(t) \binom{n}{t} \left(\frac{\theta}{1 - \theta}\right)^t \\ &= 0 \end{aligned} \tag{3.12}$$

for all $\theta \in \Theta$, *then because* $\theta/(1 - \theta)$ *can take all values in* $(0, \infty)$ *we obtain*

$$g(t) \binom{n}{t} = 0, \quad (t = 0, 1, \ldots, n),$$

which implies

$$g(t) = 0, \quad (t = 0, 1, \ldots, n).$$

Hence T *is complete.* □

Example 3.5 *Let* $\{X_j\} \sim$ *i.i.d.* $N(\theta, 1)$, $\theta \in \Theta = (-\infty, \infty)$. *Then* $T = n^{-1} \sum_{j=1}^n X_j$ *has the normal distribution* $N(\theta, n^{-1})$ *(Exercise 3.3). Suppose*

that

$$E_\theta\{g(T)\} = \int_{-\infty}^{\infty} g(t)\left(\frac{2\pi}{n}\right)^{-\frac{1}{2}} \exp\left\{-\frac{n}{2}(t-\theta)^2\right\}dt$$
$$= 0, \quad (\theta \in \Theta).$$

Then we have

$$\int_{-\infty}^{\infty} g(t)e^{-\frac{nt^2}{2}}e^{n\theta t}dt \times e^{-\frac{n\theta^2}{2}} = 0, \quad (\theta \in \Theta).$$

Since $\int_{-\infty}^{\infty} g(t)e^{-\frac{nt^2}{2}}e^{n\theta t}dt$ *is the Laplace transform of* $g(t)e^{-\frac{nt^2}{2}}$*, the above equation leads to*

$$g(t)e^{-\frac{nt^2}{2}} = 0 \quad a.e.$$

which implies $g(t) = 0$ *a.e. Thus* T *is complete.* □

Definition 3.4 *Let* $\mathbf{X} \sim \mathbb{P}_\theta$, $\theta \in \Theta$. *An unbiased estimator* $\psi_0(\boldsymbol{X})$ *of* θ *is called a* uniformly minimum variance unbiased (UMVU) estimator *of* θ *if*

$$V_\theta\{\psi_0(\mathbf{X})\} \leq V_\theta\{\psi(\mathbf{X})\} \quad (\theta \in \Theta)$$

for any other unbiased estimator $\psi(\mathbf{X})$ *of* θ.

The following theorem is a fundamental result in the unbiased estimation theory.

Theorem 3.3 (Lehmann-Scheffé theorem) *Let* $\mathbf{X} \sim \mathbb{P}_\theta$, $\theta \in \Theta$ *and let* $T = T(\mathbf{X})$ *be a complete sufficient statistic. Define*

$$\tilde{\theta}(T) \equiv E_\theta\{\hat{\theta}(\mathbf{X})|T\}$$

for any unbiased estimator $\hat{\theta}(\mathbf{X})$ *of* θ. *Then* $\tilde{\theta} = \tilde{\theta}(T)$ *is a UMVU estimator.*

PROOF It is clear that $\tilde{\theta}$ is an unbiased estimator of θ. Hence it is sufficient to show, for any unbiased estimator $u(\mathbf{X})$ of θ,

$$V_\theta(\tilde{\theta}) \leq V_\theta\{u(\mathbf{X})\}, \quad (\theta \in \Theta). \tag{3.13}$$

Let $u_0(T) \equiv E_\theta\{u(\mathbf{X})|T\}$. From Theorem 3.2 we obtain

$$\begin{aligned} &E_\theta\{u_0(T)\} = \theta, \\ &V_\theta\{u_0(T)\} \leq V_\theta\{u(\mathbf{X})\} \end{aligned} \tag{3.14}$$

for all $\theta \in \Theta$. Since

$$E_\theta\{\tilde{\theta}(T)\} = E_\theta\{u_0(T)\} \ (= \theta),$$

$$E_\theta\{\tilde{\theta}(T) - u_0(T)\} = 0, \quad (\theta \in \Theta). \tag{3.15}$$

From the completeness of T and (3.15), we have

$$\tilde{\theta}(T) = u_0(T) \quad a.e.,$$

which leads to

$$V_\theta\{u_0(T)\} = V_\theta\{\tilde{\theta}(T)\}.$$

Therefore (3.13) follows from (3.14). □

Example 3.6 *Let* $\{X_j\} \sim$ *i.i.d.* $B(1,\theta)$. *Then* $\bar{X}_n = n^{-1}\sum_{j=1}^n X_j$ *is an unbiased estimator of* θ. *Recalling* $T = \sum_{j=1}^n X_j$ *is a complete sufficient statistic, from Theorem 3.3 we can see that* $\bar{X}_n$ *is the UMVU estimator of* θ.

Example 3.7 *Let* $\{X_j\} \sim$ *i.i.d.* $N(\theta,1)$ *and let* $\bar{X}_n = n^{-1}\sum_{j=1}^n X_j$. *Then* $E_\theta(\bar{X}_n) = \theta$. *It is known that* $\bar{X}_n$ *is a complete sufficient statistic. Thus we conclude that* $\bar{X}_n$ *is the UMVU estimator of* θ.

3.3 Efficient Estimators

In this section we are interested in estimating $g(\theta)$ where $g : \Theta \to \mathbf{R}$ is a known measurable function of unknown parameter θ ($\theta \in \Theta \subset \mathbf{R}$). The next theorem gives a lower bound for the variance of unbiased estimators.

Theorem 3.4 *Suppose* $\mathbf{X} \sim \mathbb{P}_\theta$, $\theta \in \Theta \subset \mathbf{R}$ *and* $\mathbf{X}$ *has the probability density function* $f_\theta(\mathbf{x})$, $\mathbf{x} \in \mathbf{R}^n$. *Let* $T(\mathbf{X})$ *be an arbitrary unbiased estimator of* $g(\theta)$ *and denote*

$$A(\phi,\theta) \equiv V_\theta\left[\frac{f_\phi(\mathbf{X})}{f_\theta(\mathbf{X})}\right] \tag{3.16}$$

Then, the inequality

$$V_\theta(T) \geq \sup_{\phi\in\Theta} \frac{\{g(\phi) - g(\theta)\}^2}{A(\phi,\theta)} \tag{3.17}$$

holds.

PROOF

Since T is an unbiased estimator, we have

$$E_\phi\{T(\mathbf{X})\} = \int T(\mathbf{x}) f_\phi(\mathbf{x}) d\mathbf{x} = g(\phi), \quad (\phi \in \Theta). \tag{3.18}$$

Therefore, it is seen that

$$\begin{aligned}
&E_\theta\left[\{T(\mathbf{X}) - g(\theta)\}\left\{\frac{f_\phi(\mathbf{X})}{f_\theta(\mathbf{X})} - 1\right\}\right] \\
&= \int \{T(\mathbf{x}) - g(\theta)\}\left\{\frac{f_\phi(\mathbf{x})}{f_\theta(\mathbf{x})} - 1\right\} f_\theta(\mathbf{x}) d\mathbf{x} \\
&= \int T(\mathbf{x})\{f_\phi(\mathbf{x}) - f_\theta(\mathbf{x})\} d\mathbf{x} - g(\theta)\int \{f_\phi(\mathbf{x}) - f_\theta(\mathbf{x})\} d\mathbf{x} \\
&= g(\phi) - g(\theta).
\end{aligned} \tag{3.19}$$

Applying the Cauchy-Schwarz inequality to the left-hand side of (3.19), we have for any $\phi \in \Theta$,

$$V_\theta(T) V_\theta \left\{ \frac{f_\phi(\mathbf{X})}{f_\theta(\mathbf{X})} \right\} \geq \{g(\phi) - g(\theta)\}^2 . \tag{3.20}$$

Hence, we have for any $\phi \in \Theta$,

$$V_\theta(T) \geq \frac{\{g(\phi) - g(\theta)\}^2}{A(\phi, \theta)} \tag{3.21}$$

which implies the assertion of theorem. □

Henceforth we suppose $f_\theta(\mathbf{x})$ is differentiable in θ. Then, it is possible to define the most important characteristic on statistical inference as follows:

Definition 3.5 *The quantity*

$$\mathcal{F}_{\mathbf{X}}(\theta) \equiv E_\theta \left[\left\{ \frac{\partial}{\partial \theta} \log f_\theta(\mathbf{X}) \right\}^2 \right] \tag{3.22}$$

is called the Fisher information measure *(for the case that* $\mathbf{X}$ *is discrete random variables,* $f_\theta(\mathbf{x})$ *is understood as the probability function of it). If* X_j*'s are i.i.d. random variables,* X_1 *has the probability density function (probability function)* $f_\theta(x)$*, and if the expectation* (E_θ) *and differentiation by* θ $(\partial/\partial\theta)$ *are interchangeable, then we have*

$$\mathcal{F}_{\mathbf{X}}(\theta) = n\mathcal{F}(\theta), \tag{3.23}$$

where $\mathcal{F}(\theta) \equiv E_\theta \left[\{(\partial/\partial\theta) \log f_\theta(X_1)\}^2 \right]$ *(Exercise 3.4).*

If $f_\theta(x) = h(x) \exp\{c(\theta)T(x) - B(\theta)\}$ *is an exponential family and* $c(\theta)$ *has a nonvanishing continuous derivative on* Θ*, then condition for (3.23) holds (see for example Bickel and Doksum (2001)).*

If we assume the derivative $g'(\theta)$ of $g(\theta)$ and

$$\lim_{\phi \to \theta} \frac{A(\phi, \theta)}{(\phi - \theta)^2} = J(\theta) \tag{3.24}$$

exist in the inequality (3.17) of Theorem 3.4, then we obtain the inequality

$$V_\theta(T) \geq \frac{\{g'(\theta)\}^2}{J(\theta)}. \tag{3.25}$$

Next, we assume that there exists a function $G(,)$ such that for any $\theta \in \Theta$ if we take a sufficiently small $\epsilon > 0$, then for any $\phi : |\phi - \theta| < \epsilon$,

$$\left| \frac{f_\phi(\mathbf{x}) - f_\theta(\mathbf{x})}{(\phi - \theta) f_\theta(\mathbf{x})} \right| \leq G(\mathbf{x}, \theta), \tag{3.26}$$

$$E_\theta \left\{ G(\mathbf{X}, \theta)^2 \right\} < \infty. \tag{3.27}$$

Applying Lebesgue's convergence theorem (Theorem A.1) to (3.24), we obtain

$$\begin{aligned} J(\theta) &= \lim_{\phi \to \theta} \frac{A(\phi, \theta)}{(\phi - \theta)^2} \\ &= \lim_{\phi \to \theta} \int \frac{\{f_\phi(\mathbf{x}) - f_\theta(\mathbf{x})\}^2}{\{(\phi - \theta) f_\theta(\mathbf{x})\}^2} f_\theta(\mathbf{x}) d\mathbf{x} \\ &= \int \left\{ \frac{\partial}{\partial \theta} \log f_\theta(\mathbf{x}) \right\}^2 f_\theta(\mathbf{x}) d\mathbf{x} \\ &= \mathcal{F}_{\mathbf{X}}(\theta). \end{aligned} \tag{3.28}$$

Combining (3.28) with (3.25), we have the inequality

$$V_\theta(T) \geq \frac{\{g'(\theta)\}^2}{\mathcal{F}_{\mathbf{X}}(\theta)} \tag{3.29}$$

which is often referred as the *Cramér-Rao inequality*, and the right-hand side of (3.29) is called the *Cramér-Rao lower bound*.

In the above, we are led to the Cramér-Rao inequality by using Theorem 3.4. Alternatively, if an unbiased estimator $T = T(\mathbf{X})$ of $g(\theta)$ satisfies $E_\theta \left\{T^2\right\} < \infty$, we can derive it as follows:

First, from the unbiasedness, we have

$$\int T(\mathbf{x}) \frac{f_\phi(\mathbf{x}) - f_\theta(\mathbf{x})}{(\phi - \theta) f_\theta(\mathbf{x})} f_\theta(\mathbf{x}) d\mathbf{x} = \frac{g(\phi) - g(\theta)}{\phi - \theta}. \tag{3.30}$$

Furthermore, by (3.26), (3.27) and the Cauchy-Schwarz inequality, it is seen that

$$\left| T(\mathbf{x}) \frac{f_\phi(\mathbf{x}) - f_\theta(\mathbf{x})}{(\phi - \theta) f_\theta(\mathbf{x})} \right| \leq |T(\mathbf{x}) G(\mathbf{x}, \theta)|, \tag{3.31}$$

$$\left\{ E_\theta \left| T(\mathbf{X}) G(\mathbf{X}, \theta) \right| \right\}^2 \leq E_\theta \left\{ T(\mathbf{X})^2 \right\} E_\theta \left\{ G(\mathbf{X}, \theta)^2 \right\} < \infty. \tag{3.32}$$

Letting $\phi \to \theta$ on both sides of (3.30) and applying Lebesgue's convergence theorem, we have

$$\int T(\mathbf{x}) \frac{\frac{\partial}{\partial \theta} f_\theta(\mathbf{x})}{f_\theta(\mathbf{x})} f_\theta(\mathbf{x}) d\mathbf{x} = g'(\theta). \tag{3.33}$$

Note that, in particular, we can interchange differentiation and integration for a statistic $T(\mathbf{x}) \equiv 1$, namely,

$$E_\theta \left\{ \frac{\partial}{\partial \theta} \log f_\theta(\mathbf{X}) \right\} = \int \frac{\partial}{\partial \theta} f_\theta(\mathbf{x}) d\mathbf{x} = \frac{\partial}{\partial \theta} \int f_\theta(\mathbf{x}) d\mathbf{x} = 0. \tag{3.34}$$

Therefore we can see that

$$\mathrm{Cov} \left[T(\mathbf{X}), \frac{\frac{\partial}{\partial \theta} f_\theta(\mathbf{X})}{f_\theta(\mathbf{X})} \right] = g'(\theta). \tag{3.35}$$

Applying the Cauchy-Schwarz inequality to the left-hand side of (3.35), we obtain

$$V_\theta[T] \geq \frac{g'(\theta)^2}{\mathcal{F}_{\mathbf{X}}(\theta)}. \tag{3.36}$$

Definition 3.6 *A statistic T is said to be* efficient *if the equality holds in (3.36).*

Tracing the above argument, we can see the following equivalence relationship:

"T is an efficient estimator."

$\Updownarrow$

"The equality holds in the inequality (3.36) (the equality holds in the Cauchy-Schwarz inequality)."

$\Updownarrow$

"There exists a linear relationship between $T(\mathbf{x})$ and $\{(\partial/\partial\theta) f_\theta(\mathbf{x})\}/f_\theta(\mathbf{x})$, namely there exist functions $a_1(\theta)$ and $a_2(\theta)$ which are independent of $\mathbf{x}$ and satisfy

$$T(\mathbf{x}) = a_1(\theta)\frac{\frac{\partial}{\partial\theta} f_\theta(\mathbf{x})}{f_\theta(\mathbf{x})} + a_2(\theta) \quad \text{a.e.}'' \tag{3.37}$$

$\Updownarrow$

"It can be represented as

$$T(\mathbf{x}) = a_1(\theta)\frac{\partial}{\partial\theta}\{\log f_\theta(\mathbf{x})\} + a_2(\theta) \quad \text{a.e.}'' \tag{3.38}$$

$\Updownarrow$

"Upon integrating both sides of (3.38) with respect θ we get a solution in the form

$$f_\theta(\mathbf{x}) = \exp\{g_1(\theta)T(\mathbf{x}) + g_2(\theta) + U(\mathbf{x})\} \quad \text{a.e.}'' \tag{3.39}$$

Therefore, from Theorem 3.1 we can see that T becomes a sufficient statistic for $\mathcal{P} = \{\mathbb{P}_\theta\}$.

Example 3.8 *Let $\{X_j\} \sim$ i.i.d. $N(\theta, \sigma^2)$ and assume that we are interested in only the estimation of θ. The probability density function of $\mathbf{X} = (X_1, \ldots, X_n)'$ is*

$$\begin{aligned} f_\theta(\mathbf{x}) &= \prod_{j=1}^{n} f_\theta(x_j) \\ &= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_j - \theta)^2\right\}, \quad \left(\mathbf{x} = (x_1, \ldots, x_n)' \in \mathbf{R}^n\right). \end{aligned} \tag{3.40}$$

Since the Fisher information measure is

$$\begin{aligned}\mathcal{F}_{\mathbf{x}}(\theta) &= n\mathcal{F}(\theta) \\ &= nE_\theta\left\{\frac{\partial}{\partial\theta}\log f_\theta(X_1)\right\}^2 \\ &= nE_\theta\left[\frac{\partial}{\partial\theta}\left\{-\frac{(X_1-\theta)^2}{2\sigma^2}\right\}\right]^2 \\ &= nE_\theta\left\{\frac{(X_1-\theta)}{\sigma^2}\right\}^2 = \frac{n}{\sigma^2}, \end{aligned} \tag{3.41}$$

in this case, the Cramér-Rao lower bound becomes σ^2/n. *If we take* $\overline{X}_n = n^{-1}\sum_{j=1}^n X_j$ *as an estimator of* θ, *we have*

$$E_\theta\left\{\overline{X}_n\right\} = \theta, \tag{3.42}$$

$$V_\theta\left\{\overline{X}_n\right\} = \frac{\sigma^2}{n}, \tag{3.43}$$

so, the variance of $\overline{X}_n$ *attains the Cramér-Rao lower bound* σ^2/n. *Hence,* $\overline{X}_n$ *is an efficient estimator of* θ.

Before we turn to discuss the next example, we introduce a fundamental distribution. Let $\{X_j\} \sim$ *i.i.d.* $N(0,1)$, then the random variable $X \equiv X_1^2+\cdots+X_n^2$ has a χ_n^2 distribution. That is, X has the probability density function

$$f_X(x) = \frac{x^{(n-2)/2}e^{-x/2}}{2^{n/2}\Gamma(n/2)}, \quad x>0, \tag{3.44}$$

where $\Gamma(p)$ denotes the Gamma function defined by

$$\Gamma(p) = \int_0^\infty t^{p-1}e^{-t}dt. \tag{3.45}$$

Furthermore,

$$E(X) = n, \quad V(X) = 2n \tag{3.46}$$

(Exercise 3.7).

Example 3.9 *Let* $\{X_j\} \sim$ *i.i.d.* $N(\mu,\theta)$ *where both* μ *and* θ *are unknown, and assume that we are interested in only the estimation of* θ. *We consider*

$$S^2 = \frac{1}{n-1}\sum_{j=1}^n\left(X_j-\overline{X}_n\right)^2, \quad \overline{X}_n = n^{-1}\sum_{j=1}^n X_j \tag{3.47}$$

as an estimator of θ. *Then, we show that,*

(i) $\overline{X}_n$, *and* S^2 *are independent,*

(ii)

$$\overline{X}_n \sim N(\mu,\theta/n), \tag{3.48}$$

(iii)

$$\frac{n-1}{\theta}S^2 \sim \chi^2_{n-1}. \tag{3.49}$$

Construct an orthogonal matrix $T = (t_{ij})$ *whose first row is* $\mathbf{t}_1 = (n^{-1/2}, \ldots, n^{-1/2})$. *This is equivalent to finding one of the many orthogonal bases in* $\mathbf{R}^n$ *whose first member is* $\mathbf{t}_1$. *It, for instance, may be done by the Gram-Schmidt process (Exercise 3.8). Let* $\mathbf{Y} = (Y_1, \ldots, Y_n)' = T(X_1, \ldots, X_n)' = T\mathbf{X}$, *then*

$$\mathbf{Y} \sim N(T\boldsymbol{\mu}, \theta I_n), \tag{3.50}$$

where $\boldsymbol{\mu} = (\mu, \ldots, \mu)'$ *and* I_n *is identity matrix, so,* $\{Y_j\}$ *are mutually independent (Exercise 3.9). Since* $t_{i1} = 1/\sqrt{n}$, $1 \le i \le n$ *and* T *is orthogonal we see that*

$$\sum_{i=1}^{n} t_{ij} = \sqrt{n}\sum_{i=1}^{n} t_{ij}t_{i1} = 0, \quad j = 2, \ldots, n, \tag{3.51}$$

hence, $E(Y_1) = \sqrt{n}\mu$ *and* $E(Y_j) = 0$, $j = 2, \ldots, n$. *Therefore,* $\theta^{-1}\sum_{j=2}^{n} Y_j^2$ *has a* χ^2_{n-1} *distribution. Since by the definition of* T,

$$Y_1 = \sqrt{n}\,\overline{X}_n, \quad \sum_{j=1}^{n} X_j^2 = \sum_{j=1}^{n} Y_j^2, \tag{3.52}$$

hence,

$$\begin{aligned}(n-1)S^2 &= \sum_{j=1}^{n}\left(X_j - \overline{X}_n\right)^2 = \sum_{j=1}^{n} X_j^2 - n\overline{X}_n^2 \\ &= \sum_{j=1}^{n} Y_j^2 - Y_1^2 = \sum_{j=2}^{n} Y_j^2.\end{aligned} \tag{3.53}$$

This implies the assertion (3.49). From (3.46) and (3.49) we obtain

$$E_\theta(S^2) = \theta, \quad V_\theta(S^2) = \frac{2\theta^2}{n-1}. \tag{3.54}$$

Since a statistic which has one to one correspondence with a sufficient statistic is also sufficient, from Example 3.3 we see that $U(\mathbf{X}) = \left(\sum_{j=1}^{n} X_j, \sum_{j=1}^{n} X_j^2\right)$ *is a sufficient statistic for the concerned distribution family. Furthermore, we can show that* $U(\mathbf{X})$ *is complete (Exercise 3.10). Therefore, by Theorem 3.3,* S^2 *is the UMVU estimator of* θ.

Next, we calculate the Fisher information measure with respect to θ. *Since*

$$f_\theta(\mathbf{x}) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{(x_j - \mu)^2}{2\theta}\right\}, \quad (\mathbf{x} = (x_1, \ldots, x_n)'), \tag{3.55}$$

we have

$$\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{x}) = \sum_{j=1}^{n}\left\{-\frac{1}{2\theta}+\frac{(x_j-\mu)^2}{2\theta^2}\right\}. \tag{3.56}$$

Note that $(X_j-\mu)^2/\theta \sim$ *i.i.d.* χ_1^2*, so, we obtain*

$$\mathcal{F}_{\mathbf{X}}(\theta) = \sum_{j=1}^{n} E_\theta\left\{\frac{(X_j-\mu)^2}{2\theta^2}-\frac{1}{2\theta}\right\}^2 = \frac{n}{2\theta^2} \tag{3.57}$$

(Exercise 3.7), and in this case the Cramér-Rao lower bound becomes $2\theta^2/n$ *which is smaller than the variance* $2\theta/(n-1)$ *of the UMVU estimator. Therefore, in this example, the UMVU estimator is not an efficient estimator, which implies the Cramér-Rao lower bound is not necessarily an attainable bound. Table 3.4 lists the results of* 10000 *times experiments of* S^2 *values from* $N(0,\theta)$ *samples of size* n*, for* $n = 10, 50, 100$ *and* $\theta = 100, 1000, 10000$*. From these results we can conclude the variance of* S^2 *is significantly larger than the Cramér-Rao lower bound for small* n *and large* θ*.*

Table 3.4 *The results of* 10000 *times experiments of* S^2 *values from* $N(0,\theta)$ *samples of size* n*.*

(n,θ)	*mean*	$(sd)^2$	Cramér-Rao
$(10,100)$	99.79203	2209.501	2000
$(10,1000)$	1000.133	219275.4	200000
$(10,10000)$	9927.399	21574154	20000000
$(50,100)$	99.62352	407.1349	400
$(50,1000)$	999.5835	40943.39	40000
$(50,10000)$	10060.96	4168401	4000000
$(100,100)$	100.0517	204.0988	200
$(100,1000)$	1001.933	20493.08	20000
$(100,10000)$	9972.692	2004887	2000000

Until now, we have discussed the case that the unknown parameter θ is one-dimensional. However, the argument of multidimensional cases would be required. Let $\mathbf{X} = (X_1,\ldots,X_n)' \sim \mathbb{P}_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta \subset \mathbf{R}^q$ and $\boldsymbol{\theta} = (\theta_1,\ldots\theta_q)'$. We assume that $\mathbf{X}$ is continuous (discrete) random variables, which has a probability density function (probability function) $f_{\boldsymbol{\theta}}(\mathbf{x})$. Now, we are interested

in the estimation of known measurable function $g : \Theta \to \mathbf{R}^r$ $(r \leq q)$ of $\boldsymbol{\theta}$. As in the one-dimensional case, the Cramér-Rao lower bound of an unbiased estimator $\mathbf{T}$ of $g(\boldsymbol{\theta})$ is given by

$$\Delta \mathcal{F}_{\mathbf{X}}^{-1}(\boldsymbol{\theta}) \Delta', \tag{3.58}$$

where

$$\Delta = \frac{\partial}{\partial \boldsymbol{\theta}'} g(\boldsymbol{\theta}) \quad (r \times q \text{ matrix}), \tag{3.59}$$

$$\mathcal{F}_{\mathbf{X}}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\mathbf{X}) \frac{\partial}{\partial \boldsymbol{\theta}'} \log f_{\boldsymbol{\theta}}(\mathbf{X}) \right\} \quad (q \times q \text{ matrix}). \tag{3.60}$$

The matrix (3.58) is a lower bound means that, if $V_{\boldsymbol{\theta}}(\mathbf{T})$ is the variance matrix of any unbiased estimator $\mathbf{T}$ of $g(\boldsymbol{\theta})$, then $V_{\boldsymbol{\theta}}(\mathbf{T}) - \Delta \mathcal{F}_{\mathbf{X}}^{-1}(\boldsymbol{\theta}) \Delta'$ is non-negative definite. The matrix (3.60) is called the *Fisher information measure matrix*. If $\{X_j\}$ is *i.i.d.* and has a probability density function (probability function) $f_{\boldsymbol{\theta}}(x_1)$, $x_1 \in \mathbf{R}$, and the operations $\partial / \partial \boldsymbol{\theta}$ and $E_{\boldsymbol{\theta}}$ are interchangeable, then we have $\mathcal{F}_{\mathbf{X}}(\boldsymbol{\theta}) = n \mathcal{F}(\boldsymbol{\theta})$, where

$$\mathcal{F}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(X_1) \frac{\partial}{\partial \boldsymbol{\theta}'} \log f_{\boldsymbol{\theta}}(X_1) \right\}. \tag{3.61}$$

Therefore, in this case the Cramér-Rao lower bound becomes

$$\frac{1}{n} \Delta \mathcal{F}^{-1}(\boldsymbol{\theta}) \Delta'. \tag{3.62}$$

3.4 Asymptotically Efficient Estimators

As we saw in Example 3.9 of the previous section, although the unbiased estimator $S^2 = (n-1)^{-1} \sum_{j=1}^{n} (X_j - \bar{X}_n)^2$ of the variance of a normal distribution is a UMVU estimator, it is not an efficient estimator. In fact, from

$$\{\text{Cramér-Rao lower bound}\} / V_{\theta}(S^2) = \frac{n-1}{n} \tag{3.63}$$

the above equation is less than 1 for any finite sample size. However, since the above equation tends to 1 as $n \to \infty$, S^2 is a good estimator of variance for sufficiently large n. This section discusses "goodness" of estimators for large n.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables. If $\mathbf{X} = (X_1, \ldots, X_n)'$ has a probability distribution $\mathbb{P}_{n,\boldsymbol{\theta}}$ depending on an unknown parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)' \in \Theta \subset \mathbf{R}^q$, then we write $\mathbf{X} \sim \mathbb{P}_{n,\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta$. The previous section discussed the class of unbiased estimators of $\boldsymbol{\theta}$. Here we consider the following class of estimators.

Definition 3.7 *An estimator $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\mathbf{X})$ of $\boldsymbol{\theta}$ is said to be a* consistent estimator *if for every $\boldsymbol{\theta} \in \Theta$ and every $\varepsilon > 0$*

$$\lim_{n \to \infty} \mathbb{P}_{n,\boldsymbol{\theta}} \{ \| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \| > \varepsilon \} = 0, \tag{3.64}$$

where $\|\cdot\|$ is the Euclidean norm. (3.64) *implies that $\hat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}$ in probability, and as in the previous section it is expressed as $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}$.*

Let each X_j have the probability density function (probability function) $f_{\boldsymbol{\theta}}(x)$, $x \in \mathbf{R}$, and let $f_{\boldsymbol{\theta}}(x)$, $x \in \mathbf{R}$ be differentiable with respect to $\boldsymbol{\theta}$, then $\mathbf{X}$ has the joint probability density function (joint probability function) $f_{n,\boldsymbol{\theta}}(\mathbf{x}) = \prod_{j=1}^{n} f_{\boldsymbol{\theta}}(x_j)$, $\mathbf{x} = (x_1, \ldots, x_n)'$. We set

$$\mathbf{Z}_n = \frac{1}{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{n,\boldsymbol{\theta}}(\mathbf{X}). \tag{3.65}$$

Henceforth, it is assumed that the Fisher information matrix

$$\mathcal{F}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left[\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{n,\boldsymbol{\theta}}(X_1) \frac{\partial}{\partial \boldsymbol{\theta}'} \log f_{n,\boldsymbol{\theta}}(X_1) \right\} \right]$$

exists, and is a positive definite matrix.

Definition 3.8 *If a consistent estimator $\mathbf{T}_n$ of $\boldsymbol{\theta}$ satisfies*

$$\sqrt{n} \|\mathbf{T}_n - \boldsymbol{\theta} - \mathcal{F}(\boldsymbol{\theta})^{-1} \mathbf{Z}_n\| \xrightarrow{p} 0, \quad (n \to \infty), \tag{3.66}$$

then $\mathbf{T}_n$ is said to be an asymptotically efficient estimator *of $\boldsymbol{\theta}$.*

Although implications of this definition is elusive, it is known that asymptotically efficient estimator minimizes the expectation of loss function, which belongs to a rich class, of $\mathbf{T}_n - \boldsymbol{\theta}$. This discussion is omitted in this book. See e.g., Taniguchi and Kakizawa (2000) for the reader interested in more detail.

Let us see an implication of the asymptotic efficiency. From Theorem 2.7 (ii) and Remark 2.2 we have

Theorem 3.5 *If $\mathbf{T}_n$ is an asymptotically efficient estimator of $\boldsymbol{\theta}$, then*

$$\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathcal{F}(\boldsymbol{\theta})^{-1}), \quad (n \to \infty). \tag{3.67}$$

From this theorem we can see that if $\mathbf{T}_n$ is asymptotically efficient, then the covariance matrix of the asymptotic distribution of $\mathbf{T}_n$ is $\mathcal{F}(\boldsymbol{\theta})^{-1}$ and $\mathbf{T}_n$ attains the Cramér-Rao lower bound asymptotically.

Let us see an important example of an asymptotically efficient estimator. First, let $\mathbf{X} = (X_1, \ldots, X_n)' \sim \mathbb{P}_{n,\boldsymbol{\theta}}$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)' \in \Theta \subset \mathbf{R}^q$. If $\mathbb{P}_{n,\boldsymbol{\theta}}$ is a continuous (discrete) distribution, we denote the probability density function (probability function) of $\mathbf{X}$ by $f_{n,\boldsymbol{\theta}}(\mathbf{x})$, $\mathbf{x} \in \mathbf{R}$. The *likelihood function* is defined by

$$L_n(\boldsymbol{\theta}) = f_{n,\boldsymbol{\theta}}(\mathbf{X}).$$

It is the probability (density) function of the observation at $\mathbf{X}$, treated as a function of $\boldsymbol{\theta}$. Note that for the sake of simplicity, the dependence of $L_n(\boldsymbol{\theta})$ on $\mathbf{X}$ is suppressed. Denote $\log L_n(\boldsymbol{\theta})$ by $l_n(\boldsymbol{\theta})$.

Definition 3.9 *We say that* $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\mathbf{X})$ *is the* maximum likelihood estimator (MLE) *of* $\boldsymbol{\theta}$ *if* $\hat{\boldsymbol{\theta}}_n$ *satisfies*

$$l_n(\hat{\boldsymbol{\theta}}_n) = \sup_{\boldsymbol{\theta}\in\Theta} l_n(\boldsymbol{\theta}_n). \tag{3.68}$$

If the above supremum is achieved at interior points of Θ and if $l_n(\boldsymbol{\theta})$ is differentiable in $\boldsymbol{\theta}$, then $\hat{\boldsymbol{\theta}}_n$ is a solution of the *likelihood equation*

$$\frac{\partial}{\partial\boldsymbol{\theta}} l_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}. \tag{3.69}$$

The estimator satisfying the likelihood equation is called the *likelihood equation estimator*. If $l_n(\boldsymbol{\theta})$ is smooth with respect to $\boldsymbol{\theta}$, the likelihood equation estimator is often the MLE.

Example 3.10 *Let* $X_1, \ldots, X_n \sim$ *i.i.d.* $N(\mu, \sigma^2)$, $-\infty < \mu < \infty$, $0 < \sigma^2 < \infty$, *and* $\boldsymbol{\theta} = (\mu, \sigma^2)'$. *The log likelihood is*

$$l_n(\boldsymbol{\theta}_n) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(X_j - \mu)^2, \tag{3.70}$$

and the likelihood equations are

$$\frac{\partial}{\partial\mu} l_n(\boldsymbol{\theta}_n) = \frac{1}{\sigma^2}\left\{\sum_{j=1}^{n} X_j - n\mu\right\} = 0, \tag{3.71}$$

$$\frac{\partial}{\partial\sigma^2} l_n(\boldsymbol{\theta}_n) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{j=1}^{n}(X_j - \mu)^2 = 0. \tag{3.72}$$

From (3.71) *and* (3.72) *the likelihood equation estimators of* μ *and* σ^2 *are given by*

$$\hat{\mu}_n = \bar{X}_n \equiv n^{-1}\sum_{j=1}^{n} X_j, \qquad \hat{\sigma}_n^2 = n^{-1}\sum_{j=1}^{n}(X_j - \bar{X}_n)^2,$$

respectively. Putting $\hat{\boldsymbol{\theta}}_n = (\hat{\mu}_n, \hat{\sigma}_n^2)'$, *we can see that* $\hat{\boldsymbol{\theta}}_n$ *is the MLE of* $\boldsymbol{\theta}$. *In fact, note that*

$$l_n(\hat{\boldsymbol{\theta}}_n) = -\frac{n}{2}\log(2\pi) - n\log\hat{\sigma}_n - \frac{n}{2},$$

$$l_n(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{n\hat{\sigma}_n^2 + n(\bar{X}_n - \mu)^2}{2\sigma^2}.$$

we obtain, using the inequality $A - 1 - \log A \geq 0$ $(A > 0)$,

$$\begin{aligned} l_n(\hat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}) &= \frac{n}{2}\left[\frac{\hat{\sigma}_n^2}{\sigma^2} - 1 - \log\frac{\hat{\sigma}_n^2}{\sigma^2}\right] + \frac{n(\bar{X}_n - \mu)^2}{2\sigma^2} \\ &\geq \frac{n(\bar{X}_n - \mu)^2}{2\sigma^2} \geq 0 \end{aligned}$$

for all $\boldsymbol{\theta} \in \Theta$. *Hence,* $\hat{\boldsymbol{\theta}}_n$ *is the point maximizing* $l_n(\boldsymbol{\theta})$, *and is the MLE.* □

Let $X_1, \ldots, X_n$ be i.i.d. continuous (discrete) random variables with probability density function (probability function) $f_{\boldsymbol{\theta}_0}(x), \in \mathbf{R}$, $\boldsymbol{\theta}_0 \in \Theta \subset \mathbf{R}$, where $\boldsymbol{\theta}_0$ is the true value. Since the following discussion requires a description for all $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_1)' (\neq \boldsymbol{\theta}_0) \in \Theta$, we write this way. We impose the following assumption

Assumption 3.1 *(i) The derivatives* $\partial \log f_{\boldsymbol{\theta}}(x)/\partial\theta_i$, $\partial^2 \log f_{\boldsymbol{\theta}}(x)/\partial\theta_i\partial\theta_j$ *and* $\partial^3 \log f_{\boldsymbol{\theta}}(x)/\partial\theta_i\partial\theta_j\partial\theta_k$, $i, j, k = 1, \ldots, q$, $(x \in \mathbf{R})$ *exist for all* $\boldsymbol{\theta} \in \Theta$.

(ii)

$$E_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(X_1)\right] = \mathbf{0} \quad (q \times 1 \text{ vector}),$$

$$E_{\boldsymbol{\theta}}\left[\frac{1}{f_{\boldsymbol{\theta}}(X_1)} \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} f_{\boldsymbol{\theta}}(X_1)\right] = \mathbf{0} \quad (q \times q \text{ matrix}),$$

$$E_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(X_1) \frac{\partial}{\partial\boldsymbol{\theta}'} \log f_{\boldsymbol{\theta}}(X_1)\right] > 0 \quad (q \times q \text{ positive definite matrix})$$

at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

(iii) For all $\boldsymbol{\theta} \in \Theta$ *there exist a function* $M(x)$ *and a constant* K *such that*

$$\left|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k} \log f_{\boldsymbol{\theta}}(x)\right| \leq M(x), \quad (i, j, k = 1, \ldots, q),$$

$$E_{\boldsymbol{\theta}} M(X_1) \leq K < \infty.$$

(iv) If $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$, *then the set* $\{x : f_{\boldsymbol{\theta}}(x) \neq f_{\boldsymbol{\theta}^*}(x)\}$ *has a positive probability at* $\boldsymbol{\theta}$.

Theorem 3.6 *Under Assumption 3.1 there exists a likelihood equation estimator* $\hat{\boldsymbol{\theta}}_n = (\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,q})'$ *of* $\boldsymbol{\theta}_0$ *which is a consistent estimator of* $\boldsymbol{\theta}_0 = (\theta_{0,1}, \ldots, \theta_{0,q})'$.

PROOF From Exercise 3.15 and Assumption 3.1 (iv) it is seen that for any $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_q)'$, $\delta_j > 0$, $j = 1, \ldots, q$

$$E_{\boldsymbol{\theta}_0}\left\{\log \frac{f_{\boldsymbol{\theta}_0 - \boldsymbol{\delta}}(X_1)}{f_{\boldsymbol{\theta}_0}(X_1)}\right\} < 0, \qquad E_{\boldsymbol{\theta}_0}\left\{\log \frac{f_{\boldsymbol{\theta}_0 + \boldsymbol{\delta}}(X_1)}{f_{\boldsymbol{\theta}_0}(X_1)}\right\} < 0, \tag{3.73}$$

(Exercise 3.16). Hence, setting $l_n(\boldsymbol{\theta}) = \sum_{j=1}^n \log f_{\boldsymbol{\theta}}(X_j)$, we have, by the law of large numbers (Remark 2.1 and Theorem 2.8),

$$\frac{1}{n}\left[l_n(\boldsymbol{\theta}_0 - \boldsymbol{\delta}) - l_n(\boldsymbol{\theta}_0)\right] \xrightarrow{a.s.} c_1 < 0, \tag{3.74}$$

$$\frac{1}{n}\left[l_n(\boldsymbol{\theta}_0 + \boldsymbol{\delta}) - l_n(\boldsymbol{\theta}_0)\right] \xrightarrow{a.s.} c_2 < 0 \tag{3.75}$$

as $n \to \infty$. Therefore $l_n(\boldsymbol{\theta})$ has a larger value at $\boldsymbol{\theta}_0$ than at $\boldsymbol{\theta}_0 \pm \boldsymbol{\delta}$ a.s. as n tends to ∞. Because $l_n(\boldsymbol{\theta})$ is continuous with respect to $\boldsymbol{\theta}$, it has a maximum value on the line between $\boldsymbol{\theta}_0 - \boldsymbol{\delta}$ and $\boldsymbol{\theta}_0 + \boldsymbol{\delta}$. Also because $l_n(\boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$, it is seen that its first derivative is equal to $\mathbf{0}$ at maxima. Since $\boldsymbol{\delta}$ is arbitrary, the theorem follows. □

Theorem 3.7 *Under Assumption 3.1 the likelihood equation estimator $\hat{\boldsymbol{\theta}}_n$ satisfying the statement in Theorem 3.6 is an asymptotically efficient estimator and*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathcal{F}(\boldsymbol{\theta}_0)^{-1}),$$

where

$$\mathcal{F}(\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}_0}\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}_0}(X_1) \frac{\partial}{\partial \boldsymbol{\theta}'} \log f_{\boldsymbol{\theta}_0}(X_1) \right\} \ (\textit{Fisher information matrix}).$$

PROOF Since $\hat{\boldsymbol{\theta}}_n$ satisfies the likelihood equation

$$\frac{\partial}{\partial \boldsymbol{\theta}} l_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}, \tag{3.76}$$

expanding the left-hand side of (3.76) into a Taylor series at $\boldsymbol{\theta}_0$, we get

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\theta}} l_n(\boldsymbol{\theta}_0) + \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} l_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + A_n, \tag{3.77}$$

where

$$A_n = \frac{1}{2}\left[(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' \frac{\partial}{\partial \theta_1} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} l_n(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0), \ldots, \right.$$
$$\left. (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' \frac{\partial}{\partial \theta_q} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} l_n(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \right]',$$

and $\boldsymbol{\theta}^*$ is between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n$. Solving (3.77) with respect to $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0$, we obtain

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -\left[\frac{1}{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} l_n(\boldsymbol{\theta}_0) + B_n \right]^{-1} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}} l_n(\boldsymbol{\theta}_0), \tag{3.78}$$

where B_n is a $q \times q$ matrix and the absolute values of each component of B_n is less than

$$\sum_{j=1}^{q} \left| \hat{\theta}_{n,j} - \theta_{0,j} \right| \times \max_{1 \le i,k,l \le q} \frac{1}{n} \sum_{j=1}^{n} \left| \frac{\partial^3}{\partial \theta_i \partial \theta_k \partial \theta_l} f_{\boldsymbol{\theta}^*}(X_j) \right|. \tag{3.79}$$

From Assumption 3.1 (iii) and consistency of $\hat{\boldsymbol{\theta}}_n$ it follows that $B_n \xrightarrow{p} \mathbf{0}$. From (3.78) we have

$$\begin{aligned} &\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) - \mathcal{F}(\boldsymbol{\theta}_0)^{-1} \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}} l_n(\boldsymbol{\theta}_0) \\ &\quad = -\left[\left\{ \frac{1}{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} l_n(\boldsymbol{\theta}_0) + B_n \right\}^{-1} + \mathcal{F}(\boldsymbol{\theta}_0)^{-1} \right] \frac{1}{\sqrt{n}} \frac{\partial}{\partial \boldsymbol{\theta}} l_n(\boldsymbol{\theta}_0). \end{aligned} \tag{3.80}$$

From Assumption 3.1 (ii) it follows that

$$E_{\boldsymbol{\theta}_0}\left\{\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}l_n(\boldsymbol{\theta}_0)\right\} = -\mathcal{F}(\boldsymbol{\theta}_0).$$

Applying the law of large numbers and the central limit theorem (Remark 2.2) to $(1/n)(\partial^2/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}')l_n(\boldsymbol{\theta}_0)$ and $(1/\sqrt{n})(\partial/\partial\boldsymbol{\theta})l_n(\boldsymbol{\theta}_0)$, respectively, from Exercise 2.18 and 2.19, it is seen that the right-hand side of (3.80) converge to $\mathbf{0}$ in probability, which implies the assertion. □

Exercises

3.1 Prove Theorem 3.1 in the case that $\mathbf{X}$ is continuous.

3.2 Prove the statement (3.10).

3.3 If $\{X_j\} \sim$ i.i.d. $N(\theta, 1)$, then show that the distribution of $n^{-1}\sum_{j=1}^n X_j$ is $N(\theta, n^{-1})$.

3.4 Verify (3.23).

3.5 Minimize $E_\theta\left\{T(\mathbf{X}) - a(\theta)(\partial/\partial\theta)\log f_\theta(\mathbf{X}) - b(\theta)\right\}^2$ with respect to $a(\theta)$ and $b(\theta)$, where $a(\theta)$ and $b(\theta)$ are independent of $\mathbf{X}$.

3.6 Verify the equivalent relation (3.37)-(3.39).

3.7 **(i)** Verify (3.44) and (3.46).
(ii) Show (3.57).

3.8 Find an orthogonal matrix T whose first row is $\mathbf{t}_1 = \left(n^{-1/2}, \ldots, n^{-1/2}\right)$.

3.9 Verify (3.50).

3.10 Let $\{X_j\} \sim$ *i.i.d.* $N(\mu, \sigma)$. Show that $\left(\sum_{j=1}^n X_j, \sum_{j=1}^n X_j^2\right)$ is a sufficient and complete statistic for $\theta = (\mu, \sigma)$.

3.11 If $\{X_j\} \sim$ i.i.d. $P_o(\lambda)$ (Poisson distribution), then find the UMVU estimator of λ.

3.12 If $\{X_j\} \sim$ i.i.d. $Exp(1/\theta, 0)$ (exponential distribution), then find the UMVU estimator of θ.

3.13 If $\{X_j\} \sim$ i.i.d. $P_o(\lambda)$, then find the MLE of λ.

3.14 If $\{X_j\} \sim$ i.i.d. $U(0, \theta)$ (uniform distribution), then find the MLE of θ.

3.15 Let $f(x)$ and $g(x)$ be probability density functions. Then show that the following inequality holds

$$\int_{\mathbf{R}} f(x)\log\left\{\frac{f(x)}{g(x)}\right\}dx \geq 0,$$

and that the equality holds if and only if $f(x) = g(x)$ a.e.

3.16 Prove the inequality (3.73).

CHAPTER 4

Various Statistical Methods

In the previous chapter we discussed the estimation of unknown parameters in statistical models. In this chapter, we will present various statistical methods. Concretely speaking, we deal with interval estimations, testing problems and discriminant analyses. The inference in the previous chapter is the method for estimating an unknown parameter to be a certain value based on a sample (point estimation). In this chapter we discuss a method for seeking interval depending on a sample, in which the unknown parameter lies at a certain level of probability accuracy (interval estimation).

When a hypothesis formulates the probability distribution of an observed sample, the process of determining whether the observation indicates that the hypothesis is true or not is called the testing hypothesis. We use a statistic to make a decision whether the hypothesis is true or not and call this the test statistic. We describe the fundamental theory concerned with the optimality of test statistics. Furthermore, we explain various types of testing hypotheses.

There are myriads of statistical techniques and it is impossible to cover the entirety in this book. In the latter part of this chapter we introduce the discriminant analysis which is fundamental and of importance in various fields of applied statistics. We consider the case when we know a sample $\mathbf{X}$ belongs to one of several categories described by probability distributions but do not know to which it belongs. The purpose of the discriminant analysis is to find a discriminant criterion that distinguishes the category to which a sample $\mathbf{X}$ belongs, with a possibly high probability. In recent years, this has been much demanded for applications in biomedical diagnosis and, in particular, credit rating in financial engineering.

4.1 Interval Estimation

First, let $\mathbf{X} = (X_1, \ldots, X_n)' \sim \mathbb{P}_\theta$, $\theta \in \Theta$ (an open interval) $\subset \mathbf{R}$. A random closed interval $[l(\mathbf{X}), u(\mathbf{X})]$ independent of θ in Θ is said to be a level $(1-\alpha)$ *confidence interval* for θ, if, for all $\theta \in \Theta$,

$$\mathbb{P}_\theta \{l(\mathbf{X}) \leq \theta \leq u(\mathbf{X})\} \geq 1 - \alpha \tag{4.1}$$

and $l(\mathbf{X})$, $u(\mathbf{X})$ are called *confidence bounds*. For a given interval $[l(\mathbf{X}), u(\mathbf{X})]$, the confidence level is clearly not unique, since any number $(1-\alpha') \leq (1-\alpha)$

will be a confidence level, if $(1-\alpha)$ is so. In order to avoid this indefiniteness it is convenient to define the *confidence coefficient* of $[l(\mathbf{X}), u(\mathbf{X})]$ to be the largest possible confidence level, that is,

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta \{l(\mathbf{X}) \le \theta \le u(\mathbf{X})\} = 1-\alpha, \tag{4.2}$$

where the left-hand side is the minimum probability of coverage. We usually take 0.1, 0.05, 0.01 etc. as α. From the relation (4.1), we can announce that we are $100(1-\alpha)\%$ confident that an unknown parameter θ lies in the interval $[l(\mathbf{X}), u(\mathbf{X})]$ and call this inference procedure for unknown parameter the *interval estimation*. Of course the interval should be as short as possible. We can extend the notion of confidence interval for one-dimensional parameter θ to that for q-dimensional parameter vector $\boldsymbol{\theta}$. The random set $C(\mathbf{X})$ in Θ is a level $(1-\alpha)$ *confidence set* for $\boldsymbol{\theta}$, if, for all $\boldsymbol{\theta} \in \Theta \subset \mathbf{R}^q$,

$$\mathbb{P}_{\boldsymbol{\theta}} \{\boldsymbol{\theta} \in C(\mathbf{X})\} \ge 1-\alpha. \tag{4.3}$$

For simplicity, in what follows, we discuss the one-dimensional parameter case. There is a complete duality between the confidence interval (set) and testing problems described below. We will give the argument concerned with "goodness" of confidence interval (set) in Section 4.3 after describing "goodness" of tests. Therefore, in this section, we construct concrete examples of confidence interval (set) under various settings.

For construction of confidence interval we use a function $S_\theta(\mathbf{X})$ of a sample whose distribution does not depend on θ. In the following three examples we assume $\{X_j\} \sim$ *i.i.d.* $N(\mu, \sigma^2)$, $\mathbf{X} = (X_1, \ldots, X_n)'$ and give confidence intervals concerned with μ and σ^2.

Example 4.1 (σ^2 is known) *Since the sample mean $\overline{X}_n = n^{-1}\sum_{j=1}^n X_j$ is distributed as $N(\mu, \sigma^2/n)$,*

$$S_\mu(\mathbf{X}) \equiv \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \tag{4.4}$$

has the standard normal distribution $N(0,1)$, hence this distribution does not depend on μ. Therefore, for any $\alpha \in (0,1)$ we have

$$\mathbb{P}_\mu \left\{-z_{\alpha/2} \le S_\mu(\mathbf{X}) \le z_{\alpha/2}\right\} = 1-\alpha, \tag{4.5}$$

where $z_{\alpha/2}$ is the upper $100(\alpha/2)$ percent point of $N(0,1)$. Here, for concrete values of α percent points of normal distribution or other fundamental distributions will be found in the software environments for statistics (e.g. the R project). Furthermore, we selected the interval $[-z_{\alpha/2}, z_{\alpha/2}]$ among intervals $[a,b]$ satisfying $\mathbb{P}_\mu \{S_\mu(\mathbf{X}) \in [a,b]\} = 1-\alpha$, since the length $b-a$ is minimized at $b = z_{\alpha/2}$, $a = -z_{\alpha/2}$ in view of the symmetry and unimodality of normal distribution.

Rewriting the event in the left-hand side of (4.5) with respect to μ, (4.5)

becomes

$$\mathbb{P}_\mu \left\{ \overline{X}_n - \frac{\sigma z_{\alpha/2}}{\sqrt{n}} \leq \mu \leq \overline{X}_n + \frac{\sigma z_{\alpha/2}}{\sqrt{n}} \right\} = 1 - \alpha, \tag{4.6}$$

hence the confidence interval of μ *with confidence coefficient* $(1-\alpha)$ *is*

$$\left[\overline{X}_n - \frac{\sigma z_{\alpha/2}}{\sqrt{n}}, \overline{X}_n + \frac{\sigma z_{\alpha/2}}{\sqrt{n}} \right]. \tag{4.7}$$

Since it is not realistic that we know σ^2 as in the example above, we consider the case that σ^2 is unknown. For this purpose we need the following distribution:

Definition 4.1 *Let* Y *and* Z *be independent random variables with* $N(0,1)$ *and* $\chi^2(n)$ *distributions, respectively. The distribution of*

$$T \equiv \frac{Y}{\sqrt{\frac{Z}{n}}} \tag{4.8}$$

is called the t-distribution *with* n *degrees of freedom. We shall denote this by* $T \sim t(n)$*. In Exercise 4.2 we will derive the probability density function of* $t(n)$ *(see Figure 4.1). In view of Figure 4.1, we see that the tail probability of* $t(n)$ *is larger than that of* $N(0,1)$*. Furthermore, the probability density function of* $t(n)$ *tends to that of* $N(0,1)$ *as the degrees of freedom* n *of* $t(n)$ *becomes large.*

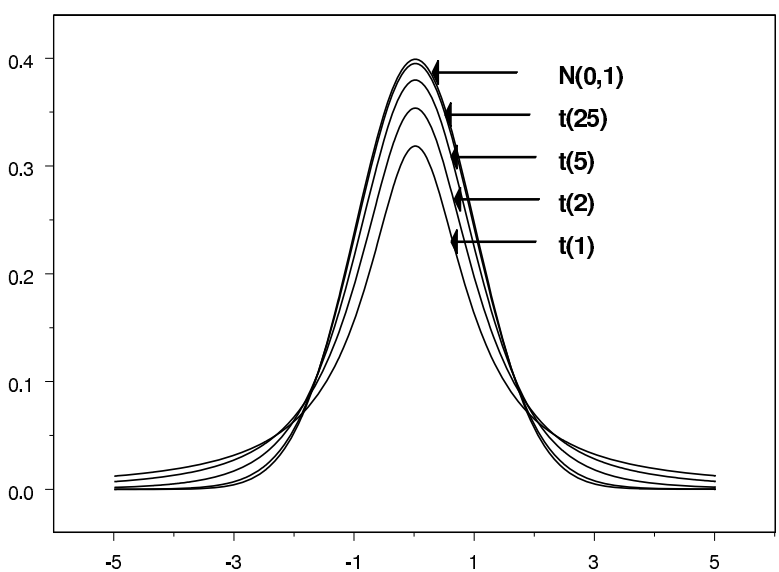


Figure 4.1 *The probability density functions of* $t(n)$*-distributions.*

Example 4.2 (σ^2 is unknown) *For this case, let us recall Example 3.9 in Chapter 3. Consider the sample variance*

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}_n\right)^2, \tag{4.9}$$

then we have

$$(n-1)\frac{\hat{\sigma}_n^2}{\sigma^2} \sim \chi^2(n-1). \tag{4.10}$$

Recall that $\sqrt{n}\left(\overline{X}_n - \mu\right)/\sigma \sim N(0,1)$ *and that it is independent of* $(n-1)\hat{\sigma}_n^2/\sigma^2$. *We can conclude from the definition of t-distribution, if we write*

$$\begin{aligned} S_\mu^* \equiv & \frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\hat{\sigma}_n} \\ = & \frac{\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma}}{\sqrt{\frac{\hat{\sigma}_n^2}{\sigma^2}}}, \end{aligned} \tag{4.11}$$

then $S_\mu^* \sim t(n-1)$. *Similarly as in Example 4.1, let* $t_{\alpha/2}(n-1)$ *denote the upper* $100(\alpha/2)$ *percent point of* $t(n-1)$*-distribution, then*

$$\mathbb{P}_{\mu,\sigma^2}\left\{-t_{\alpha/2}(n-1) \le S_\mu^* \le t_{\alpha/2}(n-1)\right\} = 1-\alpha. \tag{4.12}$$

Rewriting it with respect to μ, *we have*

$$\mathbb{P}_{\mu,\sigma^2}\left\{\overline{X}_n - \frac{t_{\alpha/2}(n-1)\hat{\sigma}_n}{\sqrt{n}} \le \mu \le \overline{X}_n + \frac{t_{\alpha/2}(n-1)\hat{\sigma}_n}{\sqrt{n}}\right\} = 1-\alpha, \tag{4.13}$$

hence the confidence interval of μ *with confidence coefficient* $(1-\alpha)$ *is*

$$\left[\overline{X}_n - \frac{t_{\alpha/2}(n-1)\hat{\sigma}_n}{\sqrt{n}}, \overline{X}_n + \frac{t_{\alpha/2}(n-1)\hat{\sigma}_n}{\sqrt{n}}\right], \tag{4.14}$$

(a concrete numerical example is given in Exercise 4.3).

Example 4.3 (Confidence interval of σ^2) *Let* $\chi_\alpha^2(n-1)$ *denote the upper* 100α *percent point of* $\chi_\alpha^2(n-1)$*-distribution. Since the sample variance* $\hat{\sigma}_n^2 = (n-1)^{-1}\sum_{j=1}^{n}\left(X_j - \overline{X}_n\right)^2$ *satisfies the relation (4.10), we have*

$$\mathbb{P}_{\mu,\sigma^2}\left\{\chi_{1-\alpha/2}^2(n-1) \le \frac{(n-1)\hat{\sigma}_n^2}{\sigma^2} \le \chi_{\alpha/2}^2(n-1)\right\} = 1-\alpha, \tag{4.15}$$

hence

$$\mathbb{P}_{\mu,\sigma^2}\left\{\frac{(n-1)\hat{\sigma}_n^2}{\chi_{\alpha/2}^2(n-1)} \le \sigma^2 \le \frac{(n-1)\hat{\sigma}_n^2}{\chi_{1-\alpha/2}^2(n-1)}\right\} = 1-\alpha. \tag{4.16}$$

Therefore, the interval

$$\left[\frac{(n-1)\hat{\sigma}_n^2}{\chi_{\alpha/2}^2(n-1)}, \frac{(n-1)\hat{\sigma}_n^2}{\chi_{1-\alpha/2}^2(n-1)}\right] \tag{4.17}$$

is the confidence interval with confidence coefficient $(1-\alpha)$.

Example 4.4 (Interval estimation for ratio) *Let* $\{X_j\} \sim$ *i.i.d.* $B(1,\theta)$ *(Bernoulli distribution). Then,* $\sum_{j=1}^{n} X_j$ *is distributed as* $B(n,\theta)$ *(binomial distribution) and* $\overline{X}_n = n^{-1}\sum_{j=1}^{n} X_j$ *has mean* θ*, variance* $n^{-1}\theta(1-\theta)$*. By the central limit theorem (Theorem 2.10), we have, as* $n \to \infty$*,*

$$S_\theta(\mathbf{X}) \equiv \frac{\sqrt{n}\left(\overline{X}_n - \theta\right)}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} N(0,1). \tag{4.18}$$

Therefore, for sufficiently large n*, using the percent point* z_α *of the normal distribution (Example 4.1), we obtain approximately*

$$\mathbb{P}\left\{-z_{\alpha/2} \le \frac{\sqrt{n}\left(\overline{X}_n - \theta\right)}{\sqrt{\theta(1-\theta)}} \le z_{\alpha/2}\right\} \approx 1-\alpha. \tag{4.19}$$

Rewriting the event in the above with respect to θ*, the approximate confidence interval of* θ *with confidence coefficient* $(1-\alpha)$ *is given by*

$$\left[\left(1+\frac{z_{\alpha/2}^2}{n}\right)^{-1}\left(\overline{X}_n + \frac{z_{\alpha/2}^2}{2n} - \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\overline{X}_n\left(1-\overline{X}_n\right)+\frac{z_{\alpha/2}^2}{4n}}\right),\right.$$
$$\left.\left(1+\frac{z_{\alpha/2}^2}{n}\right)^{-1}\left(\overline{X}_n + \frac{z_{\alpha/2}^2}{2n} + \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\overline{X}_n\left(1-\overline{X}_n\right)+\frac{z_{\alpha/2}^2}{4n}}\right)\right] \tag{4.20}$$

Since this interval seems to be a little bit complicated, we consider another statistic instead of $S_\theta(\mathbf{X})$ *in (4.18),*

$$S_\theta^*(\mathbf{X}) \equiv \frac{\sqrt{n}\left(\overline{X}_n - \theta\right)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \tag{4.21}$$

in which the dominator $\sqrt{\theta(1-\theta)}$ *of* $S_\theta(\mathbf{X})$ *is replaced by its consistent estimator* $\sqrt{\overline{X}_n(1-\overline{X}_n)}$*. Then, from Slutsky's lemma (Exercise 2.18), we obtain*

$$S_\theta^*(\mathbf{X}) \xrightarrow{d} N(0,1). \tag{4.22}$$

Based on (4.22), it is seen that the approximate confidence interval of θ *with confidence coefficient* $(1-\alpha)$ *is given by*

$$\left[\overline{X}_n - \frac{z_{\alpha/2}\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}, \overline{X}_n + \frac{z_{\alpha/2}\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}\right]. \tag{4.23}$$

4.2 Most Powerful Test

In various fields, one wishes to get yes or no answers to important questions. To try to answer questions we make a hypothesis whose outcomes have some

bearing on the question of interest. The process of deciding whether the hypothesis is correct or not is called *hypothesis testing*. The judge lies between only two decisions: accepting or rejecting the hypothesis. The decision is performed based on the sample. Let $\mathbf{X} = (X_1, \ldots, X_n)' \sim \mathbb{P}_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta \subset \mathbf{R}^q$. Suppose we think that $\boldsymbol{\theta} \in \Theta_0$ or $\boldsymbol{\theta} \in \Theta_1 \equiv \Theta \cap \Theta_0^c$ where Θ_0 is a subset of Θ. We write these hypotheses as

$$H : \boldsymbol{\theta} \in \Theta_0, \qquad A : \boldsymbol{\theta} \in \Theta_1. \tag{4.24}$$

Deciding to take H or A is called the *test* of H against A. The hypothesis H is referred to as the *null hypothesis* while A is referred to as the *alternative hypothesis*. H is called the *simple hypothesis* if Θ_0 consists of one point only, and otherwise it is called the *composite hypothesis*. The terminology is similar for A. Denote the set of all possible values of $\mathbf{X}$ by $\mathcal{X}$. As fundamental testing method we divide $\mathcal{X}$ into two regions W and W^c. If an observed value $\mathbf{x}$ of $\mathbf{X}$ falls into W, the hypothesis H is rejected; otherwise it is accepted. The set W is called the *critical region*, and the set W^c the *acceptance region*.

In the testing procedure, besides the correct decisions, the following two errors occur. (i) A *Type I error* occurs if H is rejected when it is true. (ii) A *Type II error* occurs if H is accepted when A is true. It is desirable to choose a critical region from all possible critical regions which minimizes the probabilities of the two types of error. Unfortunately, in general, both probabilities cannot be controlled simultaneously. Henceforth we define "good" test rules in a more general framework. Let $\phi(\mathbf{x})$ be a measurable function on $\mathcal{X}$ satisfying $0 \leq \phi(\mathbf{x}) \leq 1, \mathbf{x} \in \mathcal{X}$. We consider the testing rule which rejects H with probability $\phi(\mathbf{x})$ when $\mathbf{X} = \mathbf{x}$ is observed. Let $\phi(\mathbf{x})$ be the indicator function of the set W. Then the testing rule by $\phi(\mathbf{x})$ is the one mentioned above. This $\phi(\mathbf{x})$ is called the *critical function*. $\phi(\mathbf{x})$ is called the *nonrandomized test* if $\phi(\mathbf{x})$ is an indicator function of a subset of $\mathcal{X}$, otherwise it is called the *randomized test*. Henceforth we call $\phi(\mathbf{X})$ a *test statistic*. For the test ϕ the probability of a type I error is given by

$$E_{\boldsymbol{\theta}}\{\phi(\mathbf{X})\}, \quad (\boldsymbol{\theta} \in \Theta_0), \tag{4.25}$$

and the probability of a type II error is given by

$$E_{\boldsymbol{\theta}}\{1 - \phi(\mathbf{X})\} = 1 - E_{\boldsymbol{\theta}}\{\phi(\mathbf{X})\}, \quad (\boldsymbol{\theta} \in \Theta_1). \tag{4.26}$$

First, we consider the test ϕ which satisfies

$$\sup_{\boldsymbol{\theta} \in \Theta_0} E_{\boldsymbol{\theta}}\{\phi(\mathbf{X})\} \leq \alpha. \tag{4.27}$$

This means that the supremum of the probability of a type I error is bounded by $\alpha \in (0, 1)$. Here α is a sufficiently small positive number (for example, 0.1, 0.05, 0.01). We say a test ϕ is a (significance) *level* α *test* if it satisfies (4.27). Next, among these level α tests we want to select one which minimizes the probability of a type II error (4.26), i.e., we want to maximize

$$E_{\boldsymbol{\theta}}\{\phi(\mathbf{X})\}, \quad (\boldsymbol{\theta} \in \Theta_1),$$

as a function of $\boldsymbol{\theta} \in \Theta_1$. The function $\beta_\phi(\boldsymbol{\theta}) \equiv E_{\boldsymbol{\theta}}\{\phi(\mathbf{X})\}$, $\boldsymbol{\theta} \in \Theta_1$ is called the *power function* of test ϕ. Thus "good" test is defined as follows:

Definition 4.2 *A level α test ϕ is called the* uniformly most powerful (UMP) test *if*

$$\beta_\phi(\boldsymbol{\theta}) \geq \beta_{\phi^*}(\boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta} \in \Theta_1,$$

for any other level α test ϕ^*. Especially a UMP test is called the most powerful (MP) test *if both null and alternative hypotheses are simple.*

We give the fundamental theorem which provides a systematic method of finding an MP test. Henceforth, we suppose that $\mathbf{X}$ has the probability density function $f_{\boldsymbol{\theta}}(\mathbf{x})$ for simplicity. In discrete case we regard $f_{\boldsymbol{\theta}}(\mathbf{x})$ as a probability function.

Theorem 4.1 (Neyman-Pearson theorem) *Let $\mathbf{X} = (X_1, \ldots, X_n)' \sim \mathbb{P}_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta \subset \mathbf{R}^q$, and $\mathbf{X}$ have the probability density function $f_{\boldsymbol{\theta}}(\boldsymbol{x})$. For the testing problem*

$$H : \boldsymbol{\theta} = \boldsymbol{\theta}_0, \qquad A : \boldsymbol{\theta} = \boldsymbol{\theta}_1, \quad (\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0), \tag{4.28}$$

an MP test $\phi_0(\mathbf{x})$ of level α ($\in (0,1)$) is given by

$$\phi_0(\mathbf{x}) = \begin{cases} 1, & (\textit{when } f_{\boldsymbol{\theta}_1}(\mathbf{x}) > k f_{\boldsymbol{\theta}_0}(\mathbf{x})), \\ \gamma, & (\textit{when } f_{\boldsymbol{\theta}_1}(\mathbf{x}) = k f_{\boldsymbol{\theta}_0}(\mathbf{x})), \\ 0, & (\textit{when } f_{\boldsymbol{\theta}_1}(\mathbf{x}) < k f_{\boldsymbol{\theta}_0}(\mathbf{x})), \end{cases} \tag{4.29}$$

where γ and k ($0 \leq \gamma \leq 1$, $k \geq 0$) are determined by

$$E_{\boldsymbol{\theta}_0}\{\phi_0(\mathbf{X})\} = \alpha.$$

PROOF Define

$$\begin{aligned} B_1 &\equiv \{\mathbf{x} \in \mathbf{R}^n : f_{\boldsymbol{\theta}_1}(\mathbf{x}) > k f_{\boldsymbol{\theta}_0}(\mathbf{x})\}, \\ B_2 &\equiv \{\mathbf{x} \in \mathbf{R}^n : f_{\boldsymbol{\theta}_1}(\mathbf{x}) = k f_{\boldsymbol{\theta}_0}(\mathbf{x})\}, \\ B_3 &\equiv \{\mathbf{x} \in \mathbf{R}^n : f_{\boldsymbol{\theta}_1}(\mathbf{x}) < k f_{\boldsymbol{\theta}_0}(\mathbf{x})\}. \end{aligned}$$

Let $\phi(\mathbf{X})$ be a level α test of the testing problem (4.28). Then $\phi(\mathbf{X})$ satisfies

$$E_{\boldsymbol{\theta}_0}\{\phi(\mathbf{X})\} \leq \alpha. \tag{4.30}$$

From (4.29) we obtain

$$\begin{aligned}
E_{\boldsymbol{\theta}_1}\{\phi_0(\mathbf{X})\} - E_{\boldsymbol{\theta}_1}\{\phi(\mathbf{X})\} &= \int_{\mathbf{R}^n} \{\phi_0(\mathbf{x}) - \phi(\mathbf{x})\} f_{\boldsymbol{\theta}_1}(\mathbf{x}) d\mathbf{x} \\
&= \int_{B_1} \{1 - \phi(\mathbf{x})\} f_{\boldsymbol{\theta}_1}(\mathbf{x}) d\mathbf{x} \\
&\quad + \int_{B_2} \{\gamma - \phi(\mathbf{x})\} f_{\boldsymbol{\theta}_1}(\mathbf{x}) d\mathbf{x} \\
&\quad + \int_{B_3} \{-\phi(\mathbf{x})\} f_{\boldsymbol{\theta}_1}(\mathbf{x}) d\mathbf{x} \\
&\geq \int_{B_1} \{1 - \phi(\mathbf{x})\} k f_{\boldsymbol{\theta}_0}(\mathbf{x}) d\mathbf{x} \\
&\quad + \int_{B_2} \{\gamma - \phi(\mathbf{x})\} k f_{\boldsymbol{\theta}_0}(\mathbf{x}) d\mathbf{x} \\
&\quad + \int_{B_3} \{-\phi(\mathbf{x})\} k f_{\boldsymbol{\theta}_0}(\mathbf{x}) d\mathbf{x} \\
&= k \int_{B_1} \{1 - \phi(\mathbf{x})\} f_{\boldsymbol{\theta}_0}(\mathbf{x}) d\mathbf{x} \\
&\quad + k \int_{B_2} \{\gamma - \phi(\mathbf{x})\} f_{\boldsymbol{\theta}_0}(\mathbf{x}) d\mathbf{x} \\
&\quad + k \int_{B_3} \{-\phi(\mathbf{x})\} f_{\boldsymbol{\theta}_0}(\mathbf{x}) d\mathbf{x} \\
&= k \int_{\mathbf{R}^n} \{\phi_0(\mathbf{x}) - \phi(\mathbf{x})\} f_{\boldsymbol{\theta}_0}(\mathbf{x}) d\mathbf{x} \\
&= k[E_{\boldsymbol{\theta}_0}\{\phi_0(\mathbf{x})\} - E_{\boldsymbol{\theta}_0}\{\phi(\mathbf{X})\}] \\
&= k[\alpha - E_{\boldsymbol{\theta}_0}\{\phi(\mathbf{X})\}] \geq 0, \quad \text{(by (4.30))},
\end{aligned}$$

which implies $E_{\boldsymbol{\theta}_1}\{\phi_0(\mathbf{X})\} \geq E_{\boldsymbol{\theta}_1}\{\phi(\mathbf{X})\}$. Thus it follows that ϕ_0 is an MP test of level α. □

Let us see examples of MP tests.

Example 4.5 *Let $X_1, \ldots, X_n \sim$ i.i.d. $N(\theta, \sigma^2)$ where σ^2 is known. Then for the testing problem*

$$H : \theta = \theta_0, \qquad A : \theta = \theta_1, \quad (\theta_1 > \theta_0) \tag{4.31}$$

we seek an MP test. Since the probability density function of $\mathbf{X} = (X_1, \ldots, X_n)'$ is given by

$$f_\theta(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^n (x_j - \theta)^2\right], \qquad (\mathbf{x} = (x_1, \ldots, x_n)'),$$

$$\log\left[\frac{f_{\theta_1}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})}\right] = -\frac{1}{2\sigma^2}\left[\sum_{j=1}^{n}(x_j-\theta_1)^2 - \sum_{j=1}^{n}(x_j-\theta_0)^2\right]$$
$$= \frac{(\theta_1-\theta_0)}{\sigma^2}\left(\bar{x}_n - \frac{\theta_0+\theta_1}{2}\right), \quad \left(\bar{x}_n = n^{-1}\sum_{j=1}^{n} x_j\right). \tag{4.32}$$

Hence we get

$$f_{\theta_1}(\mathbf{x}) > k f_{\theta_0}(\mathbf{x}) \Longleftrightarrow \log\frac{f_{\theta_1}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})} > \log k$$
$$\Longleftrightarrow \bar{x}_n > k', \quad (by\ (4.32)).$$

Note that $\mathbf{X}$ *is a continuous random variable. Because the probability of* $\bar{X}_n(\equiv n^{-1}\sum_{j=1}^{n} X_j) = k$ *is* 0*, an MP test is*

$$\phi_0(\mathbf{x}) = \begin{cases} 1, & \bar{x}_n > k', \\ 0, & \bar{x}_n < k', \end{cases}$$

where the constant k' *satisfies*

$$\alpha = \mathrm{P}_{\theta_0}\{\bar{X}_n > k'\}. \tag{4.33}$$

(4.33) *can be rewritten as*

$$\alpha = \mathrm{P}_{\theta_0}\left\{\frac{\sqrt{n}(\bar{X}_n-\theta_0)}{\sigma} > \frac{\sqrt{n}(k'-\theta_0)}{\sigma}\right\}. \tag{4.34}$$

From $\sqrt{n}(\bar{X}_n-\theta_0)/\sigma \sim N(0,1)$ *under* $\theta=\theta_0$, (4.34) *is equivalent to*

$$\alpha = 1 - \Phi\left[\frac{\sqrt{n}(k'-\theta_0)}{\sigma}\right],$$

where

$$\Phi(z) = \int_{-\infty}^{z}\left(\frac{1}{2\pi}\right)^{1/2}\exp\left(-\frac{x^2}{2}\right)dx.$$

Denote the upper $100\alpha\%$ *point of the standard normal distribution by* z_α*. Then we obtain*

$$\frac{\sqrt{n}(k'-\theta_0)}{\sigma} = z_\alpha, \quad i.e., \quad k' = \theta_0 + \frac{z_\alpha\sigma}{\sqrt{n}}.$$

Therefore the MP test of level α *is*

$$\phi_0(\mathbf{x}) = \begin{cases} 1, & \bar{x}_n > \theta_0 + z_\alpha\sigma/\sqrt{n}, \\ 0, & \bar{x}_n < \theta_0 + z_\alpha\sigma/\sqrt{n}. \end{cases} \tag{4.35}$$

The power of ϕ_0 is given by

$$\begin{aligned}\beta_{\phi_0}(\theta_1) &\equiv E_{\theta_1}\{\phi_0(\mathbf{X})\} = \mathrm{P}_{\theta_1}\left\{\bar{X}_n > \theta_0 + \frac{z_\alpha \sigma}{\sqrt{n}}\right\} \\ &= \mathrm{P}_{\theta_1}\left\{\frac{\sqrt{n}(\bar{X}_n - \theta_1)}{\sigma} > z_\alpha - \frac{\sqrt{n}(\theta_1 - \theta_0)}{\sigma}\right\}.\end{aligned} \tag{4.36}$$

If $\theta = \theta_1$, then $\sqrt{n}(\bar{X}_n - \theta_1)/\sigma \sim N(0,1)$. From (4.36) we can see that the power increases as the mean difference $\theta_1 - \theta_0$ increases. Also we can observe that the power increases as the sample size n increases. □

The following is an example in the case of discrete distributions.

Example 4.6 *Let $X_1, \ldots, X_n \sim$ i.i.d. $B(1,\theta)$ (Bernoulli distribution). Then for the testing problem*

$$H : \theta = \theta_0, \qquad A : \theta = \theta_1, \quad (\theta_1 > \theta_0) \tag{4.37}$$

we seek an MP test. Since the probability function of $\mathbf{X} = (X_1, \ldots, X_n)'$ is given by

$$f_\theta(\mathbf{x}) = \prod_{j=1}^{n} \theta^{x_j}(1-\theta)^{1-x_j}, \qquad (\mathbf{x} = (x_1, \ldots, x_n)'), \tag{4.38}$$

$$\log \frac{f_{\theta_1}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})} = \left\{\sum_{j=1}^{n} x_j\right\} \log\left\{\frac{\theta_1(1-\theta_0)}{(1-\theta_1)\theta_0}\right\} + n \log\left\{\frac{1-\theta_1}{1-\theta_0}\right\}.$$

From $\theta_1 > \theta_0$ we have

$$\begin{aligned} f_{\theta_1}(\mathbf{x}) > k f_{\theta_0}(\mathbf{x}) &\Longleftrightarrow \log \frac{f_{\theta_1}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})} > k' \\ &\Longleftrightarrow \sum_{j=1}^{n} x_j > k''. \end{aligned}$$

Hence an MP test of level α is

$$\phi_0(\mathbf{x}) = \begin{cases} 1, & (\sum_{j=1}^{n} x_j > k''), \\ \gamma, & (\sum_{j=1}^{n} x_j = k''), \\ 0, & (\sum_{j=1}^{n} x_j < k''), \end{cases} \tag{4.39}$$

where γ and k'' are determined by

$$\begin{aligned}\alpha &= E_{\theta_0}\{\phi_0(\mathbf{X})\} \\ &= \mathrm{P}_{\theta_0}\left\{\sum_{j=1}^{n} X_j > k''\right\} + \gamma \mathrm{P}_{\theta_0}\left\{\sum_{j=1}^{n} X_j = k''\right\}.\end{aligned} \tag{4.40}$$

Note that $\sum_{j=1}^{n} X_j$ *has the binomial distribution* $B(n, \theta_0)$ *under* $\theta = \theta_0$. (4.40) *can be rewritten as*

$$\alpha = \sum_{j=k''+1}^{n} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j} + \gamma \binom{n}{k''} \theta_0^{k''} (1-\theta_0)^{n-k''}. \tag{4.41}$$

First, we find an integer k'' *which satisfies*

$$\sum_{j=k''+1}^{n} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j} \le \alpha < \sum_{j=k''}^{n} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j}. \tag{4.42}$$

Denote this integer k'' *by* k_0. *Then* γ *is given by*

$$\gamma = \left[\alpha - \sum_{j=k_0+1}^{n} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j} \right] \Bigg/ \binom{n}{k_0} \theta_0^{k_0} (1-\theta_0)^{n-k_0}. \tag{4.43}$$

Since this calculation is considerably complicated when n *is large, we formerly needed normal approximations of* $\sum_{j=1}^{n} X_j$. *However, recently, we can easily calculate the above* k'' *and* γ *using various statistical softwares.* □

In Examples 4.5 and 4.6, tests (4.35) and (4.39) are independent of the values of θ_1 in alternative hypotheses. Therefore these tests are UMP tests for the testing problem

$$H : \theta = \theta_0, \qquad A : \theta > \theta_0$$

The following theorem summarizes the above argument more generally.

Theorem 4.2 *Suppose that* $\mathbf{X} = (X_1, \ldots, X_n)'$ *has the following probability density function*

$$f_\theta(\mathbf{x}) = c(\theta) \exp\{Q(\theta) T(\mathbf{x})\} h(\mathbf{x}), \qquad (\mathbf{x} = (x_1, \ldots, x_n)'), \tag{4.44}$$

where $\theta \in \Theta$, Θ *is an interval on* $\mathbf{R}$ *and* $Q(\theta)$ *is an increasing function on* Θ. *Then for the testing problem*

$$H : \theta = \theta_0, \qquad A : \theta > \theta_0 \tag{4.45}$$

a UMP test of level α *is given by*

$$\phi_0(\mathbf{x}) = \begin{cases} 1, & (\textit{when } T(\mathbf{x}) > k), \\ \gamma, & (\textit{when } T(\mathbf{x}) = k), \\ 0, & (\textit{when } T(\mathbf{x}) < k), \end{cases} \tag{4.46}$$

where k *and* γ $(k \ge 0,\ 0 \le \gamma \le 1)$ *are determined by*

$$\mathrm{E}_{\theta_0}\{\phi_0(\mathbf{x})\} = \alpha. \tag{4.47}$$

PROOF For the testing problem of simple hypotheses

$$H : \theta = \theta_0, \qquad A' : \theta = \theta_1, \quad (\theta_1 > \theta_0) \tag{4.48}$$

an MP test is given in Theorem 4.1. Since $Q(\theta)$ is an increasing function, we have

$$f_{\theta_1}(\mathbf{x}) > k f_{\theta_0}(\mathbf{x}) \Longleftrightarrow T(\mathbf{x}) > k'.$$

Hence an MP test of (4.48) is given by (4.46), where k is determined by (4.47). Note that the test ϕ_0 is independent of θ_1. It follows that ϕ_0 is a UMP test of (4.45). □

4.3 Various Tests

The previous section discussed MP and UMP tests of specific hypotheses. It is known that there do not exist UMP tests of composite hypotheses in general. However, if we restrict the class of tests, then there exists a uniformly most powerful test in this class. This section explains these results. Let $\mathbf{X} \sim \mathbb{P}_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta \subset \mathbf{R}^q$, and Θ_0 be a subset of Θ. Consider the problem of testing

$$H : \boldsymbol{\theta} \in \Theta_0, \qquad A : \boldsymbol{\theta} \in \Theta_1 \equiv \Theta - \Theta_0. \tag{4.49}$$

A level α test ϕ of this hypothesis is said to be an *unbiased test* of level α, if the power function satisfies

$$\beta_\phi(\boldsymbol{\theta}) \equiv E_{\boldsymbol{\theta}}\{\phi(\mathbf{X})\} \geq \alpha \tag{4.50}$$

for all $\boldsymbol{\theta} \in \Theta_1$. A level α unbiased test ϕ is called the *uniformly most powerful unbiased (UMPU) test* if

$$\beta_\phi(\boldsymbol{\theta}) \geq \beta_{\phi^*}(\boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta} \in \Theta_1,$$

for any other level α unbiased test ϕ^*. UMPU tests for the means and variances of normal distributions are important in the application, and are given in the following examples, which state only the results. See, e.g., Lehmann (1986) for the reader interested in the proofs. In the following two examples, let $X_1, X_2, \ldots, X_n \sim$ i.i.d. $N(\mu, \sigma^2)$, and define

$$\begin{aligned} &\bar{X}_n = n^{-1}\sum_{j=1}^n X_j, \qquad \hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \bar{X}_n)^2, \\ &T(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\hat{\sigma}_n}, \end{aligned} \tag{4.51}$$

where μ_0 is a given constant and $\mathbf{X} = (X_1, \ldots, X_n)'$. Also, let $\mathbf{x} = (x_1, \ldots, x_n)'$.

Example 4.7 *(One-sided t-test) Suppose we want to test*

$$H : \mu \leq \mu_0, \qquad A : \mu > \mu_0. \tag{4.52}$$

The hypothesis of the form $A : \mu > \mu_0$ *or* $A : \mu < \mu_0$ *is called a* one-sided hypothesis. *For the testing problem* (4.52) *the test*

$$\phi(\mathbf{x}) = \begin{cases} 1, & (\textit{when } T(\mathbf{x}) > t_\alpha(n-1)), \\ 0, & (\textit{when } T(\mathbf{x}) < t_\alpha(n-1)) \end{cases} \tag{4.53}$$

is called a one-sided *t*-test. *Then ϕ is a UMPU test of level α, where $t_\alpha(n-1)$ is the upper $100\alpha\%$ point of the t-distribution with $n-1$ degrees of freedom.* □

Example 4.8 *(Two-sided t-test) Suppose we want to test*

$$H : \mu = \mu_0, \qquad A : \mu \neq \mu_0. \tag{4.54}$$

The hypothesis of the form $A : \mu \neq \mu_0$, or $A : \mu \leq \mu_1$ or $\mu \geq \mu_2$ $(\mu_1 < \mu_2)$ is called a two-sided hypothesis. *For the testing problem* (4.54) *the test*

$$\phi(\mathbf{x}) = \begin{cases} 1, & (\textit{when } |T(\mathbf{x})| > t_{\alpha/2}(n-1)), \\ 0, & (\textit{when } |T(\mathbf{x})| < t_{\alpha/2}(n-1)) \end{cases} \tag{4.55}$$

is called a two-sided *t*-test*, and is a UMPU test of level α.* □

We often want to compare two samples from different populations. Such a problem is called a *two sample problem*. We consider UMPU tests for two sample problems in the normal distribution. Let $X_1, X_2, \ldots, X_m \sim$ i.i.d. $N(\mu_1, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_n \sim$ i.i.d. $N(\mu_2, \sigma^2)$. It is assumed that $\mathbf{X} = (X_1, \ldots, X_m)'$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ are mutually independent. Write $\mathbf{x} = (x_1, \ldots, x_m)'$ and $\mathbf{y} = (y_1, \ldots, y_n)'$. Also define

$$\bar{X}_m = \frac{1}{m}\sum_{j=1}^{m} X_j, \qquad \bar{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j,$$

$$\hat{\sigma}_X^2 = \frac{1}{m-1}\sum_{j=1}^{m}(X_j - \bar{X}_m)^2, \qquad \hat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y}_n)^2,$$

$$T(\mathbf{X}, \mathbf{Y}) = \frac{(\bar{X}_m - \bar{Y}_n)\Big/\sqrt{\frac{1}{m} + \frac{1}{n}}}{\sqrt{\frac{(m-1)\hat{\sigma}_X^2 + (n-1)\hat{\sigma}_Y^2}{m+n-2}}}.$$

Example 4.9 *(Two sample one-sided t-test) For the testing problem*

$$H : \mu_1 \leq \mu_2, \qquad A : \mu_1 > \mu_2, \tag{4.56}$$

$$\phi(\mathbf{x}, \mathbf{y}) = \begin{cases} 1, & (\textit{when } T(\mathbf{x}, \mathbf{y}) > t_\alpha(m+n-2)), \\ 0, & (\textit{when } T(\mathbf{x}, \mathbf{y}) < t_\alpha(m+n-2)) \end{cases} \tag{4.57}$$

is a UMPU test of level α. □

Example 4.10 *(Two sample two-sided t-test) For the testing problem*

$$H : \mu_1 = \mu_2, \qquad A : \mu_1 \neq \mu_2, \tag{4.58}$$

$$\phi(\mathbf{x}, \mathbf{y}) = \begin{cases} 1, & (\textit{when } |T(\mathbf{x}, \mathbf{y})| > t_{\alpha/2}(m+n-2)), \\ 0, & (\textit{when } |T(\mathbf{x}, \mathbf{y})| < t_{\alpha/2}(m+n-2)) \end{cases} \tag{4.59}$$

is a UMPU test of level α. □

Next, we consider tests for the difference of variances between two populations. For this purpose we define the following distribution.

Definition 4.3 *If $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, and if X and Y are independent, then the distribution of*

$$F \equiv \frac{\frac{1}{m}X}{\frac{1}{n}Y}$$

is called the *F-distribution with (m, n) degrees of freedom*, and henceforth we write $F \sim F(m, n)$. See Exercise 4.5 for the probability density function of $F(m, n)$.

In the following, let $X_1, X_2, \ldots, X_m \sim$ i.i.d. $N(\mu_1, \sigma^2)$, $Y_1, Y_2, \ldots, Y_n \sim$ i.i.d. $N(\mu_2, \sigma^2)$, and let $\mathbf{X} = (X_1, \ldots, X_m)'$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ be independent.

Example 4.11 *Consider the testing problem*

$$H : \sigma_1^2 \leq \sigma_2^2, \qquad A : \sigma_1^2 > \sigma_2^2. \tag{4.60}$$

Define

$$F(\mathbf{x}, \mathbf{y}) \equiv \frac{\sum_{j=1}^m (x_j - \bar{x}_m)^2/(m-1)}{\sum_{j=1}^n (y_j - \bar{y}_n)^2/(n-1)},$$

where $\mathbf{x} = (x_1, \ldots, x_m)'$, $\mathbf{y} = (y_1, \ldots, y_n)'$, $\bar{x}_m = m^{-1}\sum_{j=1}^m x_j$ and $\bar{y}_n = n^{-1}\sum_{j=1}^n y_j$. Then the test

$$\phi(\mathbf{x}, \mathbf{y}) = \begin{cases} 1, & (\textit{when } F(\mathbf{x}, \mathbf{y}) > F_\alpha(m-1, n-1)), \\ 0, & (\textit{when } F(\mathbf{x}, \mathbf{y}) < F_\alpha(m-1, n-1)) \end{cases} \tag{4.61}$$

is a UMPU test of level α. □

Section 4.1 explained construction procedures of confidence intervals (sets). Here we discuss "goodness" of confidence intervals (sets). In fact, there exists a close relationship between UMPU tests and "goodness" of confidence intervals. First let $\mathbf{X} \sim \mathbb{P}_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \Theta \subset \mathbf{R}^q$, and $\mathcal{X}$ be the set of all possible values of $\mathbf{X}$. A confidence set $C(\mathbf{X})$ with confidence coefficient $(1-\alpha)$ was defined by

$$\mathbb{P}_{\boldsymbol{\theta}}\{\boldsymbol{\theta} \in C(\mathbf{X})\} \geq 1 - \alpha. \tag{4.62}$$

In addition, if $C(\mathbf{X})$ satisfies

$$\mathbb{P}_{\boldsymbol{\theta}'}\{\boldsymbol{\theta} \in C(\mathbf{X})\} \leq 1 - \alpha \tag{4.63}$$

for all $\boldsymbol{\theta}'$, $\boldsymbol{\theta}$ $(\boldsymbol{\theta}' \neq \boldsymbol{\theta}) \in \Theta$, then $C(\mathbf{X})$ is said to be an *unbiased confidence set*. We say that $C_0(\mathbf{X})$ is a *uniformly most powerful confidence set* if $C_0(\mathbf{X})$ minimizes (4.63) among unbiased confidence sets with confidence coefficient $(1-\alpha)$ for all $\boldsymbol{\theta}'$, $\boldsymbol{\theta}$ $(\boldsymbol{\theta}' \neq \boldsymbol{\theta}) \in \Theta$. Now we construct $C_0(\mathbf{X})$. Suppose that for the testing problem

$$H : \boldsymbol{\theta} = \boldsymbol{\theta}_0, \qquad A : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \tag{4.64}$$

a UMPU test of level α exists. Denote the corresponding acceptance region by $W^c(\boldsymbol{\theta}_0)$. Set

$$I(\mathbf{x}) \equiv \{\boldsymbol{\theta} : \mathbf{x} \in W^c(\boldsymbol{\theta})\} \tag{4.65}$$

for $\mathbf{x} \in \mathcal{X}$. Then it is seen that

$$\mathbf{x} \in W^c(\boldsymbol{\theta}) \Longleftrightarrow \boldsymbol{\theta} \in I(\mathbf{x}), \tag{4.66}$$

which implies

$$\mathbb{P}_{\boldsymbol{\theta}}\{\mathbf{X} \in W^c(\boldsymbol{\theta})\} = \mathbb{P}_{\boldsymbol{\theta}}\{\boldsymbol{\theta} \in I(\mathbf{X})\}. \tag{4.67}$$

Recalling the definition of unbiased tests, we get

$$\mathbb{P}_{\boldsymbol{\theta}}\{\mathbf{X} \in W^c(\boldsymbol{\theta})\} = \mathbb{P}_{\boldsymbol{\theta}}\{\boldsymbol{\theta} \in I(\mathbf{X})\} \geq 1 - \alpha, \tag{4.68}$$

$$\mathbb{P}_{\boldsymbol{\theta}'}\{\mathbf{X} \in W^c(\boldsymbol{\theta})\} = \mathbb{P}_{\boldsymbol{\theta}'}\{\boldsymbol{\theta} \in I(\mathbf{X})\} \leq 1 - \alpha \tag{4.69}$$

for $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$. Let $J(\mathbf{X})$ be any unbiased confidence set satisfying (4.62) and (4.63), and $U^c(\boldsymbol{\theta}_0)$ be the corresponding acceptance region. Similarly it is clear that

$$\mathbb{P}_{\boldsymbol{\theta}}\{\mathbf{X} \in U^c(\boldsymbol{\theta})\} = \mathbb{P}_{\boldsymbol{\theta}}\{\boldsymbol{\theta} \in J(\mathbf{X})\} \geq 1 - \alpha,$$
$$\mathbb{P}_{\boldsymbol{\theta}'}\{\mathbf{X} \in U^c(\boldsymbol{\theta})\} = \mathbb{P}_{\boldsymbol{\theta}'}\{\boldsymbol{\theta} \in J(\mathbf{X})\} \leq 1 - \alpha.$$

On the other hand, since $W^c(\boldsymbol{\theta})$ is the acceptance region of a level α UMPU test,

$$\mathbb{P}_{\boldsymbol{\theta}'}\{\mathbf{X} \in U(\boldsymbol{\theta})\} \leq \mathbb{P}_{\boldsymbol{\theta}'}\{\mathbf{X} \in W(\boldsymbol{\theta})\}, \qquad (\boldsymbol{\theta} \neq \boldsymbol{\theta}').$$

Hence, we have

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}'}\{\boldsymbol{\theta} \in I(\mathbf{X})\} &= \mathbb{P}_{\boldsymbol{\theta}'}\{\mathbf{X} \in W^c(\boldsymbol{\theta})\} \\ &\leq \mathbb{P}_{\boldsymbol{\theta}'}\{\mathbf{X} \in U^c(\boldsymbol{\theta})\} = \mathbb{P}_{\boldsymbol{\theta}'}\{\boldsymbol{\theta} \in J(\mathbf{X})\}\end{aligned}$$

for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}' \in \Theta$, which implies that the confidence set $I(\mathbf{x})$ corresponding to the acceptance region of a UMPU test is a uniformly most powerful confidence set.

Example 4.12 *Since the UMPU test for (4.54) was given by (4.55) in Example 4.8, a uniformly most powerful confidence set of level α for the mean μ of a normal distribution is*

$$\left[\bar{X}_n - \frac{\hat{\sigma}_n}{\sqrt{n}} t_{\alpha/2}(n-1), \bar{X}_n + \frac{\hat{\sigma}_n}{\sqrt{n}} t_{\alpha/2}(n-1)\right].$$

□

4.4 Discriminant Analysis

There are various methods and techniques in statistical analysis. In this section, we give a brief description of discriminant analysis, making a point of

multivariate normal sample. The discriminant analysis is fundamental and important in the field of financial engineering. We consider the case when we know a sample $\mathbf{X}$ belongs to one of several categories which are described by the probability distributions, but we do not know it belongs to which. We need to select the category to which $\mathbf{X}$ belongs, with possibly high probability.

First, we consider the two categories case. Assume an m-dimensional random vector $\mathbf{X}$ has the probability density function $f(\mathbf{x})$, $\mathbf{x} \in \mathbf{R}^m$ and we know $f(\mathbf{x})$ belongs to one of the following two categories Π_j, $j = 1, 2$:

$$\Pi_1 : f(\mathbf{x}) = f_1(\mathbf{x}), \quad \Pi_2 : f(\mathbf{x}) = f_2(\mathbf{x}). \tag{4.70}$$

Decompose $\mathbf{R}^m$ into two exclusive regions $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathbf{R}^m = \mathcal{R}_1 \cup \mathcal{R}_2$. When we observed $\mathbf{X} = \mathbf{x}$, if $\mathbf{x} \in \mathcal{R}_1$, we assign $\mathbf{X}$ to Π_1 and if $\mathbf{x} \in \mathcal{R}_2$, we assign $\mathbf{X}$ to Π_2. Then, we say that $\mathbf{X}$ is classified by the *classification rule* $\mathcal{R} = (\mathcal{R}_1, \mathcal{R}_2)$. Of course we want to seek a "good" classification rule. For $\mathcal{R}$, the probability

$$P(j|k) \equiv \int_{\mathcal{R}_j} f_k(\mathbf{x}) d\mathbf{x}, \quad j, k = 1, 2 \ (j \neq k) \tag{4.71}$$

is the *misclassification probability* when $\mathbf{X}$ is misclassified to Π_j although in fact it belongs to Π_k, $(k \neq j)$. Here, we seek the $\mathcal{R}$ which minimizes the sum

$$P(2|1) + P(1|2) \tag{4.72}$$

and call this the *optimal classification rule*. Since (4.72) becomes

$$\begin{aligned} &\int_{\mathcal{R}_2} f_1(\mathbf{x}) d\mathbf{x} + \int_{\mathcal{R}_1} f_2(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{R}_2} \{f_1(\mathbf{x}) d\mathbf{x} - f_2(\mathbf{x})\} + \int_{\mathbf{R}^m} f_2(\mathbf{x}) d\mathbf{x}, \end{aligned} \tag{4.73}$$

we seek the region $\mathcal{R}_2$ which minimizes the integral of the first term in the right-hand side of (4.73). Let $\mathcal{R}_2$ include all of $\mathbf{x}$ such that $f_1(\mathbf{x}) - f_2(\mathbf{x}) < 0$ and not include $\mathbf{x}$ such that $f_1(\mathbf{x}) - f_2(\mathbf{x}) \geq 0$, then $\mathcal{R}_2$ minimizes it. Hence, we have following theorem:

Theorem 4.3 *The optimal classification rule for the discriminant problem (4.70) is given by*

$$\mathcal{R}_1 = \{\mathbf{x} \in \mathbf{R}^m : f_1(\mathbf{x}) \geq f_2(\mathbf{x})\}, \tag{4.74}$$

$$\mathcal{R}_2 = \{\mathbf{x} \in \mathbf{R}^m : f_1(\mathbf{x}) < f_2(\mathbf{x})\}. \tag{4.75}$$

Now, let us see an example of the above result for concrete distribution.

Example 4.13 (Discrimination between $N(\boldsymbol{\mu}^{(1)}, \Sigma)$ and $N(\boldsymbol{\mu}^{(2)}, \Sigma)$) *Assume that the probability density functions of $f_i(\mathbf{x})$, $(i = 1, 2)$ is given by*

$$f_i(\mathbf{x}) = (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}^{(i)}\right)' \Sigma^{-1} \left(\mathbf{x} - \boldsymbol{\mu}^{(i)}\right)\right\}, \tag{4.76}$$

where Σ *is an* $m \times m$ *positive matrix and* $\boldsymbol{\mu}^{(i)} = \left(\mu_1^{(i)}, \ldots, \mu_m^{(i)}\right)'$. *In this case the optimal classification rule* $\mathcal{R}$ *given in (4.74) and (4.75) become*

$$\mathcal{R}_1 = \left\{ \mathbf{x} \in \mathbf{R}^m : \mathbf{x}'\Sigma^{-1}\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\boldsymbol{\mu}^{(1)} + \mu^{(2)}\right)'\Sigma^{-1}\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right) \geq 0 \right\}, \tag{4.77}$$

$$\mathcal{R}_2 = \left\{ \mathbf{x} \in \mathbf{R}^m : \mathbf{x}'\Sigma^{-1}\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\boldsymbol{\mu}^{(1)} + \mu^{(2)}\right)'\Sigma^{-1}\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right) < 0 \right\} \tag{4.78}$$

(Exercise 4.6). To evaluate the misclassification probability of this criterion we define the random variable

$$U = \mathbf{X}'\Sigma^{-1}\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\boldsymbol{\mu}^{(1)} + \mu^{(2)}\right)'\Sigma^{-1}\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right). \tag{4.79}$$

We call this the discriminant function. *Henceforth, we denote the expectation and variance under* $f_i(\mathbf{x})$ $i = 1, 2$ *by* $E_i(\cdot)$ *and* $V_i(\cdot)$, *respectively. Then, we have*

$$\begin{aligned} &E_1(U) \\ &= \boldsymbol{\mu}^{(1)\prime}\Sigma^{-1}\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\boldsymbol{\mu}^{(1)} + \mu^{(2)}\right)'\Sigma^{-1}\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right) \\ &= \frac{1}{2}\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right)'\Sigma^{-1}\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right), \end{aligned} \tag{4.80}$$

$$\begin{aligned} V_1(U) &= V_1\left\{\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right)'\Sigma^{-1}\mathbf{X}\right\} \\ &= \left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right)'\Sigma^{-1}V_1(\mathbf{X})\Sigma^{-1}\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right) \\ &= \left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right)'\Sigma^{-1}\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right). \end{aligned} \tag{4.81}$$

The quantity

$$\Delta^2 \equiv \left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right)'\Sigma^{-1}\left(\boldsymbol{\mu}^{(1)} - \mu^{(2)}\right) \tag{4.82}$$

is called the Mahalanobis distance *between* $N\left(\boldsymbol{\mu}^{(1)}, \Sigma\right)$ *and* $N\left(\boldsymbol{\mu}^{(2)}, \Sigma\right)$. *Therefore, if* $\mathbf{X} \sim N\left(\boldsymbol{\mu}^{(1)}, \Sigma\right)$, *then* $U \sim N\left(\frac{1}{2}\Delta^2, \Delta^2\right)$. *Similarly, it is seen that if* $\mathbf{X} \sim N\left(\boldsymbol{\mu}^{(2)}, \Sigma\right)$, *then*

$$U \sim N\left(-\frac{1}{2}\Delta^2, \Delta^2\right) \tag{4.83}$$

(Exercise 4.7). From the above the misclassification probabilities of $\mathcal{R} = \{\mathcal{R}_1, \mathcal{R}_2\}$

defined in (4.77) and (4.78) become

$$P(2|1) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\Delta} e^{-\frac{1}{2}\left(z-\frac{\Delta^2}{2}\right)^2/\Delta^2} dz = \int_{-\infty}^{-\frac{\Delta}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz, \tag{4.84}$$

$$P(1|2) = \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}\Delta} e^{-\frac{1}{2}\left(z+\frac{\Delta^2}{2}\right)^2/\Delta^2} dz = \int_{\frac{\Delta}{2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz. \tag{4.85}$$

Writing $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$, *we have*

$$P(2|1) + P(1|2) = 2\left\{1 - \Phi\left(\frac{\Delta}{2}\right)\right\}. \tag{4.86}$$

Hence, as the Mahalanobis distance tends to large, the equation (4.86) $\searrow 0$, *which implies the discriminant procedure based on* U *works well.*

Up to now we assumed that $f_i(\mathbf{x})$, $i = 1, 2$ are known, which describe each of two categories. In actual cases, however, it is more plausible that we do not know them. In such a case, if we have samples $\mathbf{X}^{(1)} \equiv \left(\mathbf{X}_1^{(1)}, \ldots, \mathbf{X}_{n_1}^{(1)}\right)'$ and $\mathbf{X}^{(2)} \equiv \left(\mathbf{X}_1^{(2)}, \ldots, \mathbf{X}_{n_2}^{(2)}\right)'$ which are known to belong to Π_1 and Π_2, respectively, then we can estimate $f_i(\mathbf{x})$ from the samples and use them in the discrimination problem above. We call such samples $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ the *training samples*. While we can use U in (4.79) for the discrimination problem under the setting in Example 4.13, in this case, we have to estimate unknown $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$ and Σ from $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. Hence, substituting

$$\overline{\mathbf{X}}^{(1)} \equiv n_1^{-1} \sum_{j=1}^{n_1} \mathbf{X}_j^{(1)}, \quad \overline{\mathbf{X}}^{(2)} \equiv n_2^{-1} \sum_{j=1}^{n_2} \mathbf{X}_j^{(2)} \tag{4.87}$$

and

$$\mathbf{S} \equiv \frac{1}{n_1 + n_2 - 2}\left[\sum_{j=1}^{n_1}\left(\mathbf{X}_j^{(1)} - \overline{\mathbf{X}}^{(1)}\right)\left(\mathbf{X}_j^{(1)} - \overline{\mathbf{X}}^{(1)}\right)' + \sum_{j=1}^{n_2}\left(\mathbf{X}_j^{(2)} - \overline{\mathbf{X}}^{(2)}\right)\left(\mathbf{X}_j^{(2)} - \overline{\mathbf{X}}^{(2)}\right)'\right] \tag{4.88}$$

into unknown parameters $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$ and Σ, respectively, we use the *plug-in discriminant function*

$$\hat{U} \equiv \mathbf{X}'\mathbf{S}^{-1}\left(\overline{\mathbf{X}}^{(1)} - \overline{\mathbf{X}}^{(2)}\right) - \frac{1}{2}\left(\overline{\mathbf{X}}^{(1)} + \overline{\mathbf{X}}^{(2)}\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{X}}^{(1)} - \overline{\mathbf{X}}^{(2)}\right). \tag{4.89}$$

Therefore, we judge $\mathbf{X}$ is a sample from Π_1 if $\hat{U} \geq 0$, and $\mathbf{X}$ is a sample from Π_2 if $\hat{U} < 0$. The exact evaluation of the misclassification probability for $\hat{U}$ is difficult, however consideration of the asymptotic distribution of $\hat{U}$ simplifies the evaluation if n_1 and n_2 are large. In fact, the law of large numbers and

Exercise 2.17 lead us to

$$\overline{\mathbf{X}}^{(1)} \xrightarrow{P} \boldsymbol{\mu}^{(1)}, \quad (n_1 \to \infty), \tag{4.90}$$

$$\overline{\mathbf{X}}^{(2)} \xrightarrow{P} \boldsymbol{\mu}^{(2)}, \quad (n_2 \to \infty), \tag{4.91}$$

$$\mathbf{S} \xrightarrow{P} \Sigma, \quad (n_1, n_2 \to \infty) \tag{4.92}$$

(Exercise 4.8). Using Slutsky's lemma, we have

$$\hat{U} \xrightarrow{d} U, \quad (n_1, n_2 \to \infty), \tag{4.93}$$

where

$$U \sim \begin{cases} N\left(\frac{1}{2}\Delta^2, \Delta^2\right), & (\mathbf{X} \in \Pi_1), \\ N\left(-\frac{1}{2}\Delta^2, \Delta^2\right), & (\mathbf{X} \in \Pi_2). \end{cases} \tag{4.94}$$

Therefore, the misclassification probability based on $\hat{U}$ tends to that based on U, asymptotically.

Up until now, we assumed that the variance matrices Σ of $\mathbf{X}$ under Π_1 and Π_2 are equivalent. We can extend the argument above to the case when the variance matrices are different, as follows. Suppose that the probability density functions of $\mathbf{X}$ described by the categories Π_i $(i = 1, 2)$ are given by

$$f_i(\mathbf{x}) = (2\pi)^{-\frac{m}{2}} |\Sigma_i|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}^{(i)}\right)' \Sigma_i^{-1}\left(\mathbf{x} - \boldsymbol{\mu}^{(i)}\right)\right\}, \tag{4.95}$$

where $\boldsymbol{\mu}^{(i)}$ and Σ_i $(i = 1, 2)$ are assumed to be known. In this case the optimal classification rule $\mathcal{R} = \{\mathcal{R}_1, \mathcal{R}_2\}$ given in Theorem 4.3 becomes

$$\mathcal{R}_1 = \left\{\mathbf{x} \in \mathbf{R}^m : \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \left(\mathbf{x} - \boldsymbol{\mu}^{(1)}\right)' \Sigma_1^{-1}\left(\mathbf{x} - \boldsymbol{\mu}^{(1)}\right) - \left(\mathbf{x} - \boldsymbol{\mu}^{(2)}\right)' \Sigma_2^{-1}\left(\mathbf{x} - \boldsymbol{\mu}^{(2)}\right) \leq 0\right\}, \tag{4.96}$$

$$\mathcal{R}_2 = \left\{\mathbf{x} \in \mathbf{R}^m : \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \left(\mathbf{x} - \boldsymbol{\mu}^{(1)}\right)' \Sigma_1^{-1}\left(\mathbf{x} - \boldsymbol{\mu}^{(1)}\right) - \left(\mathbf{x} - \boldsymbol{\mu}^{(2)}\right)' \Sigma_2^{-1}\left(\mathbf{x} - \boldsymbol{\mu}^{(2)}\right) > 0\right\}. \tag{4.97}$$

However, for the case when $\boldsymbol{\mu}^{(i)}$ and Σ_i are unknown, we need the training samples. Similarly as in the above, suppose that we have the training samples $\mathbf{X}^{(i)}$ from Π_i. Put

$$D\left(\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \Sigma_1, \Sigma_2\right) \equiv \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \left(\mathbf{X} - \boldsymbol{\mu}^{(1)}\right)' \Sigma_1^{-1}\left(\mathbf{X} - \boldsymbol{\mu}^{(1)}\right) - \left(\mathbf{X} - \boldsymbol{\mu}^{(2)}\right)' \Sigma_2^{-1}\left(\mathbf{X} - \boldsymbol{\mu}^{(2)}\right) \tag{4.98}$$

and define

$$\hat{D} \equiv D\left(\overline{\mathbf{X}}^{(1)}, \overline{\mathbf{X}}^{(2)}, \mathbf{S}_1, \mathbf{S}_2\right), \tag{4.99}$$

where $\mathbf{S}_i = (n_i - 1)^{-1} \sum_{j=1}^{n_i} \left(\mathbf{X}_j^{(i)} - \overline{\mathbf{X}}^{(i)} \right) \left(\mathbf{X}_j^{(i)} - \overline{\mathbf{X}}^{(i)} \right)'$, $i = 1, 2$. Therefore, in this case, we can classify $\mathbf{X}$ to Π_1 if $\hat{D} \leq 0$ and $\mathbf{X}$ to Π_2 if $\hat{D} > 0$.

Furthermore, Taniguchi (1994) dealt with the problem of classifying $\mathbf{X}$ into one of two categories

$$\Pi_1 : f\left(\mathbf{x}, \boldsymbol{\theta}^{(1)}\right), \quad \Pi_2 : f\left(\mathbf{x}, \boldsymbol{\theta}^{(2)}\right), \tag{4.100}$$

where the distribution of $\mathbf{X}$ belongs an exponential family which includes the normal distribution, and $f(\mathbf{x}, \boldsymbol{\theta})$ denotes the probability density function of $\mathbf{X}$. The discriminant function is given by

$$\hat{W} \equiv \log \left\{ \frac{f\left(\mathbf{X}, \hat{\boldsymbol{\theta}}^{(1)}\right)}{f\left(\mathbf{X}, \hat{\boldsymbol{\theta}}^{(2)}\right)} \right\}, \tag{4.101}$$

where under Π_i $(i = 1, 2)$, $\hat{\boldsymbol{\theta}}^{(i)}$ is a consistent estimator of $\boldsymbol{\theta}^{(i)}$ based on the training sample of size n_i. Evaluating the expectation of the misclassification probability with respect to the training sample up to the order n_i^{-1}, we see that this is minimized if we take the MLE of $\boldsymbol{\theta}^{(i)}$ as $\hat{\boldsymbol{\theta}}^{(i)}$. This result includes various results for the discriminant analysis of normal distributions as special cases.

So far we considered the discriminant problem of two categories, however we can extend it to general p categories case. Let an m-dimensional random vector $\mathbf{X}$ have the probability density function $f(\mathbf{x})$, $\mathbf{x} \in \mathbf{R}^m$. Suppose that we know $f(\mathbf{x})$ belongs to one of the following p categories

$$\Pi_i : f(\mathbf{x}) = f_i(\mathbf{x}), \quad (i = 1, \ldots, p). \tag{4.102}$$

When we observe $\mathbf{X} = \mathbf{x}$, decomposing $\mathbf{R}^m$ into p exclusive regions $\mathcal{R}_1, \ldots, \mathcal{R}_p$ $(\mathbf{R}^m = \cup_i \mathcal{R}_i)$, we classify $\mathbf{X}$ to Π_i if $\mathbf{x} \in \mathcal{R}_i$. The misclassification probability of this discriminant criterion $\mathcal{R} = \{\mathcal{R}_1, \ldots, \mathcal{R}_p\}$ is

$$M(\mathcal{R}) \equiv \sum_{i=1}^{p} \sum_{j=1, j \neq i}^{p} p(j|i), \tag{4.103}$$

where

$$p(j|i) = \int_{\mathcal{R}_j} f_i(\mathbf{x}) d\mathbf{x}. \tag{4.104}$$

Put

$$h_j(\mathbf{x}) = \sum_{j=1, j \neq i}^{p} f_i(\mathbf{x}), \tag{4.105}$$

then we have

$$M(\mathcal{R}) = \sum_{i=1}^{p} \int_{\mathcal{R}_j} h_j(\mathbf{x}) d\mathbf{x}. \tag{4.106}$$

Let the discriminant criterion $\mathcal{R}^* = \{\mathcal{R}_1^*, \ldots, \mathcal{R}_p^*\}$ be

$$\mathcal{R}_k^* = \left\{ \mathbf{x} \in \mathbf{R}^m : \sum_{i=1, i\neq k}^{p} f_i(\mathbf{x}) \leq \sum_{i=1, i\neq j}^{p} f_i(\mathbf{x}),\ j = 1, \ldots, p\ (j \neq k) \right\}, \tag{4.107}$$

then this implies

$$M(\mathcal{R}) - M(\mathcal{R}^*) = \sum_{j=1}^{p} \int_{\mathcal{R}_j} \left\{ h_j(\mathbf{x}) - \min_{1\leq i\leq p} h_i(\mathbf{x}) \right\} d\mathbf{x} \geq 0. \tag{4.108}$$

Since $\sum_{i=1, i\neq k}^{p} f_i(\mathbf{x}) = \sum_{i=1}^{p} f_i(\mathbf{x}) - f_k(\mathbf{x})$, we can rewrite

$$\mathcal{R}_k^* = \{\mathbf{x} \in \mathbf{R}^m : f_k(\mathbf{x}) \geq f_j(\mathbf{x}),\ j = 1, \ldots, p\ (j \neq k)\}. \tag{4.109}$$

Summarising the above, we obtain the following extension of Theorem 4.3 to that of p categories case.

Theorem 4.4 *For the discriminant problem (4.102), if we define the discriminant criterion $\mathcal{R}^* = \{\mathcal{R}_1^*, \ldots, \mathcal{R}_p^*\}$ as (4.109), then this is the optimal classification rule, which minimizes the misclassification probability $M(\mathcal{R})$.*

Exercises

4.1 Let $Z \sim \chi^2(n)$. Show that Z has the probability density function

$$f(z) = 2^{-\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)^{-1} z^{\frac{n}{2}-1} e^{-\frac{z}{2}}, \quad z > 0. \tag{4.110}$$

4.2 In (4.8) (i) Derive the joint probability density function of (Y, Z). (ii) Consider the transformation $(Y, Z) \to (T, Z)$ and derive the joint probability density function of (T, Z) (using the *change of variable formula* (Theorem A.6)). (iii) Show that the probability density function of T (the probability density function of $t(n)$-distribution) is given by

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < t < \infty. \tag{4.111}$$

4.3 The following data

0.53, 1.51, 1.55, 2.57, 1.16, 0.66, 1.61, 0.74, 0.59, 2.01

are from a sample of size 10, which are supposed to be independent and identically distributed as $N(\mu, \sigma^2)$. Give a confidence interval of μ with confidence coefficient 0.95.

4.4 When we tossed a coin twenty times, the following result was obtained, where 0 and 1 represent tails and heads, respectively.

1 0 1 1 1 0 1 0 1 1 1 1 0 0 1 1 1 1 1 1

Then using the test in Example 4.6, test whether the coin is fair or unfair with a significance level of 0.05.

4.5 Show that the density function of F-distribution with (m, n) degrees of freedom is given by

$$f(x) = \frac{\Gamma(\frac{m+n}{2}) m^{m/2} n^{n/2}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \frac{x^{m/2-1}}{(mx+n)^{(m+n)/2}}, \quad (x > 0).$$

4.6 Verify that the optimal $\mathcal{R} = \{\mathcal{R}_1, \mathcal{R}_2\}$ in Example 4.13 is given by (4.77) and (4.78).

4.7 Prove that (4.83) holds.

4.8 Show that $\mathbf{S} \xrightarrow{P} \Sigma$, as $n_1, n_2 \to \infty$ in (4.92).

4.9 Verify (4.108).

CHAPTER 5

Stochastic Processes

So far we have assumed that random variables $X_1, X_2, \ldots, X_n$ forming a sample are mutually independent and identically distributed. For example, in the toss of a dice, if X_t is a random variable describing the tth outcome, we may suppose that the outcome X_t does not affect any other outcomes X_s, $s \neq t$. Hence the setting of independence seems natural. If $X_1, \ldots, X_n$ show the height of n students randomly sampled in a school class, then we may also accept the setting of independence naturally.

However, if X_t is the value of a stock price at time t, it should be natural to assume that the present value X_{t_0}, the past values X_s, $s < t_0$, and the future values X_s, $s > t_0$, are dependent, i.e., interactive. Also, if X_t is the value of an hourly record of temperature at a given place at time t, or the value of an electrophysiological signal, then their past, present and future values should be interactive. The stochastic process was introduced to describe such a series of observations X_t, $t = 1, \ldots, n$, which randomly vary together with time t and are mutually dependent.

This chapter explains elements of stochastic processes, e.g., stationarity, spectral structure, ergodicity, mixing property, martingale, etc. Because the statistical analysis for stochastic processes largely relies on the asymptotic theory with respect to the length of observations, we present some useful limit theorems and central limit theorems.

5.1 Elements of Stochastic Processes

Stochastic processes were born as a mathematical model describing random quantities which vary together with time. For each time $t \in \mathbf{Z}$, suppose that there exists a random variable X_t defined on a probability space $(\Omega, \mathcal{A}, P)$, then the family of random variables $\{X_t : t \in \mathbf{Z}\}$ is called a *stochastic process*. Here we assumed that time t belongs to $\mathbf{Z}$, which is called an *index set*. However, we may take continuous sets, for example, $[0, \infty)$, $\mathbf{R}$ etc. as the index set. In what follows, because we often deal with stochastic processes with index set $\mathbf{Z}$, the index set is assumed to be $\mathbf{Z}$ if we do not mention it.

From the definition of stochastic process, $\{X_t\}$ may be a family of any random variables. But, if we want to do mathematical or statistical analysis, a sort

of regularity or invariance for $\{X_t\}$ is needed. The most fundamental one is stationarity.

Definition 5.1 *A stochastic process* $\{X_t : t \in \mathbf{Z}\}$ *is called* strictly stationary *if, for all* $n \in \mathbf{N}$, $t_1, \ldots, t_n$, $h \in \mathbf{Z}$, *the distributions of* $\mathbf{X}_0 \equiv (X_{t_1}, \ldots, X_{t_n})'$ *and of* $\mathbf{X}_h \equiv (X_{t_1+h}, \ldots, X_{t_n+h})'$ *are the same, i.e.,* $P\{\mathbf{X}_0^{-1}(\cdot)\} = P\{\mathbf{X}_h^{-1}(\cdot)\}$.

The function on $\mathbf{R}^n$

$$F_{t_1,\ldots,t_n}(x_1, \ldots, x_n) \equiv P\{X_{t_1} \le x_1, \ldots, X_{t_n} \le x_n\} \tag{5.1}$$

is called the *joint distribution function* of $X_{t_1}, \ldots, X_{t_n}$. In terms of the joint distribution function, $\{X_t\}$ is strictly stationary if, for all $n \in \mathbf{N}$, $t_1, \ldots, t_n$, $h \in \mathbf{Z}$, it holds that

$$F_{t_1,\ldots,t_n}(x_1, \ldots, x_n) = F_{t_1+h,\ldots,t_n+h}(x_1, \ldots, x_n) \tag{5.2}$$

for all $(x_1, \ldots, x_n)' \in \mathbf{R}^n$.

If $\{X_t\}$ is strictly stationary, the distribution function of X_t is the same for all $t \in \mathbf{Z}$, and the joint distribution function of $X_{t_1}, \ldots, X_{t_n}$ depends only on the time differences $t_2 - t_1, \ldots, t_n - t_{n-1}$ (Exercise 5.1). Next, let us see some examples of a strictly stationary process. If X_t, $t \in \mathbf{Z}$, are i.i.d. random variables, then $\{X_t : t \in \mathbf{Z}\}$ satisfies (5.2). Hence, sequences of i.i.d. random variables become the simplest examples of strictly stationary process.

Example 5.1 *Let* $\{X_t : t \in \mathbf{Z}\}$ *be a strictly stationary process. For a measurable function* $\phi : \mathbf{R}^q \to \mathbf{R}$, *define* Y_t *by*

$$Y_t = \phi(X_t, X_{t-1}, \ldots, X_{t-q}). \tag{5.3}$$

Then, for all $n \in \mathbf{N}$, $t_1, \ldots, t_n$, $h \in \mathbf{Z}$, *letting* $\mathbf{Y}_h \equiv (Y_{t_1+h}, \ldots, Y_{t_n+h})'$, *from the strict stationarity of* $\{X_t\}$, *we can show that*

$$P\{\mathbf{Y}_h^{-1}(B)\} = P\{\mathbf{Y}_0^{-1}(B)\}, \quad (^\forall B \in \mathcal{B}^n)$$

which implies that $\{Y_t\}$ *is a strictly stationary process. As a special case, let* $\{u_t\}$ *be a sequence of i.i.d. random variables, and define* X_t *by*

$$X_t = \alpha_0 u_t + \alpha_1 u_{t-1} + \cdots + \alpha_q u_{t-q}, \tag{5.4}$$

where $\alpha_0, \ldots, \alpha_q$ *are real constants. Then the process* $\{X_t : t \in \mathbf{Z}\}$ *is a strictly stationary process. The process* (5.4) *is called a* moving average process of order q *and is denoted by MA(q).* □

Although the concept of strict stationarity is mathematically fundamental and natural, the assumption (5.2) is too severe in view of statistical analysis. Since we often base our statistical discussion substantially on first and second order moment properties, in such situations we use another stationarity. For a stochastic process $\{X_t : t \in \mathbf{Z}\}$, define

$$R(t, s) = Cov(X_t, X_s) \equiv E[\{X_t - E(X_t)\}\{\overline{X_s - E(X_s)}\}] \tag{5.5}$$

where $\{\bar{\cdot}\}$ denotes complex conjugate of $\{\cdot\}$. However, we shall assume except where explicitly stated otherwise, that X_t is real.

Definition 5.2 *A stochastic process* $\{X_t : t \in \mathbf{Z}\}$ *is said to be* weakly (or second-order) stationary *if*

(i) $E\{|X_t|^2\} < \infty$ *for all* $t \in \mathbf{Z}$,
(ii) $E(X_t) = c$ *for all* $t \in \mathbf{Z}$, *where* c *is constant,*
(iii) $R(t, s) = R(0, s - t)$ *for all* $s, t \in \mathbf{Z}$.

If $\{X_t : t \in \mathbf{Z}\}$ *is weakly stationary, we redefine*

$$R(s - t) \equiv R(0, s - t) \quad \text{for all } s, t \in \mathbf{Z}.$$

The function $R(h)$ *is called the* autocovariance function *of* $\{X_t\}$ *at lag* h $(\in \mathbf{Z})$. *We note that a strictly stationary process with finite second-order moments is weakly stationary.*

Definition 5.3 *A stochastic process* $\{X_t : t \in \mathbf{Z}\}$ *is said to be a* Gaussian process *if for each* $t_1, \ldots, t_n \in \mathbf{Z}$, $n \in \mathbf{Z}$, *the joint distribution of* $X_{t_1}, \ldots, X_{t_n}$ *is multivariate normal.*

For a Gaussian process the mean and the autocovariance function completely determine all finite-dimensional distributions, hence weak stationarity is equivalent to strict stationarity. In what follows we will mainly deal with weakly stationary processes, henceforth, we call them *stationary processes* for simplicity.

Example 5.2 *Let a random variable* $U \sim U[-\pi, \pi]$ *(uniform distribution on* $[-\pi, \pi]$*). For real constants* A *and* λ, *define*

$$X_t = A\cos(\lambda t + U), \quad (t \in \mathbf{Z}).$$

Then it is shown that

$$\begin{aligned} &E(X_t) = 0, && (^{\forall}t \in \mathbf{Z}), \\ &Cov(X_t, X_s) = \frac{1}{2}A^2\cos\{\lambda(t - s)\}, && (^{\forall}t, s \in \mathbf{Z}), \end{aligned} \tag{5.6}$$

(Exercise 5.2). Therefore the process $\{X_t : t \in \mathbf{Z}\}$ *becomes a stationary process.* □

Example 5.3 *Let* $\{X_t : t \in \mathbf{Z}\}$ *be the MA(q) process defined in* (5.4). *If we assume that* $\{u_t\} \sim$ *i.i.d.* $(0, \sigma^2)$, *then it is seen that*

$$E(X_t) = 0,$$

$$Cov(X_t, X_s) = \begin{cases} \sigma^2 \displaystyle\sum_{j=0}^{q-|t-s|} \alpha_j \alpha_{j+|t-s|}, & \text{if } 0 \le |t - s| \le q, \\ 0, & \text{if } |t - s| > q. \end{cases} \tag{5.7}$$

(Exercise 5.2). Hence this MA(q) process is stationary with autocovariance function (5.7). □

Before this we provided examples of stationary processes. In contrast with this, we mention processes which are not stationary, i.e., *nonstationary*, below.

Example 5.4 *Assume that* $\{u_t\} \sim$ *i.i.d.* $(0, \sigma^2)$*, and define the following two stochastic processes:*

$$X_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p + u_t, \qquad (t \in \mathbf{Z}) \tag{5.8}$$

$$Y_t = \sum_{j=1}^{t} u_j, \qquad (t \in \mathbf{N}) \tag{5.9}$$

Here $\beta_0, \ldots, \beta_p$ *are real constants. It is easily seen that*

$$E(X_t) = \beta_0 + \beta_1 t + \cdots + \beta_p t^p, \tag{5.10}$$

$$Cov(X_t, X_s) = \begin{cases} \sigma^2, & (t = s), \\ 0, & (t \neq s). \end{cases} \tag{5.11}$$

from which $\{X_t\}$ *does not satisfy* (ii) *of Definition 5.2, hence,* $\{X_t\}$ *is a nonstationary process. Figure 5.1 plots* $X_1, \ldots, X_{100}$ *generated by* $X_t = 1 + 0.2t + u_t$*, what* $\{u_t\} \sim$ *i.i.d.* $N(0, 1)$*. We can observe that the data fluctuate around the trend* $1 + 0.2t$*.*

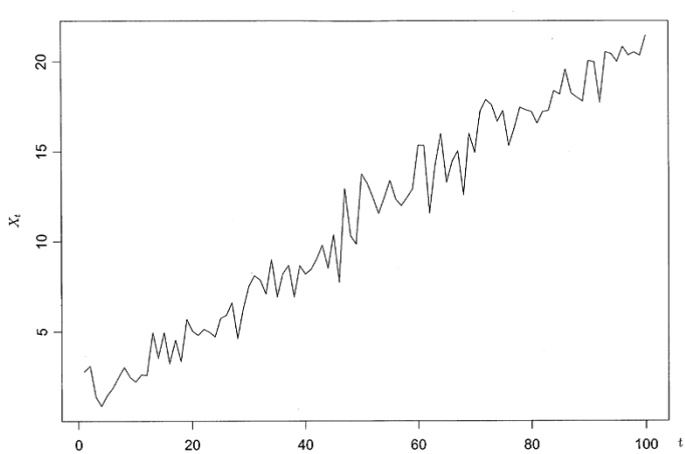


Figure 5.1 *Graph of* $X_t = 1 + 0.2t + u_t$*.*

Next, returning to $\{Y_t\}$ *defined by* (5.9) *we obtain*

$$\begin{aligned} E(Y_t) &= 0, \\ Cov(Y_t, Y_{t+h}) &= Cov\{\sum_{j=1}^{t} u_j, \sum_{k=1}^{t+h} u_k\}, \quad (h \in \mathbf{N}) \\ &= Cov\{\sum_{j=1}^{t} u_j, \sum_{k=1}^{t} u_k\}, \\ &= \sigma^2 t \end{aligned} \tag{5.12}$$

which implies that $\{Y_t\}$ *does not satisfy* (iii) *of Definition 5.2, hence, it is a nonstationary process, and is called the* random walk process. *Figure 5.2 plots* $Y_1, \ldots, Y_{100}$ *generated by* (5.9) *when* $\{u_t\} \sim$ *i.i.d.* $N(0,1)$. *The graph shows an aggregation of fluctuation of* u_t, *and the feature of nonstationarity of* $\{Y_t\}$ *is different from that of* $\{X_t\}$. □

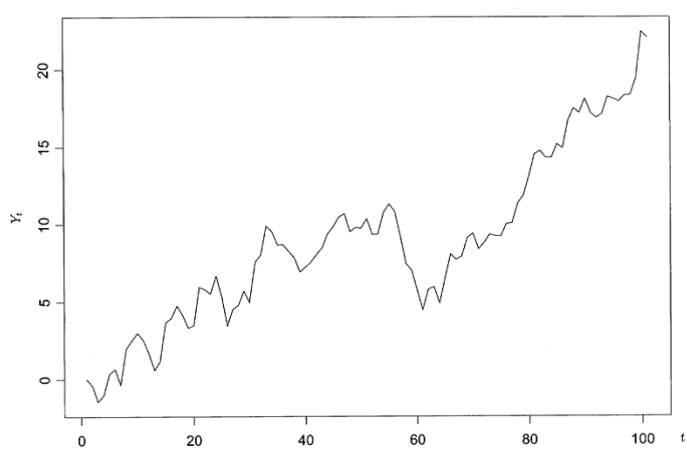


Figure 5.2 *Graph of* Y_t.

5.2 Spectral Analysis

Spectral structure is a very important and fundamental characteristic of stationary processes. To understand the concept of spectral structure, let us see the following stochastic process. Let

$$X_t = \sum_{j=1}^{n} A(\lambda_j) e^{-it\lambda_j}, \quad (i = \sqrt{-1}) \tag{5.13}$$

where $\lambda_1, \ldots, \lambda_n$ are real constants satisfying $-\pi < \lambda_1 < \lambda_2 < \cdots < \lambda_n = \pi$, and $A(\lambda_1), \ldots, A(\lambda_n)$ are complex-valued random variables with

$$E\{A(\lambda_j)\} = 0,$$
$$E\{A(\lambda_j)\overline{A(\lambda_k)}\} = \begin{cases} \sigma_j^2, & (j = k), \\ 0, & (j \neq k). \end{cases}$$

From the definition of $\{X_t\}$ it follows that

$$E(X_t) = \sum_{j=1}^{n} E\{A(\lambda_j)\}\, e^{-\boldsymbol{i}t\lambda_j} = 0,$$
$$E(X_t\overline{X_{t+h}}) = \sum_{j=1}^{n}\sum_{k=1}^{n} E\{A(\lambda_j)\overline{A(\lambda_k)}\}e^{-it\lambda_j + i(t+h)\lambda_k} \tag{5.14}$$
$$= \sum_{j=1}^{n} \sigma_j^2 e^{ih\lambda_j}.$$

Then $\{X_t\}$ is a stationary process with mean 0 and autocovariance function

$$R(h) = \sum_{j=1}^{n} \sigma_j^2 e^{ih\lambda_j}. \tag{5.15}$$

Define

$$F(\lambda) \equiv \sum_{j:\lambda_j \leq \lambda} \sigma_j^2, \tag{5.16}$$

which is a step function with jump σ_j^2 at $\lambda = \lambda_j$. Then the relation (5.15) is representable as the Lebesgue-Stieltjes integral

$$R(h) = \int_{-\pi}^{\pi} e^{ih\lambda} dF(\lambda). \tag{5.17}$$

Though the representation (5.17) is derived for the process (5.13), actually, it is possible to get the representation (5.17) for general stationary processes.

Theorem 5.1 *If $R(\cdot)$ is the autocovariance function of a stationary process $\{X_t : t \in \mathbf{Z}\}$, then*

$$R(h) = \int_{-\pi}^{\pi} e^{ih\lambda} dF(\lambda), \quad (h \in \mathbf{Z}), \tag{5.18}$$

where $F(\lambda)$ is a nondecreasing function. The function $F(\lambda)$ is uniquely defined if we require in addition that (i) $F(-\pi) = 0$ and (ii) $F(\lambda)$ is right continuous.

PROOF First, note that $R(t)$ is a non-negative definite function, that is,

$$\sum_j \sum_k \bar{\beta}_j \beta_k R(t_j - t_k) \geq 0$$

for any set of complex numbers β_j and any n integers t_j, and then form

$$F^{(n)}(\lambda) = \frac{1}{2\pi} \sum_{h=-n}^{n} R(h) \frac{e^{ih\pi} - e^{-ih\pi}}{ih} \left(1 - \frac{|h|}{n}\right),$$

taking $R(0)(\pi + \lambda)$ as the term for $h = 0$. Then it is seen that $F^{(n)}(\lambda)$ is nondecreasing, $F^{(n)}(-\pi) = 0$, $F^{(n)}(\pi) = R(0) < \infty$ for all n and that

$$\int_{-\pi}^{\pi} e^{ih\lambda} dF^{(n)}(\lambda) = \begin{cases} R(t)\left(1 - \frac{|t|}{n}\right), & |t| \le n, \\ 0, & |t| > n. \end{cases} \tag{5.19}$$

We can apply Helly's theorem (Theorem A.10) and Theorem A.9 to deduce that there is a distribution function $F(\lambda)$ and a subsequence $\{F^{(n_k)}\}$ of $\{F^{(n)}\}$ such that

$$\int_{-\pi}^{\pi} e^{it\lambda} dF^{(n_k)}(\lambda) \to \int_{-\pi}^{\pi} e^{it\lambda} dF(\lambda) \quad \text{as } k \to \infty.$$

Hence the required representation

$$R(t) = \int_{-\pi}^{\pi} e^{it\lambda}\, dF(\lambda)$$

follows from (5.19). □

The function $F(\lambda)$ is called the *spectral distribution function* of $\{X_t\}$. If $F(\lambda)$ is absolutely continuous with respect to Lebesgue measure on $[-\pi, \pi]$ so that

$$F(\lambda) = \int_{-\pi}^{\lambda} f(\mu) d\mu, \quad (dF(\lambda) = f(\lambda) d\lambda),$$

then $f(\lambda)$ is called the *spectral density function* of $\{X_t\}$. If $F(\lambda)$ has the spectral density function $f(\lambda)$, the representation (5.18) becomes

$$R(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda) d\lambda. \tag{5.20}$$

Henceforth λ is often called the *frequency*. Returning to the process (5.13), i.e., $X_t = \sum_{j=1}^{n} A(\lambda_j) e^{-it\lambda_j}$, from (5.16) we may understand that the spectral distribution shows a degree of strength (variance) of "frequency component" $A(\lambda_j)\, e^{-it\lambda_j}$ contained in X_t.

Now, in what follows, we deal with a general stationary process $\{X_t\}$ with mean 0 and autocovariance function $R(\cdot)$. Initially we make the following.

Assumption 5.1

$$\sum_{-\infty}^{\infty} |R(j)| < \infty. \tag{5.21}$$

This assumption seems natural because it implies that the correlation between X_t and X_{t+j} tends to zero as $|j| \to \infty$. Let

$$f(\lambda) \equiv \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} R(j) e^{ij\lambda}. \tag{5.22}$$

If we substitute $f(\lambda)$ into the right-hand side of (5.20), then it satisfies the relation (5.20). Therefore, under Assumption 5.1, $f(\lambda)$ is the spectral density function of $\{X_t\}$. From this we can understand that the spectral density function is nothing but the Fourier transformed autocovariance function.

For an observed stretch $X_1, \ldots, X_n$ of $\{X_t\}$, let

$$\mathcal{F}_n(\lambda) = \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} X_t e^{it\lambda}, \quad (\lambda \in [-\pi, \pi]), \tag{5.23}$$

which is called the *finite Fourier transform*. The quantity $I_n(\lambda) = |\mathcal{F}_n(\lambda)|^2$ is called the *periodogram*. The next theorem describes their fundamental properties.

Theorem 5.2 *Suppose that Assumption 5.1 holds. Then,*

(i)

$$\lim_{n\to\infty} E\{I_n(\lambda)\} = f(\lambda), \quad (\lambda \in [-\pi, \pi]) \tag{5.24}$$

(ii) for $\lambda_k = 2\pi k/n$,

$$\lim_{n\to\infty} E\{\mathcal{F}_n(\lambda_k)\overline{\mathcal{F}_n(\lambda_r)}\} = 0, \quad (k \neq r,\ k, r = 1, \ldots, n). \tag{5.25}$$

PROOF

(i) From the definition of $I_n(\lambda)$,

$$\begin{aligned} E\{I_n(\lambda)\} &= \frac{1}{2\pi n} \sum_{t=1}^{n} \sum_{s=1}^{n} E(X_t X_s) e^{-i(s-t)\lambda} \\ &= \frac{1}{2\pi n} \sum_{t=1}^{n} \sum_{s=1}^{n} R(s-t) e^{-i(s-t)\lambda} \quad \text{(by stationarity)} \\ &= \frac{1}{2\pi n} \sum_{l=-n+1}^{n-1} (n - |l|) R(l) e^{-il\lambda}, \quad (l = s - t) \\ &= \frac{1}{2\pi} \sum_{l=-n+1}^{n-1} R(l) e^{-il\lambda} - \frac{1}{2\pi n} \sum_{l=-n+1}^{n-1} |l| R(l) e^{-il\lambda} \\ &= (A) + (B), \quad \text{(say)}. \end{aligned}$$

It follows from Assumption 5.1 and (5.22) that $(A) \to f(\lambda)$ $(n \to \infty)$.

We next evaluate (B). For each $\varepsilon > 0$, from Assumption 5.1 we can choose $M_\varepsilon \in \mathbf{N}$ such that $\sum_{|l|>M_\varepsilon}|R(l)| < \varepsilon$. For this M_ε, we have

$$n^{-1}\sum_{|l|\le M_\varepsilon}|l||R(l)| \le \frac{M_\varepsilon}{n}\sum_{|l|\le M_\varepsilon}|R(l)|,$$

hence,

$$|2\pi(B)| \le \frac{1}{n}\sum_{l=-n+1}^{n-1}|l||R(l)| < \frac{M_\varepsilon}{n}\sum_{l=-\infty}^{\infty}|R(l)| + \varepsilon.$$

Since we can make M_ε/n arbitrarily near to zero if n becomes sufficiently large, we have proved $(B) \to 0$ $(n \to \infty)$, which completes the proof of (5.24).

(ii) First, let us note the following fundamental formula

$$\frac{1}{n}\sum_{s=1}^{n} e^{is\lambda_k} = \frac{e^{i\lambda_k}(1-e^{in\lambda_k})}{n(1-e^{i\lambda_k})} = \begin{cases} 1, & (k = 0, \pm n, \pm 2n, \dots), \\ 0, & (\text{otherwise}). \end{cases} \tag{5.26}$$

From the definition of $\mathcal{F}_n(\lambda)$ we obtain

$$\begin{aligned} E\{\mathcal{F}_n(\lambda_k)\overline{\mathcal{F}_n(\lambda_r)}\} &= \frac{1}{2\pi n}\sum_{t=1}^{n}\sum_{s=1}^{n} E(X_tX_s)e^{-is\lambda_r+it\lambda_k} \\ &= \frac{1}{2\pi n}\sum_{t=1}^{n}\sum_{s=1}^{n} R(s-t)e^{-\boldsymbol{i}(s-t)\lambda_r+it(\lambda_k-\lambda_r)} \\ &= \frac{1}{2\pi}\sum_{l=-n+1}^{n-1} R(l)e^{-il\lambda_r}\frac{1}{n}\sum_{\substack{1\le t\le n\\ 1\le t+l\le n}} e^{it(\lambda_k-\lambda_r)}. \end{aligned} \tag{5.27}$$

Noting

$$\left|\sum_{\substack{1\le t\le n\\ 1\le t+l\le n}} e^{it(\lambda_k-\lambda_r)} - \sum_{t=1}^{n} e^{it(\lambda_k-\lambda_r)}\right| \le |l|,$$

and using the same evaluation method as in (B) of (i), we can see that (5.27) is equal to

$$\frac{1}{2\pi}\sum_{l=-n+1}^{n-1} R(l)e^{-\boldsymbol{i}l\lambda_k}\frac{1}{n}\sum_{t=1}^{n} e^{it(\lambda_k-\lambda_r)} + o(1).$$

Hence, from (5.26) it is seen that if $k \ne r$,

$$E\{\mathcal{F}_n(\lambda_k)\overline{\mathcal{F}_n(\lambda_r)}\} \to 0, \quad (n \to \infty).$$

In Theorem 5.1 we saw that the autocovariance function of stationary processes can be expressed by the spectral distribution function. In what follows, we discuss the spectral representation for stationary processes. For this, recall the convergence in pth mean (see Section 2.4). In the case of $p = 2$, we call it the convergence in mean square. If a sequence $\{Y_n\}$ of random variables converges in mean square to a random variable Y, we write

$$l.i.m._{n\to\infty} Y_n = Y$$

For the stationary process $\{X_t\}$ satisfying Assumption 5.1, we have

$$\begin{aligned}\sum_{s=1}^{n} e^{-it(2\pi s/n)}\sqrt{\frac{2\pi}{n}}\mathcal{F}_n\left(\frac{2\pi s}{n}\right) &= \sum_{s=1}^{n} e^{-it(2\pi s/n)}\frac{1}{n}\sum_{r=1}^{n} X_r e^{ir(2\pi s/n)} \\ &= \sum_{r=1}^{n} X_r \frac{1}{n}\sum_{s=1}^{n} e^{is\{2\pi(r-t)/n\}} \\ &= X_t \qquad \text{(by (5.26))}.\end{aligned} \tag{5.28}$$

If we set $\Delta Z_n(2\pi s/n) \equiv \sqrt{2\pi/n}\mathcal{F}_n(2\pi s/n)$, the relation (5.28) is written as

$$X_t = \sum_{s=1}^{n} e^{-it(2\pi s/n)}\Delta Z_n\left(\frac{2\pi s}{n}\right) \tag{5.29}$$

From (i) of Theorem 5.2 it follows that

$$E\left[\frac{|\Delta Z_n(2\pi s/n)|^2}{\Delta\lambda}\right] - f\left(\frac{2\pi s}{n}\right) \to 0, \quad (n\to\infty) \tag{5.30}$$

where $\Delta\lambda = 2\pi/n$. We write (5.30) as

$$E\left[\left|\Delta Z_n\left(\frac{2\pi s}{n}\right)\right|^2\right] \sim f\left(\frac{2\pi s}{n}\right)\Delta\lambda \tag{5.31}$$

Also, from (ii) of Theorem 5.2 we can see that

$$E\left[\Delta Z_n\left(\frac{2\pi s}{n}\right)\overline{\Delta Z_n\left(\frac{2\pi r}{n}\right)}\right] \to 0, \quad (s\neq r). \tag{5.32}$$

Write $dZ(\lambda) \equiv l.i.m._{n\to\infty}\Delta Z_n(\lambda)$, i.e., $dZ(\lambda)$ is the limit in the sense of $l.i.m.$ of the Fourier transform of $\{X_t\}$ at frequency λ. Taking $l.i.m.$ in the equation (5.29) we obtain the spectral representation

$$X_t = \int_{-\pi}^{\pi} e^{-it\lambda} dZ(\lambda). \tag{5.33}$$

From (5.31) and (5.32) we can intuitively understand that $Z(\lambda)$ satisfies $E\{|dZ(\lambda)|^2\} = f(\lambda)d\lambda$ and $E\{dZ(\lambda)\overline{dZ(\mu)}\} = 0$, $\lambda \neq \mu \in [-\pi,\pi]$, respectively. Above we provided the representation (5.33) substantially and intuitively. The next theorem claims that the representation (5.33) is possible for

general stationary processes. Since the substantial and intuitive derivation of (5.33) seems important rather than mathematical in view of statistical analysis, we omit the proof. For mathematically rigorous proof, see, e.g., Brockwell and Davis (1991, p.145).

Theorem 5.3 *Suppose that* $\{X_t : t \in \mathbf{Z}\}$ *is a stationary process with mean* 0 *and spectral distribution function* $F(\lambda)$. *Then,* X_t *has the spectral representation*

$$X_t = \int_{-\pi}^{\pi} e^{-it\lambda} dZ(\lambda), \tag{5.34}$$

where $Z(\lambda)$ *satisfies*

(i) $E\{Z(\lambda)\} = 0$,
(ii) $E\{|dZ(\lambda)|^2\} = dF(\lambda)$, $\lambda \in [-\pi, \pi]$,
(iii) $E\{dZ(\lambda)\overline{dZ(\mu)}\} = 0$, $\lambda \neq \mu \in [-\pi, \pi]$.

If a sequence $\{u_t : t \in \mathbf{Z}\}$ of random variables satisfies

$$E(u_t) = 0,$$
$$R_u(s) \equiv E(u_t u_{t+s}) = \begin{cases} \sigma^2, & (s = 0), \\ 0, & (s \neq 0), \end{cases}$$

then it is called the *uncorrelated process*. Of course $\{u_t\}$ is a stationary process. Recalling (5.22), we can see that it has the spectral density $f_u(\lambda) = \sigma^2/2\pi$.

For a sequence $\{a_j : j = 0, 1, 2, \dots\}$ of real numbers satisfying $\sum_{j=0}^{\infty} a_j^2 < \infty$, if X_t can be expressed as

$$X_t = \sum_{j=0}^{\infty} a_j u_{t-j}, \tag{5.35}$$

then $\{X_t : t \in \mathbf{Z}\}$ is called a general linear process. Here the right-hand side of (5.35) is defined in the sense of $l.i.m._{n\to\infty}$. Further, if $\{a_j\}$ satisfies a stronger condition $\sum_{j=0}^{\infty} |a_j| < \infty$, then $\{X_t\}$ is called a *linear process*.

Let $L_2 \equiv \{Y : E\{|Y|^2\} < \infty\}$. For Y_n, $W_n \in L_2$, define the inner product by $< Y_n, W_n > \equiv E(Y_n \overline{W_n})$, and write $Y = l.i.m._{n\to\infty} Y_n$ and $W = l.i.m._{n\to\infty} W_n$. Then the inner product has a continuity in the sense that $< Y, W > = \lim_{n\to\infty} < Y_n, W_n >$, (see Exercise 5.3). From this, for the general linear process (5.35) we can show that

(1) $E(X_t) = 0$,
(2) $R_X(s) \equiv E(X_t X_{t+s}) = (\sum_{j=0}^{\infty} a_j a_{j+s})\sigma^2$,

which implies that $\{X_t\}$ is a stationary process with mean 0 and autocovariance function $R_X(s)$. Let the spectral representation of $\{u_t\}$ be

$$u_t = \int_{-\pi}^{\pi} e^{-it\lambda} dZ_u(\lambda), \tag{5.36}$$

where $E\{|dZ_u(\lambda)|^2\} = (\sigma^2/2\pi)d\lambda$. Let the spectral distribution function of $\{X_t\}$ in (5.35) be $F_X(\lambda)$, and denote the spectral representation by

$$X_t = \int_{-\pi}^{\pi} e^{-it\lambda} dZ_X(\lambda), \tag{5.37}$$

where $E\{|dZ_X(\lambda)|^2\} = dF_X(\lambda)$. Then the process (5.35) is expressed as

$$\int_{-\pi}^{\pi} e^{-it\lambda} dZ_X(\lambda) = \int_{-\pi}^{\pi} e^{-it\lambda} \left\{ \sum_{j=0}^{\infty} a_j e^{-ij\lambda} \right\} dZ_u(\lambda),$$

hence,

$$dZ_X(\lambda) = \left\{ \sum_{j=0}^{\infty} a_j e^{ij\lambda} \right\} dZ_u(\lambda). \tag{5.38}$$

From (5.38) it follows that

$$\begin{aligned} E\left\{|dZ_X(\lambda)|^2\right\} &= \left| \sum_{j=0}^{\infty} a_j e^{ij\lambda} \right|^2 E\left\{|dZ_u(\lambda)|^2\right\} \\ &= \left| \sum_{j=0}^{\infty} a_j e^{ij\lambda} \right|^2 \left(\frac{\sigma^2}{2\pi}\right) d\lambda. \end{aligned}$$

Therefore the general linear process (5.35) has the spectral density function

$$f_X(\lambda) = \frac{\sigma^2}{2\pi} \left| \sum_{j=0}^{\infty} a_j e^{ij\lambda} \right|^2. \tag{5.39}$$

Up until now we have dealt with scalar-valued stochastic processes. However, if we think of the applications, extension to the case of a vector-valued stochastic process is needed. For vector-valued stochastic processes, it is possible to develop the discussion on stationarity, spectral structure, general linear process, etc., parallel to the scalar case.

For a matrix A, define the norm by

$$||A|| = \text{ the square root of the greatest eigenvalue of } A^*A.$$

Suppose that a family $\{A(j) : j = 0, 1, 2, \dots\}$ of $m \times m$ matrices satisfies

$$\sum_{j=}^{\infty} ||A(j)||^2 < \infty. \tag{5.40}$$

Let an m-vector stochastic process $\{\mathbf{X}_t = (X_{1t}, \dots, X_{mt})'\}$ be defined by

$$\mathbf{X}_t = \sum_{j=0}^{\infty} A(j)\mathbf{U}_{t-j}, \tag{5.41}$$

where $\{\mathbf{U}_t\}$ is an m-vector uncorrelated process satisfying

(i) $E\{\mathbf{U}_t\} = \mathbf{0}$,

(ii) $E\{\mathbf{U}_t\mathbf{U}_r'\} = \begin{cases} V, & (t = r) \\ \mathbf{0}, & (t \neq r), \end{cases}$

and each component of the right-hand side of (5.41) is defined in the sense of *l.i.m.*-limit. Then the process $\{\mathbf{X}_t : t \in \mathbf{Z}\}$ is called an *m-vector general linear process*. Similarly as in the case of scalar processes, we can show that $\{\mathbf{X}_t\}$ is a stationary process, and the spectral density becomes a matrix, called the *spectral density matrix* given by

$$\mathbf{f}(\lambda) = \frac{1}{2\pi}\left\{\sum_{j=0}^{\infty} A(j)e^{ij\lambda}\right\} V \left\{\sum_{j=0}^{\infty} A(j)e^{ij\lambda}\right\}^{*}, \tag{5.42}$$

(see Exercise 5.4).

5.3 Ergodicity, Mixing and Martingale

Consider a physical process governed by a strictly stationary process. Suppose that we need to compute some statistical averages of the process and instead what is available is only a long-term observation of a single realization. In such situations it is natural to ask whether it is possible to determine the statistical average from an appropriate time average corresponding to a single realization. If the statistical (or ensemble) average of the process equals the time average, the process will be called *ergodic*.

Let $\{X_t : t \in \mathbf{Z}\}$ be a strictly stationary process defined on a probability space $(\Omega, \mathcal{A}, P)$. The σ-algebra $\mathcal{A}$ is generated by the family of all cylinder sets

$$\{\omega \in \Omega; (X_{t_1}, \ldots, X_{t_k}) \in B\},$$

where $B \in \mathcal{B}^k$. For these cylinder sets we define the shift operator $A \to TA$ where, for a set A of the form

$$A = \{\omega \in \Omega : (X_{t_1}, \ldots, X_{t_k}) \in C\}, \quad C \in \mathcal{B}^k,$$

we have

$$TA = \{\omega \in \Omega : (X_{t_1+1}, \ldots, X_{t_k+1}) \in C\}.$$

This definition extends to all sets in $\mathcal{A}$. Since $\{X_t\}$ is strictly stationary it is easily shown that A and $T^{-1}A$ always have the same probability content. Then we say that T is *measure preserving*.

Definition 5.4 *(i) Given a measure preserving transformation T, a measurable event A is said to be* invariant *if $T^{-1}A = A$. Denote the collection of invariant sets by $\mathcal{A}_I$.*

(ii) The process $\{X_t : t \in \mathbf{Z}\}$ *is said to be* ergodic *if for all* $A \in \mathcal{A}_I$*, either* $P(A) = 0$ *or* $P(A) = 1$.

A condition that implies ergodicity and whose meaning is somewhat more apparent is the condition

$$\lim_{n\to\infty} P(A \cap T^{-n}B) = P(A)P(B), \quad A, B \in \mathcal{A}. \tag{5.43}$$

If this condition holds then $\{X_t : t \in \mathbf{Z}\}$ is said to be *mixing*. That the mixing property implies ergodicity can be seen by taking $A = B \in \mathcal{A}_I$ in the relation (5.43). Then we obtain $P(A)^2 = P(A)$ so that $P(A) = 0$ or 1. A stronger condition than (5.43) is the strong mixing condition which we now define. Let there exist a positive function g satisfying $g(n) \to 0$ as $n \to \infty$ so that

$$|P(A \cap B) - P(A)P(B)| < g(r - q), \quad A \in \mathcal{A}^q_{-\infty},\ B \in \mathcal{A}^\infty_r, \tag{5.44}$$

where $\mathcal{A}^q_{-\infty} = \sigma\{X_q, X_{q-1}, \dots\}$ and $\mathcal{A}^\infty_r = \sigma\{X_r, X_{r+1}, \dots\}$. Then $\{X_t : t \in \mathbf{Z}\}$ is said to satisfy a *strong mixing condition*. Obviously this condition implies (5.43) and thus the ergodicity of $\{X_t\}$. A further stronger condition than (5.44) is given as follows. The process $\{X_t : t \in \mathbf{Z}\}$ is said to satisfy a *uniform mixing condition* if

$$\sup_{A\in\mathcal{A}^q_{-\infty},\ B\in\mathcal{A}^\infty_{t+\tau}} \frac{|P(A \cap B) - P(A)P(B)|}{P(A)} \equiv \phi(\tau) \to 0 \quad \text{as } \tau \to \infty. \tag{5.45}$$

In the real world, it is plausible that if time distance between two phenomena becomes large, their mutual interaction will become weaker. Hence the three mixing conditions (5.43)-(5.45) seem natural ones which describe the actual world.

The following theorem is useful.

Theorem 5.4 *Suppose that a process* $\{X_t : t \in \mathbf{Z}\}$ *is strictly stationary and ergodic, and that there is a measurable function* $\phi : \mathbf{R}^\infty \to \mathbf{R}$*. Let* $Y_t = \phi(X_t, X_{t-1}, \dots)$ *define* $\{Y_t : t \in \mathbf{Z}\}$*. Then the process* $\{Y_t : t \in \mathbf{Z}\}$ *is strictly stationary and ergodic.*

PROOF Strict stationarity of $\{Y_t\}$ follows from the definition. Let

$$\phi_t(\mathbf{x}) = \phi(x_t, x_{t-1}, \dots)$$

for each $\mathbf{x} = (x_t, x_{t-1}, \dots) \in \mathbf{R}^\infty$ define the function $\phi_t : \mathbf{R}^\infty \to \mathbf{R}$. Let A be an invariant set for $\{Y_t\}$, i.e., $A = \{(Y_t, Y_{t-1}, \dots) \in C\}$ for all $t \in \mathbf{N}$ and some $C \in \mathcal{B}^\infty$. Thus

$$A = [\{\phi(X_t, X_{t-1}, \dots), \phi(X_{t-1}, X_{t-2}, \dots), \dots\} \in C]$$

for all $t \in \mathbf{N}$. Let

$$C_1 = [\mathbf{x} : \{\phi_t(\mathbf{x}), \phi_{t-1}(\mathbf{x}), \dots\} \in C] \in \mathcal{B}^\infty.$$

Then $A = [(X_t, X_{t-1}, \dots) \in C_1]$ for all $t \geq 1$. Hence A is also an invariant set for $\{X_t\}$. Thus $P(A) = 0$ or $P(A) = 1$ since $\{X_t\}$ is ergodic. □

Since the i.i.d sequences are strictly stationary and ergodic we have

Theorem 5.5 *Suppose that $\{u_t : t \in \mathbf{Z}\}$ is a sequence of random variables that are independent and identically distributed. Let*

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

where $\sum_{j=0}^{\infty} |a_j|^2 < \infty$. Then $\{X_t : t \in \mathbf{Z}\}$ is strictly stationary and ergodic.

Probability theory has its roots in games of chance, and it is often profitable to interpret result in terms of a gambling situation. Martingale is such an example.

Definition 5.5 *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $\{X_1, X_2, \dots\}$ a sequence of integrable random variables on $(\Omega, \mathcal{A}, P)$, and $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \cdots$ an increasing sequence of sub σ-algebras of $\mathcal{A}$, where X_t is assumed to be $\mathcal{A}_t$-measurable. The sequence $\{X_t\}$ is said to be a* martingale relative to the $\mathcal{A}_t$ *(alternatively, we say that $\{X_t, \mathcal{A}_t\}$ is a* martingale *if for all $t \in \mathbf{N}$,*

$$E(X_{t+1}|\mathcal{A}_t) = X_t, \text{ almost everywhere (a.e.).} \tag{5.46}$$

If $\{X_t, \mathcal{A}_t\}$ is a martingale, it is automatically a martingale relative to the σ-algebra $\sigma(X_1, \dots, X_t)$ (the smallest σ-algebra generated by $X_1, \dots, X_t$). To see this, condition both sides of (5.46) with respect to $\sigma(X_1, \dots, X_t)$. If we do not mention the $\mathcal{A}_t$ explicitly, henceforth, we always mean $\mathcal{A}_t = \sigma(X_1, \dots, X_t)$.

Martingales may be considered appropriate models for fair games, in the sense that X_t signifies the amount of money that a player has at time t. The martingale property states, then, that the average amount a player will have at time $(t+1)$, given that he has amount X_t at time t, is equal to X_t regardless of what his past fortune has been.

Definition 5.6 *Let $\{X_t : t \in \mathbf{N}\}$ be a stochastic process defined on a probability space $(\Omega, \mathcal{A}, P)$ and $\{\mathcal{A}_t : t \in \mathbf{N}\}$ an increasing sequence of sub σ-algebras of $\mathcal{A}$. If X_t is $\mathcal{A}_t$-measurable for each $t \in \mathbf{N}$, the sub σ-algebras $\{\mathcal{A}_t : t \in \mathbf{N}\}$ are said to be* adapted *to the stochastic process $\{X_t : t \in \mathbf{N}\}$ and $\{X_t, \mathcal{A}_t, t \in \mathbf{N}\}$ is said to be an* adapted stochastic process.

Definition 5.7 *If an adapted stochastic process $\{X_t, \mathcal{A}, t \in \mathbf{N}\}$ satisfies*

$$E(X_t|\mathcal{A}_{t-1}) = 0 \quad \text{a.e. for each } t \in \mathbf{N},$$

then $\{X_t, \mathcal{A}_t, t \in \mathbf{N}\}$ is called a martingale difference sequence.

The following assertions are immediate from the definition.

Theorem 5.6 *Let* $\{X_t, \mathcal{A}_t, t \in \mathbf{N}\}$ *be a martingale difference sequence and let* $S_n = \sum_{t=1}^{n} X_t$, $n \in \mathbf{N}$. *Then*

(i) $\{S_n, \mathcal{A}_n, n \in \mathbf{N}\}$ *is a martingale, i.e.,*

$$E(S_n|\mathcal{A}_{n-1}) = S_{n-1}, \text{ a.e. for each } n \geq 2,$$

(ii) if $E(X_t^2) < \infty$ *for each* $t \in \mathbf{N}$, *then*

$$E(X_t X_s) = 0 \quad \text{for } t \neq s.$$

Conversely, if $\{S_n, \mathcal{A}_n, n \in \mathbf{N}\}$ *is a martingale, and if we set* $X_n = S_n - S_{n-1}$ *for* $n \in \mathbf{N}$, *then* $\{X_n, \mathcal{A}_n, n \in \mathbf{N}\}$ *is a martingale difference sequence.*

Martingales will play an important role in Chapter 6 and financial problems in Chapter 7. Here we give a fundamental example of statistical estimation.

Example 5.5 *In statistical asymptotic theory, one of the most important quantities becomes a martingale under suitable conditions. Let* $\mathbf{X}_n = (X_1, \ldots, X_n)'$ *be a sequence of random variables forming a stochastic process, and possessing the probability density* $p_\theta^n(\mathbf{x})$, $\mathbf{x} = (x_1, \ldots, x_n)'$, *where* $\theta \in \Theta \subset \mathbf{R}^1$. *Suppose that the conditional density of* $\mathbf{X}_k$ *given* $\mathbf{X}_{k-1}$ *is given by*

$$p_\theta^k(\mathbf{x}_k|\mathbf{x}_{k-1}) = \frac{p_\theta^k(\mathbf{x}_k)}{p_\theta^{k-1}(\mathbf{x}_{k-1})},$$

where

$$p_\theta^{k-1}(\mathbf{x}_{k-1}) = \int p_\theta^k(\mathbf{x}_k) dx_k.$$

We assume that $p_\theta^k(\mathbf{x}_k)$, $k = 1, 2, \ldots, n$, *are differentiable with respect to* θ *and that*

$$\frac{\partial}{\partial\theta} p_\theta^{k-1}(\mathbf{x}_{k-1}) = \int \frac{\partial}{\partial\theta} p_\theta^k(\mathbf{x}) dx_k, \quad k = 1, \ldots, n. \tag{5.47}$$

Then the log-likelihood function based on $\mathbf{X}_n$ *is*

$$L_n(\theta) = \sum_{k=1}^{n} \log p_\theta^k(\mathbf{X}_k|\mathbf{X}_{k-1}),$$

where we take $p_\theta^0(\mathbf{X}_0) = 1$. *Let*

$$S_n = \frac{\partial}{\partial\theta} L_n(\theta) = \sum_{t=1}^{n} \left\{ \frac{\frac{\partial}{\partial\theta} p_\theta^k(\mathbf{x}_k)}{p_\theta^k(\mathbf{x}_k)} - \frac{\frac{\partial}{\partial\theta} p_\theta^{k-1}(\mathbf{x}_{k-1})}{p_\theta^{k-1}(\mathbf{x}_{k-1})} \right\},$$

which is called the score function. *For* $\mathcal{A}_k = \sigma(X_1, X_2, \ldots, X_k)$, *it follows from*

(5.47) *that*

$$
\begin{aligned}
E\left\{\frac{\frac{\partial}{\partial\theta}p_\theta^k(\mathbf{X}_k)}{p_\theta^k(\mathbf{X}_k)} - \frac{\frac{\partial}{\partial\theta}p_\theta^{k-1}(\mathbf{X}_{k-1})}{p_\theta^{k-1}(\mathbf{X}_{k-1})}\Big|\mathcal{A}_{k-1}\right\}
&= \int\left\{\frac{\frac{\partial}{\partial\theta}p_\theta^k(\mathbf{x}_k)}{p_\theta^k(\mathbf{x}_k)} - \frac{\frac{\partial}{\partial\theta}p_\theta^{k-1}(\mathbf{x}_{k-1})}{p_\theta^{k-1}(\mathbf{x}_{k-1})}\right\}\frac{p_\theta^k(\mathbf{x}_k)}{p_\theta^{k-1}(\mathbf{x}_{k-1})}dx_k \\
&= \int\frac{\frac{\partial}{\partial\theta}p_\theta^k(\mathbf{x}_k)}{p_\theta^k(\mathbf{x}_{k-1})}dx_k - \frac{\frac{\partial}{\partial\theta}p_\theta^{k-1}(\mathbf{x}_{k-1})}{p_\theta^{k-1}(\mathbf{x}_{k-1})} \\
&= 0, \qquad a.e.
\end{aligned}
$$

Therefore $\{S_n, \mathcal{A}_n\}$ *is a martingale. In many regular statistical models, the main order term of the maximum likelihood estimators is expressed by* S_n *after suitable standardization.*

5.4 Limit Theorems for Stochastic Processes

For dependent observations it is difficult to use the exact distribution theory. Thus we often use the asymptotic distribution theory when the length of observation tends to infinity. For this we need limit theorems for stochastic processes. In this section we shall state some useful limit theorems and central limit theorems for sample mean of dependent observations.

Theorem 5.7 *If* $\{X_t :\in \mathbf{Z}\}$ *is strictly stationary and ergodic and* $E\{|X_t|\} < \infty$*, then*

$$\frac{1}{n}\sum_{t=1}^{n} X_t \xrightarrow{a.s.} E(X_1). \tag{5.48}$$

Also if $E\{|X_t|^2\} < \infty$*, then*

$$\frac{1}{n}\sum_{t=1}^{n} X_t X_{t+m} \xrightarrow{a.s.} E(X_1 X_{1+m}). \tag{5.49}$$

PROOF The assertion (5.48) is nothing but the pointwise ergodic theorem (see e.g., Theorem 3.5.7 of Stout (1974)). Next, consider $X_t X_{t+m}$. Fixing m and allowing t to vary, this constitutes a sequence of random variables which again constitutes a strictly stationary process. If $\{X_t\}$ is ergodic then so is $\{X_t X_{t+m}\}$. Hence the assertion (5.49) follows from the pointwise ergodic theorem. □

We may state the following martingale convergence theorem (see, e.g., Hall and Heyde (1980)).

Theorem 5.8 *(Doob's martingale convergence theorem). If* $\{S_n, \mathcal{A}_n, n \in \mathbf{N}\}$

is a martingale such that $\sup_{n\geq 1} E|S_n| < \infty$, *then there exists a random variable* S *such that* $E|S| < \infty$ *and* $S_n \xrightarrow{a.s.} S$.

Next we provide some central limit theorems which are fundamental to derive the asymptotic distribution of statistics for stochastic processes. The first one is due to Ibragimov (1963).

Theorem 5.9 *Let* $\{X_t : t \in \mathbf{Z}\}$ *be a strictly stationary and ergodic process such that*

$$E(X_t^2) = \sigma^2, \quad 0 < \sigma^2 < \infty, \quad E(X_t|\mathcal{A}_{t-1}) = 0, \quad a.e.,$$

where $\mathcal{A}_t = \sigma(X_1, \ldots, X_t)$. *Then*

$$\frac{1}{\sigma\sqrt{n}} \sum_{t=1}^{n} X_t \xrightarrow{d} N(0,1).$$

Let $\{X_{n,t} : t = 1, \ldots, k_n\}$ be an array of random variables on a probability space $(\Omega, \mathcal{A}, P)$. Let $\{\mathcal{A}_{n,t} : 0 \leq t \leq k_n\}$ be any triangular array of sub σ-algebras of $\mathcal{A}$ such that for each n and $1 \leq t \leq k_n$, $X_{n,t}$ is $\mathcal{A}_{n,t}$-measurable and $\mathcal{A}_{n,t-1} \subset \mathcal{A}_{n,t}$. We denote $S_n = \sum_{t=1}^{k_n} X_t$. The following theorem is due to Brown (1971).

Theorem 5.10 *Suppose that* $\{X_{n,t}, \mathcal{A}_{n,t}, 1 \leq t \leq k_n\}$ *is a martingale difference array satisfying*

(i)

$$\sum_{t=1}^{k_n} E\{X_{n,t}^2 \chi(|X_{n,t}| > \varepsilon)\} \to 0 \text{ as } n \to \infty \text{ for all } \varepsilon > 0, \tag{5.50}$$

(Lindeberg condition)

(ii)

$$\sum_{t=1}^{k_n} E\{X_{n,t}^2 | \mathcal{A}_{n,t-1}\} \xrightarrow{p} 1. \tag{5.51}$$

Then

$$S_n \xrightarrow{d} N(0,1), \quad (n \to \infty).$$

Regarding other types of central limit theorems we state the following for mixing sequence. For proofs, see Ibragimov and Linnik (1971). We set $S_n = \sum_{t=1}^{n} X_t$ and $\sigma_n^2 = V(S_n)$.

Theorem 5.11 *If* $\{X_t : t \in \mathbf{Z}\}$ *is zero-mean strongly mixing, and* $\sigma_n^2/n \to \sigma^2$ *as* $n \to \infty$ $(\sigma > 0)$, *then* $S_n/\sigma_n \xrightarrow{d} N(0,1)$ *if and only if*

$$\lim_{N\to\infty} \limsup_{n\to\infty} \int_{|z|>N} z^2 dF_n(z) = 0,$$

where $F_n(x)$ is the distribution function of S_n/σ_n.

Theorem 5.12 *Let $\{X_t : t \in \mathbf{Z}\}$ be a strictly stationary process with $E(X_t) = 0$. Suppose that $\{X_t\}$ satisfies the uniform mixing condition and that the mixing coefficient $\phi(\tau)$ satisfies*

$$\sum_{\tau} \psi(\tau)^{1/2} < \infty.$$

Then the sum

$$\sigma^2 = E(X_0^2) + 2\sum_{t=1}^{\infty} E(X_0 X_t)$$

converges, and if $\sigma > 0$,

$$\frac{S_n}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty.$$

Exercises

5.1. Let $\{X_t : t \in \mathbf{Z}\}$ be a strictly stationary process. Then, show that the joint distribution function of $X_{t_1}, \ldots, X_{t_n}$ $(t_1, \ldots, t_n \in \mathbf{Z})$ depends only on $t_2 - t_1, \ldots, t_n - t_{n-1}$.

5.2. Verify (5.6) and (5.7).

5.3. Let $L_2 \equiv \{Y : E(|Y|^2) < \infty\}$. For $Y_n, W_n \in L_2$, define $< Y_n, W_n > \equiv E(Y_n \overline{W}_n)$, and write $Y = l.i.m._{n\to\infty} Y_n$ and $W = l.i.m._{n\to\infty} W_n$. Then, show that

$$< Y, W > = \lim_{n\to\infty} < Y_n, W_n >, \quad \text{(continuity of inner product)}.$$

5.4. Verify that the m-vector general linear process (5.41) has the spectral density matrix (5.42).

5.5. Show that the following relation holds:

$$\text{uniform mixing} \;\Rightarrow\; \text{strong mixing} \;\Rightarrow\; \text{mixing}.$$

5.6. Verify Theorem 5.6.

5.7. Let $\{X_{n,t}, \mathcal{A}_{n,t}\}$ be a martingale difference sequence. Suppose the Lindeberg condition:

$$\sum_{t=1}^{n} E\{X_{n,t}^2 \chi(|X_{n,t}| > \varepsilon)\} \to 0, \quad (n \to \infty), \text{ for every } \varepsilon > 0.$$

Then, show the following (i) and (ii):

(i) $\max_{1\le t\le n} |X_{n,t}| \xrightarrow{p} 0$,

(ii) there exists $M > 0$ satisfying

$$E\left\{\max_{1\le t\le n}|X_{n,t}|^2\right\} \le M < \infty.$$

CHAPTER 6

Time Series Analysis

Time series analysis is statistical analysis for stochastic processes. Recently a lot of statistical methods have been introduced in this field. The fundamental thing is to estimate statistical models which describe the time series concerned. As candidates of time series models, many nonlinear models, e.g., ARCH, GARCH, etc., besides classical linear models, e.g., AR, ARMA, etc., have been proposed in financial time series analysis.

This chapter explains typical linear and nonlinear parametric models, and states the asymptotically optimal estimation for their unknown parameters. We also discuss the problem of model selection by use of some information criteria. Although the parametric approach is powerful and dominant in time series analysis, it is often difficult for parametric models to describe the real world sufficiently. For this we address nonparametric and semiparametric estimation problems for spectra of stationary processes and trend functions of time series regression models. Local polynomial fitting and local likelihood for time series are also expounded.

So far we assumed stationarity of the concerned processes. However, this assumption is often severe for actual time series data. We mention the statistical inference for locally stationary processes, which are nonstationary. Stationary processes whose autocovariance functions converge to zero with power law decay are introduced. Because this rate of convergence is slower than that of the usual AR and ARMA processes, we call them long-memory processes. The phenomenon of long-memory was observed in many fields. The asymptotic estimation and testing theory is discussed.

The problem of prediction and discriminant analysis is important in financial time series analysis. We explain the best predictor in terms of spectral density and conditional expectation, and discuss its statistical estimation. As for discriminant analysis, we give an asymptotically optimal discriminator, and evaluate the asymptotic misclassification probabilities. Discriminant analysis for time series is applied to the credit rating problem in finance.

6.1 Time Series Model

This section introduces typical time series models. First, let us see an actual real time series data. Figure 6.1 plots the velocity of wind (mile/hour) for 111 consecutive days $Y_1, Y_2, \cdots, Y_{111}$ in New York. Figure 6.2 plots the logarithm difference of Y_t i.e., $X_t \equiv \log Y_{t+1} - \log Y_t,\ t = 1, 2, \cdots, 110$.

As one of the most fundamental methods in time series, if an observed stretch $X_1, X_2, \cdots, X_n$, is available, we often take a look at the behavior of the *sample autocorrelation function*

$$SACF(l) \equiv \frac{\sum_{t=1}^{n-l}(X_{t+l} - \bar{X}_n)(X_t - \bar{X}_n)}{\sum_{t=1}^{n}(X_t - \bar{X}_n)^2}, \tag{6.1}$$

where $\bar{X}_n = n^{-1}\sum_{t=1}^{n} X_t$. This is a sort of sample correlation between X_{t+l} and X_t, and if X_t's are mutually independent or uncorrelated, then we can imagine $SACF(l) \approx 0$ for $l \neq 0$. For the data $X_1, X_2, \cdots, X_{110}$ of Figure 6.2, Figure 6.3 plots the values of the $SACF(l),\ l = 0, 1, \cdots, 20$.

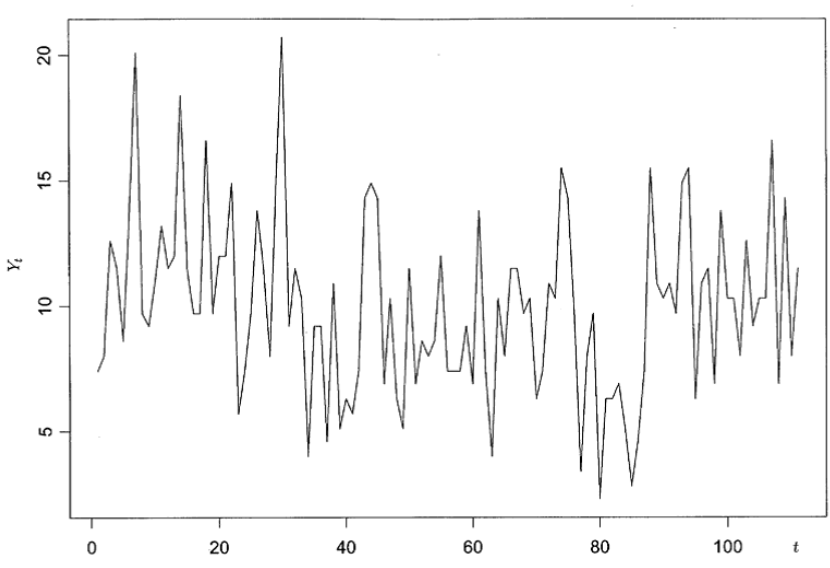


Figure 6.1 *The velocity of wind (mile/hour) for 111 consecutive days* $Y_1, Y_2, \ldots, Y_{111}$ *in New York.*

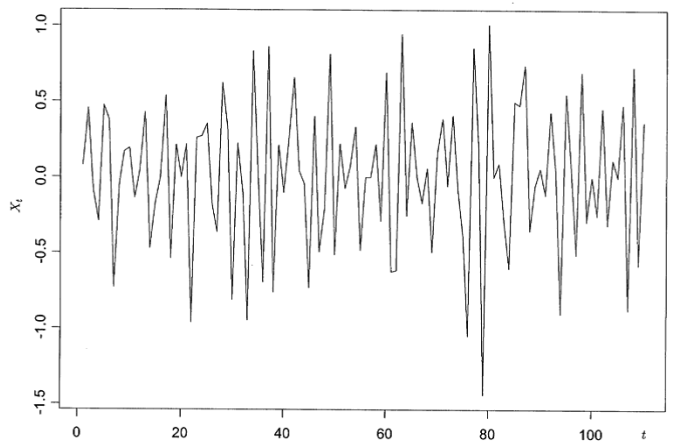


Figure 6.2 *The logarithm difference* $X_t = \log Y_{t+1} - \log Y_t$ *of* Y_t.

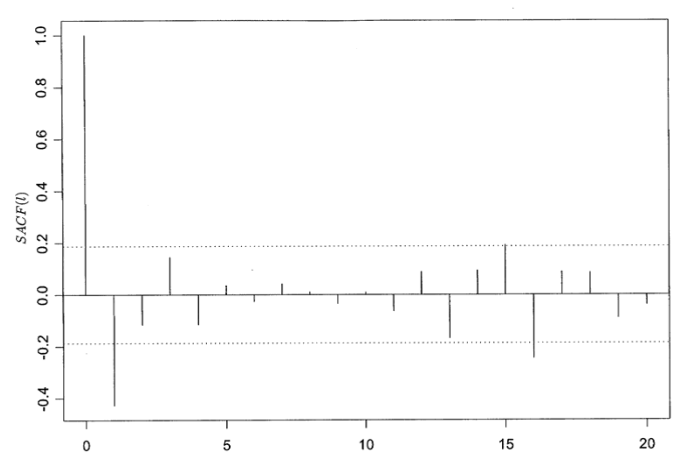


Figure 6.3 *The values of the* $SACF(l)$, $l = 0, 1, \ldots, 20$.

We observe that there exist $SACF(l),\ l \neq 0$, whose absolute values are fairly large. Thus it seems difficult to suppose that X_t's are mutually independent or uncorrelated. For such data, how should we make the time series model? From the viewpoint of regression analysis, we can think of the following model

$$X_t = -b_1 X_{t-1} - \cdots - b_p X_{t-p} + u_t, \tag{6.2}$$

where X_t is expressed as the sum of linear combination of the past values $X_{t-1}, \cdots, X_{t-p}$ and a noise u_t. Here $\{u_t\} \sim i.i.d\ (0, \sigma^2)$. Then $\{X_t\}$ defined by (6.2) is called a *pth order autoregressive model* (AR(p)), which seems the most intuitive and natural one describing dependence. For simplicity we often write $\{X_t\} \sim \mathrm{AR}(p)$.

Now we suppose that the wind velocity data $X_1, X_2, \cdots, X_{110}$ follow the model (6.2). Since the order p, the coefficients $b_1, \cdots, b_p$ and the variance σ^2 of u_t, are unknown, we have to estimate them from the data. Estimation theory for time series models will be discussed in the next section. Hence, skipping the details, we estimate the unknown parameter $(p, b_1, \cdots, b_p, \sigma^2)$ by a standard method. Then the estimated value is given by

$$(\hat{p}, \hat{b}_1, \cdots, \hat{b}_p, \hat{\sigma}^2) = (4, 0.6452, 0.5079, 0.2233, 0.1766, 0.1647),$$

which implies $\{X_t\} \sim \mathrm{AR}(4)$. From this we observe that the present value X_t is affected by the past four values $X_{t-1}, \cdots, X_{t-4}$.

Here, let us check the characteristics of AR models. First, consider the AR(1) model

$$X_t = -b_1 X_{t-1} + u_t. \tag{6.3}$$

Expressing X_{t-1} in (6.3) recursively, we obtain

$$\begin{aligned} X_t &= -b_t X_{t-1} + u_t \\ &= -b_1(-b_1 X_{t-2} + u_{t-1}) + u_t \\ &= (-b_1)^2 X_{t-2} + (-b_1) u_{t-1} + u_t \\ &\quad \vdots \\ &= (-b_1)^{s+1} X_{t-s-1} + (-b_1)^s u_{t-s} + \cdots + (-b_1) u_{t-1} + u_t. \end{aligned}$$

Hence, if b_1 satisfies

$$|b_1| < 1, \tag{6.4}$$

and if $\{X_t\}$ is stationary, then

$$E\{|X_t - \sum_{j=0}^{s} (-b_1)^j u_{t-j}|^2\} = |b_1|^{2(s+1)} E\{|X_{t-s-1}|^2\} \to 0 \quad (s \to \infty),$$

which leads to the representation

$$X_t = \sum_{j=0}^{\infty} (-b_1)^j u_{t-j}, \tag{6.5}$$

where the right-hand side is defined in the sense of $l.i.m._{n\to\infty}$. Summarizing the above, we can see that the AR(1) model (6.3) is expressed as the linear process (6.5) under the condition (6.4). It is possible to express the general AR(p) model as a linear process if some appropriate condition corresponding to (6.4) is assumed. In what follows we ascertain this. Introduce the following polynomial function of $z \in \mathbf{C}$:

$$\beta(z) = \sum_{j=0}^{p} b_j z^j, \quad (b_0 = 1). \tag{6.6}$$

Assumption 6.1 *$\beta(z) = 0$ has no roots in $\mathcal{D} \equiv \{z \in \mathbf{C} : |z| \le 1\}$.*

If $p = 1$, then it is easy to check that Assumption 6.1 is equivalent to (6.4). Let $z_1, \cdots, z_p$ be the roots of the equation $\beta(z) = 0$. Then Assumption 6.1 is equivalent that $|z_j| > 1$ for all $j = 1, \cdots, p$. Factorize $\beta(z)$ as $\beta(z) = \prod_{j=1}^{p}(1 - z_j^{-1}z)$. Using the backward shift operator $B : B^j X_t = X_{t-j},\ j \in \mathbf{Z}$, we can write (6.2) as

$$\beta(B)X_t = \prod_{j=1}^{p}(1 - z_j^{-1}B)X_t = u_t. \tag{6.7}$$

Since $|z_j| > 1$, $(1 - z_j^{-1}B)^{-1}$ can be expressed as $\sum_{l=0}^{\infty}(z_j^{-1})^l B^l$. For each j, $j = 1, \cdots, p$, operate $(1 - z_j^{-1}B)^{-1}$ on both sides of (6.7) consecutively. Then, under Assumption 6.1, it is seen that the AR(p) process (6.2) can be written as

$$X_t = \sum_{j=0}^{\infty} \rho_j u_{t-j}, \quad (\sum_{j=0}^{\infty} |\rho_j| < \infty\ ).$$

In the usual discussion for AR processes, we assume Assumption 6.1. But, when we deal with economic or financial time series, it is often to use AR models with $z_l = 1$. If an AR model has some roots with $z_l = 1$, we say that it has *unit roots*. Time series models with unit roots become nonstationary. It is known that their estimation and testing are very difficult. For $\{Y_t\}$ in Example 5.4 of Chapter 5, if we set $Y_0 = 0$, then it becomes the following AR model with unit root:

$$Y_t = Y_{t-1} + u_t\ .$$

Generating $\{Y_t\}$ from $\{u_t\} \sim i.i.d.\ N(0, 1)$, we plotted the graph in Figure 5.2, which associates the typical feature of financial data, e.g., stock price.

Although the autoregressive model is a very natural one, the following model generalizing the u_t part, is often used. If $\{X_t : t \in \mathbf{Z}\}$ is generated by

$$\sum_{j=0}^{p} b_j X_{t-j} = \sum_{j=0}^{q} a_j u_{t-j}, \quad (a_0 = b_0 = 1, a_q \neq 0, b_p \neq 0), \tag{6.8}$$

it is called an *autoregressive moving average model* (ARMA(p,q)). Simply we write $\{X_t\} \sim ARMA$(p,q). Similarly as in the case of AR(p), it is easily seen that Assumption 6.1 is a sufficient condition of stationarity for the ARMA(p,q) model. Assuming that the ARMA(p,q) in (6.8) is stationary, let us see its spectral structure. From Theorem 5.3, the processes $\{X_t\}$ and $\{u_t\}$ have the spectral representations

$$X_t = \int_{-\pi}^{\pi} e^{-it\lambda} dZ_X(\lambda),\ u_t = \int_{-\pi}^{\pi} e^{-it\lambda} dZ_u(\lambda), \tag{6.9}$$

where $E|dZ_u(\lambda)|^2 = (\sigma^2/2\pi)d\lambda$. Then the spectral representation of (6.8) is given by

$$\int_{-\pi}^{\pi} e^{-it\lambda}\beta(e^{i\lambda})dZ_X(\lambda) = \int_{-\pi}^{\pi} e^{-it\lambda}\alpha(e^{i\lambda})dZ_u(\lambda), \tag{6.10}$$

where $\alpha(e^{i\lambda}) = \sum_{j=0}^{q} \alpha_j e^{ij\lambda}$ and $\beta(e^{i\lambda}) = \sum_{j=0}^{p} b_j e^{ij\lambda}$. Hence $\beta(e^{i\lambda})dZ_X(\lambda) = \alpha(e^{i\lambda})dZ_u(\lambda)$, which yields

$$\begin{aligned} E\{|dZ_X(\lambda)|^2\} &= E\left[\left|\frac{\alpha(e^{i\lambda})}{\beta(e^{i\lambda})}dZ_u(\lambda)\right|^2\right] \\ &= \left|\frac{\alpha(e^{i\lambda})}{\beta(e^{i\lambda})}\right|^2 E[|dZ_u(\lambda)|^2] \\ &= \left|\frac{\alpha(e^{i\lambda})}{\beta(e^{i\lambda})}\right|^2 \frac{\sigma^2}{2\pi}d\lambda. \end{aligned}$$

Therefore the ARMA(p,q) defined by (6.8) has the spectral density

$$f_X(\lambda) = \frac{\sigma^2}{2\pi}\frac{|\alpha(e^{i\lambda})|^2}{|\beta(e^{i\lambda})|^2}. \tag{6.11}$$

The process (6.8) is multi-dimensionalized as follows. If $\mathbf{X}_t = (X_{1,t}, \cdots, X_{m,t})'$ is defined by the relation

$$\sum_{j=0}^{p} B(j)\mathbf{X}_{t-j} = \sum_{j=0}^{q} A(j)\mathbf{U}_{t-j}, \tag{6.12}$$

then $\{\mathbf{X}_t : t \in \mathbf{Z}\}$ is called an *m-dimensional autoregressive moving average process*, and is written as $\{\mathbf{X}_t\} \sim$ VARMA(p,q). Here $\{A(j)\}$ and $\{B(j)\}$ are sequences of $m \times m$-matrices, $A(0) = B(0) = \mathbf{I}_m$($m \times m$-identity matrix), and $\{\mathbf{U}_t\}$ is an uncorrelated process with $E\mathbf{U}_t = \mathbf{0}$ and Var$\{\mathbf{U}_t\} = V$. Write

$$B(z) = \det\left\{\sum_{j=0}^{p} B(j)z^j\right\},\quad (\ z \in \mathbf{C}\ ).$$

The assumption corresponding to Assumption 6.1 is stated as follows.

Assumption 6.2 *$B(z) = 0$ has no roots in $D = \{z \in \mathbf{C} : |z| \le 1\}$.*

Under this assumption, the VARMA(p,q) given in (6.12) becomes stationary, hence, it has the spectral density matrix

$$\mathbf{f}(\lambda) = \frac{1}{2\pi}\left\{\sum_{j=0}^{p} B(j)e^{ij\lambda}\right\}^{-1}\left\{\sum_{j=0}^{q} A(j)e^{ij\lambda}\right\}V \times \left\{\sum_{j=0}^{q} A(j)e^{ij\lambda}\right\}^{*}\left\{\sum_{j=0}^{p} B(j)e^{ij\lambda}\right\}^{*^{-1}} \tag{6.13}$$

(Exercise 6.1).

So far we saw typical linear time series models, however, it is not sufficient for linear models such as ARMA models to describe the real world. Tong (1990) discussed fitting nonlinear models to the Canadian Lynx data and the sunspot numbers data in various ways. A list of the literature for nonlinear analysis of data from solar physics, ecology, economics, medical science, hydrology, environmental sciences, and other areas is available in Tong (1990).

Using ARCH models which are typical nonlinear models, Hafner (1998) gave an extensive financial time series analysis. The results by Tong and Hafner reveal that many relationships in real data are nonlinear. Thus the analysis of nonlinear time series is becoming a central component of time series analysis. In what follows, representative nonlinear time series models are introduced.

Example 6.1 (Bilinear model) *A stochastic process* $\{X_t : t \in \mathbf{Z}\}$ *is said to follow a* bilinear model *(denoted by BL(p,q,r)) if it satisfies*

$$X_t + \sum_{j=1}^{p} a_j X_{t-j} = b_{00} + \sum_{j=1}^{q}\sum_{k=1}^{r} b_{jk} X_{t-j}u_{t-k} + u_t, \tag{6.14}$$

where a_j and b_{jk} are real constants. This model was first studied by Granger and Andersen (1978). Subba Rao (1981) showed that a bilinear model can grasp sudden large amplitude bursts and is suitable for seismological data like earthquakes and underground nuclear explosions.

Example 6.2 (SETAR model) *A* self-exciting threshold autoregressive model *(SETAR) proposed by Tong (1990) is of the form*

$$X_t = a_0^{(j)} + a_1^{(j)} X_{t-1} + \cdots + a_p^{(j)} X_{t-p} + u_t, \tag{6.15}$$

if $X_{t-d} \in \Omega_j$, $j = 1, \cdots, k$, where the Ω_j's are disjoint intervals on $\mathbf{R}$ *with $\bigcup_{j=1}^{k} \Omega_j = \mathbf{R}$, and d is called the* threshold lag. *We denote (6.15) by SETAR$(k; p, \cdots, p)$, where p is repeated k times. Using SETAR models, Tong (1990) gave an extensive study for various real data.*

Example 6.3 (EXPAR model) *Haggan and Ozaki (1981) proposed an* ex-

ponential autoregressive model *(EXPAR), which is of the form*

$$X_t = \{a_1 + b_1 \exp(-cX_{t-d}^2)\}X_{t-1} + \cdots + \{a_p + b_p \exp(-cX_{t-d}^2)\}X_{t-p} + u_t, \tag{6.16}$$

where $c \geq 0$, $d \in \mathbf{N}$, *and* $a_j, b_j, j = 1, \cdots, p$, *are real constants. The model was fitted to the Canadian lynx data, and demonstrated that this is suitable for reproduction of nonlinear phenomena like limit cycles, amplitude-dependent frequency, and jump phenomena.*

Example 6.4 (ARCH or GARCH model) *Traditional econometric models assume a constant one-period forecast variance (e.g., AR model). But empirical financial studies show that this assumption is very severe. In order to overcome this implausible assumption, Engle (1982) introduced an* autoregressive conditional heteroscedastic model *(ARCH(q)), which is defined as*

$$\begin{cases} E(X_t|\mathcal{F}_{t-1}) = 0 \qquad a.e., \\ Var(X_t|\mathcal{F}_{t-1}) = a_0 + \sum_{j=1}^{q} a_j X_{t-j}^2, \ a.e., \end{cases} \tag{6.17}$$

where $\mathcal{F}_{t-1}$ *is the* σ*-algebra generated by* $\{X_{t-1}, X_{t-2}, \cdots\}$ *and* $a_0 > 0, \ a_j \geq 0, \ j = 1, \cdots, q$. *A concrete representation of the ARCH(q) model is given by*

$$\begin{cases} X_t = u_t \sqrt{h_t}\ , \\ h_t = a_0 + \sum_{j=1}^{q} a_j X_{t-j}^2\ , \end{cases} \tag{6.18}$$

where $\{u_t\}$ *is a sequence of i.i.d.* $(0, \sigma^2)$ *random variables. Engle won the Nobel Prize in 2003 for his proposal of the ARCH model and construction of ARCH-based financial time series analysis. Bollerslev (1986) generalized (6.17) to*

$$\begin{cases} E(X_t|\mathcal{F}_{t-1}) = 0 \quad a.e., \\ Var(X_t|\mathcal{F}_{t-1}) \equiv h_t = a_0 + \sum_{j=1}^{q} a_j X_{t-j}^2 + \sum_{j=1}^{p} b_j h_{t-j}, \ a.e., \end{cases} \tag{6.19}$$

where $a_0 > 0, \ a_j \geq 0, \ j = 1, \cdots, q, \ b_j \geq 0, \ j = 1, \cdots, p$. *This is called a* generalized autoregressive conditional heteroscedastic model *(GARCH(p, q)). The ARCH and GARCH models are the most fundamental ones in financial time series analysis.*

Example 6.5 (EGARCH model) *Financial empirical studies show that stock returns are negatively correlated with changes in returns volatility, i.e., volatility tends to rise in response to 'bad news' and to fall in response to 'good news'. ARCH and GARCH models cannot describe this feature. To allow the asymmetric effects between positive and negative asset returns, Nelson (1991) proposed the following* exponential GARCH model *(EGARCH(p, q)):*

$$\begin{cases} X_t = u_t \sigma_t, \\ \log \sigma^2 = a_0 + \sum_{j=1}^{p} a_j \frac{|X_{t-j}| + \gamma_j X_{t-j}}{\sigma_{t-j}} + \sum_{j=1}^{q} b_j \log \sigma_{t-j}^2, \end{cases} \tag{6.20}$$

where a_j, b_j, γ_j *are unknown parameters, and are allowed to be negative unlike ARCH and GARCH parameters. If we assume* $a_1\gamma_1 < 0$, *then the asymmetry of the model is understandable.*

In Figure 6.4, we plot $X_1, X_2, \cdots, X_{200}$ generated by (6.20) when $\{u_t\} \sim$ i.i.d. $N(0,1)$. Here the real line shows the case of $p = 1,\ q = 1,\ a_0 = 0,\ a_1 = 0.5,\ \gamma_1 = -0.3,\ b_1 = 0.2$, and the dot line shows the case of $p = 1,\ q = 1,\ a_0 = 0,\ a_1 = 0.5,\ \gamma_1 = -0.3,\ b_1 = 0.9$. The parameter b_1 controls stationarity of $\{X_t\}$. It is seen that the amplitude of $\{X_t\}$ becomes large as b_1 tends to 1.

Figure 6.5 plots $\sigma_1^2,\ \sigma_2^2, \cdots, \sigma_{200}^2$ by real line and $X_1,\ X_2, \cdots, X_{200}$ by dot line generated by (6.20) when $p = 1,\ q = 1,\ a_0 = 0,\ a_1 = 0.5,\ \gamma_1 = -0.3,\ b_1 = 0.2$. Then it shows an EGARCH tendency that the value of σ_t^2 right after the value of X_t gets small becomes larger than that right after the value of X_t gets large.

Example 6.6 (TGARCH model) *We can combine GARCH and SETAR models. Let $\{X_t\}$ be defined by*

$$\begin{cases} X_t = u_t \sigma_t, \\ \sigma_t^2 = a_0 + \sum_{i=1}^{p} \{a_i + \gamma_i S_{t-i}\} X_{t-i}^2 + \sum_{j=1}^{q} b_j \sigma_{t-j}^2, \end{cases} \tag{6.21}$$

where the coefficients $\{a_i\}, \{b_j\}, \{\gamma_i\}$ are non-negative, $\{u_t\} \sim$ i.i.d. $(0, \sigma^2)$, and

$$S_{t-i} = \begin{cases} 1, & \text{if } X_{t-i} \le 0, \\ 0, & \text{if } X_{t-i} > 0. \end{cases}$$

This is called a threshold GARCH model *(TGARCH(p, q)).*

Example 6.7 ((Stochastic volatility model) *A stochastic process $\{X_t : t \in \mathbf{Z}\}$ is said to follow a* stochastic volatility model *(SV(m)) if it satisfies*

$$\begin{cases} X_t = \sigma_t u_t \\ \log \sigma_t^2 - \alpha_1 \log \sigma_{t-1}^2 - \cdots - \alpha_m \log \sigma_{t-m}^2 = \alpha_0 + v_t, \end{cases} \tag{6.22}$$

where $\{u_t\} \sim$ i.i.d. (0.1), $\{v_t\} \sim$ i.i.d. $(0, \sigma_v^2)$, and $\{u_t\}$ and $\{v_t\}$ are mutually independent. For stationarity of $\{\log \sigma_t^2\}$, the coefficients $\{\alpha_j\}$ are assumed to satisfy that all the roots of the equation

$$1 - \alpha_1 z - \cdots - \alpha_m z^m = 0$$

lie in $D = \{z \in \mathbf{C} : |z| > 1\}$.

Example 6.8 (ARCH(∞) model) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\{\mathcal{F}_t\}$ be a sequence of sub-σ-field of $\mathcal{F}$ satisfying $\mathcal{F}_t \subset \mathcal{F}_{t+1},\ t \in \mathbf{Z}$. Giraitis et al.(2000) introduced an* ARCH(∞) model*, which is defined by*

$$\begin{cases} X_t = \sigma_t u_t \\ \sigma_t^2 = a_0 + \sum_{j=1}^{\infty} a_j X_{t-j}^2, \end{cases} \tag{6.23}$$

where $a_0 > 0,\ a_j \ge 0,\ j = 1, 2, \cdots,\ \{u_t\} \sim$ i.i.d. $(0, 1)$, and u_t is $\mathcal{F}_t$-measurable and independent of $\mathcal{F}_{t-1}$. The class of ARCH(∞) models is larger than the class of stationary GARCH(p, q) models.

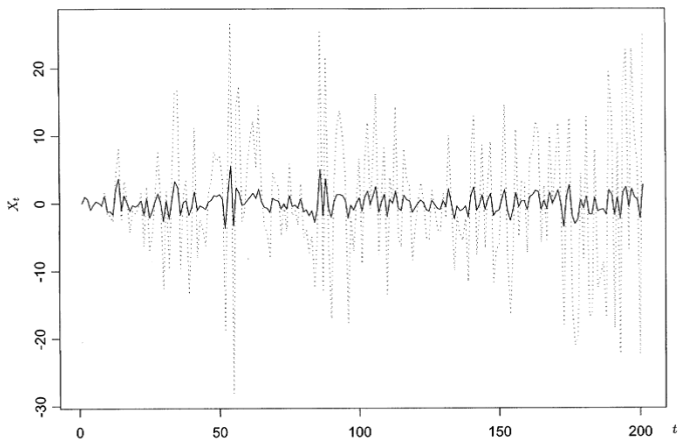


Figure 6.4 *The observed stretch* $X_1, X_2, \ldots, X_{200}$ *generated by (6.20).*

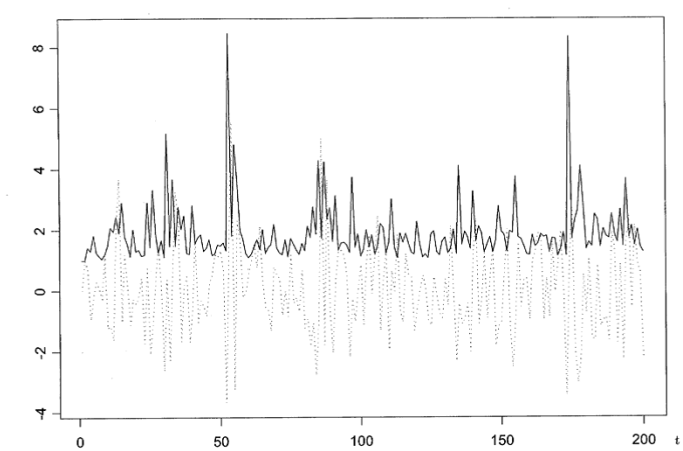


Figure 6.5 *The volatility* $\sigma_1^2,\ \sigma_2^2, \ldots, \sigma_{200}^2$ *(real line) and the observed stretch* $X_1,\ X_2, \ldots, X_{200}$ *(dot line) generated by (6.20).*

The GARCH (EGARCH, TGARCH, etc.) modeling may be applied to the innovation process instead of the initial process. Thus we may consider a linear regression model with GARCH (EGARCH, TGARCH, etc.) errors:

$$\begin{cases} Y_t = z_t\beta + X_t \\ \{X_t\} \sim \text{GARCH (EGARCH, TGARCH, etc.)}, \end{cases} \tag{6.24}$$

or an ARMA model with GARCH (EGARCH, TGARCH, etc.) errors (ARMA-GARCH (EGARCH, TGARCH, etc.)):

$$\begin{cases} Y_t + \beta_1 Y_{t-1} + \cdots + \beta_r Y_{t-r} = X_t + \alpha_1 X_{t-1} + \cdots + \alpha_s X_{t-s}, \\ \{X_t\} \sim \text{GARCH (EGARCH, TGARCH, etc.)}. \end{cases} \tag{6.25}$$

In this way we can construct infinitely many nonlinear time series models. The following example gives a very general and persuasive model which includes ARCH, AR-ARCH, SETAR, EXPAR, etc. as special cases.

Example 6.9 (CHARN model) *A stochastic process* $\{\mathbf{X}_t = (X_{1,t}, \ldots, X_{m,t})' : t \in \mathbf{Z}\}$ *is said to follow a* conditional heteroscedastic autoregressive nonlinear model *(CHARN) if it satisfies*

$$\mathbf{X}_t = \mathbf{F}_{\boldsymbol{\theta}}(\mathbf{X}_{t-1}, \cdots, \mathbf{X}_{t-p}) + \mathbf{H}_{\boldsymbol{\theta}}(\mathbf{X}_{t-1}, \cdots, \mathbf{X}_{t-q})\mathbf{U}_t, \tag{6.26}$$

where $\mathbf{U}_t = (U_{i,t}, \cdots, U_{m,t})' \sim$ *i.i.d.* $(\mathbf{0}, V)$, $\mathbf{F}_{\boldsymbol{\theta}} : \mathbf{R}^{mp} \to \mathbf{R}^m$ *and* $\mathbf{H}_{\boldsymbol{\theta}} : \mathbf{R}^{mq} \to \mathbf{R}^m \times \mathbf{R}^m$ *are measurable functions, and* $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_r)' \in \Theta \subset \mathbf{R}^r$ *is an unknown parameter (Härdle, Tsybakov and Yang (1998)). This model is very general, and can be applied to analysis of brain and muscular waves as well as financial time series analysis (Kato, Taniguchi and Honda (2006)).*

When we analyze the nonlinear time series models mentioned, their stationarity is a fundamental and important condition. In what follows, we state some sufficient conditions for typical nonlinear models to be stationary.

Theorem 6.1 (Chen and An (1998)) *If the GARCH model (6.19) satisfies*

$$\sum_{i=1}^{q} a_i + \sum_{j=1}^{p} b_j < 1, \tag{6.27}$$

then $\{X_t\}$ *is strictly stationary.*

We can give a sufficient condition for the CHARN model (6.26) to be stationary. We denote by $|A|$ the sum of the absolute values of all the elements of a matrix or vector A. Let $\mathbf{x} = (x_{11}, \ldots, x_{1m}, x_{21}, \ldots, x_{2m}, \ldots, x_{p1}, \ldots, x_{pm})' \in \mathbf{R}^{mp}$. Without loss of generality we assume $p = q$ in (6.26).

Theorem 6.2 (Lu and Jiang (2001)) *Let* $\{\mathbf{X}_t\}$ *be generated by the CHARN model (6.26). Assume the following (i)-(iv):*

(i) $\mathbf{U}_t$ *has the probability density function* $p(\mathbf{u}) > 0$ *a.e.,* $\mathbf{u} \in \mathbf{R}^m$.
(ii) There exist $a_{ij} \geq 0,\ b_{ij} \geq 0,\ 1 \leq i \leq m,\ 1 \leq j \leq p$, *such that*

$$|\mathbf{F}_\theta(\mathbf{x})| \leq \sum_{i=1}^{m}\sum_{j=1}^{p} a_{ij}|x_{ij}| + o(|\mathbf{x}|),$$

$$|\mathbf{H}_\theta(\mathbf{x})| \leq \sum_{i=1}^{m}\sum_{j=1}^{p} b_{ij}|x_{ij}| + o(|\mathbf{x}|),\ \text{as } |\mathbf{x}| \to \infty.$$

(iii) $\mathbf{H}_\theta(\mathbf{x})$ *is a continuous and symmetric function with respect to* $\mathbf{x}$, *and there exists* $\lambda > 0$ *such that*

$$\{\text{the minimum eigenvalue of } \mathbf{H}_\theta(\mathbf{x})\} \geq \lambda$$

for all $\mathbf{x} \in \mathbf{R}^{mp}$.
(iv)

$$\max_{1\leq i\leq m}\left\{\sum_{j=1}^{p} a_{ij} + E|\mathbf{U}_1|\sum_{j=1}^{p} b_{ij}\right\} < 1.$$

Then, $\{\mathbf{X}_t\}$ *is strictly stationary.*

In probability theoretical finance, continuous time stochastic processes are used to describe price processes.

Example 6.10 (Diffusion-type process) *A stochastic process* $\{\mathbf{X}_t = (X_{1,t}, \ldots, X_{m,t})' : t \in [0,T]\}$ *is called an* m-vector diffusion-type process *if it satisfies*

$$d\mathbf{X}_t = A_t dt + B_t d\mathbf{W}_t \tag{6.28}$$

where $\{\mathbf{W}_t : t \in [0,T]\}$ *is a Wiener process, and* A_t *and* B_t *are measurable with respect to* $\mathcal{F}(\mathbf{X}_s : 0 \leq s \leq t)$ *(a brief mathematical explanation of (6.28) is given in the Appendix). If* A_t *and* B_t *are* $\mathbf{X}_t$*-measurable, i.e.,*

$$d\mathbf{X}_t = A_t(\mathbf{X}_t)dt + B_t(\mathbf{X}_t)d\mathbf{W}_t, \tag{6.29}$$

then $\{\mathbf{X}_t\}$ *is called an* m-vector diffusion process.

Since actual financial time series data are essentially discrete time observations, continuous time stochastic processes are *models* to describe the phenomena. As far as we are interested in the statistical analysis, discrete time models are more convenient in many aspects. Hence, in this book, we will develop the discussion based on discrete time models mainly.

Here we mention the relation between discrete and continuous time models very briefly. Divide $[0,T]$ into n subintervals with length $s_n = T/n$, and let $t_k = ks_n,\ k = 0, 1, \ldots, n$. For $\{\epsilon_k\} \sim i.i.d.\ N(0,1)$, define

$$\xi_k = \{\mathrm{Var}(\log \epsilon_1^2)\}^{-1/2}\{\log \epsilon_k^2 - \mathrm{E}(\log \epsilon_k^2)\}.$$

Let $\{X_{n,k}\}$ be generated by the following AR-GARCH model:

$$\begin{cases} X_{n,k} - X_{n,k-1} = (\gamma_0 + \gamma_1 \sigma_{n,k}^2) s_n + \sigma_{n,k} s_n^{1/2} \epsilon_k, \\ \log \sigma_{n,k}^2 = \beta_0 s_n + (1 + \beta_1 s_n) \log \sigma_{n,k-1}^2 + \beta_2 s_n^{1/2} \xi_{k-1}. \end{cases} \tag{6.30}$$

Next, define the continuous time model $\{X_{n,t}\}$ by

$$\begin{cases} X_{n,t} = X_{n,k}, \quad t \in [t_k, t_{k+1}), \\ \sigma_{n,t}^2 = \sigma_{n,k}^2, \quad t \in [t_k, t_{k+1}). \end{cases} \tag{6.31}$$

Nelson (1990) showed that $\{X_{n,t}\}$ and $\{\sigma_{n,t}^2\}$ converge in distribution to $\{X_t\}$ and $\{\sigma_t^2\}$, as $n \to \infty$, respectively, which are defined by the stochastic differential equation

$$\begin{cases} dX_t = (\gamma_0 + \gamma_1 \sigma_t^2) dt + \sigma_t dW_{1,t}, \\ d \log \sigma_t^2 = (\beta_0 + \beta_1 \log \sigma_t^2) dt + \beta dW_{2,t}, \end{cases} \tag{6.32}$$

where $\{W_{1,t}\}$ and $\{W_{2,t}\}$ are mutually independent Wiener processes. Thus we may understand that (6.32) is a continuous time limit of model (6.30). Furthermore, for various GARCH models, Duan (1997) gave their limit processes of diffusion-type.

6.2 Estimation of Time Series Models

In the previous section we introduced a lot of time series models described by unknown parameters. Actually we need to estimate the unknown parameters from real data. First, consider the following autoregressive model (AR(p)) which is one of the most fundamental models:

$$X_t + b_1 X_{t-1} + \cdots + b_p X_{t-p} = u_t, \tag{6.33}$$

where $\{u_t\} \sim$ i.i.d. $N(0, \sigma^2)$, and the coefficients $b_1, \ldots, b_p$ are assumed to satisfy Assumption 6.1.

Now, how should we estimate the unknown parameter $\boldsymbol{\theta} = (b_1, \ldots, b_p, \sigma^2)'$? In the estimation theory for independent sample, we already saw in Section 3.4 that the maximum likelihood estimator (MLE) is asymptotically efficient. In fact, if we say from the conclusion, this claim holds true for dependent sample.

Below we think of the MLE of $\boldsymbol{\theta}$. Let $\mathbf{X} = (X_1, \ldots, X_n)'$ be an observed stretch from (6.33). Consider the transformation

$$(X_1, \ldots, X_p, u_{p+1}, \ldots, u_n)' \to (X_1, \ldots, X_p, X_{p+1}, \ldots, X_n)'. \tag{6.34}$$

Since it is possible to express (6.33) as

$$X_t = \textstyle\sum_{j=0}^{\infty} \rho_j u_{t-j},$$

$\{u_{p+1}, \ldots, u_n\}$ and $\mathbf{X}_p = (X_1, \ldots, X_p)'$ become mutually independent. Hence, the likelihood of $(X_1, \ldots, X_p, u_{p+1}, \ldots, u_n)'$ is

$$\phi_p(\mathbf{X}_p) \frac{1}{(2\pi)^{\frac{n-p}{2}} (\sigma^2)^{\frac{n-p}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{t=p+1}^{n} u_t^2 \right\}, \tag{6.35}$$

where $\phi_p(\cdot)$ is the probability density function of $\mathbf{X}_p$. From (6.33), it follows that the Jacobian of the transformation (6.34) is 1, hence, by the formula for change of variables (Theorem A.6), we can see that the likelihood of $\mathbf{X}$ is

$$\begin{aligned} L_n(\boldsymbol{\theta}) =& \phi_p(\mathbf{X}_p) \frac{1}{(2\pi)^{\frac{n-p}{2}} (\sigma^2)^{\frac{n-p}{2}}} \\ & \exp\left\{ -\frac{1}{2\sigma^2} \sum_{t=p+1}^{n} (X_t + b_1 X_{t-1} + \cdots + b_p X_{t-p})^2 \right\}. \end{aligned} \tag{6.36}$$

Thus, the logarithm of likelihood, i.e., $l_n(\boldsymbol{\theta}) = \log L_n(\boldsymbol{\theta})$ is

$$\begin{aligned} l_n(\boldsymbol{\theta}) = \log \phi_p(\mathbf{X}_p) - \frac{n-p}{2} \log 2\pi - \frac{n-p}{2} \log \sigma^2 \\ - \frac{1}{2\sigma^2} \sum_{t=p+1}^{n} (X_t + b_1 X_{t-1} + \cdots + b_p X_{t-p})^2. \end{aligned} \tag{6.37}$$

The MLE $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\sup_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}).$$

However, it is difficult to obtain $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ in a closed form even if the process is an AR(1) (Exercise 6.2). Dividing (6.36) by $\phi_p(\mathbf{X}_p)$, we get the conditional likelihood of $\mathbf{X}$ given $\mathbf{X}_p$. Taking its logarithm, and dropping the constant term we obtain

$$l_n^Q(\boldsymbol{\theta}) = -\frac{n-p}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=p+1}^{n} (X_t + b_1 X_{t-1} + \cdots + b_p X_{t-p})^2, \tag{6.38}$$

which is called a *quasi-Gaussian log likelihood*. The *quasi-Gaussian maximum likelihood estimator* (QGMLE) $\hat{\boldsymbol{\theta}}_{\mathrm{QGML}}$ of $\boldsymbol{\theta}$ is defined by

$$\hat{\boldsymbol{\theta}}_{\mathrm{QGML}} = \arg\sup_{\boldsymbol{\theta}} l_n^Q(\boldsymbol{\theta}), \tag{6.39}$$

($\hat{\boldsymbol{\theta}}_{\mathrm{QGML}}$ is also called a conditional MLE in other references). This is given as follows. Let $\mathbf{b} = (b_1, \ldots, b_p)'$ and $\widetilde{\mathbf{X}}_{t-1} = (X_{t-1}, \ldots, X_{t-p})'$. Then the

solutions $\mathbf{b} = \hat{\mathbf{b}}_{\text{QGML}}$ and $\sigma^2 = \hat{\sigma}^2_{\text{QGML}}$ of the equations

$$\frac{\partial l_n^Q(\boldsymbol{\theta})}{\partial \mathbf{b}} = -\frac{1}{\sigma^2}\sum_{t=p+1}^{n} \widetilde{\mathbf{X}}_{t-1}(X_t + b_1 X_{t-1} + \cdots + b_p X_{t-p}) = \mathbf{0}, \tag{6.40}$$

$$\frac{\partial l_n^Q(\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{n-p}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=p+1}^{n}(X_t + b_1 X_{t-1} + \cdots + b_p X_{t-p})^2 = 0, \tag{6.41}$$

become their QGMLE. Write

$$\hat{\Gamma}_p \equiv \frac{1}{n-p}\sum_{t=p+1}^{n} \widetilde{\mathbf{X}}_{t-1}\widetilde{\mathbf{X}}'_{t-1},$$

$$\hat{\boldsymbol{\gamma}}_p \equiv \frac{1}{n-p}\sum_{t=p+1}^{n} \widetilde{\mathbf{X}}_{t-1} X_t.$$

Then (6.40) and (6.41) are expressed as

$$\hat{\Gamma}_p \hat{\mathbf{b}}_{\text{QGML}} = -\hat{\boldsymbol{\gamma}}_p, \tag{6.42}$$

$$\hat{\sigma}^2_{\text{QGML}} = \frac{1}{n-p}\sum_{t=p+1}^{n}(X_t + \hat{\mathbf{b}}'_{\text{QGML}}\widetilde{\mathbf{X}}_{t-1})^2. \tag{6.43}$$

Since $\{X_t\}$ is a strictly stationary and ergodic process, Theorem 5.7 implies that

$$\hat{\Gamma}_p \xrightarrow{\text{a.s.}} \text{E}(\widetilde{\mathbf{X}}_{t-1}\widetilde{\mathbf{X}}'_{t-1}) = \Gamma_p \text{ (say)}, \tag{6.44}$$

$$\hat{\boldsymbol{\gamma}}_p \xrightarrow{\text{a.s.}} \text{E}(\widetilde{\mathbf{X}}_{t-1} X_t) = \boldsymbol{\gamma}_p \text{ (say)}. \tag{6.45}$$

Here Γ_p is a regular matrix (Exercise 6.3). Writing (6.33) as $X_t = -\widetilde{\mathbf{X}}'_{t-1}\mathbf{b} + u_t$, we have

$$\widetilde{\mathbf{X}}_{t-1} X_t = -\widetilde{\mathbf{X}}_{t-1}\widetilde{\mathbf{X}}'_{t-1}\mathbf{b} + \widetilde{\mathbf{X}}_{t-1} u_t,$$

whose expectation leads to $\boldsymbol{\gamma}_p = -\Gamma_p \mathbf{b}$. For sufficiently large n, we get $\hat{\mathbf{b}}_{\text{QGML}} = -\hat{\Gamma}_p^{-1}\hat{\boldsymbol{\gamma}}_p$. From (6.44) and (6.45) it follows that

$$\hat{\mathbf{b}}_{\text{QGML}} \xrightarrow{\text{a.s.}} \mathbf{b}. \tag{6.46}$$

On the other hand,

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{QGML}} &= \frac{1}{n-p}\sum_{t=p+1}^{n}\{u_t + (\hat{\mathbf{b}}_{\mathrm{QGML}} - \mathbf{b})'\widetilde{\mathbf{X}}_{t-1}\}^2 \\ &= \frac{1}{n-p}\sum_{t=p+1}^{n} u_t^2 + 2(\hat{\mathbf{b}}_{\mathrm{QGML}} - \mathbf{b})'\frac{1}{n-p}\sum_{t=p+1}^{n} u_t\widetilde{\mathbf{X}}_{t-1} \\ &\quad + (\hat{\mathbf{b}}_{\mathrm{QGML}} - \mathbf{b})'\left\{\frac{1}{n-p}\sum_{t=p+1}^{n}\widetilde{\mathbf{X}}_{t-1}\widetilde{\mathbf{X}}'_{t-1}\right\}(\hat{\mathbf{b}}_{\mathrm{QGML}} - \mathbf{b}) \\ &= (\mathrm{A}) + (\mathrm{B}) + (\mathrm{C}),\ (\text{say}). \end{aligned} \tag{6.47}$$

From Theorem 5.7 we can see that (A) $\xrightarrow{\text{a.s.}} \sigma^2$ and

$$\frac{1}{n-p}\sum_{t=p+1}^{n} u_t\widetilde{\mathbf{X}}_{t-1} \xrightarrow{\text{a.s.}} 0, \quad \frac{1}{n-p}\sum_{t=p+1}^{n}\widetilde{\mathbf{X}}_{t-1}\widetilde{\mathbf{X}}'_{t-1} \xrightarrow{\text{a.s.}} \Gamma_p, \tag{6.48}$$

hence, together with (6.46), (B) $\xrightarrow{\text{a.s.}} 0$ and (C) $\xrightarrow{\text{a.s.}} 0$. Therefore we have $\hat{\sigma}^2_{\mathrm{QGML}} \xrightarrow{\text{a.s.}} \sigma^2$, and

$$\begin{Bmatrix}\sqrt{n}(\hat{\mathbf{b}}_{\mathrm{QGML}} - \mathbf{b}) \\ \sqrt{n}(\hat{\sigma}^2_{\mathrm{QGML}} - \sigma^2)\end{Bmatrix} = \begin{Bmatrix}-\Gamma_p^{-1}\dfrac{1}{\sqrt{n-p}}\displaystyle\sum_{t=p+1}^{n}\widetilde{\mathbf{X}}_{t-1}u_t \\ \dfrac{1}{\sqrt{n-p}}\displaystyle\sum_{t=p+1}^{n}(u_t^2 - \sigma^2)\end{Bmatrix} + o(1),\ \text{a.s.} \tag{6.49}$$

Applying the Cramér-Wold device (Theorem A.5) and Theorem 5.9 to the right-hand side of (6.49) we obtain the following theorem.

Theorem 6.3 *For the AR(p) process given by (6.33), we have the following;*

$$(i)\quad \hat{\mathbf{b}}_{QGML} \xrightarrow{a.s.} \mathbf{b}, \quad \hat{\sigma}^2_{QGML} \xrightarrow{a.s.} \sigma^2, \tag{6.50}$$

$$(ii)\quad \begin{bmatrix}\sqrt{n}(\hat{\mathbf{b}}_{QGML} - \mathbf{b}) \\ \sqrt{n}(\hat{\sigma}^2_{QGML} - \sigma^2)\end{bmatrix} \xrightarrow{d} N\left[\mathbf{0}, \begin{pmatrix}\sigma^2\Gamma_p^{-1} & \mathbf{0} \\ \mathbf{0} & 2\sigma^2\end{pmatrix}\right]. \tag{6.51}$$

For our AR(p) process the quasi-Gaussian log-likelihood is expressed in the explicit form of (6.38). However, in the case of ARMA processes and the other processes, it is generally difficult to express their likelihoods in explicit forms like this. Therefore, let us express their approximate likelihoods in terms of spectral densities. As we saw in Section 6.1, the spectral density of (6.33) is

$$f^{\mathrm{AR}}_{\boldsymbol{\theta}}(\lambda) = \frac{\sigma^2}{2\pi}\left|\sum_{j=0}^{p} b_j e^{ij\lambda}\right|^{-2}. \tag{6.52}$$

Consider expressing (6.38) by use of (6.52) and the periodogram

$$I_n(\lambda) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} X_t e^{it\lambda}\right|^2. \tag{6.53}$$

First, we observe that

$$\begin{aligned}
&\int_{-\pi}^{\pi} f_{\boldsymbol{\theta}}^{\mathrm{AR}}(\lambda)^{-1} I_n(\lambda) d\lambda \\
&= \frac{2\pi}{\sigma^2 n}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p} b_{j_1} b_{j_2} \sum_{t=1+\max\{j_1,j_2\}}^{n+\min\{j_1,j_2\}} X_{t-j_1} X_{t-j_2} \\
&= \frac{2\pi}{\sigma^2 n}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p} b_{j_1} b_{j_2} \sum_{t=p+1}^{n} X_{t-j_1} X_{t-j_2} + O(n^{-1}) \quad \text{a.e.}, \\
&= \frac{2\pi}{\sigma^2}\frac{1}{n}\sum_{t=p+1}^{n}\left(\sum_{j=0}^{p} b_j X_{t-j}\right)^2 + O(n^{-1}) \quad \text{a.e.},
\end{aligned} \tag{6.54}$$

(see Exercise 6.5). Since the model satisfies Assumption 6.1, we can factorize (6.52) as follows:

$$f_{\boldsymbol{\theta}}^{\mathrm{AR}}(\lambda) = \frac{\sigma^2}{2\pi}\left[(1 - z_1 e^{i\lambda})(\overline{1 - z_1 e^{i\lambda}}) \cdots (1 - z_p e^{i\lambda})(\overline{1 - z_p e^{i\lambda}})\right]^{-1}, \tag{6.55}$$

where $|z_j| < 1,\ j = 1, \ldots, p$. Then,

$$\begin{aligned}
&\int_{-\pi}^{\pi} \log f_{\boldsymbol{\theta}}^{\mathrm{AR}}(\lambda) d\lambda \\
&= \int_{-\pi}^{\pi}\left[-\sum_{j=1}^{p}\{\log(1 - z_j e^{i\lambda}) + \log(1 - \bar{z_j} e^{-i\lambda})\} + \log\left(\frac{\sigma^2}{2\pi}\right)\right] d\lambda \\
&= -\sum_{j=1}^{p}\left[\int_{-\pi}^{\pi}\left\{\sum_{k=1}^{\infty}\frac{z_j^k e^{ik\lambda}}{k} + \sum_{k=1}^{\infty}\frac{\bar{z_j}^k e^{-ik\lambda}}{k}\right\} d\lambda\right] + 2\pi \log\frac{\sigma^2}{2\pi} \\
&= 2\pi \log\frac{\sigma^2}{2\pi}.
\end{aligned} \tag{6.56}$$

From (6.38), (6.54) and (6.56), it is seen that

$$l_n^Q(\boldsymbol{\theta}) = -\frac{n}{4\pi}\int_{-\pi}^{\pi}\left\{\log f_{\boldsymbol{\theta}}^{\mathrm{AR}}(\lambda) + \frac{I_n(\lambda)}{f_{\boldsymbol{\theta}}^{\mathrm{AR}}(\lambda)}\right\} d\lambda - \frac{n}{2}\log 2\pi + O(1) \quad \text{a.e.} \tag{6.57}$$

Therefore, seeking QGMLE of $\boldsymbol{\theta}$ is asymptotically equivalent to finding the value which minimizes

$$\int_{-\pi}^{\pi}\left\{\log f_{\boldsymbol{\theta}}^{\mathrm{AR}}(\lambda) + \frac{I_n(\lambda)}{f_{\boldsymbol{\theta}}^{\mathrm{AR}}(\lambda)}\right\} d\lambda \tag{6.58}$$

with respect to $\boldsymbol{\theta}$. Although the relation (6.57) was derived for AR process,

it holds for more general Gaussian stationary processes including ARMA. In fact, suppose that $\{X_t\}$ is a Gaussian stationary process with covariance function $R(\cdot)$ and spectral density $f_{\boldsymbol{\theta}}(\lambda)$, $\boldsymbol{\theta} \in \Theta$, and satisfies

(A·1) There exist positive numbers M_1 and M_2 such that

$$0 < M_1 \leq f_{\boldsymbol{\theta}}(\lambda) \leq M_2 < \infty,$$

(A·2)

$$\sum_{t=1}^{\infty} t|R(t)|^2 < \infty.$$

Then the relation (6.57) holds true, i.e., the quantity

$$D(f_{\boldsymbol{\theta}}, I_n) = \int_{-\pi}^{\pi} \left\{ \log f_{\boldsymbol{\theta}}(\lambda) + \frac{I_n(\lambda)}{f_{\boldsymbol{\theta}}(\lambda)} \right\} d\lambda \tag{6.59}$$

is the main order term of $(-4\pi/n) \times$ (log-likelihood) (e.g., Dzhaparidze (1986), Taniguchi and Kakizawa (2000)). Thus we call

$$\hat{\boldsymbol{\theta}}_{\mathrm{QGML}} \equiv \arg\min_{\boldsymbol{\theta} \in \Theta} D(f_{\boldsymbol{\theta}}, I_n), \tag{6.60}$$

a *quasi-Gaussian maximum likelihood* (QGML) estimator of $\boldsymbol{\theta}$. Here, for a given spectral density $g = g(\lambda)$, it may be noted that $D(f_{\boldsymbol{\theta}}, g) \geq D(g, g)$ where the equality holds if and only if $f_{\boldsymbol{\theta}}(\lambda) = g(\lambda)$ a.e. Hence $D(\ ,\ )$ is a sort of disparity measure. Assuming some appropriate regularity conditions (e.g., smoothness of $f_{\boldsymbol{\theta}}(\lambda)$ with respect to $\boldsymbol{\theta}$), and that $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$, we can show that

$$\text{(i)} \quad \hat{\boldsymbol{\theta}}_{\mathrm{QGML}} \xrightarrow{\text{a.s.}} \boldsymbol{\theta}_0, \tag{6.61}$$

$$\text{(ii)} \quad \sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{QGML}} - \boldsymbol{\theta}_0) \xrightarrow{\text{d}} \mathrm{N}(\mathbf{0}, \mathcal{F}(\boldsymbol{\theta}_0)^{-1}), \tag{6.62}$$

where

$$\mathcal{F}(\boldsymbol{\theta}) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\lambda) \frac{\partial}{\partial \boldsymbol{\theta}'} \log f_{\boldsymbol{\theta}}(\lambda) d\lambda, \tag{6.63}$$

which is called the *Fisher information matrix*.

We do not give the proof of (i) and (ii) because we proceed to more general discussion. Let $\{\mathbf{X}_t = (X_{1t}, \ldots, X_{mt})'\}$ be generated by

$$\mathbf{X}_t = \sum_{j=0}^{\infty} A(j)\mathbf{U}_{t-j}, \tag{6.64}$$

where $\{\mathbf{U}_t = (U_{1t}, \ldots, U_{mt})'\} \sim$ i.i.d. $(\mathbf{0}, V)$, and $\mathbf{U}_t$'s have the fourth-order cumulant. Here $A(j)$'s are $m \times m$ matrices whose elements are assumed to be summable. Then $\{\mathbf{X}_t\}$ is strictly stationary and ergodic with spectral density

matrix

$$\mathbf{f}(\lambda) = \frac{1}{2\pi}\left\{\sum_{j=0}^{\infty} A(j)e^{ij\lambda}\right\} V \left\{\sum_{j=0}^{\infty} A(j)e^{ij\lambda}\right\}^{*}. \tag{6.65}$$

Write

$$\mathbf{I}_n(\lambda) = \frac{1}{2\pi n}\left\{\sum_{t=1}^{n} \mathbf{X}_t e^{it\lambda}\right\}\left\{\sum_{t=1}^{n} \mathbf{X}_t e^{it\lambda}\right\}^{*}. \tag{6.66}$$

Lemma 6.1 *Let $\phi_j(\lambda)$, $j = 1, \ldots, r$, be $m \times m$ matrix-valued continuous functions on $[-\pi, \pi]$ such that $\phi_j(\lambda) = \phi_j(\lambda)^*$ and $\phi_j(-\lambda) = \phi_j(\lambda)'$. Then*
(i) for each $j = 1, \ldots, r$,

$$\int_{-\pi}^{\pi} \mathrm{tr}\{\phi_{\mathrm{j}}(\lambda)\mathbf{I}_{\mathrm{n}}(\lambda)\}\mathrm{d}\lambda \xrightarrow{a.s.} \int_{-\pi}^{\pi} \mathrm{tr}\{\phi_{\mathrm{j}}(\lambda)\mathbf{f}(\lambda)\}\mathrm{d}\lambda, \text{ as } \mathrm{n} \to \infty. \tag{6.67}$$

(ii) the quantities

$$\sqrt{n}\int_{-\pi}^{\pi} \mathrm{tr}\left[\phi_{\mathrm{j}}(\lambda)\{\mathbf{I}_{\mathrm{n}}(\lambda) - \mathbf{f}(\lambda)\}\right]\mathrm{d}\lambda, \ \mathrm{j} = 1, \ldots, \mathrm{r},$$

have, asymptotically, a normal distribution with zero mean vector and covariance matrix W whose (j, l)th element is

$$4\pi\int_{-\pi}^{\pi} \mathrm{tr}\{\mathbf{f}(\lambda)\phi_{\mathrm{j}}(\lambda)\mathbf{f}(\lambda)\phi_{\mathrm{l}}(\lambda)\}\mathrm{d}\lambda + 2\pi\sum_{q,s,u,v=1}^{m}\iint_{-\pi}^{\pi}\phi_{qs}^{(j)}(\lambda_1)\phi_{uv}^{(l)}(\lambda_2)Q_{qsuv}^{X}(-\lambda_1, \lambda_2, -\lambda_2)d\lambda_1 d\lambda_2,$$

where $\phi_{qs}^{(j)}(\lambda_1)$ is the (q, s)th element of $\phi_j(\lambda)$, and

$$Q_{qsuv}^{X}(\lambda_1, \lambda_2, \lambda_3) = (2\pi)^{-3}\sum_{t_1,t_2,t_3=-\infty}^{\infty}\exp\{-i(\lambda_1 t_1 + \lambda_2 t_2 + \lambda_3 t_3)\}c_{qsuv}^{X}(t_1, t_2, t_3),$$

with $c_{qsuv}^{X}(t_1, t_2, t_3) \equiv cum\{X_{q0}, X_{st_1}, X_{ut_2}, X_{vt_3}\}$.

SKETCH OF PROOF

Because the rigorous proof is very technical and complicated, we just provide an outline of the proof following Hannan and Robinson (1973).
(i) Since $\phi_j(\lambda)$ is continuous, we can introduce the Cesaro sum

$$\phi_j^{(M)}(\lambda) = \sum_{|k|\le M} a_j(k)\left(1 - \frac{|k|}{M}\right)\exp(-ik\lambda)$$

of the Fourier series of $\phi_j(\lambda)$, where $\sup_{\lambda}\left|\phi_j(\lambda) - \phi_j^{(M)}(\lambda)\right| < \epsilon$ for any $\epsilon > 0$ and M sufficiently large. Then it is seen that

$$\left|\int_{-\pi}^{\pi} \mathrm{tr}\{\phi_j(\lambda)\mathbf{I}_n(\lambda)\}d\lambda - \int_{-\pi}^{\pi} \mathrm{tr}\{\phi_j^{(M)}(\lambda)\mathbf{I}_n(\lambda)\}d\lambda\right| = O(\epsilon) \text{ a.s.} \tag{6.68}$$

and $\int_{-\pi}^{\pi} \mathrm{tr}\{\phi_j^{(M)}(\lambda)\mathbf{I}_n(\lambda)\}d\lambda$ is a linear combination of

$$\frac{1}{n}\sum_{t=1}^{n-|l|} X_{a,t}X_{b,t+|l|}, \quad l = 0, 1, \ldots, M, \ a, b = 1, \ldots, m. \tag{6.69}$$

From Theorem 5.7, we observe that

$$\frac{1}{n}\sum_{t=1}^{n-|l|} X_{a,t}X_{b,t+|l|} \xrightarrow{\text{a.s.}} \gamma_{ab}(l) \quad \text{for each} \ \ l, a, b, \tag{6.70}$$

as $n \to \infty$. Thus $\int_{-\pi}^{\pi} \mathrm{tr}\{\phi_j^{(M)}(\lambda)\mathbf{I}_n(\lambda)\}d\lambda$ converges almost surely to a value which can be arbitrarily close to $\int_{-\pi}^{\pi} \mathrm{tr}\{\phi_j(\lambda)\mathbf{f}(\lambda)\}d\lambda$ if M is sufficiently large. Hence, (6.67) is proved.

(ii) As in (i) if we replace $\phi_j(\lambda)$ by the Cesaro sum $\phi_j^{(M)}(\lambda)$, we can see that $\sqrt{n}\int_{-\pi}^{\pi} \mathrm{tr}\left[\phi_j(\lambda)\{\mathbf{I}_n(\lambda) - \mathbf{f}(\lambda)\}\right] d\lambda$ is approximated by a linear combination of

$$\sqrt{n}\left\{\frac{1}{n}\sum_{t=1}^{n-|l|} X_{a,t}X_{b,t+|l|} - \gamma_{ab}(l)\right\}, \tag{6.71}$$

where $l = 0, 1, \ldots, M, \ a, b = 1, \ldots, m$. Recalling (6.64) it is seen that (6.71)'s are approximated by linear combinations of

$$\sqrt{n}\left\{\frac{1}{n}\sum_{t=1}^{n} u_{c,t}u_{d,t+l'} - \delta(l', 0)V_{cd}\right\}, \quad l' \le l, \tag{6.72}$$

where $c, d = 1, \ldots, m$. The asymptotic normality of (6.72) follows from Theorem 5.9. The above approximations can be taken to be arbitrarily close to the corresponding quantities. Hence, we can grasp the main line of (ii). □

Next, suppose that $\mathbf{f}(\lambda)$ is parameterized as $\mathbf{f}_{\boldsymbol{\theta}}(\lambda)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)' \in \Theta$, where Θ is a compact set of $\mathbf{R}^r$. We are now interested in estimation of $\boldsymbol{\theta}$ based on $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$. For this we use the following version of the quasi-Gaussian likelihood (6.59):

$$\widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{I}_n) \equiv \int_{-\pi}^{\pi} \left[\log\det\{\mathbf{f}_{\boldsymbol{\theta}}(\lambda)\} + \mathrm{tr}\{\mathbf{I}_n(\lambda)\mathbf{f}_{\boldsymbol{\theta}}(\lambda)^{-1}\}\right] d\lambda. \tag{6.73}$$

Here it should be noted that $\widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{I}_n)$ is not an approximation for the log likelihood generally, because we do not assume Gaussianity of $\{\mathbf{X}_t\}$. Even so, we can use the QGMLE $\widetilde{\boldsymbol{\theta}} \equiv \arg\min_{\boldsymbol{\theta}\in\Theta} \widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{I}_n)$ as estimator of $\boldsymbol{\theta}$. In what follows we describe the asymptotics of $\widetilde{\boldsymbol{\theta}}$.

Assumption 6.3 *(1) The true value* $\boldsymbol{\theta}_0 \in \mathrm{lnt}\ \Theta$.

(2) $\widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{f}_{\boldsymbol{\theta}_0}) \geq \widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}_0}, \mathbf{f}_{\boldsymbol{\theta}_0})$ *for all* $\boldsymbol{\theta} \in \Theta$*, and the equality holds if and only if* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$*.*
(3) $\mathbf{f}_{\boldsymbol{\theta}}(\lambda)$ *is continuously twice differentiable with respect to* $\boldsymbol{\theta}$*.*
(4) The $r \times r$ *matrix*

$$\Gamma(\boldsymbol{\theta}_0) = \Big[\frac{1}{4\pi}\int_{-\pi}^{\pi} \mathrm{tr}\left\{\mathbf{f}_{\boldsymbol{\theta}_0}(\lambda)^{-1}\left(\frac{\partial}{\partial\theta_{\mathrm{j}}}\mathbf{f}_{\boldsymbol{\theta}_0}(\lambda)\right)\mathbf{f}_{\boldsymbol{\theta}_0}(\lambda)^{-1}\left(\frac{\partial}{\partial\theta_{\mathrm{l}}}\mathbf{f}_{\boldsymbol{\theta}_0}(\lambda)\right)\right\}\mathrm{d}\lambda\ ; \quad j,l = 1,\ldots,r\Big]$$

is nonsingular.

We now get the following results for $\widetilde{\boldsymbol{\theta}}$ (c.f., Hosoya and Taniguchi (1982)).

Theorem 6.4 *Under Assumption 6.3, the following assertions are true.*

(i) $$\widetilde{\boldsymbol{\theta}}_{QGML} \xrightarrow{a.s.} \boldsymbol{\theta}_0 \quad as\ n \to \infty. \tag{6.74}$$

(ii) The distribution of $\sqrt{n}(\widetilde{\boldsymbol{\theta}}_{QGML} - \boldsymbol{\theta}_0)$ *tends to the normal distribution with zero mean vector and covariance matrix* $\Gamma(\boldsymbol{\theta}_0)^{-1} + \Gamma(\boldsymbol{\theta}_0)^{-1}\Pi(\boldsymbol{\theta}_0)\Gamma(\boldsymbol{\theta}_0)^{-1}$*, where* $\Pi(\boldsymbol{\theta}) = \{\Pi_{jl}(\boldsymbol{\theta})\}$ *is an* $r \times r$ *matrix such that*

$$\Pi_{jl}(\boldsymbol{\theta}) = \frac{1}{8\pi}\sum_{q,s,u,v=1}^{m}\iint_{-\pi}^{\pi}\frac{\partial}{\partial\theta_j}\mathbf{f}_{\boldsymbol{\theta}}^{(q,s)}(\lambda_1)\frac{\partial}{\partial\theta_l}\mathbf{f}_{\boldsymbol{\theta}}^{(u,v)}(\lambda_2) \times Q^X_{qsuv}(-\lambda_1,\lambda_2,-\lambda_2)d\lambda_1 d\lambda_2.$$

Here $\mathbf{f}_{\boldsymbol{\theta}}^{(q,s)}(\lambda)$ *is the* (q,s)*th element of* $\mathbf{f}_{\boldsymbol{\theta}}(\lambda)^{-1}$*.*

PROOF

From (i) of Lemma 6.1 it follows that

$$\widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{I}_n) \xrightarrow{\text{a.s.}} \widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{f}_{\boldsymbol{\theta}_0}), \quad (n \to \infty). \tag{6.75}$$

For any given $\epsilon > 0$, if n is sufficiently large, because of (6.75) and $\widetilde{\boldsymbol{\theta}}_{\mathrm{QGML}}$ minimizes $\widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{I}_n)$, we have, almost surely,

$$\widetilde{D}(\mathbf{f}_{\widetilde{\boldsymbol{\theta}}_{\mathrm{QGML}}}, \mathbf{f}_{\boldsymbol{\theta}_0}) - \epsilon \leq \widetilde{D}(\mathbf{f}_{\widetilde{\boldsymbol{\theta}}_{\mathrm{QGML}}}, \mathbf{I}_n) \leq \widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}_0}, \mathbf{I}_n) \leq \widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}_0}, \mathbf{f}_{\boldsymbol{\theta}_0}) + \epsilon. \tag{6.76}$$

In view of (2) in Assumption 6.3, $\widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{f}_{\boldsymbol{\theta}_0})$ is continuous with respect to $\boldsymbol{\theta}$, and has a unique minimum at $\boldsymbol{\theta}_0$. The assertion (i) follows from (6.76). Next we prove (ii). For sufficiently large n, expanding $(\partial/\partial\boldsymbol{\theta})\widetilde{D}(\mathbf{f}_{\widetilde{\boldsymbol{\theta}}_{\mathrm{QGML}}}, \mathbf{I}_n)$ around $\boldsymbol{\theta}_0$ we obtain

$$\mathbf{0} = \frac{\partial}{\partial\boldsymbol{\theta}}\widetilde{D}(\mathbf{f}_{\widetilde{\boldsymbol{\theta}}_{\mathrm{QGML}}}, \mathbf{I}_n) = \frac{\partial}{\partial\boldsymbol{\theta}}\widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}_0}, \mathbf{I}_n) + \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\widetilde{D}(\mathbf{f}_{\overline{\boldsymbol{\theta}}}, \mathbf{I}_n)(\widetilde{\boldsymbol{\theta}}_{\mathrm{QGML}} - \boldsymbol{\theta}_0), \tag{6.77}$$

where $\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \leq \|\widetilde{\boldsymbol{\theta}}_{\mathrm{QGML}} - \boldsymbol{\theta}_0\|$. From (i) of Theorem 6.4 and (i) of Lemma 6.1 it is not difficult to see that

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\widetilde{D}(\mathbf{f}_{\overline{\boldsymbol{\theta}}}, \mathbf{I}_n) \xrightarrow{\text{a.s.}} 4\pi\Gamma(\boldsymbol{\theta}_0). \tag{6.78}$$

Noting that $(\partial/\partial\boldsymbol{\theta})\widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{f}_{\boldsymbol{\theta}_0})|_{\boldsymbol{\theta}_0} = \mathbf{0}$, (6.77) and (6.78), we can see that

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_{\mathrm{QGML}} - \boldsymbol{\theta}_0) = -\Gamma(\boldsymbol{\theta}_0)^{-1}\frac{\sqrt{n}}{4\pi}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}_0}, \mathbf{I}_n) - \frac{\partial}{\partial\boldsymbol{\theta}}\widetilde{D}(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{f}_{\boldsymbol{\theta}_0})|_{\boldsymbol{\theta}_0}\right] + o_p(1). \tag{6.79}$$

Then the assertion follows from application of (ii) of Lemma 6.1 to the right-hand side of (6.79). □

Remark 6.1 *If $\{\mathbf{X}_t\}$ is Gaussian, then (ii) of Theorem 6.4 becomes that*

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_{QGML} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \Gamma(\boldsymbol{\theta}_0)^{-1}). \tag{6.80}$$

Since $\Gamma(\boldsymbol{\theta}_0)$ is the Fisher information matrix in time series, we may say that $\widetilde{\boldsymbol{\theta}}_{QGML}$ is Gaussian asymptotically efficient *in the sense of (6.80). Later we will discuss the rigorous asymptotic efficiency which rests on the LAN results.*

The asymptotic variance of $\widetilde{\boldsymbol{\theta}}_{\mathrm{QGML}}$ in Theorem 6.4 (ii) contains the integrals of the fourth-order cumulant spectral densities $Q^X_{qsuv}(\cdot,\cdot,)$ which represent a degree of non-Gaussianity. Next we investigate when these integrals vanish. Since $\{\mathbf{U}_t\} \sim$ i.i.d. with fourth-order cumulant, the fourth-order cumulant of U_{at_1}, U_{bt_2}, U_{ct_3}, U_{dt_4}, satisfies

$$\mathrm{cum}\{U_{at_1}, U_{bt_2}, U_{ct_3}, U_{dt_4}\} = \begin{cases} \kappa_{abcd} & \text{if } t_1 = t_2 = t_3 = t_4, \\ 0 & \text{otherwise.} \end{cases} \tag{6.81}$$

Recall the spectral density matrix (6.65) of the process (6.64). Suppose that (6.65) is parameterized as

$$\mathbf{f}_{\boldsymbol{\theta}}(\lambda) = \frac{1}{2\pi}\left\{\sum_{j=0}^{\infty} A_{\boldsymbol{\theta}}(j)e^{ij\lambda}\right\} V \left\{\sum_{j=0}^{\infty} A_{\boldsymbol{\theta}}(j)e^{ij\lambda}\right\}^*, \tag{6.82}$$

with $A_{\boldsymbol{\theta}}(0) = \mathbf{I}_m$. If V is independent of $\boldsymbol{\theta}$, then we say that $\boldsymbol{\theta}$ is *innovation-free*. In what follows, we assume that $\det\left\{\sum_{j=0}^{\infty} A_{\boldsymbol{\theta}}(j)z^j\right\} \neq 0$ for all $|z| \leq 1$, and that $\boldsymbol{\theta}$ is innovation-free. Write $A_{\boldsymbol{\theta}}(\lambda) = \sum_{j=0}^{\infty} A_{\boldsymbol{\theta}}(j)e^{ij\lambda}$. Then the matrix $\Pi(\boldsymbol{\theta}_0)$ in Theorem 6.4 can be expressed as

$$\frac{1}{16\pi^2}\sum_{a,b,c,d=1}^{m}\kappa_{abcd}\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} A_{\boldsymbol{\theta}_0}(\lambda)^*\left\{\frac{\partial}{\partial\theta_j}\mathbf{f}_{\boldsymbol{\theta}_0}(\lambda)^{-1}\right\}A_{\boldsymbol{\theta}_0}(\lambda)d\lambda\right]_{ab} \times \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} A_{\boldsymbol{\theta}_0}(\lambda)^*\left\{\frac{\partial}{\partial\theta_l}\mathbf{f}_{\boldsymbol{\theta}_0}(\lambda)^{-1}\right\}A_{\boldsymbol{\theta}_0}(\lambda)d\lambda\right]_{cd}, \tag{6.83}$$

where $[\quad]_{ab}$ denotes the (a,b)th element of the matrix in the bracket (c.f. Dunsmuir (1979)). It is easy to see that

$$\begin{aligned}
&\int_{-\pi}^{\pi} A_{\boldsymbol{\theta}_0}(\lambda)^* \left\{ \frac{\partial}{\partial \theta_j} \mathbf{f}_{\boldsymbol{\theta}_0}(\lambda)^{-1} \right\} A_{\boldsymbol{\theta}_0}(\lambda) d\lambda \\
&= -\int_{-\pi}^{\pi} \left\{ \frac{\partial}{\partial \theta_j} A_{\boldsymbol{\theta}_0}(\lambda)^* \right\} \{A_{\boldsymbol{\theta}_0}(\lambda)^*\}^{-1} V^{-1} d\lambda \\
&- \int_{-\pi}^{\pi} V^{-1} \{A_{\boldsymbol{\theta}_0}(\lambda)\}^{-1} \left\{ \frac{\partial}{\partial \theta_j} A_{\boldsymbol{\theta}_0}(\lambda) \right\} d\lambda = \mathbf{0}, \qquad (6.84)
\end{aligned}$$

which together with (6.83) implies that the asymptotic distribution of $\widetilde{\boldsymbol{\theta}}_{\mathrm{QGML}}$ is then independent of the fourth-order cumulants κ_{abcd} (i.e., non-Gaussianity of the process). Henceforth we say that an estimator of $\boldsymbol{\theta}$ is *non-Gaussian robust* if the asymptotic distribution is independent of non-Gaussianity of the process.

Here we summarize the above. Although $\widetilde{\boldsymbol{\theta}}_{\mathrm{QGML}}$ is defined by an approximated Gaussian likelihood, we can use it for non-Gaussian processes. Then we observe that, even if the process concerned is non-Gaussian, $\widetilde{\boldsymbol{\theta}}_{\mathrm{QGML}}$ has good properties, i.e., Gaussian efficiency and non-Gaussian robustness if $\boldsymbol{\theta}_0$ is innovation-free.

Next, let us discuss the problem of inference for nonlinear time series models. Consider the following ARCH(q) model:

$$\begin{cases} X_t = u_t \sqrt{h_t}, \\ h_t = a_0 + \sum_{j=1}^{q} a_j X_{t-j}^2, \end{cases} \qquad (6.85)$$

where $a_0 > 0,\ a_j \geq 0,\ j = 1, \ldots, q$, and $\{u_t\} \sim$ i.i.d. $(0,1)$ with probability density function $g(u)$. As candidates for $g(u)$, we often use the following:
(N) Standard normal distribution

$$g(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right),$$

(T) T-distribution with ν degrees of freedom

$$g(u) = \frac{\Gamma((\nu+1)/2)}{(\pi\nu)^{1/2}\Gamma(\nu/2)} \left(\frac{\nu}{\nu-2}\right)^{1/2} \left(1 + \frac{u^2}{\nu-2}\right)^{-\frac{\nu+1}{2}},$$

(GG) Generalized Gaussian distribution

$$g(u) = \nu \left\{ \lambda 2^{1+\frac{1}{\nu}} \Gamma\left(\frac{1}{\nu}\right) \right\}^{-1} \exp\left\{ -\frac{1}{2} \left| \frac{u}{\lambda} \right|^{\nu} \right\}, \qquad (6.86)$$

where $\lambda = \{2^{-2/\nu}\Gamma(1/\nu)/\Gamma(3/\nu)\}^{1/2}$ with $0 < \nu < 2$. The generalized Gaussian distribution contains the following double exponential distribution as a special case:

$$g(u) = \frac{1}{\sqrt{2}} \exp\left\{ -\sqrt{2}|u| \right\}.$$

The distributions (T) and (GG) have fatter tails than that for (N).

Let us consider the maximum likelihood estimator of unknown parameter $\boldsymbol{\theta} = (a_0, \ldots, a_q)'$. As in the case of AR models, it is difficult to deal with the likelihood of $X_1, \ldots, X_n$. Hence we think of the conditional likelihood given $(X_1, \ldots, X_q)$. We call this a quasi-likelihood, which is given by

$$l_n^Q(\boldsymbol{\theta}) \equiv \sum_{t=q+1}^{n} \left\{ -\frac{1}{2} \log h_t + \log g \left(\frac{X_t}{\sqrt{h_t}} \right) \right\}. \tag{6.87}$$

The quasi-maximum likelihood (QML) estimator of $\boldsymbol{\theta}$ is defined by

$$\hat{\boldsymbol{\theta}}_{\mathrm{QML}} \equiv \arg\sup_{\boldsymbol{\theta}} l_n^Q(\boldsymbol{\theta}). \tag{6.88}$$

Similarly it is possible to estimate parameters of GARCH models. Let $\{X_t\}$ be generated by the GARCH(p,q) model:

$$\begin{cases} X_t = u_t \sqrt{h_t}, \\ h_t = a_0 + a_1 X_{t-1}^2 + \cdots + a_q X_{t-q}^2 + b_1 h_{t-1} + \cdots + b_p h_{t-p}. \end{cases} \tag{6.89}$$

If h_t can be expressed as a linear combination of X_{t-j}^2, the form will be

$$h_t = c_0 + \sum_{j=1}^{\infty} c_j X_{t-j}^2, \tag{6.90}$$

in general. However, we observe the stretch $X_1, \ldots, X_n$ of $\{X_t\}$. In order to calculate (6.87) we have to replace (6.90) by the following feasible expression

$$\tilde{h}_t = c_0 + \sum_{j=1}^{t-1} c_j X_{t-j}^2. \tag{6.91}$$

Then the QML estimator $\tilde{\boldsymbol{\theta}}_{\mathrm{QML}}$ of $\boldsymbol{\theta}$ is given by the value which minimizes

$$\tilde{l}_n^Q(\boldsymbol{\theta}) = \sum_{t=q+1}^{n} \left\{ -\frac{1}{2} \log \tilde{h}_t + \log g \left(\frac{X_t}{\sqrt{\tilde{h}_t}} \right) \right\} \tag{6.92}$$

with respect to $\boldsymbol{\theta}$. This estimation method can be applied to many other models, e.g., EGARCH(p,q), TGARCH(p,q), etc.

So far we have seen estimators of the ML type. In what follows we mention a simple and convenient estimation method. Suppose that $\{\mathbf{X}_t\}$ is an m-vector stochastic process depending on an unknown parameter $\boldsymbol{\theta} \in \Theta \subset \mathbf{R}^q$. Let $\mathcal{F}_t(l)$ be the σ-algebra generated by $\{\mathbf{X}_s : t - l \leq s \leq t - 1\}$ where l is an appropriately chosen positive integer. Letting $m_{\boldsymbol{\theta}}(t, t-1) \equiv \mathrm{E}\{\mathbf{X}_t | \mathcal{F}_t(l)\}$, define

$$\hat{\boldsymbol{\theta}}_{\mathrm{CL}} = \arg\min_{\boldsymbol{\theta} \in \Theta} Q_n^c(\boldsymbol{\theta}), \tag{6.93}$$

where

$$Q_n^c(\boldsymbol{\theta}) = \sum_{t=l+1}^{n} \{\mathbf{X}_t - m_{\boldsymbol{\theta}}(t, t-1)\}' \{\mathbf{X}_t - m_{\boldsymbol{\theta}}(t, t-1)\}. \tag{6.94}$$

This $\hat{\boldsymbol{\theta}}_{\mathrm{CL}}$ is called a *conditional least squares estimator* of $\boldsymbol{\theta}$. Although $\hat{\boldsymbol{\theta}}_{\mathrm{CL}}$ is not asymptotically efficient, it has convenience. To see this, consider the ARCH(q) model

$$X_t = u_t \sqrt{a_0 + \textstyle\sum_{j=1}^{q} a_j X_{t-j}^2}, \quad (\{u_t\} \sim \text{i.i.d. } (0,1)). \tag{6.95}$$

Since $X_t^2 = u_t^2 \{a_0 + \sum_{j=1}^{q} a_j X_{t-j}^2\}$, we regard this X_t^2 as $\mathbf{X_t}$ in (6.94). Then, setting $l = q$ and $\boldsymbol{\theta} = (a_0, \cdots, a_q)'$, we have

$$Q_n^c(\boldsymbol{\theta}) = \sum_{t=q+1}^{n} \left\{ X_t^2 - (a_0 + \sum_{j=1}^{q} a_j X_{t-j}^2) \right\}^2, \tag{6.96}$$

which implies

$$\hat{\boldsymbol{\theta}}_{CL} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}, \tag{6.97}$$

where

$$\mathbf{Y} = (X_{q+1}^2, \cdots, X_n^2)',$$

$$\mathbf{Z} = \begin{pmatrix} 1 & X_q^2 & \cdots & X_1^2 \\ \vdots & & & \\ 1 & X_{t-1}^2 & \cdots & X_{t-q}^2 \\ \vdots & & & \\ 1 & X_{n-1}^2 & \cdots & X_{n-q}^2 \end{pmatrix}.$$

Hence, $\hat{\boldsymbol{\theta}}_{CL}$ in this case has a simple and explicit form (6.97). From (6.97) we can write $\hat{\boldsymbol{\theta}}_{CL} - \boldsymbol{\theta}$ in the form

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{CL} - \boldsymbol{\theta}) = \left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1} \frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{b}, \tag{6.98}$$

and observe that each component of $\mathbf{Z}'\mathbf{b}$ is a martingale. Assume that $a_0 > 0$, $a_1, \cdots, a_q \geq 0$ and $\sum_{j=1}^{q} a_j < 1$, then $\{X_t\}$ is strictly stationary and ergodic. Thus, if the moments of $\{u_t\}$ exist up to necessary order, we can apply Theorems 5.4, 5.7, 5.9 and the Cramér-Wold device to the right-hand side of (6.98). Then it is shown that $n^{-1}\mathbf{Z}'\mathbf{Z}$ converges to a constant matrix almost surely, and that $n^{-1/2}\mathbf{Z}'\mathbf{b}$ converges to a normal distribution, which proves $\hat{\boldsymbol{\theta}}_{CL} \xrightarrow{a.s.} \boldsymbol{\theta}$ and the asymptotic normality of $\sqrt{n}(\hat{\boldsymbol{\theta}}_{CL} - \boldsymbol{\theta})$.

For general m-vector stochastic processes, Tjøstheim (1986) showed the following theorem.

Theorem 6.5 *Assume that* $\{\mathbf{X}_t\}$ *is an* m*-vector strictly stationary ergodic process with* $E\{\| \mathbf{X}_t \|^2\} < \infty$ *and that* $m_{\boldsymbol{\theta}}(t, t-1) = E_{\boldsymbol{\theta}}\{\mathbf{X}_t | \mathcal{F}_t(l)\}$ *is almost surely three times continuously differentiable in an open set* B *containing the true value* $\boldsymbol{\theta}_0$*. Moreover, suppose that the following conditions hold:*
(A1) For $j, k = 1, \cdots, q$

$$E\left\{\left\|\frac{\partial}{\partial\theta_j} m_{\boldsymbol{\theta}_0}(t, t-1)\right\|^2\right\} < \infty,$$

and

$$E\left\{\left\|\frac{\partial^2}{\partial\theta_j \partial\theta_k} m_{\boldsymbol{\theta}_0}(t, t-1)\right\|^2\right\} < \infty.$$

(A2) The vectors $(\partial/\partial\theta_j) m_{\boldsymbol{\theta}_0}(t, t-1)$, $j = 1, \cdots, q$, *are linearly independent in the sense that if* $c_1, \cdots, c_q$ *are arbitrary real numbers such that*

$$E\left\{\left\|\sum_{j=1}^{q} c_j \frac{\partial}{\partial\theta_j} m_{\boldsymbol{\theta}_0}(t, t-1)\right\|^2\right\} = 0,$$

then $c_1 = c_2 = \cdots = c_q = 0$.
(A3) For $\boldsymbol{\theta} \in B$, *there exist functions* $G_{t-1}^{ijk}(\mathbf{X}_1, \cdots, \mathbf{X}_{t-1})$ *and* $H_t^{ijk}(\mathbf{X}_1, \cdots, \mathbf{X}_t)$ *such that*

$$\left|\frac{\partial}{\partial\theta_i} m_{\boldsymbol{\theta}}(t, t-1)' \frac{\partial^2}{\partial\theta_j \partial\theta_k} m_{\boldsymbol{\theta}}(t, t-1)\right| \le G_{t-1}^{ijk}, \quad E(G_{t-1}^{ijk}) < \infty$$

and

$$\left|\{\mathbf{X}_t - m_{\boldsymbol{\theta}}(t, t-1)\}' \frac{\partial^3}{\partial\theta_i \partial\theta_j \partial\theta_k} m_{\boldsymbol{\theta}}(t, t-1)\right| \le H_t^{ijk}, \quad E(H_t^{ijk}) < \infty,$$

for $i, j, k = 1, \cdots, q$.
(A4)

$$\begin{aligned} R &= E\left[\frac{\partial}{\partial\boldsymbol{\theta}} m_{\boldsymbol{\theta}_0}(t, t-1)'\{\mathbf{X}_t - m_{\boldsymbol{\theta}_0}(t, t-1)\}\right. \\ &\times \left.\{\mathbf{X}_t - m_{\boldsymbol{\theta}_0}(t, t-1)\}' \frac{\partial}{\partial\boldsymbol{\theta}} m_{\boldsymbol{\theta}_0}(t, t-1)\right] < \infty. \end{aligned}$$

Then there exists $\hat{\boldsymbol{\theta}}_{CL}$ *such that* $\hat{\boldsymbol{\theta}}_{CL} \xrightarrow{a.s.} \boldsymbol{\theta}_0$. *Furthermore, if there exists an* $l \in \mathbf{N}$ *satisfying* $E_{\boldsymbol{\theta}}\{\mathbf{X}_t | \mathcal{F}_t(l)\} = E_{\boldsymbol{\theta}}(\mathbf{X}_t | \mathcal{F}_t)$, *where* $\mathcal{F}_t$ *is the* σ*-algebra generated by* $\{\mathbf{X}_s : s \le t-1\}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{CL} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, U^{-1} R U^{-1}), \tag{6.99}$$

where

$$U = E\left\{\frac{\partial}{\partial\boldsymbol{\theta}} m_{\boldsymbol{\theta}_0}(t, t-1)' \frac{\partial}{\partial\boldsymbol{\theta}'} m_{\boldsymbol{\theta}_0}(t, t-1)\right\}.$$

Now we turn to discuss optimality of estimation and testing. Lucien LeCam established the most important and sophisticated foundation of the general statistical asymptotic theory. He introduced the concept of *local asymptotic normality* (LAN) for the likelihood ratio of general statistical models. Once LAN is proved, the asymptotic optimality of estimators and tests is described in terms of the LAN property. The LAN approach has been developed for various stochastic processes. First, we give a historical review of the literature on this topic.

Roussas (1972) established a modern and elegant approach for statistical analysis of a Markov process. Using the concept of contiguity and LAN, he studied asymptotic optimality of sequences of estimators and tests. The description is systematic and mathematically rigorous. In (1979) Roussas further generalized the asymptotic estimation theory put forward in his (1972) book, to the case where the process concerned is not necessarily Markovian.

Hallin, Ingenbleek and Puri (1985) introduced a class of linear serial rank statistics for the problem of testing white noise against alternatives of ARMA serial dependence. The asymptotic normality of the proposed statistics was established under both the null and alternative hypotheses using the LAN results. The efficiency properties of the proposed statistics were investigated, and an explicit formulation of the asymptotically most efficient score generating functions was provided. Hallin and Puri (1994) considered ARMA processes with a linear regression trend under unspecified innovation densities, and proved a LAN result involving a rank-measurable central sequence. Then locally asymptotically most stringent aligned rank tests were derived. Furthermore, Garel and Hallin (1995) showed the LAN result for vector ARMA models with a linear trend. In contrast with the result by Hallin and Puri (1994), Garel and Hallin expressed the central sequence in terms of a generalized concept of residual cross-covariance function.

Kreiss (1987) considered the estimation problem for the parameter θ of a stationary ARMA(p, q) process with independent and identically, but not necessary normally, distributed errors. The LAN property for this model was proved. Then he constructed local asymptotically minimax (LAM) estimators, which asymptotically achieve the smallest possible covariance matrix. Next Kreiss (1990) established the LAN property for a class of non-Gaussian autoregressive models with infinite order and obtained a LAM estimator for a finite dimensional subparameter of the model.

Recently much attention has been paid to long-memory processes, which appear in many fields (e.g., hydrology and economics). For a time series regression model with fractional ARIMA disturbances, Hallin, Taniguchi, Serroukh and Choy (1999) established the LAN theorem and the optimal inference. Also Taniguchi and Kakizawa (2000) developed the LAN approach for a class of vector linear processes which exhibit long-memory dependence.

It is also possible to develop the asymptotic theory for nonlinear time series

models by use of the LAN methodology. Based on the LAN property, Benghabrit and Hallin (1992, 1996) constructed locally asymptotic optimal tests against bilinear time series dependence. Jeganathan (1995) gave an extensive review of the LAN approach for various linear and nonlinear time series models. For a class of ARCH(∞) models with stochastic mean, Lee and Taniguchi (2005) proved the LAN property, and discussed asymptotically optimal estimators for unknown parameters when the innovation density is known and when it is unknown.

In Example 6.9 we introduced CHARN models which are very general, and contain various typical financial time series models as special cases. In what follows we mention the LAN approach for CHARN models, and develop the asymptotically optimal estimation and testing theory in line with Kato, Taniguchi and Honda (2006).

Let $\{\mathbf{X}_t = (X_{1,t}, \cdots, X_{m,t})' : t \in \mathbf{Z}\}$ be generated by

$$\mathbf{X}_t = \mathbf{F}_{\boldsymbol{\theta}}(\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \cdots, \mathbf{X}_{t-p}) + \mathbf{H}_{\boldsymbol{\theta}}(\mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \cdots, \mathbf{X}_{t-q})\mathbf{U}_t, \quad (6.100)$$

where $\mathbf{F}_{\boldsymbol{\theta}} : \mathbf{R}^{mp} \to \mathbf{R}^m$ is a vector-valued measurable function, $\mathbf{H}_{\boldsymbol{\theta}} : \mathbf{R}^{mq} \to \mathbf{R}^m \times \mathbf{R}^m$ is a positive definite matrix-valued measurable function, and $\{\mathbf{U}_t = (U_{1,t}, \cdots, U_{m,t})'\}$ is a sequence of i.i.d. random vectors with $E\mathbf{U}_t = \mathbf{0}$, $E|\mathbf{U}_t| < \infty$ and $\mathbf{U}_t$ is independent of $\{\mathbf{X}_s : s < t\}$. Here, $|\mathbf{U}_t|$ is the sum of the absolute values of all entries of $\mathbf{U}_t$, and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_r)' \in \Theta \subset \mathbf{R}^r$, is a vector of unknown parameters. Recall Theorem 6.2, and assume the conditions (i)-(iv) for stationarity of $\{\mathbf{X}_t\}$. Henceforth, without loss of generality we assume $p = q$ and $\|(\cdot)\|$ is the Euclidean norm of $(\cdot)$. Further, we set down the following.

Assumption 6.4 *(A.1)*

$$E_{\boldsymbol{\theta}} \parallel \mathbf{F}_{\boldsymbol{\theta}}(\mathbf{X}_{t-1}, \cdots, \mathbf{X}_{t-p}) \parallel^2 < \infty, \; E_{\boldsymbol{\theta}} \parallel \mathbf{H}_{\boldsymbol{\theta}}(\mathbf{X}_{t-1}, \cdots, \mathbf{X}_{t-p}) \parallel^2 < \infty, \quad \textit{for all } \boldsymbol{\theta} \in \Theta.$$

(A.2) There exists $c > 0$ such that

$$c \le \left\| \mathbf{H}_{\boldsymbol{\theta}'}^{-1/2}(\mathbf{x})\, \mathbf{H}_{\boldsymbol{\theta}}(\mathbf{x})\, \mathbf{H}_{\boldsymbol{\theta}'}^{-1/2}(\mathbf{x}) \right\| < \infty,$$

for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ and for all $\mathbf{x} \in \mathbf{R}^{mp}$.
(A.3) $\mathbf{H}_{\boldsymbol{\theta}}$ and $\mathbf{F}_{\boldsymbol{\theta}}$ are continuously differentiable with respect to $\boldsymbol{\theta}$, and their derivatives $\partial_j \mathbf{H}_{\boldsymbol{\theta}}$ and $\partial_j \mathbf{F}_{\boldsymbol{\theta}}$ ($\partial_j = \partial/\partial\theta_j$, $j = 1, \cdots, r$) satisfy the condition that there exist square-integrable functions A_j and B_j such that

$$\|\partial_j \mathbf{H}_{\boldsymbol{\theta}}\| \le A_j \text{ and } \|\partial_j \mathbf{F}_{\boldsymbol{\theta}}\| \le B_j, \; j = 1, \cdots, r, \quad \textit{for all } \boldsymbol{\theta} \in \Theta.$$

(A.4) The innovation density $p(\cdot)$ satisfies

$$\lim_{\|\mathbf{u}\| \to \infty} \|\mathbf{u}\|\, p(\mathbf{u}) = 0, \qquad \int \mathbf{u}\mathbf{u}'\, p(\mathbf{u})\, d\mathbf{u} = \mathbf{I}_m,$$

where $\mathbf{I}_m$ is the $m \times m$ identity matrix.

(A.5) The continuous derivative Dp of $p(\cdot)$ exists on $\mathbf{R}^m$, and

$$\int \| p^{-1}Dp \|^4 \, p(\mathbf{u})\, d\mathbf{u} < \infty, \quad \int \| \mathbf{u} \|^2 \| p^{-1}Dp \|^2 \, p(\mathbf{u})\, d\mathbf{u} < \infty.$$

Suppose that an observed stretch $\mathbf{X}^{(n)} = (\mathbf{X}_1, \cdots, \mathbf{X}_n)$ from (6.100) is available. We denote the probability distribution of $\mathbf{X}^{(n)}$ by $P_{n,\boldsymbol{\theta}}$. For two hypothetical values $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, the log-likelihood ratio based on $\mathbf{X}^{(n)}$ is

$$\begin{aligned} \Lambda_n(\boldsymbol{\theta}, \boldsymbol{\theta}') &\equiv \log \frac{dP_{n,\boldsymbol{\theta}'}}{dP_{n,\boldsymbol{\theta}}} \\ &= \sum_{t=p}^{n} \log \frac{p\{\mathbf{H}_{\boldsymbol{\theta}'}^{-1}(\mathbf{X}_t - \mathbf{F}_{\boldsymbol{\theta}'})\} \det \mathbf{H}_{\boldsymbol{\theta}}}{p\{\mathbf{H}_{\boldsymbol{\theta}}^{-1}(\mathbf{X}_t - \mathbf{F}_{\boldsymbol{\theta}})\} \det \mathbf{H}_{\boldsymbol{\theta}'}}. \end{aligned} \tag{6.101}$$

The maximum likelihood estimator of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}}_{ML} \equiv \arg\max_{\boldsymbol{\theta}} \Lambda_n(\underline{\boldsymbol{\theta}}, \boldsymbol{\theta}), \tag{6.102}$$

where $\underline{\boldsymbol{\theta}} \in \Theta$ is some fixed value. Denote by $H(p; \boldsymbol{\theta})$ the hypothesis under which the concerned model is (6.100) with unknown parameter $\boldsymbol{\theta} \in \Theta$ and the innovation density $p(\cdot)$. Define the sequence of contiguous alternatives by

$$\boldsymbol{\theta}_n = \boldsymbol{\theta} + \frac{1}{\sqrt{n}}\mathbf{h}, \quad \mathbf{h} \in S \subset \mathbf{R}^r, \tag{6.103}$$

where $\mathbf{h} = (h_1, \cdots, h_r)'$ and S is an open subset of $\mathbf{R}^r$. Henceforth we denote by $\mathbf{R}^{\mathbf{N}}$ to express the product space $\mathbf{R} \times \mathbf{R} \times \cdots$, where component spaces correspond to the coordinate spaces of $(\mathbf{X}_1', \mathbf{X}_2', \cdots)'$, and write its Borel σ-field as $\mathcal{B}^{\mathbf{N}}$. For $\boldsymbol{\Phi}_t \equiv \mathbf{H}_{\boldsymbol{\theta}}^{-1}(\mathbf{X}_t - \mathbf{F}_{\boldsymbol{\theta}})$, we write

$$\mathbf{W}_t = \begin{bmatrix} -\mathrm{vec}'(\mathbf{H}_{\boldsymbol{\theta}}^{-1}\, \partial_1 \mathbf{H}_{\boldsymbol{\theta}}), & \partial_1 \mathbf{F}_{\boldsymbol{\theta}}' \cdot \mathbf{H}_{\boldsymbol{\theta}}^{-1} \\ \vdots & \\ -\mathrm{vec}'(\mathbf{H}_{\boldsymbol{\theta}}^{-1}\, \partial_r \mathbf{H}_{\boldsymbol{\theta}}), & \partial_r \mathbf{F}_{\boldsymbol{\theta}}' \cdot \mathbf{H}_{\boldsymbol{\theta}}^{-1} \end{bmatrix} \left(r \times (m^2 + m) - \text{matrix}\right),$$

$$\boldsymbol{\Psi}_t = \begin{bmatrix} \{\boldsymbol{\Phi}_t \otimes \mathbf{I}_m\} p^{-1}(\boldsymbol{\Phi}_t)\, Dp(\boldsymbol{\Phi}_t) + \mathrm{vec}\, \mathbf{I}_m \\ -p^{-1}(\boldsymbol{\Phi}_t)\, Dp(\boldsymbol{\Phi}_t) \end{bmatrix} \left((m^2 + m) \times 1 - \text{vector}\right),$$

$$\Delta_n = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \mathbf{W}_t \boldsymbol{\Psi}_t,$$

$$\mathcal{F}(p) = E\{\boldsymbol{\Psi}_t\, \boldsymbol{\Psi}_t'\} \qquad \left((m^2 + m) \times (m^2 + m) - \text{matrix}\right),$$

$$\Gamma(p, \boldsymbol{\theta}) = E\left[\mathbf{W}_t\, \mathcal{F}(p)\, \mathbf{W}_t'\right] \qquad (r \times r - \text{matrix}).$$

Now we state the LAN theorem for CHARN models.

Theorem 6.6 *Suppose that the conditions (i)-(iv) of Theorem 6.2 and Assumption 6.4 hold and that $\Gamma(p, \boldsymbol{\theta})$ is positive definite. Then the sequence of*

experiments

$$\mathcal{E}_n = \left\{\mathbf{R}^{\mathbf{N}},\ \mathcal{B}^{\mathbf{N}},\ \{P_{n,\boldsymbol{\theta}}\ :\ \boldsymbol{\theta} \in \Theta \subset \mathbf{R}^r\}\right\},\ n \in \mathbf{N},$$

is locally asymptotically normal and equicontinuous on the compact subset C of S. That is,
(i) For all $\boldsymbol{\theta} \in \Theta$, the log-likelihood ratio $\Lambda_n(\boldsymbol{\theta}, \boldsymbol{\theta}_n)$ admits the following asymptotic representation under $H(p;\boldsymbol{\theta})$:

$$\Lambda_n(\boldsymbol{\theta}, \boldsymbol{\theta}_n) = \mathbf{h}'\Delta_n - \frac{1}{2}\mathbf{h}'\,\Gamma(p,\boldsymbol{\theta})\,\mathbf{h} + o_p(1). \tag{6.104}$$

(ii) Under $H(p;\boldsymbol{\theta})$,

$$\Delta_n \xrightarrow{d} N(\mathbf{0}, \Gamma(p,\boldsymbol{\theta})), \qquad n \to \infty.$$

(iii) For all $n \in \mathbf{N}$ and $\mathbf{h} \in S$, the mapping $\mathbf{h} \to P_{n,\boldsymbol{\theta}_n}$ is continuous with respect to the variational distance

$$\parallel P - Q \parallel = \sup\left\{|P(A) - Q(A)| : A \in \mathcal{B}^{\mathbf{N}}\right\}.$$

PROOF

Let

$$\begin{aligned}
\mathbf{A} &= (\mathbf{H}_{\boldsymbol{\theta}'}^{-1} - \mathbf{H}_{\boldsymbol{\theta}}^{-1})\,\mathbf{H}_{\boldsymbol{\theta}}, \\
\mathbf{b} &= \mathbf{H}_{\boldsymbol{\theta}}^{-1}\,(\mathbf{F}_{\boldsymbol{\theta}'} - \mathbf{F}_{\boldsymbol{\theta}}), \\
G(\mathbf{u};\mathbf{A},\mathbf{b}) &= \frac{p^{1/2}\left[\{\mathbf{I}_m + \mathbf{A}\}\,(\mathbf{u} - \mathbf{b})\right]\det^{1/2}\{\mathbf{I}_m + \mathbf{A}\}}{p^{1/2}(\mathbf{u})} - 1, \\
DG(\mathbf{u}) &\equiv \left(\begin{array}{c} \dfrac{\partial}{\partial\,\mathrm{vec}\mathbf{A}}G(\mathbf{u};\mathbf{A},\mathbf{b}) \\ \\ \dfrac{\partial}{\partial\mathbf{b}}G(\mathbf{u};\mathbf{A},\mathbf{b}) \end{array}\right)_{\mathbf{A}=\mathbf{0},\mathbf{b}=\mathbf{0}}.
\end{aligned}$$

Then it is seen that

$$DG(\mathbf{u}) = \left(\begin{array}{c} \dfrac{1}{2}p(\mathbf{u})^{-1}\,(\mathbf{u}\otimes\mathbf{I}_m)\,Dp(\mathbf{u}) + \dfrac{1}{2}\mathrm{vec}\mathbf{I}_m \\ \\ -\dfrac{1}{2}p(\mathbf{u})^{-1}\,Dp(\mathbf{u}) \end{array}\right),\ \left((m^2+m)\times 1 - \text{vector}\right).$$

Similarly as in Garel and Hallin (1995) or Taniguchi and Kakizawa (2000, pp.40-41), we can prove the following lemma.

Lemma 6.2 *Suppose that Assumption 6.4 holds. Let $\mathbf{v} = ((\mathrm{vec}\mathbf{A})', \mathbf{b}')'$, and*

write $G(\mathbf{u};\mathbf{A},\mathbf{b})$ *as* $G(\mathbf{u};\mathbf{v})$. *Then the following statements hold.*
(i) For all $\mathbf{v}\in\mathbf{R}^{m^2+m}$,

$$\int \left[\, G(\mathbf{u};\mathbf{v}) - \mathbf{v}'\, DG(\mathbf{u}) \,\right]^2 p(\mathbf{u})\, d\mathbf{u} = O\left[\, \|\mathbf{v}\|^2 \,\right]. \tag{6.105}$$

(ii) For all $\mathbf{v}\to\mathbf{0}$ $(\,\mathbf{0}\neq\mathbf{v}\in\mathbf{R}^{m^2+m}\,)$,

$$(\mathbf{v}'\,\mathbf{v})^{-1}\int \left[\, G(\mathbf{u};\mathbf{v}) - \mathbf{v}'\, DG(\mathbf{u}) \,\right]^2 p(\mathbf{u})\, d\mathbf{u} \to 0. \tag{6.106}$$

(iii) For any $\mathbf{v}_n\to\mathbf{0}$ $(\,\mathbf{v}_n\in\mathbf{R}^{m^2+m}\,)$, *and* $c>0$,

$$\lim_{n\to\infty}\sup_{\|\mathbf{s}\|\le c}\int \left\|n^{-1/2}\mathbf{s}+\mathbf{v}_n\right\|^{-2} \times\left[\, G(\mathbf{u};n^{-1/2}\mathbf{s}+\mathbf{v}_n) - (n^{-1/2}\mathbf{s}+\mathbf{v}_n)'\, DG(\mathbf{u}) \,\right]^2 p(\mathbf{u})\, d\mathbf{u} = 0. \tag{6.107}$$

Now we return to the proof of the theorem. Let $W_t^{(n)} = \frac{1}{2\sqrt{n}}\mathbf{h}'\,\mathbf{W}_t\,\mathbf{\Psi}_t$, and let $\mathcal{F}_t$ be the σ-algebra generated by $\{\mathbf{X}_1,\cdots,\mathbf{X}_t\}$. To prove the theorem, we have only to check Swensen's conditions (S1)-(S6) (see Swensen (1985) and Taniguchi and Kakizawa (2000)).
(S1) $E\{W_t^{(n)}|\mathcal{F}_{t-1}\}=0$, *a.e.*
This condition is checked by relation

$$\int \mathbf{u}\,\{Dp(\mathbf{u})\}'\, d\mathbf{u} = -\mathbf{I}_m,$$

which follows from Assumption 6.4.
(S2)

$$\lim_{n\to\infty} E\left[\sum_{t=1}^{n}\left\{U_t^{(n)} - W_t^{(n)}\right\}^2\right] = 0,$$

where

$$U_t^{(n)} = \left[\frac{p\left\{\mathbf{H}_{\boldsymbol{\theta}_n}^{-1}(\,\mathbf{X}_t-\mathbf{F}_{\boldsymbol{\theta}_n}\,)\right\}\det\mathbf{H}_{\boldsymbol{\theta}}}{p\left\{\mathbf{H}_{\boldsymbol{\theta}}^{-1}(\,\mathbf{X}_t-\mathbf{F}_{\boldsymbol{\theta}}\,)\right\}\det\mathbf{H}_{\boldsymbol{\theta}_n}}\right]^{1/2} - 1.$$

This condition is checked by showing

$$\lim_{n\to\infty} E\left[\sum_{t=1}^{n}\left\{U_t^{(n)} - W_t^{*(n)}\right\}^2\right] = 0, \tag{6.108}$$

and

$$\lim_{n\to\infty} E\left[\sum_{t=1}^{n}\left\{W_t^{*(n)} - W_t^{(n)}\right\}^2\right] = 0, \tag{6.109}$$

where $W_t^{*(n)} = \mathbf{v}'\, DG(\phi_t)$. For every $c_1 > 0$, we have

$$\begin{aligned} & E\left[\sum_{t=1}^{n}\left\{U_t^{(n)} - W_t^{*(n)}\right\}^2\right] \\ &= \sum_{t=1}^{n} E\left[\chi\Big\{\sqrt{n}\parallel \mathbf{v}\parallel \le c_1\Big\}\Big\{U_t^{(n)} - W_t^{*(n)}\Big\}^2\right] \\ &+ \sum_{t=1}^{n} E\left[\chi\Big\{\sqrt{n}\parallel \mathbf{v}\parallel > c_1\Big\}\Big\{U_t^{(n)} - W_t^{*(n)}\Big\}^2\right] \\ &= E_1 + E_2, \qquad (\text{say}). \end{aligned}$$

From (iii) of Lemma 6.2, it follows that

$$E_1 \le \left[\sum_{t=1}^{n} E\parallel \mathbf{v}\parallel^2\right] o_{c_1}(1), \tag{6.110}$$

where $\lim_{n\to\infty} o_{c_1}(1) = 0$ for any given $c_1 > 0$. Here,

$$E\|\mathbf{v}\|^2 \le 2E \parallel \text{vec}\mathbf{A} \parallel^2 + 2E \parallel \mathbf{b} \parallel^2 . \tag{6.111}$$

By (iii) of Theorem 6.2 we observe that

$$E \parallel \text{vec}\mathbf{A}\parallel^2 = O\left[E \parallel \mathbf{H}_{\boldsymbol{\theta}_n} - \mathbf{H}_{\boldsymbol{\theta}}\parallel^2\right]. \tag{6.112}$$

Since $E \parallel \partial_j \mathbf{H}_{\boldsymbol{\theta}} \parallel^2 < \infty$ for all $\boldsymbol{\theta} \in \Theta$ (see Assumption 6.4), we obtain

$$\mathbf{H}_{\boldsymbol{\theta}_n} - \mathbf{H}_{\boldsymbol{\theta}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{r} h_i \left(\partial_i \mathbf{H}_{\boldsymbol{\theta}^*}\right), \qquad \boldsymbol{\theta} \lesseqgtr \boldsymbol{\theta}^* \lesseqgtr \boldsymbol{\theta}_n,$$

which implies $E \parallel \text{vec}\mathbf{A} \parallel^2 = O(n^{-1})$. Similarly we get $E \parallel \mathbf{b} \parallel^2 = O(n^{-1})$. Therefore $\lim_{n\to\infty} E_1 = 0$ for any given $c_1 > 0$. We next evaluate E_2. Using (i) of Lemma 6.2, we can see that

$$\begin{aligned} E_2 &\le \sum_{t=1}^{n} E\left[\chi\{\sqrt{n}\,\|\mathbf{v}\| \ge c_1\}\, O(\|\mathbf{v}\|^2)\right] \\ &= \frac{1}{n}\sum_{t=1}^{n} E\left[\chi\{\sqrt{n}\,\|\mathbf{v}\| \ge c_1\}\, O(\|\sqrt{n}\,\mathbf{v}\|^2)\right]. \end{aligned} \tag{6.113}$$

Expanding $\mathbf{v}$ in a Taylor series at $\boldsymbol{\theta}$ it is seen that $\sqrt{n}\,\mathbf{v}$ converges to a random vector $\mathbf{v}^0$ in L_2-sense (note (A.3) of Assumption 6.4). Hence, $\sqrt{n}\,\mathbf{v}$ is uniformly integrable (e.g., Ash (1972, p.297)), which implies that (6.113) converges to zero as $c_1 \to \infty$. Therefore, $E_2 \to 0$, and (6.108) is proved. The assertion (6.109) follows from the definition of $W_t^{*(n)}$ and $W_t^{(n)}$, (iii) of Theorem 6.2, (A.3) of Assumption 6.4 and the dominated convergence theorem. Hence, (S2) is established.

(S3) $\sup_n E\{\sum_{t=1}^{n} W_t^{(n)^2}\} < \infty$.

Recall the definition of $W_t^{(n)}$, i.e., $W_t^{(n)} = \frac{1}{2\sqrt{n}}\mathbf{h}'\,\mathbf{W}_t\,\mathbf{\Psi}_t$, and that $\mathbf{W}_t \in \mathcal{F}_{t-1}$ and $\mathbf{\Psi}_t$ is independent of $\mathbf{W}_t$. Since $\{\mathbf{X}_t\}$ is strictly stationary under (i)-(iv) of Theorem 6.2, we have

$$E\left\{\sum_{t=1}^{n} W_t^{(n)^2}\right\} = \frac{1}{n}\sum_{t=1}^{n} E\left\{nW_t^{(n)^2}\right\}$$

$$= \frac{1}{4}\mathbf{h}'\,E\{\,\mathbf{W}_t\,\mathcal{F}(p)\,\mathbf{W}_t'\,\}\,\mathbf{h} = \frac{\tau^2}{4} \qquad \text{(say)},$$

which, together with Assumption 6.4, implies (S3).

(S4) $\sum_{t=1}^{n} W_t^{(n)^2} \xrightarrow{p} \frac{\tau^2}{4}$.

From the definition of $W_t^{(n)}$, we can see that $\{\sqrt{n}\,W_t^{(n)}\}$ is strictly stationary and ergodic with second moment (see Stout (1974, p.182) and Lu and Jiang (2001)). The assertion (S4) follows from the ergodic theorem (e.g., Stout (1974, p.181)).

(S5) $\max_{1\le t\le n} |W_t^{(n)}| \xrightarrow{p} 0$.

Application of Markov inequality yields

$$P\left(\max_{1\le t\le n} |W_t^{(n)}| > \varepsilon\right) \le P\left[\sum_{t=1}^{n} W_t^{(n)^2}\chi\left\{|W_t^{(n)}| > \varepsilon\right\} > \varepsilon^2\right]$$

$$\le \varepsilon^{-2}n^{-1}\sum_{t=1}^{n} E\left[nW_t^{(n)^2}\chi\left\{\sqrt{n}\,|W_t^{(n)}| > \varepsilon\sqrt{n}\right\}\right], \quad (6.114)$$

for every $\varepsilon > 0$. It is easily seen that $n\,W_t^{(n)^2}$ is uniformly integrable, hence, (6.114) converges to zero as $n \to \infty$.

(S6) $\sum_{t=1}^{n} E\,[\,W_t^{(n)^2}\chi\{\,|W_t^{(n)}| > \delta\}\,\big|\,\mathcal{F}_{t-1}] \xrightarrow{p} 0$, for some $\delta > 0$.

The assertion follows from the fact $\sum_{t=1}^{n} E\,[\,W_t^{(n)^2}\chi\{\,|W_t^{(n)}| > \delta\}] \to 0$ for some $\delta > 0$, as $n \to \infty$, which was shown in the proof of (S5). Therefore (i) and (ii) of the theorem are proved. Finally, part (iii) follows from Scheffés theorem (c.f. Bhattacharya and Rao (1976)) and continuity of $p(\cdot)$. □

Next we discuss the estimation theory for $\boldsymbol{\theta}$. Henceforth the distribution law of a random vector $\mathbf{Y}_n$ under $P_{n,\boldsymbol{\theta}}$ is denoted by $\mathcal{L}\{\mathbf{Y}\big|P_{n,\boldsymbol{\theta}}\}$, and the weak convergence to $\mathbf{Y}$ is denoted by $\mathcal{L}\{\mathbf{Y}_n\big|P_{n,\boldsymbol{\theta}}\} \xrightarrow{d} \mathbf{Y}$. We define the class $\mathcal{A}$ of sequences of estimators of $\boldsymbol{\theta}$, $\{\mathbf{S}_n\}$, as follows

$$\mathcal{A} = \left[\,\{\mathbf{S}_n\}; \mathcal{L}\{\sqrt{n}\,(\mathbf{S}_n - \boldsymbol{\theta}_n)\big|P_{n,\boldsymbol{\theta}_n}\} \xrightarrow{d} \mathbf{Y}_{\boldsymbol{\theta}},\ \text{a probability distribution}\right], \quad (6.115)$$

where $\mathbf{Y}_{\boldsymbol{\theta}}$, in general, depends on $\{\mathbf{S}_n\}$. Let L be the class of all loss functions $l : \mathbf{R}^r \to [0,\infty)$ of the form $l(\mathbf{y}) = \tau(\|\mathbf{y}\|)$ which satisfies $\tau(0) = 0$ and $\tau(a) \le \tau(b)$ if $a \le b$. Typical examples are $l(\mathbf{y}) = \chi\{\|\mathbf{y}\| > a\}$ and $l(\mathbf{y}) = \|\mathbf{y}\|^p$, $p \ge 1$.

Assume the LAN theorem (Theorem 6.6). Then, a sequence $\{\hat{\boldsymbol{\theta}}_n\}$ of estimators

of $\boldsymbol{\theta}$ is said to be an *asymptotically centering estimator* if

$$\sqrt{n}\,(\,\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\,) - \Gamma^{-1}\Delta_n = o_p(1) \quad \text{in } P_{n,\boldsymbol{\theta}}, \tag{6.116}$$

where $\Gamma = \Gamma(p, \boldsymbol{\theta})$. The following theorem can be verified by following the arguments in Strasser (1985, Section 83), Jeganathan (1995), and Taniguchi and Kakizawa (2000, p.69).

Theorem 6.7 *Assume the LAN theorem (Theorem 6.6) for the CHARN model (6.100). Let $\{\mathbf{S}_n\}$ be a sequence of estimators of $\boldsymbol{\theta}$ that belongs to $\mathcal{A}$. Let Δ be a random vector distributed as $N(\mathbf{0}, \Gamma)$. Then the following statements hold.*
(i) For any $l \in L$ with $E\,l(\Delta) < \infty$,

$$\limsup_{n\to\infty} E\,\left[\,l\left\{\sqrt{n}\,(\,\mathbf{S}_n - \boldsymbol{\theta}\,)\right\}\,|P_{n,\boldsymbol{\theta}}\,\right] \geq E\left\{\,l(\Gamma^{-1}\Delta)\right\}. \tag{6.117}$$

(ii) If

$$\limsup_{n\to\infty} E[l\{\sqrt{n}(\mathbf{S}_n - \boldsymbol{\theta})\}|P_{n,\boldsymbol{\theta}}] \leq E\{l(\Gamma^{-1}\Delta)\}, \tag{6.118}$$

for a nonconstant $l \in L$ with $El(\Delta) < \infty$, then $\mathbf{S}_n$ is a sequence of asymptotically centering estimators.

In view of the above result, a sequence of estimators $\{\hat{\boldsymbol{\theta}}_n\} \in \mathcal{A}$ of $\boldsymbol{\theta}$ is *asymptotically efficient* if it is a sequence of asymptotically centering estimators. Write $\eta_t(\boldsymbol{\theta}) \equiv \log p\{\mathbf{H}_{\boldsymbol{\theta}}^{-1}(\mathbf{X}_t - \mathbf{F}_{\boldsymbol{\theta}})\}\{\det \mathbf{H}_{\boldsymbol{\theta}}\}^{-1}$. We further impose the following assumption.

Assumption 6.5 *(i) The true value $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$ is the unique maximum of*

$$\lim_{n\to\infty} n^{-1} \sum_{t=p}^{n} E_{\boldsymbol{\theta}}\{\eta_t(\boldsymbol{\theta})\}$$

with respect to $\boldsymbol{\theta} \in \Theta$.
(ii) $\eta_t(\boldsymbol{\theta})$'s are three times differentiable with respect to $\boldsymbol{\theta}$, and there exist functions $Q_{ij}^t = Q_{ij}^t(\mathbf{X}_1, \ldots, \mathbf{X}_t)$ and $T_{ijk}^t = T_{ijk}^t(\mathbf{X}_1, \ldots, \mathbf{X}_t)$ such that

$$|\partial_i \partial_j \eta_t(\boldsymbol{\theta})| \leq Q_{ij}^t, \quad EQ_{ij}^t < \infty,$$

and

$$|\partial_i \partial_j \partial_k \eta_t(\boldsymbol{\theta})| \leq T_{ijk}^t, \quad ET_{ijk}^t < \infty.$$

Using essentially the same arguments as in Theorems 6.4 and 6.5 it is not difficult to show that the MLE $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is asymptotically centering under (i)-(iv) of Theorem 6.2 and Assumptions 6.4 and 6.5, hence,

Theorem 6.8 *The MLE $\hat{\boldsymbol{\theta}}_{ML}$ for CHARN models is asymptotically efficient.*

Next we discuss the problem of testing. Denote by $\mathrm{E}_{n,\mathbf{h}}(\cdot)$ the expectation with respect to $P_{n,\boldsymbol{\theta}_n}$ with $\boldsymbol{\theta}_n = \boldsymbol{\theta} + \frac{1}{\sqrt{n}}\mathbf{h}$. Let $\mathcal{M}_0$ be a linear subspace of $\mathbf{R}^r$

with finite $k = \dim \mathcal{M}_0^{\perp}$, where $\mathcal{M}_0^{\perp}$ is the orthogonal complement space of $\mathcal{M}_0$. Consider the testing problem (H,A):

$$\text{H: } \mathbf{h} \in \mathcal{M}_0 \quad \text{vs} \quad \text{A: } \mathbf{h} \in \mathbf{R}^r - \mathcal{M}_0. \tag{6.119}$$

A sequence of tests ϕ_n, $n \in \mathbf{N}$, is asymptotically unbiased of level $\alpha \in [0,1]$ for the testing problem (H,A) if

$$\limsup_{n\to\infty} \mathrm{E}_{n,\mathbf{h}}(\phi_n) \le \alpha \quad \text{for } \mathbf{h} \in \mathcal{M}_0$$

and

$$\liminf_{n\to\infty} \mathrm{E}_{n,\mathbf{h}}(\phi_n) \ge \alpha \quad \text{for } \mathbf{h} \in \mathbf{R}^r - \mathcal{M}_0.$$

Similarly to Strasser (1985, Section 82) and Taniguchi and Kakizawa (2000, p.78), we obtain the following theorem.

Theorem 6.9 *Assume the LAN theorem (Theorem 6.6) for the CHARN model (6.100). We denote $\gamma_n = \Gamma^{-1/2}\Delta_n$. For $\alpha \in [0,1]$, choose $k_\alpha \in [0,\infty)$ such that*

$$\lim_{n\to\infty} P_{n,\boldsymbol{\theta}}\left[\|(id - \pi_{\mathcal{M}_0}) \circ \gamma_n\| > k_\alpha\right] = \alpha,$$

where id is the identity map, and $\pi_{\mathcal{M}_0}$ is the orthogonal projection of $\mathbf{R}^r$ onto $\mathcal{M}_0$. Then the following assertions hold.
(i) The sequence of tests

$$\phi_n^* = \begin{cases} 1 & \text{if } \|(id - \pi_{\mathcal{M}_0}) \circ \gamma_n\| > k_\alpha \\ 0 & \text{if } \|(id - \pi_{\mathcal{M}_0}) \circ \gamma_n\| < k_\alpha \end{cases}$$

is asymptotically unbiased of level α for (H,A).
(ii) If $\{\phi_n\}$ is another sequence that is asymptotically unbiased of level α for (H,A), then

$$\limsup_{n\to\infty} \inf_{\mathbf{h}\in B_c} E_{n,\mathbf{h}}(\phi_n) \le \lim_{n\to\infty} \inf_{\mathbf{h}\in B_c} E_{n,\mathbf{h}}(\phi_n^*),$$

where $B_c = \{\mathbf{h} \in \mathbf{R}^r : \|\mathbf{h} - \pi_{\mathcal{M}_0}(\mathbf{h})\| = c\}$, $c > 0$.

If the equality holds, we say that $\{\phi_n\}$ is *locally asymptotically optimal*.

Let $\mathcal{M}(B)$ be the linear space spanned by the columns of a matrix B. The problem consists of testing the null hypothesis H, under which

$$\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in \mathcal{M}(B)$$

for some given $r \times (r-l)$ matrix B of full rank and given vector $\boldsymbol{\theta}_0 \in \mathbf{R}^r$. Then, in view of Theorem 6.9, the test

$$T_n = n(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_0)' \left[\Gamma(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}) - \Gamma(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})B\{B'\Gamma(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})B\}^{-1}B'\Gamma(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})\right] \times (\hat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}_0), \tag{6.120}$$

which has the χ_l^2 null distribution asymptotically, is locally asymptotically optimal.

6.3 Model Selection Problems

Up until now we have assumed that the order of proposed models ($\dim \Theta$) is known. However, in the actual statistical analysis the order must be inferred from the data.

Let us consider the AR(p) model

$$X_t + b_1 X_{t-1} + \cdots + b_p X_{t-p} = u_t, \quad (\{u_t\} \sim \text{i.i.d. } (0,\sigma^2)). \tag{6.121}$$

Observing $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ from (6.121), we are going to estimate the order p of (6.121) based on $\mathbf{X}$. In Section 6.4, it will be shown that the best linear predictor of X_t on $\mathcal{F}_{t-1} = \sigma\{X_{t-1}, X_{t-2}, \ldots\}$, i.e., the linear combination $\sum_{j\geq 1} c_j X_{t-j}$ which minimizes $\mathrm{E}[\{X_t - \sum_{j\geq 1} c_j X_{t-j}\}^2]$, is given by

$$-b_1 X_{t-1} - \cdots - b_p X_{t-p}. \tag{6.122}$$

Now, let $\{Y_t\}$ be a stochastic process which has the same probability structure as $\{X_t\}$, and is mutually independent of $\{X_t\}$. Since the coefficient $\mathbf{b} = (b_1, \ldots, b_p)'$ is unknown, we estimate it by the QGMLE $\hat{\mathbf{b}} = (\hat{b}_{1,\mathrm{QGML}}, \ldots, \hat{b}_{p,\mathrm{QGML}})'$ based on $\mathbf{X}$ (recall (6.42)). Next we predict Y_t by the estimated predictor

$$-\hat{b}_{1,\mathrm{QGML}} Y_{t-1} - \cdots - \hat{b}_{p,\mathrm{QGML}} Y_{t-p}. \tag{6.123}$$

Then the prediction error is

$$\mathrm{E}_{\mathbf{Y}}[\{Y_t + \hat{b}_{1,\mathrm{QGML}} Y_{t-1} + \cdots + \hat{b}_{p,\mathrm{QGML}} Y_{t-p}\}^2], \tag{6.124}$$

where $\mathrm{E}_{\mathbf{Y}}[\cdot]$ is the expectation with respect to $\{Y_t\}$. Since (6.124) is a function of $\mathbf{X}$, further, taking the expectation $\mathrm{E}_{\mathbf{X}}$ with respect to $\mathbf{X}$ we obtain

$$\mathrm{E}_{\mathbf{X}}\mathrm{E}_{\mathbf{Y}}[\{Y_t + \hat{b}_{1,\mathrm{QGML}} Y_{t-1} + \cdots + \hat{b}_{p,\mathrm{QGML}} Y_{t-p}\}^2], \tag{6.125}$$

which is called the *final prediction error* (FPE). Akaike (1970) proposed

$$\mathrm{FPE}(p) = \hat{\sigma}^2_{\mathrm{QGML}}(p)\frac{n+p}{n-p}, \tag{6.126}$$

as an "asymptotically unbiased estimator" of (6.125) in terms of $\mathbf{X}$, where $\hat{\sigma}^2_{\mathrm{QGML}}(p)$ is the QGML estimator (6.43) of σ^2 when we fit AR(p). Then Akaike proposed to choose the order $\hat{p}$ which minimizes FPE(p) with respect to p satisfying $0 \leq p \leq L$, where L is a preassigned positive integer. Hence, from the data, we could completely specify the structure of AR(p) by $(\hat{p}, \hat{b}_{1,\mathrm{QGML}}, \ldots, \hat{b}_{\hat{p},\mathrm{QGML}})$.

FPE was derived for the index of the prediction error of AR model. If we are interested in selection of the order of general probability distribution model $p_{\boldsymbol{\theta}}(\cdot)$ ($r = \dim(\boldsymbol{\theta})$), we use the following index based on Kullback-Leibler information

$$\mathrm{KL}(p, p_{\boldsymbol{\theta}}) \equiv -\int p(\mathbf{x}) \log p_{\boldsymbol{\theta}}(\mathbf{x}) d\mathbf{x} \tag{6.127}$$

which is a measure of disparity between the true model $p(\cdot)$ and a fitted

parametric model $p_{\boldsymbol{\theta}}(\cdot)$. Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$, or an estimator which is asymptotically equivalent to MLE (e.g., QMLE). Akaike (1973) considered the disparity between $p_{\hat{\boldsymbol{\theta}}}$ and p;

$$\mathrm{E}_{\hat{\boldsymbol{\theta}}}[\mathrm{KL}(p, p_{\hat{\boldsymbol{\theta}}})], \tag{6.128}$$

where $\mathrm{E}_{\hat{\boldsymbol{\theta}}}$ is the expectation with respect to the asymptotic distribution of $\hat{\boldsymbol{\theta}}$. As an asymptotically unbiased estimator of (6.128), Akaike proposed

$$\mathrm{AIC}(r) = -2\log\{\text{maximum likelihood}\} + 2r, \tag{6.129}$$

and suggested to select the order $\hat{r}$ which minimizes $\mathrm{AIC}(r)$ with respect to r for $0 \le r \le L$ (a preassigned positive integer). (6.129) is called *Akaike's information criterion* (AIC).

In the i.i.d. case Takeuchi (1976) gave an excellent derivation of a generalized AIC (we call this Takeuchi's information criterion (TIC)), which includes the original AIC as a special case. Borrowing Takeuchi's idea we derive a generalized TIC (GTIC) by a unified method for general stochastic models. Let $\{X_t\}$ be a stochastic process, and let g be a structure specifying $\{X_t\}$. As examples of g we can take the probability distribution function for the i.i.d. case, the trend function for the regression model, the spectral density function for the stationary process, and the dynamic system function for the nonlinear model. We will fit a class of parametric models $\mathcal{P} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset \mathbf{R}^r\}$ to g by use of a measure of disparity $\mathrm{DI}(f_{\boldsymbol{\theta}}, g)$, which takes the minimum if and only if $f_{\boldsymbol{\theta}} = g$ a.e. Suppose that an observed stretch $\mathbf{X} = \{X_1, \ldots, X_n\}$ is available. We estimate $\boldsymbol{\theta}$ by the value $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(X_1, \ldots, X_n)$ which minimizes $\mathrm{DI}(f_{\boldsymbol{\theta}}, \hat{g}_n)$ with respect to $\boldsymbol{\theta}$, where $\hat{g}_n = \hat{g}_n(X_1, \ldots, X_n)$ is an appropriate nonparametric estimator of g. Nearness between the estimated model $f_{\hat{\boldsymbol{\theta}}_n}$ and the true model g is measured by $\mathrm{E}_{\mathbf{X}}\{\mathrm{DI}(f_{\hat{\boldsymbol{\theta}}_n}, g)\}$, where $\mathrm{E}_{\mathbf{X}}\{\cdot\}$ is the expectation with respect to $\mathbf{X}$. Henceforth the regularity conditions for the ordinary asymptotics of $\hat{\boldsymbol{\theta}}_n$ and $\hat{g}_n$ are assumed (e.g., smoothness of the model and the asymptotic normality for $\hat{\boldsymbol{\theta}}_n$). Define the pseudo true value $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$ by

$$\mathrm{DI}(f_{\boldsymbol{\theta}_0}, g) = \min_{\boldsymbol{\theta} \in \Theta} \mathrm{DI}(f_{\boldsymbol{\theta}}, g). \tag{6.130}$$

Expanding $\mathrm{DI}(f_{\hat{\boldsymbol{\theta}}_n}, g)$ around $\boldsymbol{\theta}_0$ we obtain

$$\begin{aligned}\mathrm{E}_{\mathbf{X}}\{\mathrm{DI}(f_{\hat{\boldsymbol{\theta}}_n}, g)\} \approx \mathrm{DI}(f_{\boldsymbol{\theta}_0}, g) &+ (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{DI}(f_{\boldsymbol{\theta}_0}, g) \\ &+ \frac{1}{2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \mathrm{DI}(f_{\boldsymbol{\theta}_0}, g)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0), \end{aligned} \tag{6.131}$$

where $\approx$ means that the left-hand side is approximated by the expectation of the right-hand side. From (6.130) it follows that

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{DI}(f_{\boldsymbol{\theta}_0}, g) = \mathbf{0},$$

which implies that

$$\mathrm{E}_{\mathbf{X}}\{\mathrm{DI}(f_{\hat{\boldsymbol{\theta}}_n}, g)\} \approx \mathrm{DI}(f_{\boldsymbol{\theta}_0}, g) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mathrm{DI}(f_{\boldsymbol{\theta}_0}, g)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0). \quad (6.132)$$

On the other hand we have

$$\begin{aligned}
&\mathrm{DI}(f_{\boldsymbol{\theta}_0}, g) \\
&= \mathrm{DI}(f_{\hat{\boldsymbol{\theta}}_n}, \hat{g}_n) + \{\mathrm{DI}(f_{\boldsymbol{\theta}_0}, \hat{g}_n) - \mathrm{DI}(f_{\hat{\boldsymbol{\theta}}_n}, \hat{g}_n)\} + \{\mathrm{DI}(f_{\boldsymbol{\theta}_0}, g) - \mathrm{DI}(f_{\boldsymbol{\theta}_0}, \hat{g}_n)\} \\
&\approx \mathrm{DI}(f_{\hat{\boldsymbol{\theta}}_n}, \hat{g}_n) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mathrm{DI}(f_{\hat{\boldsymbol{\theta}}_n}, \hat{g}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + M, \qquad (6.133)
\end{aligned}$$

where $M = \mathrm{DI}(f_{\boldsymbol{\theta}_0}, g) - \mathrm{DI}(f_{\boldsymbol{\theta}_0}, \hat{g}_n)$. From (6.132) and (6.133) it follows that

$$\mathrm{E}_{\mathbf{X}}\{\mathrm{DI}(f_{\hat{\boldsymbol{\theta}}_n}, g)\} \approx \mathrm{DI}(f_{\hat{\boldsymbol{\theta}}_n}, \hat{g}_n) + (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + M, \quad (6.134)$$

where

$$J(\boldsymbol{\theta}_0) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mathrm{DI}(f_{\boldsymbol{\theta}_0}, g).$$

Denote by $I(\boldsymbol{\theta}_0)$ the asymptotic variance of $\sqrt{n}(\partial/\partial\boldsymbol{\theta})\mathrm{DI}(f_{\boldsymbol{\theta}_0}, \hat{g}_n)$. Then we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, J(\boldsymbol{\theta}_0)^{-1} I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1}). \quad (6.135)$$

Since the asymptotic mean of M is zero, (6.134) and (6.135) validate that $T_n(r) \equiv \mathrm{DI}(f_{\hat{\boldsymbol{\theta}}_n}, \hat{g}_n) + n^{-1}\mathrm{tr}\{J(\hat{\boldsymbol{\theta}}_n)^{-1} I(\hat{\boldsymbol{\theta}}_n)\}$ is an "asymptotically unbiased estimator" of $\mathrm{E}_{\mathbf{X}}\{\mathrm{DI}(f_{\hat{\boldsymbol{\theta}}_n}, g)\}$. Multiplying $T_n(r)$ by n we call

$$\mathrm{GTIC}(r) \equiv n\mathrm{DI}(f_{\hat{\boldsymbol{\theta}}_n}, \hat{g}_n) + \mathrm{tr}\{J(\hat{\boldsymbol{\theta}}_n)^{-1} I(\hat{\boldsymbol{\theta}}_n)\}, \quad (6.136)$$

a *generalized Takeuchi's criterion* (GTIC). If $J(\hat{\boldsymbol{\theta}}_n)^{-1} I(\hat{\boldsymbol{\theta}}_n)$ depends on g, we replace $I(\cdot)$ and $J(\cdot)$ by appropriate nonparametric estimators $\hat{I}(\cdot)$ and $\hat{J}(\cdot)$ constructed from $\hat{g}_n$. By scanning r successively from 0 to some upper limit L, the order of the model is chosen by the $\hat{r}$ that gives the minimum of $\mathrm{GTIC}(r)$, $r = 0, 1, \ldots, L$. This criterion is very general and includes various criteria as special cases.

Example 6.11 *(FPE). Let $\{X_t : t \in \mathbf{N}\}$ be a Gaussian autoregressive process of order r with spectral density $g(\lambda) = (2\pi)^{-1}\sigma^2 |\sum_{j=0}^{r} \eta_j e^{ij\lambda}|^{-2}$, $\eta_0 \equiv 1$. Assuming that $\{X_t\}$ has the spectral density*

$$f_{\boldsymbol{\theta}}(\lambda) = \frac{\sigma^2}{2\pi}\left|\sum_{j=0}^{r} \theta_j e^{ij\lambda}\right|^{-2}, \quad \theta_0 \equiv 1,$$

we make the best linear predictor $\hat{X}_t = -(\theta_1 X_{t-1} + \cdots + \theta_r X_{t-r})$ of X_t. Then the prediction error is

$$E\{(X_t - \hat{X}_t)^2\} = \int_{-\pi}^{\pi} \frac{g(\lambda)}{h_{\boldsymbol{\theta}}(\lambda)} d\lambda, \quad (6.137)$$

where $h_{\boldsymbol{\theta}}(\lambda) = |\sum_{j=0}^{r} \theta_j e^{ij\lambda}|^{-2}$ (we will explain the foundation of prediction

theory in Section 6.5). In this case we set $DI(f_{\boldsymbol{\theta}}, g) = \int_{-\pi}^{\pi}\{g(\lambda)/h_{\boldsymbol{\theta}}(\lambda)\}d\lambda$ *and* $\hat{g}_n(\lambda) = (2\pi n)^{-1}|\sum_{t=1}^{n} X_t e^{it\lambda}|^2$ *(i.e., the periodogram of* $(X_1, \ldots, X_n)$*). Then* $\hat{\boldsymbol{\theta}}_n$ *becomes the QGMLE of* $\boldsymbol{\theta}$*. Since*

$$\frac{\partial}{\partial \boldsymbol{\theta}} DI(f_{\boldsymbol{\eta}}, \hat{g}_n) = -\int_{-\pi}^{\pi} \frac{\partial}{\partial \boldsymbol{\theta}} h_{\boldsymbol{\eta}}(\lambda) \frac{\hat{g}_n(\lambda)}{h_{\boldsymbol{\eta}}(\lambda)^2} d\lambda, \quad \boldsymbol{\eta} = (\eta_1, \ldots, \eta_r)', \tag{6.138}$$

application of Lemma 6.1 yields

$$I(\boldsymbol{\eta}) = 4\pi \int_{-\pi}^{\pi} \frac{\partial h_{\boldsymbol{\eta}}(\lambda)}{\partial \boldsymbol{\theta}} \cdot \frac{\partial h_{\boldsymbol{\eta}}(\lambda)}{\partial \boldsymbol{\theta}'} \frac{g(\lambda)^2}{h_{\boldsymbol{\eta}}(\lambda)^4} d\lambda.$$

Similarly we have

$$J(\boldsymbol{\eta}) = \int_{-\pi}^{\pi} \frac{\partial h_{\boldsymbol{\eta}}(\lambda)}{\partial \boldsymbol{\theta}} \cdot \frac{\partial h_{\boldsymbol{\eta}}(\lambda)}{\partial \boldsymbol{\theta}'} \frac{g(\lambda)}{h_{\boldsymbol{\eta}}(\lambda)^3} d\lambda.$$

Hence, $tr\{J(\boldsymbol{\eta})^{-1} I(\boldsymbol{\eta})\} = 2\sigma^2 r$*, and the GTIC is given by*

$$GTIC(r) = (n + 2r)\hat{\sigma}_r^2, \tag{6.139}$$

where $\hat{\sigma}_r^2 = \int_{-\pi}^{\pi} \hat{g}_n(\lambda)/h_{\hat{\boldsymbol{\theta}}_n}(\lambda) d\lambda$*. This is essentially equivalent to the original FPE criterion proposed by Akaike (see Shibata (1980)).*

Example 6.12 *(AIC and BIC). Let* $X_1, \ldots, X_n$ *be a sequence of i.i.d. random variables with probability density* $g(\cdot)$*. Assume that* g *is sufficiently approximated by a class of parametric models* $\mathcal{P} = \{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset \mathbf{R}^r\}$*. In this case we set* $DI(f_{\boldsymbol{\theta}}, g) = -\int g(x) \log f_{\boldsymbol{\theta}}(x) dx$ *and* $\int \hat{g}_n =$ *the empirical distribution function. Then* $DI(f_{\boldsymbol{\theta}}, \hat{g}_n) = -n^{-1} \sum_{t=1}^{n} \log f_{\boldsymbol{\theta}}(X_t)$*, where* $\hat{\boldsymbol{\theta}}_n$ *is the MLE of* $\boldsymbol{\theta}$*. Since*

$$J(\boldsymbol{\theta}_0) = -E_g \left\{ \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f_{\boldsymbol{\theta}_0}(X_1) \right\}$$

and

$$I(\boldsymbol{\theta}_0) = E_g \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}_0}(X_1) \frac{\partial}{\partial \boldsymbol{\theta}'} \log f_{\boldsymbol{\theta}_0}(X_1) \right\},$$

the GTIC(r) becomes Takeuchi's criterion

$$TIC(r) = -\log \prod_{t=1}^{n} f_{\hat{\boldsymbol{\theta}}_n}(X_t) + tr\{J(\hat{\boldsymbol{\theta}}_n)^{-1} I(\hat{\boldsymbol{\theta}}_n)\}. \tag{6.140}$$

Here it should be noted that $J(\boldsymbol{\theta}_0) \neq I(\boldsymbol{\theta}_0)$ *generally if* $g \notin \mathcal{P}$*, but if* $g \in \mathcal{P}$*,* $J(\boldsymbol{\theta}_0) = I(\boldsymbol{\theta}_0)$*. Therefore, if* $g \in \mathcal{P}$ *we can see that TIC(r) is equivalent to AIC(r) defined by (6.129).*

Example 6.13 *(AIC for ARMA and CHARN). (i) Suppose that* $\{X_1, \ldots, X_n\}$ *is an observed stretch from a Gaussian ARMA(p,q) process with spectral density*

$$g(\lambda) = \frac{\sigma^2}{2\pi} \frac{\left|1 + \sum_{j=1}^{q} a_j^{(0)} e^{ij\lambda}\right|^2}{\left|1 + \sum_{j=1}^{p} b_j^{(0)} e^{ij\lambda}\right|^2}.$$

Let

$$\mathcal{P} = \left\{ f_{\boldsymbol{\theta}} : f_{\boldsymbol{\theta}}(\lambda) = \frac{\sigma^2}{2\pi} \frac{\left|1 + \sum_{j=1}^{q} a_j e^{ij\lambda}\right|^2}{\left|1 + \sum_{j=1}^{p} b_j e^{ij\lambda}\right|^2},\ \boldsymbol{\theta} = (a_1, \ldots, a_q, b_1, \ldots, b_p, \sigma^2)' \right\}$$

be a class of fitted spectral density models. In this case we set

$$DI(f_{\boldsymbol{\theta}}, g) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left\{ \log f_{\boldsymbol{\theta}}(\lambda) + \frac{g(\lambda)}{f_{\boldsymbol{\theta}}(\lambda)} \right\} d\lambda$$

and

$$\hat{g}_n(\lambda) = \frac{1}{2\pi n} \left| \sum_{t=1}^{n} X_t e^{it\lambda} \right|^2.$$

Then it is not difficult to show that AIC for this case is equivalent to

$$AIC(p, q) = n \log \hat{\sigma}^2(p, q) + 2(p + q), \tag{6.141}$$

where $\hat{\sigma}^2(p, q)$ is the QGML of σ^2 defined by (6.60).
(ii) Suppose that $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ is an observed stretch from the CHARN model (6.100). Recalling (6.101), (6.102) and (6.129), we can see that the AIC for CHARN is given by

$$AIC(r) = -2 \sum_{t=\max(p,q)}^{n} \log p \left\{ \mathbf{H}_{\hat{\boldsymbol{\theta}}_{ML}}^{-1} (\mathbf{X}_t - \mathbf{F}_{\hat{\boldsymbol{\theta}}_{ML}}) \right\} \det \mathbf{H}_{\hat{\boldsymbol{\theta}}_{ML}}^{-1} + 2r. \tag{6.142}$$

Because the essential idea of GTIC is that it is an asymptotically unbiased estimator of a discrepancy of the estimated model $f_{\hat{\boldsymbol{\theta}}_n}$ from the true structure g, we can easily understand that the selected order $\hat{r}$ by GTIC(r) is not a consistent estimator of the true order, r_0 (i.e., $\hat{r} \overset{\text{p}}{\not\to} r_0$). For fitting AR($r$) models, Shibata (1976) showed that AIC has a tendency to overestimate r_0. There is another criterion which attempts to correct the overfitting nature of AIC. It is defined to be

$$\begin{aligned} \text{BIC} = {} & -2 \log(\text{maximum likelihood}) \\ & + (\log n)(\text{number of parameters}), \end{aligned} \tag{6.143}$$

where n is the sample size used for the computation of the MLE (see Akaike (1977, 1978) and Schwarz (1978)). The BIC is a consistent order selection procedure. If $\{X_1, \ldots, X_n\}$ are observations of an ARMA(p,q) process, and if $\hat{p}$ and $\hat{q}$ are the orders estimated by BIC, then Hannan (1980) showed that $\hat{p}$ and $\hat{q}$ are strongly consistent (i.e., $\hat{p} \xrightarrow{\text{a.s.}} p$ and $\hat{q} \xrightarrow{\text{a.s.}} q$ as $n \to \infty$).

For fitting AR(r) models, Hannan and Quinn (1979) suggested the criterion

$$\text{HQ}(r) = \log \hat{\sigma}^2 + n^{-1} 2rc \cdot \log \log n, \quad c > 1, \tag{6.144}$$

where $\hat{\sigma}^2$ is the QGMLE of the innovation variance. Then they proved that the estimated order given by minimizing HQ(r) is strongly consistent for the true

order. This consistency property is not shared by the AIC or FPE. But this does not mean the disadvantage of AIC or FPE. Shibata (1980) showed that order selection by minimization of the AIC or FPE is asymptotically efficient for autoregressive models, while order selection by BIC or HQ minimization is not so. Shibata's efficiency is defined as follows. Let $\{X_t\}$ be an AR(∞) process of the form

$$X_t + \sum_{j=1}^{\infty} a_j X_{t-j} = u_t,$$

where $\{u_t\} \sim$ i.i.d. $\mathrm{N}(0, \sigma^2)$. Let $(\hat{a}_{r1}, \ldots, \hat{a}_{rr})'$ be the QGMLE of the coefficients of an AR(r) model fitted to the data $\{X_1, \ldots, X_n\}$. The prediction error for an independent realization $\{Y_t\}$ of $\{X_t\}$ based on the AR(r) model fitted to $\{X_t\}$ is

$$\begin{aligned} &\mathrm{E}_Y\{(Y_t - \hat{a}_{r1}Y_{t-1} - \cdots - \hat{a}_{rr}Y_{t-r})^2\} \\ &= \sigma^2 + (\hat{\mathbf{a}}_{r,\infty} - \mathbf{a}_\infty)'\Gamma_\infty(\hat{\mathbf{a}}_{r,\infty} - \mathbf{a}_\infty) = H(r), \quad \text{(say)}, \end{aligned} \tag{6.145}$$

where $\mathrm{E}_Y(\cdot)$ is the expectation with respect to $\{Y_t\}$, Γ_∞ is the infinite-dimensional covariance matrix of $\{Y_t\}$, and $\hat{\mathbf{a}}_{r,\infty} = (\hat{a}_{r1}, \ldots, \hat{a}_{rr}, 0, 0, \ldots)'$ and $\mathbf{a}_\infty = (a_1, a_2, \ldots)'$ are the infinite-dimensional vectors. Then an order selection procedure is said to be *asymptotically efficient* if the estimated order $\hat{r}_n$ satisfies

$$\frac{H(r_n^*)}{H(\hat{r}_n)} \xrightarrow{\mathrm{p}} 1 \quad \text{as } n \to \infty, \tag{6.146}$$

where r_n^* is the value of r which minimizes $H(r)$, $0 \le r \le k_n$, and k_n is a sequence of constants tending to infinity at a suitable rate.

Because we explained the model selection procedures and the parameter estimation methods, we can now identify the statistical models from real data.

Figure 6.6 shows the graph of the daily stock returns $\{X_t\}$ of the Ford Motor Company from July 3, 1962, to December 31, 1991 (7420 trading days).

Figure 6.7 plots the sample autocorrelation function of $\{X_t\}$:

$$\mathrm{SACF}_{X_t}(l) = \frac{\sum_{t=1}^{n-l}(X_{t+l} - \bar{X}_n)(X_t - \bar{X}_n)}{\sum_{t=1}^{n}(X_t - \bar{X}_n)^2}, \tag{6.147}$$

where $\bar{X}_n = n^{-1}\sum_{t=1}^{n} X_t$.

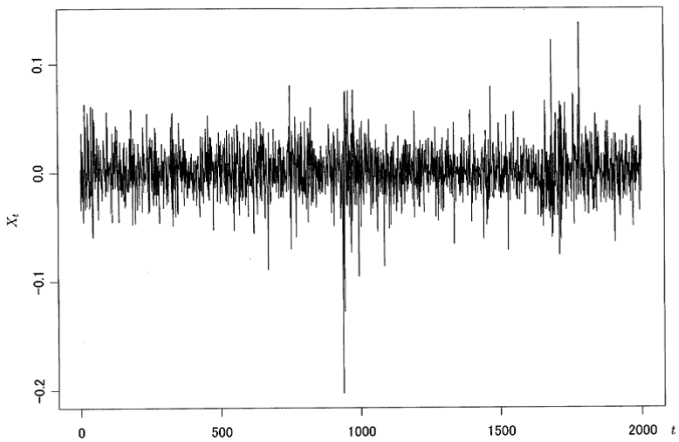


Figure 6.6 *Ford daily returns.*

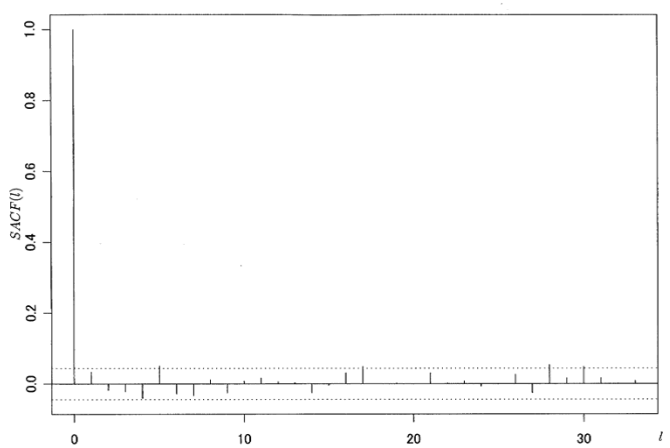


Figure 6.7 *The sample autocorrelation.*

From Figure 6.7, we observe that the values of $\mathrm{SACF}_{X_t}(l)$ are near 0 except for the case of $l = 0$, which suggests that X_t's are almost uncorrelated. Figure 6.8 plots the sample autocorrelation function of X_t^2, i.e., $\mathrm{SACF}_{X_t^2}(l)$.

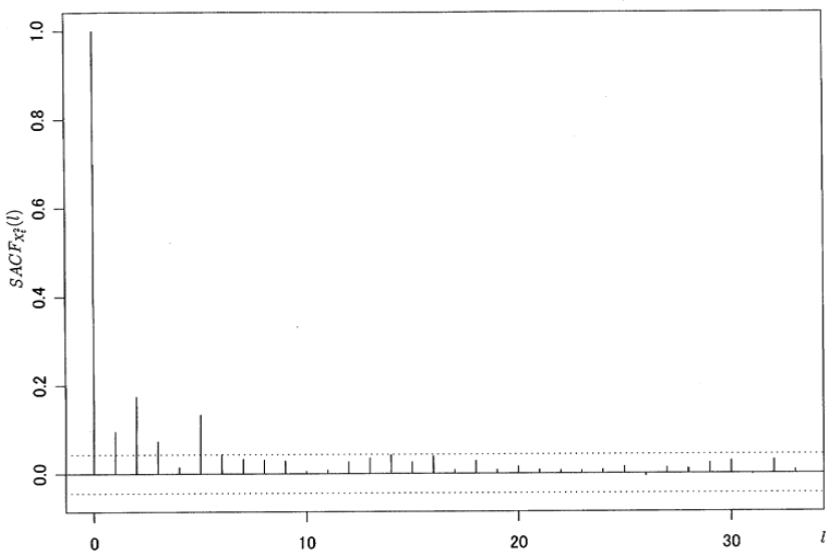


Figure 6.8 *SACF of square transformed data.*

From Figure 6.8, we observe that some values of $\mathrm{SACF}_{X_t^2}(l)$, $l \neq 0$, are fairly deviated from zero, which suggests that X_t's are not mutually independent. Summarizing the above observations we may suppose that $\{X_t\}$ is an uncorrelated, but not independent process. Also this entails that $\{X_t\}$ is not Gaussian.

Now, let us fit the following EGARCH(p,q) model to the data:

$$\begin{cases} X_t = u_t \sigma_t \\ \log \sigma_t^2 = a_0 + \displaystyle\sum_{j=1}^{p} a_j \frac{|X_{t-j}| + \gamma_j X_{t-j}}{\sigma_{t-j}} + \sum_{j=1}^{q} b_j \log \sigma_{t-j}^2. \end{cases} \tag{6.148}$$

First, assuming $\{u_t\} \sim$ i.i.d. $\mathrm{N}(0,1)$, we fit EGARCH(p,q) to the data by use of AIC. Here it should be noted that Gaussianity of $\{u_t\}$ does not imply that

of $\{X_t\}$. The results are given below. $\hat{a}_j$, $\hat{b}_j$ and $\hat{\gamma}_j$ are QMLE of a_j, b_j and γ_j, respectively.

Table 6.1

EGARCH(p,q)		
(p,q)	AIC	QMLE
(0,1)	−7601.867	$\hat{a}_0 = -0.03514$
		$\hat{b}_1 = -0.99628$
(1,0)	−10370.23	$\hat{a}_0 = -8.2313$
		$\hat{a}_1 = 0.2579$
		$\hat{\gamma}_1 = 0.1155$
(1,1)	−10497.51	$\hat{a}_0 = -0.2398$
		$\hat{a}_1 = 0.1000$
		$\hat{\gamma}_1 = 1.236 \times 10^{-9}$
		$\hat{b}_1 = 0.98$
(1,2)	−7749.906	$\hat{a}_0 = -0.2398$
		$\hat{a}_1 = 0.1000$
		$\hat{\gamma}_1 = -1.868 \times 10^{-9}$
		$\hat{b}_1 = 0.9800$
		$\hat{b}_2 = 0.010$
(2,0)	−10448.16	$\hat{a}_0 = -8.4995$
		$\hat{a}_1 = 0.2625$
		$\hat{a}_2 = 0.3123$
		$\hat{\gamma}_1 = 0.1635$
		$\hat{\gamma}_2 = -0.3724$
(2,1)	−10494.31	$\hat{a}_0 = -0.24097$
		$\hat{a}_1 = 0.10051$
		$\hat{a}_2 = 0.00055$
		$\hat{\gamma}_1 = 0.00182$
		$\hat{\gamma}_2 = -0.38790$
		$\hat{b}_1 = 0.97986$
(2,2)	−7837.20	$\hat{a}_0 = -0.2398$
		$\hat{a}_1 = 0.1000$
		$\hat{a}_2 = 0.00100$
		$\hat{\gamma}_1 = 5.333 \times 10^{-9}$
		$\hat{\gamma}_2 = 4.177 \times 10^{-7}$
		$\hat{b}_1 = 0.9800$
		$\hat{b}_2 = 0.0100$

It is seen that AIC selected EGARCH(1,1) model for the data. Next, assuming that $\{u_t\} \sim$ i.i.d. $t(\nu)$, we fitted EGARCH(p,q) to the data by AIC. Then AIC selected the following model:

Table 6.2

$$\begin{aligned} \text{EGARCH(1,2): AIC} &= -10584.62 \\ \hat{a}_0 &= -0.4476 \\ \hat{a}_1 &= 0.1767 \\ \hat{\gamma}_1 &= -0.1498 \\ \hat{b}_1 &= 0.4547 \\ \hat{b}_2 &= 0.5068 \\ \hat{\nu} &= 7.6537\ . \end{aligned}$$

Here the degree of freedom ν was estimated by QMLE as unknown parameter.

6.4 Nonparametric Estimation

In the previous sections we dealt with time series models described by unknown parameter $\boldsymbol{\theta}$, and discussed estimation of $\boldsymbol{\theta}$. In this section we discuss estimation for time series models which are not described by finite-dimensional unknown parameters. Especially we consider the problems of nonparametric estimation for spectral density of stationary processes. Spectral density is a fundamental and important index of stochastic processes.

Let $\{X_t\}$ be a Gaussian stationary process with mean zero, covariance function $R(\cdot)$ and spectral density function $f(\lambda)$. Initially, we make the following assumption.

Assumption 6.6

$$\sum_{j=-\infty}^{\infty} |j||R(j)| < \infty.$$

Suppose that an observed stretch $\{X_1, X_2, \ldots, X_n\}$ is available. In Theorem 5.2 we saw that the periodogram $I_n(\lambda) = (2\pi n)^{-1} |\sum_{t=1}^{n} X_t e^{it\lambda}|^2$ becomes an asymptotically unbiased estimator of $f(\lambda)$, i.e.,

$$\lim_{n\to\infty} \mathrm{E}\{I_n(\lambda)\} = f(\lambda). \tag{6.149}$$

Thus we might think that $I_n(\lambda)$ could be used as a nonparametric estimator of $f(\lambda)$. However, in what follows, we will show that $I_n(\lambda)$ is not suitable as an estimator of $f(\lambda)$. For this, let us see the covariance structure of $I_n(\lambda)$. For

discrete frequencies $\lambda_j = 2\pi j/n$, $j \in \mathbf{Z}$, we have

$$\begin{aligned}
&\mathrm{Cov}\{I_n(\lambda_j), I_n(\lambda_k)\}\mathrm{E}\{I_n(\lambda_j), I_n(\lambda_k)\} - \mathrm{E}\{I_n(\lambda_j)\}\mathrm{E}\{I_n(\lambda_k)\} \\
&= \left(\frac{1}{2\pi n}\right)^2 \mathrm{E}\left\{ \left(\sum_{t=1}^{n} X_t e^{it\lambda_j}\right)\left(\sum_{s=1}^{n} X_s e^{-is\lambda_j}\right) \right. \\
&\qquad \left. \left(\sum_{u=1}^{n} X_u e^{iu\lambda_k}\right)\left(\sum_{v=1}^{n} X_v e^{-iv\lambda_k}\right)\right\} \\
&\quad - \left(\frac{1}{2\pi n}\right)^2 \mathrm{E}\left\{\left(\sum_{t=1}^{n} X_t e^{it\lambda_j}\right)\left(\sum_{s=1}^{n} X_s e^{-is\lambda_j}\right)\right\} \\
&\qquad \mathrm{E}\left\{\left(\sum_{u=1}^{n} X_u e^{iu\lambda_k}\right)\left(\sum_{v=1}^{n} X_v e^{-iv\lambda_k}\right)\right\}. \qquad (6.150)
\end{aligned}$$

Since $\{X_t\}$ is a Gaussian stationary process,

$$\begin{aligned}
&\mathrm{E}\{X_t X_s X_u X_v\} \\
&= R(s-t)R(v-u) + R(u-t)R(v-s) + R(v-t)R(u-s), \qquad (6.151)
\end{aligned}$$

(see Exercise 6.7). Similarly as in Theorem 5.2 (i), we can see that (6.150) is

$$\begin{aligned}
&\left(\frac{1}{2\pi n}\right)^2 \sum_{t=1}^{n}\sum_{s=1}^{n}\sum_{u=1}^{n}\sum_{v=1}^{n}\{R(u-t)R(v-s) + R(v-t)R(u-s)\} \\
&\qquad \times e^{i\lambda_j(t-s)} \times e^{i\lambda_k(u-v)} \\
&= (\mathrm{A}){+}(\mathrm{B}), \quad (\text{say}). \qquad (6.152)
\end{aligned}$$

Letting $u - t = h$ and $v - s = l$, we obtain

$$\begin{aligned}
(\mathrm{A}) &= \left(\frac{1}{2\pi n}\right)^2 \sum_{h=-n+1}^{n-1}\sum_{l=-n+1}^{n-1} R(h)R(l) \\
&\quad \times \sum_{1\le u\le n,\ 1\le u-h\le n}\ \sum_{1\le v\le n,\ 1\le v-l\le n} e^{i\lambda_j(u-h+l-v)} \times e^{i\lambda_k(u-v)} \\
&= \frac{1}{n}\left\{\frac{1}{2\pi}\sum_{h=-n+1}^{n-1} R(h)e^{-i\lambda_j h} \sum_{1\le u\le n,\ 1\le u-h\le n} e^{i(\lambda_j+\lambda_k)u}\right\} \\
&\quad \times \frac{1}{n}\left\{\frac{1}{2\pi}\sum_{l=-n+1}^{n-1} R(l)e^{i\lambda_j l} \sum_{1\le v\le n,\ 1\le v-l\le n} e^{-i(\lambda_j+\lambda_k)v}\right\} \\
&= (\mathrm{A1}) \times (\mathrm{A2}), \quad (\text{say}). \qquad (6.153)
\end{aligned}$$

From Assumption 6.6 and $|e^{i(\lambda_j+\lambda_k)u}| = 1$, it is seen that

$$\left|(\text{A1}) - \frac{1}{2\pi n}\left\{\sum_{h=-n+1}^{n-1} R(h)e^{-i\lambda_j h}\sum_{u=1}^{n} e^{i(\lambda_j+\lambda_k)u}\right\}\right|$$
$$\leq \frac{1}{2\pi n}\sum_{h=-n+1}^{n-1}|h||R(h)| = O(n^{-1}). \tag{6.154}$$

Since it is possible to get a similar inequality for (A2), we have

$$(\text{A}) = \left\{\frac{1}{2\pi}\sum_{h=-n+1}^{n-1} R(h)e^{-i\lambda_j h}\ \frac{1}{n}\sum_{u=1}^{n} e^{i(\lambda_j+\lambda_k)u} + O(n^{-1})\right\}$$
$$\times\left\{\frac{1}{2\pi}\sum_{l=-n+1}^{n-1} R(l)e^{i\lambda_j l}\ \frac{1}{n}\sum_{v=1}^{n} e^{-i(\lambda_j+\lambda_k)v} + O(n^{-1})\right\}. \tag{6.155}$$

Then (5.22) and (5.26) yield

$$(\text{A}) = \begin{cases} f(\lambda_j)^2 + O(n^{-1}), & j+k = 0 \ (\text{mod } n), \\ O(n^{-2}), & j+k \neq 0 \ (\text{mod } n). \end{cases}$$

Similarly,

$$(\text{B}) = \begin{cases} f(\lambda_j)^2 + O(n^{-1}), & j-k = 0 \ (\text{mod } n), \\ O(n^{-2}), & j-k \neq 0 \ (\text{mod } n). \end{cases}$$

Summarizing the above we have,

Theorem 6.10 *Under Assumption 6.6, for* $\lambda_j,\ \lambda_k \in [-\pi, \pi]$,

(i)

$$Var\{I_n(\lambda_j)\} = \begin{cases} f(\lambda_j)^2 + O(n^{-1}), & (j \neq 0) \\ 2f(\lambda_j)^2 + O(n^{-1}), & (j = 0), \end{cases} \tag{6.156}$$

(ii)

$$Cov\{I_n(\lambda_j), I_n(\lambda_k)\} = O(n^{-2}), \quad (j \pm k \neq 0). \tag{6.157}$$

It may be noted that (i) of Theorem 6.10 implies

$$\lim_{n\to\infty} \text{Var}\{I_n(\lambda_j)\} = f(\lambda_j)^2 > 0, \quad (j \neq 0), \tag{6.158}$$

hence, $I_n(\lambda)$ is not a consistent estimator of $f(\lambda)$, although it is asymptotically unbiased. Therefore $I_n(\lambda)$ itself is not suitable as an estimator of $f(\lambda)$. Let us check this graphically. Suppose that $X_1, X_2, \ldots, X_{500}$'s are generated by the AR(2) model:

$$X_t - 0.9X_{t-1} + 0.14X_{t-2} = u_t, \tag{6.159}$$

where $\{u_t\} \sim$ i.i.d. $N(0,1)$. The spectral density of (6.159) is

$$f(\lambda) = \frac{1}{2\pi}|1 - 0.9e^{i\lambda} + 0.14e^{2i\lambda}|^{-2}.$$

Figure 6.9 plots the periodogram $I_{500}(\lambda)$ by real line, and $f(\lambda)$ by dotted line for $\lambda_j = 2\pi j/500$, $j = 1, \ldots, 250$.

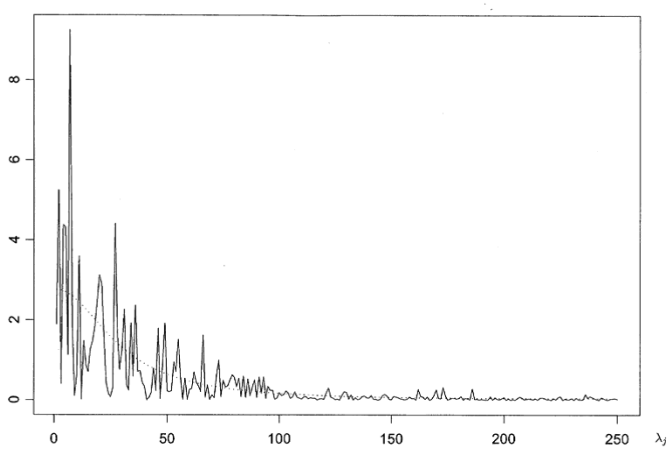
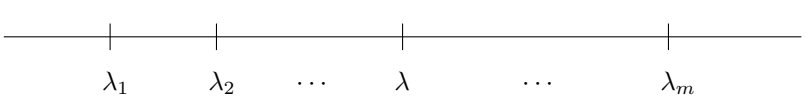


Figure 6.9 *Graphs of* $I_{500}(\lambda)$ *and* $f(\lambda)$.

From Figure 6.9 we observe that the graph of the periodogram is very toothed and variable, hence we can understand that the periodogram is not a consistent estimator of $f(\lambda)$.

Now, how should we construct a nonparametric and consistent estimator of $f(\lambda)$? We already saw in (ii) of Theorem 6.10 that $\mathrm{Cov}\{I_n(\lambda_j), I_n(\lambda_k)\} = O(n^{-2})$ if $\lambda_j \neq \lambda_k$. Hence, taking frequencies $\lambda_1, \ldots, \lambda_m$ in a neighborhood of λ such as



we make the following average of periodograms

$$\hat{s}_n(\lambda) \equiv \frac{1}{m}\sum_{j=1}^{m} I_n(\lambda_j) \tag{6.160}$$

as an estimator of $f(\lambda)$. Here, from Theorem 6.10, the variance of $\hat{s}_n(\lambda)$ is,

$$\begin{aligned}\mathrm{Var}\{\hat{s}_n(\lambda)\} &= \mathrm{Var}\left\{\frac{1}{m}\sum_{j=1}^{m} I_n(\lambda_j)\right\} \\ &= \frac{1}{m^2}\sum_{j=1}^{m}\mathrm{Var}\{I_n(\lambda_j)\} + \frac{1}{m^2}\sum_{j\neq k}\mathrm{Cov}\{I_n(\lambda_j), I_n(\lambda_k)\} \\ &= O\left(\frac{1}{m}\right) + O\left(\frac{1}{n^2}\right).\end{aligned}$$

Therefore, if $m = m(n) \to \infty$ as $n \to \infty$, then $\mathrm{Var}\{\hat{s}_n(\lambda)\} \to 0$. On the other hand, $I_n(\lambda_j)$, $j = 1, \ldots, m$, are required to be asymptotically unbiased estimators of $f(\lambda)$, hence, we assume $m/n \to 0$ as $n \to \infty$. Figure 6.10 plots the graphs of $\hat{s}_n(\lambda_j)$, $I_n(\lambda_j)$ and $f(\lambda_j)$, for $\lambda_j = 2\pi j/500$, $j = 1, \ldots, 250$, by real line, dotted line and broken line, respectively, when the data is generated by the AR(2) in (6.159) with $n = 500$ and $m = 8$.

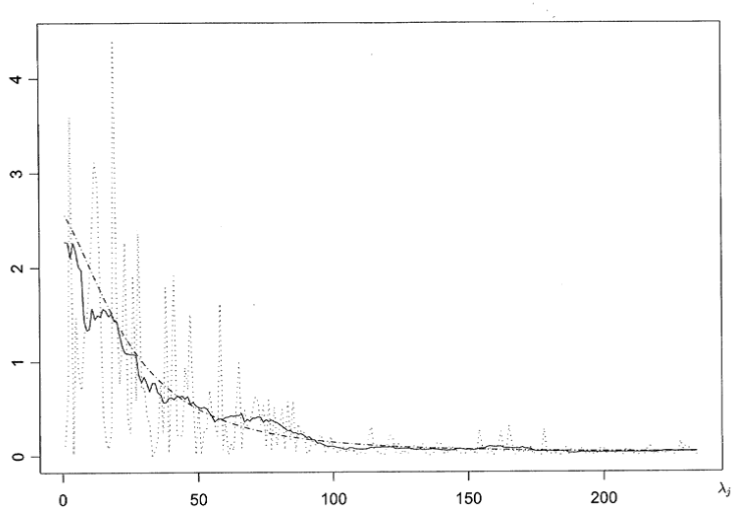


Figure 6.10 *Graphs of* $\hat{S}_n(\lambda)$, $I_n(\lambda)$ *and* $f(\lambda)$.

From this figure it is seen that $\hat{s}_n(\lambda)$ is a better estimator of $f(\lambda)$ than $I_n(\lambda)$, which suggests that, if we want to make a consistent estimator of $f(\lambda)$, we should smooth the periodogram, and further consider a weighted smoothing

$$\hat{f}_n(\lambda) = \int_{-\pi}^{\pi} W_n(\lambda - \mu) I_n(\mu) d\mu, \tag{6.161}$$

where $W_n(\lambda)$ is a weight function, and is called a *spectral window function*. $W_n(\lambda)$ is assumed to have the form

$$W_n(\lambda) = \frac{1}{2\pi} \sum_{|l| \le M} w\left(\frac{l}{M}\right) e^{-il\lambda}. \tag{6.162}$$

Here M is a positive integer, and $w(\cdot)$ is called a *lag window function*. Their prescriptions will be given later. To investigate the asymptotics of $\hat{f}_n(\lambda)$, we set down the following assumption.

Assumption 6.7 *For some positive integer q,*

$$\sum_{j=-\infty}^{\infty} |j|^q |R(j)| < \infty.$$

Next $w(\cdot)$ and M are required to satisfy the following.

Assumption 6.8 *Let q be given in Assumption 6.7.*
(i) $w(x)$ is a continuous, even function, and satisfies

$$w(0) = 1, \ |w(x)| \le 1, \ x \in [-1, 1], \ w(x) = 0, \ |x| > 1.$$

(ii) The limit

$$\lim_{x \to 0} \frac{1 - w(x)}{|x|^q} = \kappa_q < \infty, \tag{6.163}$$

exists.
(iii) $M = M(n)$ satisfies

$$M \to \infty \ \ and \ \ \frac{M^q}{n} \to 0, \quad as \ n \to \infty.$$

The asymptotics of $\hat{f}_n(\lambda)$ is given as follows.

Theorem 6.11 *Under Assumptions 6.7 and 6.8, the following statements hold true.*
(i)

$$\begin{aligned} \lim_{n \to \infty} M^q [E\{\hat{f}_n(\lambda)\} - f(\lambda)] &= -\frac{\kappa_q}{2\pi} \sum_{l=-\infty}^{\infty} |l|^q R(l) e^{-il\lambda}, \ \lambda \in [-\pi, \pi], \\ &= \kappa_q b(\lambda), \quad (say). \end{aligned} \tag{6.164}$$

(ii)

$$\lim_{n\to\infty} \frac{n}{M} Var\{\hat{f}_n(\lambda)\} = \begin{cases} 2f(\lambda)^2 \int_{-1}^{1} w(x)^2 dx, & (\lambda = 0, \pi) \\ f(\lambda)^2 \int_{-1}^{1} w(x)^2 dx, & (0 < \lambda < \pi). \end{cases}$$

PROOF

Letting $\hat{R}(l) = n^{-1} \sum_{t=1}^{n-|l|} X_t X_{t+|l|}$, we can rewrite the periodogram as

$$I_n(\lambda) = \frac{1}{2\pi} \sum_{l=-n+1}^{n-1} \hat{R}(l) e^{-il\lambda}.$$

Then (6.161) can be written as

$$\hat{f}_n(\lambda) = \frac{1}{2\pi} \sum_{l=-M}^{M} w\left(\frac{l}{M}\right) \hat{R}(l) e^{-il\lambda}, \tag{6.165}$$

(Exercise 6.8). Since $\mathrm{E}\{\hat{R}(l)\} = (1 - |l|/n)R(l)$, we obtain

$$\mathrm{E}\{\hat{f}_n(\lambda)\} = \frac{1}{2\pi} \sum_{l=-M}^{M} w\left(\frac{l}{M}\right) \left(1 - \frac{|l|}{n}\right) R(l) e^{-il\lambda}. \tag{6.166}$$

From (5.22) it follows that

$$\begin{aligned} M^q[\mathrm{E}\{\hat{f}_n(\lambda)\} - f(\lambda)] &= -\frac{M^q}{2\pi} \sum_{|l|>M} R(l) e^{-il\lambda} \\ &\quad + \frac{M^q}{2\pi} \sum_{l=-M}^{M} \left\{ w\left(\frac{l}{M}\right) - 1 \right\} R(l) e^{-il\lambda} \\ &\quad - \frac{M^q}{2\pi n} \sum_{l=-M}^{M} w\left(\frac{l}{M}\right) |l| R(l) e^{-il\lambda} \\ &= (1) + (2) + (3), \quad \text{(say)}. \end{aligned} \tag{6.167}$$

From Assumption 6.7, we have

$$|(1)| \le \frac{1}{2\pi} \sum_{|l|>M} |l|^q |R(l)| \to 0, \quad \text{as } M \to \infty. \tag{6.168}$$

Since, for each l,

$$\frac{1 - w\left(\frac{l}{M}\right)}{\left|\frac{l}{M}\right|^q} \to \kappa_q \quad \text{(by (ii) of Assumption 6.8)},$$

it is seen from Assumption 6.7 that

$$(2) = -\frac{1}{2\pi}\sum_{l=-M}^{M}\frac{1-w\left(\frac{l}{M}\right)}{\left|\frac{l}{M}\right|^q}|l|^q R(l)e^{-il\lambda}$$
$$\to -\frac{\kappa_q}{2\pi}\sum_{l=-\infty}^{\infty}|l|^q R(l)e^{-il\lambda}, \quad \text{as } M\to\infty. \tag{6.169}$$

Evidently,

$$|(3)| \le \frac{M^q}{2\pi n}\sum_{l=-\infty}^{\infty}|l||R(l)| \to 0, \quad \text{as } n\to\infty, \tag{6.170}$$

under Assumptions 6.7 and 6.8. The assertion follows from (6.167)-(6.170).

Next we give an intuitive proof of (ii) for the case of $0<\lambda<\pi$. In what follows, the notation $\sim$ means that the main order terms of both sides are equivalent. First, we have

$$\hat{f}_n(\lambda) \sim \frac{2\pi}{n}\sum_{s=-\left[\frac{n-1}{2}\right]}^{\left[\frac{n}{2}\right]} W_n\left(\lambda-\frac{2\pi s}{n}\right) I_n\left(\frac{2\pi s}{n}\right).$$

Hence,

$$\mathrm{Var}\{\hat{f}_n(\lambda)\} \sim \left(\frac{2\pi}{n}\right)^2 \sum_{s=-\left[\frac{n-1}{2}\right]}^{\left[\frac{n}{2}\right]}\sum_{t=-\left[\frac{n-1}{2}\right]}^{\left[\frac{n}{2}\right]} W_n\left(\lambda-\frac{2\pi s}{n}\right) W_n\left(\lambda-\frac{2\pi t}{n}\right)$$
$$\times \mathrm{Cov}\left\{I_n\left(\frac{2\pi s}{n}\right), I_n\left(\frac{2\pi t}{n}\right)\right\}, \tag{6.171}$$

which, together with Theorem 6.10, implies that the right-hand side of (6.171) is

$$\sim \left(\frac{2\pi}{n}\right)^2 \sum_{s=-\left[\frac{n-1}{2}\right]}^{\left[\frac{n}{2}\right]} W_n\left(\lambda-\frac{2\pi s}{n}\right)^2 \mathrm{Var}\left\{I_n\left(\frac{2\pi s}{n}\right)\right\}$$
$$+\left(\frac{2\pi}{n}\right)^2 \sum_{s=-\left[\frac{n-1}{2}\right]}^{\left[\frac{n}{2}\right]} W_n\left(\lambda-\frac{2\pi s}{n}\right) W_n\left(\lambda+\frac{2\pi s}{n}\right)$$
$$\times \mathrm{Cov}\left\{I_n\left(\frac{2\pi s}{n}\right), I_n\left(-\frac{2\pi s}{n}\right)\right\}$$
$$\sim \frac{2\pi}{n}\int_{-\pi}^{\pi} W_n(\lambda-\mu)^2 f(\mu)^2 d\mu + \frac{2\pi}{n}\int_{-\pi}^{\pi} W_n(\lambda-\mu)W_n(\lambda+\mu)f(\mu)^2 d\mu. \tag{6.172}$$

Let

$$W(\lambda) = \frac{1}{2\pi}\int_{-1}^{1} w(x)e^{-ix\lambda}dx. \tag{6.173}$$

Since $W_n(\lambda) \sim MW(M\lambda)$, the right-hand side of (6.172) is

$$\begin{aligned}
&\sim \frac{2\pi M^2}{n}\int_{-\pi}^{\pi} W\{M(\lambda-\mu)\}^2 f(\mu)^2 d\mu \\
&+\frac{2\pi M^2}{n}\int_{-\pi}^{\pi} W\{M(\lambda-\mu)\}W\{M(\lambda+\mu)\}f(\mu)^2 d\mu \\
&\qquad\qquad (\text{by } \rho = M(\lambda-\mu)) \\
&\sim \frac{2\pi M}{n}\int_{-\infty}^{\infty} W(\rho)^2 f\left(\lambda - \frac{\rho}{M}\right)^2 d\rho \\
&+\frac{2\pi M}{n}\int_{-\infty}^{\infty} W(\rho)W(2M\lambda-\rho) f\left(\lambda - \frac{\rho}{M}\right)^2 d\rho, \quad (0<\lambda<\pi).
\end{aligned} \tag{6.174}$$

By Parseval's equality (Theorem A.7), the first term of (6.174) is

$$\sim \frac{M}{n}\int_{-1}^{1} w(x)^2 dx \cdot f(\lambda)^2,$$

and, by the Riemann-Lebesgue theorem (Theorem A.8), the second term of (6.174) converges to zero as $M \to \infty$. Thus the assertion follows. □

From this theorem we observe that

$$\mathrm{E}\{\hat{f}_n(\lambda)\} - f(\lambda) = O(M^{-q}), \quad \mathrm{Var}\{\hat{f}_n(\lambda)\} = O\left(\frac{M}{n}\right), \tag{6.175}$$

which implies that $\hat{f}_n(\lambda)$ becomes a consistent estimator of $f(\lambda)$. The main order terms of (6.175) depend on the window function only through the quantities κ_q and $\int_{-1}^{1} w(x)^2 dx$. Therefore we understand that these quantities become indices showing "goodness" of the window function.

Now, let us see concrete examples of window functions.

Example 6.14 *(Bartlett window). This lag window function is defined by*

$$w(x) = \begin{cases} 1-|x|, & |x| \le 1, \\ 0, & |x| > 1. \end{cases}$$

In this case, $q = 1$, $\kappa_1 = 1$ *and* $\int_{-1}^{1} w(x)^2 dx = 2/3$.

Example 6.15 *(Hanning window). This lag window function is defined by*

$$w(x) = \begin{cases} \frac{1}{2}(1+\cos \pi x), & |x| \le 1, \\ 0, & |x| > 1. \end{cases}$$

In this case, $q = 2$, $\kappa_2 = \pi^2/4$ *and* $\int_{-1}^{1} w(x)^2 dx = 3/4$.

Example 6.16 *(Parzen window). This lag window function is defined by*

$$w(x) = \begin{cases} 1 - 6x^2 + 6|x|^3, & |x| \le 1/2, \\ 2(1 - |x|)^3, & 1/2 < |x| \le 1, \\ 0, & |x| > 1. \end{cases}$$

In this case, $q = 2$, $\kappa_2 = 6$ *and* $\int_{-1}^{1} w(x)^2 dx = 151/280$.

Example 6.17 *(Daniell window). This lag window function is defined by*

$$w(x) = \begin{cases} \frac{\sin \frac{\pi x}{2}}{\frac{\pi x}{2}}, & |x| \le 1, \\ 0, & |x| > 1. \end{cases} \tag{6.176}$$

In this case, $q = 2$, $\kappa_2 = \pi^2/6$ *and* $\int_{-1}^{1} w(x)^2 dx = 2$.
Here, recalling (6.162) we can see that the spectral window function of (6.176) is

$$W_n(\lambda) = \begin{cases} \frac{M}{\pi}, & |\lambda| \le \frac{\pi}{2M}, \\ 0, & \textit{otherwise.} \end{cases}$$

Therefore $\hat{f}_n(\lambda)$ *becomes the simple moving average* $\hat{s}_n(\lambda)$ *of the periodogram, which was given by (6.160).*

Example 6.18 (Akaike window) *This lag window function is defined by*

$$w(x) = \begin{cases} 0.6398 + 0.4802 \cos \pi x - 0.12 \cos 2\pi x, & |x| \le 1, \\ 0, & |x| > 1. \end{cases}$$

The calculation of q, κ_q *and* $\int_{-1}^{1} w(x)^2 dx$ *is left to the reader (Exercise 6.9).*

Although we saw typical window functions above, how should we choose the window function to construct a "good" spectral density estimator? Returning to Theorem 6.11, we rewrite (i) and (ii) as

$$\mathrm{E}\{\hat{f}_n(\lambda)\} - f(\lambda) = \frac{1}{M^q} \kappa_q b(\lambda) + o\left(\frac{1}{M^q}\right), \tag{6.177}$$

$$\mathrm{Var}\{\hat{f}_n(\lambda)\} = \frac{M}{n} f(\lambda)^2 \int_{-1}^{1} w(x)^2 dx + o\left(\frac{M}{n}\right). \tag{6.178}$$

Henceforth, assuming $0 < \lambda < \pi$, for simplicity, we proceed to our discussion. We measure "goodness" of $\hat{f}_n(\lambda)$ by the mean squares error:

$$\mathrm{E}[\{\hat{f}_n(\lambda) - f(\lambda)\}^2] = \mathrm{Var}\{\hat{f}_n(\lambda)\} + [\mathrm{E}\{\hat{f}_n(\lambda)\} - f(\lambda)]^2. \tag{6.179}$$

From (6.177) and (6.178) it follows that

$$\mathrm{Var}\{\hat{f}_n(\lambda)\} = O\left(\frac{M}{n}\right), \quad [\mathrm{E}\{\hat{f}_n(\lambda)\} - f(\lambda)]^2 = O\left(\frac{1}{M^{2q}}\right),$$

which implies that (6.179) is minimized if we choose the order of M so that $O(M/n) = O(1/M^{2q})$, i.e.,

$$M = cn^{\frac{1}{1+2q}}, \quad (c \text{ is a constant}). \tag{6.180}$$

For this M, from (6.177)-(6.179) we obtain

$$\begin{aligned}\lim_{n\to\infty} \mathrm{E}\left[\left\{n^{\frac{q}{1+2q}}(\hat{f}_n(\lambda) - f(\lambda))\right\}^2\right] \\ = cf(\lambda)^2 \int_{-1}^{1} w(x)^2 dx + c^{-2q}\kappa_q^2 b^2(\lambda).\end{aligned} \tag{6.181}$$

If a window function minimizes the right-hand side of (6.181) it is asymptotically optimal. However, the right-hand side of (6.181) depends on the "unknown parameter $f(\lambda)$", we cannot choose the asymptotically optimal window function generally.

The factor $n^{q/1+2q}$ in the left-hand side of (6.181) is called the *consistent order* of $\hat{f}_n(\lambda)$. If this order becomes larger, the estimator becomes better. In Section 6.2, we saw that the consistent order of parametric estimators of unknown parameter $\boldsymbol{\theta}$ is $\sqrt{n}$. Therefore, in the case of $\hat{f}_n(\lambda)$, the consistent order is lower than $\sqrt{n}$ for any $q \geq 1$. Hence, $\hat{f}_n(\lambda)$ is an inferior estimator in comparison with parametric ones. We can observe this feature from Figure 6.10.

So far we saw negative aspects of $\hat{f}_n(\lambda)$. But, are nonparametric estimators like this useless ? This is not true. For example, in econometric empirical analysis, it is very restrictive to assume finite-dimensional parametric models for the data. In such a case, nonparametric estimators become useful. Although $\hat{f}_n(\lambda)$ is not a good estimator of $f(\lambda)$ at each $\lambda \in [-\pi, \pi]$, if we think of the following "integral functional" of $\hat{f}_n(\lambda)$:

$$\int_{-\pi}^{\pi} \Phi\{\hat{f}_n(\lambda)\}d\lambda, \quad (\Phi(\cdot) \text{ is a smooth function}), \tag{6.182}$$

then we will show that it becomes a $\sqrt{n}$-consistent estimator of the corresponding quantity $\int_{-\pi}^{\pi} \Phi\{f(\lambda)\}d\lambda$. From this we can discuss the same consistent order asymptotics as in the parametric case for estimators based on the integral functional (6.182). Let us see this feature by Figure 6.10. We can grasp that the measure of area environed by the graph of $\hat{f}_n(\lambda)$, the vertical axis and the horizontal axis is approximately near to corresponding one by the graph of $f(\lambda)$, because the asperity of $\hat{f}_n(\lambda)$ is canceled around $f(\lambda)$. Since (6.182) is a sort of area measure of $\hat{f}_n(\lambda)$, it becomes a good estimator of the corresponding area measure of $f(\lambda)$, although $\hat{f}_n(\lambda)$ is not a good estimator of $f(\lambda)$ at each $\lambda \in [-\pi, \pi]$.

Now, recall Lemma 6.1. If we set $\Phi(x) = x$ in (6.182), then it follows from Lemma 6.1 that $\int_{-\pi}^{\pi} \Phi\{I_n(\lambda)\}d\lambda$ becomes a $\sqrt{n}$-consistent estimator of $\int_{-\pi}^{\pi} \Phi\{f(\lambda)\}d\lambda$. We can understand this feature by Figure 6.10 similarly as

above. However, if $\Phi(x)$ is not linear it is shown that $\int_{-\pi}^{\pi} \Phi\{I_n(\lambda)\}d\lambda$ is not a consistent estimator of $\int_{-\pi}^{\pi} \Phi\{f(\lambda)\}d\lambda$ (see Exercise 6.10). Therefore, if $\Phi(x)$ is nonlinear, it is essential to use $\hat{f}_n(\lambda)$ instead of $I_n(\lambda)$.

In what follows we will see the asymptotics of $\int_{-\pi}^{\pi} \Phi\{\hat{f}_n(\lambda)\}d\lambda$. Suppose that $\{X_t\}$ is a Gaussian stationary process with mean zero and covariance function $R(\cdot)$ satisfying Assumptions 6.7 and 6.8 with q=2. Let $M = M(n)$ satisfy

$$\frac{n^{1/4}}{M} + \frac{M}{\sqrt{n}} \to 0, \quad (n \to \infty),$$

and assume that $\Phi(x)$ is continuously three times differentiable on $(0, \infty)$. From Theorem 6.11 it follows that

$$E[\{\hat{f}_n(\lambda) - f(\lambda)\}^2] = O(M/n). \tag{6.183}$$

Expanding $\Phi(\cdot)$ around $f(\lambda)$ we obtain

$$\sqrt{n}\int_{-\pi}^{\pi} [\Phi\{\hat{f}_n(\lambda)\} - \Phi\{f(\lambda)\}]\, d\lambda \;\sim\; \sqrt{n}\int_{-\pi}^{\pi} \Phi^{(1)}\{f(\lambda)\}\{\hat{f}_n(\lambda) - f(\lambda)\}\, d\lambda, \tag{6.184}$$

where $\Phi^{(1)}(x) = \frac{d}{dx}\Phi(x)$. Then the right-hand side of (6.184) is

$$\begin{aligned}
&\sqrt{n}\int_{-\pi}^{\pi} \left[\Phi^{(1)}\{f(\lambda)\}\int_{-\pi}^{\pi} \{I_n(\mu) - f(\mu)\}W_n(\lambda - \mu)d\mu\right] d\lambda \\
&+ \sqrt{n}\int_{-\pi}^{\pi} \left[\Phi^{(1)}\{f(\lambda)\}\left\{\int_{-\pi}^{\pi} f(\mu)W_n(\lambda - \mu)d\mu - f(\lambda)\right\}\right] d\lambda \\
&= (A1) + (A2), \quad \text{(say)}.
\end{aligned}$$

Since $W_n(\cdot)$ converges to the delta function as $n \to \infty$, we observe that

$$\begin{aligned}
(A1) &\sim \sqrt{n}\int_{-\pi}^{\pi} \Phi^{(1)}\{f(\lambda)\}\{I_n(\lambda) - f(\lambda)\}d\lambda, \\
(A2) &\sim 0,
\end{aligned}$$

which implies

$$\sqrt{n}\int_{-\pi}^{\pi} \left[\Phi\{\hat{f}_n(\lambda)\} - \Phi\{f(\lambda)\}\right] d\lambda \;\sim\; \sqrt{n}\int_{-\pi}^{\pi} \Phi^{(1)}\{f(\lambda)\}\{I_n(\lambda) - f(\lambda)\}d\lambda. \tag{6.185}$$

Therefore, recalling Lemma 6.1 we have the following theorem.

Theorem 6.12 *As $n \to \infty$,*

(i)

$$\int_{-\pi}^{\pi} \Phi\{\hat{f}_n(\lambda)\}d\lambda \xrightarrow{p} \int_{-\pi}^{\pi} \Phi\{f(\lambda)\}d\lambda,$$

(ii)

$$\sqrt{n}\int_{-\pi}^{\pi}\left[\Phi\{\hat{f}_n(\lambda)\}-\Phi\{f(\lambda)\}\right]d\lambda \xrightarrow{d} N\left(0, 4\pi\int_{-\pi}^{\pi}\left[\Phi^{(1)}\{f(\lambda)\}\right]^2 f(\lambda)^2 d\lambda\right).$$

From (6.183) it is seen that $\hat{f}_n(\lambda)$ is a $\sqrt{n/M}$-consistent estimator of $f(\lambda)$, hence this order is inferior to that of the parametric estimator (i.e. $\sqrt{n}$-order). However, Theorem 6.12 claims that integral functionals of $\hat{f}_n(\lambda)$ have $\sqrt{n}$-consistency. If we sloganize this phenomenon, we may say

"Integration recovers $\sqrt{n}$-consistency."

Although Theorem 6.12 can be applied to estimation, testing, discriminant analysis and the other various statistical methods, we just mention the estimation theory briefly.

Let $\{X_t\}$ be a Gaussian stationary process with mean zero and spectral density function $g(\lambda)$. For $g(\lambda)$ we fit a class of parametric spectral density models $f_{\boldsymbol{\theta}}(\lambda)$, $\boldsymbol{\theta} \in \Theta \subset \mathbf{R}^r$, and estimate the unknown parameter $\boldsymbol{\theta}$. To do so we use the following disparity measure between $f_{\boldsymbol{\theta}}$ and g;

$$D(f_{\boldsymbol{\theta}}, g) \equiv \int_{-\pi}^{\pi} K\left\{\frac{f_{\boldsymbol{\theta}}(\lambda)}{g(\lambda)}\right\} d\lambda, \tag{6.186}$$

where $K(x)$ is continuously three times differentiable on $(0, \infty)$, and has a unique minimum at $x = 1$. We present an important example of $K(\cdot)$.

Example 6.19 (α-entropy). For $\alpha \in (0, 1)$,

$$K(x) = \log\{(1-\alpha)+\alpha x\} - \alpha \log x. \tag{6.187}$$

Now we are going to estimate the unknown parameter $\boldsymbol{\theta}$. In this case we do not assume that the true model g belongs to a class of fitted models $f_{\boldsymbol{\theta}}$. Hence we define the *pseudo true value* $\underline{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ by

$$\underline{\boldsymbol{\theta}} \equiv \arg\min_{\boldsymbol{\theta}} D(f_{\boldsymbol{\theta}}, g), \tag{6.188}$$

which implies that $f_{\underline{\boldsymbol{\theta}}}$ is the nearest model to g in view of the criterion $D(f_{\boldsymbol{\theta}}, g)$. To estimate $\underline{\boldsymbol{\theta}}$, we need to construct a sample version of $D(f_{\boldsymbol{\theta}}, g)$. Since $g(\lambda)$ is unknown, we estimate it by the nonparametric estimator $\hat{g}_n(\lambda)$ introduced above, which suggests the following estimator

$$\hat{\boldsymbol{\theta}}_n \equiv \arg\min_{\boldsymbol{\theta}} D(f_{\boldsymbol{\theta}}, \hat{g}_n). \tag{6.189}$$

$\hat{\boldsymbol{\theta}}_n$ is called a *minimum contrast estimator* of $\underline{\boldsymbol{\theta}}$. Let

$$\begin{aligned} H_g &= \int_{-\pi}^{\pi} \left[\frac{1}{g(\lambda)^2} K^{(2)}\{f_{\boldsymbol{\theta}}(\lambda)/g(\lambda)\} \frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\lambda) \frac{\partial}{\partial \boldsymbol{\theta}'} f_{\boldsymbol{\theta}}(\lambda) \right. \\ &+ \left. \frac{1}{g(\lambda)} K^{(1)}\{f_{\boldsymbol{\theta}}(\lambda)/g(\lambda)\} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f_{\boldsymbol{\theta}}(\lambda) \right]_{\boldsymbol{\theta}=\underline{\boldsymbol{\theta}}} d\lambda \end{aligned}$$

and

$$\begin{aligned} \rho_g(\lambda) = -H_g^{-1} &\left[K^{(2)}\{f_{\boldsymbol{\theta}}(\lambda)/g(\lambda)\} \frac{f_{\boldsymbol{\theta}}(\lambda)}{g(\lambda)^3} + K^{(1)}\{f_{\boldsymbol{\theta}}(\lambda)/g(\lambda)\} \frac{1}{g(\lambda)^2} \right] \\ &\times \left[\frac{\partial}{\partial \boldsymbol{\theta}} f_{\boldsymbol{\theta}}(\lambda) \right]_{\boldsymbol{\theta}=\underline{\boldsymbol{\theta}}} \end{aligned}$$

where $K^{(l)} = \frac{\partial^l}{\partial x^l} K(x),\ l = 1, 2$. Then, recalling the arguments of the proof of Theorem 6.4 and using Theorem 6.12, we have the following results;

(i) $$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \underline{\boldsymbol{\theta}}\ ,$$

(ii) $$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \underline{\boldsymbol{\theta}}) \xrightarrow{d} N(\mathbf{0}, 4\pi \textstyle\int_{-\pi}^{\pi} \rho_g(\lambda)\rho_g(\lambda)' g(\lambda)^2 d\lambda).$$

If $g(\lambda) = f_{\boldsymbol{\theta}}(\lambda)$, then $\underline{\boldsymbol{\theta}} = \boldsymbol{\theta}$, and it is seen that (ii) becomes

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathcal{F}(\boldsymbol{\theta})^{-1}),$$

where $\mathcal{F}(\boldsymbol{\theta})$ is the Fisher information matrix given in (6.63). Hence $\hat{\boldsymbol{\theta}}_n$ is (Gaussian) asymptotically efficient. As we saw in Example 6.19, candidates of $K(\cdot)$ are infinitely many. Therefore we can make infinitely many asymptotically efficient estimators by the method which is essentially different from the MLE.

6.5 Prediction of Time Series

Prediction is one of the most important problems in time series analysis. Here we state the linear prediction problem of stationary processes, and also mention the nonlinear prediction for time series. Let $\{X_t : t \in \mathbf{Z}\}$ be a stationary process with mean zero and spectral density $g(\lambda)$. We write the spectral representation as

$$X_t = \int_{-\pi}^{\pi} e^{-it\lambda} dZ(\lambda), \qquad E\{|dZ(\lambda)|^2\} = g(\lambda)\, d\lambda, \tag{6.190}$$

and assume that

$$\int_{-\pi}^{\pi} \log g(\lambda)\, d\lambda > -\infty,$$

and $g(\lambda)$ is expressed as $g(\lambda) = (1/2\pi)|A_g(e^{i\lambda})|^2$, where $A_g(z) = \sum_{j=0}^{\infty} a_j^{(g)} z^j$, $z \in \mathbf{C}$. Suppose that $X_{t-1}, X_{t-2}, \cdots$ are observable. Then we predict X_t by

the linear combination

$$\hat{X}_t = \sum_{j\geq 1} b_j X_{t-j}. \tag{6.191}$$

We call $\hat{X}_t$ the *linear predictor* of X_t. Naturally, we are interested in a "good" predictor. For this we seek the linear predictor which minimizes $E\{|X_t - \hat{X}_t|^2\}$, and call it the *best linear predictor*. The best linear predictor of X_t is given by

$$\hat{X}_t^{best} = \int_{-\pi}^{\pi} e^{-it\lambda} \frac{A_g(e^{i\lambda}) - A_g(0)}{A_g(e^{i\lambda})} dZ(\lambda). \tag{6.192}$$

To see this, let $\hat{X}_t$ be an arbitrary linear predictor of X_t given by (6.191). Its spectral representation is

$$\hat{X}_t = \int_{-\pi}^{\pi} e^{-it\lambda} B(\lambda) dZ(\lambda), \quad (\ B(\lambda) = \sum_{j\geq 1} b_j e^{ij\lambda}\ ). \tag{6.193}$$

Then

$$\begin{aligned} E[\{X_t - \hat{X}_t\}^2] &= E[\{X_t - \hat{X}_t^{best} + \hat{X}_t^{best} - \hat{X}_t\}^2] \\ &= E[\{X_t - \hat{X}_t^{best}\}^2] + 2E[\{X_t - \hat{X}_t^{best}\}\overline{\{\hat{X}_t^{best} - \hat{X}_t\}}] \\ &\quad + \ E[\{\hat{X}_t^{best} - \hat{X}_t\}^2] \\ &= (1) + (2) + (3), \quad \text{(say)}. \end{aligned} \tag{6.194}$$

From (6.190), (6.192) and (6.193) it follows that

$$\begin{aligned} (2) \ &= \ 2E\left[\int_{-\pi}^{\pi} e^{-it\lambda} \frac{A_g(0)}{A_g(e^{i\lambda})} dZ(\lambda) \right. \\ &\quad \left. \times \ \overline{\int_{-\pi}^{\pi} e^{-it\lambda} \left\{ \frac{A_g(e^{i\lambda}) - A_g(0)}{A_g(e^{i\lambda})} - B(\lambda) \right\} dZ(\lambda)} \right] \\ &= \ 2\int_{-\pi}^{\pi} A_g(0)\overline{\Gamma(\lambda)} \frac{g(\lambda)}{A_g(e^{i\lambda})}\, d\lambda \quad \text{(by Theorem 5.3)} \\ &= \ \frac{A_g(0)}{\pi} \int_{-\pi}^{\pi} \overline{\Gamma(\lambda) A_g(e^{i\lambda})}\, d\lambda \end{aligned} \tag{6.195}$$

where $\Gamma(\lambda) = \{A_g(e^{i\lambda}) - A_g(0)\}/A_g(e^{i\lambda}) - B(\lambda)$. Here it is seen that $\overline{\Gamma}(\lambda)$ is expressed as a linear combination of $\{e^{-i\lambda}, e^{-2i\lambda}, e^{-3i\lambda}, \ldots\}$, and that $\overline{A_g(e^{i\lambda})}$ is expressed as a linear combination of $\{1, e^{-i\lambda}, e^{-2i\lambda}, \ldots\}$. Hence (6.195) $= 0$. Therefore, returning to (6.194) we observe that (6.194) is minimized if $\hat{X}_t = \hat{X}_t^{best}$ *a.s.*, which implies that $\hat{X}_t^{best}$ is the best linear predictor of X_t.

Let $\{X_t\}$ be the AR(p) model

$$X_t + b_1 X_{t-1} + \cdots + b_p X_{t-p} = u_t, \qquad (\ \{u_t\} \sim i.i.d.\ (0, \sigma^2)\ )$$

satisfying Assumption 6.1. Then $A_g(e^{i\lambda}) = (\sum_{j=0}^{p} b_j e^{ij\lambda})^{-1}$, $(b_0 = 1)$, and (6.192) is

$$\begin{aligned} \hat{X}_t^{best} &= \int_{-\pi}^{\pi} e^{-it\lambda} \left(1 - \sum_{j=0}^{p} b_j e^{ij\lambda}\right) dZ(\lambda) \\ &= -b_1 X_{t-1} - \cdots - b_p X_{t-p}. \end{aligned} \tag{6.196}$$

Thus, in the case of AR(p), the best linear predictor $\hat{X}_t^{best}$ has explicit form with finite length. However, if not AR(p), e.g., ARMA(p, q) etc., the $\hat{X}_t^{best}$ does not have explicit form with finite length generally.

If the spectral density $g(\lambda)$ of $\{X_t\}$ is completely specified, $\hat{X}_t^{best}$ is constructed by (6.192). But, in the actual situation, the model selection and estimation for $g(\lambda)$ are needed. Hence the true spectral density is likely to be misspecified. This leads us to a misspecified prediction problem when a conjectured spectral density

$$f(\lambda) = \frac{1}{2\pi} |A_f(e^{i\lambda})|^2, \tag{6.197}$$

is fitted to $g(\lambda)$. From (6.192) the best linear predictor which is computed on the basis of this conjectured spectral density $f(\lambda)$ is given by

$$\hat{X}_t^f = \int_{-\pi}^{\pi} e^{-it\lambda} \frac{A_f(e^{i\lambda}) - A_f(0)}{A_f(e^{i\lambda})} dZ(\lambda). \tag{6.198}$$

The prediction error is

$$\begin{aligned} E[\{X_t - \hat{X}_t^f\}^2] &= E\left[\left| \int_{-\pi}^{\pi} e^{-it\lambda} \frac{A_f(0)}{A_f(e^{i\lambda})} dZ(\lambda) \right|^2\right] \\ &= \int_{-\pi}^{\pi} \frac{|A_f(0)|^2}{|A_f(e^{i\lambda})|^2} g(\lambda)\, d\lambda \quad \text{(by Theorem 5.3)} \\ &= \frac{|A_f(0)|^2}{2\pi} \int_{-\pi}^{\pi} \frac{g(\lambda)}{f(\lambda)}\, d\lambda \quad \text{(by (6.197))} \end{aligned} \tag{6.199}$$

It may be noted that Grenander and Rosenblatt (1957, Chap.8) first evaluated the misspecified prediction error (6.199). To recognize the importance of the misspecified prediction problem, suppose that $g(\lambda) = (2\pi)^{-1}|1 - 0.3e^{i\lambda}|^2$, and $f(\lambda) = (2\pi)^{-1}|1 - (0.3 + \theta)e^{i\lambda} + 0.3\,\theta\, e^{2i\lambda}|^2$, $|\theta| < 1$. In this case it is seen that $\int_{-\pi}^{\pi} \{g(\lambda)/f(\lambda)\} d\lambda = 2\pi/(1 - \theta^2)$, which means the prediction error (6.199) $\to \infty$ as $|\theta| \nearrow 1$ (see Exercise 6.11).

The setting of misspecification is very convenient, and can be applied to the problem of h-step ahead prediction, i.e., prediction of X_{t+h} based on linear combinations of $X_t, X_{t-1}, X_{t-2}, \ldots$ This problem can be grasped by fitting

the misspecified spectral density

$$f_{\boldsymbol{\theta}}(\lambda) = \frac{1}{2\pi} \frac{1}{|1 - \theta_1 e^{ih\lambda} - \theta_2 e^{i(h+1)\lambda} - \theta_3 e^{i(h+2)\lambda} - \cdots|^2}, \quad (\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)') \ , \tag{6.200}$$

to the true one $g(\lambda)$. The best h-step ahead linear predictor of X_{t+h} is of the form

$$\theta_1 X_t + \theta_2 X_{t-1} + \theta_3 X_{t-2} + \cdots \ , \tag{6.201}$$

and the coefficient vector $\boldsymbol{\theta}$ is determined by the value which minimizes the misspecified prediction error (6.199):

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{g(\lambda)}{f_\theta(\lambda)} \, d\lambda, \tag{6.202}$$

with respect to $\boldsymbol{\theta}$. Since $\boldsymbol{\theta}$ is actually unknown, we estimate $\boldsymbol{\theta}$ by the QGML $\hat{\boldsymbol{\theta}}_{QGML} = (\hat{\theta}_{1,QGML}, \hat{\theta}_{2,QGML}, \ldots)'$. Then the estimated predictor of (6.201) is given by

$$\hat{\theta}_{1,QGML} X_t + \hat{\theta}_{2,QGML} X_{t-1} + \hat{\theta}_{3,QGML} X_{t-2} + \cdots \ . \tag{6.203}$$

Choi and Taniguchi (2001) evaluated the prediction error of (6.203). Also, Choi and Taniguchi (2003) discussed the prediction problems of square-transformed stationary process X_t^2.

So far we discussed the prediction of X_t based on linear combinations of infinite stretch $X_{t-1}, X_{t-2}, \ldots$, generally. However, in actual problems, we are required to predict future values $X_{t+h}, \ h \in \mathbf{N}$ from an observed stretch $\mathbf{X}(1,t) \equiv \{X_1, X_2, \ldots, X_t\}$ of finite length. In what follows we discuss the problem of prediction for a future value X_{t+h} by a measurable function $\phi\{\mathbf{X}(1,t)\}$ of $\mathbf{X}(1,t)$. Here $\phi\{\cdot\}$ may be a nonlinear function. Therefore the problem becomes the h-step ahead nonlinear prediction problem. We seek $\phi = \phi_h(t)$ which minimizes

$$E[|X_{t+h} - \phi\{\mathbf{X}(1,t)\}|^2]. \tag{6.204}$$

$\phi_h(t)$ is called the *h-step ahead best predictor.*

Theorem 6.13 *The h-step ahead best predictor is given by*

$$\phi_h(t) = E\{X_{t+h}|\mathbf{X}(1,t)\}. \tag{6.205}$$

PROOF

For any given measurable function $f = f\{\mathbf{X}(1,t)\}$,

$$\begin{aligned} E[\{X_{t+h} - f\}^2] &= E[\{X_{t+h} - E(X_{t+h}|\mathbf{X}(1,t))\}^2] \\ &\quad + E[\{E(X_{t+h}|\mathbf{X}(1,t)) - f\}^2] \\ &\quad + 2E[\{E(X_{t+h}|\mathbf{X}(1,t)) - f\}\{X_{t+h} - E(X_{t+h}|\mathbf{X}(1,t))\}]. \end{aligned} \tag{6.206}$$

The last expectation in the right-hand side of (6.206) is

$$\begin{aligned} &EE[\{E(X_{t+h}|\mathbf{X}(1,t)) - f\}\{X_{t+h} - E(X_{t+h}|\mathbf{X}(1,t))\}|\mathbf{X}(1,t)] \\ &= E[\{E(X_{t+h}|\mathbf{X}(1,t)) - f\}E\{X_{t+h} - E(X_{t+h}|\mathbf{X}(1,t))|\mathbf{X}(1,t)\}] \\ &= E[\{E(X_{t+h}|\mathbf{X}(1,t)) - f\}\{E(X_{t+h}|\mathbf{X}(1,t)) - E(X_{t+h}|\mathbf{X}(1,t))\}] \\ &= 0, \end{aligned}$$

which, together with (6.206), implies

$$E[\{X_{t+h} - f\}^2] \geq E[\{X_{t+h} - E(X_{t+h}|\mathbf{X}(1,t))\}^2].$$

Hence the assertion follows. □

Henceforth, for h satisfying $-t < h \leq 0$, we set

$$\phi_h(t) = X_{t-|h|}. \tag{6.207}$$

Let $\{X_t\}$ be generated by the AR(p) process

$$X_t = -b_1 X_{t-1} - \cdots - b_p X_{t-p} + u_t, \quad \{u_t\} \sim i.i.d.\ (0, \sigma^2). \tag{6.208}$$

Here we assume Assumption 6.1. Replacing t in (6.208) by $t+h$ ($h > 0$), and taking $E\{\cdot|\mathbf{X}(1,t)\}$ of (6.208), we have

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

Since $E\{u_{t+h}|\mathbf{X}(1,t)\} = 0$, we obtain

$$\phi_h(t) = -b_1\phi_{h-1}(t) - \cdots - b_p\phi_{h-p}(t). \tag{6.209}$$

Let $h = 1$ in (6.209). Then

$$\phi_1(t) = -b_1 X_t - \cdots - b_p X_{t-p+1}, \tag{6.210}$$

which we saw in (6.196). From (6.207) and (6.209) we can get $\phi_2(t), \ldots, \phi_h(t)$, recursively. Since the coefficients $b_1, \ldots, b_p$ are unknown in actual problems, we use the QGML estimators $\hat{b}_1, \ldots, \hat{b}_p$ so that the estimated one-step ahead predictor is given by

$$\hat{\phi}_1(t) = -\hat{b}_1 X_t - \cdots - \hat{b}_p X_{t-p+1}. \tag{6.211}$$

Then, similarly as in the above, we get the estimated h-step ahead predictor $\hat{\phi}_h(t)$ from (6.209) recursively.

Next, let us see the behavior of $\hat{\phi}_h(t)$ for concrete time series models. Let $X_1, X_2, \ldots, X_{200}$ be generated by the AR(2) model

$$X_t = 0.7\, X_{t-1} - 0.21\, X_{t-2} + u_t, \quad \{u_t\} \sim i.i.d.\ N(0,1). \tag{6.212}$$

Assuming that $X_1, X_2, \ldots, X_{195}$ are observed, we predict $X_{196}, \ldots, X_{200}$ by $\hat{\phi}_1(195), \ldots, \hat{\phi}_5(195)$, respectively. Actually, we fit AR(p) model to $X_1, \ldots, X_{195}$ by AIC. AIC selected the AR(2) model with the coefficients

$$\hat{b}_1 = -0.6935, \qquad \hat{b}_2 = 0.1382.$$

Then, $\hat{\phi}_1(195) = 0.6935\, X_{195} - 0.1382\, X_{194}$, and from (6.209) we can calculate $\hat{\phi}_2(195), \ldots, \hat{\phi}_5(195)$ recursively. Figure 6.11 plots the values of $X_t,\ t = 1, \ldots, 200$, by real line, and plots $\hat{\phi}_1(195), \ldots, \hat{\phi}_5(195)$ for $t = 196, \ldots, 200$, by $\circ$.

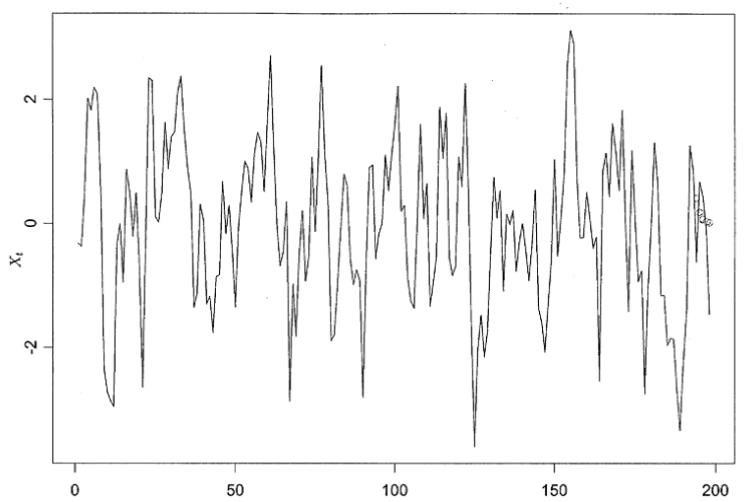


Figure 6.11 *Prediction plot.*

The best predictor $\phi_h(t)$ given in Theorem 6.13 is the optimal one among all the predictors including nonlinear ones. Therefore, regarding prediction we have only to seek $\phi_h(t)$. However, in general, since $\phi_h(t)$ is a complicated and implicit function of $\mathbf{X}(1,t)$, it is very difficult to seek $\phi_h(t)$ except for some special models. Furthermore we have to estimate the unknown parameters describing the predictor from observations.

Finally we consider a multistep prediction for the volatility of the ARCH(q) model, which is a typical nonlinear model, and is given by

$$\begin{cases} X_t = u_t \sigma_t \\ \sigma_t^2 = a_0 + \sum_{j=1}^{q} a_j X_{t-j}^2, \end{cases} \tag{6.213}$$

where $a_0 > 0$, $a_j \geq 0$, $j = 1, \ldots, q$ and $\{u_t\} \sim i.i.d.\ (0,1)$. When $\{\mathbf{X}(1,t)\}$ is observed, we are now interested in the prediction of σ_{t+h+1}^2. In this case the best predictor is given by $\phi_h(t) = E\{\sigma_{t+h+1}^2 | \mathbf{X}(1,t)\}$. Since $\sigma_{t+h+1}^2 = a_0 + \sum_{j=1}^{q} a_j X_{t+h+1-j}^2$ from (6.213), taking $E\{\cdot | \mathbf{X}(1,t)\}$ of both sides we

obtain

$$\phi_h(t) = a_0 + \sum_{j=1}^{q} a_j E\{X_{t+h+1-j}^2 | \mathbf{X}(1,t)\}$$

$$= \begin{cases} a_0 + \sum_{j=1}^{h} a_j E\{\sigma_{t+h+1-j}^2 | \mathbf{X}(1,t)\} + \sum_{j=h+1}^{q} a_j X_{t+h+1-j}^2, & \text{if } h < q, \\ a_0 + \sum_{j=1}^{q} a_j E\{\sigma_{t+h+1-j}^2 | \mathbf{X}(1,t)\}, & \text{if } h \geq q. \end{cases}$$

$$= \begin{cases} a_0 + \sum_{j=1}^{h} a_j \phi_{h-j}(t) + \sum_{j=h+1}^{q} a_j X_{t+h+1-j}^2, & \text{if } h < q, \\ a_0 + \sum_{j=1}^{q} a_j \phi_{h-j}(t), & \text{if } h \geq q. \end{cases}$$

Hence we can recursively seek $\phi_1(t), \ldots, \phi_h(t)$ from $\phi_0(t)$. Since they contain the unknown parameters $a_0, \ldots, a_q$, we estimate them by QMLE. Then we get the estimated predictor $\hat{\phi}_h(t)$. Fitting the ARCH(1) model to the daily stock returns of AMOCO company, Figure 6.12 plots $\hat{\sigma}_t$ by real line. Here the estimated values for a_0 and a_1 are $\hat{a}_0 = 0.00015$ and $\hat{a}_1 = 0.04648$, respectively. Assume that the values of $\hat{\sigma}_t$ plotted by real line are observed. Then, by use of the above method we calculate the 5-step ahead predictors $\hat{\phi}_1(t), \ldots, \hat{\phi}_5(t)$, and plot them by $\circ$ in Figure 6.12.

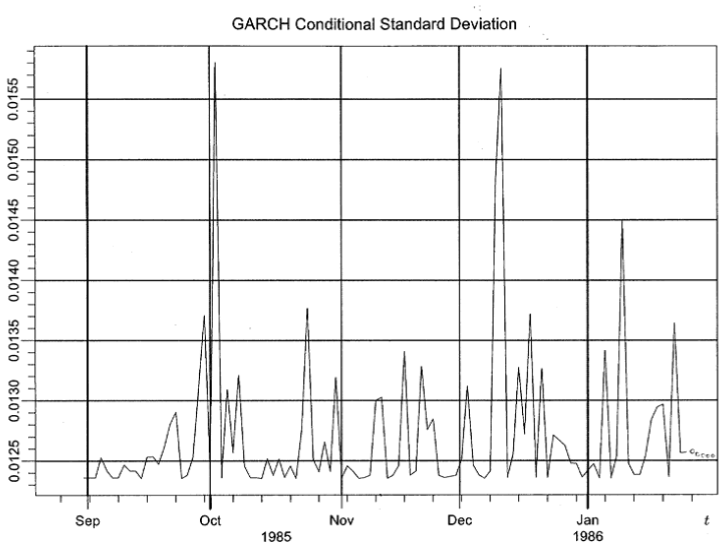


Figure 6.12 *Prediction of volatility.*

6.6 Regression for Time Series

So far we have assumed that the stochastic processes concerned have zero-means. However, if we want to apply the results to actual problems, it is natural to assume that the models have nonzero-mean function depending on time t. First, consider the following stochastic model

$$Y_t = T(t) + X_t, \tag{6.214}$$

where $\{X_t\}$ is a stationary process with mean zero spectral density $f(\lambda)$, and $T(t)$ is a nonrandom function of t. Then, $E(Y_t) = T(t)$, and $T(t)$ is called the *trend function* of $\{Y_t\}$. Since this description is indecisive, henceforth, we suppose that $T(t)$ is expressed as

$$T(t) = \mathbf{z}_t'\boldsymbol{\beta}, \tag{6.215}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is an unknown parameter vector, and $\mathbf{z}_t = (z_{t1}, \ldots, z_{tp})'$ is a known nonrandom function. That is, we consider the linear regression model

$$Y_t = \mathbf{z}_t'\boldsymbol{\beta} + X_t, \quad (t \in \mathbf{N}) \tag{6.216}$$

and discuss estimation of $\boldsymbol{\beta}$ based on an observed stretch $\mathbf{Y}_t = (Y_1, \ldots, Y_n)'$. $\{\mathbf{z_t}\}$ is called a *regressor function*. Let

$$\begin{aligned} a_{jk}^{(n)}(h) &= \sum_{t=1}^{n-h} z_{t+h,j} z_{tk}, \quad (h = 0, 1, \ldots) \\ &= \sum_{t=1-h}^{n} z_{t+h,j} z_{tk}, \quad (h = -1, -2, \ldots) \end{aligned}$$

We assume Grenander's conditions for $\{\mathbf{z}_t\}$.

Assumption 6.9 (Grenander's conditions)

(*G*1) $a_{jj}^{(n)}(0) \to \infty$ $(n \to \infty)$ $(j = 1, \ldots, p)$.

(*G*2) $\lim_{n\to\infty} \dfrac{z_{n+1,j}^2}{a_{jj}^{(n)}(0)} = 0, \quad (j = 1, \ldots, p).$

(*G*3) *The limit*

$$\lim_{n\to\infty} \frac{a_{jk}^{(n)}(h)}{\sqrt{a_{jj}^{(n)}(0) a_{kk}^{(n)}(0)}} = \rho_{jk}(h).$$

exists for $j, k = 1, \ldots, p,\ h \in \mathbf{Z}$.

(*G*4) *The* $p \times p$ *matrix* $\Phi(0) \equiv \{\rho_{jk}(0) : j, k = 1, \ldots, p\}$ *is regular.*

The point of $(G1)$ is to ensure the consistency of the least squares estimator of $\boldsymbol{\beta}$, and the point of $(G2)$ is to prevent the last $z_{n+1,j}^2$ from being an appreciable part of the sum of squares for large n. The third assumption $(G3)$ is that the correlations between regressors for all sufficiently large n are approximately fixed values. The fourth assumption $(G4)$ is for avoidance of multicollineality of the model.

Let $\Phi(h) \equiv \{\rho_{jk}(h) : j,k = 1,\ldots,p\}$. Then there exists a Hermitian matrix function $\mathbf{M}(\lambda) = \{M_{jk}(\lambda) : j,k = 1,\ldots,p\}$ with positive semidefinite increments such that

$$\Phi(h) = \int_{-\pi}^{\pi} e^{ih\lambda} d\mathbf{M}(\lambda). \tag{6.217}$$

$\mathbf{M}(\lambda)$ is called the *regression spectral measure* of $\{\mathbf{z}_t\}$. We may understand the substantial of $dM_{jk}(\lambda)$ as the limit of

$$dM_{jk}^{(n)}(\lambda) = \{a_{jj}^{(n)}(0)a_{kk}^{(n)}(0)\}^{-1/2} \left(\sum_{t=1}^{n} z_{tj}e^{-it\lambda}\right)\left(\sum_{t=1}^{n} z_{tj}e^{it\lambda}\right) d\lambda. \tag{6.218}$$

Let us see examples of $\{\mathbf{z}_t\}$.

Example 6.20 (i) *(Polynomial trend).*

Let

$$z_{tj} = t^{j-1}, \quad j = 1,\ldots,p,$$

then it is seen that

$$\rho_{jk}(h) = \frac{\sqrt{(2j-1)(2k-1)}}{j+k-1}, \quad (j,k = 1,\ldots,p,\ h = 0,\pm1,\ldots), \tag{6.219}$$

which do not depend on h. Hence, $\mathbf{M}(\lambda)$ *has only a jump at* $\lambda = 0$ *of*

$$\mathbf{M}_0 = \left\{ \frac{\sqrt{(2j-1)(2k-1)}}{j+k-1} : j,k = 1,\ldots,p \right\}, \tag{6.220}$$

(see Exercise 6.12).

(ii) *(Harmonic trend)*

Let

$$z_{tj} = \cos\nu_j t, \quad (0 < \nu_1 < \cdots < \nu_p < \pi),$$

then we obtain

$$\lim_{n\to\infty} n^{-1} \sum_{t=1}^{n-h} \cos\nu t \cos\lambda(t+h) = \begin{cases} \frac{1}{2}\cos\nu t, & (0 < \nu = \lambda < \pi) \\ 0, & (0 \le \nu \ne \lambda \le \pi), \end{cases} \tag{6.221}$$

which implies

$$\rho_{jk}(h) = \begin{cases} \cos\nu_j h, & (j = k), \\ 0, & (j \ne k). \end{cases} \tag{6.222}$$

Therefore $\mathbf{M}(\lambda)$ *has a jump* $\mathbf{M}_j = diag(0,\ldots,0,1/2,0,\ldots,0)$*,(1/2 is in the* j*th diagonal) at* $\lambda = \pm\nu_j$ *(Exercise 6.12).* □

Suppose that $\mathbf{Y} = (Y_1,\ldots,Y_n)'$ is generated by the regression model (6.216). Based on $\mathbf{Y}$ we estimate $\boldsymbol{\beta}$ by

$$\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y},$$
$$\hat{\boldsymbol{\beta}}_{BLU} = (\mathbf{Z}'\Sigma^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\Sigma^{-1}\mathbf{Y},$$

where $\mathbf{Z}' = (\mathbf{z}_1,\ldots,\mathbf{z}_n)$ and $\Sigma = \left\{\int_{-\pi}^{\pi} e^{i(l-j)\lambda} f(\lambda)\,d\lambda : l,j = 1,\ldots,n\right\}$ ($n \times n$-matrix). $\hat{\boldsymbol{\beta}}_{LS}$ and $\hat{\boldsymbol{\beta}}_{BLU}$ are called the least squares estimator and the best linear unbiased estimator of $\boldsymbol{\beta}$, respectively. In view of the general linear regression theory, the covariance matrix of $\hat{\boldsymbol{\beta}}_{BLU}$ is smaller than that of $\hat{\boldsymbol{\beta}}_{LS}$, i.e., $\hat{\boldsymbol{\beta}}_{BLU}$ is better than $\hat{\boldsymbol{\beta}}_{LS}$. However, since $\hat{\boldsymbol{\beta}}_{BLU}$ contains the unknown covariance matrix Σ, it is not a feasible estimator. But $\hat{\boldsymbol{\beta}}_{LS}$ is feasible, and has a simple form. Therefore we investigate the goodness of $\hat{\boldsymbol{\beta}}_{LS}$ in comparison with $\hat{\boldsymbol{\beta}}_{BLU}$ as follows. Letting

$$D_n = diag\left\{\left(\sum_{t=1}^{n} z_{t1}^2\right)^{1/2},\ldots,\left(\sum_{t=1}^{n} z_{tp}^2\right)^{1/2}\right\},$$

we define an asymptotic efficiency of $\hat{\boldsymbol{\beta}}_{LS}$ with respect to $\hat{\boldsymbol{\beta}}_{BLU}$ by

$$e \equiv \lim_{n\to\infty} \frac{\det[D_n E\{(\hat{\boldsymbol{\beta}}_{BLU} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{BLU} - \boldsymbol{\beta})'\}D_n]}{\det[D_n E\{(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})'\}D_n]}. \tag{6.223}$$

If $e = 1$, then $\hat{\boldsymbol{\beta}}_{LS}$ is said to be *asymptotically efficient*. The following theorem provides a foundation for discussion on the asymptotic efficiency of $\hat{\boldsymbol{\beta}}_{LS}$.

Theorem 6.14 (Grenander and Rosenblatt (1957, Chap.7)) *In the regression model (6.216), suppose that the spectral density* $f(\lambda)$ *of* $\{X_t\}$ *is continuous and* $f(\lambda) > 0$ *on* $[-\pi,\pi]$*, and that* $\{\mathbf{z}_t\}$ *satisfies Assumption 6.9. Then, the following statements hold true.*

(i)

$$\lim_{n\to\infty} D_n E\{(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})'\}D_n = 2\pi\Phi(0)^{-1}\int_{-\pi}^{\pi} f(\lambda)d\mathbf{M}(\lambda)\Phi(0)^{-1}. \tag{6.224}$$

(ii)

$$\lim_{n\to\infty} D_n E\{(\hat{\boldsymbol{\beta}}_{BLU} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{BLU} - \boldsymbol{\beta})'\}D_n = 2\pi\left[\int_{-\pi}^{\pi} f(\lambda)d\mathbf{M}(\lambda)\right]^{-1}. \tag{6.225}$$

PROOF Since the rigorous proof is very complicated, we give a simple and heuristic one. First, note that

$$E\{(\hat{\boldsymbol{\beta}}_{LS}-\boldsymbol{\beta})((\hat{\boldsymbol{\beta}}_{LS}-\boldsymbol{\beta})'\}=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\Sigma\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}, \tag{6.226}$$

$$E\{(\hat{\boldsymbol{\beta}}_{BLU}-\boldsymbol{\beta})((\hat{\boldsymbol{\beta}}_{BLU}-\boldsymbol{\beta})'\}=(\mathbf{Z}'\Sigma^{-1}\mathbf{Z})^{-1}. \tag{6.227}$$

Let $\mathbf{U}=\{n^{-1/2}e^{i2\pi kj/n};j,k=1,\ldots,n\}$. Then, from (5.26), $\mathbf{U}$ becomes a unitary matrix. Since

$$\int_{-\pi}^{\pi}e^{i(l-j)\lambda}f(\lambda)\,d\lambda\sim\frac{2\pi}{n}\sum_{t=1}^{n}e^{i(l-j)\lambda_t}f(\lambda_t),\quad(\lambda_t=\frac{2\pi t}{n}),$$

we obtain an approximation relation

$$\Sigma\sim\mathbf{U}\begin{pmatrix}2\pi f(\lambda_1)&&\mathbf{0}\\&\ddots&\\\mathbf{0}&&2\pi f(\lambda_n)\end{pmatrix}\mathbf{U}^*. \tag{6.228}$$

From (6.218) and (6.228), (i) follows from the following relation

$$\begin{aligned}&D_nE\{(\hat{\boldsymbol{\beta}}_{LS}-\boldsymbol{\beta})((\hat{\boldsymbol{\beta}}_{LS}-\boldsymbol{\beta})'\}D_n=D_n(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\Sigma\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}D_n\\&=(D_n^{-1}\mathbf{Z}'\mathbf{Z}D_n^{-1})^{-1}D_n^{-1}\mathbf{Z}'\mathbf{U}\mathbf{U}^*\Sigma\mathbf{U}\mathbf{U}^*\mathbf{Z}D_n^{-1}(D_n^{-1}\mathbf{Z}'\mathbf{Z}D_n^{-1})^{-1}\\&\sim\Phi(0)^{-1}\left\{D_n^{-1}\mathbf{Z}'\mathbf{U}\begin{pmatrix}2\pi f(\lambda_1)&&\mathbf{0}\\&\ddots&\\\mathbf{0}&&2\pi f(\lambda_n)\end{pmatrix}\mathbf{U}^*\mathbf{Z}D_n^{-1}\right\}\Phi(0)^{-1}\\&\sim2\pi\Phi(0)^{-1}\int_{-\pi}^{\pi}f(\lambda)d\mathbf{M}(\lambda)\Phi(0)^{-1}.\end{aligned} \tag{6.229}$$

The proof of (ii) follows as did that of (i). □

To compare (6.224) with (6.225) we need the following lemma, which is due to Kholevo (1969).

Lemma 6.3 *Let $A(\lambda)$ and $B(\lambda)$ be, respectively, $(r\times s)$ and $(t\times s)$ matrix-valued functions, and let $g(\lambda)$ be positive on $[-\pi,\pi]$. Then the following inequality holds if all the integrals exist;*

$$\begin{aligned}&\int_{-\pi}^{\pi}A(\lambda)A(\lambda)'g(\lambda)\,d\lambda\\&\geq\left\{\int_{-\pi}^{\pi}A(\lambda)B(\lambda)'\,d\lambda\right\}\left\{\int_{-\pi}^{\pi}B(\lambda)B(\lambda)'g(\lambda)^{-1}d\lambda\right\}^{-1}\left\{\int_{-\pi}^{\pi}A(\lambda)B(\lambda)'\,d\lambda\right\}'.\end{aligned} \tag{6.230}$$

Here the matrix inequality $\{*\}\geq\{\cdot\}$ means that the matrix $\{*\}-\{\cdot\}$ is non-

negative definite. In (6.230), the equality holds if and only if there exists a $(r \times t)$ *constant matrix* C *such that*

$$g(\lambda)A(\lambda) + CB(\lambda) = 0, \quad a.e. \ \ \lambda \in [-\pi, \pi]. \tag{6.231}$$

PROOF

Let $\mathbf{u}$ and $\mathbf{v}$ be arbitrary $(r \times 1)$ and $(t \times 1)$ vectors, respectively. Then we get

$$\left[g(\lambda)^{1/2}A(\lambda)'\mathbf{u} + \frac{B(\lambda)'}{g(\lambda)}\mathbf{v}\right]' \left[g(\lambda)^{1/2}A(\lambda)'\mathbf{u} + \frac{B(\lambda)'}{g(\lambda)}\mathbf{v}\right] \geq 0. \tag{6.232}$$

Integration of (6.232) with respect to $\lambda \in [-\pi, \pi]$ yields

$$\mathbf{u}'X\mathbf{u} + \mathbf{u}'Y\mathbf{v} + \mathbf{v}'Y'\mathbf{u} + \mathbf{v}'Z\mathbf{v} \geq 0, \tag{6.233}$$

where

$$X = \int_{-\pi}^{\pi} A(\lambda)A(\lambda)'g(\lambda)\, d\lambda, \qquad Y = \int_{-\pi}^{\pi} A(\lambda)B(\lambda)'\, d\lambda,$$
$$Z = \int_{-\pi}^{\pi} B(\lambda)B(\lambda)'g(\lambda)^{-1} d\lambda.$$

Substituting $\mathbf{v} = -Z^{-1}Y'\mathbf{u}$ into (6.233) we obtain

$$\mathbf{u}'(X - YZ^{-1}Y')\mathbf{u} \geq 0,$$

which implies $X \geq YZ^{-1}Y'$. □

We can understand the following theorem very roughly if we set

$$g(\lambda) = f(\lambda), \quad A(\lambda)\sqrt{(d\lambda)} = \{d\mathbf{M}(\lambda)\}^{1/2}, \quad B(\lambda)\sqrt{(d\lambda)} = \{d\mathbf{M}(\lambda)\}^{1/2},$$

in (6.230).

Theorem 6.15 (Grenander and Rosenblatt (1957)) *A necessary and sufficient condition under the same assumptions as in Theorem 6.14 that* $\hat{\boldsymbol{\beta}}_{LS}$ *is asymptotically efficient is that* $\mathbf{M}(\lambda)$ *increases at not more than* p *values of* λ, $0 \leq \lambda \leq \pi$, *and the sum of the ranks of the increases in* $\mathbf{M}(\lambda)$ *is* p.

In view of this theorem, recalling Example 6.20 we can see that $\hat{\boldsymbol{\beta}}_{LS}$ is asymptotically efficient for natural classes of regressors, i.e., polynomials and harmonic functions.

It is possible to replace Σ in $\hat{\boldsymbol{\beta}}_{BLU} = \hat{\boldsymbol{\beta}}_{BLU}(\Sigma)$ by a consistent estimator $\hat{\Sigma}$, i.e., $\hat{\hat{\boldsymbol{\beta}}}_{BLU} = \hat{\boldsymbol{\beta}}_{BLU}(\hat{\Sigma})$. The asymptotics of $\hat{\hat{\boldsymbol{\beta}}}_{BLU}$ will be discussed in Section 6.10.

Finally we briefly mention the regression model (6.216) when the regressor $\{\mathbf{z}_t\}$ is a random process. For such a model we have the following fundamental results.

Theorem 6.16 *In the regression model (6.216), assume*

(i) $\{(\mathbf{z}_t', X_t)\}$ *is strictly stationary and ergodic,*

(ii) $E(\mathbf{z}_t X_t) = \mathbf{0}$ *and* $E|\mathbf{z}_{tj} X_t| < \infty,\ j = 1, \ldots, p,$

(iii) $\mathbf{M} \equiv E(\mathbf{z}_t \mathbf{z}_t')$ *exists, and is positive definite.*

Then

$$\hat{\boldsymbol{\beta}}_{LS} \overset{a.s.}{\to} \boldsymbol{\beta}.$$

PROOF

Note that

$$\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta} = \left(\frac{1}{n} \sum_{t=1}^{n} \mathbf{z}_t \mathbf{z}_t' \right)^{-1} \frac{1}{n} \sum_{t=1}^{n} \mathbf{z}_t X_t. \tag{6.234}$$

From (i) $\sim$ (iii) and Theorems 5.4 and 5.7 it follows that

$$\frac{1}{n} \sum_{t=1}^{n} \mathbf{z}_t \mathbf{z}_t' \overset{a.s.}{\to} \mathbf{M}, \qquad \frac{1}{n} \sum_{t=1}^{n} \mathbf{z}_t X_t \overset{a.s.}{\to} \mathbf{0},$$

which implies that the right-hand side of (6.234) converges to 0 almost surely. Hence the assertion follows. □

We can generalize the model (6.216) in a multivariate form, and discuss the asymptotic normality of $\sqrt{n}(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})$ (e.g., see White (2000)).

6.7 Long Memory Processes

As we saw in (6.3)-(6.5), if $\{X_t\}$ is generated by

$$X_t = -b_1 X_{t-1} + u_t, \quad (\ |b_1| < 1,\ \{u_t\} \sim i.i.d.\ (0, 1)\ ), \tag{6.235}$$

then X_t is expressed as

$$X_t = \sum_{j=0}^{\infty} (-b_1)^j u_{t-j},$$

and the autocorrelation function is

$$\rho(k) \equiv \frac{E\{X_t X_{t+k}\}}{E\{X_t^2\}} = b_1^{|k|}, \quad (k \in \mathbf{Z}), \tag{6.236}$$

(Exercise 6.14). Since $|b_1| < 1$, $\rho(k)$ converges to zero exponentially as $|k| \to \infty$. For general AR(p) and ARMA(p, q) models satisfying Assumption 6.1, it is seen that there exists c ($|c| < 1$) such that

$$\rho(k) = O\left(c^{|k|}\right), \tag{6.237}$$

(Exercise 6.14). For stationary processes $\{X_t\}$ satisfying (6.237), the autocovariance function $R(k) = E(X_t X_{t+k})$ fulfills

$$\sum_{k=-\infty}^{\infty} |R(k)| < \infty. \tag{6.238}$$

Henceforth, stationary processes whose covariance function satisfies (6.238) are called *short-memory processes*. Evidently the usual AR, MA, ARMA processes are short-memory.

However we often observe time series whose autocovariance function is presumed to converge to zero with power law decay satisfying

$$\sum_{k=-\infty}^{\infty} |R(k)| = \infty, \tag{6.239}$$

in many fields such as hydrology, economics, engineering, environmental science and physics. If stationary processes satisfy (6.239), we call them *long-memory processes* (or processes with long-range dependence). Hence, the rate of convergence for the autocovariance function of long-memory processes is slower than that of the usual AR, MA and ARMA processes.

Figure 6.13 plots the daily returns of the S&P 500 Index $\{X_t : t = 1, \ldots, 17054\}$ from January 4, 1928 to August 31, 1991. Figure 6.14 plots the sample autocorrelation function $\hat{\rho}(k)$ of the square-transformed data $Y_t = X_t^2$.

From Figure 6.14 we observe that the decay of $\hat{\rho}(k)$ is very weak. Therefore we may say that the process has long-range dependence.

The phenomenon of long-range dependence was known long before suitable statistical models were introduced. Hurst (1951) studied the records of water flows through the Nile and through other rivers, the price of wheat, and meteorological series such as rainfall, temperature, and so on. His empirical conclusion was that the range of the records shows long-range dependence corresponding to

$$R(k) = O\left(k^{2H-2}\right), \quad (\frac{1}{2} < H < 1). \tag{6.240}$$

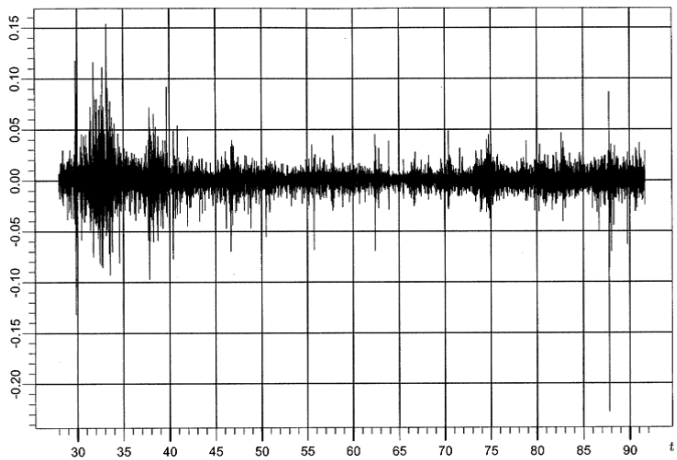


Figure 6.13 *Daily returns of the S&P 500 index.*

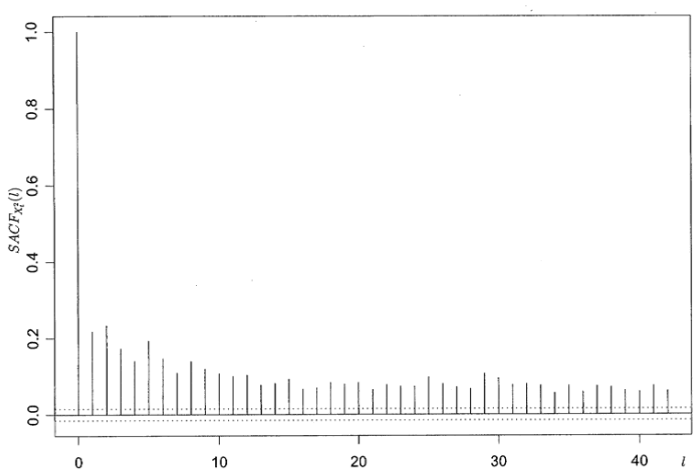


Figure 6.14 *Sample autocorrelation.*

The index H stems from Hurst's name, and phenomena satisfying $1/2 < H < 1$ are called Hurst phenomena. In what follows, we call stationary processes with autocovariance function (6.240) *long-memory processes*. The condition (6.240) can be equivalently paraphrased in terms of the spectral density $f(\lambda)$, which fulfills

$$f(\lambda) \sim O\left(|\lambda|^{-(2H-1)}\right), \quad (\lambda \to 0,\ 0 < 2H - 1 < 1), \tag{6.241}$$

(e.g., Zygmund (1959), Chap.V.2). Thus we may call stationary processes with spectral density (6.241) long-memory processes.

A concrete spectral density model satisfying (6.241) is given by

$$f_{\boldsymbol{\theta}}(\lambda) = \frac{\sigma^2}{2\pi}|1 - e^{e\lambda}|^{-2d}\frac{|\alpha(e^{i\lambda})|^2}{|\beta(e^{i\lambda})|^2}, \quad (0 < d < \frac{1}{2}), \tag{6.242}$$

where $\alpha(e^{i\lambda}) = \sum_{j=0}^{q} a_j e^{ij\lambda}$, $\beta(e^{ij\lambda}) = \sum_{j=0}^{p} b_j e^{ij\lambda}$ ($a_0 = 1$, $b_0 = 1$) and $\boldsymbol{\theta} = (a_1, \ldots, a_q, b_1, \ldots, b_p, \sigma^2, d)'$. Here it is assumed that $\alpha(z) \neq 0$ and $\beta(z) \neq 0$ for all $z \in \{z \in \mathbf{C} : |z| \leq 1\}$. Noting that $|1 - e^{i\lambda}| = 2\sin(\lambda/2)$ and $\{2\sin(\lambda/2)\}/\lambda \to 1$ ($\lambda \to 0$), we obtain

$$f_{\boldsymbol{\theta}}(\lambda) \sim \frac{\sigma^2}{2\pi}|\lambda|^{-2d}\frac{|\alpha(1)|^2}{|\beta(1)|^2}, \quad (\lambda \to 0), \tag{6.243}$$

which is a spectral density of a long-memory process. The relation between d and Hurst's H is

$$d = H - \frac{1}{2}. \tag{6.244}$$

If $\{X_t\}$ is a stationary process with spectral density (6.242), then it is expressed as

$$\beta(B)(1 - B)^d X_t = \alpha(B)u_t, \tag{6.245}$$

where B is the backward shift operator and $\{u_t\} \sim i.i.d.\ (0, \sigma^2)$. We call the $\{X_t\}$ an *autoregressive fractionally integrated moving average* (ARFIMA(p, d, q) or FARIMA(p, d, q)) process.

Let $\{X_t : t \in \mathbf{Z}\}$ be a Gaussian stationary process with mean μ and spectral density $f_{\boldsymbol{\theta}}(\lambda)$, $\lambda \in [-\pi, \pi]$, where μ and $\boldsymbol{\theta}$ ($\in \Theta \subset \mathbf{R}^q$) are unknown parameters which have to be estimated. We suppose that

$$f_{\boldsymbol{\theta}}(\lambda) \sim |\lambda|^{-\alpha(\boldsymbol{\theta})}L_{\boldsymbol{\theta}}(\lambda), \quad \lambda \to 0, \tag{6.246}$$

where $0 < \alpha(\boldsymbol{\theta}) < 1$ and $L_{\boldsymbol{\theta}}(\lambda)$ varies slowly at zero. Dahlhaus (1989) considered estimating $\boldsymbol{\theta}$ by the value $\hat{\boldsymbol{\theta}}_{QGML}$ that minimizes

$$\mathcal{L}_n^{(1)}(\boldsymbol{\theta}) \equiv \frac{1}{4\pi}\int_{-\pi}^{\pi}\left\{\log f_{\boldsymbol{\theta}}(\lambda) + \frac{\tilde{I}_n(\lambda)}{f_{\boldsymbol{\theta}}(\lambda)}\right\} d\lambda \tag{6.247}$$

with respect to $\boldsymbol{\theta} \in \Theta$, where $\tilde{I}_n(\lambda) = \frac{1}{2\pi n}|\sum_{t=1}^{n}(X_t - \bar{X}_n)|^2$ with $\bar{X}_n = n^{-1}\sum_{t=1}^{n} X_t$. $\mathcal{L}_n^{(1)}(\boldsymbol{\theta})$ is an approximation of $-n^{-1} \times \log$ (exact likelihood):

$$\mathcal{L}_n^{(2)}(\boldsymbol{\theta}) = \frac{1}{2\pi}\log\det T_n(f_{\boldsymbol{\theta}}) + \frac{1}{2\pi}(\mathbf{X}_n - \tilde{\mu}_n\mathbf{1})'T_n(f_{\boldsymbol{\theta}})^{-1}(\mathbf{X}_n - \tilde{\mu}\mathbf{1}), \tag{6.248}$$

where $\mathbf{X}_n = (X_1, \ldots, X_n)'$, $\mathbf{1} = (1, \ldots, 1)'$ and $T_n(f_{\boldsymbol{\theta}})$ is the $n \times n$ Toeplitz matrix of $f_{\boldsymbol{\theta}}$ whose (s,t)th element is $\int_{-\pi}^{\pi} f_{\boldsymbol{\theta}}(\lambda) \exp\{i\lambda(s-t)\} d\lambda$, and $\tilde{\mu}_n$ is a consistent estimator of μ (e.g., $\tilde{\mu}_n = \bar{X}_n$). An exact maximum likelihood estimator of $\boldsymbol{\theta}$ is given by the value $\hat{\boldsymbol{\theta}}_{ML}$ that minimizes $\mathcal{L}_n^{(2)}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta} \in \Theta$. The following theorem is due to Dahlhaus (1989).

Theorem 6.17 *Under appropriate regularity conditions on $f_{\boldsymbol{\theta}}$, the estimators $\hat{\boldsymbol{\theta}}_{QGML}$ and $\hat{\boldsymbol{\theta}}_{ML}$ are consistent, and for $* = QGML$ and ML,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_* - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \Gamma(\boldsymbol{\theta})^{-1}), \quad \text{as} \quad n \to \infty, \tag{6.249}$$

where

$$\Gamma(\boldsymbol{\theta}) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(\lambda) \, \frac{\partial}{\partial \boldsymbol{\theta}'} \log f_{\boldsymbol{\theta}}(\lambda) \, d\lambda.$$

Hence, $\hat{\boldsymbol{\theta}}_{QGML}$ and $\hat{\boldsymbol{\theta}}_{ML}$ are asymptotically Gaussian efficient.

It may be noted that the asymptotics of maximum likelihood type estimators for spectral parameter of long-memory processes are the same as those of short-memory processes (recall (6.62)).

In Theorem 6.17 we assumed that the process is Gaussian. Without Gaussianity of the process, assuming that the spectral density $f_{\boldsymbol{\theta}}(\lambda)$ satisfies

$$\int_{-\pi}^{\pi} \log f_{\boldsymbol{\theta}}(\lambda) \, d\lambda = 0, \tag{6.250}$$

Giraitis and Surgaillis (1990) showed that $\hat{\boldsymbol{\theta}}_{QGML}$ has the same asymptotics as in Theorem 6.17. Hosoya (1997) generalized the results above to the case when the processes concerned are multivariate, possibly non-Gaussian long-memory, and their spectral density matrices may have singularities not restricted at the origin. He showed that $\hat{\boldsymbol{\theta}}_{QGML}$ of a spectral parameter $\boldsymbol{\theta}$ satisfies

$$\hat{\boldsymbol{\theta}}_{QGML} \xrightarrow{p} \boldsymbol{\theta} \quad \text{and} \quad \sqrt{n}(\hat{\boldsymbol{\theta}}_{QGML} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, V),$$

where the asymptotic variance matrix V depends on non-Gaussianity of the process.

As the final topic of this section we discuss the estimation theory for regression models with long-memory disturbances. It will be seen that the asymptotics of estimators for the regression coefficients are different from those for short-memory disturbances.

Suppose that we observe $\mathbb{Y}^{(n)} = (Y_1, \ldots, Y_n)'$ generated by

$$Y_t = \mathbf{z}_t' \boldsymbol{\beta} + u_t, \quad t \in \mathbf{Z}, \tag{6.251}$$

where $\mathbf{z}_t = (z_{t1}, \ldots, z_{tq})'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$, and $\{u_t\}$ is generated by the following FARIMA (p_1, d, q_1) model

$$\sum_{k=0}^{p_1} \phi_k B^k (1-B)^d u_t = \sum_{k=0}^{q_1} \eta_k B^k \varepsilon_t, \quad t \in \mathbf{Z}, \quad \phi_0 = \eta_0 = 1, \tag{6.252}$$

where B is the backward shift operator and $\{\varepsilon_t\}$ is a sequence of *i.i.d.* $(0, \sigma^2)$ random variables with nonvanishing probability density $g = g(\cdot)$. Let

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p_1+q_1+1})' \equiv (d, \phi_1, \ldots, \phi_{p_1}, \eta_1, \ldots, \eta_{q_1})'.$$

Initially, we make the following assumption.

Assumption 6.10

(R1) *The characteristic polynomials*

$$\phi(z) \equiv \sum_{k=0}^{p_1} \phi_k z^k \quad \textit{and} \quad \eta(z) \equiv \sum_{k=0}^{q_1} \eta_k z^k$$

have no roots within the unit disc $D = \{z \in \mathbf{C} : |z| \le 1\}$.

(R2) $0 < d < 1/2$.

(R3) *The innovation probability density* g *is absolutely continuous with a.e. derivative* g' *satisfying*

$$0 < \mathcal{I}(g) \equiv \int \left\{ \frac{g'(z)}{g(z)} \right\}^2 g(z) dz < \infty$$

and

$$\int \left\{ \frac{g'(z)}{g(z)} \right\}^4 g(z) dz < \infty.$$

Under the assumption, $\psi(z) \equiv \phi(z)\eta(z)^{-1}(1-z)^d$ admits the absolutely convergent series

$$\psi(z) = \sum_{k=0}^{\infty} \psi_k z^k, \quad z \in D,$$

which implies that $\{u_t\}$ has the AR(∞) representation

$$\sum_{k=0}^{\infty} \psi_k u_{t-k} = \varepsilon_t, \quad t \in \mathbf{Z}. \tag{6.253}$$

Similarly , letting $\xi(z) \equiv \sum_{k=0}^{\infty} \xi_k z^k \equiv \{\psi(z)\}^{-1}, z \in D$, $\{u_t\}$ has the MA(∞) representation

$$u_t = \sum_{k=0}^{\infty} \xi_k \varepsilon_{t-k}, \quad t \in \mathbf{Z} \tag{6.254}$$

where $\{\xi_k\}$ is a square-summable sequence. Here $\{u_t\}$ has the spectral density

$$f_\theta(\lambda) = \frac{1}{2\pi|1-e^{i\lambda}|^{2d}} \frac{|\sum_{k=0}^{q_1} \eta_k e^{ik\lambda}|^2}{|\sum_{k=0}^{p_1} \phi_k e^{ik\lambda}|^2}. \tag{6.255}$$

The regressors $\mathbf{z}_t$'s are supposed to satisfy Assumption 6.9 (Grenander's conditions) with the regression spectral measure $\mathbf{M}(\lambda) = \{M_{jk}(\lambda)\}$ and $a_{jk}^{(n)}(h) = \sum_{t=1}^{n-h} z_{t+h,j} z_{tk}$. Further we assume,

Assumption 6.11 *For some $\delta > 1 - 2d$,*

$$\max_{1\le t\le n} \frac{z_{tj}^2}{a_{jj}^{(n)}(0)} = o(n^{-\delta}), \quad j = 1, \ldots, q.$$

Assumption 6.12 *1* $z_{tj} = t^{j-1}, j = 1, \ldots, l,\ 0 \le l \le m$, *hence*

$$M_{jj}(0+) - M_{jj}(0) = 1, \quad j = 1, \ldots, l$$

2 $0 < M_{jj}(0+) - M_{jj}(0) < 1, \quad j = l+1, \ldots, m$

3 $M_{jj}(0+) - M_{jj}(0) = 0, \quad j = m+1, \ldots, q.$

In what follows, we write

$$\tilde{D}_n = \operatorname{diag}\left(n^{-d}\sqrt{a_{11}^{(n)}(0)}, \ldots, n^{-d}\sqrt{a_{ll}^{(n)}(0)}, \sqrt{a_{l+1,l+1}^{(n)}(0)}, \ldots, \sqrt{a_{qq}^{(n)}(0)}\right),$$

$$W_1 = \left\{ \frac{2\pi(\sum_{k=0}^{p_1}\phi_k)^2\Gamma(j-d)\Gamma(k-d)\{(2j-1)(2k-1)\}^{1/2}}{(\sum_{k=0}^{p_1}\eta_k)^2\Gamma(j-2d)\Gamma(k-2d)(j+k-1-2d)}; j,k = 1,\ldots,l \right\},$$

and

$$W_2 = \left\{ \int_{-\pi}^{\pi} f_\theta(\lambda)^{-1} dM_{j+l,k+l}(\lambda);\ j,k = 1,\ldots,q-l \right\}.$$

Consider the local sequences

$$\boldsymbol{\theta}^{(n)} = \boldsymbol{\theta} + n^{-1/2}\mathbf{h}, \quad \boldsymbol{\beta}^{(n)} = \boldsymbol{\beta} + \tilde{D}_n^{-1}\mathbf{k} \tag{6.256}$$

where $\mathbf{h} \in \mathbf{R}^{p_1+q_1+1}$, $\mathbf{k} \in \mathbf{R}^q$, and $\mathbf{u} \equiv (\mathbf{h}',\ \mathbf{k}')'$ belongs to an open subset $\mathcal{H}$ of $\mathbf{R}^{q+p_1+q_1+1}$. We denote by $\{\psi_k^{(n)}\}$ and $\{\xi_k^{(n)}\}$ the sequences resulting from substituting $\boldsymbol{\theta}^{(n)}$ for $\boldsymbol{\theta}$ in the definition of ψ and ξ ((6.253) and (6.254)), respectively.

The sequence of statistical experiments is

$$\mathcal{E}_n = \left\{ \mathbf{R}^{\mathbf{Z}}, \mathcal{B}^{\mathbf{Z}}, \{P_{\boldsymbol{\theta},\boldsymbol{\beta}}^{(n)} | (\boldsymbol{\theta},\boldsymbol{\beta}) \in \mathbf{R}^{q+p_1+q_1+1}\} \right\}, \quad n \in \mathbf{N},$$

where $\mathcal{B}^{\mathbf{Z}}$ denotes the Borel σ-field on $\mathbf{R}^{\mathbf{Z}}$, and $P_{\boldsymbol{\theta},\boldsymbol{\beta}}^{(n)}$ the joint distribution of $\{\varepsilon_s : s \le 0; Y_1, \ldots, Y_n\}$ characterized by the parameter value $(\boldsymbol{\theta},\boldsymbol{\beta})$ and the

innovation density g. Denote by $H_g^{(n)}(\boldsymbol{\theta},\boldsymbol{\beta})$ the sequence of simple hypotheses $\left\{\{P_{\boldsymbol{\theta},\boldsymbol{\beta}}^{(n)}\}, n\in\mathbf{N}\right\}$. The log-likelihood ratio of $H_g^{(n)}(\boldsymbol{\theta}^{(n)},\boldsymbol{\beta}^{(n)})$ with respect to $H_g^{(n)}(\boldsymbol{\theta},\boldsymbol{\beta})$ is of the form (under $H_g^{(n)}(\boldsymbol{\theta},\boldsymbol{\beta})$),

$$\begin{aligned}\Lambda_g^{(n)}(\boldsymbol{\theta},\boldsymbol{\beta}) &\equiv \log\frac{dP^{(n)}_{\boldsymbol{\theta}^{(n)},\boldsymbol{\beta}^{(n)}}}{dP^{(n)}_{\boldsymbol{\theta},\boldsymbol{\beta}}}\\ &= \sum_{t=1}^{n}\Big[\log g\{\varepsilon_t+\sum_{\nu=0}^{t-1}\psi_\nu^{(n)}(Y_{t-\nu}-\mathbf{z}'_{t-\nu}\boldsymbol{\beta}^{(n)})\\ &\quad-\sum_{\nu=0}^{t-1}\psi_\nu(Y_{t-\nu}-\mathbf{z}'_{t-\nu}\boldsymbol{\beta})\\ &\quad+\sum_{\gamma=0}^{\infty}\sum_{\mu=0}^{\gamma}(\psi_{\mu+t}^{(n)}\xi_{\gamma-\mu}^{(n)}-\psi_{\mu+t}\xi_{\gamma-\mu})\varepsilon_{-\gamma}\}-\log g(\varepsilon_t)\Big] \qquad (6.257)\end{aligned}$$

The following LAN result is due to Hallin, Taniguchi, Serroukh and Choy (1999).

Theorem 6.18 *In the regression model (6.251), suppose that Assumptions 6.9 to 6.12 hold. Then the sequence of experiments $\mathcal{E}_n, n\in\mathbf{N}$, is locally asymptotically normal and equicontinuous on compact subsets $\mathcal{C}$ of $\mathcal{H}$. That is,*
(i) For all $\boldsymbol{\theta},\boldsymbol{\beta}$, the log-likelihood ratio (6.257) has, under $H_g^{(n)}(\boldsymbol{\theta},\boldsymbol{\beta})$, as $n\to\infty$, the stochastic expansion

$$\begin{aligned}\Lambda_g^{(n)}(\boldsymbol{\theta},\boldsymbol{\beta}) &= (\mathbf{h}',\mathbf{k}')\Delta_g^{(n)}(\boldsymbol{\theta},\boldsymbol{\beta})\\ &\quad-\frac{1}{2}\{\sigma^2\mathcal{I}(g)\mathbf{h}'Q(\boldsymbol{\theta})\mathbf{h}+\mathcal{I}(g)\mathbf{k}'W(\boldsymbol{\theta})\mathbf{k}\}+o_p(1), \qquad (6.258)\end{aligned}$$

with the $(q+p_1+q_1+1)$-dimensional central sequence

$$\Delta_g^{(n)}(\boldsymbol{\theta},\boldsymbol{\beta})\equiv n^{-1/2}\begin{pmatrix}\sum_{t=1}^{n}\frac{g'(\gamma_t)}{g(\gamma_t)}\sum_{\nu=1}^{t-1}\frac{\partial}{\partial\theta}(\psi_\nu)u_{t-\nu}\\ -\tilde{D}_n^{-1}\sum_{t=1}^{n}\frac{g'(\gamma_t)}{g(\gamma_t)}\sum_{\nu=0}^{t-1}\psi_\nu\mathbf{z}_{t-\nu}\end{pmatrix}, \qquad (6.259)$$

the $(p_1+q_1+1)\times(p_1+q_1+1)$ matrix

$$Q(\boldsymbol{\theta})=(4\pi)^{-1}\int_{\pi}^{\pi}\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{\theta}}(\lambda)\frac{\partial}{\partial\boldsymbol{\theta}'}\log f_{\boldsymbol{\theta}}(\lambda)d\lambda \qquad (6.260)$$

and the $q\times q$ matrix

$$W(\boldsymbol{\theta})=(2\pi)^{-1}\begin{pmatrix}W_1 & \mathbf{0}\\ \mathbf{0} & W_2\end{pmatrix}.$$

Here $\gamma_t=\gamma_t(\theta,\boldsymbol{\beta}), t=1,\ldots,n$, stand for the approximate residual

$$\gamma_t(\boldsymbol{\theta},\boldsymbol{\beta})\equiv\sum_{k=0}^{t-1}\psi_k(Y_{t-k}-z'_{t-k}\boldsymbol{\beta}).$$

(ii) Under $H_g^{(n)}(\boldsymbol{\theta}, \boldsymbol{\beta})$*, as* $n \to \infty$*,*

$$\Delta_g^{(n)}(\boldsymbol{\theta}, \boldsymbol{\beta}) \quad \xrightarrow{d} \quad N(\mathbf{0}, \Gamma_g(\boldsymbol{\theta}, \boldsymbol{\beta})), \tag{6.261}$$

where

$$\Gamma_g(\boldsymbol{\theta}, \boldsymbol{\beta}) \equiv \begin{pmatrix} \sigma^2 \mathcal{I}(g) Q(\boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{0} & \mathcal{I}(g) Q(\boldsymbol{\theta}) \end{pmatrix}.$$

(iii) For all $n \in \mathbf{N}$ *and all* $(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathcal{H}$*, the mapping*

$$\mathbf{u} = (\mathbf{h}, \mathbf{k}) \quad \longrightarrow \quad P^{(n)}_{\theta^{(n)}, \beta^{(n)}}$$

is continuous with respect to the variational distance.

As mentioned in Section 6.2, this theorem is the key result for virtually all problems in asymptotic inference and testing connected with (6.251). Here we just illustrated an application of the testing problem.

Suppose that the regression model (6.251) satisfies Assumptions 6.9-6.12, and that the innovation density g is Gaussian with $\sigma^2 = 1$. Further, we assume that m in Assumption 6.12 is equal to 0, and that the regression spectral measure $M(\lambda)$ increases at not more than q values of λ, $0 \le \lambda \le \pi$, and the sum of the ranks at the increases in $M(\lambda)$ is q. Then it follows from Theorem 6.15 that the LSE $\hat{\boldsymbol{\beta}}_{LSE}$ of $\boldsymbol{\beta}$ is asymptotically efficient. The MLE $\hat{\boldsymbol{\theta}}_{ML}$ of $\boldsymbol{\theta}$ is obtained by maximizing

$$l_n(\boldsymbol{\theta}) \equiv -\frac{1}{2} \log |\Sigma_n(\boldsymbol{\theta})| - \frac{1}{2} (\mathbb{Y}_n - \mathbb{Z}_n \hat{\boldsymbol{\beta}}_{LSE})' \Sigma_n^{-1}(\boldsymbol{\theta}) (\mathbb{Y}_n - \mathbb{Z}_n \hat{\boldsymbol{\beta}}_{LSE})$$

with respect to $\boldsymbol{\theta}$, where

$$\mathbb{Z}_n = \{z_{tj} : t = 1, \ldots, n, j = 1, \ldots, q\}, \quad (n \times q \text{ matrix})$$

and $\Sigma_n(\boldsymbol{\theta})$ is the covariance matrix of $(u_1, \ldots, u_n)'$.

Denote by $M(B)$ the linear space spanned by the columns of a matrix B. We are now interested in the hypothesis

$$H_0^{(n)} : \sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \in M(B_1) \quad \text{and} \quad \tilde{D}_n(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \in M(B_2) \tag{6.262}$$

for some given $(p_1 + q_1 + 1) \times (p_1 + q_1 + 1 - l_1)$ and $q \times (q - l_2)$ matrices B_1 and B_2 of full ranks, respectively. Here $\boldsymbol{\theta}_0 \in \mathbf{R}^{p_1 + q_1 + 1}$ and $\boldsymbol{\beta}_0 \in \mathbf{R}^q$ are given hypothetical vectors. Then the test rejecting $H_0^{(n)}$ whenever

$$\begin{aligned} \varphi_n^* =& n(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)' \left[Q(\hat{\boldsymbol{\theta}}_{ML}) - Q(\hat{\boldsymbol{\theta}}_{ML}) B_1 \{B_1' Q(\hat{\boldsymbol{\theta}}_{ML}) B_1\}^{-1} B_1' Q(\hat{\boldsymbol{\theta}}_{ML})' \right] \\ & \times (\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0) \\ & + (\hat{\boldsymbol{\beta}}_{LSE} - \boldsymbol{\beta}_0)' \tilde{D}_n \left\{ \frac{1}{2\pi} W_2 - \tilde{D}_n B_2 (B_2' \tilde{D}_n (2\pi)^{-1} W_2 \tilde{D}_n B_2)^{-1} B_2' \tilde{D}_n \right\} \tilde{D}_n \\ & \times (\hat{\boldsymbol{\beta}}_{LSE} - \boldsymbol{\beta}_0) \end{aligned}$$

exceeds the α-quantile $\chi^2_{l;\alpha}$ of a chi-square distribution with $l = l_1 + l_2$ degrees of freedom has asymptotic level α and is locally asymptotically optimal in the sense of Theorem 6.9.

6.8 Local Whittle Likelihood Approach

The study of spectral density functions of stationary processes provides an effective approach to estimate their underlying models and various methods for testing problems. There are two ways to estimate the spectral density directly. The one is the parametric method which is studied in the literature (e.g., Brockwell and Davis (1991)). Also Hosoya and Taniguchi (1982) constructed a very general framework of the Whittle estimator for a class of non-Gaussian linear processes. Further, Giraitis and Robinson (2001) proposed to use the Whittle estimation procedure for the squared ARCH processes. The other is the nonparametric method based on smoothed periodogram (e.g., Hannan (1970)).

For i.i.d. observations, Hjort and Jones (1996) proposed a new probability density estimator $f_{\hat{\theta}(x)}(x)$ which minimizes a local likelihood for f_θ around x. They showed that $f_{\hat{\theta}(x)}(x)$ has the same asymptotic variance as the ordinary nonparametric kernel estimator but potentially a smaller bias. For a Gaussian stationary process, Fan and Kreutzberger (1998) proposed a local polynomial estimator based on the Whittle likelihood. Then it was shown that it has advantages over the least-squares based on log-periodogram.

In this section, for a class of non-Gaussian linear processes, we introduce a local Whittle likelihood of the spectral density $f_\theta(\lambda)$ and propose the local Whittle estimator $f_{\hat{\theta}(\lambda)}(\lambda)$ around each frequency $\lambda \in [-\pi, \pi]$. Then we elucidate the asymptotics of $\hat{\theta}(\lambda)$ and $f_{\hat{\theta}(\lambda)}(\lambda)$.

Next we consider the problem of testing whether the spectral density of a class of stationary processes belongs to a parametric family or not. For this testing problem, Fan and Yao (2003, Section 9.3.2) and Fan and Zhang (2004) proposed generalized likelihood ratio tests based on the Whittle likelihood and a local Whittle estimator. Then they elucidated the asymptotics of the generalized likelihood ratio tests under the null hypothesis.

Because the results above rely on Gaussianity of the process concerned, here, we drop this assumption, and discuss the problem of testing whether the spectral density of a class of non-Gaussian linear processes belongs to a parametric family or not. A local Whittle likelihood ratio test is proposed. Then it is shown that the asymptotic distribution of the test converges in distribution to a normal distribution.

An interesting feature is that the asymptotics of the estimator and test statistic do not depend on non-Gaussianity of the process. Because we do not assume

Gaussianity of the process concerned, we can apply the results to stationary nonlinear processes which include GARCH processes. Some examples of a class of GARCH models are given. Numerical studies for the local Whittle likelihood ratio test are also provided. In what follows, we develop our discussion based on the results by Naito, Asai and Taniguchi (2006).

Assume that $\{z(n) : n \in \mathbf{Z}\}$ is a general linear process defined by

$$z(n) = \sum_{j=0}^{\infty} G(j)e(n-j), \quad n \in \mathbf{Z}, \tag{6.263}$$

where $\{e(n)\}$ is a white noise process satisfying

$$E[e(n)] = 0$$
$$E[e(n)e(m)] = \delta(m,n)\sigma^2, \quad E[e(n)^4] < \infty.$$

Furthermore, we assume that

$$\sum_{j=0}^{\infty} G(j)^2 < \infty. \tag{6.264}$$

Then $\{z(n)\}$ is a second-order stationary process with spectral density

$$f(w) = \frac{\sigma^2}{2\pi}\left|\sum_{j=0}^{\infty} G(j)e^{-iwj}\right|^2. \tag{6.265}$$

Henceforth we denote by $\mathcal{P}$ the set of all spectral density functions of the form (6.265). Write the autocovariance function of $z(n)$ as $\gamma(\cdot)$. Here we assume,

Assumption 6.13

$$\sum_{n=-\infty}^{\infty} n^2|\gamma(n)| < \infty. \tag{6.266}$$

Let $z(1), z(2), \ldots, z(N)$ be an observed stretch of $\{z(n)\}$, and denote the periodogram of $\{z(n)\}$ by

$$I_N(\lambda) = \frac{1}{2\pi N}\left|\sum_{n=1}^{N} z(n)e^{in\lambda}\right|^2.$$

Assumption 6.14 *Let $K(\cdot)$ be a kernel function which satisfies:*

(1) *K is a real bounded non-negative even function with a bounded support.*

(2) *$\int_{-\infty}^{\infty} K(t)dt = 1$, $\int_{-\infty}^{\infty} t^2K(t)dt = \kappa_2 < \infty$, $\int_{-\infty}^{\infty} t^jK^2(t)dt < \infty$ for $j = 0, 1, 2$.*

(3) *For $k(x) = \int_{-\infty}^{\infty} K(t)e^{itx}dt$, there exists an even integrable monotone decreasing function $\bar{k}(x)$ such that*

$$|k(x)| \le \bar{k}(x) \quad \text{on } [0, \infty).$$

Assumption 6.15 *For a bandwidth h decreasing as $N \to \infty$, we assume that*

$$\frac{1}{N^{\frac{1}{2}}h} + N^{\frac{1}{5}}h \to 0. \tag{6.267}$$

Let $\{f_\theta(\lambda) : \theta \in \Theta \subset R^q\}$ be a parametric family of spectral density functions of $\mathcal{P}$ where Θ is a compact set. Here we impose the following assumption.

Assumption 6.16 *$f_\theta(w)$ is three times continuously differentiable with respect to w and θ.*

We define a local distance function $D_\lambda(\cdot,\cdot)$ around a given local point λ for spectral densities $\{f_t\}$ by

$$D_\lambda(f_t, f) = \int_{-\pi}^{\pi} K_h(\lambda - w)\left\{\log f_t(w) + \frac{f(w)}{f_t(w)}\right\} dw,$$

where $K_h(x) = \frac{1}{h}K(x/h)$. If we replace f by I_N, we call $D_\lambda(f_\theta, I_N)$ the local Whittle likelihood function under f_θ.

Define $T_\lambda(f) \in \Theta$ by

$$D_\lambda(f_{T_\lambda(f)}, f) = \min_{t\in\theta} D_\lambda(f_t, f).$$

Henceforth we sometimes write $T_\lambda(f)$ as $\theta_0(\lambda)$ which is called a pseudo-true value of θ. As an estimator of $\theta_0(\lambda)$, we use $\hat{\theta}(\lambda)$ defined by

$$\hat{\theta}(\lambda) = T_\lambda(I_N) = \arg\min_{t\in\Theta} D_\lambda(f_t, I_N).$$

We can use $f_{\hat{\theta}(\lambda)}(\lambda)$ as a local estimator of f and call this the local Whittle likelihood estimator. For simplicity we sometimes write $f_{\theta(\lambda)}(\lambda)$ as $f_\theta(\lambda)$.

First, we investigate the asymptotics of $\hat{\theta}(\lambda)$ and $f_{\hat{\theta}}(\lambda)$. Fan and Kreutzberger (1998) showed the asymptotics of a local polynomial estimator of the spectral density based on the Whittle likelihood for Gaussian linear processes. We set down the following assumption.

Assumption 6.17 $\sum_{j_1,j_2,j_3=-\infty}^{\infty} |Q_e(j_1, j_2, j_3)| < \infty$ *where $Q_e(j_1, j_2, j_3)$ is the joint fourth-order cumulant of $e(n), e(n + j_1), e(n + j_2), e(n + j_3)$.*

Next we show the asymptotic distribution of $\hat{\theta}(\lambda)$ as $N \to \infty$.

Theorem 6.19 *Suppose that the $\{z(n)\}$ given in (6.263) and $K(\cdot)$ satisfy Assumptions 6.13 - 6.17, $T_\lambda(f)$ exists uniquely and lies in Int Θ, and that*

$$M(\lambda) = \frac{\partial^2}{\partial\theta\partial\theta'}(f_{\theta_0}^{-1}(\lambda)f(\lambda) + \log f_{\theta_0}(\lambda))$$

is a nonsingular matrix. Then if $N \to \infty$,

$$\sqrt{Nh}\left\{\hat{\theta}(\lambda) - \theta_0(\lambda) - \frac{1}{2}h^2\sigma_K^2 M(\lambda)^{-1}\frac{\partial^2}{\partial\lambda^2}\{f(\lambda)\frac{\partial}{\partial\theta}f_{\theta_0}^{-1}(\lambda)\}\right\}$$
$$\xrightarrow{d} N(0, M(\lambda)^{-1}\tilde{V}(M(\lambda)^{-1})'),$$

where

$$\tilde{V} = \begin{cases} 2\pi f^2(\lambda)\int_{-\infty}^{\infty} K^2(x)dx\frac{\partial}{\partial\theta}f_{\theta_0}^{-1}(\lambda)\frac{\partial}{\partial\theta'}f_{\theta_0}^{-1}(\lambda) & (\lambda \neq 0) \\ 4\pi f^2(\lambda)\int_{-\infty}^{\infty} K^2(x)dx\frac{\partial}{\partial\theta}f_{\theta_0}^{-1}(\lambda)\frac{\partial}{\partial\theta'}f_{\theta_0}^{-1}(\lambda) & (\lambda = 0) \end{cases}$$

and $\sigma_K^2 = \int t^2 K(t)dt$.

For a spectral density f, we say that a sequence of spectral densities $\{f_N\}$ convergences to f weakly if

$$\int_{-\pi}^{\pi}\psi(\omega)f_N(\omega)d\omega \to \int_{-\pi}^{\pi}\psi(\omega)f(\omega)d\omega \quad as\ N \to \infty$$

for every continuous function $\psi(w)$, and denote $f_N \xrightarrow{w} f$.

To prove the Theorem 6.19, we first show two lemmas.

Lemma 6.4 *Suppose that* f_θ *and* K *satisfy Assumption 6.16 and Assumption 6.14 respectively, and that* $T_\lambda(f)$ *exists uniquely and lies in Int* Θ, *and*

$$\left[\int_{-\pi}^{\pi} K_h(\lambda - t)\frac{\partial^2}{\partial\theta\partial\theta'}\left\{f_\theta^{-1}(t)f(t) + \log f_\theta(t)\right\}|_{\theta=T_\lambda(f)}dt\right]$$

is a nonsingular matrix for every $h > 0$. *Then for* $f_N \xrightarrow{w} f$, *it holds that*

$$\begin{aligned} &T_\lambda(f_N) - T_\lambda(f) \\ &\quad = -\left[\int_{-\pi}^{\pi} K_h(\lambda - t)\frac{\partial^2}{\partial\theta\partial\theta'}\left\{f_\theta^{-1}(t)f(t) + \log f_\theta(t)\right\}|_{\theta=T_\lambda(f)}dt\right]^{-1} \\ &\qquad \times \int_{-\pi}^{\pi} K_h(\lambda - t)\frac{\partial}{\partial\theta}f_\theta^{-1}(t)|_{\theta=T_\lambda(f)}\{f_N(t) - f(t)\}dt \\ &\qquad + \alpha_{N,h}\int_{-\pi}^{\pi} K_h(\lambda - t)\frac{\partial}{\partial\theta}f_\theta^{-1}(t)|_{\theta=T_\lambda(f)}\{f_N(t) - f(t)\}dt \end{aligned}$$

as $N \to \infty$, *where* $\alpha_{N,h} \to 0,\ N \to \infty$.

PROOF OF LEMMA 6.4 From the definition of $T_\lambda(\cdot)$,

$$\int_{-\pi}^{\pi} K_h(\lambda - t)\left\{\frac{\partial}{\partial\theta}f_\theta^{-1}|_{\theta=T_\lambda(f)}f(t) + \frac{\partial}{\partial\theta}\log f_\theta(t)|_{\theta=T_\lambda(f)}\right\}dt = 0, \quad (6.268)$$

$$\int_{-\pi}^{\pi} K_h(\lambda - t)\left\{\frac{\partial}{\partial\theta}f_\theta^{-1}|_{\theta=T_\lambda(f_N)}f_N(t) + \frac{\partial}{\partial\theta}\log f_\theta(t)|_{\theta=T_\lambda(f_N)}\right\}dt = 0. \quad (6.269)$$

Using Taylor expansion around $\theta = T_\lambda(f)$ in the left-hand side of (6.269), we can derive that

$$0 = \int_\pi^\pi K_h(\lambda - t) \left\{ \frac{\partial}{\partial\theta} f_\theta^{-1}(t)|_{\theta=T_\lambda(f)} f_N(t) + \frac{\partial}{\partial\theta} \log f_\theta(t)|_{\theta=T_\lambda(f)} \right.$$
$$+ \frac{\partial^2}{\partial\theta\partial\theta'} f_\theta^{-1}(t)|_{\theta=\theta_1} f_N(t)\big(T_\lambda(f_N) - T_\lambda(f)\big)$$
$$\left. + \frac{\partial^2}{\partial\theta\partial\theta'} \log f_\theta(t)|_{\theta=\theta_1} \big(T_\lambda(f_N) - T_\lambda(f)\big) \right\} dt,$$

where $\theta_1 = T_\lambda(f) + \alpha_1(T_\lambda(f_N) - T_\lambda(f)),\ \alpha_1 \in (0,1)$. On the other hand, by (6.268)

$$-\int_{-\pi}^\pi K_h(\lambda - t) \left\{ \frac{\partial}{\partial\theta} f_\theta^{-1}|_{\theta=T_\lambda(f)} f(t) \right\} dt$$
$$= \int_{-\pi}^\pi K_h(\lambda - t) \left\{ \frac{\partial}{\partial\theta} \log f_\theta(t)|_{\theta=T_\lambda(f)} \right\} dt$$

holds. Then we obtain

$$T_\lambda(f_N) - T_\lambda(f)$$
$$= -\left[\int_{-\pi}^\pi K_h(\lambda - t) \left\{ \frac{\partial^2}{\partial\theta\partial\theta'} f_\theta^{-1}(t)|_{\theta=\theta_1} f_N(t) + \frac{\partial^2}{\partial\theta\partial\theta'} \log f_\theta(t)|_{\theta=\theta_1} \right\} dt \right]^{-1}$$
$$\times \int_{-\pi}^\pi K_h(\lambda - t) \left\{ \frac{\partial}{\partial\theta} f_\theta^{-1}(t)|_{\theta=T_\lambda(f)} \right\} \big(f_N(t) - f(t)\big) dt.$$

Because of Theorem 1 of Taniguchi (1987), $\theta_1 \to T_\lambda(f)$ as $N \to \infty$. Since the derivatives $\frac{\partial^2}{\partial\theta\partial\theta'} f_\theta(t)$ and $\frac{\partial}{\partial\theta} f_\theta$ are continuous with respect to θ, the proof is completed. □

Lemma 6.5 *Suppose that the $\{z(n)\}$ and $K(\cdot)$ satisfy Assumptions 6.13 - 6.17. Then for $N \to \infty$, it holds that*

$$\sqrt{Nh}\left(\int_{-\pi}^\pi K_h(\lambda - t) \frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(t)(I_N(t) - f(t)) dt \right.$$
$$\left. + \frac{1}{2} h^2 \sigma_K^2 \frac{\partial^2}{\partial\lambda^2} \{ f(\lambda) \frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(\lambda) \} \right) \xrightarrow{d} N(0, \tilde{V}_\lambda),$$

where

$$\tilde{V}_\lambda = \begin{cases} 2\pi f(\lambda)^2 \int_{-\infty}^\infty K^2(x) dx \frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(\lambda) \frac{\partial}{\partial\theta'} f_{\theta_0}^{-1}(\lambda) & (\lambda \neq 0) \\ 4\pi f(\lambda)^2 \int_{-\infty}^\infty K^2(x) dx \frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(\lambda) \frac{\partial}{\partial\theta'} f_{\theta_0}^{-1}(\lambda) & (\lambda = 0). \end{cases}$$

PROOF OF LEMMA 6.5 For a given smooth function g, it is easily seen that as $h \to 0$

$$\int K_h(\lambda - t) g(t) dt = g(\lambda) + \frac{1}{2} \sigma_K^2 h^2 g''(\lambda) + O(h^4),$$

by using Taylor expansion. Then we have

$$\int_{-\pi}^{\pi} K_h(\lambda - t)\frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(t) f(t) dt$$
$$= \frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(\lambda) f(\lambda) + \frac{1}{2} h^2 \sigma_K^2 \frac{\partial^2}{\partial\lambda^2}\{f(\lambda)\frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(\lambda)\} + O(h^4).$$

Therefore we have only to show that

$$\sqrt{Nh}\left(\int_{-\pi}^{\pi} K_h(\lambda - t)\frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(t) I_N(t) dt - \frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(\lambda) f(\lambda)\right) \xrightarrow{d} N(0, \tilde{V}_\lambda). \quad (6.270)$$

Note that

$$\int_{-\pi}^{\pi} K_h(\lambda - t)\frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(t) I_N(t) dt - \frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(\lambda) f(\lambda)$$
$$= \frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(\lambda)\left\{\int_{-\pi}^{\pi} K_h(\lambda - t) I_N(t) dt - f(\lambda)\right\}$$
$$+ \int_{-\pi}^{\pi} K_h(\lambda - t)\left(\frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(t) - \frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(\lambda)\right) I_N(t) dt$$
$$= A_1 + A_2 \quad \text{(say)}.$$

For A_1 term, it is known that

$$\sqrt{Nh}\left(\int_{-\pi}^{\pi} K_h(\lambda - t) I_N(t) dt - f(\lambda)\right) dt \xrightarrow{d} N(0, V_\lambda),$$

where

$$V_\lambda = \begin{cases} 2\pi f(\lambda)^2 \int_{-\infty}^{\infty} K^2(x) dx & (\lambda \neq 0) \\ 4\pi f(\lambda)^2 \int_{-\infty}^{\infty} K^2(x) dx & (\lambda = 0) \end{cases}$$

(see Theorems 9, 10, 11 of Hannan (1970)). From Assumption 6.14 (1), it follows that

$$\int_{-\infty}^{\infty} K_h(\lambda - t)^2 \{\frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(t) - \frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(\lambda)\}^2 dt$$
$$= \frac{1}{h}\{\frac{\partial}{\partial\lambda}\frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(\lambda)\}^2 \int_{-\infty}^{\infty} s^2 K^2(s) ds + o(h) < \infty. \quad (6.271)$$

Next, regarding A_2 term, from Lemma A2.2 of Hosoya and Taniguchi (1982) and (6.271), we observe that

$$\lim_{N\to\infty} Nh\mathrm{Var}\left\{\int_{-\pi}^{\pi} K_h(\lambda - t)(\frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(t) - \frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(\lambda)) I_N(t) dt\right\}$$
$$= \lim_{N\to\infty}\left\{4\pi h \int_{-\pi}^{\pi} K_h^2(\lambda - t)(\frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(t) - \frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(\lambda))^2 f^2(t) dt\right.$$
$$+ 2\pi h \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} K_h(\lambda - t) K_h(\lambda - s)(\frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(t) - \frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(\lambda))$$
$$\left.\times(\frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(-s) - \frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(\lambda))\tilde{Q}_z(t, s, -s) dt ds\right\}$$

$$
\begin{aligned}
&= \lim_{N\to\infty} \Big\{ 4\pi \int_{-\pi}^{\pi} K^2(z) \{ \frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(\lambda - hz) - \frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(\lambda) \}^2 f^2(\lambda - hz) dz \\
&\qquad + 2\pi h \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} K(z_1)K(z_2) \{ \frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(\lambda - hz_1) - \frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(\lambda) \} \\
&\qquad\qquad \times \{ \frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(-\lambda + hz_2) - \frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(\lambda) \} \\
&\qquad\qquad \times \tilde{Q}_z(\lambda - hz_1, -\lambda - hz_2, \lambda - hz_2) dz_1 dz_2 \Big\} \\
&\longrightarrow 0,
\end{aligned}
$$

by the dominated convergence theorem, where $\tilde{Q}_z(\lambda_1, \lambda_2, \lambda_3)$ is the fourth-order spectral density of $\{z(n)\}$. Therefore

$$
\int_{-\pi}^{\pi} K_h(\lambda - t) (\frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(t) - \frac{\partial}{\partial\theta_j} f_{\theta_0}^{-1}(\lambda)) I_N(t) dt
$$

is of order $o_p((Nh)^{-1/2})$, which completes the proof. □

PROOF OF THEOREM 6.19 By Lemma 6.4

$$
\begin{aligned}
&\sqrt{Nh}(\hat{\theta}(\lambda) - \theta_0(\lambda)) \\
&\quad = -\sqrt{Nh} \left[\int_{-\pi}^{\pi} K_h(\lambda - t) \frac{\partial^2}{\partial\theta\partial\theta'} (f_{\theta_0}^{-1}(t) I_N(t) + \log f_{\theta_0}(t)) dt \right]^{-1} \\
&\qquad \times \int_{-\pi}^{\pi} K_h(\lambda - t) \frac{\partial}{\partial\theta} f_{\theta}^{-1}(t)|_{\theta = T_\lambda(f)} (I_N(t) - f(t)) dt \\
&\qquad + \alpha_{N,h} \sqrt{Nh} \int_{-\pi}^{\pi} K_h(\lambda - t) \frac{\partial}{\partial\theta} f_{\theta}^{-1}(t)|_{\theta = T_\lambda(f)} (I_N(t) - f(t)) dt \\
&\quad = -M_h^{-1}(\lambda) \sqrt{Nh} \int_{-\pi}^{\pi} K_h(\lambda - t) \frac{\partial}{\partial\theta} f_{\theta_0}^{-1}(t) (I_N(t) - f(t)) dt + o_p(1),
\end{aligned}
$$

where

$$
M_h(\lambda) = \int_{-\pi}^{\pi} K_h(\lambda - t) \frac{\partial^2}{\partial\theta\partial\theta'} (f_{\theta_0}^{-1}(t) I_N(t) + \log f_{\theta_0}(t)) dt.
$$

Since

$$
M_h(\lambda) \to M(\lambda) = \frac{\partial^2}{\partial\theta\partial\theta'} (f_{\theta_0}^{-1}(\lambda) f(\lambda) + \log f_{\theta_0}(\lambda)),
$$

the result follows from Lemma 6.5. □

The following theorem establishes the asymptotic normality of local Whittle estimator $f_{\hat{\theta}(\lambda)}(\lambda)$.

Theorem 6.20 *Under the same conditions as in Theorem 6.19*

$$
\sqrt{Nh} \Big(f_{\hat{\theta}(\lambda)}(\lambda) - f(\lambda) \Big) \longrightarrow N\Big(0, (\frac{\partial}{\partial\theta'} f_{\theta_0}(\lambda)) M(\lambda)^{-1} \tilde{V} M(\lambda)^{-1} (\frac{\partial}{\partial\theta} f_{\theta_0}(\lambda)) \Big).
$$

PROOF The bias term of Theorem 6.19 is negligible. In fact,

$$\begin{aligned} 0 &= \int_{-\pi}^{\pi} K_h(\lambda - t) f_{\theta_0}^{-2}(t) \frac{\partial}{\partial \theta} f_{\theta_0}(t)(f_{\theta_0}(t) - f(t))dt \\ &= f_{\theta_0}^{-2}(\lambda) \frac{\partial}{\partial \theta} f_{\theta_0}(\lambda)\{f_{\theta_0}(\lambda) - f(\lambda)\} \\ &\quad + \frac{1}{2} h^2 \sigma_K^2 \frac{\partial^2}{\partial \lambda^2} \{f_{\theta_0}^{-2}(\lambda) \frac{\partial}{\partial \theta} f_{\theta_0}(\lambda)(f_{\theta_0}(\lambda) - f(\lambda))\} + O(h^4) \end{aligned}$$

holds by the definition of $\theta_0(\lambda)$. Because $Nh^5 \to 0$ as $N \to \infty$, we can see $f_{\theta_0}(\lambda) - f(\lambda) = o((Nh)^{-1/2})$. Hence Theorem 6.20 follows from the delta method (see Proposition 6.4.3 of Brockwell and Davis (1991)). □

Remark 6.2 *In Theorem 6.20, we can see that the asymptotic variance and bias of the local Whittle estimator depend on only $f(\lambda)$, $f_{\theta_0}(\lambda)$ and K. Thus the asymptotic distribution of the local Whittle estimator does not depend on non-Gaussianity of the process.*

When we estimate the spectral density of an observed time series, it is a significant problem whether the spectral density is parametric or not. For this, Fan and Zhang (2004) applies local linear polynomial techniques to the log-periodogram of Gaussian process.

Consider the problem of testing whether $f(\lambda)$ belongs to a specific parametric family $\{f_\theta(\cdot) : \theta \in \Theta\}$ or not, i.e.,

$$H_0 : f(\cdot) = f_\theta(\cdot) \quad \text{v.s.} \quad H_1 : f(\cdot) \neq f_\theta(\cdot). \tag{6.272}$$

Although we do not assume Gaussianity of $\{X_t\}$, the Whittle likelihood function under H_0 is expressed as

$$l(\theta) = -\sum_{k=1}^{N} \left\{ \log f_\theta(\lambda_k) + \frac{I_N(\lambda_k)}{f_\theta(\lambda_k)} \right\}, \ \lambda_k = -\pi + 2\pi k/N \ (k = 1, \ldots, N).$$

We call $\hat{\theta}_{\mathrm{WH}} = \arg\max_{\theta \in \Theta} l(\theta)$ the Whittle likelihood estimator of θ.

For H_1, we use the following local Whittle likelihood function around $\lambda \in [-\pi, \pi]$:

$$l^{\mathrm{loc}}(\theta) = -\sum_{k=1}^{N} \left\{ \log f_\theta(\lambda_k) + \frac{I_N(\lambda_k)}{f_\theta(\lambda_k)} \right\} K_h(\lambda - \lambda_k), \tag{6.273}$$

where $K_h(\cdot)$ is an appropriate kernel function. Let $\hat{\theta}_{\mathrm{LW}}(\lambda) = \mathrm{argmax}_{\theta \in \Theta} l^{\mathrm{loc}}(\theta)$, and we regard $f_{\hat{\theta}_{\mathrm{LW}}(\lambda)}(\lambda)$ as a sort of nonparametric estimator of the spectral density $f(\lambda)$. For the testing problem (6.272), we use the following likelihood

ratio test statistic

$$\begin{aligned} T_{\mathrm{LW}} &= l(\hat{\theta}_{\mathrm{WH}}) - l(\hat{\theta}_{\mathrm{LW}}(\lambda)) \\ &= -\sum_{k=1}^{N}\left\{\log f_{\hat{\theta}_{\mathrm{WH}}}(\lambda_k) + \frac{I_N(\lambda_k)}{f_{\hat{\theta}_{\mathrm{WH}}}(\lambda_k)}\right\} \\ &\qquad + \sum_{k=1}^{N}\left\{\log f_{\hat{\theta}_{\mathrm{LW}(\lambda_k)}}(\lambda_k) + \frac{I_N(\lambda_k)}{f_{\hat{\theta}_{\mathrm{LW}(\lambda_k)}}(\lambda_k)}\right\} \\ &= \sum_{k=1}^{N}\Big\{\log f_{\hat{\theta}_{\mathrm{LW}(\lambda_k)}}(\lambda_k) - \log f_{\hat{\theta}_{\mathrm{WH}}}(\lambda_k) \\ &\qquad + I_N(\lambda_k)(f_{\hat{\theta}_{\mathrm{LW}(\lambda_k)}}(\lambda_k)^{-1} - f_{\hat{\theta}_{\mathrm{WH}}}(\lambda_k)^{-1})\Big\}. \end{aligned}$$

Actually if $T_{\mathrm{LW}} > z_\alpha$, a selected level, we reject H_0, otherwise, accept it.

We derive the asymptotics of T_{LW} under H_0 of (6.272). Write T_{LW} as

$$\begin{aligned} T_{\mathrm{LW}} &= \{l(\theta) - l(\hat{\theta}_{\mathrm{LW}}(\lambda))\} - \{l(\theta) - l(\hat{\theta}_{\mathrm{WH}})\} \\ &= T_{\mathrm{LW},1} - T_{\mathrm{LW},2}. \end{aligned}$$

It is known that $T_{\mathrm{LW},2} = O_p(1)$ under appropriate regularity conditions (e.g., Taniguchi and Kakizawa (2000, Section 3.1)). In what follows, it is seen that $T_{\mathrm{LW},1}$ is asymptotically of order in probability tending to ∞. Hence, in order to derive the asymptotic distribution of T_{LW} we have only to derive that of $T_{\mathrm{LW},1}$. For this, furthermore, we impose the following assumption.

Assumption 6.18 *(i) $\{z(t)\}$ is kth order stationary with all of whose moments exist.*
(ii) The joint kth order cumulant $Q_z(j_1, \ldots, j_{k-1})$ of $z(t), z(t+j_1), \ldots, z(t+j_{k-1})$ satisfies

$$\sum_{j_1,\ldots,j_{k-1}=-\infty}^{\infty} (1+|j_l|)|Q_z(j_1,\ldots,j_{k-1})| < \infty$$

for $l = 1, \ldots, k-1$ and any k, $k = 2, 3, \ldots$

Then we get the following theorem.

Theorem 6.21 *Suppose that Assumptions 6.13 - 6.16 and 6.18 hold. Then, under H_0,*

$$\sigma_N^{-1}(T_{LW} - \mu_N) \xrightarrow{d} N(0,1),$$

where

$$\mu_N = \frac{1}{h}\left[-K(0)\int_{-\pi}^{\pi} F(\lambda)d\lambda + \frac{1}{2}\int_{-\pi}^{\pi} F(\lambda)^2 d\lambda \int_{-\infty}^{\infty} K(\omega)^2 d\omega\right],$$
$$\sigma_N^2 = \frac{2}{h}\int_{-\pi}^{\pi} F(\lambda)^2 d\lambda \int_{-\infty}^{\infty} K(\omega)^2 d\omega,$$
$$F(\lambda) = \frac{\partial}{\partial\theta'} f_{\theta(\lambda)}(\lambda)^{-1} M(\lambda)^{-1} \frac{\partial}{\partial\theta} f_{\theta(\lambda)}(\lambda)^{-1} f_{\theta(\lambda)}(\lambda)^2,$$
$$M(\lambda) = \frac{\partial^2}{\partial\theta\partial\theta'} \log f_{\theta(\lambda)}(\lambda).$$

PROOF Noting the fact $T_{\mathrm{LW},2} = O_p(1)$ and the proof of Theorem 6.20, Taylor expansion around $\theta(\lambda)$ yields

$$\begin{aligned} T_{\mathrm{LW}} = \sum_{k=1}^{N} \Bigg[& \frac{\partial}{\partial\theta'} \log f_{\theta(\lambda_k)}(\lambda_k)(\hat{\theta}(\lambda_k) - \theta(\lambda_k)) \\ & + \frac{1}{2}(\hat{\theta}(\lambda_k) - \theta(\lambda_k))' \frac{\partial^2}{\partial\theta\partial\theta'} \log f_{\theta(\lambda_k)}(\lambda_k)(\hat{\theta}(\lambda_k) - \theta(\lambda_k)) \\ & + I_N(\lambda_k)\left\{ \frac{\partial}{\partial\theta'} f_{\theta(\lambda_k)}(\lambda_k)^{-1}(\hat{\theta}(\lambda_k) - \theta(\lambda_k)) \right. \\ & \left. + \frac{1}{2}(\hat{\theta}(\lambda_k) - \theta(\lambda_k))' \frac{\partial^2}{\partial\theta\partial\theta'} f_{\theta(\lambda_k)}(\lambda_k)^{-1}(\hat{\theta}(\lambda_k) - \theta(\lambda_k)) \right\} \Bigg] \\ & + \text{ lower order terms.} \end{aligned} \tag{6.274}$$

The validity for the lower terms can be found in Theorem 5.1.7. of Fuller (1996). Since

$$\begin{aligned} \frac{\partial^2}{\partial\theta\partial\theta'} \log f_{\theta(\lambda)}(\lambda) + f_{\theta(\lambda)}(\lambda) \frac{\partial^2}{\partial\theta\partial\theta'} f_{\theta(\lambda)}(\lambda)^{-1} & \\ = f_{\theta(\lambda)}(\lambda)^{-2} \frac{\partial}{\partial\theta} f_{\theta(\lambda)}(\lambda) \frac{\partial}{\partial\theta'} f_{\theta(\lambda)}(\lambda), & \\ E\{I_N(\lambda)\} = f_{\theta(\lambda)}(\lambda) + O(N^{-1}), & \end{aligned}$$

it is not difficult to show that (6.274) is equal to

$$\begin{aligned} T_{\mathrm{LW}} = \sum_{k=1}^{N} \Bigg[& f_{\theta(\lambda_k)}(\lambda_k)^{-2}\{f_{\theta(\lambda_k)}(\lambda_k) - I_N(\lambda_k)\} \\ & \times \frac{\partial}{\partial\theta'} f_{\theta(\lambda_k)}(\lambda_k)(\hat{\theta}(\lambda_k) - \theta_0(\lambda_k)) \\ & + \frac{1}{2} f_{\theta(\lambda_k)}(\lambda_k)^{-2}(\hat{\theta}(\lambda_k) - \theta_0(\lambda_k))' \frac{\partial}{\partial\theta} f_{\theta(\lambda_k)}(\lambda_k) \\ & \times \frac{\partial}{\partial\theta'} f_{\theta(\lambda_k)}(\lambda_k)(\hat{\theta}(\lambda_k) - \theta_0(\lambda_k)) \Bigg] \\ & + \text{lower order terms.} \end{aligned} \tag{6.275}$$

From Theorem 6.19, it follows that

$$\begin{aligned}
&\hat{\theta}(\lambda) - \theta(\lambda) \\
&\quad = -M(\lambda)^{-1} \frac{\partial}{\partial \theta} f_{\theta(\lambda)}(\lambda)^{-1} [\hat{f}_N(\lambda) - E\{\hat{f}_N(\lambda)\}] + o_p(N^{-\frac{1}{2}} h^{-\frac{1}{2}}) \\
&\quad = -M(\lambda)^{-1} \frac{\partial}{\partial \theta} f_{\theta(\lambda)}(\lambda)^{-1} \frac{2\pi}{N} \sum_{l=1}^{N} K_h(\lambda - \lambda_l) \{I_N(\lambda_l) - f_{\theta(\lambda_l)}(\lambda_l)\} \\
&\qquad + o_p(N^{-\frac{1}{2}} h^{-\frac{1}{2}}),
\end{aligned} \tag{6.276}$$

where

$$\hat{f}_N(\lambda) = \int_{-\pi}^{\pi} K_h(\lambda - \omega) I_N(\omega) d\omega.$$

Substituting (6.276) into (6.275), we obtain

$$\begin{aligned}
&\sqrt{h} T_{\mathrm{LW}} \\
&= \frac{1}{\sqrt{N}} \sum_{k=1}^{N} f_{\theta(\lambda_k)}(\lambda_k)^{-2} \{f_{\theta(\lambda_k)}(\lambda_k) - I_N(\lambda_k)\} \frac{\partial}{\partial \theta'} f_{\theta(\lambda_k)}(\lambda_k) M(\lambda_k)^{-1} \\
&\quad \times \sqrt{Nh} \frac{2\pi}{N} \sum_{l=1}^{N} K_h(\lambda_k - \lambda_l) \frac{\partial}{\partial \theta} f_{\theta(\lambda_k)}(\lambda_k)^{-1} \{I_N(\lambda_l) - f_{\theta(\lambda_l)}(\lambda_l)\} \\
&\quad + \sqrt{h} \sum_{k=1}^{N} \frac{1}{2} f_{\theta(\lambda_k)}(\lambda_k)^{-2} \left\{ M(\lambda_k)^{-1} \frac{\partial}{\partial \theta} f_{\theta(\lambda_k)}(\lambda_k)^{-1} (\hat{f}_N(\lambda_k) - E(\hat{f}_N(\lambda_k))) \right\}' \\
&\quad \times \frac{\partial}{\partial \theta} f_{\theta(\lambda_k)}(\lambda_k) \frac{\partial}{\partial \theta'} f_{\theta(\lambda_k)}(\lambda_k) \\
&\quad \times \left\{ M(\lambda_k)^{-1} \frac{\partial}{\partial \theta} f_{\theta(\lambda_k)}(\lambda_k)^{-1} (\hat{f}_N(\lambda_k) - E(\hat{f}_N(\lambda_k))) \right\} + \text{lower order terms} \\
&= -\frac{2\pi\sqrt{h}}{N} \sum_{k,l=1}^{N} K_h(\lambda_k - \lambda_l) A(\lambda_k) \{I_N(\lambda_k) - f_{\theta(\lambda_k)}(\lambda_k)\} \{I_N(\lambda_l) - f_{\theta_0(\lambda_l)}(\lambda_l)\} \\
&\quad + \frac{\sqrt{h}}{2} \sum_{k=1}^{N} f_{\theta(\lambda_k)}(\lambda_k)^2 A(\lambda_k)^2 \{\hat{f}_N(\lambda_k) - E(\hat{f}_N(\lambda_k))\}^2 + \text{lower order terms} \\
&= (\mathrm{B1}) + (\mathrm{B2}) + \text{lower order terms} \quad \text{(say)},
\end{aligned} \tag{6.277}$$

where

$$A(\lambda_k) = \frac{\partial}{\partial \theta'} f_{\theta(\lambda_k)}(\lambda_k)^{-1} M(\lambda_k)^{-1} \frac{\partial}{\partial \theta} f_{\theta(\lambda_k)}(\lambda_k)^{-1}.$$

The expectation of (B1) is written as

$$E(\mathrm{B1}) = -\frac{2\pi\sqrt{h}}{N}\sum_{k,l=1}^{N} K_h(\lambda_k-\lambda_l)A(\lambda_k) \times E\{(I_N(\lambda_k)-f_{\theta(\lambda_k)}(\lambda_k))(I_N(\lambda_l)-f_{\theta_0(\lambda_l)}(\lambda_l))\}. \tag{6.278}$$

By Theorem 5.2.4 of Brillinger (1981) and noting $K_h(\cdot)=\frac{1}{h}K(\cdot/h)$, it is seen that

$$E(\mathrm{B1}) = -\frac{K(0)}{\sqrt{h}}\int_{-\pi}^{\pi} A(\lambda) f_{\theta_0(\lambda)}(\lambda)^2 d\lambda + \text{lower order terms}. \tag{6.279}$$

Next, we evaluate the expectation of (B2). Corollary 5.6.2 of Brillinger (1981) leads to

$$E(\mathrm{B2}) = \frac{\sqrt{h}}{2}\sum_{k=1}^{N} f_{\theta(\lambda_k)}(\lambda_k)^2 A(\lambda_k)^2 \ \mathrm{Var}(\hat{f}_N(\lambda_k)) \tag{6.280}$$

$$= \frac{\sqrt{h}}{2}\sum_{k=1}^{N} f_{\theta(\lambda_k)}(\lambda_k)^2 A(\lambda_k)^2 \left\{\frac{2\pi}{Nh} f_{\theta(\lambda_k)}(\lambda_k)^2 \int_{-\infty}^{\infty} K(\omega)^2 d\omega\right\} + \text{lower order terms}$$

$$= \frac{1}{2\sqrt{h}}\int_{-\pi}^{\pi} f_{\theta_0(\lambda)}(\lambda)^4 A(\lambda)^2 d\lambda \int_{-\infty}^{\infty} K(\omega)^2 d\omega + \text{lower order terms}. \tag{6.281}$$

Hence, the main order term of $E(\sqrt{h}T_{\mathrm{LW}})$ is given by (6.279) and (6.281). In what follows, we evaluate the variance of (B1) and (B2). Let $H(\lambda) = I_N(\lambda) - f_{\theta(\lambda)}(\lambda)$. Then we have

$$\mathrm{Var}(\mathrm{B1}) = \left(\frac{2\pi\sqrt{h}}{N}\right)^2 \sum_{k,l,k',l'} K_h(\lambda_k-\lambda_l)K_h(\lambda_{k'}-\lambda_{l'})A(\lambda_k)A(\lambda_{k'}) \times \mathrm{cum}\{H(\lambda_k)H(\lambda_l), H(\lambda_{k'})H(\lambda_{l'})\}. \tag{6.282}$$

Using the formula

$$\begin{aligned}\mathrm{cum}\{H(\lambda_k)H(\lambda_l), H(\lambda_{k'})H(\lambda_{l'})\} &= \mathrm{cum}\{H(\lambda_k), H(\lambda_l), H(\lambda_{k'}), H(\lambda_{l'})\} \\ &\quad + \mathrm{cum}\{H(\lambda_k)H(\lambda_{k'})\}\mathrm{cum}\{H(\lambda_l)H(\lambda_{l'})\} \\ &\quad + \mathrm{cum}\{H(\lambda_k)H(\lambda_{l'})\}\mathrm{cum}\{H(\lambda_l)H(\lambda_{k'})\},\end{aligned} \tag{6.283}$$

we can see that the second and third terms in the right-hand side of (6.283) contribute the main order terms V_2 and V_3, (say), respectively, of (6.282). In

fact, it is seen that

$$\begin{aligned} V_2 =& \left(\frac{2\pi\sqrt{h}}{N}\right)^2 \sum_{k,l} K_h(\lambda_k-\lambda_l)^2 A(\lambda_k)^2 f_{\theta(\lambda_k)}(\lambda_k)^2 f_{\theta(\lambda_l)}(\lambda_l)^2 \\ &+ \text{lower order terms} \\ &\text{(by transformation } (\lambda_k-\lambda_l)/h \to \omega_k,\ \lambda_l \to \lambda_l) \\ =& \int_{-\infty}^{\infty} K(\omega)^2 d\omega \int_{-\pi}^{\pi} A(\lambda)^2 f_{\theta(\lambda)}(\lambda)^4 d\lambda + \text{lower order terms.} \end{aligned} \tag{6.284}$$

Similarly we can show the main order term of V_3 is equal to that of V_2. Hence we get

$$\text{Var(B1)} = 2\int_{-\infty}^{\infty} K(\omega)^2 d\omega \int_{-\pi}^{\pi} A(\lambda)^2 f_{\theta(\lambda)}(\lambda)^4 d\lambda + \text{lower order terms.} \tag{6.285}$$

Next, writing $L(\lambda) = \hat{f}_N(\lambda) - E\{\hat{f}_N(\lambda)\}$, we obtain

$$\begin{aligned} \text{Var(B2)} =& h\sum_{k,l} \frac{1}{4} f_{\theta(\lambda_k)}(\lambda_k)^2 A(\lambda_k)^2 f_{\theta(\lambda_l)}(\lambda_l)^2 A(\lambda_l)^2 \\ &\times \left[\text{cum}\{L(\lambda_k), L(\lambda_k), L(\lambda_l), L(\lambda_l)\} + 2\text{cum}\{L(\lambda_k), L(\lambda_l)\}^2\right] \\ &\text{(by formula (6.283))} \\ =& \text{(C1)} + \text{(C2) (say).} \end{aligned} \tag{6.286}$$

It can be shown that the main order term of (6.286) is (C2). From Corollary 5.6.2 of Brillinger (1981), it follows that (C2) $\to 0$, hence, Var(B2) $\to 0$ as $n \to \infty$. Therefore, it suffices to show the asymptotic normality of (B1). For this we evaluate Jth order cumulant of (B1), i.e.,

$$\begin{aligned} \kappa_J =& \text{cum}(\underbrace{(B1), (B1), \ldots, (B1)}_{\text{J}}) \\ =& \left(\frac{2\pi\sqrt{h}}{N}\right)^J \sum_{k_1}\sum_{l_1}\cdots\sum_{k_J}\sum_{l_J} K_h(\lambda_{k_1}-\lambda_{l_1})\cdots K_h(\lambda_{k_J}-\lambda_{l_J}) \\ &\times A(\lambda_{k_1})\cdots A(\lambda_{k_J})\text{cum}\{H(\lambda_{k_1})H(\lambda_{l_1}), \ldots, H(\lambda_{k_J})H(\lambda_{l_J})\}. \end{aligned} \tag{6.287}$$

Thanks to Theorem 5.2.8 of Brillinger (1981), it is not difficult to show

$$\kappa_J = O(h^{\frac{J}{2}})$$

which proves the asymptotic normality of $\sqrt{h}T_{\text{LW}}$, hence, completes the proof of Theorem 6.21. □

Remark 6.3 *It should be noted that the asymptotics of T_{LW} also do not depend on non-Gaussianity of the process. This seems interesting.*

Next we study performance of the local Whittle likelihood estimators for simulated data and feature of the asymptotic distribution of T_{LW}. Consider the

following GARCH(1, 1) model:

$$\begin{aligned} x(t) &= \sigma(t)\varepsilon(t) \\ \sigma^2(t) &= c + bx^2(t-1) + a\sigma^2(t-1) \end{aligned} \tag{6.288}$$

where $c > 0,\ a \geq 0,\ b \geq 0$ and $a + b < 1$, and $\varepsilon(t) \sim$ i.i.d. $(0, 1)$. Let $z(t) = x^2(t)$, then we can write

$$z(t) = c + (b + a)z(t-1) + e(t) - ae(t-1),$$

where

$$e(t) = z^2(t) - \sigma^2(t) = \{\varepsilon^2(t) - 1\}\{c + bx^2(t-1) + a\sigma^2(t-1)\}.$$

Here $\{e(t)\}$ becomes a white noise process with $\sigma_e^2 = \text{Var}[e(t)]$. Therefore, $\{z(t)\}$ becomes an ARMA(1, 1) model, and the spectral density $f_z(\lambda)$ of $\{z(t)\}$ is given by

$$f_z(\lambda) = \frac{\sigma_e^2}{2\pi}\left|\frac{1 - ae^{i\lambda}}{1 - (b + a)e^{i\lambda}}\right|^2.$$

To estimate f_z, we consider the local Whittle likelihood estimator by fitting a family of spectral densities $\{f_\theta, \theta \in \Theta\}$ given by

$$f_\theta(\lambda) = \frac{\sigma^2}{2\pi}|1 - \alpha e^{i\lambda}|^{-2} = \frac{\sigma^2}{2\pi}(1 - 2\alpha\cos\lambda + \alpha^2)^{-1}, \quad \theta = (\alpha, \sigma^2)'.$$

The integral of local Whittle likelihood is approximated by the sum of Fourier frequencies $w_n = 2\pi n/N$. For $I_N(\lambda) = \frac{1}{2\pi N}|\sum_{n=1}^N z(n)e^{in\lambda}|^2$, we calculate $D_\lambda(f_\theta, I_N)$ over grid points on $[0.05, 0.95] \times [1, 25]$ for θ to derive $\hat{\theta}(\lambda)$ for each λ. We compare performance of the local Whittle estimator with that of the traditional estimator,

$$\hat{f}(\lambda) = \int_{-\pi}^{\pi} W_M(\lambda - t)I_N(t)dt$$

where $W_M(\cdot)$ is the Daniell window (see Example 6.17). For spectral density estimator $\tilde{f}(\lambda)$, we define the residual sum of squares (RSS) by

$$\text{RSS}(\tilde{f}) = \sum_{j=0}^{p-1}(f(\lambda_j) - \tilde{f}(\lambda_j))^2$$

where $\lambda_j = j\pi/p$.

The estimated $f_{\hat{\theta}(\lambda)}(\lambda)$ and periodogram $I_N(\lambda)$ are plotted in Figure 6.15. Figure 6.16 provides the graphs of $f_{\hat{\theta}}$ (solid line), the smoothed periodogram $\hat{f}$ (dashed line) and the true spectral density f (dotted line) with $c = 1,\ a = 0.1,\ b = 0.2,\ h = 0.5$ and $M = 4$. Then we observe that $\text{RSS}(f_{\hat{\theta}}) = 1.15409$ and $\text{RSS}(\hat{f}) = 1.54844$, which implies that the local Whittle estimator is better than the traditional one. Further, we repeated this experiment ten times, and calculated the sample mean of RSS, say $\overline{\text{RSS}}$. Then we get $\overline{\text{RSS}}(f_{\hat{\theta}}) = 2.8789$ and $\overline{\text{RSS}}(\hat{f}) = 2.9991$. Hence the local Whittle estimator $f_{\hat{\theta}}$ is recommendable.

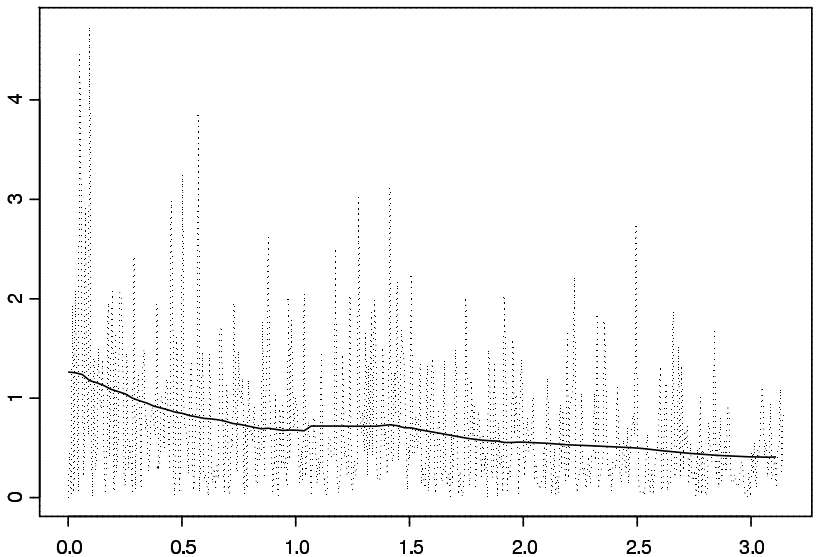


Figure 6.15 *Local Whittle estimator (solid) and periodogram (dotted).*

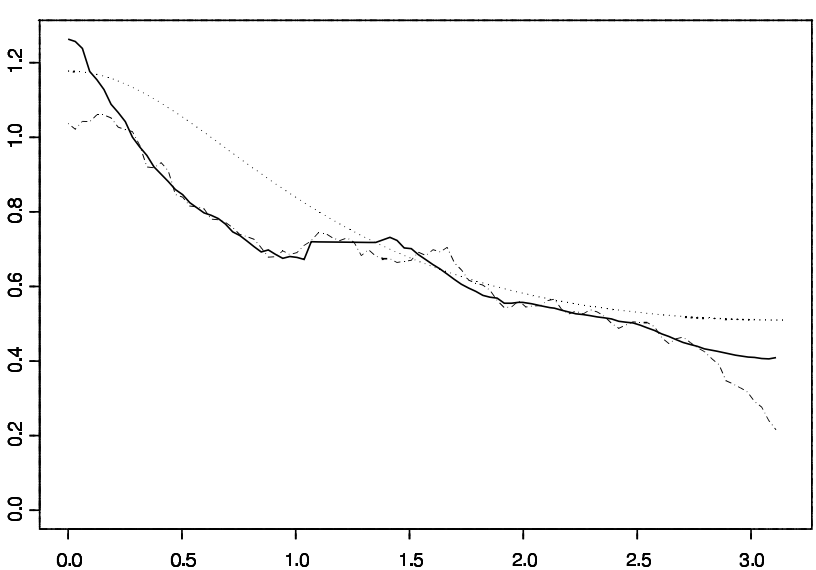


Figure 6.16 *Local Whittle estimator (solid), smoothed periodogram (dashed) and true spectral density (dotted).*

Recall that $\hat{\theta}_{\mathrm{LW}}(\lambda)$ is the minimizer of (6.273), and that $T_{\mathrm{LW},1} = l(\theta) - l(\hat{\theta}_{\mathrm{LW}}(\lambda))$. Let $c = 1.0,\ a = 0.1,\ b = 0.3$ in (6.288), then generate $x(1), \ldots, x(500)$. Based on them we can calculate $T_{\mathrm{LW},1}$. Here we consider the local Whittle likelihood estimator by fitting a family of MA(1) spectral densities to estimate $f_z(\lambda)$, i.e., $\{f_\theta, \theta \in \Theta\}$ given by

$$f_\theta(\lambda) = \frac{\sigma^2}{2\pi}|1 + \alpha e^{i\lambda}|^2$$

where $\theta = (\alpha, \sigma^2)'$. We repeated this procedure 1000 times. Then, Figures 6.17 and 6.18 gives the empirical distributions of $T_{\mathrm{LW},1}$ for the cases of (i) $\varepsilon(t) \sim$ i.i.d. $N(0,1)$ and (ii) $\varepsilon(t) = \eta(t)/\sqrt{5/4} \sim$ i.i.d. $(0,1)$ with $\eta(t) \sim$ i.i.d. $t(10)$, respectively.

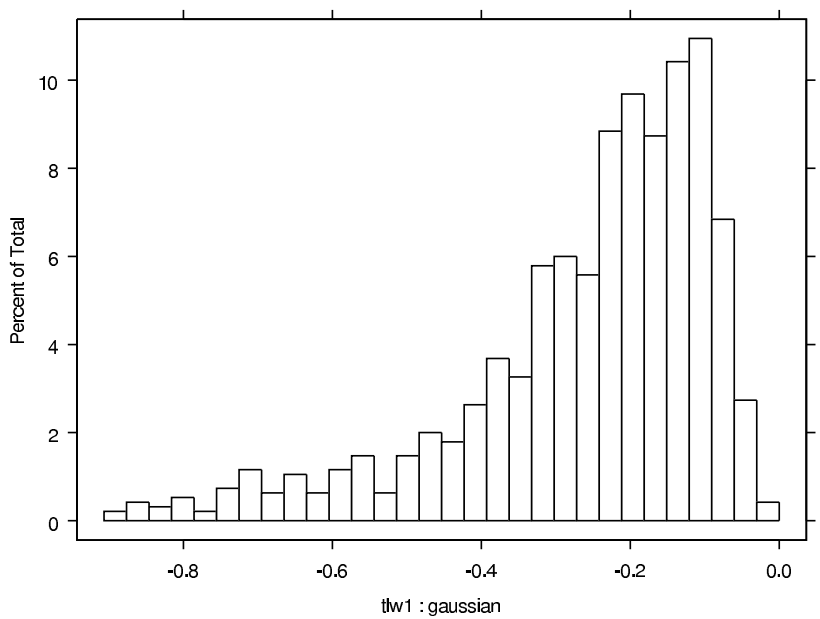


Figure 6.17 *Empirical distribution of $T_{LW,1}$ when the innovation process is Gaussian.*

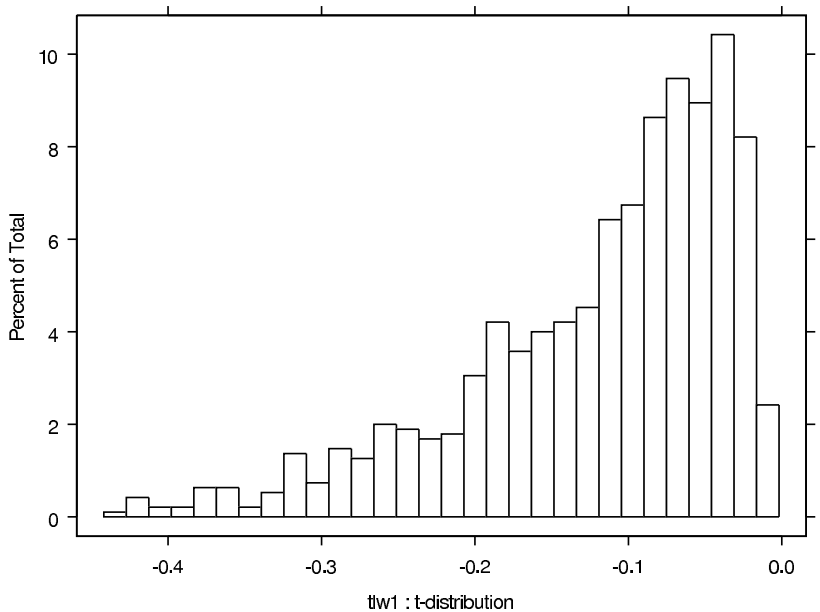


Figure 6.18 *Empirical distribution of $T_{LW,1}$ when the innovation process is t-distributed.*

These figures are similar, hence, imply Remark 6.3.

6.9 Nonstationary Processes

So far we assumed stationarity of the processes concerned. However, empirical studies show that most time series data such as financial and biological time series exhibit nonstationary behavior. The random walk process

$$Y_t = \sum_{j=1}^{t} u_j, \quad \left(\{u_j\} \sim i.i.d.\ (0, \sigma^2)\right), \tag{6.289}$$

is the most fundamental nonstationary process. This process will be reasonable for economic indices in which we suppose a value at the present time t is represented by the sum of random shocks over $t = 1, 2, \ldots, t$. We can rewrite the equation (6.289) as

$$Y_t = Y_{t-1} + u_t, \quad (Y_0 = 0), \tag{6.290}$$

which corresponds to AR(1) process

$$Y_t = bY_{t-1} + u_t, \tag{6.291}$$

with $b = 1$. In this case we say that AR model (6.291) has a unit root. Moreover, if we recall FARIMA models, it corresponds to FARIMA(0,1,0) process (6.245). Although the unit root problem seems to be an extremely special topic from a mathematical point of view, an enormous number of works have recently appeared in the field of econometrics and it forms a huge region of research area. We recommend Tanaka (1996) to the readers who are interested in a comprehensive study of this field. It is known that the local asymptotic normality (LAN) does not hold for models including the unit root. Phillips (1989) showed that the log-likelihood ratio between hypothetical value θ and contiguous alternative θ_n in these models has stochastic expansion in the form

$$\Lambda_n(\theta, \theta_n) = hU_n - \frac{h^2}{2} S_n + o_p(1) \tag{6.292}$$

and called the statistical models the *limiting Gaussian functional* (LGF). The difference from LAN is that both U_n and S_n are random variables and their limit distributions become complicated forms.

We now turn to discuss nonstationary models which are regular in the sense that they satisfy the LAN property. At the sight of actual time series data we often find that they seem locally stationary and contain several changes of the structure in its entirety. So they fall into categories of nonstationary time series. One major difficulty in developing the general nonstationary theory for such processes is the problem of asymptotics. But the asymptotic theory is needed since investigation of e.g., the maximum likelihood estimator for a fixed sample size is too complicated and will not lead to any satisfactory results. On the other hand the classical asymptotic theory with the assumption that more and more observations of future become available does not make sense since future observations of general nonstationary processes do not necessarily contain any information on the structure at present. To meet this Dahlhaus (1996a, 1996b) introduced an important class of nonstationary processes with rigorous asymptotic framework, called locally stationary processes. We give the precise definition which is due to Dahlhaus (1996a, 1996b).

Definition 6.1 *A sequence of stochastic processes* $X_{t,T}$ $(t = 1, \ldots, T; T \geq 1)$ *is called* locally stationary *with* transfer function A° *if there exists a representation*

$$X_{t,T} = \int_{-\pi}^{\pi} \exp(i\lambda t) A^\circ_{t,T}(\lambda) d\xi(\lambda), \tag{6.293}$$

where

(i) $\xi(\lambda)$ *is a stochastic process on* $[-\pi, \pi]$ *with* $\overline{\xi(\lambda)} = \xi(-\lambda)$ *and*

$$\operatorname{cum}\{d\xi(\lambda_1), \ldots, d\xi(\lambda_k)\} = \eta\left(\sum_{j=1}^{k} \lambda_j\right) h_k(\lambda_1, \ldots \lambda_{k-1}) d\lambda_1 \cdots d\lambda_{k-1}, \tag{6.294}$$

where $\operatorname{cum}\{\ldots\}$ *denotes the cumulant of kth order,* $h_1 = 0$, $h_2(\lambda) = (2\pi)^{-1}$, $|h_k(\lambda_1, \ldots, \lambda_{k-1})| \leq \text{const}_k$ *for all* $k \geq 3$ *and* $\eta(\lambda) = \sum_{j=-\infty}^{\infty} \delta(\lambda + 2\pi j)$ *is the period* 2π *extension of the Dirac delta function.*

(ii) *There exists a constant* K *and a* 2π*-periodic function* $A : [0,1] \times \mathbf{R} \to \mathbf{C}$ *with* $\overline{A(u,\lambda)} = A(u,-\lambda)$ *and*

$$\sup_{t,\lambda} \left| A^{\circ}_{t,T}(\lambda) - A\left(\frac{t}{T}, \lambda\right) \right| \leq KT^{-1} \tag{6.295}$$

for all T. $A(u,\lambda)$ *is assumed to be continuous in* u *and* $f(u,\lambda) := |A(u,\lambda)|^2$ *is called the time varying spectral density of the process.*

Here we consider the parametric case. Let $\mathbf{X}_T = (X_{1,T}, \ldots, X_{T,T})'$ be a realization of a locally stationary process with transfer function A°_{θ} where the corresponding A_θ is uniformly bounded from above and below and time varying density $f_\theta(u,\lambda)$ depends on a parameter vector $\theta = (\theta_1, \ldots, \theta_r)' \in \Theta \subset \mathbf{R}^r$. Introducing the notations $\nabla_i = \frac{\partial}{\partial \theta_i}$, $\nabla = (\nabla_1, \ldots, \nabla_r)'$, $\nabla_{ij} = \frac{\partial}{\partial \theta_i}\frac{\partial}{\partial \theta_j}$ and $\nabla^2 = (\nabla_{ij})_{i,j=1,\ldots,r}$, we make the following assumption:

Assumption 6.19 (A1) *There exists a constant* K *with*

$$\sup_{t,\lambda} \left| \nabla^s \left\{ A^{\circ}_{\theta,t,T} - A_\theta\left(\frac{t}{T}, \lambda\right) \right\} \right| \leq KT^{-1} \tag{6.296}$$

for $s = 0, 1, 2$. *The components of* $A_\theta(u,\lambda)$, $\nabla A_\theta(u,\lambda)$ *and* $\nabla^2 A_\theta(u,\lambda)$ *are differentiable in* u *and* λ *with uniformly continuous derivatives* $\frac{\partial}{\partial u}\frac{\partial}{\partial \lambda}$.

Write

$$\varepsilon_t = \int_{-\pi}^{\pi} \exp(i\lambda t) d\xi(\lambda), \tag{6.297}$$

then ε_t becomes the innovation of the process. We assume the following assumptions on $\{\varepsilon_t\}$.

Assumption 6.20 (B1) $h_k(\lambda_1, \ldots, \lambda_{k-1}) = \frac{h_k}{(2\pi)^{k-1}}$ *for all* $k \geq 3$.

(B2) *The distribution of innovation* ε_t *is absolutely continuous with respect to Lebesgue measure and the probability density* $p(\cdot)$ *of* ε_t *satisfies* $p(z) > 0$ *on* $\mathbf{R}$ *and*

$$\lim_{|z| \to \infty} zp(z) = 0. \tag{6.298}$$

(B3) *The continuous derivatives* Dp *and* $D^2 p \equiv D(Dp)$ *of* $p(\cdot)$ *exist on* $\mathbf{R}$ *and* $D^2 p$ *satisfies the Lipschitz condition.*

(B4)

$$\mathcal{F}(p) = \int (\phi(z))^2 p(z) dz < \infty, \tag{6.299}$$

$$E\left\{\varepsilon_t \phi^2(\varepsilon_t)\right\} < \infty, \quad E\left\{\varepsilon_t^2 \phi^2(\varepsilon_t)\right\} < \infty, \quad E\left\{\phi^4(\varepsilon_t)\right\} < \infty \tag{6.300}$$

and

$$\int D^2 p(z)dz = 0, \quad \lim_{|z|\to\infty} z^2 Dp(z) = 0, \tag{6.301}$$

where $\phi(\cdot) = \frac{Dp(\cdot)}{p(\cdot)}$.

Here we are concerned with the structure of the linear process:

Assumption 6.21 (C1) $\{X_{t,T}\}$ *has the MA*(∞) *and AR*(∞) *representations*

$$X_{t,T} = \sum_{j=0}^{\infty} a^{\circ}_{\theta,t,T}(j)\varepsilon_{t-j}, \tag{6.302}$$

$$a^{\circ}_{\theta,t,T}(0)\varepsilon_t = \sum_{j=0}^{\infty} b^{\circ}_{\theta,t,T}(j)X_{t-j,T}, \tag{6.303}$$

where $a^{\circ}_{\theta,t,T}(j), b^{\circ}_{\theta,t,T}(j) \in \mathbf{R}$, $b^{\circ}_{\theta,t,T}(0) \equiv 1$ *and* $a^{\circ}_{\theta,t,T}(j) = a^{\circ}_{\theta,0,T}(j) = a^{\circ}_{\theta}(j)$ *for* $t \le 0$.

(C2) *Every* $a^{\circ}_{\theta,t,T}(j)$ *is continuously three times differentiable with respect to* θ *and the derivatives satisfy*

$$\sup_{t,T}\left\{\sum_{j=0}^{\infty}(1+j)\left|\nabla_{i_1}\cdots\nabla_{i_s}a^{\circ}_{\theta,t,T}(j)\right|\right\} < \infty \quad \text{for } s = 0,1,2,3. \tag{6.304}$$

(C3) *Every* $b^{\circ}_{\theta,t,T}(j)$ *is continuously three times differentiable with respect to* θ *and the derivatives satisfy*

$$\sup_{t,T}\left\{\sum_{j=0}^{\infty}(1+j)\left|\nabla_{i_1}\cdots\nabla_{i_s}b^{\circ}_{\theta,t,T}(j)\right|\right\} < \infty \quad \text{for } s = 0,1,2,3. \tag{6.305}$$

(C4)

$$a^{\circ}_{\theta,t,T}(0) = \exp\left[\frac{1}{4\pi}\int_{-\pi}^{\pi}\log\left\{f^{\circ}_{\theta,t,T}(\lambda)\right\}d\lambda\right], \tag{6.306}$$

where $f^{\circ}_{\theta,t,T}(\lambda) = \left|A^{\circ}_{\theta,t,T}(\lambda)\right|^2$.

By (6.302) and (6.303) we have

$$a^{\circ}_{\theta,t,T}(0)\varepsilon_t = \sum_{j=0}^{t-1} b^{\circ}_{\theta,t,T}(j)X_{t-j,T} + \sum_{r=0}^{\infty} c^{\circ}_{\theta,t,T}(r)\varepsilon_{-r}, \tag{6.307}$$

where

$$c^{\circ}_{\theta,t,T}(r) = \sum_{s=0}^{r} b^{\circ}_{\theta,t,T}(t+s)a^{\circ}_{\theta}(r-s) \tag{6.308}$$

and from Assumption 6.21 it follows that

$$\sum_{r=0}^{\infty} \left|c^{\circ}_{\theta,t,T}(r)\right| \leq \sum_{j=t}^{\infty}\sum_{l=0}^{\infty} \left|b^{\circ}_{\theta,t,T}(j)\right| \left|a^{\circ}_{\theta}(l)\right| \leq \sum_{l=0}^{\infty} \left|a^{\circ}_{\theta}(l)\right| t^{-1} \sum_{j=t}^{\infty} j \left|b^{\circ}_{\theta,t,T}(j)\right| = O(t^{-1}). \tag{6.309}$$

Let $\mathcal{P}_{\theta,T}$ and $\mathcal{P}_{\varepsilon}$ be the probability distributions of $(\varepsilon_s, s \leq 0, X_{1,T}, \ldots, X_{T,T})$ and $(\varepsilon_s, s \leq 0)$, respectively. It is easy to see that the linear transformation L_{θ} exists (whose Jacobian is $\prod_{t=1}^{T} a^{\circ}_{\theta,t,T}(0)^{-1}$), which maps $(\varepsilon_s, s \leq 0, X_{1,T}, \ldots, X_{T,T})$ into $(\varepsilon_s, s \leq T)$. Then recalling (6.307), we obtain

$$d\mathcal{P}_{\theta,T} = \prod_{t=1}^{T} \frac{1}{a^{\circ}_{\theta,t,T}(0)} p\left\{ \frac{\sum_{j=0}^{t-1} b^{\circ}_{\theta,t,T}(j) X_{t-j,T} + \sum_{r=0}^{\infty} c^{\circ}_{\theta,t,T}(r)\varepsilon_{-r}}{a^{\circ}_{\theta,t,T}(0)} \right\} d\mathcal{P}_{\varepsilon}. \tag{6.310}$$

Let $a^{\circ}_{\theta,t,T}(0)\varepsilon_t \equiv z_{\theta,t,T}$, then $z_{\theta,t,T}$ has p.d.f.

$$g_{\theta,t,T}(\cdot) = \frac{1}{a^{\circ}_{\theta,t,T}(0)} p\left(\frac{\cdot}{a^{\circ}_{\theta,t,T}(0)} \right). \tag{6.311}$$

Denote by $H(p;\theta)$ the hypothesis under which the underlying parameter is $\theta \in \Theta$ and the probability density of innovation ε_t is $p = p(\cdot)$. We define the contiguous alternative hypothesis

$$\theta_T = \theta + \frac{1}{\sqrt{T}} h, \quad h = (h_1, \ldots, h_r)' \in \mathcal{H} \subset \mathbf{R}^r. \tag{6.312}$$

For two hypothetical values $\theta, \theta_T \in \Theta$, the log-likelihood ratio is

$$\Lambda_T(\theta, \theta_T) \equiv \log \frac{d\mathcal{P}_{\theta_T,T}}{d\mathcal{P}_{\theta,T}} = 2\sum_{t=1}^{T} \log \Phi_{t,T}(\theta, \theta_T), \tag{6.313}$$

where

$$\Phi_{t,T}(\theta, \theta_T)^2 = \frac{g_{\theta_T,t,T}\left(\sum_{j=0}^{t-1} b^{\circ}_{\theta_T,t,T}(j) X_{t-j,T} + \sum_{r=0}^{\infty} c^{\circ}_{\theta_T,t,T}(r)\varepsilon_{-r}\right)}{g_{\theta,t,T}\left(\sum_{j=0}^{t-1} b^{\circ}_{\theta,t,T}(j) X_{t-j,T} + \sum_{r=0}^{\infty} c^{\circ}_{\theta,t,T}(r)\varepsilon_{-r}\right)} = \frac{g_{\theta_T,t,T}\left(z_{\theta,t,T} + q_{t,T}\right)}{g_{\theta,t,T}\left(z_{\theta,t,T}\right)} \tag{6.314}$$

with

$$\begin{aligned} q_{t,T} &= \sum_{j=1}^{t-1} \left\{ b^{\circ}_{\theta_T,t,T}(j) - b^{\circ}_{\theta,t,T}(j) \right\} X_{t-j,T} + \sum_{r=0}^{\infty} \left\{ c^{\circ}_{\theta_T,t,T}(r) - c^{\circ}_{\theta,t,T}(r) \right\} \varepsilon_{-r} \\ &= \sum_{j=1}^{t-1} \left\{ \frac{1}{\sqrt{T}} h' \nabla b^{\circ}_{\theta,t,T}(j) + \frac{1}{2T} h' \nabla^2 b^{\circ}_{\theta^*,t,T}(j) h \right\} X_{t-j,T} \\ &\quad + \sum_{r=0}^{\infty} \frac{1}{\sqrt{T}} h' \nabla c^{\circ}_{\theta^{**},t,T}(r) \varepsilon_{-r}. \end{aligned} \tag{6.315}$$

Here θ^* and θ^{**} are points on the segment between θ and θ_T. Now we can state the local asymptotic normality (LAN) for the class of locally stationary processes.

Theorem 6.22 (LAN) *Suppose that Assumptions 6.19-6.21 hold. Then the sequence of experiments*

$$\mathcal{E}_T = \left[\mathbf{R}^{\mathbf{Z}}, \mathcal{B}^{\mathbf{Z}}, \{ \mathcal{P}_{\theta,T} : \theta \in \Theta \subset \mathbf{R}^r \} \right], \quad T \in \mathbf{N}, \tag{6.316}$$

where $\mathcal{B}^{\mathbf{Z}}$ denotes the Borel σ-field on $\mathbf{R}^{\mathbf{Z}}$, is locally asymptotically normal and equicontinuous on compact subset $\mathcal{C}$ of $\mathcal{H}$. That is,

(i) *For all $\theta \in \Theta$, the log-likelihood ratio $\Lambda_T(\theta, \theta_T)$ admits, under $H(p;\theta)$, as $T \to \infty$, the asymptotic representation*

$$\Lambda_T(\theta, \theta_T) = h' \Delta_T(\theta) - \frac{1}{2} h' \Gamma(\theta) h + o_{\mathcal{P}}(1), \tag{6.317}$$

where

$$\begin{aligned} \Delta_T(\theta) = \sum_{t=1}^{T} \Bigg[& \frac{\phi(\varepsilon_t)}{\sqrt{T} a^{\circ}_{\theta,t,T}(0)} \sum_{j=1}^{t-1} \nabla b^{\circ}_{\theta,t,T}(j) X_{t-j,T} \\ & - \frac{\nabla a^{\circ}_{\theta,t,T}(0)}{\sqrt{T} a^{\circ}_{\theta,t,T}(0)} \left\{ 1 + \phi(\varepsilon_t) \varepsilon_t \right\} \Bigg] \end{aligned} \tag{6.318}$$

and

$$\begin{aligned} \Gamma(\theta) = \int_0^1 \Bigg[& \frac{\mathcal{F}(p)}{4\pi} \int_{-\pi}^{\pi} \frac{\{\nabla f_\theta(u,\lambda)\} \{\nabla f_\theta(u,\lambda)\}'}{|f_\theta(u,\lambda)|^2} d\lambda \\ & + \frac{1}{16\pi^2} \left[E\left\{ \varepsilon_t^2 \phi(\varepsilon_t)^2 \right\} - 2\mathcal{F}(p) - 1 \right] \\ & \left\{ \int_{-\pi}^{\pi} \frac{\nabla f_\theta(u,\lambda)}{f_\theta(u,\lambda)} d\lambda \right\} \left\{ \int_{-\pi}^{\pi} \frac{\nabla f_\theta(u,\lambda)}{f_\theta(u,\lambda)} d\lambda \right\}' \Bigg] du. \end{aligned} \tag{6.319}$$

(ii) *Under $H(p;\theta)$,*

$$\Delta_T(\theta) \xrightarrow{d} N(0, \Gamma(\theta)). \tag{6.320}$$

(iii) *For all $T \in \mathbf{N}$ and all $h \in \mathcal{H}$, the mapping $h \to \mathcal{P}_{\theta_T,\mathcal{T}}$ is continuous with respect to the variational distance*

$$\|\mathcal{P} - \mathcal{Q}\| = \sup\left\{|\mathcal{P}(A) - \mathcal{Q}(A)| : A \in \mathcal{B}^{\mathbf{Z}}\right\}. \tag{6.321}$$

Although Dahlhaus (1996b) proved the LAN theorem for Gaussian locally stationary processes, our LAN theorem elucidates various non-Gaussian asymptotics, and the proof contains a lot of different parts from that of the Gaussian case. However, we omit the proof because the theorem can be proved by checking Swensen's conditions (S1)-(S6) in line with the LAN theorem for CHARN models (Theorem 6.6).

As we saw in Section 6.2, once LAN is proved, the asymptotic optimality of estimators and tests is described in terms of the LAN property, namely is described in terms of the central sequence $\Delta_T(\theta)$. Now we construct an asymptotically efficient estimator in the sense of (6.116). Since $(\varepsilon_t, t \leq 0)$ are unobservable we use the "quasi-likelihood"

$$\mathcal{L}_T(\theta) = \prod_{t=1}^{T} \frac{1}{a^\circ_{\theta,t,T}(0)} p\left\{\frac{\sum_{j=0}^{t-1} b^\circ_{\theta,t,T}(j) X_{t-j,T}}{a^\circ_{\theta,t,T}(0)}\right\} \tag{6.322}$$

for estimation of θ. A quasi-likelihood estimator $\hat{\theta}_{QML}$ of θ is defined as a solution of the equation

$$\nabla\left[\sum_{t=1}^{T} \log p\left\{\frac{\sum_{j=0}^{t-1} b^\circ_{\theta,t,T}(j) X_{t-j,T}}{a^\circ_{\theta,t,T}(0)}\right\} - \log a^\circ_{\theta,t,T}(0)\right] = \mathbf{0} \tag{6.323}$$

with respect to θ. Then, the QMLE $\hat{\theta}_{QML}$ is asymptotically centering under Assumptions 6.19-6.21, hence we have the following result:

Theorem 6.23 *The QMLE $\hat{\theta}_{QML}$ for the locally stationary process is asymptotically efficient.*

As non-Gaussian innovation densities, a typical example is a logistic distribution. Consider the following logistic distribution $LG\left(0, \frac{\sqrt{3}}{\pi}\right)$, whose density is given by

$$p(x) = \frac{\pi \exp\left(\frac{-\pi x}{\sqrt{3}}\right)}{\sqrt{3}\left\{1 + \exp\left(\frac{-\pi x}{\sqrt{3}}\right)\right\}^2}. \tag{6.324}$$

To test the performance of the QMLE $\hat{\theta}_{QML}$, we carry out the simulation for the following model:

$$X_{t,T} + b_\theta\left(\frac{t}{T}\right) X_{t-1,T} = \varepsilon_t, \tag{6.325}$$

where $b_\theta(u) = \frac{1}{2}\cos\{(u-\theta)\pi\}$, $\theta \in (0,1)$ and ε_t's are i.i.d. $LG\left(0, \frac{\sqrt{3}}{\pi}\right)$ random variables. On the other hand the Gaussian quasi-likelihood estimator (GQMLE) $\hat{\theta}_{GQML}$ is defined by

$$\begin{aligned}\hat{\theta}_{GQML} &:= \arg\min_{\theta\in\Theta} \mathcal{G}_T(\theta) \\ &= \arg\min_{\theta\in\Theta}\left[\frac{1}{4\pi}\sum_{t=1}^{T}\int_{-\pi}^{\pi}\left\{\log f_\theta\left(t/T,\lambda\right) + \frac{\tilde{I}_T\left(t/T,\lambda\right)}{f_\theta\left(t/T,\lambda\right)}\right\}d\lambda\right], \quad (6.326)\end{aligned}$$

where

$$\tilde{I}_T\left(u,\lambda\right) := \frac{1}{2\pi}\sum_{j:1\le[uT+1/2\pm j/2]\le T} X_{[uT+1/2+j/2],T}X_{[uT+1/2-j/2],T}\exp(-i\lambda j), \quad (6.327)$$

and is given in Dahlhaus (2000) as a local version of the periodogram. Here $[x]$ denotes the largest integer less than or equal to x. Note that for this time varying AR(1) model, $\mathcal{G}_T(\theta)$ becomes

$$\sum_{t=1}^{T}\left\{\frac{1+b_\theta\left(\frac{t}{T}\right)^2}{2}X_{t,T}^2 + b_\theta\left(\frac{t}{T}\right)X_{t,T}X_{t+1,T} + \frac{1}{4\pi}\int_{-\pi}^{\pi}\log f_\theta\left(\frac{t}{T},\lambda\right)d\lambda\right\}. \quad (6.328)$$

The mean squared errors (MSE) of QMLE and GQMLE for $T = 2^8$, $\theta = 0.3$ and 100 times experiments are given in Table 6.3. From Table 6.3, it is seen that the MSE of QMLE is smaller than that of GQMLE if the innovation density is non-Gaussian.

Table 6.3 *The mean squared errors (MSE) of QMLE and GQMLE for* $T = 2^8$, $\theta = 0.3$ *and* 100 *times experiments.*

	$\hat{\theta}_{QML}$	$\hat{\theta}_{GQML}$
MSE	0.00263382	0.003906631

Next, we give an explicit example of locally asymptotically optimal test (c.f. Sakiyama and Taniguchi (2003)). Let $\mathcal{M}_0 \equiv \left\{\left(h^{(1)\prime}, \mathbf{0}'_{r-k}\right)' : h^{(1)} \in \mathbf{R}^k\right\}$. Consider the problem of testing the composite hypothesis

$$H : h = \left(h^{(1)\prime}, h^{(2)\prime}\right)' \in \mathcal{M}_0 \quad \text{against} \quad A : h \in \mathbf{R}^r - \mathcal{M}_0, \quad (6.329)$$

where $\mathbf{0}_{r-k}$ is the $(r-k)$-dimensional zero vector, that is,

$$H : \theta = \left(\theta_T^{(1)\prime}, \theta_0^{(2)\prime}\right)' \quad \text{against} \quad A : \theta = \left(\theta_T^{(1)\prime}, \theta_T^{(2)\prime}\right)' \quad (6.330)$$

with $\theta_T^{(i)} = \theta_0^{(i)} + \frac{1}{\sqrt{T}} h^{(i)}$, $i = 1, 2$.

Write $\hat{\theta}_{QML} = \left(\hat{\theta}_{QML}^{(1)\prime}, \hat{\theta}_{QML}^{(2)\prime} \right)'$ and we use the Wald test

$$W = T \left(\hat{\theta}_{QML}^{(2)} - \theta_0^{(2)} \right)' \left\{ B\Gamma(\hat{\theta}_{QML})^{-1} B' \right\}^{-1} \left(\hat{\theta}_{QML}^{(2)} - \theta_0^{(2)} \right), \tag{6.331}$$

where $B = (\mathbf{0}, I_{r-k})$ is a $(r-k) \times r$ matrix with I_{r-k}; the $(r-k) \times (r-k)$ identity matrix. Denote

$$\Gamma(\theta) = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{pmatrix}, \tag{6.332}$$

then

$$\begin{aligned} W &= \sqrt{T} \left(\hat{\theta}_{QML}^{(2)} - \theta_0^{(2)} \right)' \Gamma_{22.1}^{1/2\prime} \sqrt{T} \Gamma_{22.1}^{1/2} \left(\hat{\theta}_{QML}^{(2)} - \theta_0^{(2)} \right) + o_{\mathcal{P}_{\theta_0,T}}(1) \\ &\equiv K_W' K_W + o_{\mathcal{P}_{\theta_0,T}}(1), \end{aligned} \tag{6.333}$$

where $\Gamma_{22/1} \equiv \Gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12}$. We denote $\mathcal{P}_{\theta_T,T}$ with $\theta_T = \theta_0 + \frac{h}{\sqrt{T}}$ by $\mathcal{P}_{T,h}$ and the distribution law of a random vector Y_T under $\mathcal{P}_{T,h}$ by $L(Y_T|\mathcal{P}_{T,h})$. Since under $\mathcal{P}_{\theta_0,T}$,

$$\left(\sqrt{T} \left(\hat{\theta}_{QML} - \theta_0 \right)', \Lambda_T(\theta_0, \theta_T) \right)' \xrightarrow{d} N(\boldsymbol{m}, \Sigma), \tag{6.334}$$

where $\boldsymbol{m} = \left(\mathbf{0}', -\frac{1}{2} h'\Gamma(\theta_0)h \right)'$ and

$$\Sigma = \begin{pmatrix} \Gamma(\theta_0)^{-1} & h \\ h' & h'\Gamma(\theta_0)h \end{pmatrix}, \tag{6.335}$$

using LeCam's third lemma, we obtain

$$L \left(\sqrt{T} \left(\hat{\theta}_{QML} - \theta_0 \right) | \mathcal{P}_{T,h} \right) \xrightarrow{d} N \left(\left(h^{(1)\prime}, h^{(2)\prime} \right)', \Gamma(\theta_0)^{-1} \right). \tag{6.336}$$

Therefore, under $\mathcal{P}_{T,h}$,

$$K_W = \Gamma_{22.1}^{1/2} B \sqrt{T} \left(\hat{\theta}_{QML} - \theta_0 \right) \xrightarrow{d} N \left(\Gamma_{22.1}^{1/2} h^{(2)}, I_{r-k} \right), \tag{6.337}$$

hence, in terms of the original tests we have

$$W \xrightarrow{d} \chi_{r-k}^2 \left(h^{(2)\prime} \Gamma_{22.1} h^{(2)} \right), \tag{6.338}$$

where $\chi_q^2(\delta^2)$ is a non-central χ^2 distribution with q degree of freedom and non-centrality parameter δ^2. Thus, we have, under H,

$$W \xrightarrow{d} \chi_{r-k}^2 \tag{6.339}$$

and, under A,

$$W \xrightarrow{d} \chi_{r-k}^2 \left(h^{(2)\prime} \Gamma_{22.1} h^{(2)} \right). \tag{6.340}$$

From the construction of W, it is seen that W is asymptotically optimal in the sense of Theorem 6.9.

Because the asymptotics depend on the non-Gaussianity of the process, in the following we turn our attention to a non-Gaussian robustness. We discuss the non-Gaussian robustness for a class of statistics which have quadratic forms. For $\mathbf{X}_T = (X_{1,T}, \ldots, X_{T,T})'$, the concerned class of statistics is

$$\mathcal{F} = \left\{ \mathcal{B}_T; \mathcal{B}_T = \frac{1}{\sqrt{T}} \left\{ \mathbf{X}_T' \mathbf{B}_T \mathbf{X}_T - E\left(\mathbf{X}_T' \mathbf{B}_T \mathbf{X}_T \right) \right\} \right\}, \tag{6.341}$$

where $\mathbf{B}_T$ is a $T \times T$ matrix whose elements are

$$\left\{ \mathbf{B}_T \right\}_{st} = \frac{1}{2\pi} \int_{-\pi}^{\pi} B_{s,T}^{\circ}(\lambda) B_{t,T}^{\circ}(-\lambda) \exp\left\{ i(s-t)\lambda \right\} d\lambda \tag{6.342}$$

and B° fulfills Assumption 6.19. We assume that $\mathbf{X}_T$ and $\mathbf{B}_T$ have time varying spectral density $f_\theta(u,\lambda) = \frac{1}{2\pi} \left| A_\theta(u,\lambda) \right|^2$ and $f_B(u,\lambda) = \frac{1}{2\pi} \left| B_\theta(u,\lambda) \right|^2$, respectively. This class includes the main order term of QMLE, tests and discriminant statistics etc. (See Dahlhaus (1997), Sakiyama and Taniguchi (2004)). In this sense, it contains a lot of important statistics. Hence, it is a sufficiently rich class.

In addition to Assumption 6.21, let $\{X_{t,T}\}$ be the linear process defined by

$$X_{t,T} = \sum_{s \in \mathbf{Z}} \alpha_{\theta,t,T}^{\circ}(s) \varepsilon_{t-s}, \quad t = 1 \ldots, T, \tag{6.343}$$

where $\alpha_{\theta,t,T}^{\circ}(s)$ are real for all $t = 1 \ldots, T$, $s \in \mathbf{Z}$ and $\sum_{s \in \mathbf{Z}} |s| \left| \alpha_{\theta,t,T}^{\circ}(s) \right| < \infty$. Here ε_t, $t \in \mathbf{Z}$ is an i.i.d. sequence with mean 0, variance 1 and finite cumulants of all order, and satisfies Assumption 6.20. Write

$$A_{\theta,t,T}^{\circ}(\lambda) = \sum_{s \in \mathbf{Z}} \alpha_{\theta,t,T}^{\circ}(s) \exp(-is\lambda), \tag{6.344}$$

then

$$\alpha_{\theta,t,T}^{\circ}(s) = \frac{1}{2\pi} \int_{-\pi}^{\pi} A_{\theta,t,T}^{\circ}(\lambda) \exp(is\lambda) d\lambda. \tag{6.345}$$

First, we have the following result:

Theorem 6.24 *For $\mathcal{B}_T \in \mathcal{F}$, we have, under $H(p;\theta)$,*

$$\mathcal{B}_T \xrightarrow{d} N(0, \sigma^2), \tag{6.346}$$

where

$$\begin{aligned} \sigma^2 =& 16\pi^3 \int_0^1 \int_{-\pi}^{\pi} \left\{ f_\theta(u,\lambda) f_B(u,\lambda) \right\}^2 d\lambda du \\ &+ 4\pi^2 \kappa_4 \int_0^1 \left\{ \int_{-\pi}^{\pi} f_\theta(u,\lambda) f_B(u,\lambda) d\lambda \right\}^2 du. \end{aligned} \tag{6.347}$$

Here κ_4 is the fourth-order cumulant of ε_t.

PROOF

We define $H^{\circ}_{t,T}(\lambda)$, $H(u,\lambda)$, $\lambda \in [-\pi,\pi]$ and $h^{\circ}_{t,T}(l)$ as

$$H^{\circ}_{t,T}(\lambda) = A^{\circ}_{\theta,t,T}(\lambda) B^{\circ}_{t,T}(-\lambda), \tag{6.348}$$

$$H(u,\lambda) = A_{\theta}(u,\lambda) B(u,-\lambda) \tag{6.349}$$

and

$$h^{\circ}_{t,T}(l) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H^{\circ}_{t,T}(\lambda) \exp(il\lambda) d\lambda, \tag{6.350}$$

respectively. Let $\tilde{X}_{t,T} = \sum_{l \in \mathbf{Z}} h^{\circ}_{t,T}(l) \varepsilon_{t-l}$ and

$$S_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left\{ \tilde{X}^2_{t,T} - E\left(\tilde{X}^2_{t,T} \right) \right\},$$

then Theorem 6.24 follows from the lemmas below.

Lemma 6.6

$$\mathrm{Var}\left(\mathcal{B}_T - S_T \right) = o(1). \tag{6.351}$$

Lemma 6.7

$$S_T \xrightarrow{d} N(0, \sigma^2). \tag{6.352}$$

PROOF OF LEMMA 6.6

Consider an infinite sum

$$S = \sum_{l_1, l_2 \in \mathbf{Z}} c(l_1, l_2) \left\{ \varepsilon_{l_1} \varepsilon_{l_2} - E\left(\varepsilon_{l_1} \varepsilon_{l_2} \right) \right\} \tag{6.353}$$

with real coefficients $c(l_1, l_2)$. Then

$$\mathrm{Var}(S) = \sum_{l_1 \neq l_2} 2c(l_1, l_2)^2 + (2 + \kappa_4) \sum_{l \in \mathbf{Z}} c(l,l)^2 \tag{6.354}$$

where C is some constant. Set

$$d_{1,T}(l_1, l_2) = \sum_{s,t=1}^{T} \left\{ B^T \right\}_{st} \alpha^{\circ}_{\theta,s,T}(s - l_1) \alpha^{\circ}_{\theta,t,T}(t - l_2) \tag{6.355}$$

and

$$\begin{aligned} d_{2,T}(l_1, l_2) &= \sum_{s,t=1}^{T} \delta_{s,t} h^{\circ}_{s,T}(s - l_1) h^{\circ}_{t,T}(t - l_2) \\ &= \sum_{t=1}^{T} h^{\circ}_{t,T}(t - l_1) h^{\circ}_{t,T}(t - l_2), \end{aligned} \tag{6.356}$$

then, by (6.354), we have

$$\mathrm{Var}\left(\mathcal{B}_T - S_T\right) \leq \frac{C}{T} \sum_{l_1,l_2 \in \mathbf{Z}} \left\{d_{1,T}(l_1,l_2) - d_{2,T}(l_1,l_2)\right\}^2. \tag{6.357}$$

Writing

$$\begin{aligned} D_{1,T}(\lambda,\mu) &= \sum_{s,t=1}^{T} \left\{B^T\right\}_{st} A^{\circ}_{\theta,s,T}(\lambda) A^{\circ}_{\theta,t,T}(\mu) \exp\left\{i(s\lambda + t\mu)\right\} \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{s,t=1}^{T} B^{\circ}_{s,T}(\nu) B^{\circ}_{t,T}(-\nu) A^{\circ}_{\theta,s,T}(\lambda) A^{\circ}_{\theta,t,T}(\mu) \\ &\quad \exp\left[i\left\{s(\lambda+\nu) + t(\mu-\nu)\right\}\right] d\nu, \end{aligned} \tag{6.358}$$

$$\begin{aligned} D_{2,T}(\lambda,\mu) &= \sum_{s,t=1}^{T} \delta_{s,t} H^{\circ}_{s,T}(\lambda) H^{\circ}_{t,T}(\mu) \exp\left\{i(s\lambda + t\mu)\right\} \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{s,t=1}^{T} H^{\circ}_{s,T}(\lambda) H^{\circ}_{t,T}(\mu) \exp\left[i\left\{s(\lambda+\nu) + t(\mu-\nu)\right\}\right] d\nu \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{s,t=1}^{T} A^{\circ}_{\theta,s,T}(\lambda) B^{\circ}_{s,T}(-\lambda) A^{\circ}_{\theta,t,T}(\mu) B^{\circ}_{t,T}(-\mu) \\ &\quad \exp\left[i\left\{s(\lambda+\nu) + t(\mu-\nu)\right\}\right] d\nu, \end{aligned} \tag{6.359}$$

we can see that

$$\begin{aligned} &d_{1,T}(l_1,l_2) - d_{2,T}(l_1,l_2) \\ &= \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \left\{D_{1,T}(\lambda,\mu) - D_{2,T}(\lambda,\mu)\right\} \exp\left\{-i(l_1\lambda + l_2\mu)\right\} d\lambda d\mu. \end{aligned} \tag{6.360}$$

By Parseval's identity it follows that

$$\begin{aligned} &\sum_{l_1,l_2 \in \mathbf{Z}} \left\{d_{1,T}(l_1,l_2) - d_{2,T}(l_1,l_2)\right\}^2 \\ &= \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \left\{D_{1,T}(\lambda,\mu) - D_{2,T}(\lambda,\mu)\right\}^2 d\lambda d\mu. \end{aligned} \tag{6.361}$$

Let $L_T : \mathbf{R} \to \mathbf{R}$, $T \in \mathbf{N}$ be the periodic extension (with period 2π) of

$$L^*_T(\lambda) := \begin{cases} T & 0 \leq |\lambda| \leq \frac{\pi}{T}, \\ \frac{\pi}{\lambda} & \frac{\pi}{T} \leq |\lambda| \leq \pi. \end{cases} \tag{6.362}$$

According to Lemma 4.2 of Dahlhaus 1996a and using the fact $|\lambda| L_T(\lambda) \leq C$,

we can see that

$$\begin{aligned} & D_{1,T}(\lambda,\mu) - D_{2,T}(\lambda,\mu) \\ & \le \int_{-\pi}^{\pi} L_T(\lambda+\nu) L_T(\mu-\nu) \left\{ |\nu+\lambda| + |\mu-\nu| \right\} d\nu \\ & \le C' \int_{-\pi}^{\pi} \left\{ L_T(\lambda+\nu) + L_T(\mu-\nu) \right\} d\nu = O\left(\log T\right), \end{aligned} \tag{6.363}$$

hence, $\mathrm{Var}\left(\mathcal{B}_T - S_T\right) = O\left(T^{-1}(\log T)^2\right) = o(1)$. □

PROOF OF LEMMA 6.7

First we prove

$$\mathrm{Var}(S_T) \to \sigma^2, \quad \text{as } T \to \infty. \tag{6.364}$$

Note that

$$\begin{aligned} \mathrm{Var}(S_T) &= \frac{2}{T} \sum_{l_1, l_2 \in \mathbf{Z}} \left\{ d_{2,T}(l_1, l_2) \right\}^2 + \frac{\kappa_4}{T} \sum_{l \in \mathbf{Z}} \left\{ d_{2,T}(l,l) \right\}^2 \\ &\equiv E^{(1)} + E^{(2)}. \end{aligned} \tag{6.365}$$

$$\begin{aligned} E^{(1)} =& \frac{1}{2\pi^2 T} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \left\{ D_{2,T}(\lambda,\mu) \right\}^2 d\lambda d\mu \\ =& \frac{1}{2\pi^2 T} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \sum_{s,t=1}^{T} H^{\circ}_{s,T}(\lambda) H^{\circ}_{s,T}(\omega-\lambda) H^{\circ}_{t,T}(-\lambda) H^{\circ}_{t,T}(\lambda-\omega) \\ & \exp\left\{ i(t-s)\omega \right\} d\lambda d\omega \\ =& \frac{1}{2\pi^2 T} \int_{-\pi}^{\pi} \sum_{s,t=1}^{T} \left| H^{\circ}_{s,T}(\lambda) \right|^2 \left| H^{\circ}_{t,T}(\lambda) \right|^2 d\lambda \int_{-\pi}^{\pi} \exp\left\{ i(t-s)\omega \right\} d\omega \\ & + O\left(\frac{|\omega| L_T(\omega)^2}{T} d\omega \right) \\ =& \frac{1}{\pi T} \int_{-\pi}^{\pi} \sum_{t=1}^{T} \left| H^{\circ}_{t,T}(\lambda) \right|^4 d\lambda + O\left(\frac{\log T}{T} \right) \\ =& \frac{1}{\pi T} \sum_{t=1}^{T} \int_{-\pi}^{\pi} \left| A_\theta\left(\frac{t}{T}, \lambda \right) B\left(\frac{t}{T}, \lambda \right) \right|^4 d\lambda + O\left(T^{-1} + \frac{\log T}{T} \right) \\ =& 16\pi^3 \int_0^1 \int_{-\pi}^{\pi} \left\{ f_\theta(u,\lambda) f_B(u,\lambda) \right\}^2 d\lambda du + o\left(1\right). \end{aligned} \tag{6.366}$$

On the other hand, we can rewrite (6.359) as

$$D_{2,T}(\lambda) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{t=1}^{T} H^{\circ}_{t,T}(\mu) H^{\circ}_{t,T}(\lambda-\mu) \exp(it\lambda) d\mu. \tag{6.367}$$

Since

$$d_{2,T}(l,l) = \frac{1}{2\pi}\int_{-\pi}^{\pi} D_{2,T}(\lambda)\exp(-il\lambda)d\lambda, \tag{6.368}$$

we observe that

$$\begin{aligned}
E^{(2)} =& \frac{\kappa_4}{2\pi T}\int_{-\pi}^{\pi}\{D_{2,T}(\lambda)\}^2 d\lambda \\
=& \frac{\kappa_4}{8\pi^3 T}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\sum_{s,t=1}^{T} H^{\circ}_{s,T}(\mu_1)H^{\circ}_{s,T}(\lambda-\mu_1)H^{\circ}_{t,T}(-\mu_2)H^{\circ}_{t,T}(\mu_2-\lambda) \\
& \exp\{i(t-s)\lambda\}\, d\lambda d\mu_1 d\mu_2 \\
=& \frac{\kappa_4}{8\pi^3 T}\sum_{s,t=1}^{T}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}|H^{\circ}_{s,T}(\mu_1)|^2|H^{\circ}_{t,T}(\mu_2)|^2 d\mu_1 d\mu_2 \int_{-\pi}^{\pi}\exp\{i(t-s)\lambda\}\, d\lambda \\
& + O\left(\int_{-\pi}^{\pi}\frac{|\lambda| L_T^2(\lambda)}{T}d\lambda\right) \\
=& 4\pi^2\kappa_4\int_0^1\left\{\int_{-\pi}^{\pi} f_\theta(u,\mu)f_B(u,\mu)d\mu\right\}^2 du + o(1).
\end{aligned} \tag{6.369}$$

Next we prove, for $k \geq 3$

$$\operatorname{cum}(S_T,\ldots,S_T) = T^{-\frac{k}{2}}\sum_{t_1,\ldots,t_k=1}^{T}\operatorname{cum}\left(\tilde{X}^2_{t_1,T},\ldots,\tilde{X}^2_{t_k,T}\right) = o(1). \tag{6.370}$$

Using the product theorem for cumulants, we have

$$\begin{aligned}
&\operatorname{cum}(\tilde{X}^2_{t_1,T},\ldots,\tilde{X}^2_{t_k,T}) \\
&= \sum_{\boldsymbol{\nu}}\operatorname{cum}(\tilde{X}_{(t_i,j),T};(t_i,j)\in\nu_1)\cdots\operatorname{cum}(\tilde{X}_{(t_i,j),T};(t_i,j)\in\nu_p),
\end{aligned} \tag{6.371}$$

where $\tilde{X}_{(t_i,j),T} = \tilde{X}_{t_i,T}$, $j = 1,2$, for all $i = 1,\ldots,k$ and the summation is over all indecomposable partitions $\boldsymbol{\nu} = \nu_1 \cup \cdots \cup \nu_p$ with $|\nu_r| \geq 2$ of the table

$$\begin{matrix}(t_1,1) & (t_1,2)\\ \vdots & \vdots \\ (t_k,1) & (t_k,2).\end{matrix} \tag{6.372}$$

Note that, for $s \geq 2$,

$$\begin{aligned}
\operatorname{cum}(\widetilde{X}_{l_1,T},\ldots,\widetilde{X}_{l_s,T}) =& \kappa_s\sum_{l\in\mathbf{Z}} h^{\circ}_{l_1,T}(l_1-l)\cdots h^{\circ}_{l_s,T}(l_s-l) \\
=& \frac{\kappa_s}{(2\pi)^{s-1}}\int_{\Pi^{s-1}} H^{\circ}_{l_1,T}(\lambda_1)\cdots H^{\circ}_{l_{s-1},T}(\lambda_{s-1})H^{\circ}_{l_s,T}(-\lambda_1-\cdots-\lambda_{s-1}) \\
& \exp\{i(l_1\lambda_1+\cdots+l_{s-1}\lambda_{s-1})\}\exp\{il_s(-\lambda_1-\cdots-\lambda_{s-1}\}\, d\lambda_1\cdots d\lambda_{s-1},
\end{aligned} \tag{6.373}$$

where κ_s is the sth order cumulant of ε_t. Then we can see that

$$T^{-\frac{k}{2}} \sum_{t_1,\ldots,t_k=1}^{T} \mathrm{cum}(\tilde{X}^2_{t_1,T},\ldots,\tilde{X}^2_{t_k,T}) = O\left(\frac{(\log T)^{k-2}}{T^{\frac{k}{2}}}\int_{-\pi}^{\pi} L_T^2(\lambda)d\lambda\right) = O\left(\frac{(\log T)^{k-2}}{T^{\frac{k-2}{2}}}\right) = o(1), \tag{6.374}$$

which implies the asymptotic normality of S_T. □

We are often interested in the local asymptotics under $\theta_T = \theta + \frac{h}{\sqrt{T}}$, for $\mathcal{B}_T \in \mathcal{F}$. Since under $H(p;\theta)$,

$$(\mathcal{B}_T, \Lambda_T)' \xrightarrow{d} N(\nu, \Sigma), \tag{6.375}$$

where

$$\nu = \left(0, \frac{1}{2}h'\Gamma(\theta)h\right)', \tag{6.376}$$

$$\Sigma = \begin{pmatrix} \sigma & m \\ m & h'\Gamma(\theta)h \end{pmatrix} \tag{6.377}$$

and

$$\begin{aligned} m =& \frac{3}{2}\left[E\left\{\phi(\varepsilon_t)\varepsilon_t\right\} - E\left\{\phi(\varepsilon_t)\varepsilon_t^3\right\}\right] \\ & \int_0^1 \left\{\int_{-\pi}^{\pi} h'\nabla \log f_\theta(u,\lambda)d\lambda \int_{-\pi}^{\pi} f_\theta(u,\lambda)f_B(u,\lambda)d\lambda\right\} du \\ & - 4\pi E\left\{\phi(\varepsilon_t)\varepsilon_t\right\}\int_0^1\int_{-\pi}^{\pi} f_\theta(u,\lambda)f_B(u,\lambda)\frac{h'\nabla A_\theta(u,\lambda)}{A_\theta(u,\lambda)}d\lambda du, \end{aligned} \tag{6.378}$$

using LeCam's third lemma, we obtain the following theorem.

Theorem 6.25 *Under $H(p;\theta_T)$, $\theta_T = \theta + \frac{h}{\sqrt{T}}$, the limiting distribution of $\mathcal{B}_T \in \mathcal{F}$ is given by*

$$\mathcal{B}_T \xrightarrow{d} N(m, \sigma^2). \tag{6.379}$$

The above theorems show that the asymptotics of $\mathcal{B}_T \in \mathcal{F}$ depend on non-Gaussian quantities κ_4, $E\left\{\phi(\varepsilon_t)\varepsilon_t\right\}$ and $E\left\{\phi(\varepsilon_t)\varepsilon_t^3\right\}$. Henceforth we say that $\mathcal{B}_T \in \mathcal{F}$ is *non-Gaussian robust* if the asymptotics are independent of the non-Gaussian quantities. To describe the non-Gaussian robustness we introduce the following concept. We say that θ is *innovation-free* if

$$\int_{-\pi}^{\pi} \frac{\nabla f_\theta(u,\lambda)}{f_\theta(u,\lambda)}d\lambda = \mathbf{0}. \tag{6.380}$$

This condition is satisfied if the time varying innovation $a^\circ_{\theta,t,T}(0)\varepsilon_t$ is independent of θ.

Theorem 6.26 *Assume that (i) θ is innovation-free, (ii) $f_B(u,\lambda) = \frac{h'\nabla f_\theta(u,\lambda)}{|f_\theta(u,\lambda)|}$ and (iii) $\lim_{|z|\to\infty} zp(z) = 0$. Then the asymptotic mean and variance of $\mathcal{B}_T \in \mathcal{F}$ become*

$$m = 2\pi \int_0^1 \int_{-\pi}^{\pi} \left\{ \frac{h'\nabla f_\theta(u,\lambda)}{f_\theta(u,\lambda)} \right\}^2 d\lambda du \tag{6.381}$$

and

$$\sigma^2 = 16\pi^3 \int_0^1 \int_{-\pi}^{\pi} \left\{ \frac{h'\nabla f_\theta(u,\lambda)}{f_\theta(u,\lambda)} \right\}^2 d\lambda du, \tag{6.382}$$

respectively, hence $\mathcal{B}_T$ is non-Gaussian robust.

To observe the non-Gaussian effect of $\mathcal{B}_T$, we consider the following model:

$$X_{t,T} + b_\theta\left(\frac{t}{T}\right) X_{t-1,T} = a_\theta\left(\frac{t}{T}\right)\varepsilon_t, \quad t = 1,\ldots,T, \tag{6.383}$$

where $a_\theta(u) = a\exp\left\{-\frac{(u-\theta)^2}{2}\right\}$, $b_\theta(u) = b\cos(u\theta)$, $|a|,|b| < 1$ and ε_t's are i.i.d. bilateral exponential random variable whose probability density is

$$p(z) = \frac{1}{\sqrt{2}} \exp\left(-\sqrt{2}|z|\right). \tag{6.384}$$

Then the time varying spectral density is given by

$$f_\theta(u,\lambda) = \frac{1}{2\pi}\left|\frac{a_\theta(u)}{1 + b_\theta(u)e^{-i\lambda}}\right|^2. \tag{6.385}$$

Let $f_B(u,\lambda) = \frac{h'\nabla f_\theta(u,\lambda)}{|f_\theta(u,\lambda)|}$, then

$$\begin{aligned}\sigma^2 =& 16\pi^3 \int_0^1 \int_{-\pi}^{\pi} \left\{ \frac{h'\nabla f_\theta(u,\lambda)}{f_\theta(u,\lambda)} \right\}^2 d\lambda du \\ &+ 24\pi^2 \int_0^1 \left\{ \int_{-\pi}^{\pi} \frac{h'\nabla f_\theta(u,\lambda)}{f_\theta(u,\lambda)} d\lambda \right\}^2 du \\ =& \sigma_G^2 + 384\pi^4 \int_0^1 \left\{ \frac{h'\nabla a_\theta(u)}{a_\theta(u)} \right\}^2 du = \sigma_G^2 + \sigma_{NG}^2, \quad \text{(say)}. \end{aligned} \tag{6.386}$$

The quantity σ_{NG}^2 is non-Gaussian effect on the asymptotic variance σ^2 of $\mathcal{B}_\mathcal{T}$. σ_{NG}^2 for model (6.383) is plotted in Figure 6.19. From the figure we observe that the non-Gaussian effect becomes large as $|h|$ and $\left|\theta - \frac{1}{2}\right|$ tend to large.

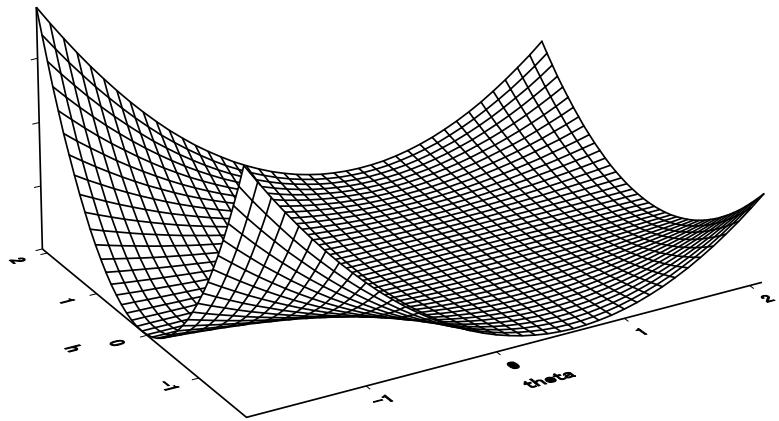


Figure 6.19 $\sigma_{NG}^2 = 384\pi^4 h^2(\theta^2 - \theta + 1/3)$ *for model (6.383).*

Another application of the LAN is adaptive estimation, namely we are led to construct asymptotic efficient estimators when the innovation density $p(\cdot)$ is unknown. Henceforth we denote the true value of θ by θ_0. For the sake of simplicity we assume that $a^{\circ}_{\theta,t,T} \equiv 1$, for all $t = 1, \ldots, T$, namely

$$\varepsilon_t(\theta) = \sum_{j=0}^{\infty} b^{\circ}_{\theta,t,T}(j) X_{t-j,T}. \tag{6.387}$$

Then, the LAN property of Theorem 6.22 is rewritten as

Theorem 6.27 (i) *For all $\theta \in \Theta$, the log-likelihood ratio $\Lambda_T(\theta, \theta_T)$ admits, under $H(p;\theta)$, as $T \to \infty$, the asymptotic representation*

$$\Lambda_T(\theta, \theta_T) = h' \Delta_T(\theta) - \frac{1}{2} h' \mathcal{F}(p) \Gamma(\theta) h + o_{\mathcal{P}}(1), \tag{6.388}$$

where

$$\Delta_T(\theta) = \sum_{t=1}^{T} \frac{\phi(\varepsilon_t(\theta))}{\sqrt{T} a^{\circ}_{\theta,t,T}(0)} \sum_{j=1}^{t-1} \nabla b^{\circ}_{\theta,t,T}(j) X_{t-j,T} \tag{6.389}$$

and

$$\Gamma(\theta) = \frac{1}{4\pi} \int_0^1 \int_{-\pi}^{\pi} \frac{\{\nabla f_\theta(u,\lambda)\} \{\nabla f_\theta(u,\lambda)\}'}{|f_\theta(u,\lambda)|^2} d\lambda du. \tag{6.390}$$

(ii) *Under $H(p;\theta)$,*

$$\Delta_T(\theta) \xrightarrow{d} N(0, \mathcal{F}(p)\Gamma(\theta)). \tag{6.391}$$

We first construct asymptotically efficient estimators of θ_0. For this purpose, the existence of $\sqrt{T}$-consistent initial estimators $\left\{\tilde{\theta}_T\right\}$ is essential.

Assumption 6.22 *There exists a sequence of estimators* $\left\{\tilde{\theta}_T\right\}$ *which satisfies*

$$\sqrt{T}\left(\tilde{\theta}_T - \theta_0\right) = O_{\mathcal{P}_{\theta_0}}(1).$$

For technical reasons we restrict ourselves to discretized estimators:

Definition 6.2 *For any sequence of estimators* $\left\{\tilde{\theta}_T\right\}$*, define the discretized estimator* $\left\{\overline{\theta}_T\right\}$ *to be the nearest vertices of* $\left\{\theta : \theta = T^{-1/2}(i_1, \ldots, i_r), i_j \in \mathbf{Z}\right\}$.

The reason for introducing this concept is that using discretized estimators, we can establish the validity of the Newton-Raphson type estimators without introducing additional differentiability or boundedness assumptions. The great advantage of discretized estimators is the following result, which goes back to LeCam (See e.g. Kreiss (1987) and Linton (1993)).

Lemma 6.8 *Assume that* $\{S_T(\theta), T \in \mathbf{N}\}$ *is a sequence of random variables which depends on* $\theta \in \Theta$*. If for each sequence* $\theta_T \in \Theta$ *satisfying that*

$$\sqrt{T}\left(\theta_T - \theta_0\right) \text{ is bounded by a constant } c > 0, \tag{6.392}$$

we have $S_T\left(\theta_T\right) = o_{\mathcal{P}_{\theta_0}}(1)$*, then* $S_T\left(\overline{\theta}_T\right) = o_{\mathcal{P}_{\theta_0}}(1)$ *holds for any discrete estimators* $\left\{\overline{\theta}_T\right\}$ *which are* $\sqrt{T}$*-consistent.*

Next we make the following assumption for the score function $\phi = \frac{Dp}{p}$.

Assumption 6.23 (i)

$$\lim_{u \to 0} \int \left\{\phi(z+u) - \phi(z)\right\}^2 p(z) dz = 0, \tag{6.393}$$

(ii)

$$\lim_{u \to 0} \int \frac{\phi(z-u) - \phi(z)}{u} p(z) dz = \mathcal{F}(p). \tag{6.394}$$

Now, we can establish the following result:

Theorem 6.28 *Assume that Assumptions 6.19-6.23 hold and that* $\left\{\overline{\theta}_T\right\}$ *is discretized and* $\sqrt{T}$*-consistent for* $\theta_0 \in \Theta$*. Then the estimator* $\hat{\theta}_T$ *defined by (6.395) below is asymptotically efficient:*

$$\hat{\theta}_T = \overline{\theta}_T + \frac{\hat{\Gamma}_T\left(\overline{\theta}_T\right)^{-1}}{\sqrt{T}\mathcal{F}(p)} \Delta_T\left(\overline{\theta}_T\right), \tag{6.395}$$

where

$$\hat{\Gamma}_T\left(\overline{\theta}_T\right) = \frac{1}{T} \sum_{t=1}^{T} \tilde{W}_{t,T}\left(\overline{\theta}_T\right) \tilde{W}_{t,T}\left(\overline{\theta}_T\right)' \tag{6.396}$$

and

$$\tilde{W}_{t,T}\left(\overline{\theta}_T\right) = \sum_{j=1}^{t-1} \nabla b^{\circ}_{\theta,t,T}(j) X_{t-j,T}. \tag{6.397}$$

Until up to now we have dealt with the asymptotically efficient estimators when innovation density is known. Here we discuss estimation of θ_0 in the case when $p(\cdot)$ is unknown. To simplify the problem, we further assume the following:

Assumption 6.24 **(i)** *p is symmetric about the origin,*
(ii) $\int z^4 p(z)dz < \infty$.

Introduce the following nonparametric density estimator:

$$\begin{aligned}\hat{p}_{\tau,t}(z;\theta) &= \frac{1}{2(T-1)} \sum_{s=1,s\neq t}^{T} \left[g\left(z+\hat{\varepsilon}_t(\theta);\tau\right) + g\left(z-\hat{\varepsilon}_t(\theta);\tau\right)\right] \\ &= \frac{1}{2(T-1)} \sum_{s=1,s\neq t}^{T} \left[g\left(z+\varepsilon_t(\theta);\tau\right) + g\left(z-\varepsilon_t(\theta);\tau\right)\right] + o_{\mathcal{P}_\theta}(1), \end{aligned} \tag{6.398}$$

for $t = 1, \ldots, T$, where

$$g(z;\tau) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{z^2}{2\tau^2}\right), \quad z \in \mathbf{R} \tag{6.399}$$

and

$$\begin{aligned}\hat{\varepsilon}_t(\theta) &= \sum_{j=0}^{t-1} b^{\circ}_{\theta,t,T}(j) X_{t-j,T} \\ &= \varepsilon_t(\theta) - \sum_{s=0}^{\infty} c^{\circ}_{\theta,t,T}(s)\varepsilon_{-s} \\ &= \varepsilon_t(\theta) + O_{\mathcal{P}_\theta}\left(t^{-1}\right). \end{aligned} \tag{6.400}$$

Denote

$$p_\tau(z) = E_\theta\left\{\hat{p}_{\tau,t}(z;\theta)\right\} = \int_{-\infty}^{\infty} g(z-y;\tau)p(y)dy + o(1). \tag{6.401}$$

As an estimator of the score function ϕ, define

$$\hat{q}_{t,T}(z;\theta) = \frac{D\hat{p}_{\tau(T),t}(z;\theta)}{\hat{p}_{\tau(T),t}(z;\theta)}, \tag{6.402}$$

if

$$\hat{p}_{\tau(T),t}(z;\theta) \geq c_T, \quad \left|D\hat{p}_{\tau(T),t}(z;\theta)\right| \leq d_T \hat{p}_{\tau(T),t}(z;\theta) \quad \text{and} \quad |z| \leq e_T, \tag{6.403}$$

and $\hat{q}_{t,T}(z;\theta) = 0$, otherwise, where $d_T \to \infty$, $e_T \to \infty$, $\tau(T) \to 0$ and $c_T \to 0$. Let

$$\tilde{\Delta}_T(\theta) = \sum_{t=1}^{T} \frac{\hat{q}_{t,T}\left\{\hat{\varepsilon}_t(\theta);\theta\right\}}{\sqrt{T}} \tilde{W}_{t,T}(\theta) \tag{6.404}$$

be an estimated version of the central sequence $\Delta_T(\theta)$, and

$$\hat{\mathcal{F}}_T = \frac{1}{T}\sum_{t=1}^{T} \hat{q}_{t,T}\left\{\hat{\varepsilon}_t(\overline{\theta}_T);\overline{\theta}_T\right\}^2, \tag{6.405}$$

(c.f., Kreiss (1987)), which is a consistent estimator of $\mathcal{F}(p)$. Let

$$\hat{\overline{\theta}}_T = \overline{\theta}_T + \frac{\hat{\Gamma}_T\left(\overline{\theta}_T\right)^{-1}}{\sqrt{T}\hat{\mathcal{F}}_T} \tilde{\Delta}_T\left(\overline{\theta}_T\right), \tag{6.406}$$

where $\left\{\overline{\theta}_T\right\}$ is a discretized and $\sqrt{T}$-consistent sequence of estimators of θ_0.

Theorem 6.29 *Assume that Assumptions 6.19-6.24 hold and that* $\left\{\overline{\theta}_T\right\}$ *is a discretized and* $\sqrt{T}$*-consistent sequence of estimators of* θ_0*. Then*

$$\sqrt{T}\left(\hat{\overline{\theta}}_T - \theta_0\right) = \frac{\Gamma(\theta_0)^{-1}}{\mathcal{F}(p)} \Delta_T(\theta_0) + o_{\mathcal{P}_{\theta_0}}(1) \tag{6.407}$$

holds, if $d_T \to \infty$, $e_T \to \infty$, $\tau(T) \to 0$, $c_T \to 0$, $\tau(T)d_T \to 0$, $e_T\{\tau(T)\}^{-4}T^{-1} \to 0$ *and* $T\{\tau(T)\}^9$ *stays bounded.*

Evidently $\hat{\overline{\theta}}_T$ is asymptotically efficient although the density p is unknown. We call $\hat{\overline{\theta}}_T$ an *adaptive estimator.*

Note that from (6.391), $\Delta_T(\theta_0) \xrightarrow{d} N(0, \mathcal{F}(p)\Gamma(\theta_0))$, hence, in view of (6.407) the asymptotic variance of the normalized adaptive estimator, i.e., $\sqrt{T}\left(\hat{\overline{\theta}}_T - \theta_0\right)$, is $\{\mathcal{F}(p)\Gamma(\theta_0)\}^{-1}$, which can be estimated by $\left\{\hat{\mathcal{F}}_T\hat{\Gamma}_T\left(\overline{\theta}_T\right)\right\}^{-1}$.

6.10 Semiparametric Estimation

The problem of efficiently estimating the coefficients in a linear regression model has been investigated widely. When the error covariance matrix depends on unknown parameters, the regression coefficients are often estimated by generalized least squares (GLS), using appropriate consistent estimators of the residual parameters. It is well known that standardized GLS estimators have the same limiting distribution as the best linear unbiased estimator. Rothenberg (1984) gave higher order approximations to the distribution of GLS estimators. Toyooka (1985,1986) derived the asymptotic expansion of the mean squared errors (MSE). Since these methods are parametric, the

standard root N asymptotics hold for time domain GLS estimators, where N is the sample size.

If the autocorrelation structure of the unobservable residuals is not parameterized, we then construct efficient estimators by spectral methods. This technique is semiparametric since it relies on a nonparametric spectral estimator of the residuals.

The semiparametric method of a linear regression model was introduced by Hannan (1963), who showed that a frequency domain GLS estimator achieves asymptotically the Gauss-Markov efficiency bound under smoothness and Grenander's conditions on the residual spectral density and the regressor sequence, respectively.

There are differences between parametric and nonparametric estimation technique that are often given in terms of consistency and rates of convergence. Velasco and Robinson (2001) derived Edgeworth expansions for the distribution of nonparametric estimates. Taniguchi et al. (2003) discussed higher order asymptotic theory for minimum contrast estimators of spectral parameters. They established that for semiparametric estimation it does not hold in general that first-order efficiency implies second-order efficiency.

The semiparametric estimation entails the problem of the bandwidth selection. Applications of higher order asymptotic expansions to this problem have been studied by many authors. Robinson (1991) studied frequency domain inference on semiparametric and nonparametric models in the presence of a data dependent bandwidth. Linton (1995) investigated the second-order properties of various quantities in a partially linear model. Xiao and Phillips (1998) gave higher order approximations of the MSE of the frequency domain GLS estimators. Linton and Xiao (2001) derived asymptotic expansions for semiparametric adaptive regression estimators. They discussed the bandwidth selection based on minimizing the integrated MSE. Also Xiao and Phillips (2002) discussed higher order approximations for Wald statistics in frequency domain regressions with integrated processes.

Taniguchi et al. (1996) established the root N asymptotic theory for functionals of nonparametric spectral density estimators. This is due to the fact that integration of nonparametric spectral density estimators recovers root N consistency. Since the Hannan estimator is based on integral functionals of nonparametric estimators, it may be expected that the Hannan estimator has attractive properties in higher order asymptotic theory.

In this section, we will develop the second-order asymptotic theory for the frequency domain GLS estimator proposed by Hannan (1963). First, we give the second-order Edgeworth expansion of the distribution of the Hannan estimator. Next, we show that the bias-adjusted version of the Hannan estimator is not second-order asymptotically Gaussian efficient in general. Of course, if the residual is Gaussian, it is second-order asymptotically efficient. As in

Xiao and Phillips (1998), if the error is a Gaussian process, then it holds that first-order efficiency implies second-order efficiency.

An interesting result is that the second-order asymptotic properties are independent of the bandwidth choice for the residual spectral estimator. This implies that the Hannan estimator has the same rate of convergence as in regular parametric estimation. This is a sharp contrast with the general semiparametric estimation theory, where it is known that the second-order asymptotic properties are strongly influenced by the bandwidth (e.g., Taniguchi et al. (2003)). In what follows, we develop our discussion based on the results by Tamaki (2007).

Now we consider the following linear regression model

$$\mathbf{y}(t) = \mathbf{B}'\mathbf{x}(t) + \mathbf{u}(t), \quad t = 1, \ldots, N, \tag{6.408}$$

where $\mathbf{x}(t) = (x_1(t), \ldots, x_q(t))'$ is a known vector and nonrandom design sequence, $\mathbf{B} = [\beta_{jk}]$ is a $(q \times p)$-matrix of unknown regression parameters, and $\mathbf{u}(t) = (u_1(t), \ldots, u_p(t))'$ is an unobserved stationary residual.

The vector process $\{\mathbf{u}(t)\}$ is supposed to satisfy the following assumption.

Assumption 6.25 *(1)* $\{\mathbf{u}(t)\}$ *is a linear process generated by*

$$\mathbf{u}(t) = \sum_{s=-\infty}^{\infty} \mathbf{A}(s)\boldsymbol{\varepsilon}(t-s),$$

where $\boldsymbol{\varepsilon}(t) = (\varepsilon_1(t), \ldots, \varepsilon_r(t))'$ *are independent identically distributed random vectors with* $E[\boldsymbol{\varepsilon}(t)] = \mathbf{0}$, $E[\boldsymbol{\varepsilon}(t)\boldsymbol{\varepsilon}(t)'] = \mathbf{G}$ *and all finite absolute moments.*

(2) The $(p \times r)$*-matrices* $\mathbf{A}(s), s = 0, \pm 1, \ldots,$ *satisfy*

$$\sum_{s=-\infty}^{\infty} (1 + |s|^2)\|\mathbf{A}(s)\| < \infty,$$

where $\|\mathbf{A}\|$ *is the square root of the greatest eigenvalue of* $\mathbf{A}^*\mathbf{A}$*, and* $\mathbf{A}^*$ *is the conjugate transpose of a matrix* $\mathbf{A}$*.*

Then $\{\mathbf{u}(t)\}$ *has the spectral density matrix*

$$\mathbf{F}(\lambda) = \frac{1}{2\pi} \sum_{s=-\infty}^{\infty} \boldsymbol{\Gamma}(s)e^{-is\lambda},$$

where $\boldsymbol{\Gamma}(s) = E[\mathbf{u}(t)\mathbf{u}(t+s)']$.

(3) There exists a positive constant γ_1 *such that*

$$\det\{\mathbf{F}(\lambda)\} \geq \gamma_1 > 0$$

for all $\lambda \in (-\pi, \pi]$.

Remark 6.4 *Assumption 6.25* (1) *and* (2) *are satisfied by a wide class of*

time series models which contains the usual VARMA processes. Under Assumption 6.25 (1) *and* (2)*, the joint k-th order cumulants of* $u_{j_1}(s)$*,* $u_{j_2}(s+s_1),\ldots,u_{j_k}(s+s_{k-1})$

$$\Gamma_{j_1\ldots j_k}(s_1,\ldots,s_{k-1}) = \mathrm{cum}^{(k)}[u_{j_1}(s), u_{j_2}(s+s_1),\ldots,u_{j_k}(s+s_{k-1})]$$

exist and satisfy

$$\sum_{s_1,\ldots,s_{k-1}=-\infty}^{\infty}(1+|s_l|^2)|\Gamma_{j_1\ldots j_k}(s_1,\ldots,s_{k-1})|<\infty,\quad j_1,\ldots,j_k=1,\ldots,p$$

for $l=1,\ldots,k-1$*. Then* $\{\mathbf{u}(t)\}$ *has the k-th order cumulant spectral density*

$$F_{j_1\ldots j_k}(\lambda_1,\ldots,\lambda_{k-1}) = \left(\frac{1}{2\pi}\right)^{k-1}\sum_{s_1,\ldots,s_{k-1}=-\infty}^{\infty}\Gamma_{j_1\ldots j_k}(s_1,\ldots,s_{k-1})e^{-i(s_1\lambda_1+\cdots+s_{k-1}\lambda_{k-1})}.$$

Assumption 6.25 (1)-(3) *imply that* $\mathbf{F}(\lambda)^{-1}$ *exists and has the Fourier series representation*

$$\mathbf{F}(\lambda)^{-1}=\frac{1}{2\pi}\sum_{s=-\infty}^{\infty}\mathbf{\Delta}(s)e^{is\lambda},\qquad \sum_{s=-\infty}^{\infty}(1+|s|^2)\|\mathbf{\Delta}(s)\|<\infty.$$

This follows from an application of a famous theorem due to Wiener (see, for example, Wiener (1933, Section 12)).

Let $d_j(N)$ be the positive square root of $\sum_{t=1}^{N}\{x_j(t)\}^2$ for $j=1,\ldots,q$ and

$$\mathbf{D}_N=\mathrm{diag}\{d_1(N),\ldots,d_q(N)\}.$$

We impose some assumptions on $\{\mathbf{x}(t)\}$.

Assumption 6.26 *(1)* $\{\mathbf{x}(t)\}$ *is uniformly bounded; that is, there exists a positive constant* γ_2 *such that*

$$\sup_{t\in\mathbf{Z}}|x_j(t)|<\gamma_2,\quad j=1,\ldots,q.$$

(2) There exists $\gamma_3>0$ *such that* $\{d_j(N)\}^2\geq\gamma_3 N$ *for* $j=1,\ldots,q$*.*

(3) There exist η_j *such that*

$$\sum_{t=1}^{N}\frac{x_j(t)}{d_j(N)}=N^{1/2}\eta_j+O(N^{-1/2}),\quad j=1,\ldots,q.$$

(4) There exist regression spectral measures $M_{j_1\dots j_k}(\lambda_1,\dots,\lambda_{k-1})$ *such that*

$$\sum_{t=1}^{N}\frac{x_{j_1}(t)x_{j_2}(t+l_1)\cdots x_{j_k}(t+l_{k-1})}{d_{j_1}(N)\cdots d_{j_k}(N)} = N^{-k/2+1}\int_{-\pi}^{\pi}\cdots\int_{-\pi}^{\pi}e^{i(l_1\lambda_1+\cdots+l_{k-1}\lambda_{k-1})}dM_{j_1\dots j_k}(\lambda_1,\dots,\lambda_{k-1}) + O(N^{-k/2})$$

for $k=2,3,\dots$

(5) $\mathbf{R}(0)$ *is nonsingular. Here* $\mathbf{R}(0)$ *is the* $(q\times q)$*-matrix given by*

$$\mathbf{R}(l)=\int_{-\pi}^{\pi}e^{il\lambda}d\mathbf{M}(\lambda),\quad l=0,\pm1,\dots,$$

where $\mathbf{M}(\lambda)=[M_{jk}(\lambda)]$.

Remark 6.5 *Assumption 6.26 is a higher order version of Grenander's conditions. For example, linear combinations of harmonic functions satisfy Assumption 6.26. Let us consider an example of* η_j *and* $M_{j_1\dots j_k}(\lambda_1,\dots,\lambda_{k-1})$.

Example 6.21 (Harmonic trend) *Suppose* $x_j(t)=\cos\nu_j t$, $j=1,\dots,q$, *where* $0<\nu_1<\cdots<\nu_q<\pi$. *From the relation*

$$\sum_{t=1}^{N}\cos\nu t=\frac{1}{2}\left\{\frac{\sin(N+1/2)\nu}{\sin\nu/2}-1\right\},\quad \nu\neq0,\pm2\pi,\dots,$$

it is seen that

$$\sum_{t=1}^{N}\frac{x_j(t)}{d_j(N)}=\frac{1}{\sqrt{2}}N^{-1/2}\left\{\frac{\sin(N+1/2)\nu_j}{\sin\nu_j/2}-1\right\}+O(N^{-3/2}),$$

which means $\eta_j=0$.

It is well known that $\mathbf{M}(\lambda)$ *has a jump* $\mathrm{diag}(0,\dots,0,1/2,0,\dots,0)$ *(1/2 is in the* j*-th diagonal) at* $\lambda=\pm\nu_j$.

To construct the Hannan estimator, we use the spectral window $W_N(\cdot)$ and the lag window $w(\cdot)$ which satisfy the following assumption.

Assumption 6.27 *(1) The function* $W_N(\lambda)$ *can be expanded as*

$$W_N(\lambda)=\frac{1}{2\pi}\sum_{l=-M}^{M}w\left(\frac{l}{M}\right)e^{-il\lambda}.$$

(2) $w(x)$ *is a continuous, even function with* $w(0)=1$ *and* $w(x)=0$ *for* $|x|\geq1$, *and satisfies*

$$|w(x)|\leq1,$$
$$\lim_{x\to0}\frac{1-w(x)}{|x|^2}<\infty.$$

(3) $M = M(N)$ satisfies

$$M/N^{1/3} + N^{1/4}/M \to 0 \quad \text{as } N \to \infty.$$

Remark 6.6 *It is easy to see that the Hanning window and Parzen window satisfy Assumption 6.27* (1) *and* (2) *(see Examples 6.15 and 6.16).*

As in Hannan (1963), we define for two sequences $y(t)$ and $x(t)$ of N scalars

$$\hat{F}_{yx}(\lambda) = \frac{1}{2\pi N} \sum_{l=-M}^{M} w\left(\frac{l}{M}\right) \sum_{m=1+\underline{l}}^{N-\bar{l}} y(m)x(m+l)e^{-il\lambda},$$

where $\underline{l} = \max(0, -l)$ and $\bar{l} = \max(0, l)$ for $l \in \mathbf{Z}$.

This serves to define all such functions as

$$\hat{F}_{y_j y_k}(\lambda), \qquad \hat{F}_{x_j x_k}(\lambda), \qquad \hat{F}_{u_j u_k}(\lambda), \qquad \hat{F}_{y_j x_k}(\lambda), \qquad \hat{F}_{u_j x_k}(\lambda).$$

We also use the matrix notation

$$\hat{\mathbf{F}}_{yy}(\lambda) = \big[\hat{F}_{y_j y_k}(\lambda)\big], \quad \hat{\mathbf{F}}_{xx}(\lambda) = \big[\hat{F}_{x_j x_k}(\lambda)\big], \quad \hat{\mathbf{F}}_{uu}(\lambda) = \big[\hat{F}_{u_j u_k}(\lambda)\big],$$
$$\hat{\mathbf{F}}_{yx}(\lambda) = \big[\hat{F}_{y_j x_k}(\lambda)\big], \quad \hat{\mathbf{F}}_{ux}(\lambda) = \big[\hat{F}_{x_j u_k}(\lambda)\big].$$

It is not assumed that all of them are estimates of well-defined spectral density matrices. Indeed $\hat{\mathbf{F}}_{uu}(\lambda)$ is constructed from the actual $\mathbf{u}(t)$ and not estimates of them.

We consider a frequency domain version of (6.408), viz.

$$\hat{\mathbf{F}}_{yx}(\lambda) = \mathbf{B}'\hat{\mathbf{F}}_{xx}(\lambda) + \hat{\mathbf{F}}_{ux}(\lambda),$$

which we rewrite in the tensor notation

$$\hat{\mathbf{f}}_{yx}(\lambda) = \big\{\mathbf{I}_p \otimes \hat{\mathbf{F}}_{xx}(\lambda)'\big\}\boldsymbol{\beta} + \hat{\mathbf{f}}_{ux}(\lambda),$$

where $\hat{\mathbf{f}}_{yx}(\lambda) = \mathrm{vec}\big[\hat{\mathbf{F}}_{yx}(\lambda)'\big]$, $\hat{\mathbf{f}}_{ux}(\lambda) = \mathrm{vec}\big[\hat{\mathbf{F}}_{ux}(\lambda)'\big]$, $\boldsymbol{\beta} = \mathrm{vec}[\mathbf{B}]$, and $\mathbf{I}_p$ is the $(p \times p)$ identity matrix.

The Hannan estimator of $\boldsymbol{\beta}$ in an integration version is given by

$$\hat{\boldsymbol{\beta}} = \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \tilde{\mathbf{F}}_{uu}(\lambda)^{-1} \otimes \hat{\mathbf{F}}_{xx}(\lambda)' d\lambda\right]^{-1} \times \left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \big\{\tilde{\mathbf{F}}_{uu}(\lambda) \otimes \mathbf{I}_q\big\}^{-1} \hat{\mathbf{f}}_{yx}(\lambda) d\lambda\right]. \tag{6.409}$$

Since the actual $\mathbf{u}(t)$ is unobservable, the quantity $\hat{\mathbf{F}}_{uu}(\lambda)$ is infeasible. Therefore, we use $\tilde{\mathbf{F}}_{uu}(\lambda)$ for the estimate of $\mathbf{F}(\lambda)$ obtained from the residuals, $\tilde{\mathbf{u}}(t) = \mathbf{y}(t) - \hat{\mathbf{B}}_{\mathrm{LS}}'\mathbf{x}(t)$, from the least squares regression. Then $\tilde{\mathbf{F}}_{uu}(\lambda)$ can be calculated directly as

$$\tilde{\mathbf{F}}_{uu}(\lambda) = \hat{\mathbf{F}}_{yy}(\lambda) - \hat{\mathbf{F}}_{yx}(\lambda)\hat{\mathbf{B}}_{\mathrm{LS}} - \hat{\mathbf{B}}_{\mathrm{LS}}'\hat{\mathbf{F}}_{xy}(\lambda) + \hat{\mathbf{B}}_{\mathrm{LS}}'\hat{\mathbf{F}}_{xx}(\lambda)\hat{\mathbf{B}}_{\mathrm{LS}}.$$

Hannan (1963) show that under very general conditions, $\hat{\boldsymbol{\beta}}$ is first-order asymptotically Gaussian efficient; that is, the distribution of $(\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges as $N \to \infty$ to the multivariate normal distribution with zero mean vector and covariance matrix given by

$$\boldsymbol{\mathcal{I}}^{-1} = \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} \mathbf{F}(\lambda)^{-1} \otimes d\mathbf{M}(\lambda)' \right]^{-1}$$

(see also Hannan (1970)).

It is well known that integration of nonparametric estimators recovers root N consistency (cf. Taniguchi et al. (1996)). Since $\hat{\boldsymbol{\beta}}$ in (6.409) is based on integral functionals of nonparametric estimators, it may be expected that $\hat{\boldsymbol{\beta}}$ has attractive properties in higher order asymptotic theory. Thus we consider the second order asymptotic properties of the estimator $\hat{\boldsymbol{\beta}}$. First, we give the following theorem.

Theorem 6.30 *The stochastic expansion for* $(\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ *is given by*

$$\begin{aligned}(\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \boldsymbol{\mathcal{I}}^{-1}\mathbf{Z}_1 - N^{-1/2}\boldsymbol{\mathcal{I}}^{-1}(\mathbf{Z}_2 - E[\mathbf{Z}_2]) - N^{-1/2}\boldsymbol{\mathcal{I}}^{-1}E[\mathbf{Z}_2] \\ + N^{-1/2}\boldsymbol{\mathcal{I}}^{-1}\mathbf{Z}_3\boldsymbol{\mathcal{I}}^{-1}\mathbf{Z}_1 + o_p(N^{-1/2}),\end{aligned}$$

where

$$\begin{aligned}\mathbf{Z}_1 &= \frac{N}{2\pi} \int_{-\pi}^{\pi} \{\mathbf{F}(\lambda)^{-1} \otimes \mathbf{D}_N{}^{-1}\} \hat{\mathbf{f}}_{ux}(\lambda) d\lambda, \\ \mathbf{Z}_2 &= \frac{N^{3/2}}{2\pi} \int_{-\pi}^{\pi} \{\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1} \otimes \mathbf{D}_N{}^{-1}\} \hat{\mathbf{f}}_{ux}(\lambda) d\lambda, \\ \mathbf{Z}_3 &= \frac{N^{3/2}}{2\pi} \int_{-\pi}^{\pi} \{\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\} \otimes \{\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1}\} d\lambda, \\ \mathbf{V}_1(\lambda) &= \hat{\mathbf{F}}_{uu}(\lambda) - E[\hat{\mathbf{F}}_{uu}(\lambda)].\end{aligned}$$

We decompose $\tilde{\mathbf{F}}_{uu}(\lambda)$ as follows:

$$\tilde{\mathbf{F}}_{uu}(\lambda) = \mathbf{F}(\lambda) + \sum_{j=1}^{4} \mathbf{V}_j(\lambda), \tag{6.410}$$

where

$$\begin{aligned}\mathbf{V}_2(\lambda) &= \int_{-\pi}^{\pi} W_N(\lambda - \mu)\mathbf{F}(\mu) d\mu - \mathbf{F}(\lambda), \\ \mathbf{V}_3(\lambda) &= \tilde{\mathbf{F}}_{uu}(\lambda) - \hat{\mathbf{F}}_{uu}(\lambda), \\ \mathbf{V}_4(\lambda) &= E\big[\hat{\mathbf{F}}_{uu}(\lambda)\big] - \int_{-\pi}^{\pi} W_N(\lambda - \mu)\mathbf{F}(\mu) d\mu.\end{aligned}$$

To prove Theorem 6.30, we first state three lemmas.

Lemma 6.9 $\mathbf{V}_1(\lambda) = O_p((M/N)^{1/2})$, $\mathbf{V}_2(\lambda) = O(M^{-2})$, $\mathbf{V}_3(\lambda) = O_p(M/N)$, *and* $\mathbf{V}_4(\lambda) = O(N^{-1})$.

PROOF OF LEMMA 6.9 The proof is given by the standard texts (e.g., Anderson (1971), Brillinger (1981), and Hannan (1970)). □

The order of magnitude for each of these terms in our decomposition (6.410) is given in Lemma 6.9.

Lemma 6.10

$$\mathbf{Z}_1 = O_p(1), \qquad \mathbf{Z}_2 = O_p(M^{1/2}),$$
$$\frac{N}{2\pi}\int_{-\pi}^{\pi}\left[\left\{\mathbf{F}(\lambda)^{-1}\mathbf{V}_2(\lambda)\mathbf{F}(\lambda)^{-1}\right\}\otimes \mathbf{D}_N{}^{-1}\right]\hat{\mathbf{f}}_{ux}(\lambda)d\lambda = O_p(M^{-2}),$$
$$\frac{N}{2\pi}\int_{-\pi}^{\pi}\left[\left\{\mathbf{F}(\lambda)^{-1}\mathbf{V}_3(\lambda)\mathbf{F}(\lambda)^{-1}\right\}\otimes \mathbf{D}_N{}^{-1}\right]\hat{\mathbf{f}}_{ux}(\lambda)d\lambda = O_p(M/N),$$
$$\frac{N}{2\pi}\int_{-\pi}^{\pi}\left[\left\{\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\right\}\otimes \mathbf{D}_N{}^{-1}\right]\hat{\mathbf{f}}_{ux}(\lambda)d\lambda = o_p(N^{-1/2}).$$

PROOF OF LEMMA 6.10 The proofs of the first four equalities follow directly by evaluating the first- and second-order moments. Hence, we only give the proof of the last equality.

Note that

$$E\left[\int_{-\pi}^{\pi}\|(\mathbf{I}_p\otimes \mathbf{D}_N{}^{-1})\hat{\mathbf{f}}_{ux}(\lambda)\|^2 d\lambda\right] = O(M/N^2)$$

and

$$E\left[\|\mathbf{V}_1(\lambda)\|^4\right] = E\left[\left\|\hat{\mathbf{F}}_{uu}(\lambda) - E\left[\hat{\mathbf{F}}_{uu}(\lambda)\right]\right\|^4\right] = O((M/N)^2), \tag{6.411}$$

(see the proof of Theorem 7.4.4 in Brillinger (1981)). We have

$$\begin{aligned}
&\left\|\frac{N}{2\pi}\int_{-\pi}^{\pi}\left[\left\{\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\right\}\otimes \mathbf{D}_N{}^{-1}\right]\hat{\mathbf{f}}_{ux}(\lambda)d\lambda\right\| \\
&\quad\le \frac{N}{2\pi}\int_{-\pi}^{\pi}\|\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\otimes \mathbf{I}_q\| \\
&\qquad\times\|(\mathbf{I}_p\otimes \mathbf{D}_N{}^{-1})\hat{\mathbf{f}}_{ux}(\lambda)\|d\lambda \\
&\quad\le \frac{N}{2\pi}\left\{\int_{-\pi}^{\pi}\|\mathbf{F}(\lambda)^{-1}\|^6\|\mathbf{V}_1(\lambda)\|^4 d\lambda\right\}^{1/2}\left\{\int_{-\pi}^{\pi}\|(\mathbf{I}_p\otimes \mathbf{D}_N{}^{-1})\hat{\mathbf{f}}_{ux}(\lambda)\|^2 d\lambda\right\}^{1/2} \\
&\quad= N\times O_p(M/N)\times O_p(M^{1/2}/N) \\
&\quad= o_p(N^{-1/2}).
\end{aligned}$$

□

Lemma 6.11

$$\frac{N}{2\pi}\int_{-\pi}^{\pi}\mathbf{F}(\lambda)^{-1}\otimes\left\{\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1}\right\}d\lambda=\boldsymbol{\mathcal{I}}+o(N^{-1/2}),$$

$$\mathbf{Z}_3=O_p(M^{1/2}),$$

$$\frac{N}{2\pi}\int_{-\pi}^{\pi}\left\{\mathbf{F}(\lambda)^{-1}\mathbf{V}_2(\lambda)\mathbf{F}(\lambda)^{-1}\right\}\otimes\left\{\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1}\right\}d\lambda=O(M^{-2}),$$

$$\frac{N}{2\pi}\int_{-\pi}^{\pi}\left\{\mathbf{F}(\lambda)^{-1}\mathbf{V}_3(\lambda)\mathbf{F}(\lambda)^{-1}\right\}\otimes\left\{\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1}\right\}d\lambda=O_p(M/N),$$

$$\frac{N}{2\pi}\int_{-\pi}^{\pi}\left\{\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\right\}\otimes\left\{\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1}\right\}d\lambda$$
$$=o_p(N^{-1/2}).$$

PROOF OF LEMMA 6.11 Similarly to Lemma 6.10, we only give the proofs of the first and last equalities. The first one is evaluated as follows:

$$\begin{aligned}
&\frac{N}{2\pi}\int_{-\pi}^{\pi}\mathbf{F}(\lambda)^{-1}\otimes\left\{\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1}\right\}d\lambda\\
&\quad=\left(\frac{1}{2\pi}\right)^2\sum_{l=-M}^{M}w\left(\frac{l}{M}\right)\boldsymbol{\Delta}(l)\otimes\left\{\mathbf{R}(l)'+O\left(\frac{1+|l|}{N}\right)\right\}\\
&\quad=\left(\frac{1}{2\pi}\right)^2\sum_{l=-\infty}^{\infty}\boldsymbol{\Delta}(l)\otimes\mathbf{R}(l)'+O(M^{-2})\\
&\quad=\frac{1}{2\pi}\int_{-\pi}^{\pi}\mathbf{F}(\lambda)^{-1}\otimes d\mathbf{M}(\lambda)'+o(N^{-1/2}).
\end{aligned}$$

The last one is evaluated as follows: From (6.411) and

$$\int_{-\pi}^{\pi}\|\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1}\|^2d\lambda=O(M/N^2),$$

we have

$$\begin{aligned}
&\left\|\frac{N}{2\pi}\int_{-\pi}^{\pi}\left\{\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\right\}\otimes\left\{\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1}\right\}d\lambda\right\|\\
&\quad\le\frac{N}{2\pi}\int_{-\pi}^{\pi}\|\mathbf{F}(\lambda)^{-1}\|^3\|\mathbf{V}_1(\lambda)\|^2\|\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1}\|d\lambda\\
&\quad\le\frac{N}{2\pi}\left\{\int_{-\pi}^{\pi}\|\mathbf{F}(\lambda)^{-1}\|^6\|\mathbf{V}_1(\lambda)\|^4d\lambda\right\}^{1/2}\left\{\int_{-\pi}^{\pi}\|\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1}\|^2d\lambda\right\}^{1/2}\\
&\quad=o_p(N^{-1/2}).
\end{aligned}$$

Thus we complete the proof of Lemma 6.11. □

PROOF OF THEOREM 6.30 Expanding $\tilde{\mathbf{F}}_{uu}(\lambda)^{-1}$ about $\mathbf{F}(\lambda)^{-1}$, we

obtain, after application of Lemma 6.9,

$$\begin{aligned}\tilde{\mathbf{F}}_{uu}(\lambda)^{-1} = \mathbf{F}(\lambda)^{-1} - \mathbf{F}(\lambda)^{-1}\sum_{j=1}^{3}\mathbf{V}_j(\lambda)\mathbf{F}(\lambda)^{-1} \\ + \mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1} \\ + O_p(M^{-3/2}N^{-1/2}).\end{aligned} \tag{6.412}$$

Let

$$\tilde{\mathbf{Z}} = \frac{N}{2\pi}\int_{-\pi}^{\pi}\{\tilde{\mathbf{F}}_{uu}(\lambda)^{-1} \otimes \mathbf{D}_N{}^{-1}\}\hat{\mathbf{f}}_{ux}(\lambda)d\lambda. \tag{6.413}$$

We then have

$$(\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \tilde{\boldsymbol{\mathcal{I}}}^{-1}\tilde{\mathbf{Z}},$$

where

$$\tilde{\boldsymbol{\mathcal{I}}} = \frac{N}{2\pi}\int_{-\pi}^{\pi}\tilde{\mathbf{F}}_{uu}(\lambda)^{-1} \otimes \{\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1}\}d\lambda.$$

Inserting (6.412) into (6.413) we have

$$\begin{aligned}\tilde{\mathbf{Z}} = \mathbf{Z}_1 - N^{-1/2}\mathbf{Z}_2 \\ - \frac{N}{2\pi}\int_{-\pi}^{\pi}\Big[\Big\{\mathbf{F}(\lambda)^{-1}\mathbf{V}_2(\lambda)\mathbf{F}(\lambda)^{-1} + \mathbf{F}(\lambda)^{-1}\mathbf{V}_3(\lambda)\mathbf{F}(\lambda)^{-1} \\ - \mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\Big\} \otimes \mathbf{D}_N{}^{-1}\Big]\hat{\mathbf{f}}_{ux}(\lambda)d\lambda \\ + o_p(N^{-1/2}),\end{aligned} \tag{6.414}$$

where we used the fact that $(\mathbf{I}_p \otimes \mathbf{D}_N{}^{-1})\hat{\mathbf{f}}_{ux}(\lambda) = O_p(M/N)$. The order of magnitude for each of these terms in (6.414) is given in Lemma 6.10.

Inserting (6.412) into $\tilde{\boldsymbol{\mathcal{I}}}$ we have

$$\begin{aligned}\tilde{\boldsymbol{\mathcal{I}}} = \frac{N}{2\pi}\int_{-\pi}^{\pi}\mathbf{F}(\lambda)^{-1} \otimes \{\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1}\}d\lambda - N^{-1/2}\mathbf{Z}_3 \\ - \frac{N}{2\pi}\int_{-\pi}^{\pi}\{\mathbf{F}(\lambda)^{-1}\mathbf{V}_2(\lambda)\mathbf{F}(\lambda)^{-1} + \mathbf{F}(\lambda)^{-1}\mathbf{V}_3(\lambda)\mathbf{F}(\lambda)^{-1} \\ - \mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\mathbf{V}_1(\lambda)\mathbf{F}(\lambda)^{-1}\} \\ \otimes \{\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1}\}d\lambda \\ + o_p(N^{-1/2}),\end{aligned} \tag{6.415}$$

where we used the fact that $\mathbf{D}_N{}^{-1}\hat{\mathbf{F}}_{xx}(\lambda)'\mathbf{D}_N{}^{-1} = O(M/N)$. The order of magnitude for each of these terms in (6.415) is given in Lemma 6.11. Hence, Theorem 6.30 follows from Lemmas 6.10 and 6.11. □

Next, we evaluate the asymptotic cumulants of $\mathbf{Z}_j$, $j = 1, 2, 3$ given in Theorem

6.30. Denote by $Z_1(jk)$ and $Z_2(jk)$ the $(j-1)q+k$-th component of the vectors $\mathbf{Z}_1$ and $\mathbf{Z}_2$, respectively. Similarly, denote by $\mathbf{Z}_3(j_1k_1, j_2k_2)$ the $((j_1-1)q+k_1, (j_2-1)q+k_2)$-th element of the matrix $\mathbf{Z}_3$. Then we have the following lemma.

Lemma 6.12 *(i)* $E[\mathbf{Z}_1] = \mathbf{0}$,
(ii) $E[Z_2(jk)] = F^{jj_1}(0)F^{j_2j_3}(0)F_{j_1j_2j_3}(0,0)\eta_k + o(N^{-1/2})$,
(iii) $E[\mathbf{Z}_3] = \mathbf{0}$,
(iv) $\mathrm{Cov}[\mathbf{Z}_1] = \boldsymbol{\mathcal{I}} + o(N^{-1/2})$,
(v) $\mathrm{Cov}[\mathbf{Z}_1, \mathbf{Z}_2] = O(M/N^{1/2})$,
(vi)

$$\begin{aligned}&\mathrm{Cov}[Z_1(j_1k_1), Z_3(j_2k_2, j_3k_3)]\\ &\quad = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathcal{K}_{j_1j_2j_3}(-\lambda,\lambda)\eta_{k_1} dM_{k_2k_3}(\lambda) + o(N^{-1/2}),\end{aligned}$$

(vii)

$$\begin{aligned}&\mathrm{cum}[Z_1(j_1k_1), Z_1(j_2k_3), Z_1(j_3k_3)]\\ &\quad = N^{-1/2}\frac{1}{2\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \mathcal{K}_{j_1j_2j_3}(\lambda_1,\lambda_2) dM_{k_1k_2k_3}(\lambda_1,\lambda_2) + o(N^{-1}),\end{aligned}$$

where

$$\mathcal{K}_{jkl}(\lambda_1,\lambda_2) = F^{jj'}(-\lambda_1-\lambda_2)F^{kk'}(\lambda_1)F^{ll'}(\lambda_2)F_{j'k'l'}(-\lambda_1,-\lambda_2),$$

and $F^{jk}(\lambda)$ *is the* (j,k)*-th element of the matrix* $\mathbf{F}(\lambda)^{-1}$. *Here we use the Einstein summation convention.*

PROOF OF LEMMA 6.12 The proof of Lemma 6.12 is quite technical and lengthy. Hence we only give the proofs of Lemma 6.12 (i)-(iii). The proofs of Lemma 6.12 (iv)-(vii) are similar to that of Lemma 6.12 (ii). The detailed proofs are given in Tamaki (2007).

(i) Because $E[\hat{\mathbf{f}}_{ux}(\lambda)] = 0$, we obtain (i).

(ii) Note that $Z_2(jk)$ is written as

$$\begin{aligned}&Z_2(jk)\\ &= \left(\frac{1}{2\pi}\right)^4 N^{-1/2}\sum_{s_1,s_2=-\infty}^{\infty}\Delta^{jj_1}(s_1)\Delta^{j_2j_3}(s_2)\sum_{l_1=-M}^{M} w\left(\frac{l_1}{M}\right)w\left(\frac{l_2}{M}\right)\\ &\quad\times\sum_{m_1=1+\underline{l_1}}^{N-\overline{l_1}}\sum_{m_2=1+\underline{l_2}}^{N-\overline{l_2}}\left[u_{j_1}(m_1)u_{j_2}(m_1+l_1) - \Gamma_{j_1j_2}(l_1)\right]u_{j_3}(m_2)\\ &\quad\times\frac{x_k(m_2+l_2)}{d_k(N)},\end{aligned} \tag{6.416}$$

where $l_2 = s_1 + s_2 - l_1$. Then we have

$$
\begin{aligned}
&E[Z_2(jk)] \\
&\quad = \left(\frac{1}{2\pi}\right)^4 N^{-1/2} \sum_{s_1,s_2=-\infty}^{\infty} \Delta^{jj_1}(s_1)\Delta^{j_2j_3}(s_2) \sum_{l_1=-M}^{M} w\left(\frac{l_1}{M}\right) w\left(\frac{l_2}{M}\right) \\
&\qquad \times \sum_{m_1=1+\underline{l_1}}^{N-\overline{l_1}} \sum_{m_2=1+\underline{l_2}}^{N-\overline{l_2}} \Gamma_{j_1j_2j_3}(l_1, m_2 - m_1) \frac{x_k(m_2 + l_2)}{d_k(N)}.
\end{aligned} \tag{6.417}
$$

From Remark 6.4, it is easily seen that the summation $\sum_{m_i=1+\underline{l_i}}^{N-\overline{l_i}}$, $i = 1, 2$ in (6.417) can be replaced by $\sum_{m_i=1}^{N}$, $i = 1, 2$, respectively. Hence,

$$
\begin{aligned}
&E[Z_2(jk)] \\
&\quad = \left(\frac{1}{2\pi}\right)^4 N^{-1/2} \sum_{s_1,s_2=-\infty}^{\infty} \Delta^{jj_1}(s_1)\Delta^{j_2j_3}(s_2) \sum_{l_1=-M}^{M} w\left(\frac{l_1}{M}\right) w\left(\frac{l_2}{M}\right) \\
&\qquad \times \sum_{m_1,m_2=1}^{N} \Gamma_{j_1j_2j_3}(l_1, m_2 - m_1) \frac{x_k(m_2 + l_2)}{d_k(N)} + O(N^{-1}) \\
&\quad = \left(\frac{1}{2\pi}\right)^4 N^{-1/2} \sum_{s_1,s_2=-\infty}^{\infty} \Delta^{jj_1}(s_1)\Delta^{j_2j_3}(s_2) \sum_{l_1=-M}^{M} w\left(\frac{l_1}{M}\right) w\left(\frac{l_2}{M}\right) \\
&\qquad \times \sum_{m_1=-(N-1)}^{N-1} \Gamma_{j_1j_2j_3}(l_1, m_1) \sum_{m_2=1+\underline{m_1}}^{N-\overline{m_1}} \frac{x_k(m_1 + m_2 + l_2)}{d_k(N)} + O(N^{-1}).
\end{aligned} \tag{6.418}
$$

Recalling Assumption 6.26 (1)-(3), we get

$$
\begin{aligned}
&N^{-1/2} \sum_{m_2=1+\underline{m_1}}^{N-\overline{m_1}} \frac{x_k(m_1 + m_2 + l_2)}{d_k(N)} \\
&\qquad = N^{-1/2} \sum_{t=1}^{N} \frac{x_k(t)}{d_k(N)} + O\left(\frac{|m_1| + |l_2| + 1}{N}\right) \\
&\qquad = \eta_k + O\left(\frac{|m_1| + |l_2| + 1}{N}\right).
\end{aligned} \tag{6.419}
$$

Inserting (6.419) into (6.418), we obtain

$$\begin{aligned}
&E[Z_2(jk)] \\
&\quad = \left(\frac{1}{2\pi}\right)^4 \eta_k \sum_{s_1,s_2=-\infty}^{\infty} \Delta^{jj_1}(s_1)\Delta^{j_2j_3}(s_2) \sum_{l_1=-M}^{M} w\left(\frac{l_1}{M}\right) w\left(\frac{l_2}{M}\right) \\
&\qquad \times \sum_{m_1=-(N-1)}^{N-1} \Gamma_{j_1j_2j_3}(l_1, m_1) + O(N^{-1}),
\end{aligned}$$

which, by Assumption 6.27 (2), leads to

$$\begin{aligned}
&E[Z_2(jk)] \\
&\quad = \left(\frac{1}{2\pi}\right)^4 \eta_k \sum_{s_1,s_2=-\infty}^{\infty} \Delta^{jj_1}(s_1)\Delta^{j_2j_3}(s_2) \sum_{l_1,m_1=-\infty}^{\infty} \Gamma_{j_1j_2j_3}(l_1, m_1) \\
&\qquad + O(M^{-2}) \\
&\quad = F^{jj_1}(0)F^{j_2j_3}(0)F_{j_1j_2j_3}(0,0)\eta_k + o(N^{-1/2}).
\end{aligned}$$

(iii) The proof follows from $E[\mathbf{V}_1(\lambda)] = 0$. □

Denote by $\mathcal{I}^{j_1k_1,j_2k_2}$ the $((j_1-1)q+k_1, (j_2-1)q+k_2)$-th element of the matrix $\boldsymbol{\mathcal{I}}^{-1}$. From Theorem 6.30 and Lemma 6.12 the asymptotic cumulants of $(\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{jk} = d_k(N)(\hat{\beta}_{kj} - \beta_{kj})$ are evaluated as follows:

$$\begin{aligned}
&E[(\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{jk}] \\
&\quad = -N^{-1/2}\mathcal{I}^{jk,j_1k_1}F^{jj_1}(0)F^{j_2j_3}(0)F_{j_1j_2j_3}(0,0)\eta_k \\
&\qquad + N^{-1/2}\frac{1}{2\pi}\mathcal{I}^{jk,j_1k_1}\mathcal{I}^{j_2k_2,j_3k_3}\int_{-\pi}^{\pi} \mathcal{K}_{j_3j_1j_2}(-\lambda,\lambda)\eta_{k_3}dM_{k_1k_2}(\lambda) \\
&\qquad + o(N^{-1}) \\
&\quad = N^{-1/2}C^{jk} + o(N^{-1}), \quad \text{(say)},
\end{aligned}$$

$$\mathrm{Cov}[(\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{j_1k_1}, (\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{j_2k_2}] = \mathcal{I}^{j_1k_1,j_2k_2} + o(N^{-1/2}),$$

$$\begin{aligned}
&\mathrm{cum}[(\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{j_1k_1}, (\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{j_2k_2}, (\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{j_3k_3}] \\
&\quad = N^{-1/2}\frac{1}{2\pi}\mathcal{I}^{j_1k_1,j_1'k_1'}\mathcal{I}^{j_2k_2,j_2'k_2'}\mathcal{I}^{j_3k_3,j_3'k_3'} \\
&\qquad \times \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \mathcal{K}_{j_1'j_2'j_3'}(\lambda_1,\lambda_2)dM_{k_1'k_2'k_3'}(\lambda_1,\lambda_2) + o(N^{-1/2}), \\
&\quad = N^{-1/2}C^{j_1k_1,j_2k_2,j_3k_3} + o(N^{-1/2}), \quad \text{(say)}.
\end{aligned}$$

The L-th order cumulants of $(\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{jk}$ satisfy

$$\mathrm{cum}^{(L)}[(\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{j_1k_1}, \ldots, (\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})_{j_Lk_L}] = O(N^{-L/2+1})$$

for each $L \geq 3$.

From the general Edgeworth expansion formula (e.g., Taniguchi and Kakizawa (2000, p.169)) we get the following theorem.

Theorem 6.31

$$P[(\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \le \mathbf{z}]$$
$$= \int_{-\infty}^{\mathbf{z}} N(\mathbf{w} : \boldsymbol{\mathcal{I}}^{-1}) \Big[1 + N^{-1/2} C^{jk} H_{jk}(\mathbf{w}) + \frac{1}{6} N^{-1/2} C^{j_1k_1, j_2k_2, j_3k_3} H_{j_1k_1, j_2k_2, j_3k_3}(\mathbf{w}) \Big] d\mathbf{w} + o(N^{-1/2}),$$

where $\mathbf{z}$ *and* $\mathbf{w}$ *are the* pq*-vectors with* z_{jk} *and* w_{jk} *in* $(j-1)q + k$*-th place, respectively,*

$$N(\mathbf{w} : \boldsymbol{\mathcal{I}}^{-1}) = (2\pi)^{-pq/2} |\boldsymbol{\mathcal{I}}|^{1/2} \exp\left(-\frac{1}{2} \mathbf{w}' \boldsymbol{\mathcal{I}} \mathbf{w} \right),$$

the multivariate normal distribution, and multivariate Hermite polynomials:

$$H_{j_1k_1, \ldots, j_sk_s}(\mathbf{w}) = \frac{(-1)^s}{N(\mathbf{w} : \boldsymbol{\mathcal{I}}^{-1})} \frac{\partial^s}{\partial w_{j_1k_1} \cdots \partial w_{j_sk_s}} N(\mathbf{w} : \boldsymbol{\mathcal{I}}^{-1}).$$

The preceding results are unexpected.

Remark 6.7 *In the context of semiparametric estimation, it is known that root-N asymptotics in general do not hold (e.g., Taniguchi et al. (2003)). However, our results claim that, in a linear regression model, the standard root-N asymptotics hold up to second order. This means that the Hannan estimator has the same rate of convergence as regular parametric estimation. Moreover, it is seen that our Edgeworth expansion is independent of the bandwidth and the window type function for the residual spectra. This is in sharp contrast to the general semiparametric estimation theory.*

We examine the performance of the second-order Edgeworth expansion given in Theorem 6.31. The model used for data generation is the following:

$$y(t) = \beta x(t) + u(t), \quad (p = q = 1)$$
$$u(t) = a u(t-1) + \varepsilon(t),$$

where $|a| < 1$, $\varepsilon(t)$'s are i.i.d. $Exp(0, 1)$ random variables with probability density

$$p(z) = \exp\{-(z+1)\}, \quad z > -1.$$

In the following Figures 6.20-6.23, we plotted the first (solid) and the second (dotted) order approximations for the distribution function of the normalized $\hat{\beta}$ given in Theorem 6.31, and the empirical distribution (dashed) which is obtained by 10000 times replications. From Figures 6.20-6.23, we observe that the second-order Edgeworth expansions are quite accurate in the neighborhood of $z = 0$.

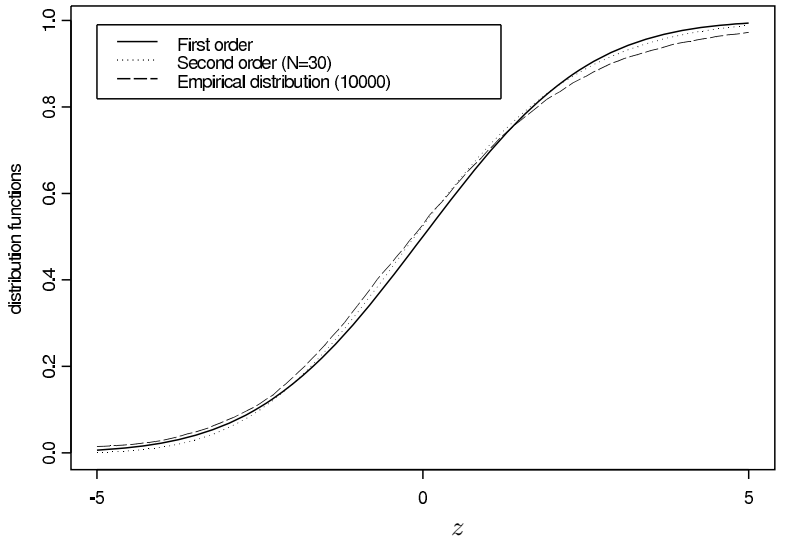


Figure 6.20 $a = 0.5$ *and* $x(t) = 1$.

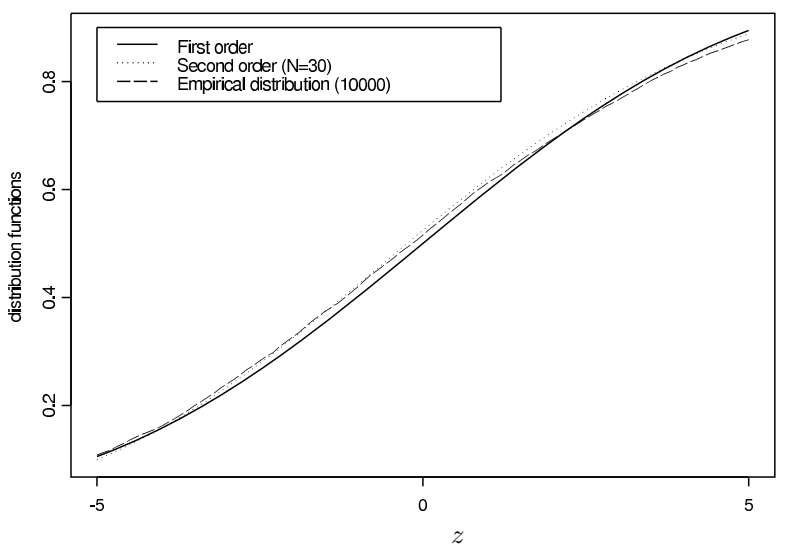


Figure 6.21 $a = 0.75$ *and* $x(t) = 1$.

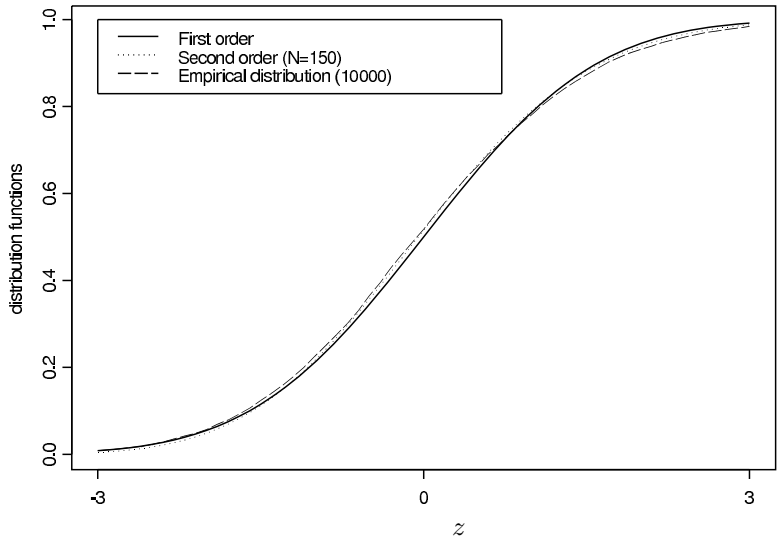


Figure 6.22 $a = 0.25$ *and* $x(t) = 1 + \cos t$.

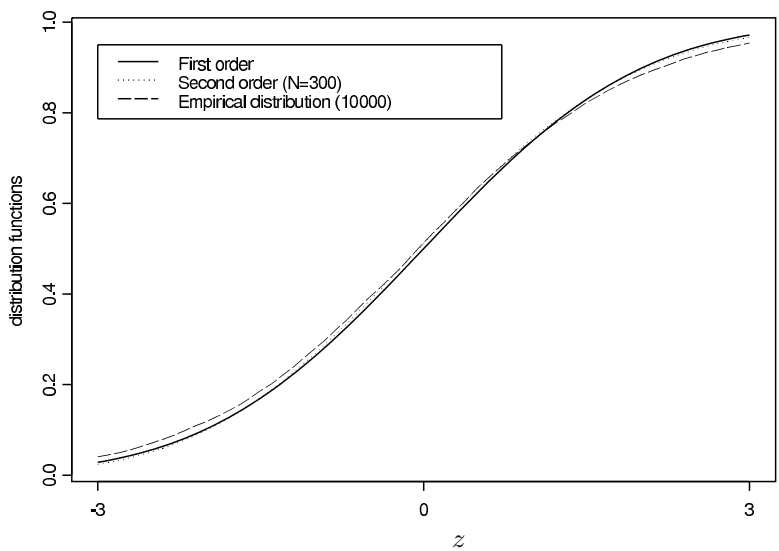


Figure 6.23 $a = 0.5$ *and* $x(t) = 1 + \cos t$.

Next, we discuss higher order asymptotic efficiency of the Hannan estimator $\hat{\boldsymbol{\beta}}$ defined by (6.409). To discuss higher order efficiency and establish unified higher order results we need to restrict the class of estimators to second-order asymptotically median unbiased (AMU).

From Theorem 6.31, it can be seen that $\hat{\boldsymbol{\beta}}$ is not second-order AMU. Thus we modify $\hat{\boldsymbol{\beta}}$ as follows:

$$\hat{\beta}^*_{kj} = \hat{\beta}_{kj} - N^{-1/2}\frac{\tilde{C}^{jk}}{d_k(N)} + \frac{1}{6}N^{-1/2}\frac{(\tilde{\mathcal{I}}^{jk,jk})^{-1}\tilde{C}^{jk,jk,jk}}{d_k(N)},$$

where $\tilde{\mathcal{I}}^{j_1k_1,j_2k_2}$ is the $((j_1-1)q+k_1,(j_2-1)q+k_2)$-th element of the matrix $\tilde{\boldsymbol{\mathcal{I}}}^{-1}$ and, $\tilde{C}^{jk}$ and $\tilde{C}^{jk,jk,jk}$ are the quantities replacing the cumulant spectrum by the nonparametric spectral estimator in C^{jk} and $C^{jk,jk,jk}$, respectively.

Then we have the following theorem.

Theorem 6.32 *(i) The estimator $\hat{\beta}^*_{kj}$ is second-order AMU.*

(ii) The second-order asymptotic distribution of $\hat{\boldsymbol{\beta}}^* = [\hat{\beta}^*_{kj}]$ is

$$\begin{aligned} &P[(\mathbf{I}_p \otimes \mathbf{D}_N)(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}) \le \mathbf{z}] \\ &\quad = \int_{-\infty}^{\mathbf{z}} N(\mathbf{w} : \boldsymbol{\mathcal{I}}^{-1})\Big[1 + \frac{1}{6}N^{-1/2}C^{jk,jk,jk}H_{jk}(\mathbf{w}) \\ &\qquad + \frac{1}{6}N^{-1/2}C^{j_1k_1,j_2k_2,j_3k_3}H_{j_1k_1,j_2k_2,j_3k_3}(\mathbf{w})\Big]d\mathbf{w} + o(N^{-1/2}). \end{aligned}$$

PROOF It is sufficient to show that $\boldsymbol{\mathcal{I}}_{\mathrm{B}} = \boldsymbol{\mathcal{I}} + o(N^{-1/2})$. The proof is substantially a modification of that of Theorem 5 in Hannan (1970, p.427), see also Theorem 10.2.7 in Anderson (1971, p.575).

From Assumption 6.25, we can find spectral matrices $\mathbf{F}_1(\lambda)^{-1}$ and $\mathbf{F}_2(\lambda)^{-1}$ of moving average processes of order M such that

$$\begin{aligned} &0 < \mathbf{F}_2(\lambda)^{-1} \le \mathbf{F}(\lambda)^{-1} \le \mathbf{F}_1(\lambda)^{-1}, \\ &\mathbf{F}_1(\lambda)^{-1} - \mathbf{F}_2(\lambda)^{-1} < \delta\mathbf{I}_p, \end{aligned} \tag{6.420}$$

where $\delta = O(M^{-2})$. Here these inequalities are to be interpreted in the usual way as between Hermitian matrices. In fact, let

$$\mathbf{F}_1(\lambda)^{-1} = \frac{1}{2\pi}\sum_{s=-M}^{M}\boldsymbol{\Delta}(s)e^{is\lambda} + \frac{K_1}{M^2}\mathbf{I}_p,$$

$$\mathbf{F}_2(\lambda)^{-1} = \frac{1}{2\pi}\sum_{s=-M}^{M}\boldsymbol{\Delta}(s)e^{is\lambda} - \frac{K_1}{M^2}\mathbf{I}_p,$$

then we can choose a constant $K_1 > 0$ such that (6.420) holds. Thus we have approximated $\mathbf{F}(\lambda)$ by autoregressive processes of order M. Let $\{\mathbf{w}(t)\}$ satisfy

the equation

$$\sum_{s=0}^{M} \mathbf{C}_1(s)\mathbf{w}(t-s) = \mathbf{v}(t),$$

where $\mathbf{C}_1(s)$ are the autoregressive matrices corresponding to $\mathbf{F}_1(\lambda)$ and the $\mathbf{v}(t)$ are independent and identically distributed random vectors with mean zero and covariance matrix unity. Let $\tilde{\mathbf{w}}$ have $w_k(t)$ in the $(t-1)p+k$-th place and $\mathbf{\Gamma}^{(1)} = \mathrm{Cov}[\tilde{\mathbf{w}}\tilde{\mathbf{w}}']$. Then, we obtain

$$\begin{aligned}
&(\mathbf{D}_N{}^{-1} \otimes \mathbf{I}_p)(\mathbf{X}' \otimes \mathbf{I}_p)\mathbf{\Gamma}^{(1)-1}(\mathbf{X} \otimes \mathbf{I}_p)(\mathbf{D}_N{}^{-1} \otimes \mathbf{I}_p) \\
&\quad = \sum_{t=M+1}^{N} \sum_{j_1,j_2=1}^{M} \{\mathbf{D}_N{}^{-1}\mathbf{x}(t-j_1)\mathbf{x}(t-j_2)'\mathbf{D}_N{}^{-1}\} \otimes \mathbf{C}_1(j_1)'\mathbf{C}_1(j_2) \\
&\qquad + O(M/N) \\
&\quad = \sum_{j_1,j_2=1}^{M} \{\mathbf{R}(j_1-j_2) + O(M/N)\} \otimes \mathbf{C}_1(j_1)'\mathbf{C}_1(j_2) + O(M/N) \\
&\quad = \frac{1}{2\pi}\int_{-\pi}^{\pi} d\mathbf{M}(\lambda) \otimes \mathbf{F}_1(-\lambda)^{-1} + O(M/N).
\end{aligned}$$

Reversing the order of the indices of the tensors we obtain

$$\boldsymbol{\mathcal{I}}_{\mathrm{B}} \leq \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathbf{F}_1(\lambda)^{-1} \otimes d\mathbf{M}(\lambda)' + o(N^{-1/2}).$$

Similarly,

$$\boldsymbol{\mathcal{I}}_{\mathrm{B}} \geq \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathbf{F}_2(\lambda)^{-1} \otimes d\mathbf{M}(\lambda)' + o(N^{-1/2}).$$

Thus we have $\boldsymbol{\mathcal{I}}_{\mathrm{B}} = \boldsymbol{\mathcal{I}} + o(N^{-1/2})$. □

Since $\hat{\boldsymbol{\beta}}$ is first-order asymptotically efficient under Gaussian errors, we concentrate our attention only on the Gaussian efficiency. From Akahira and Takeuchi (1981), the second-order Gaussian efficient bound distribution of jk-component is given by

$$\mathrm{P}[d_k(N)(\tilde{\beta}_{kj} - \beta_{kj}) \leq z] = \Phi((\mathcal{I}_{\mathrm{B}}^{jk,jk})^{-1/2}z) + o(N^{-1/2}),$$

where $\mathcal{I}_{\mathrm{B}}^{j_1k_1,j_2k_2}$ is (j_1k_1, j_2k_2)-component of the covariance matrix $\boldsymbol{\mathcal{I}}_{\mathrm{B}}^{-1}$ of the best linear unbiased estimator. Hence, we have the following result.

Theorem 6.33 *The bias-corrected estimator $\hat{\beta}_{kj}^{*}$ is second-order asymptotically Gaussian efficient, if and only if*

$$C^{jk,jk,jk} = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \mathcal{K}_{jjj}(\lambda_1,\lambda_2)dM_{kkk}(\lambda_1,\lambda_2) = 0. \tag{6.421}$$

Remark 6.8 *If the residual $\{\mathbf{u}(t)\}$ is a Gaussian process, then* (6.421) *holds.*

However, in general, the bias-corrected estimator $\hat{\boldsymbol{\beta}}^*$ is not second-order asymptotically Gaussian efficient.

Remark 6.9 *Theorem 6.33 can be employed to check whether the Hannan estimator leads to a second-order Gaussian efficient estimator. Since we do not assume the normality of the error process, in general we have $\mathcal{K}_{jjj}(\lambda_1, \lambda_2) \neq 0$. Here, we give four examples of the regressor $\{\mathbf{x}(t)\}$ in the case of $p = q = 1$.*

(i) $x_1(t) = 1$ for $t = 1, 2, \ldots$. Then $\eta_1 = 1$, $M_{11}(\lambda)$ has the jump 1 at $\lambda = 0$ and $M_{111}(\lambda_1, \lambda_2)$ has the jump 1 at $\lambda_1 = \lambda_2 = 0$. Hence, the Hannan estimator is second-order Gaussian efficient if and only if $F_{111}(0) = 0$.

(ii) $x_1(t) = \cos \nu t$, $\nu \in (0, 2\pi/3)$ for $t = 1, 2, \ldots$. Then $M_{111}(\lambda_1, \lambda_2)$ has the jump $O_p(N^{-3/2})$. Hence, the Hannan estimator is always second-order Gaussian efficient.

(iii) $x_1(t) = 1 + \cos \nu t$ for $t = 1, 2, \ldots$. Then $\eta_1 = (2/3)^{1/2}$, $M_{11}(\lambda)$ has the jump $2/3$ and $1/6$ at $\lambda = 0$ and $\lambda = \pm\nu$, respectively, and $M_{111}(\lambda_1, \lambda_2)$ has the jump $(2/3)^{3/2}$ and $(2/3)^{3/2}/4$ at $\lambda_1 = \lambda_2 = 0$ and $(\lambda_1, \lambda_2) = (0, \pm\nu)$, $(\pm\nu, 0)$, $(\nu, -\nu)$, $(-\nu, \nu)$, respectively. Hence, the Hannan estimator is not second-order Gaussian efficient.

(iv) $x_1(t) = t/N$ for $t = 1, 2, \ldots$. Then $\eta_1 = \sqrt{3}/2$, $M_{11}(\lambda)$ has the jump 1 at $\lambda = 0$ and $M_{111}(\lambda_1, \lambda_2)$ has the jump $3^{3/2}/4$ at $\lambda_1 = \lambda_2 = 0$. Hence, the Hannan estimator is second-order Gaussian efficient if and only if $F_{111}(0) = 0$.

6.11 Discriminant Analysis for Time Series

In Section 4.4 we discussed the problem of discriminant analysis for independent observations. In the field of financial engineering the problem of credit rating is addressed. Usually the credit rating has been done by use of the i.i.d. settings. Recently, the discriminant analysis for dependent observations has been done in many fields and it will be much sought after. Therefore, in this section we investigate the discriminant analysis for time series in view of differences between spectral structures.

First, we assume that $\{X_t\}$ is a Gaussian stationary process with mean 0, and we know it belongs to one of two categories which are described by two hypotheses Π_1 and Π_2. Analysts are interested in classifying $\{X_t\}$ into one of these categories, when they have the sequence of observations $\mathbf{X}_T = (X_1, \ldots, X_T)'$. Let Π_1 and Π_2 be the hypotheses in which $\{X_t\}$ has the spectral density $f(\lambda)$ and $g(\lambda)$, respectively. Henceforth we write this as

$$\Pi_1 : f(\lambda), \quad \Pi_2 : g(\lambda). \tag{6.422}$$

The general discriminant rule is described as follows. We decompose $\mathbf{R}^T$ into exhaustive and exclusive regions A_1 and A_2, that is, $A_1 \cup A_2 = \mathbf{R}^T$ and

$A_1 \cap A_2 = \emptyset$. If $\mathbf{X}_T$ falls in region A_1, then we assign $\{X_t\}$ into Π_1, and otherwise we assign $\{X_t\}$ into Π_2. If $\mathbf{X}_T$ has a probability density of the form $p_i(\mathbf{X}_T)$ under hypothesis Π_i, the probability of misclassifying $\mathbf{X}_T$ into the category Π_j is written as

$$P(j|i) = \int_{A_j} p_i(\mathbf{X}_T) d\mathbf{X}_T, \quad (i, j = 1, 2,\ i \neq j). \tag{6.423}$$

Therefore, the "good" discriminant rule requires the regions A_1 and A_2 to minimize

$$P(2|1) + P(1|2). \tag{6.424}$$

We have already seen in Theorem 4.3 that the discriminant regions based on the likelihood ratio become optimal. Namely, the optimal discriminant regions are given by

$$A_i = \left\{ \mathbf{X}_T : \text{LLR} = \frac{1}{T} \log \frac{p_i(\mathbf{X}_T)}{p_j(\mathbf{X}_T)} > 0, \quad j \neq i \right\}, \quad (i = 1, 2). \tag{6.425}$$

Since the exact likelihood ratio of time series is not handy in general, instead, we use the following approximation of the log-likelihood ratio between Π_1 and Π_2:

$$I(f : g) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left[\log \frac{g(\lambda)}{f(\lambda)} + I_T(\lambda) \left\{ \frac{1}{g(\lambda)} - \frac{1}{f(\lambda)} \right\} \right] d\lambda, \tag{6.426}$$

where $I_T(\lambda) = (2\pi T)^{-1} \left| \sum_{t=1}^{T} X_t e^{it\lambda} \right|^2$ is the periodogram of $\{X_t\}$. Then the proposed discriminant rule is to classify $\{X_t\}$ into Π_1 or Π_2 according to $I(f : g) > 0$ or $I(f : g) \leq 0$. The misclassification probabilities become

$$P(2|1) = P\{I(f : g) \leq 0 | \Pi_1\}, \quad P(1|2) = P\{I(f : g) > 0 | \Pi_2\}. \tag{6.427}$$

Now, we make the following assumption.

Assumption 6.28 **(i)** *$f(\lambda)$ and $g(\lambda)$ are continuous on $[-\pi, \pi]$ and satisfy*

$$M_1 \leq f(\lambda), g(\lambda) \leq M_2, \quad \lambda \in [-\pi, \pi], \tag{6.428}$$

for some $M_1 > 0$ and $M_2 < \infty$.

(ii) *$R_f(s) = \int_{-\pi}^{\pi} e^{is\lambda} f(\lambda) d\lambda$ and $R_g(s) = \int_{-\pi}^{\pi} e^{is\lambda} g(\lambda) d\lambda$ satisfy*

$$\sum_{s=1}^{\infty} s \left| R_f(s) \right|^2 < \infty, \quad \sum_{s=1}^{\infty} s \left| R_g(s) \right|^2 < \infty. \tag{6.429}$$

(iii) *There exists an open interval (a, b), $(a < b)$ on $[-\pi, \pi]$, such that $f(\lambda) \neq g(\lambda)$ for any $\lambda \in (a, b)$.*

Under Assumption 6.28, from Section 6.2 it follows that

(1) Under Π_1,

$$I(f:g) \xrightarrow{p} E\left\{I(f:g)|\Pi_1\right\}, \tag{6.430}$$

(2) Under Π_2,

$$I(f:g) \xrightarrow{p} E\left\{I(f:g)|\Pi_2\right\}, \tag{6.431}$$

(3) Under Π_1,

$$\sqrt{T}\left[I(f:g) - E\left\{I(f:g)|\Pi_1\right\}\right] \xrightarrow{d} N\left(0, \sigma^2(f,g)\right), \tag{6.432}$$

(4) Under Π_2,

$$\sqrt{T}\left[I(f:g) - E\left\{I(f:g)|\Pi_2\right\}\right] \xrightarrow{d} N\left(0, \sigma^2(g,f)\right). \tag{6.433}$$

Then, we have the following result:

Theorem 6.34 *The misclassification probabilities (6.427) satisfy*

$$\lim_{T\to\infty} P(2|1) = 0, \quad \lim_{T\to\infty} P(1|2) = 0. \tag{6.434}$$

That is, the discriminant criterion $I(f:g)$ is consistent.

PROOF

By Theorem 5.2, it is seen that

$$\begin{aligned} E\left\{I(f:g)|\Pi_1\right\} &\to \frac{1}{4\pi}\int_{-\pi}^{\pi}\left[\log\frac{g(\lambda)}{f(\lambda)} + \left\{\frac{f(\lambda)}{g(\lambda)} - 1\right\}\right] d\lambda \\ &= m(f,g), \quad \text{(say)}. \end{aligned} \tag{6.435}$$

Here, note that $\log x + (1/x) - 1 \geq 0$ and the equality holds if and only if $x = 1$. From Assumption 6.28 (iii), we can see that the right-hand side of (6.435) is positive. Therefore, by (6.430), under Π_1, $I(f:g)$ converges in probability to positive value, which implies $\lim_{T\to\infty} P(2|1) = 0$. Similarly, we can show that $\lim_{T\to\infty} P(1|2) = 0$. □

Theorem 6.34 implies the discriminant criterion based on $I(f:g)$ has the property that two misclassification probabilities converge to 0, hence has at least a fundamental " goodness".

Next, we evaluate more delicate "goodness" of $I(f:g)$. For this we assume the spectral density $f(\lambda)$ depends on r-dimensional parameter vector and two hypotheses Π_1 and Π_2 are contiguous, that is,

$$\Pi_1 : f(\lambda) = f_\theta(\lambda), \quad \Pi_2 : f(\lambda) = f_{\theta + \frac{1}{\sqrt{T}}h}(\lambda), \tag{6.436}$$

where $\theta \in \Theta \subset \mathbf{R}^r$ and $h \in \mathbf{R}^r$. Then from (6.432) the discriminant error of

$I(f:g)$ becomes

$$\begin{aligned} P(2|1) &= P\{I(f:g) \le 0|\Pi_1\} \\ &= P\left[\frac{\sqrt{T}\{I(f:g) - m(f,g)\}}{\sigma(f,g)} \le -\frac{\sqrt{T}m(f,g)}{\sigma(f,g)}|\Pi_1\right] \\ &\xrightarrow{d} \Phi\left\{-\sqrt{T}\frac{m(f,g)}{\sigma(f,g)}\right\}, \quad (T\to\infty). \end{aligned} \tag{6.437}$$

We can evaluate $P(1|2)$, similarly. Here $\Phi(\cdot)$ is the probability distribution function of $N(0,1)$. Let f_θ be two times differentiable with respect to θ, then under the contiguous condition (6.436), we obtain

$$m(f,g) = \frac{1}{2T}h'\mathcal{F}(\theta)h + o(T^{-1}), \quad \sigma^2(f,g) = \frac{1}{T}h'\mathcal{F}(\theta)h + o(T^{-1}), \tag{6.438}$$

where $\mathcal{F}(\theta)$ is the Fisher information matrix of time series:

$$\mathcal{F}(\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\frac{\partial}{\partial\theta}f_\theta(\lambda)\frac{\partial}{\partial\theta'}f_\theta(\lambda)\{f_\theta(\lambda)\}^{-2}\,d\lambda. \tag{6.439}$$

Summarizing the above, we have the following result:

Theorem 6.35 *Under the contiguous condition (6.436),*

$$\lim_{T\to\infty}P(2|1) = \lim_{T\to\infty}P(1|2) = \Phi\left[-\frac{1}{2}\sqrt{h'\mathcal{F}(\theta)h}\right]. \tag{6.440}$$

Up to now, we assumed that $\{X_t\}$ is a scalar valued Gaussian stationary process. However, we can extend this to an m-dimensional vector valued non-Gaussian general linear process. In such a case, we can use the following discriminant statistic as a natural multidimensional extension of $I(f:g)$:

$$I(\mathbf{f}:\mathbf{g}) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left[\log\frac{|\mathbf{g}(\lambda)|}{|\mathbf{f}(\lambda)|} + \mathrm{tr}\left[\mathbf{I}_T(\lambda)\left\{\mathbf{g}(\lambda)^{-1} - \mathbf{f}(\lambda)^{-1}\right\}\right]\right]d\lambda, \tag{6.441}$$

where $|\mathbf{A}|$ denotes the determinant of matrix $\mathbf{A}$, $\mathbf{I}_T(\lambda)$ is the periodogram matrix, and $\mathbf{f}(\lambda)$ and $\mathbf{g}(\lambda)$ are the spectral density matrices which describe the hypotheses of the discriminant problem:

$$\Pi_1 : \mathbf{f}(\lambda), \quad \Pi_2 : \mathbf{g}(\lambda). \tag{6.442}$$

Similarly, the discriminant rule is to choose Π_1 or Π_2 according to $I(\mathbf{f}:\mathbf{g}) > 0$ or $I(\mathbf{f}:\mathbf{g}) \le 0$. Since we do not assume the normality of the process, $I(\mathbf{f}:\mathbf{g})$ is no longer the approximation of the log-likelihood ratio. However, Zhang and Taniguchi (1994) showed Theorems 6.34 and 6.35 under suitable regularity condition, in which the right-hand side of (6.440) depends on non-Gaussian quantities of the process.

Before this, we assumed the spectral structures which describe the hypotheses of the discriminant problem are known. If we have training samples $\mathbf{X}_{T_1}^{(1)}$ and $\mathbf{X}_{T_2}^{(2)}$ which are known to belong to the categories Π_1 and Π_2, respectively, then

using parametric or nonparametric estimators of the spectral density matrices describing the hypotheses, say, $\hat{\mathbf{f}}$ and $\hat{\mathbf{g}}$, we can construct the discriminant statistic

$$\hat{I} = I(\hat{\mathbf{f}}, \hat{\mathbf{g}}), \tag{6.443}$$

which enables us to perform the same analysis. Furthermore, we can extend the discriminant analysis between two categories Π_1 and Π_2 to that among p categories Π_i, $(i = 1, \ldots, p)$.

All of the $I(f : g)$, $I(\mathbf{f} : \mathbf{g})$ and $I(\hat{\mathbf{f}} : \hat{\mathbf{g}})$ are discriminant statistics based on approximation of the likelihood ratio. We can also construct discriminant statistics not based on the likelihood ratio. Let $\{X_t\}$ and $\{Y_t\}$ be stationary processes which have spectral densities $f_X(\lambda)$ and $f_Y(\lambda)$, respectively. Consider the discriminant problem

$$\Pi_1 : f_X(\lambda), \quad \Pi_2 : f_Y(\lambda). \tag{6.444}$$

Assume that we obtain an observation of new time series $\{Z_t\}$ which is known to belong to one of two categories Π_1 and Π_2, but we do not know to which category it belongs. We deal with the problem of assigning it into either of them. Then we can also consider the discriminant problem based on α-entropy criterion:

$$D_\alpha(f, g) = \int \left[\log \left\{ (1 - \alpha) + \alpha \frac{f(\lambda)}{g(\lambda)} \right\} - \alpha \log \frac{f(\lambda)}{g(\lambda)} \right] d\lambda, \quad (\alpha \in (0, 1)). \tag{6.445}$$

Let $\hat{f}_X(\lambda)$, $\hat{f}_Y(\lambda)$ and $\hat{f}_Z(\lambda)$ be estimators of $f(\lambda)$, when we use the observations of $\{X_t\}$, $\{Y_t\}$ and $\{Z_t\}$, respectively. Since $D_\alpha(f, g)$ measures a sort of distance between f and g, we propose the discriminant statistic

$$\hat{B}_\alpha \equiv D_\alpha(\hat{f}_Y, \hat{f}_Z) - D_\alpha(\hat{f}_X, \hat{f}_Z) \tag{6.446}$$

and assign $\{Z_t\}$ into Π_1 and Π_2 according to $\hat{B}_\alpha > 0$ and $\hat{B}_\alpha \leq 0$. As in the case of $I(f : g)$, Kakizawa et al. (1998) extended the criterion $D_\alpha(,)$ to the case that all $\{X_t\}$, $\{Y_t\}$ and $\{Z_t\}$ are non-Gaussian vector valued stationary processes. Furthermore, using nonparametric spectral density matrix estimators of them, they gave applications to the problems of classifying earthquakes and mining explosions.

Empirical time series data recorded in real phenomena such as seismic record and financial time series, are often nonstationary and non-Gaussian. The discriminant analysis of such multivariate nonstationary and non-Gaussian time series data is increasing importance. In what follows we investigate the problem of classifying a multivariate non-Gaussian locally stationary process $\{\mathbf{X}_{t,T}\}$ into one of two categories described by two hypotheses $\Pi_i : \mathbf{f}_i(u, \lambda)$, $i = 1, 2$, where $\mathbf{f}_i(u, \lambda)$ are time varying spectral density matrices. In line with Kakizawa et al. (1998), we generalize the α-entropy criterion to a nonlinear

integral functional of time varying spectral densities, which includes $I(\mathbf{f} : \mathbf{g})$ and $D_\alpha(f, g)$.

We first give multivariate extension of the definition of locally stationary processes in Section 6.9, which is due to Dahlhaus (2000).

Definition 6.3 *A sequence of multivariate stochastic processes* $\mathbf{X}_{t,T} = \left(X_{t,T}^{(1)}, \ldots, X_{t,T}^{(m)}\right)'$, $(t = 2-N/2, \ldots, 1, \ldots, T, \ldots, T+N/2; T, N \geq 1)$ *is called locally stationary with mean vector* $\mathbf{0}$ *and transfer function matrix* $\mathbf{A}^\circ$ *if there exists a representation*

$$\mathbf{X}_{t,T} = \int_{-\pi}^{\pi} \exp(i\lambda t)\mathbf{A}_{t,T}^\circ(\lambda)d\boldsymbol{\xi}(\lambda), \tag{6.447}$$

where

(i) $\boldsymbol{\xi}(\lambda) = (\xi_1(\lambda), \ldots, \xi_m(\lambda))'$ *is a complex valued stochastic vector process on* $[-\pi, \pi]$ *with* $\overline{\xi_a(\lambda)} = \xi_a(-\lambda)$ *and*

$$\mathrm{cum}\{d\xi_{a_1}(\lambda_1), \ldots, d\xi_{a_k}(\lambda_k)\} = \eta\left(\sum_{j=1}^{k}\lambda_j\right)\frac{\kappa_{a_1\ldots a_k}}{(2\pi)^{k-1}}d\lambda_1\cdots d\lambda_{k-1}, \tag{6.448}$$

for $k \geq 2$, $a_1, \ldots, a_k = 1, \ldots, m$, *where* $\mathrm{cum}\{\ldots\}$ *denotes the cumulant of* k*th order, and* $\eta(\lambda) = \sum_{j=-\infty}^{\infty}\delta(\lambda + 2\pi j)$ *is the period* 2π *extension of the Dirac delta function.*

(ii) *There exists a constant* K *and a* 2π*-periodic matrix valued function* $\mathbf{A} : [0, 1] \times \mathbf{R} \to \mathbf{C}^{m\times m}$ *with* $\mathbf{A}(u, -\lambda) = \overline{\mathbf{A}(u, \lambda)}$ *and*

$$\sup_{t,\lambda}\left|\mathbf{A}_{t,T}^\circ(\lambda)_{ab} - \mathbf{A}\left(\frac{t}{T}, \lambda\right)_{ab}\right| \leq KT^{-1} \tag{6.449}$$

for all $a, b = 1, \ldots, m$ *and* $T \in \mathbf{N}$. $\mathbf{A}(u, \lambda)$ *is assumed to be continuous in* u.

$\mathbf{f}(u, \lambda) := \mathbf{A}(u, \lambda)\boldsymbol{\Omega}\mathbf{A}(u, \lambda)^*$ is called the time varying spectral density matrix of the process, where $\boldsymbol{\Omega} = (\kappa_{ab})_{a,b=1,\ldots,m}$ and $\mathbf{D}^*$ denotes the complex conjugate of matrix $\mathbf{D}$. Write

$$\boldsymbol{\varepsilon}_t := \int_{-\pi}^{\pi}\exp(i\lambda t)d\boldsymbol{\xi}(\lambda), \tag{6.450}$$

then $E(\boldsymbol{\varepsilon}_t) = \mathbf{0}$, $E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \boldsymbol{\Omega}$ and $E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_s')$ for $t \neq s$ is the zero matrix. Here we assume the following linear representation for $\{\mathbf{X}_{t,T}\}$.

Assumption 6.29 $\mathbf{X}_{t,T}$ *has the MA*(∞) *representation*

$$\mathbf{X}_{t,T} = \sum_{s=-\infty}^{\infty}\mathbf{a}_{t,T}(s)\boldsymbol{\varepsilon}_{t-s}, \tag{6.451}$$

that is,

$$\mathbf{A}^{\circ}_{t,T}(\lambda) = \sum_{s=-\infty}^{\infty} \mathbf{a}_{t,T}(s) \exp(-i\lambda s), \tag{6.452}$$

where the coefficients fulfill

$$\sup_{t} \sum_{s=-\infty}^{\infty} \left| \left\{ \mathbf{a}_{t,T}(s) - \mathbf{a}_s \left(\frac{t}{T} \right) \right\}_{cd} \right| = O(T^{-1}) \tag{6.453}$$

for all $c, d = 1, \ldots m$*, with continuous matrix function* $\mathbf{a}_s(u)$*. Then, the condition (6.449) is fulfilled for*

$$\mathbf{A}(u, \lambda) = \sum_{s=-\infty}^{\infty} \mathbf{a}_s(u) \exp(-i\lambda s). \tag{6.454}$$

Furthermore we make the following assumption on the transfer function matrix $\mathbf{A}(u, \lambda)$.

Assumption 6.30 (i) *The transfer function matrix* $\mathbf{A}(u, \lambda)$ *is* 2π*-periodic in* λ *and the periodic extension is twice differentiable in* u *and* λ *with uniformly bounded continuous derivatives* $\frac{\partial^2}{\partial u^2}\mathbf{A}$*,* $\frac{\partial^2}{\partial \lambda^2}\mathbf{A}$ *and* $\frac{\partial^2}{\partial u^2}\frac{\partial}{\partial \lambda}\mathbf{A}$*.*

(ii) *All the eigenvalues of* $\mathbf{f}(u, \lambda)$ *are bounded from below and above by some constants* $\delta_1, \delta_2 > 0$ *uniformly in* u *and* λ*.*

As an estimator of $\mathbf{f}(u, \lambda)$, we use the nonparametric estimator of kernel type defined by

$$\hat{\mathbf{f}}_T(u, \lambda) = \int_{-\pi}^{\pi} W_T(\lambda - \mu) \mathbf{I}_N(u, \mu) d\mu, \tag{6.455}$$

where $W_T(\omega) = M \sum_{\nu=-\infty}^{\infty} W\{M(\omega + 2\pi\nu)\}$ is a weight function and $M > 0$ depends on T, and $\mathbf{I}_N(u, \lambda)$ is the data tapered periodogram matrix over the segment $\{[uT] - N/2 + 1, [uT] + N/2\}$ defined as

$$\mathbf{I}_N(u, \lambda) = \frac{1}{2\pi H_{2,N}} \left\{ \sum_{s=1}^{N} h\left(\frac{s}{N} \right) \mathbf{X}_{[uT]-N/2+s,T} \exp\{i\lambda s\} \right\} \left\{ \sum_{r=1}^{N} h\left(\frac{r}{N} \right) \mathbf{X}_{[uT]-N/2+r,T} \exp\{i\lambda r\} \right\}^{*}. \tag{6.456}$$

Here $h : [0, 1] \to \mathbf{R}$ is a data taper and $H_{2,N} = \sum_{s=1}^{N} h\left(\frac{s}{N}\right)^2$. It should be noted that $\mathbf{I}_N(u, \lambda)$ is not a consistent estimator of the spectral density. To make a consistent estimator of $\mathbf{f}(u, \lambda)$ we have to smooth it over neighboring frequencies.

Now we impose the following assumptions on $W(\cdot)$ and $h(\cdot)$.

Assumption 6.31 *The weight function* $W : \mathbf{R} \to [0, \infty]$ *satisfies* $W(x) = 0$ *for* $x \notin [-1/2, 1/2]$*, and is a continuous and even function satisfying* $\int_{-1/2}^{1/2} W(x)dx = 1$ *and* $\int_{-1/2}^{1/2} x^2 W(x)dx < \infty$.

Assumption 6.32 *The data taper* $h : \mathbf{R} \to \mathbf{R}$ *satisfies (i)* $h(x) = 0$ *for all* $x \notin [0, 1]$ *and* $h(x) = h(1-x)$*, (ii)* $h(x)$ *is continuous on* $\mathbf{R}$*, twice differentiable at all* $x \notin U$ *where* U *is a finite set of* $\mathbf{R}$*, and* $\sup_{x \notin U} |h''(x)| < \infty$.

Write

$$K_t(x) := \left\{ \int_0^1 h(x)^2 dx \right\}^{-1} h(x + 1/2)^2, \quad x \in [-1/2, 1/2], \tag{6.457}$$

which plays a role of kernel in the time domain.

Furthermore, we assume that

Assumption 6.33 *parameters* $M = M(T)$ *and* $N = N(T)$*,* $(M \ll N \ll T)$ *satisfy*

$$\frac{\sqrt{T}}{M^2} = o(1), \ \frac{N^2}{T^{\frac{3}{2}}} = o(1), \ \frac{\sqrt{T} \log N}{N} = o(1). \tag{6.458}$$

The following lemmas are a multivariate version of Theorem 2.2 of Dahlhaus (1996c) and Theorem A.2 of Dahlhaus (1997) (See also Sakiyama and Taniguchi (2003)).

Lemma 6.13 *Assume that Assumptions 6.29-6.33 hold. Then*

(i)

$$\begin{aligned} E\left\{\mathbf{I}_N(u, \lambda)\right\} =& \mathbf{f}(u, \lambda) + \frac{N^2}{2T^2} \int_{-1/2}^{1/2} x^2 K_t(x)^2 dx \frac{\partial^2}{\partial u^2} \mathbf{f}(u, \lambda) \\ &+ o\left(\frac{N^2}{T^2}\right) + O\left(\frac{\log N}{N}\right), \end{aligned} \tag{6.459}$$

(ii)

$$\begin{aligned} E\left\{\hat{\mathbf{f}}(u, \lambda)\right\} =& \mathbf{f}(u, \lambda) + \frac{N^2}{2T^2} \int_{-1/2}^{1/2} x^2 K_t(x)^2 dx \frac{\partial^2}{\partial u^2} \mathbf{f}(u, \lambda) \\ &+ \frac{1}{2M^2} \int_{-1/2}^{1/2} x^2 W(x)^2 dx \frac{\partial^2}{\partial \lambda^2} \mathbf{f}(u, \lambda) \\ &+ o\left(\frac{N^2}{T^2} + M^{-2}\right) + O\left(\frac{\log N}{N}\right), \end{aligned} \tag{6.460}$$

(iii)

$$\sum_{i,j=1}^{m} \mathrm{Var}\left\{\hat{\mathbf{f}}_{ij}(u,\lambda)\right\} = \frac{M}{N}\sum_{i,j=1}^{m} \mathbf{f}_{ij}(u,\lambda)^2 \int_{-1/2}^{1/2} K_t(x)^2 dx \int_{-1/2}^{1/2} W(x)^2 dx \left(2\pi + 2\pi\{\lambda \equiv 0 \ mod\pi\}\right) + o\left(\frac{M}{N}\right). \tag{6.461}$$

Hence, we have

$$E\left\|\hat{\mathbf{f}}(u,\lambda) - \mathbf{f}(u,\lambda)\right\|^2 = O\left(\frac{M}{N}\right) + O\left(M^{-2} + N^2T^{-2}\right)^2 = O\left(\frac{M}{N}\right), \tag{6.462}$$

where $\|D\|$ denotes the Euclidean norm of the matrix D; $\|D\| = \{tr\,(DD^*)\}^{1/2}$.

Lemma 6.14 *Assume that Assumptions 6.29-6.33 hold. Let $\phi_j(u,\lambda)$, $j = 1,\ldots,k$ be an $m \times m$ matrix-valued continuous function on $[0,1] \times [-\pi,\pi]$ which satisfies the same conditions as those for the transfer function matrix $\mathbf{A}(u,\lambda)$ in Assumption 6.30, $\phi_j(u,\lambda)^* = \phi_j(u,\lambda)$ and $\phi_j(u,-\lambda) = \phi_j(u,\lambda)'$. Then*

$$L_T(\phi_j) = \sqrt{T}\left[\frac{1}{T}\sum_{t=1}^{T}\int_{-\pi}^{\pi} \mathrm{tr}\left\{\phi_j\left(\frac{t}{T},\lambda\right)\mathbf{I}_N\left(\frac{t}{T},\lambda\right)\right\}d\lambda - \int_0^1\int_{-\pi}^{\pi} \mathrm{tr}\left\{\phi_j(u,\lambda)\,\mathbf{f}(u,\lambda)\right\}d\lambda du\right], \quad j = 1,\ldots,k \tag{6.463}$$

have, asymptotically, a normal distribution with zero mean vector and covariance matrix V whose (i,j)th element is

$$4\pi\int_0^1\left[\int_{-\pi}^{\pi}\mathrm{tr}\left\{\phi_i(u,\lambda)\mathbf{f}(u,\lambda)\phi_j(u,\lambda)\mathbf{f}(u,\lambda)\right\}d\lambda + \frac{1}{4\pi^2}\sum_{a_1,a_2,a_3,a_4}\sum_{b_1,b_2,b_3,b_4}\kappa_{b_1b_2b_3b_4}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\phi_i(u,\lambda)_{a_1a_2}\phi_j(u,\mu)_{a_4a_3} \mathbf{A}(u,\lambda)_{a_2b_1}\mathbf{A}(u,-\lambda)_{a_1b_2}\mathbf{A}(u,-\mu)_{a_4b_3}\mathbf{A}(u,\mu)_{a_3b_4}d\lambda d\mu\right]du. \tag{6.464}$$

Assumption 6.33 does not coincide with Assumption A.1 (ii) of Dahlhaus (1997). As mentioned in Remark A.3 of Dahlhaus (1997), Assumption A.1 (ii) of Dahlhaus (1997) is required because of the $\sqrt{T}$-unbiasedness at boundaries 0 and 1. If we assume that $\{\mathbf{X}_{2-N/2,T},\ldots,\mathbf{X}_{0,T}\}$ and $\{\mathbf{X}_{T+1,T},\ldots,\mathbf{X}_{T+N/2,T}\}$

are available under Assumption 6.33, then from Lemma 6.13 (i)

$$\begin{aligned} E\left\{L_T(\phi_j)\right\} =& \sqrt{T}E\left[\frac{1}{T}\sum_{t=1}^{T}\int_{-\pi}^{\pi} tr\left\{\phi_j\left(\frac{t}{T},\lambda\right)\mathbf{I}_N\left(\frac{t}{T},\lambda\right)\right\}d\lambda\right. \\ & \left. -\int_0^1\int_{-\pi}^{\pi} tr\left\{\phi_j\left(u,\lambda\right)\mathbf{f}\left(u,\lambda\right)\right\}d\lambda du\right] \\ =& O\left(\sqrt{T}\left(\frac{N^2}{T^2}+\frac{\log N}{N}+\frac{1}{T}\right)\right) = o\left(1\right). \end{aligned} \tag{6.465}$$

Now, we suppose that we have a collection of zero-mean m-dimensional vector locally stationary time series $\mathbf{X}_{t,T} = \left(X_{t,T}^{(1)}, X_{t,T}^{(2)}, \ldots, X_{t,T}^{(m)}\right)', \ t = 1, 2, \ldots, T$. Denote by $p_i(\mathbf{x}),\ i = 1, 2$, the probability density functions of the $mT \times 1$ vector $\mathbf{X}_T = \left(\mathbf{X}_{1,T}', \mathbf{X}_{2,T}', \ldots, \mathbf{X}_{T,T}'\right)'$ under two hypotheses $\Pi_i,\ i = 1, 2$, respectively. In the case of locally stationary processes, $\Pi_i,\ i = 1, 2$ are, respectively, described by the time varying spectral density matrices $\mathbf{f}_i(u, \lambda),\ i = 1, 2$ corresponding to $mT \times mT$ matrices $\mathbf{\Sigma}_T(p_i),\ i = 1, 2$. Although the processes concerned later are not restricted to be normal, it is convenient to use the normal assumption temporarily to motivate measures of disparity between the densities $p_i(\cdot),\ i = 1, 2$.

One classical measure of disparity between two multivariate densities is the *Kullback Leibler (KL) discriminant information*, defined by

$$K(p_j; p_k) = E_{p_j}\left\{\log\frac{p_j(\mathbf{x})}{p_k(\mathbf{x})}\right\}, \tag{6.466}$$

where E_p denotes the expectation under the density $p(\cdot)$. The KL discriminant information takes the form

$$K(p_j; p_k) = \frac{1}{2}\left[\mathrm{tr}\left\{\mathbf{\Sigma}_T(p_j)\mathbf{\Sigma}_T(p_k)^{-1}\right\} - \log\frac{|\mathbf{\Sigma}_T(p_j)|}{|\mathbf{\Sigma}_T(p_k)|} - mT\right] \tag{6.467}$$

when $p_i(\mathbf{x})$ correspond to two hypothetical zero-mean multivariate normal distributions. The $mT \times mT$ covariance matrices $\mathbf{\Sigma}_T(p_i)$ contain the $m \times m$ matrices $\mathbf{\Sigma}_{st}^T(p_i),\ s, t = 1, \ldots, T$ as blocks, where

$$\mathbf{\Sigma}_{st}^T(p_i) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\exp\left\{i\lambda(s-t)\right\}\mathbf{A}_{s,T}^{\circ}(\lambda)\mathbf{\Omega}\mathbf{A}_{t,T}^{\circ}(-\lambda)'d\lambda. \tag{6.468}$$

Parzen (1992) proposed to use the *Chernoff (CH) information measure*

$$B_\alpha(p_j; p_k) = -\log E_{p_j}\left[\left\{\frac{p_k(\mathbf{x})}{p_j(\mathbf{x})}\right\}^\alpha\right], \tag{6.469}$$

as a measure of disparity between the two densities, where the measure is indexed by α, $0 < \alpha < 1$. For $\alpha = \frac{1}{2}$, the Chernoff information measure is the symmetric divergence measure. For two normal random vectors differing only

in the covariance structure, the measure (6.469) takes the form

$$B_\alpha(p_j;p_k)=\frac{1}{2}\left\{\log\frac{|\alpha\boldsymbol{\Sigma}_T(p_j)+(1-\alpha)\boldsymbol{\Sigma}_T(p_k)|}{|\boldsymbol{\Sigma}_T(p_k)|}-\alpha\log\frac{|\boldsymbol{\Sigma}_T(p_j)|}{|\boldsymbol{\Sigma}_T(p_k)|}\right\}. \tag{6.470}$$

It is of interest to note the antisymmetry property $B_\alpha(p_j;p_k)=B_{1-\alpha}(p_k;p_j)$ and that $B_\alpha(p_j;p_k)$, scaled by $\alpha(1-\alpha)$ converges to $K(p_j;p_k)$ for $\alpha\to 0$ and to $K(p_k;p_j)$ for $\alpha\to 1$. Hence the Cernoff measure behaves like the two Kullback-Leibler measures when α is close to boundaries 0 and 1.

It should be recognized that the information measures (6.467) and (6.470) both involve $mT\times mT$ matrices whose dimensions tend to infinity as $T\to\infty$. As in the case of stationary processes, it is natural to use spectral approximations in terms of the time varying spectral density matrices $\mathbf{f}_i(u,\lambda)$, $i=1,2$. The appropriate versions of (6.467) and (6.470) are

$$\begin{aligned}K(\mathbf{f}_j;\mathbf{f}_k)&=\lim_{T\to\infty}T^{-1}K(p_j;p_k)\\&=\frac{1}{4\pi}\int_0^1\int_{-\pi}^{\pi}\left[\mathrm{tr}\left\{\mathbf{f}_j(u,\lambda)\mathbf{f}_k^{-1}(u,\lambda)\right\}-\log\frac{|\mathbf{f}_j(u,\lambda)|}{|\mathbf{f}_k(u,\lambda)|}-m\right]d\lambda du\end{aligned} \tag{6.471}$$

and

$$\begin{aligned}B_\alpha(\mathbf{f}_j;\mathbf{f}_k)&=\lim_{T\to\infty}T^{-1}B_\alpha(p_j;p_k)\\&=\frac{1}{4\pi}\int_0^1\int_{-\pi}^{\pi}\left\{\log\frac{|\alpha\mathbf{f}_j(u,\lambda)+(1-\alpha)\mathbf{f}_k(u,\lambda)|}{|\mathbf{f}_k(u,\lambda)|}-\alpha\log\frac{|\mathbf{f}_j(u,\lambda)|}{|\mathbf{f}_k(u,\lambda)|}\right\}d\lambda du.\end{aligned} \tag{6.472}$$

Note here that the time varying spectral matrices $\mathbf{f}_i(u,\lambda)$ correspond to the multivariate densities $p_i(\mathbf{x})$. The advantage of (6.471) and (6.472) is that the evaluation problem is reduced to inverting $m\times m$ matrices. Both forms (6.471) and (6.472) are integral functionals of the matrix product $\mathbf{f}_j(u,\lambda)\mathbf{f}_k^{-1}(u,\lambda)$ and can be generalized to the following disparity measure

$$D_H(\mathbf{f}_j;\mathbf{f}_k)=\frac{1}{4\pi}\int_0^1\int_{-\pi}^{\pi}H\left\{\mathbf{f}_j(u,\lambda)\mathbf{f}_k^{-1}(u,\lambda)\right\}d\lambda du \tag{6.473}$$

where $H(\cdot)$ is some smooth function. To ensure that $D_H(\mathbf{f}_j;\mathbf{f}_k)$ has the quasi-distance property, we require $D_H(\mathbf{f}_j;\mathbf{f}_k)\geq 0$, and that the equality holds if and only if $\mathbf{f}_j=\mathbf{f}_k$ almost everywhere. The function $H(\mathbf{Z})$ must have a unique minimum at $\mathbf{Z}=\mathbf{E}_m$, the identity matrix. There are many possible choices of $H(\mathbf{Z})$ such that $D_H(\cdot;\cdot)$ satisfies the quasi-distance property, but we consider only the two cases corresponding to (6.471) and (6.472):

$$H_K(\mathbf{Z})=\mathrm{tr}\{\mathbf{Z}\}-\log|\mathbf{Z}|-m \tag{6.474}$$

and

$$H_{B_\alpha}(\mathbf{Z}) = \log|\alpha\mathbf{Z} + (1-\alpha)\mathbf{E}_m| - \alpha\log|\mathbf{Z}|. \tag{6.475}$$

Note that another possible choice is the quadratic function

$$H_Q(\mathbf{Z}) = \frac{1}{2}\mathrm{tr}\,(\mathbf{Z} - \mathbf{E}_m)^2. \tag{6.476}$$

Generally, $D_H(\cdot;\cdot)$ is not symmetric but can easily be made so by defining

$$\widetilde{H}(\mathbf{Z}) = H(\mathbf{Z}) + H(\mathbf{Z}^{-1}). \tag{6.477}$$

The general form (6.473) can be approximated by sum over frequencies of the form $\lambda_s = 2\pi s/T$ and time $u_t = t/T$, $s, t = 1, 2, \ldots T$, i.e.,

$$D_H(\mathbf{f}_j;\mathbf{f}_k) \approx \frac{1}{2T^2}\sum_{s,t=1}^{T} H\left\{\mathbf{f}_j(u_t,\lambda_s)\mathbf{f}_k^{-1}(u_t,\lambda_s)\right\}. \tag{6.478}$$

Suppose that we wish to investigate the problem of classifying a realization $\mathbf{X}_T = \left(\mathbf{X}'_{2-N/2,T}, \ldots, \mathbf{X}'_{1,T}, \ldots, \mathbf{X}'_{T,T}, \ldots, \mathbf{X}'_{T+N/2,T}\right)'$ into one of two known categories Π_j, $j = 1, 2$, where Π_j is described by the time varying spectral density matrix $\mathbf{f}_j(u,\lambda)$. Let $\hat{\mathbf{f}}_T(u,\lambda)$ be the nonparametric time varying spectral density estimator given by (6.455), which is based on observation to be classified. We measure the disparity between the sample spectrum of $\mathbf{X}_T$ and category Π_j by $D_H(\hat{\mathbf{f}}_T;\mathbf{f}_j)$. Then the proposed rule is to classify $\mathbf{X}_T$ into Π_1 or Π_2 according to $D_H > 0$ or $D_H \le 0$, where D_H is the discriminant function defined by

$$D_H = D_H(\hat{\mathbf{f}}_T;\mathbf{f}_2) - D_H(\hat{\mathbf{f}}_T;\mathbf{f}_1). \tag{6.479}$$

In the following we examine the asymptotic properties of discriminant function (6.479). Assume that the category Π_j is an m-variate linear process of the form $\mathbf{X}_{t,T} = \sum_{k=-\infty}^{\infty}\mathbf{a}_{t,T}^{(j)}(k)\boldsymbol{\varepsilon}_{t-k}$, where $m \times m$ matrices $\mathbf{a}_{t,T}^{(j)}(k)$'s and i.i.d. $m \times 1$ zero mean vectors $\boldsymbol{\varepsilon}_t$'s satisfy Assumptions 6.29 and 6.30. The use of $\hat{\mathbf{f}}_T(u,\lambda)$ instead of the data tapered periodogram $\mathbf{I}_N(u,\lambda)$ is essential, because $D_H(\mathbf{I}_N;\mathbf{g})$ does not converge in probability to $D_H(\mathbf{f}_j;\mathbf{g})$ under Π_j if $D_H(\mathbf{I}_N;\mathbf{g})$ is nonlinear with respect to $\mathbf{I}_N$ (See Taniguchi and Kakizawa (2000)). We discuss the performance of the discriminant function (6.479). First, we evaluate the asymptotics of the misclassification probabilities based on D_H;

$$P_{D_H}(2|1) = \mathrm{P}(D_H \le 0|\Pi_1) \tag{6.480}$$

and

$$P_{D_H}(1|2) = \mathrm{P}(D_H > 0|\Pi_2). \tag{6.481}$$

Concretely speaking we will show that $P_{D_H}(2|1)$ and $P_{D_H}(1|2)$ converge to

zero as $T \to \infty$ if $\mathbf{f}_1(u,\lambda) \neq \mathbf{f}_2(u,\lambda)$. Next assuming that Π_1 is contiguous to Π_2, the limit of the two misclassification probabilities will be evaluated. Then we will elucidate the asymptotic optimality and robustness.

In what follows, set $(j,k) = (1,2)$ or $(2,1)$. We give the following assumptions on $H(\mathbf{Z})$.

Assumption 6.34 **(i)** *$H : \mathbf{C}^{m\times m} \to \mathbf{R}$ is a real-valued holomorphic function defined on an open set D in $\mathbf{C}^{m\times m}$.*

(ii) *$H(\mathbf{E}_m) = 0$ and $H^{(1)}(\mathbf{E}_m) = \mathbf{0}$ ($m \times m$ zero matrix), where $H^{(1)}\{(\cdot)\}$ is the first derivative of $\mathbf{Z}$ at $(\cdot)$ whose (a,b)th element is $\frac{\partial}{\partial Z_{ab}} H(\mathbf{Z})$. The $m^2 \times m^2$ Hessian matrix of $H(\mathbf{Z})$, defined by*

$$\frac{\partial}{\partial (\mathrm{vec}\mathbf{Z})'} \left(\frac{\partial}{\partial (\mathrm{vec}\mathbf{Z})'} H(\mathbf{Z}) \right)' \tag{6.482}$$

is positive definite at $\mathbf{Z} = \mathbf{E}_m$. That is, $H(\mathbf{Z})$ has a unique minimum zero at $\mathbf{Z} = \mathbf{E}_m$.

(iii) *The $m \times m$ matrix $\mathbf{Q}_{j,k}(u,\lambda)$ defined by*

$$\mathbf{Q}_{j,k}(u,\lambda) = \mathbf{f}_k^{-1}(u,\lambda) \left[H^{(1)} \left\{ \mathbf{f}_j(u,\lambda) \mathbf{f}_k^{-1}(u,\lambda) \right\} \right]' \tag{6.483}$$

satisfies $\mathbf{Q}_{j,k}(u,\lambda)^* = \mathbf{Q}_{j,k}(u,\lambda)$ and $\mathbf{Q}_{j,k}(u,-\lambda) = \mathbf{Q}_{j,k}(u,\lambda)'$.

Theorem 6.36 *Under the Assumptions 6.29-6.34, suppose that $\mathbf{f}_1(u,\lambda) \neq \mathbf{f}_2(u,\lambda)$ on a set of positive Lebesgue measures. Then under Π_j,*

$$D_H \xrightarrow{\mathcal{P}} (-1)^{j+1} D_H(\mathbf{f}_j; \mathbf{f}_k) \tag{6.484}$$

and

$$\sqrt{T} \left\{ D_H + (-1)^j D_H(\mathbf{f}_j; \mathbf{f}_k) \right\} \xrightarrow{d} N(0, V_H^2(j,k)), \quad \text{as } T \to \infty, \tag{6.485}$$

where $D_H(\mathbf{f}_j; \mathbf{f}_k)$ is the integral disparity (6.473) and

$$V_H^2(j,k) = \int_0^1 \left[\frac{1}{4\pi} \int_{-\pi}^{\pi} \mathrm{tr} \left\{ \mathbf{Q}_{j,k}(u,\lambda) \mathbf{f}_j(u,\lambda) \right\}^2 d\lambda + \frac{1}{64\pi^4} \sum_{a,b,c,d} \kappa_{abcd} \gamma_{ab}^{(j,k)}(u) \gamma_{cd}^{(j,k)}(u) \right] du, \tag{6.486}$$

with

$$\mathbf{\Gamma}_H^{(j,k)}(u) = \left\{ \gamma_{ab}^{(j,k)}(u) \right\}_{a,b=1,\ldots,m} = \int_{-\pi}^{\pi} \mathbf{A}_j(u,\lambda)^* \mathbf{Q}_{j,k}(u,\lambda) \mathbf{A}_k(u,\lambda) d\lambda. \tag{6.487}$$

PROOF

Let

$$\begin{aligned}\hat{H}_j(u,\lambda) \equiv & H\left\{\hat{\mathbf{f}}_T(u,\lambda)\mathbf{f}_k^{-1}(u,\lambda)\right\} - H\left\{\mathbf{f}_j(u,\lambda)\mathbf{f}_k^{-1}(u,\lambda)\right\} \\ & - \operatorname{tr}\left[\mathbf{Q}_{j,k}(u,\lambda)\left\{\hat{\mathbf{f}}_T(u,\lambda) - \mathbf{f}_j(u,\lambda)\right\}\right], \end{aligned} \tag{6.488}$$

then from Lemma 6.13, the same argument as in Theorem 1 of Taniguchi et al. (1996), leads to, under Π_j

$$\hat{H}_j(u,\lambda) = O_{\mathcal{P}}\left(\frac{M}{N}\right) \tag{6.489}$$

and

$$H\left\{\hat{\mathbf{f}}_T(u,\lambda)\mathbf{f}_j^{-1}(u,\lambda)\right\} = O_{\mathcal{P}}\left(\frac{M}{N}\right), \tag{6.490}$$

uniformly in λ and u. Since, D_H is written as

$$D_H = \frac{1}{4\pi}\int_0^1\int_{-\pi}^{\pi}\left[H\left\{\hat{\mathbf{f}}_T(u,\lambda)\mathbf{f}_2^{-1}(u,\lambda)\right\} - H\left\{\hat{\mathbf{f}}_T(u,\lambda)\mathbf{f}_1^{-1}(u,\lambda)\right\}\right]d\lambda du, \tag{6.491}$$

it follows from (6.489) and (6.490), under Π_j

$$\begin{aligned}&\sqrt{T}\left\{D_H + (-1)^j D_H(\mathbf{f}_j;\mathbf{f}_k)\right\} \\ &= \frac{(-1)^{j+1}\sqrt{T}}{4\pi}\int_0^1\int_{-\pi}^{\pi}\operatorname{tr}\left[\mathbf{Q}_{j,k}(u,\lambda)\left\{\hat{\mathbf{f}}_T(u,\lambda) - \mathbf{f}_j(u,\lambda)\right\}\right]d\lambda du + o_{\mathcal{P}}(1) \\ &= \frac{(-1)^{j+1}}{4\pi}S_T + o_{\mathcal{P}}(1) \quad \text{(say)}. \end{aligned} \tag{6.492}$$

Moreover, according to Lemma 6.14,

$$\begin{aligned}&L_T\left\{\frac{(-1)^{j+1}}{4\pi}\mathbf{Q}_{j,k}\right\} \\ &= \frac{(-1)^{j+1}\sqrt{T}}{4\pi}\int_0^1\int_{-\pi}^{\pi}\operatorname{tr}\left[\mathbf{Q}_{j,k}(u,\lambda)\left\{\mathbf{I}_N(u,\lambda) - \mathbf{f}_j(u,\lambda)\right\}\right]d\lambda du + O_{\mathcal{P}}(T^{-\frac{1}{2}}) \\ &= \frac{(-1)^{j+1}}{4\pi}L_T + o_{\mathcal{P}}(1) \quad \text{(say)} \end{aligned} \tag{6.493}$$

have, asymptotically, a normal distribution with zero mean vector and covariance matrix $V_H^2(j,k)$. Thus, the proof of Theorem 6.36 is complete if we show

$S_T - L_T = o_{\mathcal{P}}(1)$. From the definition of $\hat{\mathbf{f}}$, it follows that

$$\begin{aligned}
&S_T - L_T \\
&= \sqrt{T}\int_0^1\int_{-\pi}^{\pi} \operatorname{tr}\left[\mathbf{Q}_{j,k}(u,\lambda)\left\{\int_{-\pi}^{\pi} W_T(\lambda-\mu)\mathbf{f}_j(u,\mu)d\mu - \mathbf{f}_j(u,\lambda)\right\}\right] d\lambda du \\
&\quad + \sqrt{T}\int_0^1\int_{-\pi}^{\pi} \operatorname{tr}\left[\mathbf{Q}_{j,k}(u,\lambda)\int_{-\pi}^{\pi}\{\mathbf{I}_N(u,\mu) - \mathbf{f}_j(u,\mu)\} W_T(\lambda-\mu)d\mu\right] d\lambda du \\
&\quad - \sqrt{T}\int_0^1\int_{-\pi}^{\pi} \operatorname{tr}\left[\mathbf{Q}_{j,k}(u,\mu)\{\mathbf{I}_N(u,\mu) - \mathbf{f}_j(u,\mu)\}\right] d\mu du \\
&= \sqrt{T}\int_0^1\int_{-\pi}^{\pi} \operatorname{tr}\left[\mathbf{Q}_{j,k}(u,\lambda)\left\{\int_{-\pi}^{\pi} W_T(\lambda-\mu)\mathbf{f}_j(u,\mu)d\mu - \mathbf{f}_j(u,\lambda)\right\}\right] d\lambda du \\
&\quad + \sqrt{T}\int_0^1\int_{-\pi}^{\pi} \operatorname{tr}\left[\mathbf{D}(u,\mu)\{\mathbf{I}_N(u,\mu) - \mathbf{f}_j(u,\mu)\}\right] d\mu du \\
&= L_T^{(1)} + L_T^{(2)} \quad \text{(say)},
\end{aligned} \tag{6.494}$$

where

$$\mathbf{D}(u,\mu) = \int_{-\infty}^{\infty}\left\{\mathbf{Q}_{j,k}\left(u,\mu+\frac{x}{M}\right) - \mathbf{Q}_{j,k}(u,\mu)\right\} W(x)dx. \tag{6.495}$$

By the dominated convergence theorem,

$$\lim_{T\to\infty}\|\mathbf{D}(u,\mu)\| = 0 \quad \text{a.e. } (\mu\in[-\pi,\pi]), \tag{6.496}$$

therefore, from Lemma 6.14, $\operatorname{Var}\left\{L_T^{(2)}\right\} = o(1)$, which implies $L_T^{(2)} = o_{\mathcal{P}}(1)$. On the other hand, by Assumption 6.31, we have

$$\int_{-\pi}^{\pi}\mathbf{f}(u,\mu)W_T(\lambda-\mu)d\mu - \mathbf{f}(u,\lambda) = O(M^{-2}), \tag{6.497}$$

hence $L_T^{(1)} = O\left(\frac{\sqrt{T}}{M^2}\right) = o(1)$. □

In view of Theorem 6.36, the limiting forms of misclassification probabilities (6.480) and (6.481) satisfy $\lim_{T\to\infty} P_{D_H}(k|j) = 0$ for $(j,k) = (1,2),(2,1)$. This shows that the discriminant D_H is consistent. From (6.485), one may also approximate them as the normal integrals

$$P_{D_H}(k|j) \approx \Phi\left(-\sqrt{T}\frac{D_H(\mathbf{f}_j;\mathbf{f}_k)}{V_H(j,k)}\right), \tag{6.498}$$

which depend on the fourth-order cumulants unless (6.487) is a zero matrix. To see robustness, we assume that the hypothetical m-variate linear process is generated by

$$\mathbf{X}_{t,T} = \sum_{s=-\infty}^{\infty}\mathbf{a}_{t,T}^{(1)}(s)\boldsymbol{\varepsilon}_{t-s} \tag{6.499}$$

under Π_1 and by

$$\mathbf{X}_{t,T} = \sum_{s=-\infty}^{\infty} \left\{ \mathbf{a}_{t,T}^{(1)}(s) + T^{-1/2}\mathbf{a}_{t,T}^{(2)}(s) \right\} \boldsymbol{\varepsilon}_{t-s} \tag{6.500}$$

under Π_2. Thus, the time varying spectral densities associated with Π_1 and Π_2 are

$$\mathbf{f}_1(u,\lambda) = \mathbf{A}^{(1)}(u,\lambda)\boldsymbol{\Omega}\mathbf{A}^{(1)}(u,\lambda)^* \tag{6.501}$$

and

$$\begin{aligned}\mathbf{f}_2(u,\lambda) = &\left\{ \mathbf{A}^{(1)}(u,\lambda) + T^{-1/2}\mathbf{A}^{(2)}(u,\lambda) \right\} \\ &\boldsymbol{\Omega} \left\{ \mathbf{A}^{(1)}(u,\lambda) + T^{-1/2}\mathbf{A}^{(2)}(u,\lambda) \right\}^*, \end{aligned} \tag{6.502}$$

with $\mathbf{A}^{(i)}(u,\lambda) = \sum_{s=-\infty}^{\infty} \mathbf{a}_s^{(i)}(u)\exp(-i\lambda s)$, $i = 1, 2$. The quantities $D_H(\mathbf{f}_j;\mathbf{f}_k)$ and $V_H(j,k)$ are determined by local property of the function $H(\mathbf{Z})$ at $\mathbf{Z} = \mathbf{E}_m$.

Assumption 6.35 *The $m^2 \times m^2$ Hessian matrix of $H(\mathbf{Z})$ at $\mathbf{E}_m$ is $c\mathbf{K}_m$, where $\mathbf{K}_m$ is the commutation matrix (e.g., Magnus and Neudecker (1988)) and $c > 0$.*

Note that H_K, H_{B_α} and H_Q in (6.474), (6.475) and (6.476) satisfy Assumptions 6.34 and 6.35.

Theorem 6.37 *Let $\mathbf{f}_1$ and $\mathbf{f}_2$, defined by (6.501) and (6.502), be the hypothetical time varying spectral density matrices of m-variate linear processes (6.499) and (6.500), respectively. Under Assumptions 6.29-6.35, if $\mathbf{a}_0^{(2)}(u) = \mathbf{0}$ ($m \times m$ zero matrix), the asymptotic misclassification probabilities are independent of non-Gaussianity of the process, and are given by*

$$\begin{aligned}&\lim_{T\to\infty} P_{D_H}(2|1) = \lim_{T\to\infty} P_{D_H}(1|2) \\ &= \Phi\left(-\frac{1}{2}\sqrt{\frac{1}{4\pi}\int_0^1 \int_{-\pi}^{\pi} \mathrm{tr}\{\boldsymbol{\Delta}(u,\lambda)\}^2 d\lambda du} \right), \end{aligned} \tag{6.503}$$

with

$$\begin{aligned}\boldsymbol{\Delta}(u,\lambda) = &\left\{ \mathbf{A}^{(1)}(u,\lambda)\boldsymbol{\Omega}\mathbf{A}^{(2)}(u,\lambda)^* + \mathbf{A}^{(2)}(u,\lambda)\boldsymbol{\Omega}\mathbf{A}^{(1)}(u,\lambda)^* \right\} \\ &\left\{ \mathbf{A}^{(1)}(u,\lambda)\boldsymbol{\Omega}\mathbf{A}^{(1)}(u,\lambda)^* \right\}^{-1}. \end{aligned} \tag{6.504}$$

PROOF

From Assumptions 6.34, 6.35 and

$$\mathbf{f}_j(u,\lambda)\mathbf{f}_k(u,\lambda)^{-1} = \mathbf{E}_m + \frac{(-1)^j}{\sqrt{T}}\boldsymbol{\Delta}(u,\lambda) + O(T^{-1}), \tag{6.505}$$

it is seen that

$$D_H(\mathbf{f}_j;\mathbf{f}_k)=\frac{c}{8\pi T}\int_0^1\int_{-\pi}^{\pi}\mathrm{tr}\{\mathbf{\Delta}(u,\lambda)\}^2d\lambda du+o(T^{-1}) \tag{6.506}$$

and

$$\begin{aligned}V_H^2(j,k)=&\frac{c^2}{4\pi T}\int_0^1\int_{-\pi}^{\pi}\mathrm{tr}\{\mathbf{\Delta}(u,\lambda)\}^2d\lambda du\\&+\frac{c^2}{64\pi^4T}\sum_{a_1,a_2,a_3,a_4=1}^{m}\kappa_{a_1a_2a_3a_4}\gamma_{a_1a_2}\gamma_{a_3a_4}+o(T^{-1}),\end{aligned} \tag{6.507}$$

where γ_{ab} is the (a,b)th element of the $m\times m$ matrix

$$\begin{aligned}\mathbf{\Gamma}_H=&\int_0^1\int_{-\pi}^{\pi}\mathbf{A}^{(1)}(u,\lambda)\left\{\mathbf{A}^{(1)}(u,\lambda)\mathbf{\Omega}\mathbf{A}^{(1)}(u,\lambda)^*\right\}^{-1}\\&\left\{\mathbf{A}^{(1)}(u,\lambda)\mathbf{\Omega}\mathbf{A}^{(2)}(u,\lambda)^*+\mathbf{A}^{(2)}(u,\lambda)\mathbf{\Omega}\mathbf{A}^{(1)}(u,\lambda)^*\right\}\\&\left\{\mathbf{A}^{(1)}(u,\lambda)\mathbf{\Omega}\mathbf{A}^{(1)}(u,\lambda)^*\right\}^{-1}\mathbf{A}^{(1)}(u,\lambda)d\lambda du.\end{aligned} \tag{6.508}$$

If $\mathbf{\Gamma}_H=\mathbf{0}$, substituting (6.506) and (6.507) into (6.498), then the asymptotic misclassification probabilities are given by

$$\begin{aligned}&\lim_{T\to\infty}P_{D_H}(2|1)=\lim_{T\to\infty}P_{D_H}(1|2)\\&=\Phi\left(-\frac{1}{2}\sqrt{\frac{1}{4\pi}\int_0^1\int_{-\pi}^{\pi}\mathrm{tr}\{\mathbf{\Delta}(u,\lambda)\}^2d\lambda du}\right).\end{aligned} \tag{6.509}$$

Since

$$\begin{aligned}\mathbf{\Gamma}_H=&\int_0^1\int_{-\pi}^{\pi}\left\{\mathbf{A}^{(2)}(u,\lambda)^*\mathbf{A}^{(1)}(u,\lambda)^{*-1}\mathbf{\Omega}^{-1}\right.\\&\left.+\mathbf{\Omega}^{-1}\mathbf{A}^{(1)}(u,\lambda)^{-1}\mathbf{A}^{(2)}(u,\lambda)\right\}d\lambda du\\=&2\pi\int_0^1\left\{\mathbf{a}_0^{(2)}(u)'\mathbf{a}_0^{(1)}(u)^{-1'}\mathbf{\Omega}^{-1}+\mathbf{\Omega}^{-1}\mathbf{a}_0^{(1)}(u)^{-1}\mathbf{a}_0^{(2)}(u)\right\}du,\end{aligned} \tag{6.510}$$

$$\begin{aligned}\mathrm{vec}\,(\mathbf{\Gamma}_H)&=2\pi\int_0^1\mathrm{vec}\left\{\left(\mathbf{K}_m+\mathbf{E}_m\right)\mathbf{\Omega}^{-1}\mathbf{a}_0^{(1)}(u)^{-1}\mathbf{a}_0^{(2)}(u)\right\}du\\&=4\pi\int_0^1\left\{\mathbf{E}_m\otimes\mathbf{\Omega}^{-1}\mathbf{a}_0^{(1)}(u)^{-1}\right\}\mathrm{vec}\left\{\mathbf{a}_0^{(2)}(u)\right\}du,\end{aligned} \tag{6.511}$$

which implies that $\mathbf{\Gamma}_H=\mathbf{0}$ is equivalent to $\mathbf{a}_0^{(2)}(u)\equiv\mathbf{0}$. □

If the process concerned is Gaussian, the exact Gaussian log-likelihood ratio

is

$$\Lambda_T(p_1,p_2) \equiv \frac{1}{T}\ \textit{Gaussian log likelihood ratio}$$
$$= \frac{1}{2T}\mathbf{X}_T'\left\{\mathbf{\Sigma}^T(p_1)^{-1}-\mathbf{\Sigma}^T(p_2)^{-1}\right\}\mathbf{X}_T - \frac{1}{2T}\log\frac{\left|\mathbf{\Sigma}^T(p_2)\right|}{\left|\mathbf{\Sigma}^T(p_1)\right|}. \tag{6.512}$$

According to Proposition 2.5 and Lemma A.8 of Dahlhaus (2000), it is seen that, under Π_j, for each $\epsilon > 0$,

$$\begin{aligned}
&E\left\{\sqrt{T}\Lambda_T(p_1,p_2)\right\}\\
&= \frac{\sqrt{T}}{2T}\left[\mathrm{tr}\left[\mathbf{\Sigma}^T(p_j)\left\{\mathbf{\Sigma}^T(p_1)^{-1}-\mathbf{\Sigma}^T(p_2)^{-1}\right\}\right]-\log\frac{\left|\mathbf{\Sigma}^T(p_2)\right|}{\left|\mathbf{\Sigma}^T(p_1)\right|}\right]\\
&= \frac{\sqrt{T}}{4\pi}\int_0^1\int_{-\pi}^{\pi}\left[\mathrm{tr}\left[\mathbf{f}_j(u,\lambda)\left\{\mathbf{f}_2^{-1}(u,\lambda)-\mathbf{f}_1^{-1}(u,\lambda)\right\}\right]-\log\frac{|\mathbf{f}_2(u,\lambda)|}{|\mathbf{f}_1(u,\lambda)|}\right]d\lambda du\\
&\quad + O\left(T^{-\frac{1}{2}+\epsilon}+T^{-\frac{1}{2}}\log^{11}T\right)\\
&= (-1)^{j+1}\sqrt{T}D_{H_K}(\mathbf{f}_j;\mathbf{f}_k)+o(1)
\end{aligned} \tag{6.513}$$

and

$$\begin{aligned}
&\mathrm{Var}\left\{\sqrt{T}\Lambda_T(p_1,p_2)\right\} = \frac{1}{2T}\mathrm{tr}\left[\mathbf{\Sigma}^T(p_j)\left\{\mathbf{\Sigma}^T(p_1)^{-1}-\mathbf{\Sigma}^T(p_2)^{-1}\right\}\right]^2\\
&= \frac{1}{4\pi}\int_0^1\int_{-\pi}^{\pi}\mathrm{tr}\left[\mathbf{f}_j(u,\lambda)\left\{\mathbf{f}_2^{-1}(u,\lambda)-\mathbf{f}_1^{-1}(u,\lambda)\right\}\right]^2 d\lambda du\\
&\quad + O\left(T^{-1}\log^{23}T\right).
\end{aligned} \tag{6.514}$$

Therefore,

$$\begin{aligned}
&\lim_{T\to\infty}P_{GLR}(2|1) = \lim_{T\to\infty}P_{GLR}(1|2)\\
&= \lim_{T\to\infty}P_{D_H}(2|1) = \lim_{T\to\infty}P_{D_H}(1|2)\\
&= \Phi\left(-\frac{1}{2}\sqrt{\frac{1}{4\pi}\int_0^1\int_{-\pi}^{\pi}\mathrm{tr}\{\mathbf{\Delta}(u,\lambda)\}^2 d\lambda du}\right),
\end{aligned} \tag{6.515}$$

where

$$P_{GLR}(2|1) = \Pr\left\{\Lambda_T(p_1,p_2)\le 0|\Pi_1\right\} \tag{6.516}$$

and

$$P_{GLR}(1|2) = \Pr\left\{\Lambda_T(p_1,p_2) > 0|\Pi_2\right\}, \tag{6.517}$$

that is, the discriminant criterion based on D_H is asymptotically Gaussian optimal.

Remark 6.10 (Peak Robustness of B_α) *Next, we consider the case when*

the time varying spectral density of $\mathbf{X}_T$ *is contaminated by a sharp peak. In this case, we can see that* $B_\alpha(\mathbf{f}_j;\mathbf{f}_k)$ *is robust with respect to peak, but* $K(\mathbf{f}_j;\mathbf{f}_k)$ *is not so. Define*

$$\bar{\mathbf{f}}_i(u,\lambda) = \begin{cases} \mathbf{f}_i(u,\lambda) & on\ \Omega = [-\pi,\pi] - \Omega_\epsilon; \\ \mathbf{f}_i(u,\lambda)/\epsilon^r & on\ \Omega_\epsilon, \end{cases} \tag{6.518}$$

where $\Omega_\epsilon = [\lambda_0, \lambda_0 + \epsilon]$ *is an interval in* $[-\pi,\pi]$ *for sufficiently small* $\epsilon > 0$ *and* $r > 1$. *Suppose that* $\mathbf{f}_1(u,\lambda) \not\equiv \mathbf{f}_2(u,\lambda)$ *on a set of a positive Lebesgue measure. Then, under Assumption 6.30, it can be shown that*

$$\lim_{\epsilon\to 0} \left| B_\alpha(\bar{\mathbf{f}}_j,\mathbf{f}_j,\mathbf{f}_k) + (-1)^j B_\alpha(\mathbf{f}_j,\mathbf{f}_k) \right| = 0 \quad for \quad \alpha \in (0,1), \tag{6.519}$$

$$\lim_{\epsilon\to 0} \left| K(\bar{\mathbf{f}}_j,\mathbf{f}_j,\mathbf{f}_k) + (-1)^j K(\mathbf{f}_j,\mathbf{f}_k) \right| = \infty, \tag{6.520}$$

where

$$B_\alpha(\bar{\mathbf{f}}_j,\mathbf{f}_j,\mathbf{f}_k) = B_\alpha(\bar{\mathbf{f}}_j,\mathbf{f}_k) - B_\alpha(\bar{\mathbf{f}}_j,\mathbf{f}_j) \tag{6.521}$$

and

$$\begin{aligned} &K(\bar{\mathbf{f}}_j,\mathbf{f}_j,\mathbf{f}_k) = K(\bar{\mathbf{f}}_j,\mathbf{f}_k) - K(\bar{\mathbf{f}}_j,\mathbf{f}_j) \\ &= \frac{1}{4\pi}\int_0^1\int_{-\pi}^{\pi} \left[\log\frac{|\mathbf{f}_k(u,\lambda)|}{|\mathbf{f}_j(u,\lambda)|} + \mathrm{tr}\left[\bar{\mathbf{f}}_j(u,\lambda)\left\{\mathbf{f}_k^{-1}(u,\lambda) - \mathbf{f}_j^{-1}(u,\lambda)\right\}\right] \right] d\lambda du. \end{aligned} \tag{6.522}$$

That is, $B_\alpha(\mathbf{f}_j;\mathbf{f}_k)$ *is insensitive to a sharp peak in the spectral density, while* $K(\mathbf{f}_j;\mathbf{f}_k)$ *is sensitive. Thus, the discriminant statistic* $B_\alpha(\mathbf{f}_j;\mathbf{f}_k)$ *is better than* $K(\mathbf{f}_j;\mathbf{f}_k)$ *if the time varying spectral density of* $\mathbf{X}_T$ *is contaminated by a sharp peak.*

Up to now our discriminant criteria have only Gaussian optimality, therefore in the following, we discuss genuine *non-Gaussian optimal discriminant criterion*. Suppose that $\mathbf{X}_T = (X_{1,T},\ldots,X_{T,T})'$ is a realization of a scalar-valued locally stationary process with transfer function A_θ° and that time varying spectral density $f_\theta(u,\lambda) := |A_\theta(u,\lambda)|^2$ depends on a parameter vector $\theta = (\theta_1,\ldots,\theta_q) \in \Theta \subset \mathbf{R}^q$.

Let Π_1 and Π_2 be two categories with probability density functions $p_1(\mathbf{X})$ and $p_2(\mathbf{X})$, respectively. We investigate the problems of classifying a locally stationary process $\{\mathbf{X}_T\}$ into one of two categories described by two hypotheses:

$$\begin{aligned} \Pi_1 &: f_1(u,\lambda) = f_\theta(u,\lambda) \\ \Pi_2 &: f_2(u,\lambda) = f_{\theta_T}(u,\lambda), \end{aligned} \tag{6.523}$$

where $\theta_T = \theta + \frac{h}{\sqrt{T}}$, $\theta \in \Theta \subset \mathbf{R}^q$ and $h = (h_1,\ldots,h_q)'$. We assign the observed stretch $\mathbf{X}_T$ to category Π_1 if $\mathbf{X}_T$ falls in the region R_1; otherwise we assign it to Π_2, where R_1 and R_2 are exclusive and exhaustive regions in $\mathbf{R}^T$. It is well

known that the classification regions defined by

$$R_1 = \left[\mathbf{X}_T : \Lambda_T(p_1, p_2) = \log \frac{p_1(\mathbf{X}_T)}{p_2(\mathbf{X}_T)} > 0 \right] \tag{6.524}$$

give the optimal classification (See Anderson (1984)).

Besides Assumptions 6.19-6.21 in Section 6.9, we assume the following assumption on the density $p(\cdot)$ of innovation ε_t.

Assumption 6.36 **(i)** *Continuous derivatives* Dp, $D^2p \equiv D(Dp)$ *and* $D^3p \equiv D(D^2p)$ *of* $p(\cdot)$ *exist on* $\mathbf{R}$, *and* D^3p *satisfies the Lipschitz condition.*

(ii) $\int \{D^s\phi(z)\}^2 p(z)dz < \infty$, $s = 1, 2$.

By Theorem 6.22 in Section 6.9, it is seen that for all $\theta \in \Theta$, under Π_1, as $T \to \infty$, the log-likelihood ratio $\Lambda_T(p_1, p_2)$ has, asymptotically, normal distribution $N\left(\frac{1}{2}h'\Gamma(\theta)h, h'\Gamma(\theta)h\right)$.

Furthermore, under Π_2, $\Lambda_T(p_1, p_2) \xrightarrow{d} N\left(-\frac{1}{2}h'\Gamma(\theta)h, h'\Gamma(\theta)h\right)$, hence

$$\begin{aligned} &\lim_{T\to\infty} P_{LR}(2|1) = \lim_{T\to\infty} P_{LR}(1|2) \\ &= \Phi\left(-\frac{1}{2}\sqrt{h'\Gamma(\theta)h}\right), \end{aligned} \tag{6.525}$$

where

$$P_{LR}(2|1) = \Pr\{\Lambda_T(p_1, p_2) \le 0|\Pi_1\} \tag{6.526}$$

and

$$P_{LR}(1|2) = \Pr\{\Lambda_T(p_1, p_2) > 0|\Pi_2\}. \tag{6.527}$$

Since $(\varepsilon_s, s \le 0)$ are unobservable, instead of $\Lambda_T(p_1, p_2)$ we use the "quasi-log-likelihood ratio"

$$F_T(p_1, p_2) = \log \frac{F_T(\theta)}{F_T(\theta_T)} \tag{6.528}$$

with

$$F_T(\theta) = \prod_{t=1}^{T} \frac{1}{a^\circ_{\theta,t,T}(0)} p\left\{ \frac{\sum_{j=0}^{t-1} b^\circ_{\theta,t,T}(j) X_{t-j,T}}{a^\circ_{\theta,t,T}(0)} \right\}, \tag{6.529}$$

for classification criterion.

Theorem 6.38 *The discriminant criterion based on the quasi-log-likelihood ratio is asymptotically optimal.*

PROOF Under Π_1, it is seen that

$$\begin{aligned}
&\Lambda_T(p_1,p_2)-F_T(p_1,p_2)\\
&=\sum_{t=1}^{T}\left[\log p\left(\varepsilon_t\right)-\log p\left\{\frac{\sum_{j=0}^{t-1}b^{\circ}_{\theta_T,t,T}(j)X_{t-j,T}+\sum_{r=0}^{\infty}c^{\circ}_{\theta_T,t,T}(r)\varepsilon_{-r}}{a^{\circ}_{\theta_T,t,T}(0)}\right\}\right.\\
&\quad\left.-\log p\left\{\frac{\sum_{j=0}^{t-1}b^{\circ}_{\theta,t,T}(j)X_{t-j,T}}{a^{\circ}_{\theta,t,T}(0)}\right\}+\log p\left\{\frac{\sum_{j=0}^{t-1}b^{\circ}_{\theta_T,t,T}(j)X_{t-j,T}}{a^{\circ}_{\theta_T,t,T}(0)}\right\}\right]\\
&=\sum_{t=1}^{T}\left[q_{t,T}\phi\left(\varepsilon_t\right)+\frac{q_{t,T}^2}{2}D\phi\left(\varepsilon_t\right)-r_{t,T}\phi\left\{\frac{\sum_{j=0}^{t-1}b^{\circ}_{\theta,t,T}(j)X_{t-j,T}}{a^{\circ}_{\theta,t,T}(0)}\right\}\right.\\
&\quad\left.-\frac{r_{t,T}^2}{2}D\phi\left\{\frac{\sum_{j=0}^{t-1}b^{\circ}_{\theta,t,T}(j)X_{t-j,T}}{a^{\circ}_{\theta,t,T}(0)}\right\}\right]+O_{\mathcal{P}}\left(T^{-1/2}\right)\\
&=\sum_{t=1}^{T}\left\{q_{t,T}\phi\left(\varepsilon_t\right)+\frac{q_{t,T}^2}{2}D\phi\left(\varepsilon_t\right)-r_{t,T}\phi\left(\varepsilon_t\right)-\frac{r_{t,T}^2}{2}D\phi\left(\varepsilon_t\right)\right\}\\
&\quad+O_{\mathcal{P}}\left(T^{-1/2}\right)+o_{\mathcal{P}}\left(1\right),
\end{aligned}\tag{6.530}$$

where

$$\begin{aligned}
q_{t,T}&=r_{t,T}+\frac{\sum_{r=0}^{\infty}c^{\circ}_{\theta,t,T}(r)\varepsilon_{-r}}{a^{\circ}_{\theta,t,T}(0)}-\frac{\sum_{r=0}^{\infty}c^{\circ}_{\theta_T,t,T}(r)\varepsilon_{-r}}{a^{\circ}_{\theta_T,t,T}(0)}\\
&=r_{t,T}+\frac{h'}{\sqrt{T}}\sum_{r=0}^{\infty}\left\{\frac{\nabla c^{\circ}_{\theta^*,t,T}(r)}{a^{\circ}_{\theta,t,T}(0)}+\frac{c^{\circ}_{\theta_T,t,T}(r)\nabla a^{\circ}_{\theta^{**},t,T}(0)}{a^{\circ}_{\theta,t,T}(0)a^{\circ}_{\theta_T,t,T}(0)}\right\}\varepsilon_{-r}\\
&=r_{t,T}+O_{\mathcal{P}}\left(T^{-1/2}t^{-1}\right)
\end{aligned}\tag{6.531}$$

and

$$\begin{aligned}
r_{t,T}&=\frac{\sum_{j=0}^{t-1}b^{\circ}_{\theta,t,T}(j)X_{t-j,T}}{a^{\circ}_{\theta,t,T}(0)}-\frac{\sum_{j=0}^{t-1}b^{\circ}_{\theta_T,t,T}(j)X_{t-j,T}}{a^{\circ}_{\theta_T,t,T}(0)}\\
&=\frac{h'}{\sqrt{T}}\left\{\sum_{j=1}^{t-1}\frac{\nabla b^{\circ}_{\theta^{***},t,T}(j)}{a^{\circ}_{\theta,t,T}(0)}X_{t-j,T}+\sum_{j=0}^{t-1}\frac{b^{\circ}_{\theta_T,t,T}(j)\nabla a^{\circ}_{\theta^{***},t,T}(0)}{a^{\circ}_{\theta,t,T}(0)a^{\circ}_{\theta_T,t,T}(0)}X_{t-j,T}\right\}.
\end{aligned}\tag{6.532}$$

Here θ^*, θ^{**} and θ^{***} are points on the segment between θ and $\theta_T=\theta+h/\sqrt{T}$. From (6.530), (6.531) and (6.532) we can see that $\Lambda_T(p_1,p_2)-F_T(p_1,p_2)=o_{\mathcal{P}}(1)$ under Π_1. Similarly, we have $\Lambda_T(p_1,p_2)-F_T(p_1,p_2)=o_{\mathcal{P}}(1)$ under Π_2. Therefore, $F_T(p_1,p_2)$ has the same limit distribution of $\Lambda_T(p_1,p_2)$ under both Π_1 and Π_2. □

Exercises

6.1 Let the VARMA(p, q) process $\{\mathbf{X}_t\}$ defined in (6.12) satisfy Assumption 6.2. Show that the spectral density function matrix of $\{\mathbf{X}_t\}$ is given by (6.13).

6.2 For an autoregressive model (6.33) with $p = 1$, give the log-likelihood $l_n(\boldsymbol{\theta})$ of (6.37) in an explicit form and simplify the likelihood equation

$$\frac{\partial l_n(\boldsymbol{\theta})}{\partial b_1} = 0, \quad \frac{\partial l_n(\boldsymbol{\theta})}{\partial \sigma^2} = 0 \tag{6.533}$$

as simple as possible.

6.3 Show that the matrix Γ_p defined in (6.44) is positive definite.

6.4 Verify the assertion (ii) of Theorem 6.3.

6.5 Verify (6.54).

6.6 Show that AIC for AR(p) model (6.121) is given by

$$AIC(p) = n \log \hat{\sigma}^2_{QGML}(p) + 2p. \tag{6.534}$$

6.7 Show that the equation (6.151) holds.

6.8 Show that the spectral density estimator $\hat{f}_n(\lambda)$ given in (6.161) can be written as in the form (6.165).

6.9 Calculate the values of q, κ_q and $\int_{-1}^{1} w(x)^2 dx$ for the Akaike window of Example 6.18.

6.10 Let $\{X_t\}$ be a Gaussian stationary process satisfying the assumption in Theorem 6.10. Then, show that

$$E\left[\log\left\{I_n(\lambda)\right\}\right] \not\to E\left\{\log f(\lambda)\right\}, \quad (n \to \infty) \tag{6.535}$$

and

$$E\int_{-\pi}^{\pi} \log\left\{I_n(\lambda)\right\} d\lambda \not\to \int_{-\pi}^{\pi} \log f(\lambda) d\lambda, \quad (n \to \infty). \tag{6.536}$$

6.11 Let $\{X_t\}$ be a stationary process with mean zero and spectral density function $g(\lambda) = (1/2\pi)|1 - 0.3\, e^{i\lambda}|^2$. Compute the prediction error $PE(g, f)$ ((6.199)) of predictor (6.198) constructed on the basis of incorrectly conjectured spectral density

$$f(\lambda) = \frac{1}{2\pi}\left|1 - (0.3 + \theta)e^{i\lambda} + 0.3\,\theta\, e^{i2\lambda}\right|^2, \quad (|\theta| < 1) \tag{6.537}$$

and confirm $\lim_{|\theta| \nearrow 1} PE(g, f) = \infty$.

6.12 Verify conclusions of (i) and (ii) in Example 6.20.

6.13 Let the regressor function of Theorem 6.14 have trend (i) or (ii) in Example 6.20. Then, show that $\hat{\boldsymbol{\beta}}_{LS}$ is asymptotically efficient.

6.14 Confirm assertions (6.236) and (6.237).

6.15 Let $\mathbf{X}_n = (X_1, \ldots, X_n)'$ be generated from the AR(1) process

$$X_t = \theta X_{t-1} + u_t, \quad (|\theta| < 1), \tag{6.538}$$

where $\{u_t\} \sim i.i.d.\ N(0,1)$. Then, the spectral density function of $\{X_t\}$ is $f_\theta(\lambda) = (1/2\pi)|1 - \theta e^{i\lambda}|^{-2}$. Here we consider the discriminant problem described by the following categories

$$\Pi_1 : f(\lambda) = f_\theta(\lambda), \quad \Pi_2 : f(\lambda) = f_\mu(\lambda), \quad (\mu \neq \theta). \tag{6.539}$$

(i) For $n = 512$, give the discriminant statistics $I(f : g)$ of (6.426) in an explicit form and simplify it as much as possible.

(ii) Let μ and θ be contiguous to each other with the relation of $\mu = \theta + 1/\sqrt{512}$ and generate 100 times iterated samples of $\mathbf{X}_{512}$ for $\theta = 0.3, 0.6, 0.9$, respectively. Then, calculate the ratio of $I(f : g) > 0$ in these 100 times experiments for each $\theta = 0.3, 0.6, 0.9$.

CHAPTER 7

Introduction to Statistical Financial Engineering

Financial engineering is the construction of various financial positions to manage financial risks due to changes in the price of assets such as stocks, bonds, etc. Behavior of assets is modelled by stochastic processes, whose estimation problems were discussed in Chapter 6. This book uses the terminology "statistical financial engineering" because we will develop the arguments based on statistical inference of stochastic processes, i.e., time series analysis.

Section 7.1 describes an introduction to option pricing theory, and provides a Monte Carlo simulation method for option pricing by use of CHARN. The classical theory by Black and Scholes assumes the stock prices follow a geometric Brownian motion with a constant volatility. However, empirical studies show that the return processes are often dependent and non-Gaussian. In view of this, Section 7.2 discusses higher order option pricing by using Edgeworth expansion when the return processes are non-Gaussian linear processes.

In the theory of portfolio analysis, optimal portfolios are determined by the mean and variance of the portfolio return. Estimators of the optimal portfolios were proposed as functions of the sample mean and the sample variance assuming that the returns are i.i.d. Because this setting is not natural, Section 7.3 addresses the problem of estimating the optimal portfolios when the portfolio returns are non-Gaussian and dependent. Then asymptotically efficient estimators are constructed.

Section 7.4 discusses some problems of existing methods for calculating the value-at-risk (VaR) in an ARCH setting. It should be noted that the commonly used approaches often confuse the true innovations with the empirical residuals, i.e., estimation errors for unknown ARCH parameters are ignored. We adjust this by using the asymptotics of the residual empirical process, and propose a feasible VaR which keeps the assets away from a specified risk with high confidence level.

7.1 Option Pricing Theory

Assets are defined as contracts that give the right to receive or obligation to

provide monetary cash flows (e.g., bank account, bond, stock). Bank accounts and bonds are called *risk-free assets*, and stocks are called *risky assets*.

Let $\mathbf{S}_t = (S_t^0, S_t^1, \ldots, S_t^d)'$ be the price of $(d+1)$ assets at time t ($t = 0, 1, \ldots, T$). Usually S_t^0 is taken to be a risk-free asset, and $(S_t^1, \ldots, S_t^d)'$ is taken to be a collection of d risky assets. However, in what follows, we assume that $S_t^0, \ldots, S_t^d$ may be any assets, i.e., S_t^0 may be risky etc.

Now we give the mathematical description. Let $(\Omega, \mathcal{A}, P)$ be a probability space, and let $\{\mathcal{A}_t\}$ be a family of sub σ-fields of $\mathcal{A}$ satisfying $\mathcal{A}_s \subset \mathcal{A}_r$ $(s \leq r)$. We assume that $\{\mathbf{S}_t\}$ is a stochastic process on $(\Omega, \mathcal{A}, P)$, and that each $\mathbf{S}_t$ is $\mathcal{A}_t$-measurable.

If one invests the assets S_t^i with fraction weights w_{it} $(i = 0, 1, \ldots, d)$ satisfying $\sum_{i=0}^d w_{it} = 1$, the fraction vector $\mathbf{w}_t = (w_{0t}, w_{1t}, \ldots, w_{dt})'$ is called the *portfolio*. Here $\mathbf{w}_t$ is assumed to be $\mathcal{A}_{t-1}$-measurable. Then the total investment at time t is

$$V_t(\mathbf{w}_t) = \sum_{i=0}^d w_{it} S_t^i, \tag{7.1}$$

which is called the *value process*.

Definition 7.1 *If a portfolio* $(w_{0t}, \ldots, w_{dt})'$ *satisfies*

$$\sum_{t=0}^d w_{i,t-1} S_t^i = \sum_{i=0}^d w_{i,t} S_t^i, \tag{7.2}$$

then it is said to be self-financing.

"Self-financing" means that after the initial investment no further capital is either invested or withdrawn.

Definition 7.2 *A collection of assets* $\mathbf{S}_t = (S_t^0, S_t^1, \ldots, S_t^d)'$ *is said to admit an* arbitrage opportunity *if there exists a self-financing portfolio* $\mathbf{w}_t$ *such that*

$$\begin{aligned} &V_0(\mathbf{w}_0) = 0, \quad V_T(\mathbf{w}_T) \geq 0, \quad (P\text{-}a.s.), \\ &P\{V_T(\mathbf{w}_T) > 0\} > 0. \end{aligned} \tag{7.3}$$

If there is no self-financing portfolio for which (7.3) *holds, then the collection of assets* $\mathbf{S}_t$ *is said to be* arbitrage-free.

"Arbitrage-free" assumes the impossibility of achieving a sure, strictly positive gain with a zero initial endowment. Thus it implies that we cannot get a "free lunch" without risk.

If $V_T(\mathbf{w}_T)$ is the value process for a self-financing portfolio $\mathbf{w}_t$, then $V_t(\mathbf{w}_t) =$

$V_t(\mathbf{w}_{t-1})$, hence, for any $m < T$, we have

$$\begin{aligned} V_T(\mathbf{w}_T) &= V_m(\mathbf{w}_m) + \sum_{t=m+1}^{T} [V_t(\mathbf{w}_{t-1}) - V_{t-1}(\mathbf{w}_{t-1})] \\ &= V_m(\mathbf{w}_m) + \sum_{t=m+1}^{T} \sum_{i=0}^{d} w_{i,(t-1)}(S_t^i - S_{t-1}^i). \end{aligned} \tag{7.4}$$

Let Q be the probability distribution of $\mathcal{S}_T \equiv \{\mathbf{S}_t : t = 0, 1, \ldots, T\}$. We assume that there exists another probability distribution Q^* of $\mathcal{S}_T$ which satisfies

(i) Q^* is equivalent to Q (i.e., $Q(A) = 0 \Leftrightarrow Q^*(A) = 0$ for any $A \in \mathcal{A}$),

(ii) $\{\mathbf{S}_t\}$ is a martingale with respect to Q^*, i.e.,

$$E^*\{\mathbf{S}_t|\mathcal{A}_{t-1}\} = \mathbf{S}_{t-1} \quad \text{a.e.},$$

where $E^*\{\cdot\}$ is the expectation with respect to Q^*.

From (7.4) and Exercises 7.1 and 7.2 it follows that

$$\begin{aligned} &E^*\{V_T(\mathbf{w}_T)|\mathcal{A}_m\} \\ &= V_m(\mathbf{w}_m) + \sum_{t=m+1}^{T} \sum_{i=0}^{d} E^*\{w_{i,(t-1)}(S_t^i - S_{t-1}^i)|\mathcal{A}_m\} \\ &= V_m(\mathbf{w}_m) + \sum_{t=m+1}^{T} \sum_{i=0}^{d} E^*[E^*\{w_{i,(t-1)}(S_t^i - S_{t-1}^i)|\mathcal{A}_{t-1}\}|\mathcal{A}_m] \\ &= V_m(\mathbf{w}_m) + \sum_{t=m+1}^{T} \sum_{i=0}^{d} E^*[w_{i,(t-1)}E^*\{(S_t^i - S_{t-1}^i)|\mathcal{A}_{t-1}\}|\mathcal{A}_m] \quad \text{a.e.} \end{aligned} \tag{7.5}$$

Since $\{\mathbf{S}_t\}$ is a martingale with respect to Q^*, we have $E^*\{(S_t^i - S_{t-1}^i)|\mathcal{A}_{t-1}\} = 0$ a.e., hence, for any $m < T$,

$$E^*\{V_T(\mathbf{w}_T)|\mathcal{A}_m\} = V_m(\mathbf{w}_m) \quad Q^*\text{-a.e.}, \tag{7.6}$$

which implies that $V_m(\mathbf{w}_m)$ is a martingale with respect to Q^*. Next we will show that $\{\mathbf{S}_t\}$ is arbitrage-free. If we assume that $\{\mathbf{S}_t\}$ is not arbitrage-free, then (7.3) holds. Hence

$$\begin{aligned} &Q(V_0(\mathbf{w}_0) = 0) = 1, \\ &Q(V_T(\mathbf{w}_T) \geq 0) = 1, \\ &Q(V_T(\mathbf{w}_T) > 0) > 0. \end{aligned} \tag{7.7}$$

Since Q^* is equivalent to Q, (7.7) implies

$$\begin{aligned} &Q^*(V_0(\mathbf{w}_0) = 0) = 1, \\ &Q^*(V_T(\mathbf{w}_T) \geq 0) = 1, \\ &Q^*(V_T(\mathbf{w}_T) > 0) > 0. \end{aligned} \tag{7.8}$$

If we set $m = 0$ in (7.6), it holds that

$$E^*\{V_T(\mathbf{w}_T)|\mathcal{A}_0\} = V_0(\mathbf{w}_0) \quad Q^*\text{-a.e.} \tag{7.9}$$

However, the relation (7.8) implies that (7.9) can not hold. Thus $\{\mathbf{S}_t\}$ must be arbitrage-free. Summarizing the above we obtain,

Theorem 7.1 *If a collection of assets* $\mathbf{S}_t = (S_t^0, S_t^1, \ldots, S_t^d)'$ $(t = 0, 1, \ldots .T)$ *is a martingale with respect to a probability distribution* Q^* *which is equivalent to the probability distribution* Q *of* $\{\mathbf{S}_0, \mathbf{S}_1, \ldots, \mathbf{S}_T\}$*, then the collection of assets* $\mathbf{S}_t$ *is arbitrage-free.*

Although the statement of Theorem 7.1 is for a general stochastic process $\{\mathbf{S}_t\}$, if the components include some risk-free assets, for example, $S_t^0 = \exp(rt)$, (a bank deposit with compound interest rate r), then $E^*\{S_t^0|\mathcal{A}_{t-1}\} = S_t^0 \neq S_{t-1}^0$ for any Q^*. Hence, S_t^0 in this case cannot be a martingale. To accommodate this case to Theorem 7.1 we consider the following standardized quantities:

$$\begin{aligned} \tilde{S}_t^0 &\equiv 1 \\ \tilde{S}_t^i &\equiv \frac{S_t^i}{S_t^0} \quad (i = 1, 2, \ldots, d), \\ \tilde{V}_T(\mathbf{w}_T) &\equiv \frac{V_T(\mathbf{w}_T)}{S_t^0} \end{aligned} \tag{7.10}$$

when $S_t^0 > 0$ a.e., for all t. Then, $\tilde{S}_t^0$ is trivially a martingale. Similarly as in Theorem 7.1 we can prove the following theorem (Exercise 7.3).

Theorem 7.1′ *If* $\tilde{\mathbf{S}}_t \equiv (\tilde{S}_t^1, \tilde{S}_t^2, \ldots, \tilde{S}_t^d)$ *is a martingale with respect to an equivalent probability distribution* Q^**, then* $\mathbf{S}_t = (S_t^0, S_t^1, \ldots, S_t^d)'$ *is arbitrage-free.*

Next we give some fundamental concepts of finance.

Definition 7.3 *(i) A* contingent claim *is a non-negative random variable* X *representing a payoff at some future time* T*. We may regard it as a contract that an investor makes at time* $t < T$ *(e.g., option).*

(ii) For a contingent claim X*, if there exists a self-financing portfolio* $\mathbf{w}_T$ *such that*

$$X = V_T(\mathbf{w}_T), \tag{7.11}$$

then $\mathbf{w}_T$ *is called a* replicating portfolio *of* X*.*

(iii) For any contingent claim, if there exists the replicating portfolio, then the market of financial assets is said to be complete*. In the definition 3.3, we introduced the terminology "complete" for statistics. Henceforth we use this word in a different meaning i.e., in the sense of* (iii)*.*

Remark 7.1 *(i) Q^* in Theorems 7.1 and 7.1′ is called the* equivalent martingale measure*, and the reverse statement "if the collection of assets is arbitrage-free, there exists an equivalent martingale measure" holds.*

(ii) In the case of arbitrage-free, the statement "the market is complete if and only if there exists a unique equivalent martingale measure" holds.

Options are one example of many derivatives on the market. A *derivative* is a financial instrument whose value is derived from the value of some underlying instrument such as stock price, interest rate, or foreign exchange rate. A *call option* gives one the right to buy the underlying asset by a certain date, called the *maturity*, for a certain price, called the *strike price*, while a *put option* gives one the right to sell the underlying asset by a maturity for a strike price. *American options* can be exercised at any time up to the maturity, but *European options* can be exercised only at their maturity. European options are easier to price than American options since one does not need to consider the possibility of early exercise. Most options traded on exchanges are American.

Now let us explain these fundamental concepts concretely. Let S_t be the price of an underlying asset at time t. A *European call* with maturity T and strike price K on this asset is theoretically equivalent to a contingent claim

$$C_T = \max(S_T - K, 0). \tag{7.12}$$

If the price S_T is greater than the strike price K at maturity T, we will buy the asset for the strike price K and sell it immediately in the market, whence the gain is $S_T - K$. On the other hand, if $S_T \leq K$, we will not exercise the right, whence the gain is 0.

A *European put* with maturity T and strike price K is theoretically equivalent to a contingent claim

$$P_T \equiv \max(K - S_T, 0). \tag{7.13}$$

Contrary to a European call, it gives the gain $K - S_T$ if $S_T < K$ and 0 otherwise. Although the payoffs of options, e.g., (7.12) and (7.13), are always non-negative, we have to pay the initial cost of purchasing the options, which is called the *premium*. Therefore, in the call of (7.12), the essential gain is C_T-(premium).

European options can only be exercised at the maturity date T, but American options can be exercised at any time before or at the maturity date T. An *American call option* is equivalent to a contingent claim

$$C_{n^*} \equiv \max(S_{n^*} - K, 0) \tag{7.14}$$

when n^* is not determined in advance but depends on the path of the underlying process. More precisely, n^* is a random variable such that the event $\{n^* = n\}$ belongs to the σ-algebra $\mathcal{A}_n^s$ generated by $\{S_0, S_1, \ldots, S_n\}$, $n \leq T$. Similarly we can define an *American put option*. Although there are various

options, as an example we just mention an *Asian call option* whose payoff function is given by

$$\max(\bar{S}_T - K, 0) \tag{7.15}$$

where $\bar{S}_T = T^{-1} \sum_{t=1}^{T} S_t$.

Let us consider the problem of pricing options. For notation, the present time is denoted by t and the maturity date of options is denoted by T. For a contingent claim X, assume that there exists a replicating portfolio $\mathbf{w}_T$ of X constructed on an asset process $\mathbf{S}_t$, i.e.,

$$X = V_T(\mathbf{w}_T). \tag{7.16}$$

Suppose that $\mathbf{S}_t$ is arbitrage-free. Then recalling (7.5), Theorem 7.1 and Remark 7.1, and converting values of future payments to their present values by risk-free interest rate r we obtain

$$E^*\{e^{-r(T-t)}X|\mathcal{A}_t\} = \tilde{V}_t(\mathbf{w}_t) = V_t(\mathbf{w}_t), \tag{7.17}$$

where $E^*\{\cdot\}$ is the expectation with respect to an equivalent martingale measure Q^*. Therefore, if we let X be an option, and if there exists a replicating portfolio of it, then (7.17) implies that the initial capital of the portfolio is equal to

$$E^*\{e^{-r(T-t)}X|\mathcal{A}_t\}. \tag{7.18}$$

Hence, the reasonable price of the option X should be (7.18). Concrete valuation of (7.18) has been often done for the geometric Brownian motion

$$S_t = S_0 \exp\{\mu t + \sigma \int_0^t dW_u\}, \quad (t \in [0, T]) \tag{7.19}$$

where $\{W_t\}$ is the Wiener process (see Definition A.2 in the Appendix). Dividing $[0, T]$ into N subintervals with length h, Kariya and Liu (2003) introduced

$$S_n = S_0 \exp\{\mu n h + \sigma \sum_{k=1}^{n} u_k \sqrt{h}\}, \quad (\{u_k\} \sim \text{i.i.d. } N(0,1)), \tag{7.20}$$

as a discretized version of (7.19) for $n = 0, 1, \ldots, N$ and $Nh = T$. Then they evaluated the price of a European call option (7.12) at time $t = nh$;

$$C = \exp\{-r(T-t)\}E^*\{\max(S_N - K, 0)|\mathcal{A}_n\} \tag{7.21}$$

as follows. (7.20) is written as

$$S_n = S_{n-1} \exp(\mu h + \sigma\sqrt{h} u_n). \tag{7.22}$$

Further we can rewrite (7.22) as

$$\log S_n - \log S_{n-1} = \mu h + \sigma\sqrt{h} u_n \tag{7.23}$$

in terms of the log-return. This is a special case of ARCH + mean models or

CHARN models. From (7.22) it follows that

$$\frac{S_n}{\exp(rnh)} = \frac{S_{n-1}}{\exp\{r(n-1)h\}} \exp\{(\mu - r)h + \sigma\sqrt{h}u_n\}. \tag{7.24}$$

For the discounted process $S_n/\exp(rnh)$ to be a martingale with respect to an equivalent martingale measure Q^*, it should hold that

$$E^*[\exp\{(\mu - r)h + \sigma\sqrt{h}u_n\}|\mathcal{A}_{n-1}] = 1 \quad \text{a.e.} \tag{7.25}$$

For this, letting the distribution of u_n under Q^* be $N(m, 1)$, we observe that the left-hand side of (7.25) is

$$\exp\{(\mu - r)h\} \times \exp\{m\sigma\sqrt{h} + \frac{1}{2}\sigma^2 h\}. \tag{7.26}$$

Therefore, if we take

$$m = -\frac{1}{\sigma\sqrt{h}}\{(\mu - r)h + \frac{\sigma^2 h}{2}\}, \tag{7.27}$$

then (7.25) holds. Here, if we set $u_n^* = u_n - m$, then $u_n^* \sim N(0, 1)$ under Q^*, hence, (7.22) can be expressed as

$$S_n = S_{n-1} \exp\{(rh - \frac{\sigma^2 h}{2}) + \sigma\sqrt{h}u_n^*\}, \tag{7.28}$$

under Q^*. From (7.28) it follows that

$$\begin{aligned} S_N =& S_n \exp\{(r - \frac{\sigma^2}{2})(N - n)h + \sigma\sqrt{h} \sum_{j=n+1}^{N} u_j^*\} \\ =& S_n \exp\{A + BZ\}, \end{aligned} \tag{7.29}$$

where

$$\begin{aligned} A &= (r - \frac{\sigma^2}{2})(N - n)h, \\ B &= \sqrt{(N - n)h}, \\ Z &= \frac{1}{\sqrt{N - n}} \sum_{j=n+1}^{N} u_j^* \sim N(0, 1) \quad \text{under } Q^*. \end{aligned}$$

From (7.29) we can evaluate the call option C defined by (7.21) as follows (Exercise 7.3).

Black-Scholes formula (Black and Scholes (1973)).

$$C = S_n\Phi(d_t) - \exp\{-(T - t)\}K\Phi(d_t - \sigma\sqrt{T - t}), \tag{7.30}$$

where $d_t = \{\log\frac{S_n}{K} + (r + \frac{\sigma^2}{2})(T - t)\}/(\sigma\sqrt{T - t})$, $T = Nh$, $t = nh$ and $\Phi(\cdot)$ is the distribution function of $N(0, 1)$.

Until now μ and σ have been assumed to be constants, in what follows, we suppose that they may depend on $S_{n-1}, \ldots, S_{n-\max(p,q)}$, i.e., $\mu_{n-1} =$

$\mu(S_{n-1}, \ldots, S_{n-p})$ and $\sigma_{n-1} = \sigma(S_{n-1}, \ldots, S_{n-q})$. Hence the model is the following CHARN model;

$$S_n = S_{n-1} \exp\{\mu_{n-1} \cdot h + \sigma_{n-1}\sqrt{h}u_n\}. \tag{7.31}$$

It is possible for us to evaluate the European call option (7.21) based on (7.31) numerically. Similarly as in (7.25), under Q^* which makes the process S_n/e^{rnh} a martingale, the process concerned is expressed as

$$S_n = S_{n-1} \exp\{(rh - \frac{\sigma_{n-1}^2 h}{2}) + \sigma_{n-1}\sqrt{h}u_n^*\} \tag{7.32}$$

where $u_n^* \sim N(0,1)$ given $\mathcal{A}_{n-1}$. Therefore,

$$S_N = S_n \exp[\sum_{j=n+1}^{N} \{(rh - \frac{\sigma_{j-1}^2 h}{2}) + \sigma_{j-1}\sqrt{h}u_j^*\}]. \tag{7.33}$$

In this case it is difficult to express (7.21) in explicit analytical form. Thus we evaluate (7.21) numerically.

In fact, because the value of σ_n is specified from $S_n, \ldots, S_{n-q+1}$, it is known at the present time n. If we generate a random number $u_{n+1}^* \sim N(0,1)$, then we get a simulated value of S_{n+1} from the formula (7.32), hence the value of σ_{n+1} is specified. If we generate a random number $u_{n+2}^* \sim N(0,1)$, the value of S_{n+2} is specified from (7.32). Repeating this procedure, we get a simulated value $\hat{z}_1$ of the superfix of the exponential in (7.33). Similarly, repeating the simulation L times we obtain simulated values $\hat{z}_2, \ldots, \hat{z}_L$. Finally, the Monte Carlo valuation of (7.21) based on the CHARN model (7.31) is

$$\hat{C}_{\text{CHARN}} \equiv \exp\{-r(T-t)\}\frac{1}{L}\sum_{l=1}^{L} \max\{S_n \exp(\hat{z}_l) - K, 0\}. \tag{7.34}$$

7.2 Higher Order Asymptotic Option Valuation for Non-Gaussian Dependent Returns

Black and Scholes (1973) provided the foundation of modern option pricing theory. Despite its usefulness, however, the Black and Scholes theory entails some inconsistencies. It is well known that the model frequently misprices deep in-the-money and deep out-of-the-money options. This result is generally attributed to the unrealistic assumptions used to derive the model. In particular, the Black and Scholes model assumes that stock prices follow a geometric Brownian motion with a constant volatility under an equivalent martingale measure.

In order to avoid this drawback, Jarrow and Rudd (1982) proposed a semiparametric option pricing model to account for non-normal skewness and kurtosis in stock returns. This approach aims to approximate the risk-neutral density by a statistical series expansion. Jarrow and Rudd (1982) approximated

the density of the state price by an Edgeworth series expansion involving the log-normal density. Corrado and Su (1996a) implemented Jarrow and Rudd's formula to price options. Corrado and Su (1996b, 1997) considered Gram-Charlier expansions for the stock log return rather than the stock price itself. Rubinstein (1998) used the Edgeworth expansion for the stock log return. Jurczenko et al. (2002) compared these different multi-moment approximate option pricing models. Also they investigated in particular the conditions that ensure the martingale restriction.

As in Kariya (1993) and Kariya and Liu (2003), the time series structure of return series does not always admit a measure which makes the discounted process a martingale. Hence, we are not able to develop an arbitrage pricing theory by forming an equivalent portfolio. In such a case, we often regard the expected value of the present value of a contingent claim as a proxy for pricing with the help of a risk neutrality argument. In view of this, Kariya (1993) considered pricing problems with no martingale property and approximated the density of the state price by the Gram-Charlier expansion for the stock log return.

In this section, we consider option pricing problems by using Kariya's approach. First, we derive the Edgeworth expansion for the stock log return via extracting dynamics structure of time series. Using this result, we investigate influences of the non-Gaussianity and the dependency of log return processes for option pricing. Numerical studies illuminate some interesting features of the influences. Next, we give option prices based on the risk neutrality argument. Also, we discuss a consistent estimator of the quantities in our results.

Let $\{S_t; t \geq 0\}$ be the price process of an underlying security at trading time t. The j-th period log return X_j is defined as

$$\log S_{T_0+j\Delta} - \log S_{T_0+(j-1)\Delta} = \Delta\mu + \Delta^{1/2} X_j, \quad j = 1, 2, \ldots, N, \tag{7.35}$$

where T_0 is present time, $N = \tau/\Delta$ is the number of unit time intervals of length Δ during a period $\tau = T - T_0$ and T is the maturity date. Then the terminal price S_T of the underlying security is given by

$$S_T = S_{T_0} \exp\left\{ \tau\mu + \left(\frac{\tau}{N}\right)^{1/2} \sum_{j=1}^{N} X_j \right\}. \tag{7.36}$$

Remark 7.2 *In the Black and Scholes option theory the price process is assumed to be a geometric Brownian motion*

$$S_T = S_{T_0} \exp\left(\tau\mu + \sigma W_\tau\right), \tag{7.37}$$

where the process $\{W_t; t \in \mathbf{R}\}$ *is a Wiener process with drift* 0 *and variance* t. *From* (7.37), *the log return at a discretized time point can be written as*

$$\log S_{t+j\Delta} - \log S_{t+(j-1)\Delta} = \Delta\mu + \Delta^{1/2}\sigma\nu_j, \quad \nu_j \sim \textit{i.i.d. } N(0,1). \tag{7.38}$$

The expression of (7.35) *is motivated from* (7.38).

First, we derive an analytical expression for the density function of S_T. Since from (7.36) the distribution of S_T depends on that of $Z_N = N^{-1/2}\sum_{j=1}^{N} X_j$, we consider the Edgeworth expansion of the density function of Z_N. If we assume that X_j's are independently and identically distributed random variables with mean zero and finite variance, it is easy to give the Edgeworth expansion for Z_N (the classical Edgeworth expansion).

However, a lot of financial empirical studies show that X_j's are not independent. Thus we suppose that $\{X_j\}$ is a dependent process which satisfies the following assumption.

Assumption 7.1 *(i)* $\{X_t; t \in \mathbf{Z}\}$ *is fourth-order stationary in the sense that*

$$\begin{aligned}
&\mathrm{E}(X_t) = 0,\\
&\mathrm{cum}(X_t, X_{t+u}) = c_{X,2}(u),\\
&\mathrm{cum}(X_t, X_{t+u_1}, X_{t+u_2}) = c_{X,3}(u_1, u_2),\\
&\mathrm{cum}(X_t, X_{t+u_1}, X_{t+u_2}, X_{t+u_3}) = c_{X,4}(u_1, u_2, u_3),
\end{aligned}$$

for any $t, u_1, \ldots, u_3 \in \mathbf{Z}$.

(ii) The cumulants $c_{X,k}(u_1, \ldots, u_{k-1})$, $k = 2, 3, 4$, *satisfy*

$$\sum_{u_1,\ldots,u_{k-1}=-\infty}^{\infty} \left(1 + |u_j|^{2-k/2}\right) |c_{X,k}(u_1, \ldots, u_{k-1})| < \infty$$

for $j = 1, \ldots, k-1$.

(iii) J-th order $(J \geq 5)$ *cumulants of* Z_N *are all* $O(N^{-J/2+1})$.

Under Assumption 7.1 (ii), $\{X_t; t \in \mathbf{Z}\}$ has the k-th order cumulant spectral density. Let $f_{X,k}$ be the k-th order cumulant spectral density evaluated at frequency $\mathbf{0}$

$$f_{X,k} = (2\pi)^{-(k-1)} \sum_{u_1,\ldots,u_{k-1}=-\infty}^{\infty} c_{X,k}(u_1, \ldots, u_{k-1})$$

for $k = 2, 3, 4$.

First, we state the following result.

Theorem 7.2 *Suppose that Assumption 7.1* (i)-(iii) *hold. The third-order Edgeworth expansion of the density function of* $Z = (2\pi f_{X,2})^{-1/2} Z_N$ *is given by*

$$\begin{aligned}
g(z) = \phi(z)\Bigg\{1 &+ \frac{(2\pi)^{1/2}}{6} N^{-1/2} \frac{f_{X,3}}{(f_{X,2})^{3/2}} H_3(z) - \frac{1}{4\pi} N^{-1} \frac{f'_{X,2}}{f_{X,2}} H_2(z)\\
&+ \frac{\pi}{12} N^{-1} \frac{f_{X,4}}{(f_{X,2})^2} H_4(z) + \frac{\pi}{36} N^{-1} \frac{(f_{X,3})^2}{(f_{X,2})^3} H_6(z)\Bigg\} + o(N^{-1}),
\end{aligned} \tag{7.39}$$

where $\phi(\cdot)$ is the standard normal density function, $H_k(\cdot)$ is the k-th order Hermite polynomial and

$$f'_{X,2} = \sum_{u=-\infty}^{\infty} |u| c_{X,2}(u).$$

PROOF First, we evaluate the asymptotic cumulants of Z_N. From Assumption 7.1 (i) and (ii), $E(Z_N) = 0$,

$$\begin{aligned}
\mathrm{cum}(Z_N, Z_N) &= N^{-1} \sum_{j=-(N-1)}^{N-1} (N - |j|) c_{X,2}(j) \\
&= 2\pi f_{X,2} - N^{-1} f'_{X,2} + o(N^{-1}), \\
\mathrm{cum}(Z_N, Z_N, Z_N) &= N^{-3/2} \sum_{j_1, j_2 = -(N-1)}^{N-1} (N - S_{j_1 j_2}) c_{X,3}(j_1, j_2) \\
&= N^{-1/2} (2\pi)^2 f_{X,3} + o(N^{-1}),
\end{aligned}$$

where

$$S_{j_1 j_2} = \begin{cases} \max(|j_1|, |j_2|) & \text{if } \mathrm{sign}(j_1) = \mathrm{sign}(j_2), \\ \min(|j_1| + |j_2|, N) & \text{if } \mathrm{sign}(j_1) = -\mathrm{sign}(j_2), \end{cases}$$

and

$$\begin{aligned}
\mathrm{cum}(Z_N, Z_N, Z_N, Z_N) &= N^{-2} \sum_{j_1, j_2, j_3 = -(N-1)}^{N-1} (N - S_{j_1 j_2 j_3}) c_{X,4}(j_1, j_2, j_3) \\
&= N^{-1} (2\pi)^3 f_{X,4} + o(N^{-1}),
\end{aligned}$$

where

$$S_{j_1 j_2 j_3} = \begin{cases} \max(|j_1|, |j_2|, |j_3|) & \text{if } \mathrm{sign}(j_1) = \mathrm{sign}(j_2) = \mathrm{sign}(j_3), \\ \min\{\max(|j_1|, |j_2|) + |j_3|, N\} & \text{if } \mathrm{sign}(j_1) = \mathrm{sign}(j_2) = -\mathrm{sign}(j_3). \end{cases}$$

Applying the general formula for the Edgeworth expansion (e.g., Taniguchi and Kakizawa (2000, pp.168–170)), we obtain (7.39). □

Many authors have proposed to use different statistical series expansions to price options (see Jarrow and Rudd (1982), Corrado and Su (1996a, 1996b, 1997), Rubinstein (1998), and Kariya (1993)). Here we give the Edgeworth expansion for the stock log return in powers of $N^{-1/2}$.

A European call option can be viewed as a security which pays at time T its holder the amount

$$X_T^* = \max(S_T - K, 0),$$

where K is the exercise or strike price. As in Kariya (1993), we price X_T^* by

its discounted expected value;

$$C = \exp(-r\tau)E_{T_0}(X_T^*), \tag{7.40}$$

where r is the interest rate which is regarded as a constant for the remaining period τ and $E_{T_0}(\cdot)$ is evaluated at T_0. Evaluate (7.40) based on the density in (7.39). Then writing

$$\begin{aligned} d_1 &= (\log S_{T_0}/K + \tau\mu + 2\pi\tau f_{X,2})/(2\pi\tau f_{X,2})^{1/2}, \\ d_2 &= d_1 - (2\pi\tau f_{X,2})^{1/2}, \end{aligned}$$

we obtain the following theorem.

Theorem 7.3 *Let* $a_1 = \exp(-r\tau)$ *and* $a_2 = \exp(\tau\mu + \pi\tau f_{X,2})$. *Then*

$$\begin{aligned} C = G_0 &+ \frac{(2\pi)^{1/2}}{6}N^{-1/2}\frac{f_{X,3}}{(f_{X,2})^{3/2}}G_3 - \frac{1}{4\pi}N^{-1}\frac{f'_{X,2}}{f_{X,2}}G_2 \\ &+ \frac{\pi}{12}N^{-1}\frac{f_{X,4}}{(f_{X,2})^2}G_4 + \frac{\pi}{36}N^{-1}\frac{(f_{X,3})^2}{(f_{X,2})^3}G_6 + o(N^{-1}), \end{aligned} \tag{7.41}$$

where

$$\begin{aligned} G_0 &= a_1\{a_2 S_{T_0}\Phi(d_1) - K\Phi(d_2)\}, \\ G_k &= a_1 a_2 S_{T_0}\left\{\sum_{j=1}^{k-1}(2\pi\tau f_{X,2})^{j/2}H_{k-j-1}(-d_2)\phi(d_1) + (2\pi\tau f_{X,2})^{k/2}\Phi(d_1)\right\}, \end{aligned}$$

for $k = 2, 3, 4, 6$, *where* $\Phi(\cdot)$ *is the standard normal distribution function.*

PROOF From Theorem 7.2 and (7.40),

$$C = e^{-r\tau}\int_{-d_2}^{\infty}\left[S_{T_0}\exp\left\{\mu\tau + (2\pi\tau f_{X,2})^{1/2}z\right\} - K\right]g(z)dz. \tag{7.42}$$

Integrating by parts and using the following equality

$$\exp\{-(2\pi\tau f_{X,2})^{1/2}d_2\}\phi(-d_2) = \exp(\pi\tau f_{X,2})\phi(d_1),$$

yield

$$\begin{aligned} &\int_{-d_2}^{\infty}\left[S_{T_0}\exp\left\{\mu\tau + (2\pi\tau f_{X,2})^{1/2}z\right\} - K\right]H_k(z)\phi(z)dz \\ &\quad = a_2 S_{T_0}\left\{\sum_{j=1}^{k-1}(2\pi\tau f_{X,2})^{j/2}H_{k-j-1}(-d_2)\phi(d_1)\right. \\ &\qquad\qquad \left. + (2\pi\tau f_{X,2})^{k/2}\Phi(d_1)\right\} \end{aligned} \tag{7.43}$$

for $k = 2, 3, 4, 6$. Inserting (7.43) in (7.42), we obtain (7.41). □

From (7.41) it is seen that the asymptotic expansion of the option price depends on $f_{X,2}$, $f'_{X,2}$, $f_{X,3}$ and $f_{X,4}$. Hence, we can see influences of the non-Gaussianity and the dependency of the log return processes for the higher order option valuation.

Corollary 7.1 *Write*

$$C = G_0 + N^{-1/2}C_{G,2} + N^{-1}C_{G,3} + N^{-1}C_{D,3} + o(N^{-1}),$$

where

$$C_{G,2} = \frac{(2\pi)^{1/2}}{6}\frac{f_{X,3}}{(f_{X,2})^{3/2}}G_3,$$
$$C_{G,3} = \frac{\pi}{12}\frac{f_{X,4}}{(f_{X,2})^2}G_4 + \frac{\pi}{36}\frac{(f_{X,3})^2}{(f_{X,2})^3}G_6,$$
$$C_{D,3} = -\frac{1}{4\pi}\frac{f'_{X,2}}{f_{X,2}}G_2.$$

If X_t's are mutually independent, then $C_{D,3} = 0$. If $\{X_t; t \in \mathbf{Z}\}$ is a Gaussian process, then $C_{G,2} = C_{G,3} = 0$.

PROOF If X_t's are mutually independent, then $f'_{X,2} = 0$. If $\{X_t; t \in \mathbf{Z}\}$ is a Gaussian process, then $f_{X,3} = f_{X,4} = 0$. Hence, Corollary 7.1 follows. □

Example 7.1 *Suppose that X_j, $j = 1, \ldots, N$, are independently and identically distributed random variables. Let $c_{X,k} = c_{X,k}(\mathbf{0})$, $k = 2, 3, 4$. Note that $f'_{X,2} = 0$ and $f_{X,k} = (2\pi)^{-(k-1)}c_{X,k}$, $k = 2, 3, 4$. The price of a European call option C_{IID} is given by*

$$C_{\mathrm{IID}} = G_0 + \frac{1}{6}N^{-1/2}\frac{c_{X,3}}{(c_{X,2})^{3/2}}G_3 + \frac{1}{24}N^{-1}\frac{c_{X,4}}{(c_{X,2})^2}G_4 + \frac{1}{72}N^{-1}\frac{(c_{X,3})^2}{(c_{X,2})^3}G_6 + o(N^{-1}),$$

where G_k, $k = 0, 3, 4, 6$, are defined in Theorem 7.2 with $f_{X,2} = (2\pi)^{-1}c_{X,2}$. If $\mu = r - c_{X,2}/2$, then $a_1a_2 = 1$ so that G_0 equals the Black and Scholes formula.

Example 7.2 *In Example 7.1, suppose that X_j, $j = 1, \ldots, N$, are distributed as a t-distribution with ν degrees of freedom. Then, for $\nu > 4$*

$$C_t = G_{t,0} + N^{-1}G_{t,3} + o(N^{-1}),$$

where

$$G_{t,0} = a_1\{a_2 S_{T_0}\Phi(d_1) - K\Phi(d_2)\},$$
$$a_2 = \exp\left\{\tau\mu + \frac{\tau\nu}{2(\nu-2)}\right\},$$
$$d_1 = \left(\log S_{T_0}/K + \tau\mu + \frac{\tau\nu}{\nu-2}\right) \Big/ \left(\frac{\tau\nu}{\nu-2}\right)^{1/2},$$
$$d_2 = d_1 - \left(\frac{\tau\nu}{\nu-2}\right)^{1/2},$$

and

$$G_{t,3} = \frac{a_1 a_2 S_{T_0}}{4(\nu-4)}\left\{\sum_{j=1}^{3}\left(\frac{\tau\nu}{\nu-2}\right)^{j/2} H_{3-j}(-d_2)\phi(d_1) + \left(\frac{\tau\nu}{\nu-2}\right)^{2}\Phi(d_1)\right\}.$$

In order to show influences of higher order terms, in Figure 7.1, we plotted $C_{t,1} = G_{t,0}$ (dotted line) and $C_{t,3} = G_{t,0} + N^{-1}G_{t,3}$ (solid line) of Example 7.2 with $S_{T_0} = K = 100$, $\tau = 30/365$, $N = 30$ ($\Delta = 1/365$), $r = \mu = 0.05$ and $4 < \nu < 9$. From this, we observe that $C_{t,3}$ diverges as $\nu \to 4$.

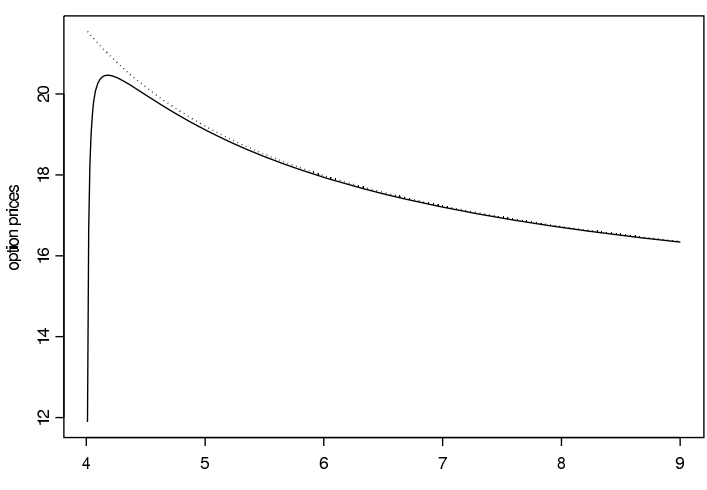


Figure 7.1 *For the t-distribution with ν degrees of freedom in Example 7.2, the approximations up to the first ($C_{t,1}$, dotted line) and third order ($C_{t,3}$, solid line) of the option price are plotted with $S_{T_0} = K = 100$, $\tau = 30/365$, $N = 30$, $r = \mu = 0.05$ and $4 < \nu < 9$.*

Example 7.3 *Let $\{X_t : t \in \mathbf{Z}\}$ be the ARCH(1) process*

$$X_t = h_t^{1/2}\eta_t \quad \text{and} \quad h_t = \psi_0 + \psi_1 X_{t-1}^2,$$

where $\psi_0 > 0$, $\psi_1 \geq 0$, $\{\eta_t : t \in \mathbf{Z}\}$ *is a sequence of independently and identically distributed random variables with*

$$E(\eta_t) = 0, \qquad E(\eta_t^2) = 1,$$
$$E(\eta_t^3) = 0, \qquad E(\eta_t^4) = m, \quad m > 1,$$

and η_t *is independent of* X_{t-s}, $s > 0$. *Then*

$$f_{X,2} = \frac{1}{2\pi}\frac{\psi_0}{1-\psi_1}, \qquad f_{X,3} = 0, \qquad f'_{X,2} = 0,$$
$$f_{X,4} = \frac{1}{(2\pi)^3}\frac{\psi_0^2(m-3+5m\psi_1-3\psi_1+2m\psi_1^2-2m\psi_1^3)}{(1-\psi_1)^3(1-m\psi_1^2)}$$

for $ma_1^2 < 1$. *Hence,*

$$C_{\mathrm{ARCH}(1)} = G_{\mathrm{ARCH}(1),0} + N^{-1}G_{\mathrm{ARCH}(1),3} + o(N^{-1}),$$

where

$$G_{\mathrm{ARCH}(1),0} = a_1 a_2 S_{T_0}\Phi(d_1) - a_1 K\Phi(d_2),$$
$$a_2 = \exp\left\{\tau\mu + \frac{\tau\psi_0}{2(1-\psi_1)}\right\},$$
$$d_1 = \left(\log S_{T_0}/K + \tau\mu + \frac{\tau\psi_0}{1-\psi_1}\right) \Big/ \left(\frac{\tau\psi_0}{1-\psi_1}\right)^{1/2},$$
$$d_2 = d_1 - \left(\frac{\tau\psi_0}{1-\psi_1}\right)^{1/2}$$

and

$$G_{\mathrm{ARCH}(1),3} = \frac{a_1 a_2 S_{T_0}}{24}\frac{m-3+5m\psi_1-3\psi_1+2m\psi_1^2-2m\psi_1^3}{(1-\psi_1)(1-m\psi_1^2)}$$
$$\times\left\{\sum_{j=1}^{3}\left(\frac{\tau\psi_0}{1-\psi_1}\right)^{j/2} H_{3-j}(-d_2)\phi(d_1) + \left(\frac{\tau\psi_0}{1-\psi_1}\right)^{2}\Phi(d_1)\right\}.$$

Figure 7.2 shows $C_{\mathrm{ARCH}(1),1} = G_{\mathrm{ARCH}(1),0}$ (dotted line) and $C_{\mathrm{ARCH}(1),3} = G_{\mathrm{ARCH}(1),0} + N^{-1}G_{\mathrm{ARCH}(1),3}$ (solid line) of Example 7.3 with $S_{T_0} = K = 100$, $\tau = 30/365$, $N = 30$ ($\Delta = 1/365$), $r = \mu = 0.05$, $m = 3$, $\psi_0 = 0.5$ and $-1/\sqrt{3} < \psi_1 < 1/\sqrt{3}$. Figure 7.2 illuminates influences of higher order terms under Gaussian innovations. From this, we can see that $C_{\mathrm{ARCH}(1),3}$ diverges as $\psi_1 \to \pm 1/\sqrt{3}$.

In Figure 7.3, we plotted $C_{\mathrm{ARCH}(1),1}$ (dotted line) and $C_{\mathrm{ARCH}(1),3}$ (solid line) of Example 7.3 with $S_{T_0} = 100$, $K = 95$, $\tau = 30/365$, $N = 30$, $r = \mu = 0.05$, $\psi_0 = 0.5$, $\psi_1 = 0.3$ and $1 < m < 9$. Figure 7.3 illuminates influences of non-Gaussian innovations. From this, we observe that $C_{\mathrm{ARCH}(1),3}$ decreases as $m \to 9$. The first-order term $C_{\mathrm{ARCH}(1),1}$ is a constant because of independence from m.

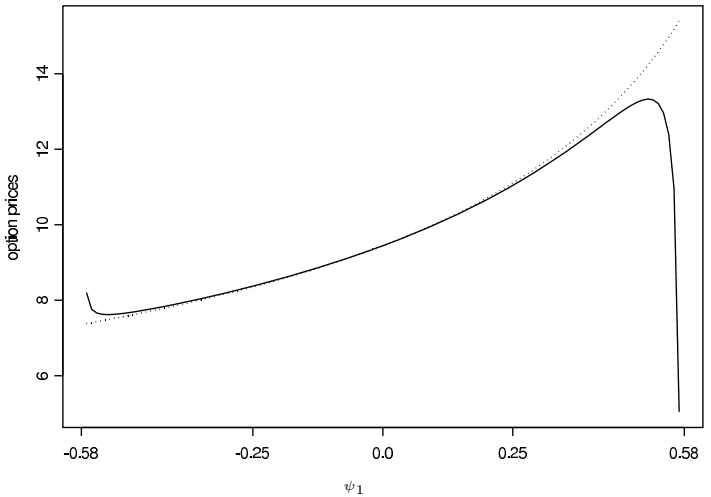


Figure 7.2 *For* ARCH(1) *in Example 7.3, the approximations up to the first* ($C_{\mathrm{ARCH}(1),1}$*, dotted line*) *and third order* ($C_{\mathrm{ARCH}(1),3}$*, solid line*) *of the option price are plotted with* $S_{T_0} = K = 100$*,* $\tau = 30/365$*,* $N = 30$*,* $r = \mu = 0.05$*,* $m = 3$*,* $\psi_0 = 0.5$ *and* $-1/\sqrt{3} < \psi_1 < 1/\sqrt{3}$*.*

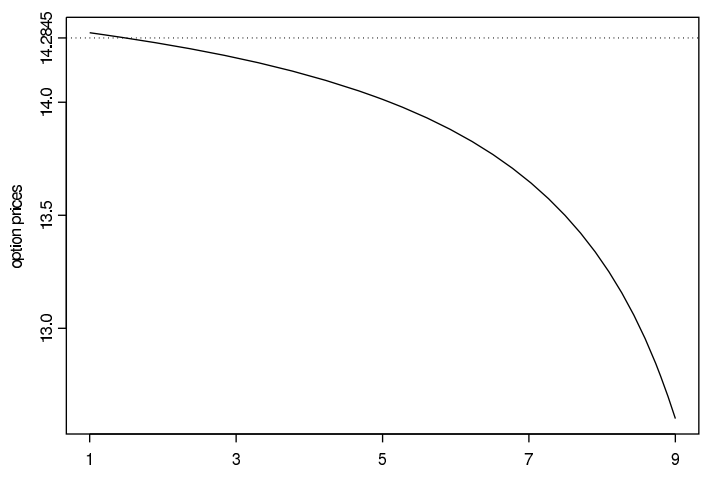


Figure 7.3 *For* ARCH(1) *in Example 7.3, the approximations up to the first* ($C_{\mathrm{ARCH}(1),1}$*, dotted line*) *and third order* ($C_{\mathrm{ARCH}(1),3}$*, solid line*) *of the option price are plotted with* $S_{T_0} = 100$*,* $K = 95$*,* $\tau = 30/365$*,* $N = 30$*,* $r = \mu = 0.05$*,* $\psi_0 = 0.5$*,* $\psi_1 = 0.3$ *and* $1 < m < 9$*.*

Next we consider option pricing problems for a class of processes generated by uncorrelated random variables, which includes the linear process and an important class in time series analysis. Here we are concerned with the following process

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j}, \quad t \in \mathbf{Z}, \tag{7.44}$$

where $\{\varepsilon_t; t \in \mathbf{Z}\}$ is a sequence of uncorrelated random variables. Instead of Assumption 7.1 (i) and (ii) we make the following assumption.

Assumption 7.2 *(i)* $\{\varepsilon_t; t \in \mathbf{Z}\}$ *is fourth-order stationary in the sense that*

$$\begin{aligned}
&\mathrm{E}(\varepsilon_t) = 0, \\
&\mathrm{Var}(\varepsilon_t) = \sigma^2, \\
&\mathrm{cum}(\varepsilon_t, \varepsilon_{t+u_1}, \varepsilon_{t+u_2}) = c_{\varepsilon,3}(u_1, u_2), \\
&\mathrm{cum}(\varepsilon_t, \varepsilon_{t+u_1}, \varepsilon_{t+u_2}, \varepsilon_{t+u_3}) = c_{\varepsilon,4}(u_1, u_2, u_3),
\end{aligned}$$

for any $t, u_1, \ldots, u_3 \in \mathbf{Z}$.

(ii) The cumulants $c_{\varepsilon,k}(u_1, \ldots, u_{k-1})$, $k = 3, 4$, *satisfy*

$$\sum_{u_1,\ldots,u_{k-1}=-\infty}^{\infty} \left(1 + |u_j|^{2-k/2}\right) |c_{\varepsilon,k}(u_1, \ldots, u_{k-1})| < \infty,$$

for $j = 1, \ldots, k-1$.

(iii) $\{a_j; j \in \mathbf{Z}\}$ *satisfies*

$$\sum_{j=0}^{\infty} (1 + |j|)|a_j| < \infty.$$

Under Assumption 7.2 (ii), $\{\varepsilon_t; t \in \mathbf{Z}\}$ has the k-th order cumulant spectral density. Let $f_{\varepsilon,k}$ be the k-th order cumulant spectral density evaluated at frequency $\mathbf{0}$

$$f_{\varepsilon,k} = (2\pi)^{-(k-1)} \sum_{u_1,\ldots,u_{k-1}=-\infty}^{\infty} c_{\varepsilon,k}(u_1, \ldots, u_{k-1})$$

for $k = 2, 3, 4$. The frequency response function of $\{a_j; j \in \mathbf{Z}\}$ is defined by

$$A(\lambda) = \sum_{j=0}^{\infty} a_j e^{-ij\lambda}$$

for $-\infty < \lambda < \infty$.

Under Assumption 7.2 (i)-(iii), Assumption 7.1 (i) and (ii) hold. Hence, from Theorem 7.2, we have

Corollary 7.2 *Suppose that Assumption 7.2* (i)-(iii) *and Assumption 7.1* (iii) *hold. Let* $a_1 = \exp(-r\tau)$ *and* $a_2 = \exp(\tau\mu + \tau\sigma^2 A^2/2)$. *Then*

$$C = G_0 + \frac{2\pi^2 A^3}{3\sigma^3|A|^3} N^{-1/2} f_{\varepsilon,3} G_3 - \frac{1}{2A^2} N^{-1} f'_{\varepsilon,2} G_2 + \frac{\pi^3}{3\sigma^4} N^{-1} f_{\varepsilon,4} G_4 + \frac{2\pi^4}{9\sigma^6} N^{-1} f^2_{\varepsilon,3} G_6 + o(N^{-1}),$$

where $A = A(0)$,

$$f'_{\varepsilon,2} = 2 \sum_{j_1,j_2=0}^{\infty} |j_2| a_{j_1} a_{j_1+j_2},$$

G_k, $k = 0, 2, 3, 4, 6$, *are given in Theorem 7.2 with*

$$f_{X,2} = \frac{\sigma^2}{2\pi} A^2.$$

PROOF From Assumption 7.2 (i)-(iii) and (7.44),

$$f_{X,k} = A^k f_{\varepsilon,k}, \quad k = 2, 3, 4,$$

and Assumption 7.1 (ii) holds. Note that

$$c_{X,2}(u) = Var\left(\sum_{j_1=0}^{\infty} a_{j_1} \varepsilon_{t-j_1}, \sum_{j_2=0}^{\infty} a_{j_2} \varepsilon_{t+u-j_2} \right) = \sigma^2 \sum_{j=0}^{\infty} a_j a_{|u|+j}.$$

We can see $f'_{X,2} = \sigma^2 f'_{\varepsilon,2}$. From the above arguments Corollary 7.2 follows. □

Example 7.4 *Let* $\{X_t; t \in \mathbf{Z}\}$ *be the AR(1) process*

$$X_t = \rho X_{t-1} + \varepsilon_t, \quad |\rho| < 1.$$

Note that

$$A = \frac{1}{1-\rho}, \qquad f'_{\varepsilon,2} = \frac{2\rho}{(1+\rho)(1-\rho)^3}.$$

The price of a European call option $C_{\mathrm{AR}(1)}$ *is given by*

$$C_{\mathrm{AR}(1)} = G_{AR(1),0} + N^{-1/2} G_{\mathrm{AR}(1),2} + N^{-1} G_{\mathrm{AR}(1),3} + o(N^{-1}),$$

where

$$G_{\mathrm{AR}(1),0} = a_1\{a_2 S_{T_0}\Phi(d_1) - K\Phi(d_2)\}, \qquad G_{\mathrm{AR}(1),2} = \frac{2\pi^2}{3\sigma^3} f_{\varepsilon,3} G_3,$$

$$G_{\mathrm{AR}(1),3} = -\frac{\rho}{1-\rho^2} G_2 + \frac{\pi^3}{3\sigma^4} f_{\varepsilon,4} G_4 + \frac{2\pi^4}{9\sigma^6}(f_{\varepsilon,3})^2 G_6,$$

$$a_2 = \exp\left\{\tau\mu + \frac{\tau\sigma^2}{2(1-\rho)^2}\right\},$$

$$d_1 = \left\{\log S_{T_0}/K + \tau\mu + \frac{\tau\sigma^2}{(1-\rho)^2}\right\} \Big/ \left(\frac{\tau^{1/2}\sigma}{1-\rho}\right), \qquad d_2 = d_1 - \left(\frac{\tau^{1/2}\sigma}{1-\rho}\right),$$

and

$$G_k = a_1 a_2 S_{T_0}\left\{\sum_{j=1}^{k-1}\left(\frac{\tau^{1/2}\sigma}{1-\rho}\right)^j H_{k-j-1}(-d_2)\phi(d_1) + \left(\frac{\tau^{1/2}\sigma}{1-\rho}\right)^k \Phi(d_1)\right\},$$

for $k = 2, 3, 4, 6$.

In order to show the influences of higher order terms, in Figure 7.4, we plotted $C_{\mathrm{AR}(1),1} = G_{\mathrm{AR}(1),0}$ (dotted line), $C_{\mathrm{AR}(1),2} = G_{\mathrm{AR}(1),0} + N^{-1/2}G_{\mathrm{AR}(1),2}$ (dashed line) and $C_{\mathrm{AR}(1),3} = G_{\mathrm{AR}(1),0} + N^{-1/2}G_{\mathrm{AR}(1),2} + N^{-1}G_{\mathrm{AR}(1),3}$ (solid line) of Example 7.4 with $S_{T_0} = K = 100$, $\tau = 30/365$, $N = 30$ ($\Delta = 1/365$), $r = \mu = 0.05$, $\sigma = 1$, $f_{X,3} = -0.1$, $f_{X,4} = 0.2$ and $-1 < \rho < 0.75$. From this, we observe that $C_{\mathrm{AR}(1),k}$, $k = 1, 2, 3$ diverge as $\rho \to 1$.

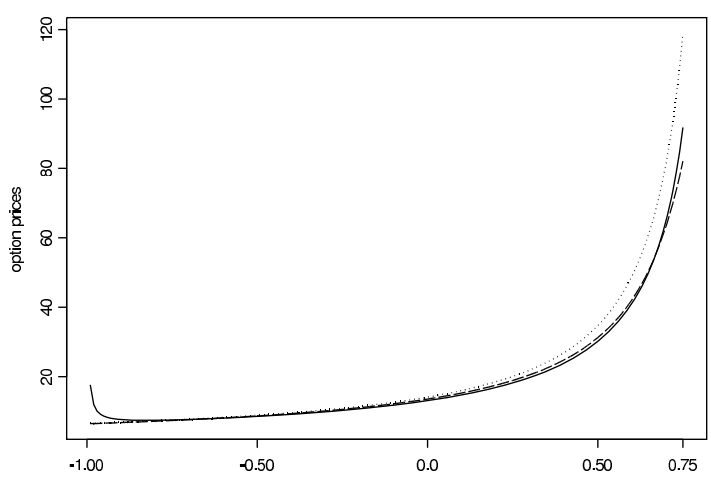


Figure 7.4 *For* AR(1) *in Example 7.4, the approximations up to the first* ($C_{\mathrm{ARCH}(1),1}$, *dotted line*), *second* ($C_{\mathrm{ARCH}(1),2}$, *dashed line*) *and third order* ($C_{\mathrm{ARCH}(1),3}$, *solid line*) *of the option price are plotted with* $S_{T_0} = K = 100$, $\tau = 30/365$, $N = 30$, $r = \mu = 0.05$, $\sigma = 1$, $f_{X,3} = -0.1$, $f_{X,4} = 0.2$ *and* $-1 < \rho < 0.75$.

In Examples 7.2 and 7.3, although the third-order terms diverge, the first-order terms do not diverge. On the other hand, in Example 7.4, even the first-order term does not converge if $\rho \to 1$. This fact is attributed to finiteness of the variances.

Before this we considered pricing problems with no martingale property. Now we recall that the theoretical price of an option is based on the risk neutrality argument. To investigate influences of the martingale restriction. we derive the option price based on the risk neutrality argument (see Cox and Ross (1976) and Longstaff (1995)).

Let

$$d_1^* = (\log S_{T_0}/K + r\tau + \pi\tau f_{X,2})/(2\pi\tau f_{X,2})^{1/2},$$
$$d_2^* = d_1^* - (2\pi\tau f_{X,2})^{1/2}.$$

Then we have

Theorem 7.4 *The fair price C^* of a European call option is given by*

$$C^* = G_0^* + \frac{(2\pi)^{1/2}}{6} N^{-1/2} \frac{f_{X,3}}{(f_{X,2})^{3/2}} G_3^* - \frac{1}{4\pi} N^{-1} \frac{f_{X,2}^*}{f_{X,2}} G_2^* + \frac{\pi}{12} N^{-1} \frac{f_{X,4}}{(f_{X,2})^2} G_4^* + \frac{\pi}{36} N^{-1} \frac{(f_{X,3})^2}{(f_{X,2})^3} G_6^* + o(N^{-1}),$$

where

$$G_0^* = S_{T_0}\Phi(d_1^*) - e^{-r\tau} K \Phi(d_2^*),$$
$$G_k^* = S_{T_0} \sum_{j=1}^{k-1} (2\pi\tau f_{X,2})^{j/2} H_{k-j-1}(-d_2^*)\phi(d_1^*),$$

for $k = 2, 3, 4$ and

$$G_6^* = S_{T_0} \left[\sum_{j=1}^{2} (2\pi\tau f_{X,2})^{j/2} \{ H_{5-j}(-d_2^*) - 2\pi\tau f_{X,2} H_{3-j}(-d_2^*) \} \right] \phi(d_1^*).$$

PROOF From the martingale restriction,

$$\begin{aligned} S_{T_0} &= e^{-r\tau} E_{T_0}[S_T], \\ &= e^{-r\tau} \int_{-\infty}^{\infty} S_{T_0} \exp\left\{ \tau\mu + (2\pi\tau f_{X,2})^{1/2} z \right\} g(z) dz. \end{aligned} \tag{7.45}$$

Note that

$$\int_{-\infty}^{\infty} \exp\left\{ (2\pi\tau f_{X,2})^{1/2} z \right\} H_k(z)\phi(z) dz = (2\pi\tau f_{X,2})^{k/2} \exp(\pi\tau f_{X,2})$$

for $k = 2, 3, 4, 6$. The equation (7.45) implies that

$$\begin{aligned} 1 = \exp\left(-r\tau + \tau\mu + \pi\tau f_{X,2}\right) \Bigg\{ & 1 + \frac{2}{3}\pi^2\tau^{3/2}N^{-1/2}f_{X,3} \\ & - \frac{1}{2}\tau N^{-1}f'_{X,2} + \frac{1}{3}\pi^3\tau^2 N^{-1}f_{X,4} + \frac{2}{9}\pi^4\tau^3 N^{-1}(f_{X,3})^2 \Bigg\} \\ & + o(N^{-1}). \end{aligned} \tag{7.46}$$

Taking the logarithm of the equation (7.46) and using Taylor expansion, yield

$$\begin{aligned} \mu = r - \pi f_{X,2} - \frac{2}{3}\pi^2\tau^{1/2}N^{-1/2}f_{X,3} \\ + \frac{1}{2}N^{-1}f'_{X,2} - \frac{1}{3}\pi^3\tau N^{-1}f_{X,4} + o(N^{-1}). \end{aligned} \tag{7.47}$$

Substituting (7.47) into G_k, $k = 0, 2, 3, 4, 6$ in Theorem 7.3, further expansion and collection of terms, we obtain

$$\begin{aligned} G_0 = G_0^* & - \frac{2}{3}\pi^2\tau^{3/2}S_{T_0}N^{-1/2}f_{X,3}\Phi(d_1^*) \\ & + S_{T_0}N^{-1}\left\{\frac{1}{2}\tau f^*_{X,2} - \frac{1}{3}\pi^3\tau^2 f_{X,4} + \frac{2}{9}\pi^4\tau^3(f_{X,3})^2\right\}\Phi(d_1^*) \\ & + \frac{\pi}{36}S_{T_0}N^{-1}\frac{(f_{X,3})^2}{(f_{X,2})^3}(2\pi\tau f_{X,2})^{5/2}\phi(d_1^*) + o(N^{-1}), \end{aligned} \tag{7.48}$$

$$\begin{aligned} G_3 = G_3^* & + S_{T_0}(2\pi\tau f_{X,2})^{3/2}\Phi(d_1^*) \\ & - \frac{(2\pi)^{7/2}}{6}\tau^3 S_{T_0}N^{-1/2}f_{X,3}(f_{X,2})^{3/2}\Phi(d_1^*) \\ & - \frac{(2\pi)^{1/2}}{6}S_{T_0}N^{-1/2}\frac{f_{X,3}}{(f_{X,2})^{3/2}}\sum_{j=1}^{3}(2\pi\tau f_{X,2})^{j/2+1}H_{3-j}(-d_2^*)\phi(d_1^*) \\ & + o(N^{-1/2}), \end{aligned} \tag{7.49}$$

and

$$\begin{aligned} G_k = S_{T_0} & \left\{\sum_{j=1}^{k-1}(2\pi\tau f_{X,2})^{j/2}H_{k-j-1}(-d_2^*)\phi(d_1^*) + (2\pi\tau f_{X,2})^{k/2}\Phi(d_1^*)\right\} \\ & + o(1) \end{aligned} \tag{7.50}$$

for $k = 2, 4, 6$. From (7.48)-(7.50), Theorem 7.4 follows. □

Example 7.5 *Suppose that $\{X_t; t \in \mathbf{Z}\}$ is the AR(1) process in Example 7.4. Then the fair price of a European call option $C^*_{\mathrm{AR}(1)}$ is given by*

$$C^*_{\mathrm{AR}(1)} = G^*_{\mathrm{AR}(1),0} + N^{-1/2}G^*_{\mathrm{AR}(1),2} + N^{-1}G^*_{\mathrm{AR}(1),3} + o(N^{-1}),$$

where

$$G^*_{\mathrm{AR}(1),0} = S_{T_0}\Phi(d^*_1) - e^{-r\tau}K\Phi(d^*_2), \quad G^*_{\mathrm{AR}(1),2} = \frac{2\pi^2}{3\sigma^3}f_{\varepsilon,3}G^*_3,$$

$$G^*_{\mathrm{AR}(1),3} = -\frac{\rho}{1-\rho^2}G^*_2 + \frac{\pi^3}{3\sigma^4}f_{\varepsilon,4}G_4 + \frac{2\pi^4}{9\sigma^6}(f_{\varepsilon,3})^2G^*_6,$$

$$d^*_1 = \left\{\log S_{T_0}/K + r\tau + \frac{\tau\sigma^2}{2(1-\rho)^2}\right\} \Big/ \left(\frac{\tau^{1/2}\sigma}{1-\rho}\right), \quad d^*_2 = d^*_1 - \left(\frac{\tau^{1/2}\sigma}{1-\rho}\right),$$

$$G^*_k = S_{T_0}\left\{\sum_{j=1}^{k-1}\left(\frac{\tau^{1/2}\sigma}{1-\rho}\right)^j H_{k-j-1}(-d^*_2)\phi(d^*_1)\right\},$$

for $k = 2, 3, 4$ *and*

$$G^*_6 = S_{T_0}\left[\sum_{j=1}^{2}\left(\frac{\tau^{1/2}\sigma}{1-\rho}\right)^j \left\{H_{5-j}(-d^*_2) - \frac{\tau\sigma^2}{(1-\rho)^2}H_{3-j}(-d^*_2)\right\}\right]\phi(d^*_1).$$

Figure 7.5 shows $C^*_{\mathrm{AR}(1),1} = G^*_{\mathrm{AR}(1),0}$ (dotted line), $C^*_{\mathrm{AR}(1),2} = G^*_{\mathrm{AR}(1),0} + N^{-1/2}G^*_{\mathrm{AR}(1),2}$ (dashed line) and $C^*_{\mathrm{AR}(1),3} = G^*_{\mathrm{AR}(1),0} + N^{-1/2}G^*_{\mathrm{AR}(1),2} + N^{-1}G^*_{\mathrm{AR}(1),3}$ (solid line) of Example 7.5 with $S_{T_0} = K = 100$, $\tau = 30/365$, $N = 30$ ($\Delta = 1/365$), $r = 0.05$, $\sigma = 1$, $f_{X,3} = -0.1$, $f_{X,4} = 0.2$ and $-1 < \rho < 1$. Unlike Example 7.4, we observe that $C_{\mathrm{AR}(1),k}$, $k = 1, 2, 3$ converge to $S_{T_0}(= 100)$ as $\rho \to 1$.

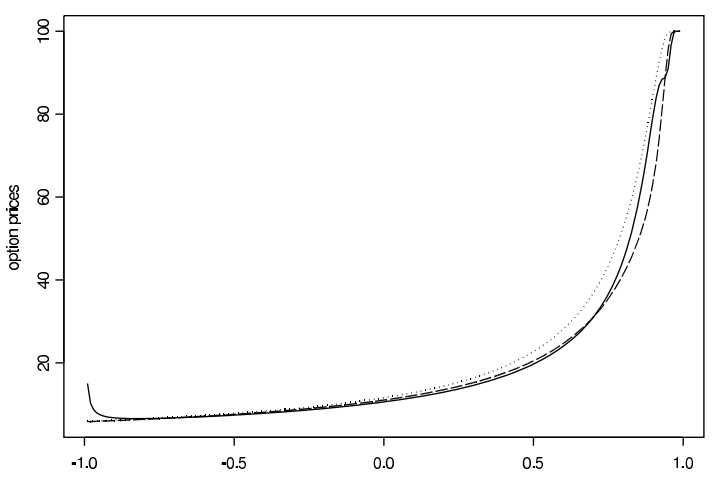


Figure 7.5 *For* AR(1) *in Example 7.5, the approximations up to the first* ($C^*_{\mathrm{ARCH}(1),1}$, *dotted line*), *second* ($C^*_{\mathrm{ARCH}(1),2}$, *dashed line*) *and third order* ($C^*_{\mathrm{ARCH}(1),3}$, *solid line*) *of the option price are plotted with* $S_{T_0} = K = 100$, $\tau = 30/365$, $N = 30$, $r = 0.05$, $\sigma = 1$, $f_{X,3} = -0.1$, $f_{X,4} = 0.2$ *and* $-1 < \rho < 1$.

From (7.35), X_{j-N_0}, $j = 1, \ldots, N_0$, are available, where $N_0 = T_0/\Delta$. Therefore, we consider to estimate μ, $f_{X,2}$, $f'_{X,2}$, $f_{X,3}$ and $f_{X,4}$ in Theorems 7.2 and 7.3 consistently based on the past observations. From Assumption 7.1 (i), $\Delta\mu$ is the mean of stock log returns. Hence, a natural unbiased estimator of μ is the sample mean

$$\hat{\mu} = \frac{1}{\Delta N_0} \sum_{j=1}^{N_0} \{\log S_{j\Delta} - \log S_{(j-1)\Delta}\}. \tag{7.51}$$

The variance of $\hat{\mu}$ is given by

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{\Delta N_0} \sum_{u=-(N_0-1)}^{N_0-1} \left(1 - \frac{|u|}{N_0}\right) c_{X,2}(u).$$

Hence, under Assumption 7.1 (ii), $\hat{\mu}$ given in (7.51) is a consistent estimator of μ.

Moreover in order to construct a consistent estimator of $f'_{X,2}$, we define the lag window function $w(\cdot)$ which is an even and piecewise continuous function satisfying the conditions,

$$\begin{aligned} &w(0) = 1, \\ &|w(x)| \le 1, \quad \text{for all } x, \\ &w(x) = 0, \quad \text{for } |x| > 1. \end{aligned} \tag{7.52}$$

Let

$$\hat{f}'_{X,2} = \sum_{u=-(N_0-1)}^{N_0-1} |u| \hat{c}_{X,2}(u) w\left(B_{N_0} u\right),$$

where $\hat{c}_{X,2}(u)$ is the sample autocovariance function at lag u

$$\begin{aligned} \hat{c}_{X,2}(u) = \frac{1}{\Delta N_0} \sum_{j=1}^{N_0-|u|} &\{\log S_{(j+|u|)\Delta} - \log S_{(j+|u|-1)\Delta} - \Delta\hat{\mu}\} \\ &\times \{\log S_{j\Delta} - \log S_{(j-1)\Delta} - \Delta\hat{\mu}\}, \end{aligned} \tag{7.53}$$

and $B_{N_0} \to 0$ as $N_0 \to \infty$, but $(B_{N_0})^3 N_0 \to \infty$. Then we can easily see that under Assumption 7.1 (ii), $\hat{f}'_{X,2}$ given in (7.53) is a consistent estimator of $f'_{X,2}$.

Since $f_{X,k}$, $k = 2, 3, 4$, are the k-th order cumulant spectral density evaluated at frequency $\mathbf{0}$, using the Brillinger and Rosenblatt (1967a, 1967b) formula, we construct consistent estimators $\hat{f}_{X,k}$ of $f_{X,k}$ $(k = 2, 3, 4)$. See also Brillinger (1981). Thus we can consistently estimate all the quantities in Theorems 7.2 and 7.3 (e.g., G_j, $j = 0, 2, 3, 4, 6.$) by the corresponding quantities replacing μ, $f'_{X,2}$ and $f_{X,k}$ by $\hat{\mu}$, $\hat{f}'_{X,2}$ and $\hat{f}_{X,k}$ $(k = 2, 3, 4)$, respectively.

For example, we discuss a consistent estimator for New York Stock Exchange data. The data are daily returns of AMOCO, Ford, HP, IBM and Merck companies. The individual time series are the last 1024 data points from stocks, representing the daily returns for the five companies from February 2, 1984, to December 31, 1991. We used the window functions

$$W(u_1, \ldots, u_{k-1}) = \begin{cases} 2^{-(k-1)} & \text{If } |u_1|, \ldots, |u_{k-1}| \leq 1, \\ 0 & \text{otherwise} \end{cases}$$

for $\hat{f}_{X,k}$ $(k = 2, 3, 4)$ and let $w(u) = 1$ for $|u| \leq 1$, where $w(u)$ is defined in (7.52). Also we used the bandwidth in frequency direction with $B_{N_0} = 1/50$ for $\hat{f}_{X,2}$, $B_{N_0} = 1/30$ for $\hat{f}_{X,3}$ and $B_{N_0} = 1/10$ for $\hat{f}_{X,4}$ and $\hat{f}'_{X,2}$ (see Brillinger and Rosenblatt (1967a, 1967b), and Brillinger (1981)).

Table 7.1 *Values of consistent estimators*

	AMOCO	Ford	HP	IBM	Merck
$\hat{\mu}$	0.235103	0.045337	0.133815	0.017165	0.481340
$\hat{f}_{X,2}$	0.002937	0.016006	0.016202	0.003085	0.004534
$\frac{\hat{f}_{X,3}}{(\hat{f}_{X,2})^{3/2}}$	-0.706149	-3.078889	8.501363	0.470144	2.419969
$\frac{\hat{f}_{X,4}}{(\hat{f}_{X,2})^{2}}$	2.278478	-0.280973	8.651378	15.0914	-2.249174
$\frac{\hat{f}'_{X,2}}{\hat{f}_{X,2}}$	-22.78799	-5.520428	0.169291	27.18047	-37.3221

Table 7.1 show the values of consistent estimators of μ, $f'_{X,2}$ and $f_{X,k}$ $(k = 2, 3, 4)$ for the five companies. From these results, we can see that the quantities involved in higher order terms are quite different from the Black and Scholes model. Therefore, in general the assumptions of the Gaussianity and mutual independence of stock log returns will not hold.

Table 7.2 *Option prices*

	AMOCO	Ford	HP	IBM	Merck
C_1	2.776419	4.031663	4.472833	1.699889	4.689151
C_2	2.809884	3.979554	4.434833	1.700269	4.495491
C_3	2.881406	4.345765	6.392765	1.374588	4.650024

Table 7.2 show the values of the approximation up to the first C_1, second C_2

and third-order C_3 of the option prices with $S_{T_0} = K = 100$, $\tau = 30/365$, $N = 30$, $r = 0.05$. From these results, we observe that option prices are strongly affected by third-order terms except for AMOCO and Merck.

Table 7.3 *Fair prices*

	AMOCO	Ford	HP	IBM	Merck
C_1^*	1.764254	3.827175	3.849221	1.80241	2.138307
C_2^*	1.769475	3.784867	3.954549	1.798532	2.124842
C_3^*	1.83751	4.111153	6.09142	1.481998	2.459177

Table 7.3 show the values of the approximation up to the first C_1^*, second C_2^* and third order C_3^* of the fair prices with $S_{T_0} = K = 100$, $\tau = 30/365$, $N = 30$, $r = 0.05$. From these results, we observe that option prices are strongly affected by third-order terms.

The Black and Scholes model assumes the Gaussianity and mutual independence of stock log returns. Empirical studies, however, show that they are not Gaussian nor independent. In this section, dropping these two assumptions, we derived a European option pricing. Then, we observed that option prices are strongly affected by the non-Gaussianity and dependence of stock log returns. Hence, it should be noted that we use option pricing models taking account of the non-Gaussianity and dependence of stock log returns.

7.3 Estimation of Portfolio

In the theory of portfolio analysis, optimal portfolios are determined by the mean $\boldsymbol{\mu}$ and variance Σ of the portfolio return. Several authors proposed estimators of optimal portfolios as functions of the sample mean $\hat{\boldsymbol{\mu}}$ and the sample variance $\hat{\Sigma}$ for independent returns of assets (e.g., Jobson and Korkie (1980, 1989), and Lauprete et al. (2002)). However, empirical studies show that financial return processes are often dependent and non-Gaussian. The following figure shows East Japan Railway Company's stock return $\{X_t\}$ from 1993/10/27 to 2005/01/28.

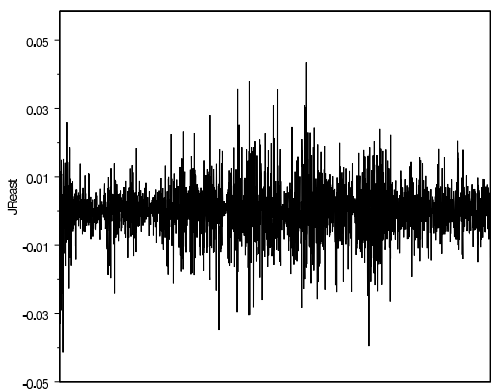


Figure 7.6 $\{X_t\}$

The sample autocorrelation function ($\mathrm{acf}_{X_t}(l)$) of $\{X_t\}$ is given in Figure 7.7.

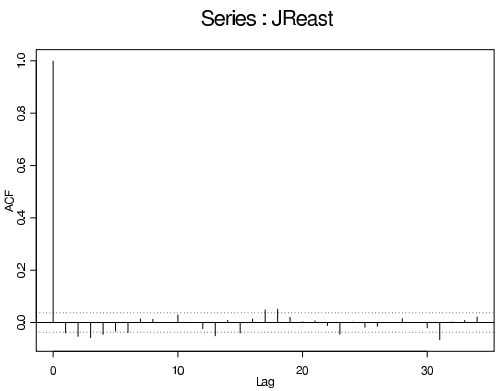


Figure 7.7 $\mathrm{acf}_{X_t}(l)$

From this we can observe that $\{X_t\}$ is almost uncorrelated, i.e.,

$$\mathrm{acf}_{X_t}(l) \approx 0 \ (l \neq 0).$$

Next we plot the autocorrelation function ($\mathrm{acf}_{X_t^2}(l)$) of $\{X_t^2\}$ in Figure 7.8.

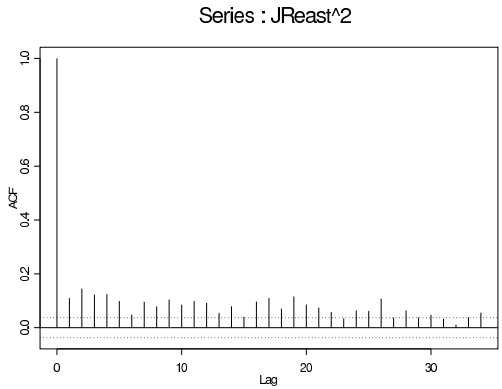


Figure 7.8 $\mathrm{acf}_{X_t^2}(l)$

From this we can see that the squared process $\{X_t^2\}$ is correlated. This symptom leads us to the assumption that financial return processes are dependent and non-Gaussian. From this point of view, Basak et al. (2002) showed the consistency of optimal portfolio estimators when portfolio returns are stationary processes. However, in the literature there has been no study on the asymptotic efficiency of estimators for optimal portfolios. Therefore, in this section, denoting optimal portfolios by a function $g = g(\boldsymbol{\mu}, \Sigma)$ of $\boldsymbol{\mu}$ and Σ, we discuss the asymptotic efficiency of estimators $\hat{g} = g(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ when the return is a vector-valued non-Gaussian stationary process $\{\mathbf{X}(t)\}$. Then it is shown that $\hat{g}$ is not asymptotically efficient generally even if $\{\mathbf{X}(t)\}$ is Gaussian, which gives a strong warning for use of the usual estimator $\hat{g}$. We also show that there are some cases when the asymptotic variance $V_{NG}(\hat{g})$ of $\hat{g}$ under non-Gaussianity can be smaller than that under Gaussianity $V_G(\hat{g})$. Numerical studies are given to illuminate the results above. For non-Gaussian dependent return processes, we propose to use maximum likelihood type estimators for g, which are asymptotically efficient. Furthermore, we investigate the problem of predicting the one step ahead optimal portfolio return by the estimated portfolio based on $\hat{g}$, and evaluate the mean squares prediction error. Numerical examples for actual financial data are provided. As a conclusion it seems very important to make the consideration for non-Gaussianity and dependence of return processes.

Throughout this section, $\|\mathbf{A}\|$ denotes the Euclidean norm of a matrix $\mathbf{A}$ and $|\mathbf{A}|$ denotes the sum of the absolute values of all entries of $\mathbf{A}$. We write $X_n \xrightarrow{\mathcal{L}} X$ if $\{X_n\}$ converges in distribution to X. The 'vec' operator transforms

a matrix into a vector by stacking columns, and the 'vech' operator transforms a symmetric matrix into a vector by stacking elements on and below the main diagonal. For matrices $\mathbf{A} = \{a_{j_1 j_2}\}$ and $\mathbf{B}$, $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$, whose (j_1, j_2)th block is $a_{j_1 j_2}\mathbf{B}$.

In what follows we develop our discussion mainly based on Shiraishi and Taniguchi (2005). Suppose the existence of a finite number of assets indexed by $i, (i = 1, \ldots, m)$. Let $\mathbf{X}(t) = (X_1(t), \ldots, X_m(t))'$ denote the random returns on m assets at time t. Assuming the stationarity of $\{\mathbf{X}(t)\}$, write $\boldsymbol{\mu} = E\{\mathbf{X}(t)\}$ and $\Sigma = \mathrm{Cov}(\mathbf{X}(t))$. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ be the vector of portfolio weights. Then the return of portfolio is $\mathbf{X}(t)'\boldsymbol{\alpha}$, and the expectation and variance are, respectively, given by $\mu(\boldsymbol{\alpha}) = \boldsymbol{\mu}'\boldsymbol{\alpha}$ and $\eta^2(\boldsymbol{\alpha}) = \boldsymbol{\alpha}'\Sigma\boldsymbol{\alpha}$. Optimal portfolio weights have been proposed by various criteria (see Jobson and Korkie (1980), and Gouriéroux (1997), etc.). The following are typical ones.

Consider the optimization problem

$$\begin{cases} \max\limits_{\boldsymbol{\alpha}} \left\{\mu(\boldsymbol{\alpha}) - \beta\eta^2(\boldsymbol{\alpha})\right\}, \\ \text{subject to } \mathbf{e}'\boldsymbol{\alpha} = 1, \end{cases}$$

where $\mathbf{e} = (1, \ldots, 1)'$ ($m \times 1$-vector), and β is a given positive number. The solution for $\boldsymbol{\alpha}$ is given by

$$\boldsymbol{\alpha}_{\mathrm{I}} = \frac{1}{2\beta}\left\{\Sigma^{-1}\boldsymbol{\mu} - \frac{\mathbf{e}'\Sigma^{-1}\boldsymbol{\mu}}{\mathbf{e}'\Sigma^{-1}\mathbf{e}}\Sigma^{-1}\mathbf{e}\right\} + \frac{\Sigma^{-1}\mathbf{e}}{\mathbf{e}'\Sigma^{-1}\mathbf{e}}. \tag{7.54}$$

Next we consider

$$\begin{cases} \min\limits_{\boldsymbol{\alpha}} \eta^2(\boldsymbol{\alpha}), \\ \text{subject to } \mathbf{e}'\boldsymbol{\alpha} = 1. \end{cases}$$

The solution for $\boldsymbol{\alpha}$ is given by

$$\boldsymbol{\alpha}_{\mathrm{II}} = \frac{\Sigma^{-1}\mathbf{e}}{\mathbf{e}'\Sigma^{-1}\mathbf{e}}. \tag{7.55}$$

Let us now suppose that there exists a risk-free asset. We denote by R_0 its return, and denote by α_0 the amount. The problem to be solved is given by

$$\begin{cases} \max\limits_{\alpha_0, \boldsymbol{\alpha}} \left\{\mu(\boldsymbol{\alpha}) + R_0\alpha_0 - \beta\eta^2(\boldsymbol{\alpha})\right\}, \\ \text{subject to } \sum_{j=0}^{m} \alpha_j = 1. \end{cases} \tag{7.56}$$

Then the solution for $\boldsymbol{\alpha}$ and α_0 are

$$\boldsymbol{\alpha}_{\mathrm{III}} = \frac{1}{2\beta}\Sigma^{-1}(\boldsymbol{\mu} - R_0\mathbf{e}), \tag{7.57}$$

$$\alpha_{0\mathrm{III}} = 1 - \frac{1}{2\beta}\mathbf{e}'\Sigma^{-1}(\boldsymbol{\mu} - R_0\mathbf{e}). \tag{7.58}$$

Therefore optimal portfolios can be considered as smooth functions of $\boldsymbol{\mu}$ and

Σ, i.e., we may put

$$g_1(\boldsymbol{\mu}, \Sigma) = \frac{1}{2\beta}\left\{\Sigma^{-1}\boldsymbol{\mu} - \frac{\mathbf{e}'\Sigma^{-1}\boldsymbol{\mu}}{\mathbf{e}'\Sigma^{-1}\mathbf{e}}\Sigma^{-1}\mathbf{e}\right\} + \frac{\Sigma^{-1}\mathbf{e}}{\mathbf{e}'\Sigma^{-1}\mathbf{e}}, \tag{7.54}'$$

$$g_2(\boldsymbol{\mu}, \Sigma) = \frac{\Sigma^{-1}\mathbf{e}}{\mathbf{e}'\Sigma^{-1}\mathbf{e}}, \tag{7.55}'$$

$$g_3(\boldsymbol{\mu}, \Sigma) = \frac{1}{2\beta}\Sigma^{-1}(\boldsymbol{\mu} - R_0\mathbf{e}), \tag{7.57}'$$

$$g_4(\boldsymbol{\mu}, \Sigma) = 1 - \frac{1}{2\beta}\mathbf{e}'\Sigma^{-1}(\boldsymbol{\mu} - R_0\mathbf{e}). \tag{7.58}'$$

Unifying the above we consider to estimate a general function $g(\boldsymbol{\mu}, \Sigma)$ of $\boldsymbol{\mu}$ and Σ. Here it should be noted that the coefficient $\boldsymbol{\alpha}$ satisfies the restriction $\mathbf{e}'\boldsymbol{\alpha} = 1$. Then we have only to estimate the subvector $(\alpha_1, \ldots, \alpha_{m-1})'$. Hence we assume that the function $g(\cdot)$ is $(m-1)$-dimensional, i.e.,

$$g : (\boldsymbol{\mu}, \Sigma) \to \mathbf{R}^{m-1}. \tag{7.59}$$

This section addresses the problem of statistical estimation for $g(\boldsymbol{\mu}, \Sigma)$, which describes various optimal portfolios.

As we said in the above, empirical studies show that financial return processes are often dependent and non-Gaussian. So it is natural to suppose that the return process concerned is dependent and non-Gaussian. Henceforth we assume that the return process $\{\mathbf{X}(t) = (X_1(t), \ldots, X_m(t))'; t \in \mathbf{Z}\}$ is an m-vector non-Gaussian stationary process with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)'$ and autocovariance matrix $\mathbf{R}(k)$.

Initially, we assume that the return process $\{\mathbf{X}(t) = (X_1(t), \ldots, X_m(t))'; t \in \mathbf{Z}\}$ is an m-vector linear process

$$\mathbf{X}(t) = \sum_{j=0}^{\infty} \mathbf{A}(j)\mathbf{U}(t-j) + \boldsymbol{\mu}, \quad t \in \mathbf{Z}, \tag{7.60}$$

where $\{\mathbf{U}(t) = (u_1(t), \ldots, u_m(t))'\}$ is a sequence of independent and identically distributed (i.i.d.) m-vector random variables with $E\mathbf{U}(t) = \mathbf{0}$, $\mathrm{Var}\{\mathbf{U}(t)\} = \mathbf{K}$ (for short $\{\mathbf{U}(t)\} \sim$ i.i.d. $(\mathbf{0}, \mathbf{K})$), and fourth-order cumulants. Here

$$\mathbf{K} = \{K_{ab}; a, b = 1, \ldots, m\},$$
$$\boldsymbol{\mu} = \{\mu_a; a = 1, \ldots, m\},$$
$$\mathbf{A}(j) = \{A_{ab}(j); a, b = 1, \ldots, m\}, \quad j \in \mathbf{Z}, \quad \mathbf{A}(0) = \mathbf{I}_m.$$

We make the following assumption.

Assumption 7.3

(i) $\sum_{j=0}^{\infty} |j|^{1+\delta} \|\mathbf{A}(j)\| < \infty$ *for some* $\delta > 0$,

(ii) $\det\left\{\sum_{j=0}^{\infty} \mathbf{A}(j) z^j\right\} \neq 0$ *on* $\{z; |z| \leq 1\}$.

The class of $\{\mathbf{X}(t)\}$ includes that of non-Gaussian vector-valued causal ARMA models. Hence the class is sufficiently rich. The process $\{\mathbf{X}(t)\}$ is a second-order stationary process with spectral density matrix

$$\mathbf{f}(\lambda) = \{f_{ab}(\lambda); a, b = 1, \ldots, m\} = (2\pi)^{-1}\mathbf{A}(\lambda)\mathbf{K}\mathbf{A}(\lambda)^*,$$

where $\mathbf{A}(\lambda) = \sum_{j=0}^{\infty} \mathbf{A}(j)e^{ij\lambda}$. Writing

$$\mathbf{R}(k) = E\left\{(\mathbf{X}(t) - \boldsymbol{\mu})\,(\mathbf{X}(t+k) - \boldsymbol{\mu})'\right\}, \quad (\mathbf{R}(0) = \Sigma),$$

we define $(m + r)$-vector parameter $\boldsymbol{\theta}$ by

$$\boldsymbol{\theta} = (\boldsymbol{\mu}', \text{vech}\{\mathbf{R}(0)\}')'$$

where $r = m(m + 1)/2$. From the partial realization $\{\mathbf{X}(1), \ldots, \mathbf{X}(n)\}$, we introduce

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{X}(t), \tag{7.61}$$

$$\hat{\mathbf{R}}(k) = \frac{1}{n-k}\sum_{t=1}^{n-k} \left\{(\mathbf{X}(t) - \hat{\boldsymbol{\mu}})\,(\mathbf{X}(t+k) - \hat{\boldsymbol{\mu}})'\right\}, \tag{7.62}$$

$$\mathbf{R}^*(k) = \frac{1}{n-k}\sum_{t=1}^{n-k} \left\{(\mathbf{X}(t) - \boldsymbol{\mu})\,(\mathbf{X}(t+k) - \boldsymbol{\mu})'\right\}, \tag{7.63}$$

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}', \text{vech}\{\hat{\mathbf{R}}(0)\}')'. \tag{7.64}$$

Denote the σ-field generated by $\{\mathbf{X}(s); s \le t\}$ by $\mathcal{F}_t$. Also we introduce matrices;

$$\Omega_1 = 2\pi\mathbf{f}(0), \quad (m \times m)\text{-matrix},$$

$$\Omega_2 = \left\{ 2\pi \int_{-\pi}^{\pi} \{f_{a_1a_3}(\lambda)\overline{f_{a_2a_4}(\lambda)} + f_{a_1a_4}(\lambda)\overline{f_{a_2a_3}(\lambda)}\} d\lambda \right.$$
$$+ \frac{1}{(2\pi)^2} \sum_{b_1,\ldots,b_4=1}^{m} c^{\mathbf{U}}_{b_1,\ldots,b_4}$$
$$\times \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} A_{a_1b_1}(\lambda_1)A_{a_2b_2}(-\lambda_1)A_{a_3b_3}(\lambda_2)A_{a_4b_4}(-\lambda_2)d\lambda_1 d\lambda_2$$
$$\left. ; a_1, a_2, a_3, a_4 = 1, \ldots, m, a_1 \ge a_2 \text{ and } a_3 \ge a_4 \right\}, \quad (r \times r)\text{-matrix},$$

$$\Omega_3 = \Bigg\{ \frac{1}{(2\pi)^2} \sum_{b_1,b_2,b_3=1}^{m} c^{\mathbf{U}}_{b_1,b_2,b_3} \times \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} A_{a_1b_1}(\lambda_1+\lambda_2)A_{a_2b_2}(-\lambda_1)A_{a_3b_3}(-\lambda_2)d\lambda_1 d\lambda_2 ; a_1, a_2, a_3 = 1, \ldots, m, a_2 \geq a_3 \Bigg\}, \quad (m \times r)\text{-matrix},$$

where $c^{\mathbf{U}}_{b_1,\ldots,b_j}$'s are jth order cumulants of $\mathbf{U}_{b_1}(t), \ldots, \mathbf{U}_{b_j}(t)$ $(j = 3, 4)$. Then we have the following result. For the proof, see Shiraishi and Taniguchi (2005).

Theorem 7.5 *Under Assumption 7.3,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \Omega_{NG}),$$

where

$$\Omega_{NG} = \begin{pmatrix} \Omega_1 & \Omega_3 \\ \Omega_3' & \Omega_2 \end{pmatrix}.$$

For g given by (7.59) we impose the following.

Assumption 7.4 *The function $g(\boldsymbol{\theta})$ is continuously differentiable with respect to $\boldsymbol{\theta}$.*

As a unified estimator for optimal portfolios we introduce $g(\hat{\boldsymbol{\theta}})$. From Theorem 7.5 and the δ-method (e.g., Brockwell and Davis (1991, Proposition 6.4.3)) we have the following result.

Theorem 7.6 *Under Assumptions 7.3 and 7.4,*

$$\sqrt{n}(g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})) \xrightarrow{\mathcal{L}} N\left(\mathbf{0}, \left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right) \Omega_{NG} \left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)'\right),$$

where $\partial g/\partial \boldsymbol{\theta}'$ is the vector differentiation (see Magnus and Neudecker (1988)).

Here we investigate the problem of predicting the one step ahead optimal portfolio return by the estimated portfolio based on $\hat{g}$.

Assume that $\{\mathbf{X}(1), \ldots, \mathbf{X}(n)\}$ is a realization of the m-vector linear process (7.60), and let $\{\mathbf{Y}(1), \ldots, \mathbf{Y}(n)\}$ be an independent realization of the same process. If $\hat{\boldsymbol{\theta}}$ is defined by (7.64) and if we use the following

$$\widehat{\mathrm{PR}}(n) = \mathbf{Y}(n)' g(\boldsymbol{\theta})$$

as a predictor of $\mathrm{PR}(n+1) \equiv \mathbf{Y}(n+1)' g(\boldsymbol{\theta})$, then the mean-square prediction

error (PE) is

$$\begin{aligned}\mathrm{PE} &= E\{\mathrm{PR}(n+1) - \widehat{\mathrm{PR}}(n)\}^2 \\ &= E\left[\mathbf{Y}(n)'\left\{g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})\right\}\right]^2 + E\left[\{\mathbf{Y}(n) - \mathbf{Y}(n+1)\}' g(\boldsymbol{\theta})\right]^2 \\ &\quad + 2E\left[\mathbf{Y}(n)'\left\{g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})\right\}\{\mathbf{Y}(n) - \mathbf{Y}(n+1)\}' g(\boldsymbol{\theta})\right] \\ &\equiv (1) + (2) + (3).\end{aligned}$$

Noting independence of $\{\mathbf{X}(t)\}$ and $\{\mathbf{Y}(t)\}$, we obtain

$$\begin{aligned}(1) &= \frac{1}{n}\mathrm{tr}\left[\{\boldsymbol{\mu}\boldsymbol{\mu}' + \mathbf{R}(0)\} C_n\right] \\ (2) &= 2g(\boldsymbol{\theta})'\{\mathbf{R}(0) - \mathbf{R}(1)\} g(\boldsymbol{\theta}) \\ (3) &= \frac{2}{\sqrt{n}} B_n' \{\mathbf{R}(0) - \mathbf{R}(1)\} g(\boldsymbol{\theta}),\end{aligned}$$

where

$$\begin{aligned}B_n &= \sqrt{n} E\left\{g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})\right\} \\ C_n &= nE\left\{g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})\right\}\left\{g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})\right\}'.\end{aligned}$$

Recalling that $\sqrt{n}\{g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})\}$ has the asymptotic distribution

$$N\left(\mathbf{0}, \left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)\Omega_{NG}\left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)'\right),$$

we can see that

$$\begin{aligned}B_n &= o(1) \\ C_n &= \left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)\Omega_{NG}\left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)' + o(1).\end{aligned}$$

We evaluate PE for various spectral structures, numerically.

Example 7.6 ***(Prediction Error (PE))*** *Let* $\mathbf{X}(1), \ldots, \mathbf{X}(100)$ *be an observed stretch from the return process* $\{\mathbf{X}(t) = (X_1(t), X_2(t))'; t \in \mathbf{Z}\}$ *generated by*

$$\begin{aligned}(1 - \alpha B)X_1(t) &= (1 - \beta B)U(t) + \mu_{X_1} \\ X_2(t) &= \mu_{X_2},\end{aligned}$$

where $U(t) \sim$ *i.i.d.* $t(10)$ *and* $t(10)$ *is a t-distribution with 10 degrees of freedom. Let the portfolio function be defined by*

$$g(\mu_{X_1}, \mu_{X_2}, R_{X_1}(0)) = \frac{\mu_{X_1} - \mu_{X_2}}{2R_{X_1}(0)}.$$

This portfolio is one of the solutions of (7.56). *In this case we estimate the* PE *by*

$$\begin{aligned}\widehat{\mathrm{PE}} &= 2g(\boldsymbol{\theta})' \{R_{X_1}(0) - R_{X_1}(1)\} g(\boldsymbol{\theta}) \quad (\equiv \mathrm{PE1}) \\ &\quad + \frac{2}{\sqrt{100}} \hat{B}'_{100} \{R_{X_1}(0) - R_{X_1}(1)\} g(\boldsymbol{\theta}) \quad (\equiv \mathrm{PE2}) \\ &\quad + \frac{1}{100} tr \left[\{\boldsymbol{\mu}\boldsymbol{\mu}' + R_{X_1}(0)\} \hat{C}_{100} \right], \quad (\equiv \mathrm{PE3})\end{aligned}$$

where

$$\hat{B}_{100} = \frac{1}{\sqrt{100}} \sum_{t=1}^{100} \left\{ g(\hat{\boldsymbol{\theta}}_t) - g(\boldsymbol{\theta}) \right\}$$

$$\hat{C}_{100} = \sum_{t=1}^{100} \left\{ g(\hat{\boldsymbol{\theta}}_t) - g(\boldsymbol{\theta}) \right\}^2$$

In Figure 7.9 we plotted the graph of PE2 + PE3 *for* $\alpha = -0.8(0.2)0.8$, $\beta = -0.8(0.2)0.8$, $\mu_{X_1} - \mu_{X_2} = 0.3$. *We observe that if* $\beta = -0.2$, *and if* $\alpha \nearrow 1$, *then* $\widehat{\mathrm{PE}}$ *increases.*

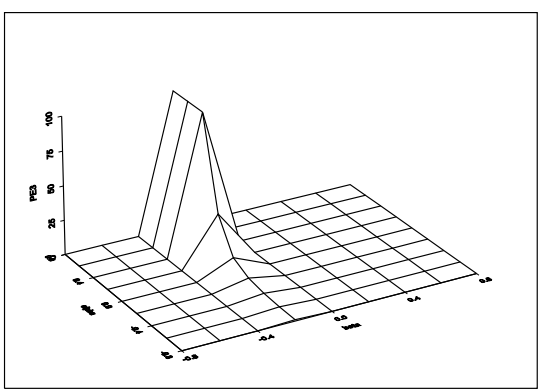


Figure 7.9 *PE2+PE3*

Next, we discuss the problem of asymptotic efficiency for the class of estimators $\hat{g}$. First, we compare the asymptotic variance of $\hat{g}$ under non-Gaussianity with that under Gaussianity. Second, we discuss the asymptotic efficiency of $\hat{g}$ when the return process is Gaussian.

If $\{\mathbf{U}(t)\} \sim$ i.i.d. $N(0, \mathbf{K})$, i.e., $\{\mathbf{X}(t)\}$ is Gaussian, then we get the following corollary from Theorem 7.6.

Corollary 7.3

$$\sqrt{n}(g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})) \xrightarrow{\mathcal{L}} N\left(\mathbf{0}, \left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right) \Omega_G \left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)'\right),$$

where

$$\Omega_G = \begin{pmatrix} \Omega_1 & \mathbf{0} \\ \mathbf{0} & \tilde{\Omega}_2 \end{pmatrix},$$

and

$$\tilde{\Omega}_2 = \left\{ 2\pi \int_{-\pi}^{\pi} \{ f_{a_1 a_3}(\lambda)\overline{f_{a_2 a_4}(\lambda)} + f_{a_1 a_4}(\lambda)\overline{f_{a_2 a_3}(\lambda)} \} d\lambda \right.$$
$$\left. ; a_1, a_2, a_3, a_4 = 1, \ldots, m, a_1 \geq a_2 \text{ and } a_3 \geq a_4 \right\}, \quad (r \times r)\text{-matrix.}$$

We evaluate

$$\boldsymbol{\mu}'(V_{NG} - V_G)\boldsymbol{\mu},$$

where

$$V_{NG} = \left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right) \Omega_{NG} \left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)',$$
$$V_G = \left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right) \Omega_G \left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)'$$

for various optimal portfolios and spectral structures.

Example 7.7 ***(VMA(1) model)*** *Let the return process be generated by*

$$\mathbf{X}(t) = \begin{pmatrix} 1 - xB & 0 \\ 0 & 1 - yB \end{pmatrix} \mathbf{U}(t) + \boldsymbol{\mu}, \quad (|x| < 1, |y| < 1),$$

where

$$\mathbf{U}(t) \equiv \begin{pmatrix} u_1(t) - \kappa_1 \\ u_2(t) - \kappa_2 \end{pmatrix}.$$

Here $u_i(t) \sim$ i.i.d. $Exp(1, \kappa_i)$, $(i = 1, 2)$, $Exp(1, \kappa_i)$ is the exponential distribution with mean κ_i, and B is the lag operator. For $x = 0.4$, $y = 0.6$, $\mu_1 = 0.1$, $\mu_2 = 0.3$, $\beta = 0.5$, $R_0 = 0.01$, $\kappa_1, \kappa_2 = -2.0(1.0)2.0$ in the case of g_3 $((7.57)')$ we calculated $\boldsymbol{\mu}'(V_{NG} - V_G)\boldsymbol{\mu}$. From Table 7.4 it is seen that, for some values of κ_1 and κ_2, $\boldsymbol{\mu}'(V_{NG} - V_G)\boldsymbol{\mu} < 0$.

Table 7.4 *VMA(1) model (the case of g_3)*

$\kappa_1 \backslash \kappa_2$	-2.0	-1.0	0	1.0	2.0
-2.0	0.00589	0.05738	0.00000	-0.00769	-0.00224
-1.0	0.00812	0.05961	0.00000	-0.00547	-0.00001
0.0	0.00000	0.00000	0.00000	0.00000	0.00000
1.0	0.00380	0.05529	0.00000	-0.00978	-0.00433
2.0	0.00535	0.05684	0.00000	-0.00823	-0.00278

Example 7.8 ***(VAR(1) model)*** *Let the return process be generated by*

$$\begin{pmatrix} 1 - xB & 0 \\ 0 & 1 - yB \end{pmatrix} \mathbf{X}(t) = \mathbf{U}(t) + \boldsymbol{\mu}$$

where $\{\mathbf{U}(t)\}$ is the same process as in Example 7.7. For $x = 0.4$, $y = 0.6$, $\mu_1 = 0.1$, $\mu_2 = .3$, $\beta = 0.5$, $R_0 = 0.01$, $\kappa_1, \kappa_2 = -2.0(1.0)2.0$ in the cases of g_1 ($(7.54)'$) and g_3 we calculated $\boldsymbol{\mu}'(V_{NG} - V_G)\boldsymbol{\mu}$.

Table 7.5 *VAR(1) model (the case of g_1)*

$\kappa_1 \backslash \kappa_2$	-2.0	-1.0	0	1.0	2.0
-2.0	0.06708	0.04353	0.00000	0.02037	0.02800
-1.0	0.03903	0.10637	0.00000	0.01853	-0.00009
0.0	0.00000	0.00000	0.00000	0.00000	0.00000
1.0	0.02328	0.04374	0.00000	-0.04410	-0.01584
2.0	0.03171	0.00641	0.00000	-0.01675	-0.00736

Table 7.6 *VAR(1) model (the case of g_3)*

$\kappa_1 \backslash \kappa_2$	-2.0	-1.0	0	1.0	2.0
-2.0	0.01909	0.15941	0.00000	-0.10870	-0.01442
-1.0	0.02488	0.16520	0.00000	-0.10291	-0.00863
0.0	0.00000	0.00000	0.00000	0.00000	0.00000
1.0	0.01243	0.15274	0.00000	-0.11537	-0.02109
2.0	0.01754	0.15785	0.00000	-0.11026	-0.01598

From Tables 7.5 and 7.6 we can see that, for some values of κ_1 *and* κ_2, $\boldsymbol{\mu}'(V_{NG} - V_G)\boldsymbol{\mu} < 0$.

The above examples illuminate an interesting feature of Gaussian and non-Gaussian asymptotics of $g(\hat{\boldsymbol{\theta}})$.

Next we discuss the asymptotic Gaussian efficiency of $g(\hat{\boldsymbol{\theta}})$. Fundamental results concerning the asymptotic efficiency of sample autocovariance matrices of vector Gaussian processes were obtained by Kakizawa (1999). He compared the asymptotic variance (AV) of sample autocovariance matrices with the inverse of the corresponding Fisher information matrix $\mathcal{F}^{-1}$, and gave the condition for the asymptotic efficiency (i.e., condition for AV $= \mathcal{F}^{-1}$). Based on this we will discuss the asymptotic efficiency of $\hat{g}$.

Suppose that $\{X(t)\}$ is a zero-mean Gaussian m-vector stationary process with spectral density matrix $\mathbf{f}(\lambda)$, and satisfies the following assumptions.

Assumption 7.5

(i) $\mathbf{f}(\lambda)$ *is parameterized by* $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)' \in \mathcal{H} \subset \mathbf{R}^{\Pi}$ *i.e.,* $\mathbf{f}_{\boldsymbol{\eta}} = \mathbf{f}_{\boldsymbol{\eta}}(\lambda)$.

(ii) For $\mathbf{A}^{(j)}(l) \equiv \int_{-\pi}^{\pi} \partial \mathbf{f}_{\boldsymbol{\eta}}(\lambda)/\partial\eta_j d\lambda$, $j = 1, \ldots, q$, $l \in \mathbf{Z}$, *it holds that* $\sum_{l=-\infty}^{\infty} \|\mathbf{A}^{(j)}(l)\| < \infty$.

(iii) $q \geq m(m+1)/2$.

Assumption 7.6 *There exists a positive constant* c *(independent of* λ*) such that* $\mathbf{f}_{\boldsymbol{\eta}}(\lambda) - c\mathbf{I}_m$ *is positive semi-definite, where* $\mathbf{I}_m$ *is an* $m \times m$ *identity matrix.*

The limit of averaged Fisher information matrix is given by

$$\mathcal{F}(\boldsymbol{\eta}) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \Delta(\lambda)^* [\{\mathbf{f}_{\boldsymbol{\eta}}(\lambda)^{-1}\}' \otimes \mathbf{f}_{\boldsymbol{\eta}}(\lambda)^{-1}] \Delta(\lambda) d\lambda$$

where

$$\Delta(\lambda) = (\mathrm{vec}\{\partial \mathbf{f}_{\boldsymbol{\eta}}(\lambda)/\partial\eta_1\}, \ldots, \mathrm{vec}\{\partial \mathbf{f}_{\boldsymbol{\eta}}(\lambda)/\partial\eta_q\}) \quad (m^2 \times q)\text{-matrix.}$$

Assumption 7.7 *The matrix* $\mathcal{F}(\boldsymbol{\eta})$ *is positive definite.*

We introduce an $m^2 \times m(m+1)/2$ matrix

$$\Phi = (\mathrm{vec}\{\psi_{11}\}, \ldots, \mathrm{vec}\{\psi_{m1}\}, \mathrm{vec}\{\psi_{22}\}, \ldots, \mathrm{vec}\{\psi_{mm}\}),$$

where $\psi_{ab} = (\mathbf{e}_a \mathbf{e}_b' + \mathbf{e}_b \mathbf{e}_a')/2$ and $\mathbf{e}_a = (1, \ldots, 1)'$ ($m \times 1$-vector). Then we have the following theorem.

Theorem 7.7 *Under Assumptions 7.3-7.7,* $g(\hat{\boldsymbol{\theta}})$ *is asymptotically efficient if and only if there exists a matrix* C *(independent of* λ*) such that*

$$\{\mathbf{f}_{\boldsymbol{\eta}}(\lambda)' \otimes \mathbf{f}_{\boldsymbol{\eta}}(\lambda)\}\Phi = (\mathrm{vec}\{\partial \mathbf{f}_{\boldsymbol{\eta}}(\lambda)/\partial\eta_1\}, \ldots, \mathrm{vec}\{\partial \mathbf{f}_{\boldsymbol{\eta}}(\lambda)/\partial\eta_q\})C. \qquad (7.65)$$

PROOF Recall Corollary 7.3. In the asymptotic variance matrix of $\sqrt{n}\{g(\hat{\boldsymbol{\theta}})-g(\boldsymbol{\theta})\}$, the transformation matrix $(\partial g/\partial \boldsymbol{\theta}')$ does not depend on the goodness of estimation. Thus, if the matrix Ω_G is minimized in the sense of matrix, the estimator $g(\hat{\boldsymbol{\theta}})$ becomes asymptotically efficient. Regarding $\tilde{\Omega}_2$-part, Kakizawa (1999) gave a necessary and sufficient condition for $\tilde{\Omega}_2$ to attain the lower bound matrix $\mathcal{F}(\boldsymbol{\eta})^{-1}$, which is given by the condition (7.65). Regarding Ω_1-part, it is known that $\Omega_1 = 2\pi \mathbf{f}(0)$ is equal to the asymptotic variance of the BLUE estimator of $\boldsymbol{\mu}$ (e.g., Hannan (1970, Chap.VII)). Hence, Ω_1 attains the lower bound matrix, which completes the proof. □

Theorem 7.7 implies that if (7.65) is not satisfied, the estimator $g(\hat{\boldsymbol{\theta}})$ is not asymptotically efficient. This is a strong warning to use the ordinary portfolio estimators for even Gaussian dependent returns. The interpretation of (7.65) is difficult. But, Kakizawa (1999) showed that (7.65) is satisfied by vector AR(p) models with coefficients $\boldsymbol{\eta}$. The following are examples, which do not satisfy (7.65).

Example 7.9 ***(VARMA(p_1, p_2) process)*** *Consider the $m \times m$ spectral density matrix of m-vector ARMA(p_1, p_2) process,*

$$\mathbf{f}(\lambda) = \frac{1}{2\pi}\Theta\{\exp(i\lambda)\}^{-1}\Psi\{\exp(i\lambda)\}\Sigma\Psi\{\exp(i\lambda)\}^{*}\Theta\{\exp(i\lambda)\}^{-1*} \tag{7.66}$$

where $\Psi(z) = I_m - \Psi_1 z - \cdots - \Psi_{p_2} z^{p_2}$, and $\Theta(z) = I_m - \Theta_1 z - \cdots - \Theta_{p_1} z^{p_1}$ satisfy $\det \Psi(z) \neq 0$, $\det \Theta(z) \neq 0$ for all $|z| \leq 1$. From Kakizawa (1999) it follows that (7.66) does not satisfy (7.65), if $p_1 < p_2$, hence $g(\hat{\boldsymbol{\theta}})$ is not asymptotically efficient if $p_1 < p_2$.

Example 7.10 ***(An exponential model)*** *Consider the $m \times m$ spectral density matrix of exponential type,*

$$\mathbf{f}(\lambda) = \exp\left\{\sum_{j\neq 0} A_j \cos(j\lambda)\right\} \tag{7.67}$$

where A_j's are $m \times m$-matrices, and $\exp\{\cdot\}$ is the matrix exponential (for the definition, see Bellman (1960, p.169)). Since (7.67) does not satisfy (7.65), $g(\hat{\boldsymbol{\theta}})$ is not asymptotically efficient.

In view of the above we should be careful when we use the usual portfolio estimators $g(\hat{\boldsymbol{\theta}})$ even if the return process is Gaussian.

Next we are interested in the degree of inefficiency of $g(\hat{\boldsymbol{\theta}})$. Note that

$$\begin{aligned} V_G - &\text{ minimum variance of } g(\hat{\boldsymbol{\theta}}) \\ &= \left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)\begin{pmatrix}\Omega_1 & \mathbf{0}\\ \mathbf{0} & \tilde{\Omega}_2\end{pmatrix}\left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)' - \left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)\begin{pmatrix}\Omega_1 & \mathbf{0}\\ \mathbf{0} & \mathcal{F}(\boldsymbol{\eta})^{-1}\end{pmatrix}\left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)' \\ &= \left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & (\tilde{\Omega}_2 - \mathcal{F}(\boldsymbol{\eta})^{-1})\end{pmatrix}\left(\frac{\partial g}{\partial \boldsymbol{\theta}'}\right)'. \end{aligned}$$

In what follows we numerically evaluate

$$\text{INE} \equiv \det\left[\tilde{\Omega}_2 - \mathcal{F}(\boldsymbol{\eta})^{-1}\right]$$

for various spectra.

Model I (VMA(1) model). Let the return process be generated by

$$\mathbf{X}(t) = \begin{pmatrix} 1-\eta_1 B & 0 \\ 0 & 1-\eta_1 B \end{pmatrix} \mathbf{U}(t), \quad \mathbf{U}(t) \sim \text{ i.i.d. } N\left(\mathbf{0}, \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}\right).$$

Figure 7.10 shows the graph of INE = INE(VMA(1)) for $\eta_1 = -0.8(0.2)0.8$. We can see that, as $|\eta_1|$ tends to 1, INE increases.

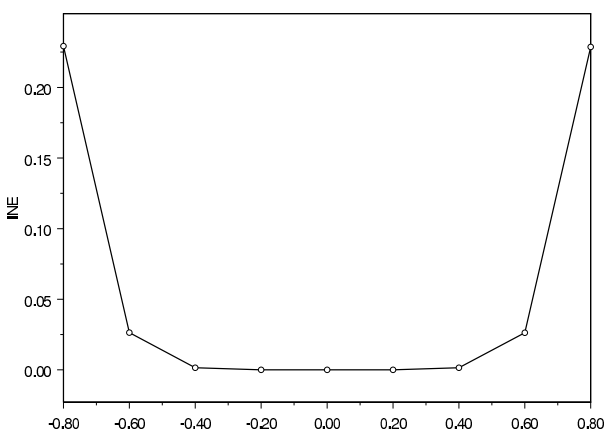


Figure 7.10 *Model I:* $\eta_1 = -0.8(0.2)0.8$

Model II (VARMA(1,2) model). Let the return process be generated by

$$\begin{pmatrix} 1-\eta_1 B & 0 \\ 0 & 1-\eta_1 B \end{pmatrix} \mathbf{X}(t) = \begin{pmatrix} (1-\eta_2 B)(1-\eta_3 B) & 0 \\ 0 & (1-\eta_2 B)(1-\eta_3 B) \end{pmatrix} \mathbf{U}(t)$$

$$\mathbf{U}(t) \sim \text{ i.i.d. } N\left(\mathbf{0}, \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}\right). \tag{7.68}$$

In Figure 7.11 we plotted the graph of INE = INE(VARMA(1, 2)) for $\eta_1 = -0.8(0.2)0.8, \eta_2 = 0.01, \eta_3 = 0.5$. We can see that if $\eta_1 \searrow -1$, INE becomes quite large.

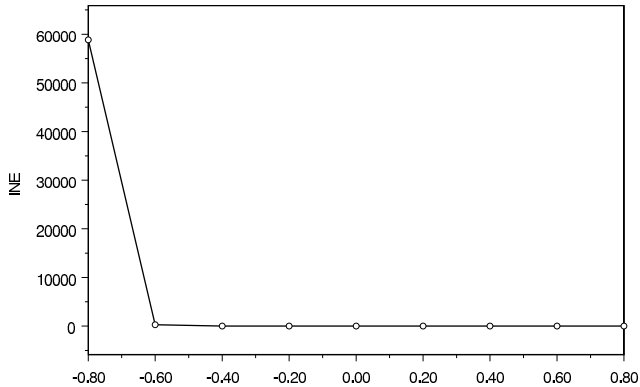


Figure 7.11 *Model II:* $\eta_1 = -0.8(0.2)0.8, \eta_2 = 0.01, \eta_3 = 0.5$

Model III (VARMA(1,2) model). Let the model be generated by (7.68) with $\eta_1 = 0.01$ and $\eta_2 = 0.5$. In Figure 7.12 we plotted the graph of INE = INE(VARMA$(1, 2)$) for $\eta_3 = -0.8(0.2)0.8$. We can see that as $\eta_3 \nearrow 1$, INE increases.

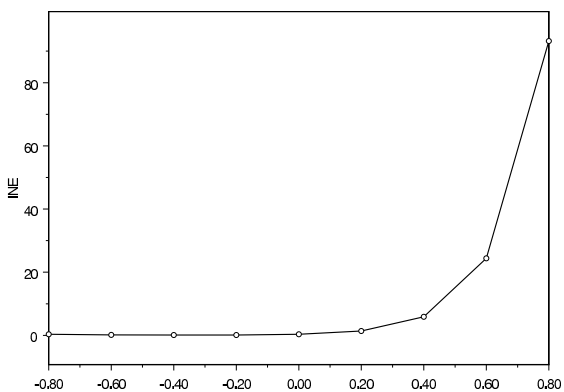


Figure 7.12 *Model III:* $\eta_1 = 0.01, \eta_2 = 0.5, \eta_3 = -0.8(0.2)0.8$

Summarizing the above we observe that

(i) For the VMA(1) model, INE increases as the absolute value of the MA coefficient η_1 tends to 1.

(ii) For the VARMA(1,2) model with the MA coefficient $\eta_2 \approx 0$, INE increases as the AR coefficient η_1 tends to -1.

(iii) For the VARMA(1,2) model with the AR coefficient $\eta_1 \approx 0$, INE increases as the MA coefficient η_3 tends to 1.

Although we just examined a few examples of dependent returns, the above studies show inefficiency of the usual portfolio estimators. Also it should be noted that the degree of inefficiency becomes quite large if some parameters tend to boundary values.

Finally we discuss construction of an efficient estimator when $\{\mathbf{X}(t)\}$ is non-Gaussian. Suppose that $\{\mathbf{X}(t)\}$ is generated by

$$\mathbf{X}(t) = \mathbf{F}_\theta(\mathbf{X}(t-1), \ldots, \mathbf{X}(t-p_1), \boldsymbol{\mu}) + \mathbf{H}_\theta(\mathbf{X}(t-1), \ldots, \mathbf{X}(t-p_2))\mathbf{U}(t), \tag{7.69}$$

where $\mathbf{F}_\theta : \mathbf{R}^{m(p_1+1)} \to \mathbf{R}^m$ is a vector-valued measurable function, $H_\theta : \mathbf{R}^{mp_2} \to \mathbf{R}^m \times \mathbf{R}^m$ is a positive definite matrix-valued measurable function, and $\{\mathbf{U}(t) = (u_1(t), \ldots, u_m(t))'\}$ is a sequence of i.i.d. random variables with $E\mathbf{U}(t) = \mathbf{0}$, $E|\mathbf{U}(t)| < \infty$ and $\mathbf{U}(t)$ is independent of $\{\mathbf{X}(s), s < t\}$. Henceforth, without loss of generality we assume $p_1 + 1 = p_2 (= p)$, and Assumptions 6.4 and 6.5 in Section 6.2. Then it was shown in Theorem 6.8 that the MLE $\hat{\theta}_{\mathrm{ML}}$ of θ is asymptotically efficient. Hence we can construct asymptotically efficient estimators of optimal portfolios when the return is generated by (7.69).

7.4 VaR Problems

Recently, Value-at-Risk (VaR) has become a widely used measure of market risk in risk management. The measure can be used by financial institutions to assess their risks or by a regulatory committee to set margin requirements. In either case, VaR is used to ensure that the financial institutions can still be in business after a catastrophic event. There are several methods for calculating VaR, namely, quantile estimation, extreme value theory and econometric approaches etc. (cf. Tsay (2002)). Here we focus on the econometric approach which contains the RiskMetrics by J. P. Morgan (1996), perhaps the most celebrated methods for calculating VaR. As RiskMetrics assumes IGARCH(1,1) model for the return process, the econometric approach first fits a suitable time series model to the return process, and then evaluates VaR through the quantile function.

In this section we assume that the return process follows an ARCH(k) process which was introduced by Engle (1982) and is known to capture the behavior of financial time series. One of the reasons to take the ARCH process is that it has the well-tailored residual empirical process (REP) theory. The asymptotics of the REP in a time series setting have been developed by many authors. For example, Boldin (1982) for the AR(k) model, Bai (1994) for the ARMA(k, l) model, Boldin (1998, 2000) for ARCH(1), Horváth et al. (2001) for the squared ARCH(k) and Lee and Taniguchi (2005) for ARCH(k).

As described in Huschens (1998), the determination of VaR is, from a statistical point of view, not a simple computation but a statistical point estimation of an unknown parameter in the underlying model. Point estimation for VaR

should be complemented by interval estimates because sampling errors are not negligible for realistic volatilities. Here we note that the tolerance region (cf. Guttman (1970)) is more appropriate than the confidence interval to assess the validity of VaRs. Since VaRs are calculated to prepare for some catastrophic event, they should serve to determine whether or not we are actually in such a catastrophe. By using the tolerance region argument, we can construct a VaR which does alarm us in, at least, specified proportion of trials. On the other hand, commonly used VaRs, which are point estimates, may fail to alarm us, for example, in half of the trials if the innovation density is symmetric. This is why we propose a new VaR with consideration for estimation errors. Also, it is worthwhile to note that our VaR under ARCH returns will exhibit the essential difference from those under ARMA returns.

In this section, based on Taniai and Taniguchi (2007) we address a reconsideration for commonly used econometric approaches, and also propose a feasible VaR based on REP.

Value-at-Risk (VaR) is an economic terminology which is defined as the worst loss of a financial position over a given time horizon at a given risk probability. Here we make this concept more actual, and reconsider commonly used approaches, especially in view of estimation of unknown parameters.

Suppose that we are in "long position", e.g., we are holding some share of a stock to sell, so that the risk of our interest occurs with smaller value of returns, and we will set the time horizon to be 1. Suppose further, the return process $\{Y_t\}$ follows the ARCH(k) model characterized by

$$Y_t = \sigma_t \varepsilon_t, \quad \sigma_t^2 = \beta_0 + \sum_{j=1}^{k} \beta_j Y_{t-j}^2, \tag{7.70}$$

where ε_t's are i.i.d. random variables with

$$E[\varepsilon_t] = 0, \quad E[\varepsilon_t^2] = 1,$$

and $\beta := (\beta_0, \beta_1, \ldots, \beta_k)' \in \Theta \subset \mathrm{R}^{k+1}$ is an unknown parameter vector satisfying

$$\beta_0 > 0, \ \beta_j \geq 0, \quad 1 \leq j \leq k.$$

Furthermore, we assume that $\beta_1 + \cdots + \beta_k < 1$ for strict stationarity (cf. Giraitis et al. (2000)) and $E[\varepsilon_t^4] < \infty$ for use later. Although we can model also the conditional mean of the process by introducing the ARMA form in the first equation of (7.70), we stick to this simple setting since, as it will be noted, ARMA parameters will not play an important role in our result.

To define the VaR, we need the 1-step ahead forecast of volatility σ_t. For this purpose, suppose that an observed stretch $\{y_t\}_{t=1}^n$ is available. Let $\hat{\beta} := (\hat{\beta}_0, \ldots, \hat{\beta}_k)'$ denote an estimator of β satisfying

$$\sqrt{n}(\hat{\beta} - \beta) = O_p(1). \tag{7.71}$$

and define the residuals as

$$\hat{\varepsilon}_t := \frac{y_t}{\hat{\sigma}_t}, \qquad k < t \le n,$$

where

$$\hat{\sigma}_t^2 := \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j y_{t-j}^2.$$

If we follow the commonly used approach for VaR, VaR (for next period from time n) will be computed as

$$\mathrm{VaR}_p^{(C)} := F^{\leftarrow}(p)\hat{\sigma}_{n+1}, \qquad \text{(Commonly used approach)} \tag{7.72}$$

where $F^{\leftarrow}(p)$ denotes the p-th quantile:

$$F^{\leftarrow}(p) := \inf\{x; F(x) \ge p\}, \qquad 0 < p < 1,$$

of the unknown distribution function $F(\cdot)$ of ε_t. Note that if we take IGARCH (1,1) setting and if F follows the standard normal, we see that this is the method known as RiskMetrics. The above computation is based on the equation:

$$P\left\{\frac{Y_{n+1}}{\hat{\sigma}_{n+1}} \le F^{\leftarrow}(p)\right\} = p, \tag{7.73}$$

that is, the idea that $\{Y_t/\hat{\sigma}_t\}$ must follow the distribution function F. But this is not true because $\hat{\sigma}_t$ includes estimators. This confusion between true innovations and empirical residuals can often be found in the econometric literature. To rectify this, first we may want to use the empirical distribution $\mathbb{F}_n(\cdot)$ of $\{\varepsilon_t\}_{t=1}^n$:

$$\mathbb{F}_n(x) := \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{\{\varepsilon_t \le x\}},$$

since we do not know the true distribution $F(\cdot)$. However, because the innovations $\{\varepsilon_t\}$ are unobservable, we should estimate them with the residuals $\{\hat{\varepsilon}_t\}$. Hence we define $\hat{\mathbb{F}}_n(\cdot)$ as

$$\hat{\mathbb{F}}_n(x) := \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_{\{\hat{\varepsilon}_t \le x\}},$$

to state, instead of (7.73), the valid equation:

$$P\left\{\frac{Y_{n+1}}{\hat{\sigma}_{n+1}} \le \hat{\mathbb{F}}_n^{\leftarrow}(p)\right\} = p. \tag{7.74}$$

Again we emphasize that the feasible VaR must be computed from this $\hat{\mathbb{F}}_n(\cdot)$, not from $F(\cdot)$. Noting that $\hat{\mathbb{F}}_n^{\leftarrow}(p)$ is a random variable which depends on the sampling $\{Y_t\} = \{y_t\}$, we will see in a later section that the residual quantile process $\sqrt{n}\{\hat{\mathbb{F}}_n^{\leftarrow}(p) - F^{\leftarrow}(p)\}$ converges to some Gaussian process with ARCH

parameter depending terms. Hence we cannot ignore the difference between commonly used approaches and the one which takes estimation effect into consideration. This motivates the problems and makes a strong warning to the usual approaches. Although the discussion about the confidence intervals for VaR can be found in Huschens (1998), our approach essentially differs from it.

Let us suppose that we obtain the distribution of $\hat{\mathbb{F}}_n^{\leftarrow}(p)$, denoted by $\Psi_{\hat{\varepsilon},p}(\cdot)$, and that we are in long position. Then, for $0 < \alpha < 0.5$, we can compute the upper $(1-\alpha)$-confidence bound $\Psi_{\hat{\varepsilon},p}^{\leftarrow}(1-\alpha)$ characterized by the following equation:

$$P\{\hat{\mathbb{F}}_n^{\leftarrow}(p) \leq \Psi_{\hat{\varepsilon},p}^{\leftarrow}(1-\alpha)\} = 1-\alpha.$$

Then we can claim that, with $(1-\alpha)$-confidence,

$$P\{Y_{n+1} \geq \hat{\sigma}_{n+1}\Psi_{\hat{\varepsilon},p}^{\leftarrow}(1-\alpha)\} \leq 1-p.$$

The above equation corresponds to the p-content tolerance region at confidence level $(1-\alpha)$, and we give the illustrative interpretation of it as follows: Suppose that we decide to sell the stock whenever the (standardized) return observations fall below the VaR, i.e., to give away when the situation becomes unacceptably risky. The commonly used approaches, (7.72), suggest that we can ignore the returns which is greater than $F^{\leftarrow}(p)$. However, considering the sampling errors, the returns greater than $F^{\leftarrow}(p)$ but also smaller $\hat{\sigma}_t\Psi_{\hat{\varepsilon},p}^{\leftarrow}(1-\alpha)$ can be a signal for selling. So we cannot make a certain decision from these. On the other hand, with high confidence, we definitely can ignore the returns greater than $\hat{\sigma}_t\Psi_{\hat{\varepsilon},p}^{\leftarrow}(1-\alpha)$. In other words, we can be $(1-\alpha)$-confident never to go into the risks which are rarer than probability p, while commonly used approaches may take us there in half of the trials. Hence, it is reasonable to define the feasible VaR (for long position) as

$$\mathrm{VaR}_p^{(REP)} = \hat{\sigma}_{n+1}\Psi_{\hat{\varepsilon},p}^{\leftarrow}(1-\alpha). \tag{7.75}$$

We call this as "VaR based on Residual Empirical Process (REP)". Note that the perturbation of $\hat{\sigma}_{n+1}$ is already considered here through equation (7.74). Also, it should be noted that, for fixed p, $\mathrm{VaR}_p^{(C)}$ based on the commonly used approaches is smaller than $\mathrm{VaR}_p^{(REP)}$ even when the innovation distribution function is known. This means that the commonly used approaches underestimate the risk probability in our sense, and may serve as a signal for riskier event than it says.

Now, we study the ARCH effect on the residual quantile process and the problem of VaR. We also construct the confidence intervals for VaR based on REP. Initially we assume

Assumption 7.8

(i) The distribution function $F(\cdot)$ has the probability density function $f(x)$,

(ii) $f(\cdot)$ *is positive and differentiable,*

Assumption 7.9

(i) $f(\cdot)$ *is "increasing and convex" on* $(-\infty, -M]$ *and "decreasing and convex" on* $[M, \infty)$ *for some positive number* M*,*

(ii) $|x|f(x) \to 0$ *as* $x \to \infty$ *and* $\sup_x x^2|f'(x)| < \infty$.

Under Assumptions 7.8 and 7.9, Lee and Taniguchi (2005) showed that the following result holds.

$$\sqrt{n}\{\hat{\mathbb{F}}_n(x) - F(x)\} = \mathbb{G}_{n,F}(x) + \frac{xf(x)}{2}A + \eta_n(x), \tag{7.76}$$

where

$$\mathbb{G}_{n,F}(x) := \sqrt{n}\{\mathbb{F}_n(x) - F(x)\}, \qquad A := \sqrt{n}(\hat{\beta} - \beta)'\tau,$$

and $\eta_n(x) = o_p(1)$ with

$$\tau := (\tau_0, \ldots, \tau_k)', \quad \tau_0 := E\left[\frac{1}{\sigma_t^2}\right], \quad \tau_j := E\left[\frac{Y_{t-j}^2}{\sigma_t^2}\right], \quad 1 \le j \le k.$$

The presence of the second term in (7.76), which does not appear for ARMA setting (cf. Bai (1994)), will play a prominent role for our $\mathrm{VaR}^{(REP)}$. This dependence of ARCH parameters was also known in the context of squared residuals by Horváth et al. (2001).

Henceforth, as the estimator $\hat{\beta}$, we use the conditional least squares (CLS) estimator $\hat{\beta}^{CL}$, i.e.,

$$\begin{aligned}\hat{\beta}^{CL} &:= \mathrm{argmin}_\beta \sum_{t=1}^{n}\{y_t^2 - E[Y_t^2 \mid \mathscr{F}_{t-1}]\}^2 \\ &= \mathrm{argmin}_\beta \sum_{t=1}^{n}\{y_t^2 - (\beta_0 + \sum_{j=1}^{k}\beta_j y_{t-j}^2)\}^2.\end{aligned}$$

By using the standard argument of asymptotic theory (cf. Chapter 6), it can be shown that $\hat{\beta}^{CL}$ admits the following representation:

$$\hat{\beta}_j^{CL} - \beta_j = \frac{1}{n}\sum_{t=1}^{n} U_{j,t}(\varepsilon_t^2 - 1) + o_p(n^{-1/2}), \qquad 0 \le j \le k \tag{7.77}$$

where $U_{j,t}$ are the j-th elements of $\sigma_t^2 \mathcal{U}^{-1} W_{t-1}$, with

$$W_t := (1, Y_t^2, \ldots, Y_{t-k+1}^2)' \qquad \text{and} \qquad \mathcal{U} := E[W_{t-1}W_{t-1}'].$$

Note also that we know the $\sqrt{n}$-consistency (7.71) and asymptotic normality of $\hat{\beta}^{CL}$ by Tjøstheim (1986), i.e., under suitable conditions,

$$\sqrt{n}(\hat{\beta}^{CL} - \beta) \xrightarrow{\mathcal{L}} N(0, \mathcal{U}^{-1}\mathcal{R}\mathcal{U}^{-1}), \tag{7.78}$$

where $\mathcal{R} := 2E[\sigma_t^4 W_{t-1} W_{t-1}']$. For the remainder of this section, we will write $\hat{\beta}^{CL}$ as $\hat{\beta}$ for simplicity.

From (7.76) and the functional delta-method (cf. Gill (1989) or van der Vaart (1998)), we have the following result.

Theorem 7.8 *Suppose that Assumptions 7.8 and 7.9 hold, and let* $\xi_p := F^{\leftarrow}(p)$. *Then the following statements hold true.*

(i)

$$\sqrt{n}\{\hat{\mathbb{F}}_n^{\leftarrow}(p) - \xi_p\} = \frac{-\{\mathbb{G}_{n,F}(\xi_p) + \frac{1}{2}\xi_p f(\xi_p) A\}}{f(\xi_p)} + o_p(1); \tag{7.79}$$

(ii)

$$\frac{\sqrt{n}\{\hat{\mathbb{F}}_n^{\leftarrow}(p) - \xi_p\}}{\sigma_{Q(\hat{\varepsilon},p)}} \xrightarrow{\mathcal{L}} N(0,1), \tag{7.80}$$

where

$$\sigma^2_{Q(\hat{\varepsilon},p)} := \frac{1}{f(\xi_p)^2}\left[p(1-p) + \xi_p f(\xi_p)\left\{\int_{-\infty}^{\xi_p} u^2 f(u)du - p\right\} \cdot \tau' \mathcal{U}^{-1}\mathcal{V} + \frac{1}{4}\xi_p^2 f(\xi_p)^2 \tau' \mathcal{U}^{-1}\mathcal{R}\mathcal{U}^{-1}\tau\right], \tag{7.81}$$

with $\mathcal{V} := E[\sigma_t^2 W_{t-1}]$.

PROOF (i) Since $\hat{\mathbb{F}}_n^{\leftarrow}(p)$ is a functional of the empirical distribution function $\hat{\mathbb{F}}_n(x)$, we can apply the functional Delta method (see Theorem 20.8 and Corollary 21.5 of van der Vaart (1998)) to $\sqrt{n}\{\hat{\mathbb{F}}_n^{\leftarrow}(p) - \xi_p\}$. Note that the existence of its distributional limit is provided by the statement (ii), which will be proved below. Then we obtain (7.79) as follows:

$$\begin{aligned}\sqrt{n}\{\hat{\mathbb{F}}_n^{\leftarrow}(p) - \xi_p\} &= \frac{-1}{f(\xi_p)}[\sqrt{n}\{\hat{\mathbb{F}}_n(\xi_p) - F(\xi_p)\}] + o_p(1) \\ &= \frac{-1}{f(\xi_p)}\{\mathbb{G}_{n,F}(\xi_p) + \frac{1}{2}\xi_p f(\xi_p) A\} + o_p(1), \quad \text{(by (7.76))}.\end{aligned}$$

(ii) The asymptotic normality follows from (7.76), (7.77) and (7.78). Also the first and third terms in the bracket of r.h.s. at (7.81) are, respectively, the asymptotic variances of $\mathbb{G}_{n,F}(\xi_p)$ and $\frac{1}{2}\xi_p f(\xi_p) A$. Thus we evaluate the asymptotic covariance between $\mathbb{G}_{n,F}(\xi_p)$ and $\frac{1}{2}\xi_p f(\xi_p) A$. From (7.77) it follows that

$$\begin{aligned}&\mathrm{E}[\sqrt{n}\{\mathbb{F}_n(\xi_p) - p\}\xi_p f(\xi_p)\sqrt{n}(\hat{\beta}^{CL} - \beta)'\tau] \\ &\quad = \xi_p f(\xi_p)\mathrm{E}[\sqrt{n}(\frac{1}{n}\sum_{s=1}^{n} \mathbb{1}_{\{\varepsilon_s \le \xi_p\}} - p)\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \sigma_t^2 W_{t-1}' \mathcal{U}^{-1}\tau(\varepsilon_t^2 - 1)].\end{aligned} \tag{7.82}$$

The above terms corresponding to the sum $\sum_{s\neq t}$ are zero because ε_t's are mutually independent, and because $\sigma_t^2 W_{t-1}$ is $\mathscr{F}_{t-1}$-measurable. Hence we have only to evaluate the sum $\sum_s \sum_t$ in (7.82) for the case of $s = t$. With strict stationarity of $\{y_t^2\}$, (7.82) becomes

$$\xi_p f(\xi_p) \mathrm{E}[\sigma_t^2 W_{t-1}' \mathcal{U}^{-1} \tau] \mathrm{E}\left[\frac{1}{n}\sum_{t=1}^{n} \left(\mathbb{1}_{\{\varepsilon_t \le \xi_p\}} - p\right)(\varepsilon_t^2 - 1)\right],$$

which, together with $\mathrm{E}[\varepsilon_t^2] = 1$, implies (ii). □

Now, in order to construct the $\mathrm{VaR}_p^{(REP)}$ from data, we make the sample version of quantities appeared in Theorem 7.8. Since τ, $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{R}$ contain the unknown parameter β, we introduce the concept of discretized estimator of β.

Definition 7.4 *For any sequence of estimators $\hat{\theta}_n$, the* ***discretized estimator*** *$\bar{\theta}$ is defined to be the nearest vertex of $\{\theta : \theta = (i_1, i_2, \ldots, i_k)'/\sqrt{n},\ i_j : \text{integers}\}$.*

Define $\bar{\tau}$, $\bar{\mathcal{U}}$, $\bar{\mathcal{R}}$, and $\bar{\mathcal{V}}$ by

$$\bar{\tau} := (\bar{\tau}_0, \ldots \bar{\tau}_k), \quad \bar{\tau}_0 := \frac{1}{n}\sum_{t=1}^{n} \frac{1}{\bar{\sigma}_t^2}, \quad \bar{\tau}_j := \frac{1}{n}\sum_{t=1}^{n} \frac{y_{t-j}^2}{\bar{\sigma}_t^2}, \quad 1 \le j \le k,$$

$$\bar{\mathcal{U}} := \frac{1}{n}\sum_{t=1}^{n} W_{t-1} W_{t-1}', \qquad \bar{\mathcal{R}} := \frac{2}{n}\sum_{t=1}^{n} \bar{\sigma}_t^4 W_{t-1} W_{t-1}',$$

$$\bar{\mathcal{V}} := \frac{1}{n}\sum_{t=1}^{n} \bar{\sigma}_t^2 W_{t-1}, \qquad \text{with} \quad \bar{\sigma}_t^2 := \bar{\beta}_0 + \sum_{j=1}^{k} \bar{\beta}_j y_{t-j}^2,$$

i.e., we replaced the expectations by the sample averages, and $\hat{\beta}$ by its discretized version $\bar{\beta}$. If we use the estimators defined by $\hat{\sigma}_t$ instead of $\bar{\sigma}_t$, then it is difficult to show the convergence in probability. This motivates the introduction of discretized estimators. In what follows, if the convergence in probability is established when the value of the parameter is β, then we write "$\xrightarrow{P_\beta}$".

Lemma 7.1 *Under Assumptions 7.8, $(\bar{\tau},\ \bar{\mathcal{U}},\ \bar{\mathcal{R}},\ \bar{\mathcal{V}}) \xrightarrow{P_\beta} (\tau,\ \mathcal{U},\ \mathcal{R},\ \mathcal{V})$.*

PROOF In the manner of Theorem 2 in Linton (1993), we will prove

$$\bar{\tau} \xrightarrow{P_\beta} \tau. \tag{7.83}$$

Since $\{Y_{t-j}^2/\sigma_t^2\}$ is a bounded stationary ergodic process, we have

$$\frac{1}{n}\sum_{t=1}^{n} \frac{y_{t-j}^2}{\sigma_t^2} \xrightarrow{P_\beta} E\left[\frac{Y_{t-j}^2}{\sigma_t^2}\right] = \tau_j, \qquad 0 \le j \le k,$$

by the ergodic theorem. We must state this for a "discretely estimated" version, i.e., σ_t^2 in the summand replaced by $\bar{\sigma}_t^2$. But the great advantage of the discretized estimator is that, thanks to Lemma 6.8, we have only to show the result for some nonrandom sequence $\beta^{(b)}$ such that $\sqrt{n}(\beta^{(b)} - \beta)$ stays bounded. Denoting $\{\sigma_t^{(b)}\}^2 := \beta_0^{(b)} + \sum_{j=1}^k \beta_j^{(b)} y_{t-j}^2$, we can, by the definition of $\beta^{(b)}$, be sure that $1/\{\sigma_t^{(b)}\}^2$ is positive and finite for suitably large n. Hence, for suitably large n, there exists $C < \infty$ such that

$$\begin{aligned}\left| \frac{y_{t-j}^2}{\sigma_t^2} - \frac{y_{t-j}^2}{\{\sigma_t^{(b)}\}^2} \right| &= \frac{y_{t-j}^2}{\sigma_t^2} \times \frac{1}{\{\sigma_t^{(b)}\}^2} \times \left| \{\sigma_t^{(b)}\}^2 - \sigma_t^2 \right| \\ &\le \frac{y_{t-j}^2}{\sigma_t^2} \times \frac{1}{\{\sigma_t^{(b)}\}^2} \times \left(|\beta_0^{(b)} - \beta_0| + \sum_{j=1}^k |\beta_j^{(b)} - \beta_j| y_{t-j}^2 \right) \\ &\le C \times ||\beta^{(b)} - \beta||.\end{aligned}$$

Now, (7.83) follows since r.h.s. of the above tends to 0 as $n \to \infty$. The results for $\bar{\mathcal{U}}$, $\bar{\mathcal{V}}$, and $\bar{\mathcal{R}}$ can be proved similarly. □

Next, to estimate $f(\cdot)$, we need to assume that $f(\cdot)$ belongs to some class of functions described by assumptions below.

Assumption 7.10 *The density f has the finite Fisher information for scale parameters,*

$$0 < I(f) := \int \psi(x)^2 f(x) dx < \infty,$$

where $\psi(x) := xf'(x)/f(x) + 1$.

Assumption 7.11 *The score function satisfies the following conditions.*

(i) $\int \{\psi(\frac{x+m}{1+s}) - \psi(x)^2\} f(x) dx \to 0, \quad m, s \to 0;$
(ii) $\frac{1}{s} \int \psi(\frac{x+m}{1+s}) f(x) dx \to -I(f), \quad m, s \to 0;$
(iii) $\frac{1}{m} \int \psi(\frac{x+m}{1+s}) f(x) dx \to 0, \quad m, s \to 0.$

Assumption 7.12 *The innovation density f satisfies*

(i) $\int x^4 f(x) dx < \infty;$
(ii) $\int \psi(x)^4 f(x) dx < \infty.$

Denote Y_0 for the initial value, and $f_0(Y_0; \cdot)$ for the unconditional density of Y_0.

Assumption 7.13 *Let $\beta^{(n)} := \beta + h/\sqrt{n}$, and, for $^\forall h \in \mathbb{R}^{k+1}$ and $^\forall \beta \in \Theta$, $f_0(Y_0; \cdot)$ satisfies*

$$f_0(Y_0; \beta^{(n)}) \xrightarrow{P_\beta} f_0(Y_0; \beta), \qquad n \to \infty.$$

Note that these assumptions are posed to ensure the locally asymptotic normality of the model, which is proved by Linton (1993). Other descriptions can also be found in e.g., Example 4.4 of Drost et al. (1997).

Consider to estimate f and ξ_p by

$$\bar{f}_{B_n}(x) := \frac{1}{n}\sum_{t=1}^{n} \phi_{B_n}(x - \bar{\varepsilon}_t), \tag{7.84}$$

and

$$\bar{\xi}_p := \bar{F}^{-1}_{B_n}(p), \quad \text{with} \quad \bar{F}_{B_n}(x) := \int_{-\infty}^{x} \bar{f}_{B_n}(u)du,$$

where $B_n > 0$ is the bandwidth and

$$\phi_{B_n}(x) := \frac{1}{\sqrt{2\pi}B_n} \exp\left(-\frac{x^2}{2B_n^2}\right), \qquad \bar{\varepsilon}_t := \frac{y_t}{\bar{\sigma}_t}.$$

This bandwidth is required to satisfy the following assumption.

Assumption 7.14 *The bandwidth B_n satisfies that, as $n \to \infty$,*

$$B_n \to 0 \quad \textit{and} \quad nB_n \to \infty. \tag{7.85}$$

Denoting $\sigma^2_{Q(\hat{\varepsilon},p)} = \sigma^2_{Q(\hat{\varepsilon},p)}(\xi_p, f, \tau, \mathcal{U}, \mathcal{V}, \mathcal{R})$, the following statement holds true.

Theorem 7.9 *Suppose that Assumptions 7.8, 7.10-7.14 hold, and let*

$$\bar{\sigma}^2_{Q(\hat{\varepsilon},p)} := \sigma^2_{Q(\hat{\varepsilon},p)}(\bar{\xi}_p, \bar{f}_{B_n}, \bar{\tau}, \bar{\mathcal{U}}, \bar{\mathcal{V}}, \bar{\mathcal{R}})$$

Then

$$\bar{\sigma}^2_{Q(\hat{\varepsilon},p)} \xrightarrow{P_\beta} \sigma^2_{Q(\hat{\varepsilon},p)}.$$

PROOF By Lemma 7.1, it remains to show that

$$\bar{f}_{B_n}(\cdot) \xrightarrow{P_\beta} f(\cdot), \tag{7.86}$$

and

$$\bar{\xi}_p \xrightarrow{P_\beta} \xi_p. \tag{7.87}$$

First, we show (7.86). Denote

$$\hat{f}_{B_n}(x|\theta) := \frac{1}{n}\sum_{t=1}^{n} \phi_{B_n}[x - g_t(\theta)],$$

where

$$\theta := (\theta_0, \theta_1, \ldots, \theta_k)' \quad \text{and} \quad g_t(\theta) := \frac{y_t}{\sqrt{\theta_0 + \sum_{j=1}^{k} \theta_j y_{t-j}^2}}.$$

As (6.12) of Bickel (1982), we can show that, for any θ fixed,

$$\hat{f}_{B_n}(\cdot|\theta) \xrightarrow{P_\theta} f(\cdot|\theta) \quad \text{if} \quad nB_n \to \infty,$$

where $f(\cdot|\theta)$ denotes the density of $g_t(\theta)$. Substituting the nonrandom sequence $\beta^{(b)}$ used in the proof of Lemma 7.1 into θ, and by the continuity of f, we have

$$\hat{f}_{B_n}(\cdot|\beta^{(b)}) \xrightarrow{P_{\beta^{(b)}}} f(\cdot|\beta) = f(\cdot)$$

for $\beta^{(b)}$ fixed. Furthermore, we can replace $P_{\beta^{(b)}}$ above with P_β because the following statement by Linton (1993):

$$\log \frac{P_{\beta^{(b)}}}{P_\beta} \xrightarrow{\mathcal{L}} \mathrm{N}\left(-\frac{1}{2}h' I_{\beta,f} h,\ h' I_{\beta,f} h\right) \quad \text{under } P_\beta,$$

for some $I_{\beta,f}$, is necessary and sufficient for the mutual contiguity of $P_{\beta^{(b)}}$ and P_β by Le Cam's first lemma (cf. Example 6.5 of van der Vaart (1998)). Here "mutual contiguous" means that we can interchange the two measures when we make statements about convergence to zero in probability: For any event S, we have $P_{\beta^{(b)}}(S) \to 0$ if and only if $P_\beta(S) \to 0$. Hence, we have

$$\hat{f}_{B_n}(\cdot|\beta^{(b)}) \xrightarrow{P_\beta} f(\cdot),$$

which, together with Lemma 6.8, implies (7.86).

(7.87) can be proved by (7.86) and Proposition 0.1 of Resnick (1987) which asserts that if $H_n, n \geq 0$ are nondecreasing functions and $H_n \to H_0$, then $H_n^{\leftarrow} \to H_0^{\leftarrow}$. □

Now that we have the asymptotic distribution of $\hat{\mathbb{F}}_n^{\leftarrow}(p)$, we can state an asymptotic version of the feasible VaR based on REP (7.75) as

$$\mathrm{VaR}_p^{(REP)} \approx \hat{\sigma}_{n+1}\left\{\xi_p + \frac{\hat{\sigma}_{Q(\hat{\varepsilon},p)}\Phi^{\leftarrow}(1-\alpha)}{\sqrt{n}}\right\}, \tag{7.88}$$

where $\Phi(\cdot)$ denotes the standard normal distribution function. Here we repeat that there is no need to go further, such as considering the perturbation of $\hat{\sigma}_{n+1}$ in (7.88), because it will only break the equation (7.74). The second term of (7.88) is the contribution of this paper, which has been ignored by commonly used approaches. We will examine its importance by numerical studies below.

First, we check that $\mathrm{VaR}^{(REP)}$ does convince us to avoid the specified risk with high confidence. We also examine the behavior of $\hat{\sigma}_Q^2$, which reveals that the ARCH affection and the influence caused by the heavy-tailedness of innovation density will reflect on $\mathrm{VaR}^{(REP)}$.

Consider the ARCH(1) process with standard normal innovations:

$$y_t = \varepsilon_t \sqrt{\beta_0 + \beta_1 y_{t-1}^2}, \qquad \varepsilon_t \sim \text{ i.i.d. } N(0,1), \qquad t = 1, \ldots, n.$$

Denoting $\mathrm{VaR}_{p,t}^{(\cdot)}$ for $\mathrm{VaR}_{p}^{(\cdot)}$ calculated by $\hat{\sigma}_t$, we counted the number:

$$N_i^{(C)} := \#\{y_{t_i} < \mathrm{VaR}_{p,t_i}^{(C)}\} \text{ and } N_i^{(REP)} := \#\{y_{t_i} < \mathrm{VaR}_{p,t_i}^{(REP)}\},$$

$(t_i = 1, \ldots, 100,)$ for 100 trials $(i = 1, \ldots, 100)$ with $p = 0.1$ and $\alpha = 0.05$. Here $\#S$ denotes the number of elements in the set S. If this number is less than $10(= p \times n)$ for some i, then it means that the calculated VaR failed to alarm for the specified risk during the i-th trial. $L := \#\{N_i^{(\cdot)} < 10\}$ in Table 7.7 below shows these undesirable lack of exceedances, and $\bar{N} := \sum_{i=1}^{100} N_i^{(\cdot)}/100$ in brackets is the averaged number of exceedance. There we set β_0 to be equal to $1 - \beta_1$ in order to keep the unconditional variance to be 1 regardless of the changes of ARCH parameters (this will not have any substantial effects on the result).

Table 7.7 *Exceedance numbers*

		L ($\bar{N}$)			
		$\mathrm{VaR}_p^{(C)}$		$\mathrm{VaR}_p^{(REP)}$	
β_1	0.1	42	(10.25)	5	(15.02)
	0.3	40	(10.15)	6	(15.04)
	0.5	48	(9.76)	5	(15)
	0.7	56	(8.81)	9	(14.15)
	0.9	67	(7.8)	29	(12.78)

As expected, $\mathrm{VaR}^{(C)}$ lacked in almost a half of the trials and hence we cannot expect $\mathrm{VaR}^{(C)}$ will keep us away from specified risk at the next future period. On the other hand, our $\mathrm{VaR}^{(REP)}$ restricts such undesirable lack with confidence arguments. Yet, unfortunately, the unexpected lack that appeared in Table 7.7 often occurred more than 5 $(= n \times \alpha)$ times especially when estimation performance was not good, i.e., when the averaged number for $\mathrm{VaR}_p^{(C)}$ is far from its expected value 10. This is understandable because, by our definition (7.88), $\mathrm{VaR}^{(REP)}$ essentially contains $\mathrm{VaR}^{(C)}$ in its first term and hence it suffers from estimation difficulties as much as $\mathrm{VaR}^{(C)}$ does. But how to obtain a good estimator for ARCH parameters is actually not our issue here, and when estimation goes well $\mathrm{VaR}^{(REP)}$ seems to return the number close enough to 5.

Next, we will investigate the difference between $\mathrm{VaR}_p^{(C)}$ and $\mathrm{VaR}_p^{(REP)}$ in substantial sense, i.e., focus on $\hat{\sigma}^2_{Q(\hat{\varepsilon},p)}$ and see how it behaves. We generated 10 stretches of $\{y_t\}$ with length $n = 100$, and evaluated the simulated mean values of $\hat{\sigma}^2_Q/n$ for various parameter values and innovation distributions. Actually, in the following, we will set $\beta_0 = 1$ and move only β_1 because β_0 is related to overall effect and does not have as much contribution to the change of asymptotic variance as β_1 does. The changes of $\hat{\sigma}^2_Q/n$ with ARCH parameter β_1 and risk probability p are displayed in Figure 7.13.

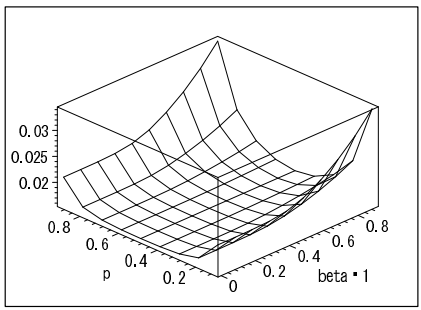


Figure 7.13 $\hat{\sigma}_Q^2/n$ *for* p *and* β_1

Recalling that VaR is usually calculated for rare events, the obvious increase of the asymptotic variances for the smaller p tells us the importance of this quantity. Furthermore, in view of its increase in ARCH parameter value, we should pay attention especially to the RiskMetrics which admits IGARCH setting, i.e., when unit roots can exist in a volatility equation (in our case, whenever $\beta_1 = 1$). This may be displayed clearly if we decompose the $\hat{\sigma}_Q^2$ into 3 terms as in (7.81). The first term of (7.81) which corresponds to the usual Brownian bridge term is independent of ARCH parameters. The dependency of the second term on ARCH parameters is through $(\hat{\beta} - \beta)$ so that the differences of the value β will not make a great contribution. Therefore, the dominant factor which causes such behavior of asymptotic variance is the third term and this dependence feature on ARCH parameters clarifies the specific importance of our $\mathrm{VaR}^{(REP)}$ in ARCH setting. As mentioned, even if we do the same thing in ARMA setting, nothing but the first term can remain in the asymptotic variance. Because the third term is essentially the asymptotic variance of estimator for ARCH parameters, the increase of $\hat{\sigma}_Q$ in ARCH parameters can be thought of as the reflection of, again, the difficulties of parameter estimation.

Finally, we will see how our $\mathrm{VaR}^{(REP)}$ may be disturbed if the innovation density exhibits heavy-tailed property. To see this, we used the Student-t distribution with 5 degrees of freedom. Here the standardized Student-t distribution with ν degrees of freedom, satisfying $E[\varepsilon_t] = 0$ and $E[\varepsilon_t^2] = 1$, is defined by its probability density function:

$$f(x|\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\nu-2)\pi}} \left(1 + \frac{x^2}{\nu-2}\right)^{-\frac{\nu+1}{2}}, \qquad \nu > 2,$$

and is written as Student-$t(\nu)$. The plot of $\hat{\sigma}_Q^2/n$ for Student-$t(5)$ is displayed in Figure 7.14.

About $\hat{\sigma}_Q^2$ decomposed into 3 terms, we can easily imagine that the first term will become larger than standard normal innovation case for very small p and

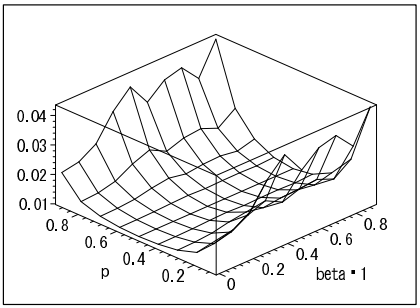


Figure 7.14 $\hat{\sigma}_Q^2/n$ *with Student-t*(5)

the second term will not play an important role as before. So the change of shape from standard normal case to heavy-tailed case is due to the third term, i.e., by ARCH sensitiveness of $\mathrm{VaR}^{(REP)}$. The influence of heavy-tailedness seems to be enlarged by this third term, and hence if heavy-tailed density appears to be appropriate for the innovations we should be more careful about the presence of ARCH parameters even if they are in small values: In fact, we may expect the gradual increase and pay attention to only larger β_1 for standard normal innovations as displayed in Figure 7.13. But, in the case of Student-t innovations one of the peaks appeared for $\beta_1 = 0.4$ as in Figure 7.14, which was not found in the standard normal case. The additional influence of heavy-tailedness through the third term, together with those through the first term, gives bigger magnitude to $\hat{\sigma}_Q^2$ and results in almost twice in comparison with the standard normal case.

In view of the figures, we observe that our $\mathrm{VaR}^{(REP)}$ can adjust ARCH effects and enhanced heavy-tail affections properly to serve as a well-grounded criterion to keep the assets away from the specified risk with high confidence level. Recalling that VaR is used to ensure that the financial institutions can still be in business, our $\mathrm{VaR}^{(REP)}$ can be useful for that purpose.

Exercises

7.1 Let $\{X_t, \mathcal{A}_t\}$ be a martingale. Then, show that

$$E\{X_m|\mathcal{A}_k\} \;=\; X_k \quad a.e., \qquad \text{for any} \;\; m > k.$$

7.2 Let X be a random variable on $(\Omega, \mathcal{A}, P)$, and let $\mathcal{A}_1$ and $\mathcal{A}_2$ be sub σ-fields of $\mathcal{A}$ satisfying $\mathcal{A}_1 \subset \mathcal{A}_2$. Then, show that

$$E\{E(X|\mathcal{A}_2|\mathcal{A}_1\} \;=\; E\{X|\mathcal{A}_1\} \quad a.e.$$

7.3 From (7.29), derive the Black-Scholes formula (7.30).

7.4 Verify that portfolio weights (7.54), (7.55), (7.57) and (7.58) are the solutions of the corresponding criteria of utility.

7.5 (i) Show that the spectral density (7.66) does not satisfy (7.65), if $p_1 < p_2$. (ii) Show that the spectral density (7.67) does not satisfy (7.65).

7.6 Verify the relation (7.77).

CHAPTER 8

Term Structure

If we buy a bond we are loaning money to a corporation. The corporation is obligated to pay back the principal, called the face value, and promises to pay a stream of certain payments, called coupons. Hence we receive a fixed stream of income. Thus bonds are called fixed-income securities. Although bonds seem to be risk-free, they are not so. Our income from the bond is guaranteed only if we keep the bond to maturity. If we sell the bond before maturity, our return will depend on changes in the price of the bond. Bond prices vary in the opposite direction to interest rates. Although the interest rate of our bond is fixed, that in the cash market varies. Therefore, the price of our bond in the cash market varies.

In financial markets, the term structure of interest rates is crucial to pricing of fixed income securities. This chapter introduces some models for interest rates and discount bonds, and discusses their no-arbitrage pricing theory. Numerical examples of actual data are also provided. More concretely, Section 8.1 explains the spot interest rates and forward rates, and introduces a class of CHARN models to describe them. Using a relation between the spot interest rates in the cash market and discount bond prices, we discuss a no-arbitrage pricing theory on discount bonds. Section 8.2 provides some empirical studies for the term structure of discount bond, yield-to-maturity and forward rate.

8.1 Spot Rates and Discount Bonds

Suppose that a time interval $[0, \tau]$ is divided into L intervals with length h. Let B_n ($n \in \mathbf{N}$) be the amount of a bank deposit after nh years from 0 with initial principal one dollar ($B_0 = 1$). If the interest $B_{n+1} - B_n$ in the period $(nh, (n+1)h]$ is proportional to B_n, i.e.,

$$B_{n+1} - B_n = rB_n, \quad n = 0, 1, 2, \ldots, \tag{8.1}$$

then r (> 0) is called the *compounded interest rate*. From (8.1),

$$B_n = (1 + r)^n, \quad n = 0, 1, 2, \ldots \tag{8.2}$$

If the interest rate paid is continuously compounding, letting $t = nh$ and $h \to 0$, we can see that the amount of deposit at time t is

$$B(t) = e^{rt}, \quad t \geq 0. \tag{8.3}$$

It is natural to consider the case when the interest rate varies with respect to time t, i.e., $r = r(t)$, which is called the *instantaneous interest rate* (or *spot rate*). A continuous version of (8.1) may be written as

$$\frac{dB(t)}{B(t)} = r(t)\,dt, \quad B(0) = 1, \tag{8.4}$$

whose solution is

$$B(t) = \exp\left\{\int_0^t r(u)\,du\right\}, \quad t \geq 0. \tag{8.5}$$

If one dollar in the cash market is rolled by the rate $r(u)$ over the period $[t, T]$, it grows to

$$B(t, T) = \exp\left\{\int_t^T r(u)\,du\right\}. \tag{8.6}$$

Conversely, given one dollar at a future time T, its value at present time t is

$$P^*(t, T) = \exp\left\{-\int_t^T r(u)\,du\right\}. \tag{8.7}$$

In the above we have dealt with the deterministic interest rate function $r(t)$. However, deterministic interest rates are inadequate to capture interest rate movements in the actual cash market. A natural approach is to consider their stochastic models. Vasicek (1977) introduced the following stochastic differential equation model

$$dr(t) = a(b - r(t))\,dt + \sigma\,dW_t, \tag{8.8}$$

where a, b, σ are non-negative constants, and $\{W_t\}$ is a Wiener process. Cox et al. (1985) suggested

$$dr(t) = (a - b\,r(t))\,dt + \sigma\sqrt{r(t)}\,dW_t, \tag{8.9}$$

where a, b and σ are non-negative constants. Hull and White (1990) proposed

$$dr(t) = [\theta(t) - a(t)\,\{b(t) - r(t)\}]\,dt + \sigma(t)\,dW_t, \tag{8.10}$$

where $\theta(t), a(t), b(t)$ and $\sigma(t)$ are deterministic functions. The models (8.8)-(8.10) are a special case of the following stochastic differential equation

$$dr(t) = \alpha(t, r(t))\,dt + \beta(t, r(t))\,dW_t. \tag{8.11}$$

A fixed income security can promise a stream of future payments of any form, but there are two cases as follows. *Zero-coupon bonds*, also called *discount bonds*, make a single payment at a date in the future known as the *maturity date*. The size of this payment is called the *face value* of the bond. The length of time to the maturity date is the *maturity* of the bond. US Treasury bills take this form. *Coupon bonds* make *coupon payments* of a given fraction of the face value at equally spaced dates up to and including the maturity date,

when the face value is also paid. US Treasury notes and bonds take this form. Coupon payments on Treasury notes and bonds are every six months.

As an example, consider a 20-year coupon bond with face value F and annual interest rate r with semi-annual coupon payments. Each coupon payment will be C. Thus, the bond holder receives 40 payments of C, and F after 20 years. If the current market price of the bond is PM, it is common to define r as a solution of the equation

$$\sum_{t=1}^{40} \frac{C}{(1+r/2)^t} + \frac{F}{(1+r/2)^{40}} = PM. \tag{8.12}$$

Let $P(t,T), (t \le T)$, be the time t price of a discount bond which pays one dollar at maturity T. The function $P(t,T)$ with respect to T is called the *term structure* of the discount bond. Let

$$Y(t,T) = -\frac{\log P(t,T)}{T-t}, \quad t \le T, \tag{8.13}$$

which is called the *yield-to-maturity* (or *yield*). Also the limit

$$r_B(t) = \lim_{T \to t} Y(t,T), \tag{8.14}$$

is called the *spot rate* of the discount bond at time t. Since (8.14) is the yield of the discount bond with infinitesimal maturity, the function $r_B(t)$ is a conceptual one. From (8.13), (8.14) and $P(t,t) = 1$, it follows that

$$r_B(t) = -\lim_{T \to t} \frac{\log P(t,T)}{T-t} = -\frac{\partial}{\partial T} \log P(t,T)\Big|_{T=t,} \tag{8.15}$$

which implies that the discount bond price $P(t,T)$ cannot be recovered from $r_B(t)$ alone.

Next, let $f(t,T,\tau)$ be defined by

$$f(t,T,\tau) = -\frac{\log P(t,\tau) - \log P(t,T)}{\tau - T}, \quad t \le T < \tau, \tag{8.16}$$

which is called the *forward yield*, and is regarded as the time t yield of the discount bond over the future time interval $[T,\tau]$. Let

$$\begin{aligned} f(t,T) &= \lim_{\tau \to T} f(t,T,\tau) \\ &= -\frac{\partial}{\partial T} \log P(t,T), \quad t \le T, \end{aligned} \tag{8.17}$$

which is called the *forward rate*. Although $f(t,T)$ cannot be observed in the cash market, from (8.17), we have

$$P(t,T) = \exp\left\{ -\int_t^T f(t,s)\, ds \right\}, \quad t \le T, \tag{8.18}$$

which implies that $P(t,T)$ is recovered from the forward rate $f(t,s)$. This is

an advantage of the forward rate. However, as we saw in (8.15), the spot rate does not have this property.

From a statistical point of view, it is convenient to develop the discussion in a discrete time approach. In what follows we proceed in this way. Suppose that a given time interval $[0, \tau]$ is divided into subintervals with length $h\,(> 0)$. The n-th interval is $(nh, (n+1)h]$, and the n-th time corresponds to nh years from 0. Let r_n be a continuously compounded interest rate at time n, and is the interest rate for the time interval $(nh, (n+1)h]$. The rate r_n is called an *h-year spot rate*, and guarantees that one dollar at time n grows to $\exp(r_n h)$ dollars at time $n+1$. If one dollar at time n is compoundedly rolled over up to time N, then it grows to

$$B_{n,N} = \exp\left(\sum_{j=n}^{N-1} r_j h\right). \tag{8.19}$$

Conversely, if one dollar at a future time N is given, then the present value at time n is

$$P^*_{n,N} = \exp\left(-\sum_{j=n}^{N-1} r_j h\right). \tag{8.20}$$

Here $B_{n,N}$ and $P^*_{n,N}$ depend on the future spot rates $\{r_m : m > n\}$, which are unobservable random variables at n.

In the cash market, there are many other longer term spot rates. For $m = 1, 2, \ldots, M$, let $r_n(m)$ be a continuously compounded interest rate at time n for the period $(nh, (n+m)h]$. The rate $r_n(m)$ is called an *mh-year spot rate*, and guarantees that one dollar at time n grows to $\exp\{r_n(m)mh\}$ dollars at time $n+m$. Then the h-year spot rate is the special case of $m = 1$, i.e.,

$$r_n = r_n(1),$$

and is called the *shortest spot rate* in distinction from $\{r_n(m), m \geq 2\}$. Although all of $r_n(m), m = 1, \ldots, M$, may not exist actually, we assume that all of them exist from a theoretical point of view.

At time n, $r_n(m)$ is realized, and the set

$$\{r_n(1), r_n(2), \ldots, r_n(M)\} \tag{8.21}$$

gives the *term structure of spot rates* at n. If $N = km$, the value at n of one dollar at N is evaluated by the mh-year rate as

$$P^*_{n,N}(m) = \exp\left\{-\sum_{j=0}^{k-1} r_{n+jm}(m)mh\right\}. \tag{8.22}$$

Evidently $P^*_{n,N}(m)$ is an unrealized random variable at n, unless $N = m$ $(k = 1)$.

Kariya and Liu (2003) discussed no-arbitrage pricing of discount bonds assuming the following K-factor discrete time diffusion model

$$r_n(m) - r_{n-1}(m) = \alpha_{n-1}(m)h + \sum_{k=1}^{K} \beta_{k,n-1}(m)\sqrt{h}\, u_{k,n}, \tag{8.23}$$

with affine structure

$$r_n(m) = a(m) + b(m)r_n.$$

Here $u_{k,n}$'s are i.i.d. $N(0,1)$, and $\alpha_{n-1}(m)$ and $\beta_{k,n-1}(m)$ may depend on the past term structures, and $a(m)$ and $b(m)$ are constants. In what follows, we assume a class of CHARN models introduced in Section 6.2 for spot rates. CHARN models are more general than ones in (8.23). Then the problem of pricing discount bonds is addressed.

Let $P_{n,N}$ be the price at n of a discount bond which guarantees one dollar at maturity N. Of course $P_{N,N} = 1$. For $n < N$, $P_{n,N}$'s are random variables. Recall $P^*_{n,N}$ in (8.20) which is the value at n of one dollar at N in the cash market. Let us see the relation between $P_{n,N}$ and $P^*_{n,N}$. In the case of $N = n+1$, it should hold that $P_{n,n+1} = P^*_{n,n+1}$. However, in the case of $N = n+k$ with $k \geq 2$, because $P^*_{n,n+k}$ depends on future spot rates which are random, it will hold that $P_{n,n+k} \neq P^*_{n,n+k}$, $(k \geq 2)$. Similarly to this argument, recalling $P^*_{n,N}(m)$ in (8.22) and mh-year spot rate $r_n(m)$, we have

$$P_{n,n+m} = P^*_{n,n+m}(m) \quad (m = 1, 2, \ldots, M). \tag{8.24}$$

If we assume the affine structure $r_n(m) = a(m) + b(m)r_n$, then the price of bond $P_{n,n+m}$ is expressed as a function of the shortest spot rate r_n.

In what follows, by use of the no-arbitrage theory we derive the no-arbitrage price of discount bonds in terms of spot rates $r_0, r_1, \ldots, r_n$ in the cash market. Let

$$B_n = \exp\left\{\sum_{j=0}^{n-1} r_j h\right\},$$

i.e., one dollar at time 0 grows to B_n at time n. Suppose that the spot rates in cash market and the bond spot rates are of an arbitrage relation. For $P_{n,N}$ and B_n to be arbitrage-free, it is required that the process

$$Y_{n,N} \equiv P_{n,N}/B_n, \quad (n = 0, 1, \ldots, N) \tag{8.25}$$

is a martingale with respect to a probability measure Q^* equivalent to the probability measure of $\{r_n\}$. Thus,

$$P_{n,N}/B_n = E^*_n\{P_{N,N}/B_N\} = E^*_n\{1/B_N\}, \tag{8.26}$$

where $E^*_n(\cdot)$ is the conditional expectation of $(\cdot)$ with respect to Q^* given $\{r_0, \ldots, r_n\}$. Since $P_{n,N}/B_n = E^*_n\{1/B_N\}$ is automatically a martingale (see Exercise 8.1) for any Q^*, we usually use the original probability measure

Q of $\{r_0, ..., r_n\}$ as Q^*, and denote by $E_n(\cdot)$ the conditional expectation under Q. From (8.26) it follows that

$$P_{n,N} = E_n\{B_n/B_N\} = E_n\{P^*_{n,N}\} = E_n\left\{\exp\left(-\sum_{j=n}^{N-1} r_j h\right)\right\}, \qquad (8.27)$$

which is the no-arbitrage price of a discount bond by spot rates. Suppose that $\{r_j\}$ is generated by the following CHARN type model

$$r_j - r_{j-1} = \alpha(j-1)h + \beta(j-1)\sqrt{h}\, u_j, \qquad (8.28)$$

where $\alpha(j-1)$ and $\beta(j-1) \in \sigma\{u_{j-1}, u_{j-2}, \ldots\}$, and $\{u_j\} \sim$ i.i.d. $(0,1)$ with moment generating function $\phi(\cdot)$. From (8.27) we obtain

$$P_{N-1,N} = E_{N-1}\left\{\exp\left(-r_{N-1}h\right)\right\} = \exp\left(-r_{N-1}h\right). \qquad (8.29)$$

Next, we evaluate

$$\begin{aligned} P_{N-2,N} &= E_{N-2}\left[\exp\left\{-r_{N-1}h - r_{N-2}h\right\}\right] \\ &= \exp\{-2r_{N-2}h\}\, E_{N-2}\left[\exp\left\{-(r_{N-1} - r_{N-2})h\right\}\right]. \end{aligned} \qquad (8.30)$$

From (8.28) we obtain

$$\begin{aligned} P_{N-2,N} &= \exp\{-2r_{N-2}h\} E_{N-2}[\exp\{-(\alpha(N-2)h + \beta(N-2)\sqrt{h}u_{N-1})h\}] \\ &= \exp\{-2r_{N-2}h - \alpha(N-2)h^2\}\phi\{-\beta(N-2)h^{3/2}\}. \end{aligned} \qquad (8.31)$$

If the structures of $\alpha(\cdot), \beta(\cdot)$ and $\phi(\cdot)$ are simple and manageable, we can proceed to get a recursive formula for $P_{N-k,N}$. For example, assuming that

$$\{u_j\} \;\sim\; i.i.d.\; N(0,1), \qquad (8.32)$$

$$\alpha(j) = \theta_1 + \theta_2\gamma_j, \qquad \beta(j) = \sqrt{\theta_3 + \theta_4\gamma_j}, \qquad (8.33)$$

Kariya and Liu (2003) derived a recursive formula for $P_{N-k,N}$ (Exercise 8.2).

However, if the structures of $\alpha(\cdot), \beta(\cdot)$ and $\phi(\cdot)$ are intractable, we need some numerical method to valuate (8.27) as in Section 7.1.

8.2 Estimation Procedures for Term Structure

This section explains how to interpolate term structure data for discount bonds, spot rates and forward rates, which are observed at discrete maturity points. Also, a pricing via forward rates is discussed in terms of the CHARN model.

Let $P(t,T)$, $(t \le T)$, be the time t price of a discount bond which pays one dollar at maturity T. Recall (8.13) and (8.17). The yield-to-maturity $Y(t,T)$

and the forward rate $f(t,T)$ were defined by

$$Y(t,T) = -\frac{\log P(t,T)}{T-t}, \tag{8.34}$$

$$f(t,T) = -\frac{\partial}{\partial T}\log P(t,T). \tag{8.35}$$

From (8.34) and (8.35) the following relations hold:

$$P(t,t+\tau) = \exp\left\{-\int_0^\tau f(t,t+u)\ du\right\}, \tag{8.36}$$

$$Y(t,t+\tau) = \frac{1}{\tau}\int_0^\tau f(t,t+u)\ du. \tag{8.37}$$

Hence, at a given time t, any one of the following three curves:

$$x \longrightarrow P(t,t+x), \tag{8.38}$$

$$x \longrightarrow Y(t,t+x), \tag{8.39}$$

$$x \longrightarrow f(t,t+x), \tag{8.40}$$

can describe the term structure of discount bond, yield-to-maturity and forward rate. Figure 8.1 shows the yield curve and the forward rate curve of the U.S. zero coupon for January 1991, corresponding to (8.39) and (8.40), respectively.

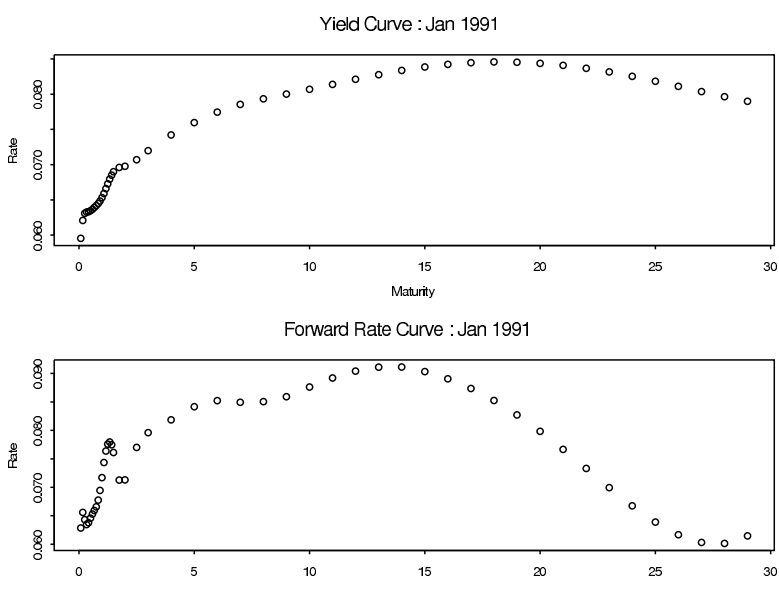


Figure 8.1 *Yield and forward rate curves*

Since they are observed at discrete maturity points, we next discuss estimation of the curves.

Nelson and Siegel (1987) proposed the following model

$$f_{NS}(x:\theta) = \theta_0 + \theta_1 e^{-x/\theta_3} + \theta_2 \frac{x}{\theta_3} e^{-x/\theta_3} \tag{8.41}$$

for the forward rate curve (8.40), where $\theta = (\theta_0, \theta_1, \theta_2, \theta_3)'$. It is assumed that $\theta_0 > 0$ and $\theta_3 > 0$. Hence, θ_0 is the asymptotic value of the forward rate, and $\theta_0 + \theta_1$ is the forward rate at the present time. Appropriate choice of θ can describe typical shapes of forward rate curves, such as upward sloping, downward sloping, humped, or inverted humped. The three components in (8.41) can be interpreted as the long-term, short-term and medium-term configuration. From (8.37) the corresponding yield curve is

$$Y_{NS}(x:\theta) = \theta_0 + \theta_1 \frac{1 - e^{-x/\theta_3}}{x/\theta_3} + \theta_2 \left\{ \frac{1 - e^{-x/\theta_3}}{x/\theta_3} - e^{-x/\theta_3} \right\}. \tag{8.42}$$

We fit $Y_{NS}(x:\theta)$ to the yield data in Figure 8.1 by use of the least squares method. Figure 8.2 plots the estimated $Y_{NS}(x:\theta)$ by solid line.

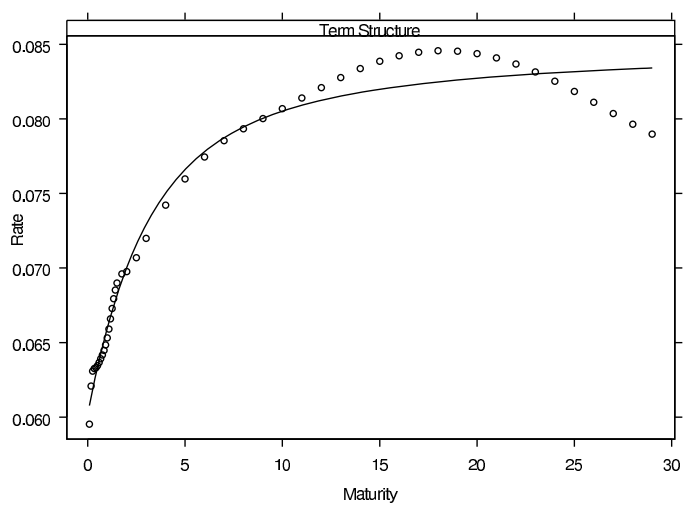


Figure 8.2 *Yield curve*

However, the fit is not so good. In view of this, Svensson (1994) extended the Nelson-Siegel forward function $f_{NS}(x:\theta)$ to

$$f_S(x) = f_{NS}(x:\theta) + \theta_4 \frac{x}{\theta_5} e^{-x/\theta_5}. \tag{8.43}$$

The corresponding yield function is

$$Y_S(x) = \theta_0 + \theta_1 \frac{1 - e^{-x/\theta_3}}{x/\theta_3} + \theta_2 \left\{ \frac{1 - e^{-x/\theta_3}}{x/\theta_3} - e^{-x/\theta_3} \right\} + \theta_4 \left\{ \frac{1 - e^{-x/\theta_5}}{x/\theta_4} - e^{-x/\theta_5} \right\}. \tag{8.44}$$

Fitting $Y_S(x)$ to the yield data by the least squares method, we plot the estimated $Y_S(x)$ in Figure 8.3 by real line.

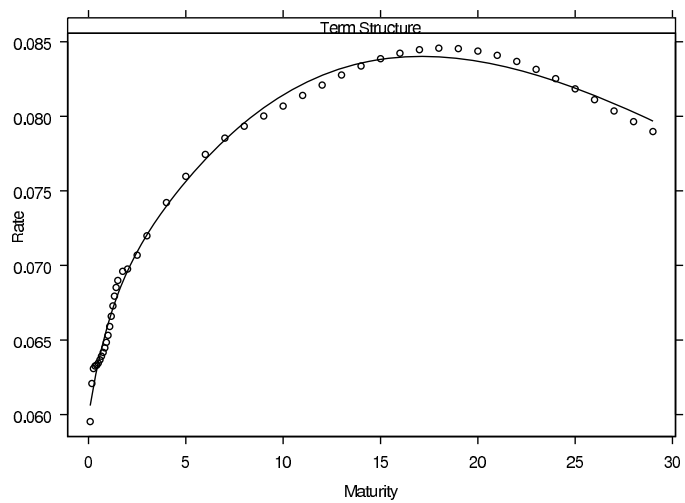


Figure 8.3 *Yield curve*

From Figure 8.3 we observe that the fit by (8.44) is good.

There are many other methods, which estimate the term structure curves (8.38) - (8.40). McCulloch (1971, 1975) proposed a spline method to estimate the term structure of a discount bond. For the price of discount bond $P(t, t+x)$, he fits

$$p(x) = a_0 + \sum_{j=1}^{k} a_j s_j(x), \qquad x \in [0, M], \tag{8.45}$$

to the data by use of the least squares method. Here $s_j(x)$'s are known functions of maturity x, and a_j's are unknown parameters. Because of $P(t,t) = 1$, we set $a_0 = 1$ and $s_j(0) = 0$, $j = 1, \ldots, k$. Concrete forms of $s_j(x)$ are given in McCulloch (1971) as follows. Divide $[0, M]$ into subintervals $(d_{j-1}, d_j]$ satisfying $0 = d_1 < d_2 < \cdots < d_n = M$. Then the $s_j(x)$'s are defined as

$$s_1(x) = \begin{cases} x - \frac{1}{2d_2}x^2, & 0 \le x \le d_2, \\ \frac{1}{2}d_2, & d_2 < x \le d_n, \end{cases}$$

$$s_j(x) = \begin{cases} 0, & 0 < x < d_{j-1}, \\ \frac{(x-d_{j-1})^2}{2(d_j - d_{j-1})}, & d_{j-1} < x \le d_j, \\ \frac{d_j - d_{j-1}}{2} + (x - d_j) - \frac{(x-d_j)^2}{2(d_{j+1}-d_j)}d_2, & d_j < x \le d_{j+1}, \\ \frac{d_{j+1}-d_j}{2}, & d_{j+1} < x \le d_n, \end{cases} \quad (8.46)$$

$$(j = 2, \ldots, k-1)$$

$$s_k(x) = \begin{cases} 0, & 0 \le x \le d_{k-1}, \\ \frac{(x-d_{k-1})^2}{2(d_n - d_{k-1})}, & d_{k-1} < x \le d_n, \end{cases}$$

which are called quadratic spline functions.

Furthermore, McCulloch (1975) proposed a cubic spline form of $s_j(x)$ based on piecewise polynomials of degree 3. For the U.S. zero coupon data, we fitted the $p(x)$ based on the quadratic spline functions (8.46) by the least squares method. Figure 8.4 plots the discount bond price and the estimated $p(x)$ by dot points and solid line, respectively. The fitting curve is good.

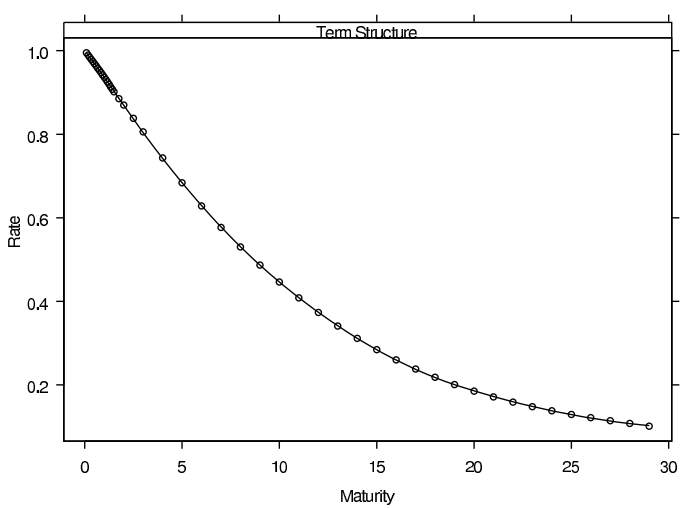


Figure 8.4 *Discount function*

There are many other methods to estimate the term structure curves, e.g., smoothing spline method etc. For details, we refer the reader to, e.g., Zivot

and Wang (2006), Carmona (2004), Houglet (1980) and Vasicek and Fong (1982).

In the previous section the no-arbitrage price of a bond by spot rates was given by use of the conditional expectation (see (8.26)). In what follows we discuss pricing a bond by the forward rates. Let r_n be a shortest spot rate over the time interval $(nh, (n+1)h]$, and let $f_{n,N}$ be a forward rate at n on $r_N, (n < N)$. Then, without taking conditional expectation, the theoretical price at n of a discount bond with maturity N is given by

$$P_{n,N} = \exp\left[-\sum_{m=n}^{N-1} f_{n,m}\, h\right], \tag{8.47}$$

which is a discretized version of (8.18) with setting $t = nh$ and $T = Nh$. Assuming that $\{f_{n,m}\}$ is generated by a K-factor discrete diffusion process, Kariya and Liu (2003) derived a no-arbitrage condition on the discount bond. It is possible to derive such a condition for a class of CHARN type models, which includes the K-factor discrete diffusion process. In fact, let $\{f_{n.m}\}$ be generated by

$$f_{n,m} - f_{n-1,m} = \alpha(n-1,m)h + \boldsymbol{\beta}'(n-1,m)\sqrt{h}\,\mathbf{U}(n) \quad (m \geq n) \tag{8.48}$$

where $\boldsymbol{\beta}(n-1,m)$'s are K-dimensional random vectors, and $\mathbf{U}(n)$'s are i.i.d. K-dimensional random vectors with moment generating function $\phi_K(\cdot)$, zero mean and variance matrix I_K ($K \times K$ identity matrix). Furthermore, it is assumed that $\alpha(n-1,m)$ and all the components of $\boldsymbol{\beta}(n-1,m) \in \mathcal{F}_{n-1} \equiv \sigma\{\mathbf{U}(n-1), \mathbf{U}(n-2), \dots\}$. Note that the shortest spot rate r_m is equal to $f_{m,m}$. Hence, from (8.48) it follows that

$$f_{n,m} = f_{0,m} + \sum_{j=1}^{n} \alpha(j-1,m)h + \sum_{j=1}^{n} \boldsymbol{\beta}'(j-1,m)\sqrt{h}\,\mathbf{U}(j), \tag{8.49}$$

$$r_m = f_{0,m} + \sum_{j=1}^{m} \alpha(j-1,m)h + \sum_{j=1}^{m} \boldsymbol{\beta}'(j-1,m)\sqrt{h}\,\mathbf{U}(j), \tag{8.50}$$

The price of cash rolled over with spot rate r_m up to n is

$$B_n = \exp\left\{\sum_{m=0}^{n-1} r_m h\right\}, \tag{8.51}$$

which forms the relative price of $P_{n,N}$ given by (8.47) if we consider a no-arbitrage condition. Let the relative price be

$$Y_{n,N} = \frac{P_{n,N}}{B_n}. \tag{8.52}$$

We can derive a martingale condition for $Y_{n,N}$ under an equivalent measure

Q^*. From (8.49) and (8.50) it is seen that

$$Y_{n,N} = Y_{n-1,N} \exp\left\{ -\sum_{m=n}^{N-1} \alpha(n-1,m)h^2 - \sum_{m=n}^{N-1} \boldsymbol{\beta}'(n-1,m)h^{3/2}\mathbf{U}(n) \right\} \tag{8.53}$$

(Exercise 8.4). For $Y_{n,N}$ to be a martingale under Q^*, it is required that

$$E^*\left[\exp\left\{ -\sum_{m=n}^{N-1} \alpha(n-1,m)h^2 - \sum_{m=n}^{N-1} \boldsymbol{\beta}'(n-1,m)h^{3/2}\mathbf{U}(n) \right\} \middle| \mathcal{F}_{n-1} \right] = 1 \quad a.e. \tag{8.54}$$

Let $\mathbf{U}^*(n) \equiv \mathbf{U}(n) - \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is a K-dimensional constant vector. Denoting the moment generating function of $\mathbf{U}^*(n)$ by $\phi_K^*(\cdot)$, we can see that (8.54) is equivalent to

$$\exp\left\{ -\sum_{m=n}^{N-1} \alpha(n-1,m)h^2 \right\} \phi_K^*\left\{ -\sum_{m=n}^{N-1} \boldsymbol{\beta}'(n-1,m)h^{3/2} \right\} = 1, \quad a.e. \tag{8.55}$$

If $\phi_K^*(\cdot)$ is given in explicit form, from (8.55) we can write a no-arbitrage condition in terms of $\alpha(\ ,\ ), \boldsymbol{\beta}(\ ,\ )$ and $\boldsymbol{\xi}$ (see Exercise 8.5).

Exercises

8.1 Let X be an integrable random variable on $(\Omega, \mathcal{A}, P)$, and let $\{\mathcal{A}_n\}$ be an increasing sequence of sub σ-fields of $\mathcal{A}$, and $X_n = E(X|\mathcal{A}_n)$. Then show that $\{X_n, \mathcal{A}_n\}$ is a martingale.

8.2 Under (8.32) and (8.33), derive the recursive formula of $P_{N-k,N}$, $k = 1, 2, \cdots, N-1$, for the model (8.28).

8.3 Check the formulas (8.42) and (8.44).

8.4 Prove the equation (8.53).

8.5 If $\phi_K(\mathbf{u}) = \exp\left\{\frac{1}{2}\mathbf{u}'\mathbf{u}\right\}$ in (8.55), then give a no-arbitrage condition for $Y_{n,N}$ given by (8.53) in terms of $\alpha(\ ,\ ), \boldsymbol{\beta}(\ ,\ )$ and $\boldsymbol{\xi}$ (Kariya and Liu (2003)).

CHAPTER 9

Credit Rating

One of the most interesting topics in the field of financial engineering is the problem of credit rating. Usually credit rating has been done by use of i.i.d. settings. In this chapter we investigate problems of credit rating, based on a methodology of time series analysis discussed in Chapter 6. We develop discriminant and cluster analysis for financial time series in the case where concerned time series are locally stationary processes, and suggest applying the results to the problem of credit rating. Section 9.1 discusses a clustering problem of stock data on the New York Stock Exchange, using the parametric approach for estimation of time varying spectral densities. In Section 9.2 we develop discrimination and clustering techniques based on nonparametric time varying spectral density estimators. So far, we assumed the mean vectors are zero. However, actual financial time series shows that the mean of data smoothly changes even after taking log-returns. Therefore, Section 9.3 suggests a credit rating based on taking into account not only covariance structures but also mean structures.

9.1 Parametric Clustering for Financial Time Series

Time series analysis under stationarity has been well established. However, stationary time series models are not plausible to describe the real world. Empirical studies show that most time series data are nonstationary. Despite the fact that concerned time series has a nonstationary behavior, many researchers used the ordinary autoregressive (AR) or autoregressive moving average (ARMA) models. For this problem, Sakiyama (2002) suggests fitting a time varying autoregressive (TVAR) model of order p to data whose coefficients are polynomials with respect to time. A favorable model is selected by use of AIC and the results are applied to discriminant and cluster analysis for financial time series:

We consider the following time varying AR model

$$\sum_{j=0}^{p} a_j^{\theta}\left(\frac{t}{T}\right) X_{t-j,T} = \sigma\left(\frac{t}{T}\right) e_t, \tag{9.1}$$

where $a_0^{\theta}(u) \equiv 1$ and $\{e_t\}$ is a sequence of i.i.d. random variables with mean

zero and variance 1. Suppose that coefficient functions $a_\theta(u) = (a_1^\theta(u), \ldots, a_p^\theta(u))'$ depend on a finite dimensional parameter $\theta \in \Theta$. We further suppose that $a_j^\theta(u)$'s are parametrized as

$$a_j^\theta(u) = \sum_{k=1}^{K} \theta_{jk} u^{k-1}. \tag{9.2}$$

Let $\theta = (\theta_{11}, \ldots, \theta_{1K}, \ldots, \theta_{p1}, \ldots \theta_{pK})'$ and $b_k(u) = u^{k-1}$. Let $A \otimes B$ denote the Kronecker product of the matrices A and B.

Write $B(u) = \{b_k(u)b_l(u)\}_{k,l=1,\ldots,K}$ and $b(u) = (b_1(u), \ldots, b_K(u))'$. Letting

$$d_N^{(a)}(u, \lambda) = \sum_{t=1}^{N-1} X_{[uT]-N/2+t+1,T}^{(a)} \exp(-it\lambda), \tag{9.3}$$

we introduce a periodogram matrix $I_N(u, \lambda) = \left\{ I_N^{(ab)}(u, \lambda) : a, b = 1, \ldots, d \right\}$ over a segment of length N with midpoint $[uT]$, where

$$I_N^{(ab)}(u, \lambda) = \frac{1}{2\pi N} d_N^{(a)}(u, \lambda) d_N^{(b)}(u, -\lambda). \tag{9.4}$$

The shift from segment to segment is denoted by N. $I_N(u_s, \lambda)$ is calculated over segments with midpoints $[u_s T] = t_s = N(s - 1/2)$, $(s = 1, \ldots, M)$ where $T = NM$. We define the quasi-Gaussian likelihood as

$$L_T(\theta) = \frac{1}{4\pi M} \sum_{s=1}^{M} \int_{-\pi}^{\pi} \left[\log \{ f(u_s, \lambda) \} + I_N(u_s, \lambda) f^{-1}(u_s, \lambda) \right] d\lambda. \tag{9.5}$$

It is easily seen that the quasi-likelihood estimator $\hat{\theta}$ of θ satisfying $L_T(\hat{\theta}) = \max_{\theta \in \Theta} L_T(\theta)$ is given by

$$\begin{aligned} &\hat{\theta} \left\{ \sigma(u_1)^2, \ldots, \sigma(u_M)^2 \right\} \\ &= - \left\{ \frac{1}{M} \sum_{s=1}^{M} \sigma(u_s)^{-2} \Sigma_N(u_s) \otimes B(u_s) \right\}^{-1} \left\{ \frac{1}{M} \sum_{s=1}^{M} \sigma(u_s)^{-2} C_N(u_s) \otimes b(u_s) \right\}, \end{aligned} \tag{9.6}$$

where

$$\begin{aligned} c_N(u, j) &= \int_{-\pi}^{\pi} I_N(u, j) \exp(i\lambda j) d\lambda \\ &= \frac{1}{N} \sum_{s,t=0}^{N-1} \delta \{ s - t = j \} X_{[Tu]-N/2+s+1,T} X_{[Tu]-N/2+t+1,T}, \end{aligned} \tag{9.7}$$

$C_N(u) = (c_N(u, 1), \ldots, c_N(u, p))'$ and $\Sigma_N(u) = \{c_N(u, i - j)\}_{i,j=1,\ldots,p}$. In the definition of $\hat{\theta}$ we need a knowledge of the innovation variance $\sigma(u_s)^2$. Since we may suppose a locally stationary process $\{X_{t,T}\}$ is stationary on each time

segment with midpoint $u_s T$, it is possible to estimate $\sigma(u_s)^2$ by

$$\hat{\sigma}(u_s)^2 = 2\pi \exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \log \beta I_N(u_s, \lambda) d\lambda\right], \tag{9.8}$$

where $\beta = \exp\gamma$ ($\gamma \approx 0.57721$ Euler's constant) (see Taniguchi (1980).) Then we can estimate θ by $\hat{\theta} \equiv \theta\left\{\hat{\sigma}(u_1)^2, \ldots, \hat{\sigma}(u_M)^2\right\}$. On each time segment we can calculate the residual variance $\hat{\sigma}(u_s, p, K)^2$ of time varying AR model by means of the relation (9.1) with coefficients $a_j^{\hat{\theta}}\left(\frac{t}{T}\right)$. For a concerned time series we fit the time varying AR model (9.1) which minimizes the AIC criterion

$$\mathrm{AIC}(p, K) = \frac{1}{M}\sum_{s=1}^{M} \log \hat{\sigma}(u_s, p, K)^2 + 2\left\{p + p(K-1)\right\}/T, \tag{9.9}$$

where

$$\hat{\sigma}(u_s, p, K)^2 \approx \frac{1}{N}\sum_{t=N(s-1)+1}^{Ns}\left\{X_{t,T} + \sum_{j=1}^{p} a_j^{\hat{\theta}}(u_s) X_{t-j,T}\right\}^2. \tag{9.10}$$

Now, we discuss a clustering problem for New York Stock Exchange data. The data are daily returns of AMOCO, Ford, HP, IBM and Merck companies. The individual time series are the last 1024 data points of the daily returns for the five companies from February 2, 1984 to December 31, 1991. In Figure 1.1 we plotted the graph of the data of Hewlett-Packard. In the figure we can find changes of variance (volatility) with time. For such data, Engle (1982) introduced an autoregressive conditionally heteroscedastic (ARCH) model of order p defined as

$$X_t = e_t\sqrt{u_t}, \tag{9.11}$$

where $\{e_t\}$ is a sequence of i.i.d. $(0, 1)$ random variables, e_t is independent of X_s, $s < t$ and u_t evolves according to

$$u_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2. \tag{9.12}$$

The model in (9.11) can be rewritten in the form

$$\begin{aligned} Y_t &= u_t + \eta_t, \\ u_t &= \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i}, \end{aligned} \tag{9.13}$$

where $Y_t = X_t^2$ and $\eta_t = u_t(e_t^2 - 1)$. Henceforth, $\mathcal{F}_t$ denotes the σ-field generated by $\{X_s : s \le t\}$. We note that the disturbance term η_t in (9.13) is a martingale difference since $E\left(\eta_t|\mathcal{F}_{t-1}\right) = u_t E\left\{(e_t^2 - 1)|\mathcal{F}_{t-1}\right\} = 0$. In Figures 1.2 and 1.3, we plotted the sample autocorrelations of Hewlett-Packard data and the square transformed data, respectively. From these figures we may suppose that the time series is uncorrelated, on the other hand the square

transformed data is correlated. By using AIC we attempt to fit a stationary AR(p) model to the square transformed data. The results are shown in Table 9.1.

Table 9.1 *Suitable stationary AR model.*

	AMOCO	Ford	Hewlett-Packard	IBM	Merck
order p	2	1	23	0	8

Next, we plot the local sample mean $\hat{\mu}_T(t/T)$ and the local sample autocovariances $\hat{c}_T(t/T, k)$ of the data and the square transformed data in Figures 9.1-9.3, respectively. Here the local sample mean $\hat{\mu}_T(t/T)$ and local sample autocovariance $\hat{c}_T(t/T, k)$ are given by

$$\hat{\mu}\left(\frac{t}{T}\right) = \frac{1}{b_T T} \sum_{s=[t-b_T T/2]+1}^{[t+b_T T/2]} K\left(\frac{t-s}{b_T T}\right) X_{s,T} \tag{9.14}$$

and

$$\hat{c}\left(\frac{t}{T}, k\right) = \frac{1}{b_T T} \sum_{s=[t-b_T T/2]+1}^{[t+b_T T/2]} K\left(\frac{t-s-k/2}{b_T T}\right) \left\{X_{s,T} - \hat{\mu}\left(\frac{s}{T}\right)\right\} \left\{X_{s+k,T} - \hat{\mu}\left(\frac{s+k}{T}\right)\right\}, \tag{9.15}$$

respectively, where $K : \mathbf{R} \to [0, \infty)$ is a kernel function and b_T is a bandwidth. To simplify, we often take $K \equiv 1$, for $[-1/2, 1/2]$, and $K \equiv 0$, otherwise. From these figures we can find that all of them are time varying.

Therefore, we try to use TVAR models for the square transformed data. Selected parameters are $T = 2^{10} = 1024$, $M = 2^6 = 64$ and $N = 2^4 = 16$. Assuming that innovation variances of the data are constant over time, we choose the orders of models by minimizing the AIC criterion. The selected orders of TVAR models p and those of polynomials K are listed in Table 9.2. From the results we can see that the TVAR (1) model is preferred. Moreover, polynomial of order 4 is good for AMOCO, IBM and Merck. On the other hand, polynomials of order 6 and 2 are better for Ford and HP, respectively. For locally stationary data, it seems that the TVAR (1) model with parametric polynomial function of time shows a good fit, which makes a sharp contrast with the usual AR fitting (see Table 9.1).

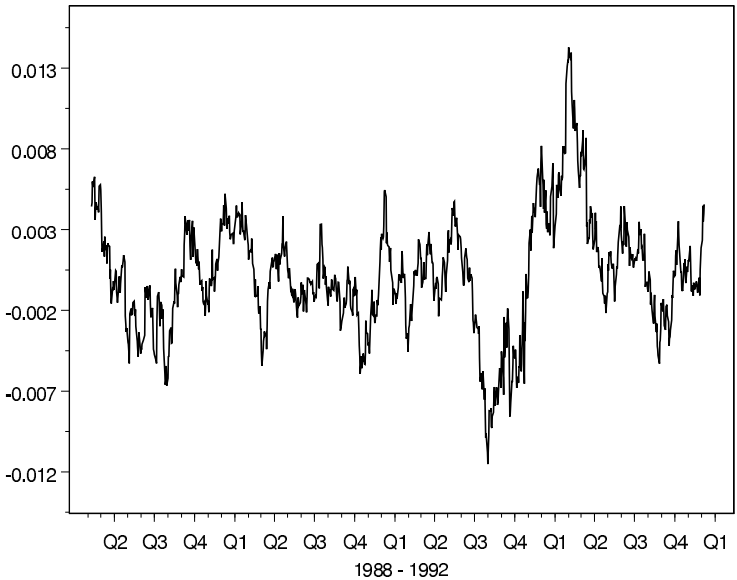


Figure 9.1 *The local sample mean (Hewlett-Packard).*

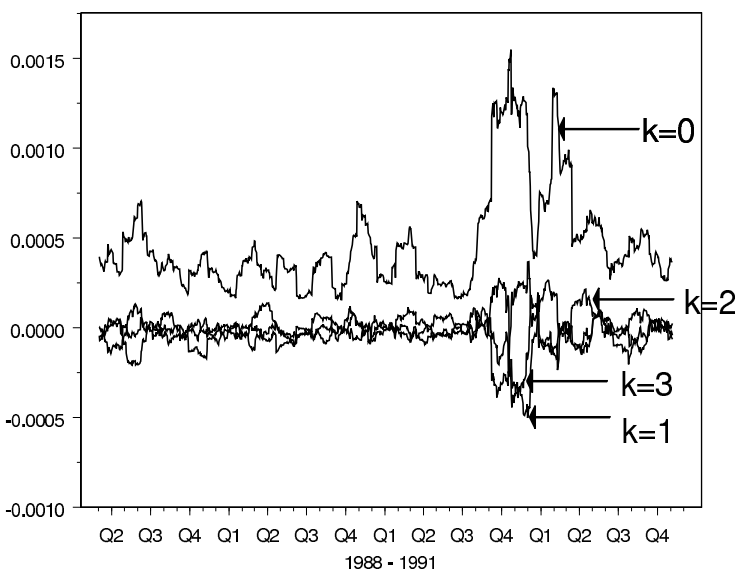


Figure 9.2 *The local sample covariance (Hewlett-Packard).*

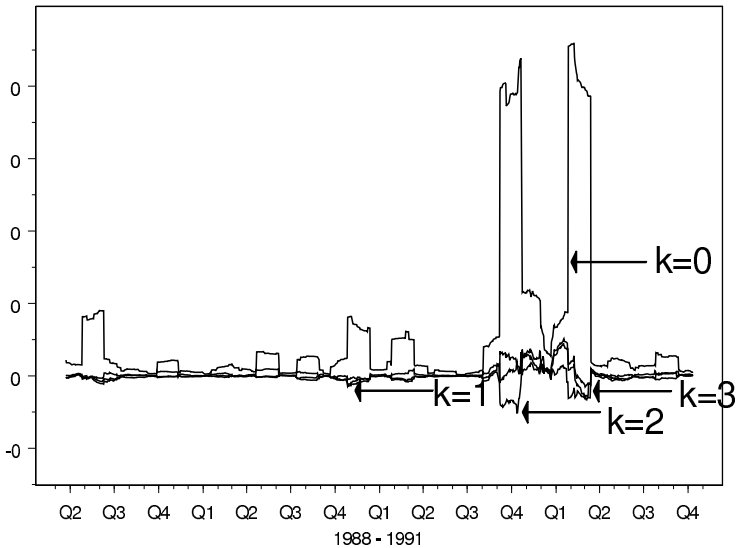


Figure 9.3 *The local sample covariance of the squared data (Hewlett-Packard).*

Table 9.2 *Suitable TVAR model.*

	AMOCO	Ford	Hewlett-Packard	IBM	Merck
order (p, K)	(1,4)	(1,6)	(1,2)	(1,4)	(1,4)

Now we discuss a clustering problem for the five companies on the New York Stock Exchange. Sakiyama and Taniguchi (2004) considered the problems of classifying a multivariate locally stationary process $\{X_{t,T}\}$ into one of two categories described by two hypotheses:

$$\Pi_1 : f(u, \lambda), \quad \Pi_2 : g(u, \lambda), \tag{9.16}$$

where $f(u, \lambda)$ and $g(u, \lambda)$ are $d \times d$ time varying spectral density matrices. For this discriminant problem, they employed the following Gaussian Kullback-Leibler information measure

$$\begin{aligned} D(f : g) = & \frac{1}{4\pi M} \sum_{s=1}^{M} \int_{-\pi}^{\pi} \Big[\log \frac{|g(u_s, \lambda)|}{|f(u_s, \lambda)|} \\ & + \operatorname{tr} \left[I_N(u_s, \lambda) \left\{ g(u_s, \lambda)^{-1} - f(u_s, \lambda)^{-1} \right\} \right] \Big] d\lambda \end{aligned} \tag{9.17}$$

as a classification statistic. That is, if $D(f : g) > 0$ we choose category Π_1. Otherwise we choose category Π_2.

In what follows, we discuss the discriminant problem in the parametric form:

$$\Pi_1 : f_\theta(u, \lambda), \quad \Pi_2 : g_\theta(u, \lambda), \tag{9.18}$$

where $f_\theta(u, \lambda)$ and $g_\theta(u, \lambda)$ are parametric time varying spectral densities. We can construct the estimated spectral density $h_{\hat{\theta}}(u, \lambda)$ from $\{X_{t,T}\}$. Let

$$D(h_\theta : g_\theta) = \frac{1}{4\pi M} \sum_{s=1}^{M} \int_{-\pi}^{\pi} \left[\log \frac{|g_\theta(u_s, \lambda)|}{|h_\theta(u_s, \lambda)|} + \operatorname{tr} \left\{ h_\theta(u_s, \lambda) g_\theta(u_s, \lambda)^{-1} \right\} - d \right] d\lambda. \tag{9.19}$$

Then the inequality

$$D(h_{\hat{\theta}} : g_\theta) > D(h_{\hat{\theta}} : f_\theta) \tag{9.20}$$

implies that $h_{\hat{\theta}}(u, \lambda)$ is nearer to $f_\theta(u, \lambda)$ than $g_\theta(u, \lambda)$ in the sense of spectral divergence measure $D(:)$. Write

$$\begin{aligned} D(h_{\hat{\theta}}) \equiv & D(h_{\hat{\theta}} : g_{\hat{\theta}}) - D(h_{\hat{\theta}} : f_{\hat{\theta}}) \\ = & \frac{1}{4\pi M} \sum_{s=1}^{M} \int_{-\pi}^{\pi} \Big[\log \frac{\left| g_{\hat{\theta}}(u_s, \lambda) \right|}{\left| f_{\hat{\theta}}(u_s, \lambda) \right|} \\ & + \operatorname{tr} \left[h_{\hat{\theta}}(u_s, \lambda) \left\{ g_{\hat{\theta}}(u_s, \lambda)^{-1} - f_{\hat{\theta}}(u_s, \lambda)^{-1} \right\} \right] \Big] d\lambda. \end{aligned} \tag{9.21}$$

We propose a rule to classify $\{X_{t,T}\}$ into Π_1 or Π_2 according as $D(h_{\hat{\theta}}) > 0$ or $D(h_{\hat{\theta}}) \leq 0$, respectively.

This measure of disparity between spectral densities can be used for clustering locally stationary processes. For example, let $f_{\hat{\theta}}$ and $g_{\hat{\theta}}$ be estimated spectral densities for two different processes. The symmetric measure is more convenient for clustering. Although $D(f_{\hat{\theta}} : g_{\hat{\theta}})$ is not symmetric with respect to $f_{\hat{\theta}}$ and $g_{\hat{\theta}}$, it can easily be made so by defining

$$\begin{aligned} \overline{D}(f_{\hat{\theta}} : g_{\hat{\theta}}) & \equiv \frac{1}{2} \left\{ D(f_{\hat{\theta}} : g_{\hat{\theta}}) + D(f_{\hat{\theta}} : g_{\hat{\theta}}) \right\} \\ & = \frac{1}{4\pi M} \sum_{s=1}^{M} \int_{-\pi}^{\pi} \operatorname{tr} \left\{ f_{\hat{\theta}}(u_s, \lambda) g_{\hat{\theta}}(u_s, \lambda)^{-1} + g_{\hat{\theta}}(u_s, \lambda) f_{\hat{\theta}}(u_s, \lambda)^{-1} \right\} d\lambda - 2d. \end{aligned} \tag{9.22}$$

Now we discuss hierarchical clustering techniques for locally stationary processes as follows. First, find the two elements which are closest in the sense of the distance (9.22). Then these two items become a cluster. Next the distance between nonclustered items and a current cluster is calculated as the average of the distances to elements in the cluster (note that many different ideas of defining the "distance between two clusters" are possible). Again, we combine the objects that are closest together, and then compute the new distance

between two different clusters. Repeat the above step until all the items are merged into one cluster. This is a hierarchical clustering method. Table 9.3 shows the results of hierarchical clustering based on (9.22) for the daily return data.

Table 9.3 *Results of hierarchical clustering based on* $\overline{D}(\cdot:\cdot)$.

No.	Minimum Distance	Hierarchical Clustering
4		(AMOCO, IBM, Ford, HP, Merck)
3	7.15908	(AMOCO, IBM, Ford, HP), (Merck)
2	6.01698	(AMOCO, IBM, Ford), (HP), (Merck)
1	3.7082	(AMOCO, IBM), (Ford), (HP), (Merck)
0	1.42466	(AMOCO), (Ford), (HP), (IBM), (Merck)

We estimate the group (cluster $\left\{f_{\hat{\theta}}^{(1)}, \ldots, f_{\hat{\theta}}^{(k)}\right\}$) spectral density by

$$\frac{1}{k}\sum_{l=1}^{k} f_{\hat{\theta}}^{(l)}. \tag{9.23}$$

Namely, it is obtained by averaging the estimators.

For the clustering problem for seismic data by use of stationary modeling, Kakizawa et al. (1998) considered the following distance measures computed from the symmetric Chernoff information divergence

$$JB_{\alpha}(f:g) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left[\log\left\{\frac{\alpha f+(1-\alpha)g}{g}\right\}+\log\left\{\frac{\alpha g+(1-\alpha)f}{f}\right\}\right]d\lambda, \tag{9.24}$$

where $f \equiv f(\lambda)$ and $g \equiv g(\lambda)$ are spectral densities for two different stationary processes. Similarly, for locally stationary processes, we can introduce a distance measure

$$DJ_{\alpha}(f_{\hat{\theta}}:g_{\hat{\theta}}) = \frac{1}{4\pi M}\sum_{s=1}^{M}\int_{-\pi}^{\pi}\left[\log\left\{\frac{\alpha f_{\hat{\theta}}(u_s,\lambda)+(1-\alpha)g_{\hat{\theta}}(u_s,\lambda)}{g_{\hat{\theta}}(u_s,\lambda)}\right\}\right.$$
$$\left.+\log\left\{\frac{\alpha g_{\hat{\theta}}(u_s,\lambda)+(1-\alpha)f_{\hat{\theta}}(u_s,\lambda)}{f_{\hat{\theta}}(u_s,\lambda)}\right\}\right]d\lambda. \tag{9.25}$$

For $\alpha = 0.5$ Table 9.4 shows the result of hierarchical clustering based on (9.25) for the daily returns data. From Tables 9.3 and 9.4, we can see that

the result for $\overline{D}(\cdot:\cdot)$ is different from that for $DJ_\alpha(\cdot:\cdot)$. Namely, in No.3 of Table 9.3 we get the cluster of AMOCO, IBM, Ford and HP. However, in No.3 of Table 9.4, AMOCO, IBM, Ford and Merck are clustered.

Table 9.4 *Results of hierarchical clustering based on* $DJ_\alpha(\cdot:\cdot)$ *(*$\alpha = 0.5$*).*

No.	Minimum Distance	Hierarchical Clustering
4		(AMOCO, IBM, Ford, Merck, HP)
3	0.335568	(AMOCO, IBM, Ford, Merck), (HP)
2	0.202399	(AMOCO, IBM, Ford), (HP), (Merck)
1	0.061422	(AMOCO, IBM), (Ford), (HP), (Merck)
0	0.020970	(AMOCO), (Ford), (HP), (IBM), (Merck)

9.2 Nonparametric Clustering for Financial Time Series

Discrimination and clustering of locally stationary processes, that can be characterized by difference in covariance or time varying spectral structures, are important in applications occurring in analysis of financial data, seismic records and biometric data, etc. Sakiyama and Taniguchi (2004) investigated the problem of classifying a multivariate non-Gaussian locally stationary process into one of two categories in which hypotheses are described in terms of time varying spectral density matrices. They used an approximation of a Gaussian Kullback-Leibler information measure as a classification statistic. In Section 6.11 we generalized this measure to nonlinear integral functional measures of time varying spectral density matrices which include Gaussian Kullback-Leibler and Chernoff information measures. In empirical studies, we find time series in real phenomena such as financial time series data and seismic record are often nonstationary and non-Gaussian. To investigate the actual performance of the clustering of such nonstationary and non-Gaussian time series will be of increasing importance. Section 9.1 discussed a clustering problem of stock returns of 5 companies on the New York Stock Exchange and employed the parametric approach for estimation of time varying spectral densities. Alternatively we consider a nonparametric approach in this section. Shumway (2003) exploited the use of Gaussian locally stationary Kullback-Leibler discrimination measure of distance for clustering earthquakes and mining explosions at regional distances. In this section, we employ generalized nonlinear integral functional symmetric measures of time varying spectra introduced in Section 6.11. Thus our clustering methods are based on nonparametric estimators of time varying spectral densities. Our nonlinear integral functional

measure includes a Gaussian Kullback-Leibler information measure as a special case and we will see that it has actually better performance when a time varying spectrum is contaminated by a sharp peak. Since our measure is a nonlinear integral functional of time varying spectral density, we have to use a kernel type nonparametric estimator. Otherwise, the integral no longer recovers $\sqrt{T}$-consistency of our estimator. Consequently, we observe that the clustering results well extract features of relationships among companies. Furthermore, we discuss robustness of the Chernoff information measure for a sharp peak of sample time varying spectrum in time domain, which corresponds to a rapid change in the spectral structure.

First, we make the following assumption on time varying spectral density $f(u, \lambda)$.

Assumption 9.1 *The time varying spectral density $f(u, \lambda)$ is bounded from below and above by some constants $\delta_1, \delta_2 > 0$ uniformly in u and λ.*

As an estimator of the time varying spectral density $f(u, \lambda)$, we use the nonparametric estimator of kernel type defined in (6.455):

$$\hat{f}_T(u, \lambda) = \int_{-\pi}^{\pi} W_{M_T}(\lambda - \mu) I_N(u, \mu) d\mu, \tag{9.26}$$

where $W_{M_T}(\omega) = M_T \sum_{\nu=-\infty}^{\infty} W\{M_T(\omega + 2\pi\nu)\}$ is a weight function.

To execute classifying and clustering given observations of locally stationary processes, first we have to decide how to measure the disparity between different locally stationary processes which have time varying spectral densities $f_i(u, \lambda)$, $i = 1, 2$. According to Kakizawa et al. (1998), in Section 6.11 we employed a generalized nonlinear integral functional disparity measure of the time varying spectral densities, defined as

$$D_H(f_j; f_k) = \frac{1}{4\pi} \int_0^1 \int_{-\pi}^{\pi} H\left\{f_j(u, \lambda) f_k^{-1}(u, \lambda)\right\} d\lambda du, \tag{9.27}$$

where $H(\cdot)$ is a smooth function of $f_j(u, \lambda) f_k^{-1}(u, \lambda)$ and $(j, k) = (1, 2)$ or $(2, 1)$. To ensure that $D_H(f_j; f_k)$ has the quasi-distance property, we require $D_H(f_j; f_k) \geq 0$ and that the equality holds if and only if $f_j(u, \lambda) = f_k(u, \lambda)$, almost everywhere.

Denote by $p_i(\mathbf{x})$, $i = 1, 2$, the probability density functions of observed time series $\mathbf{X}_T = (X_{1,T}, \ldots, X_{T,T})'$ under two hypotheses Π_i, $i = 1, 2$, respectively. Suppose that $\mathbf{X}_T$ comes from a zero mean Gaussian locally stationary process with time varying spectral density $f_i(u, \lambda)$ under Π_i. Then, it is seen that disparity measures $D_H(f_j; f_k)$ include a Gaussian Kullback-Leibler discriminant

information ratio

$$D_{H_K}(f_j; f_k) = \lim_{T\to\infty} T^{-1} \int \log p_j(\mathbf{x}) \frac{p_j(\mathbf{x})}{p_k(\mathbf{x})} d\mathbf{x}$$
$$= \frac{1}{4\pi} \int_0^1 \int_{-\pi}^{\pi} H_K \left\{ f_j(u,\lambda) f_k^{-1}(u,\lambda) \right\} d\lambda du \tag{9.28}$$

and Chernoff information measure

$$D_{H_{B_\alpha}}(f_j; f_k) = -\lim_{T\to\infty} T^{-1} \log \int p_j(\mathbf{x}) \left\{ \frac{p_k(\mathbf{x})}{p_j(\mathbf{x})} \right\}^{\alpha} d\mathbf{x}$$
$$= \frac{1}{4\pi} \int_0^1 \int_{-\pi}^{\pi} H_{B_\alpha} \left\{ f_j(u,\lambda) f_k^{-1}(u,\lambda) \right\} d\lambda du, \tag{9.29}$$

where

$$H_K(z) = z - \log(z) - 1 \tag{9.30}$$

and

$$H_{B_\alpha}(z) = \log(\alpha z + 1 - \alpha) - \alpha \log(z). \tag{9.31}$$

Note that another possible choice of functional $H(\cdot)$ is a quadratic function

$$H_Q(z) = \frac{1}{2}(z-1)^2. \tag{9.32}$$

The disparity measure $D_H(f_j; f_k)$ is not a mathematical distance because it is not symmetric and does not satisfy triangle inequality. For cluster problems, as we mentioned in the previous section, it is more convenient to use the symmetric information divergence

$$\tilde{D}_H(f_j; f_k) = D_H(f_j; f_k) + D_H(f_k; f_j)$$
$$= \frac{1}{4\pi} \int_0^1 \int_{-\pi}^{\pi} \tilde{H} \left\{ f_j(u,\lambda) f_k^{-1}(u,\lambda) \right\} d\lambda du, \tag{9.33}$$

where

$$\tilde{H}_K(z) = z + z^{-1} - 2, \quad \tilde{H}_{B_\alpha}(z) = \log(\alpha z + 1 - \alpha) + \log(\alpha z^{-1} + 1 - \alpha) \tag{9.34}$$

and

$$\tilde{H}_Q(z) = \frac{1}{2}(z-1)^2 + \frac{1}{2}(z^{-1}-1)^2. \tag{9.35}$$

Now, we execute the clustering of stock returns on the Tokyo Stock Exchange. The data are daily log-returns of 13 companies: 1. HITACHI, 2. MATSUSHITA, 3. SHARP, 4. SONY, 5. DENSO, 6. KYOCERA, 7. NISSAN, 8. TOYOTA, 9. HONDA, 10. CANON, 11. NTT, 12. KDDI, 13. NTTDOCOMO. The individual time series are 1174 data points between December 28, 1999 and October 1, 2004. Table 9.5 describes the details of stock data of the Tokyo

Stock Exchange. The last column of Table 9.5 is the stock prices of each company on October 1, 2004, which consist of (stock price)×(stock unit), so they are minimum unit prices which one can buy.

Table 9.5 *Stock data of the Tokyo Stock Exchange.*

	Company Code	Name	Industry	Price
1	6501	HITACHI, LTD.	Electric Appliances	664 ×1000
2	6752	MATSUSHITA ELECTRIC IND. CO., LTD.	Electric Appliances	1487 ×1000
3	6753	SHARP CORP.	Electric Appliances	1512 ×1000
4	6758	SONY CORP.	Electric Appliances	3780 ×100
5	6902	DENSO CORP.	Transportation Equipment	2655 ×100
6	6971	KYOCERA CORP.	Electric Appliances	7880 ×100
7	7201	NISSAN MOTOR CO., LTD.	Transportation Equipment	1210 ×100
8	7203	TOYOTA MOTOR CORP.	Transportation Equipment	4210 ×100
9	7267	HONDA MOTOR CO., LTD.	Transportation Equipment	5420 ×100
10	7751	CANON INC.	Electric Appliances	5210 ×100
11	9432	NIPPON TELEGRAPH & TELEPHONE CORP.	Information & Communication	442000
12	9433	KDDI CORP.	Information & Communication	535000
13	9437	NTT DOCOMO, INC.	Information & Communication	192000

For nonparametric estimators of the time varying spectral densities, we use the following weight function

$$W(x) = \begin{cases} \frac{\pi}{2}\cos(\pi x) & x \in [-\frac{1}{2}, \frac{1}{2}] \\ 0 & \text{otherwise.} \end{cases} \tag{9.36}$$

To simplify the calculation, we set $h(x) \equiv 1$. The nonparametric time varying spectral density estimator of MATSUSHITA is plotted in Figure 9.4, where the selected parameters are $T = 1000$, $N = 175$ and $M = 8$. From this figure it is seen that the spectral structure changes as time changes.

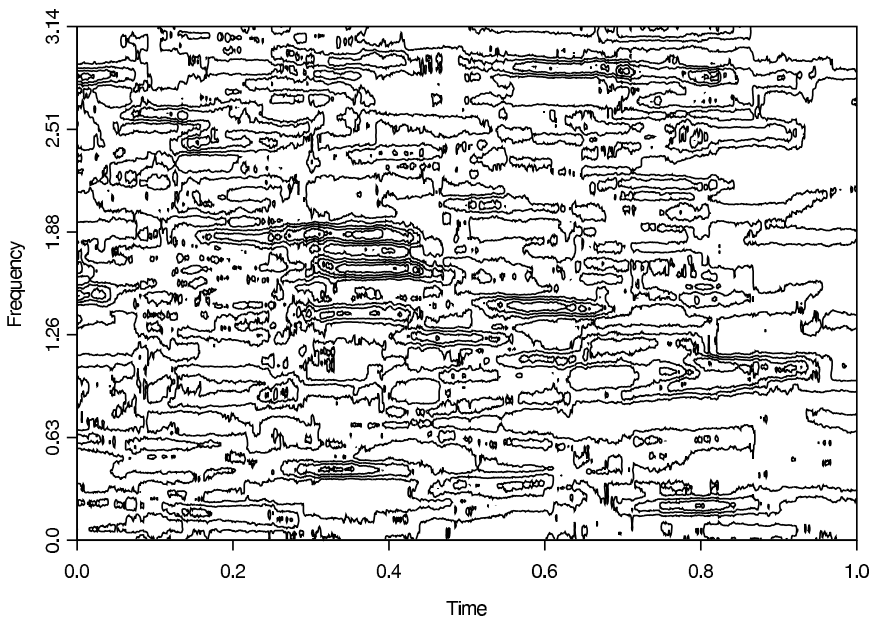


Figure 9.4 *Nonparametric time varying spectral estimator of MATSUSHITA.*

We compute distance between two different stock returns via the measures of disparity given by integral functional of the nonparametric time varying spectral density estimators:

$$\begin{aligned} \tilde{D}_H(\hat{f}_j; \hat{f}_k) &= \frac{1}{4\pi} \int_0^1 \int_{-\pi}^{\pi} \tilde{H} \left\{ \frac{\hat{f}_j(u, \lambda)}{\hat{f}_k(u, \lambda)} \right\} d\lambda du \\ &\approx \frac{1}{2T^2} \sum_{t=1}^{T} \sum_{s=1}^{T} \tilde{H} \left\{ \frac{\hat{f}_j(\frac{t}{T}, \frac{2\pi s}{T})}{\hat{f}_k(\frac{t}{T}, \frac{2\pi s}{T})} \right\}. \end{aligned} \tag{9.37}$$

Table 9.6 shows the quasi-distance matrix for a Kullback-Leibler measure. First, we assign two elements of minimum distance into one cluster. In this case, the first cluster is {2. MATSUSHITA, 3. SHARP} with distance 0.280. Next we can iterate this procedure and it is seen that the second cluster is {7. NISSAN, 9. HONDA} with distance 0.290. Furthermore, we define the distance between clusters as the average of the distances between each element of the clusters. For instance, the distance between {2. MATSUSHITA, 3. SHARP} and {1. HITACHI, 4. SONY} is given by the sum divided by 4 of the {(1,2), (1,3), (2,4) and (3,4)}th elements of matrix, and is 0.328. Then, we can iteratively define the distances between all the clusters. In Figures 9.5 and 9.6, the clusters identified by distance values of the Gaussian Kullback-Leibler disparity measure and Chernoff measures with $\alpha = 0.3$, respectively, are displayed as dendrograms. In these figures, each height denotes the distance between clusters.

Table 9.6 *The Kullback-Leibler quasi-distance matrix.*

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.000	0.363	0.300	0.296	0.699	0.420	0.458	0.554	0.382	0.431	0.444	0.664	0.414
2	0.363	0.000	0.280	0.347	0.458	0.603	0.508	0.412	0.417	0.422	0.507	1.070	0.653
3	0.300	0.280	0.000	0.302	0.510	0.381	0.391	0.466	0.311	0.316	0.360	0.695	0.438
4	0.296	0.347	0.302	0.000	0.712	0.404	0.359	0.517	0.318	0.381	0.389	0.588	0.388
5	0.699	0.458	0.510	0.712	0.000	0.818	0.599	0.437	0.481	0.557	0.616	1.290	0.877
6	0.420	0.603	0.381	0.404	0.818	0.000	0.513	0.867	0.448	0.398	0.396	0.437	0.292
7	0.458	0.508	0.391	0.359	0.599	0.513	0.000	0.537	0.290	0.376	0.422	0.633	0.487
8	0.554	0.412	0.466	0.517	0.437	0.867	0.537	0.000	0.408	0.520	0.611	1.291	0.807
9	0.382	0.417	0.311	0.318	0.481	0.448	0.290	0.408	0.000	0.314	0.398	0.670	0.421
10	0.431	0.422	0.316	0.381	0.557	0.398	0.376	0.520	0.314	0.000	0.387	0.619	0.412
11	0.444	0.507	0.360	0.389	0.616	0.396	0.422	0.611	0.398	0.387	0.000	0.518	0.292
12	0.664	1.070	0.695	0.588	1.290	0.437	0.633	1.291	0.670	0.619	0.518	0.000	0.357
13	0.414	0.653	0.438	0.388	0.877	0.292	0.487	0.807	0.421	0.412	0.292	0.357	0.000

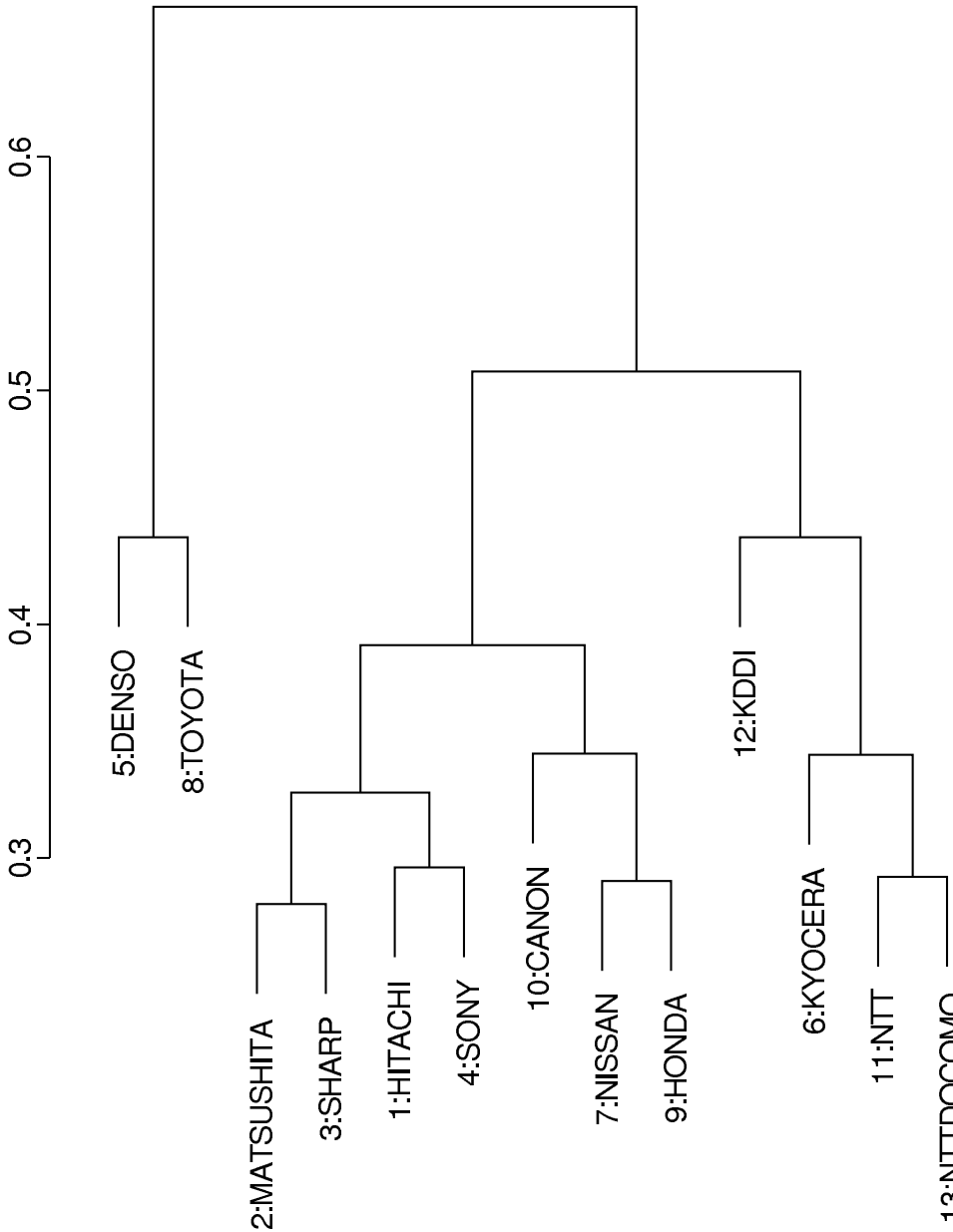


Figure 9.5 *Result of average distance hierarchical clustering using a Kullback-Leibler disparity measure.*

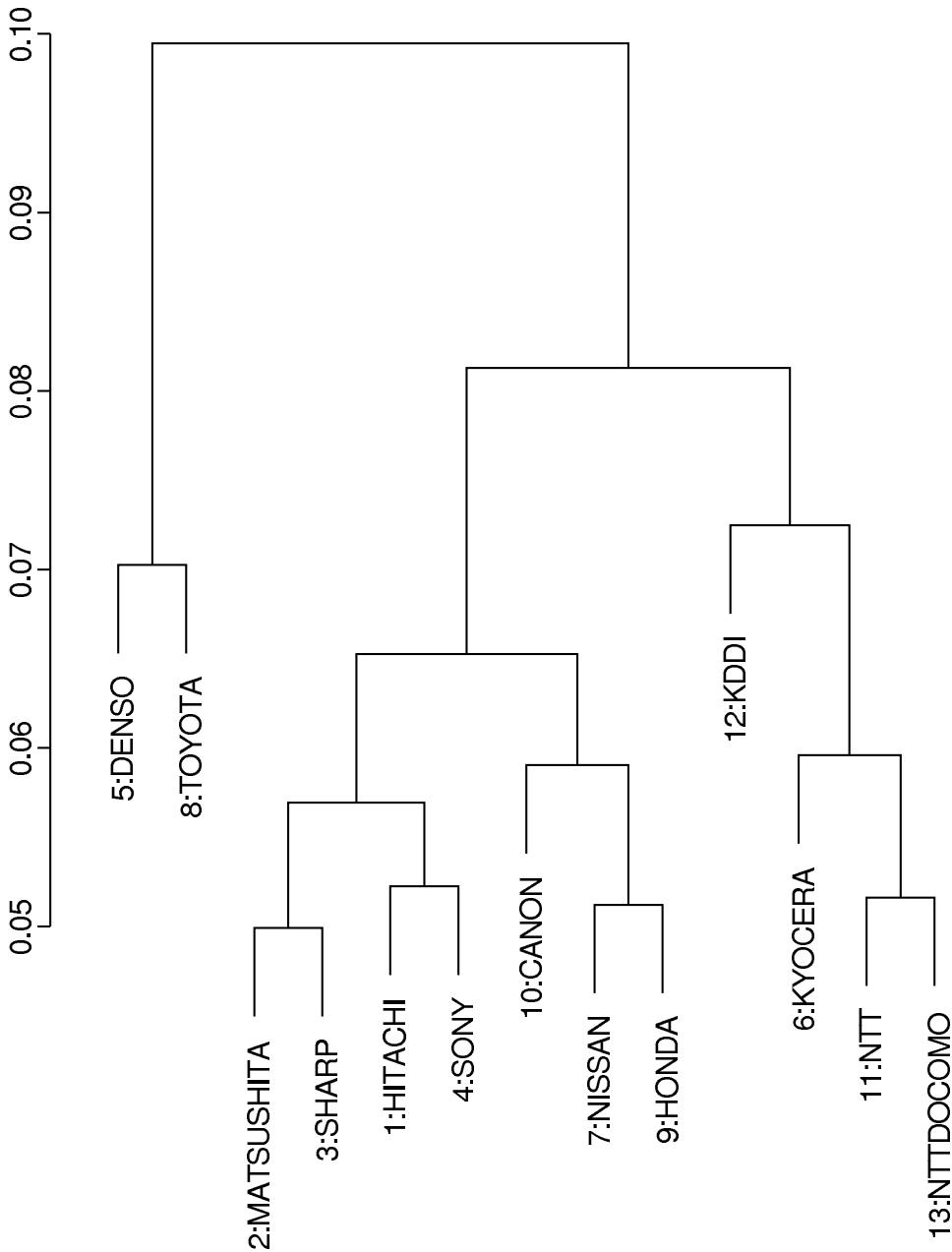


Figure 9.6 *Result of average distance hierarchical clustering using a Chernoff disparity measure with* $\alpha = 0.3$.

Both results show that the first main cluster is {1:HITACHI, 2:MATSUSHITA, 3:SHARP, 4:SONY}, which are electric appliances companies. Furthermore, the distances between NISSAN and HONDA, which are transportation equipment companies, and between NTT and NTTDOCOMO, which are information and communication companies, are close, respectively. Note that DENSO and TOYOTA, which are a kind of affiliated companies, are classified into one category. Taking into account all the above results, we can conclude our clustering methods work well.

Next, we turn to discuss peak robustness of the Chernoff measure $\tilde{D}_{H_{B_\alpha}}$. We consider the case that sample time varying spectral density of $\mathbf{X}_T$ is contaminated by a sharp peak in the time domain. Such a case corresponds to one of the phenomena of rapid changes in spectral structure. We shall prove that $\tilde{D}_{H_{B_\alpha}}(f_j; f_k)$ is robust with respect to peak, but $\tilde{D}_{H_K}(f_j; f_k)$ is not so. Define

$$\bar{f}_i(u,\lambda) = \begin{cases} f_i(u,\lambda) & \text{if } u \in \Omega = [0,1] - \Omega_\epsilon; \\ f_i(u,\lambda)/\epsilon^r & \text{if } u \in \Omega_\epsilon, \end{cases} \tag{9.38}$$

where $\Omega_\epsilon = [u_0, u_0 + \epsilon]$ is an interval in $[0,1]$ for sufficiently small $\epsilon > 0$ and $r > 1$. Suppose that $f_1(u,\lambda) \not\equiv f_2(u,\lambda)$ on a set of a positive Lebesgue measure. Then, under Assumption 9.1, it is seen that

$$\begin{aligned} &\tilde{D}_{H_{B_\alpha}}(\bar{f}_j; f_k) - \tilde{D}_{H_{B_\alpha}}(f_j; f_k) \\ &= \frac{1}{4\pi}\int_{\Omega_\epsilon}\int_{-\pi}^{\pi}\Big[\log\{f_j(u,\lambda)f_k^{-1}(u,\lambda) + (1-\alpha)\epsilon^r\} - r\log(\epsilon) \\ &+ \log\{\epsilon^r f_k(u,\lambda)f_j^{-1}(u,\lambda) + 1 - \alpha\}\Big]d\lambda du - \frac{1}{4\pi}\int_{\Omega_\epsilon}\int_{-\pi}^{\pi}\tilde{H}_{B_\alpha}(f_j; f_k)d\lambda du, \end{aligned} \tag{9.39}$$

which converges to zero, as $\epsilon \to 0$ (Exercise 9.1). On the other hand we have

$$\begin{aligned} &\tilde{D}_{H_K}(\bar{f}_j; f_k) - \tilde{D}_{H_K}(f_j; f_k) \\ &= \frac{1}{4\pi}\int_{\Omega_\epsilon}\int_{-\pi}^{\pi}\left\{\epsilon^{-r}f_j(u,\lambda)f_k^{-1}(u,\lambda) + \epsilon^r f_k(u,\lambda)f_j^{-1}(u,\lambda)\right\}d\lambda du \\ &\quad - \frac{1}{4\pi}\int_{\Omega_\epsilon}\int_{-\pi}^{\pi}\tilde{H}_K(f_j; f_k)d\lambda du, \end{aligned} \tag{9.40}$$

which diverges, as $\epsilon \to 0$ (Exercise 9.2). Therefore we can see that $\tilde{D}_{H_{B_\alpha}}$ is insensitive to a peak in the spectrum, while $\tilde{D}_{H_K}$ is sensitive. Thus, $\tilde{D}_{H_{B_\alpha}}$ is better than $\tilde{D}_{H_K}$ if the sample spectrum is contaminated by a sharp peak.

Now, we assume that the sample spectrum of MATSUSHITA is contaminated by a sharp peak, that is, the sample spectrum of MATSUSHITA is given by (9.38) with $u_0 = 0.5$, $\epsilon = 0.001$ and $r = 1.5$, which is plotted in Figure 9.7. For this case, the clusters identified by each distance value are displayed as dendrograms in Figures 9.8-9.11. It is seen that a Chernoff measure with α not close to zero works well, while a Kullback-Leibler measure does not work.

These results agree with the theoretical result, because the Chernoff measure tends to the Kullback-Leibler measure as $\alpha \to 0$.

Usually, the credit rating has been done by use of i.i.d. settings. Our clustering method, in contrast, suggests credit rating based on non-Gaussian and nonstationary settings. This would have high potential for the future developments of this field.

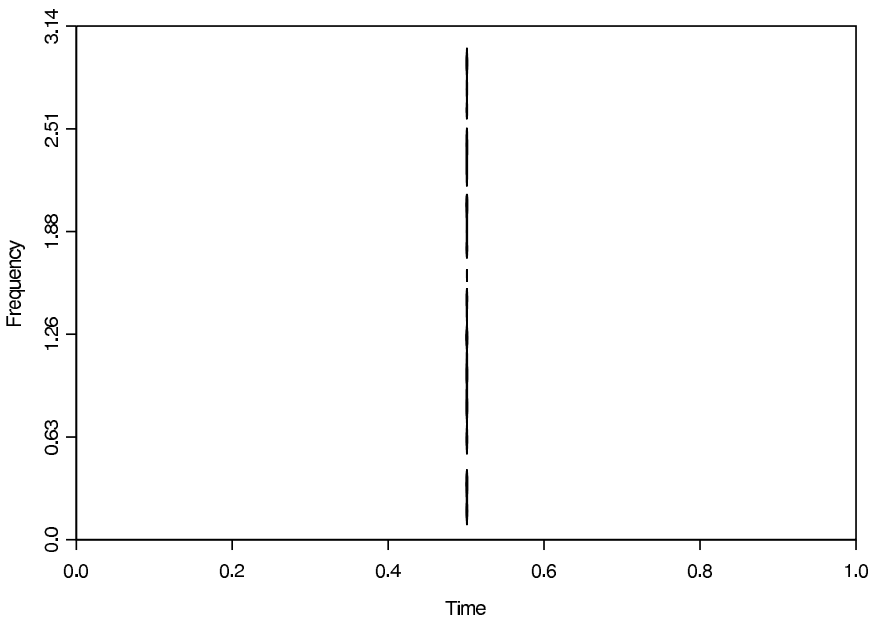


Figure 9.7 *The sample spectrum of MATSUSHITA contaminated by a sharp peak.*

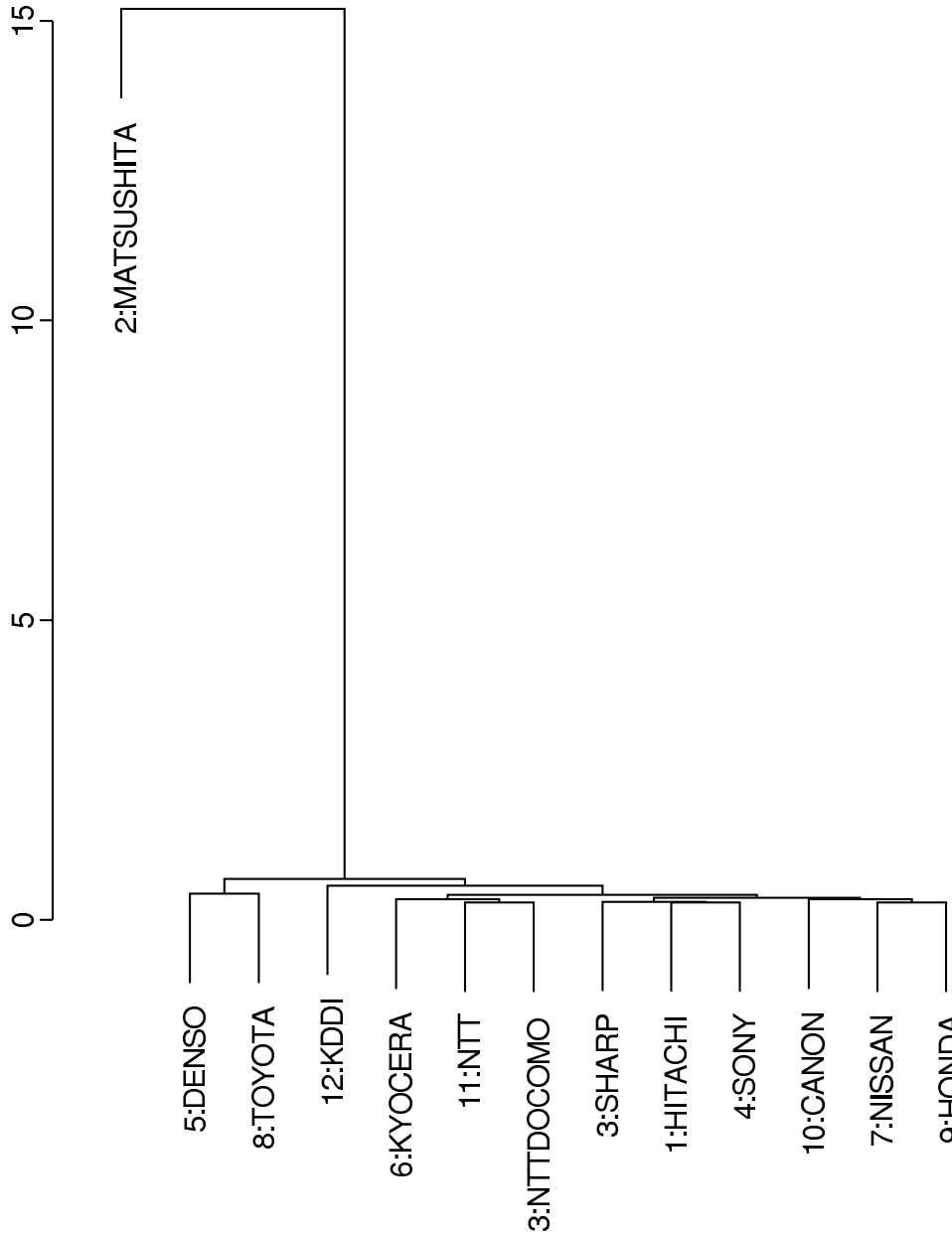


Figure 9.8 *Result of average distance hierarchical clustering using a Kullback-Leibler disparity measure (MATSUSHITA is contaminated by a sharp peak).*

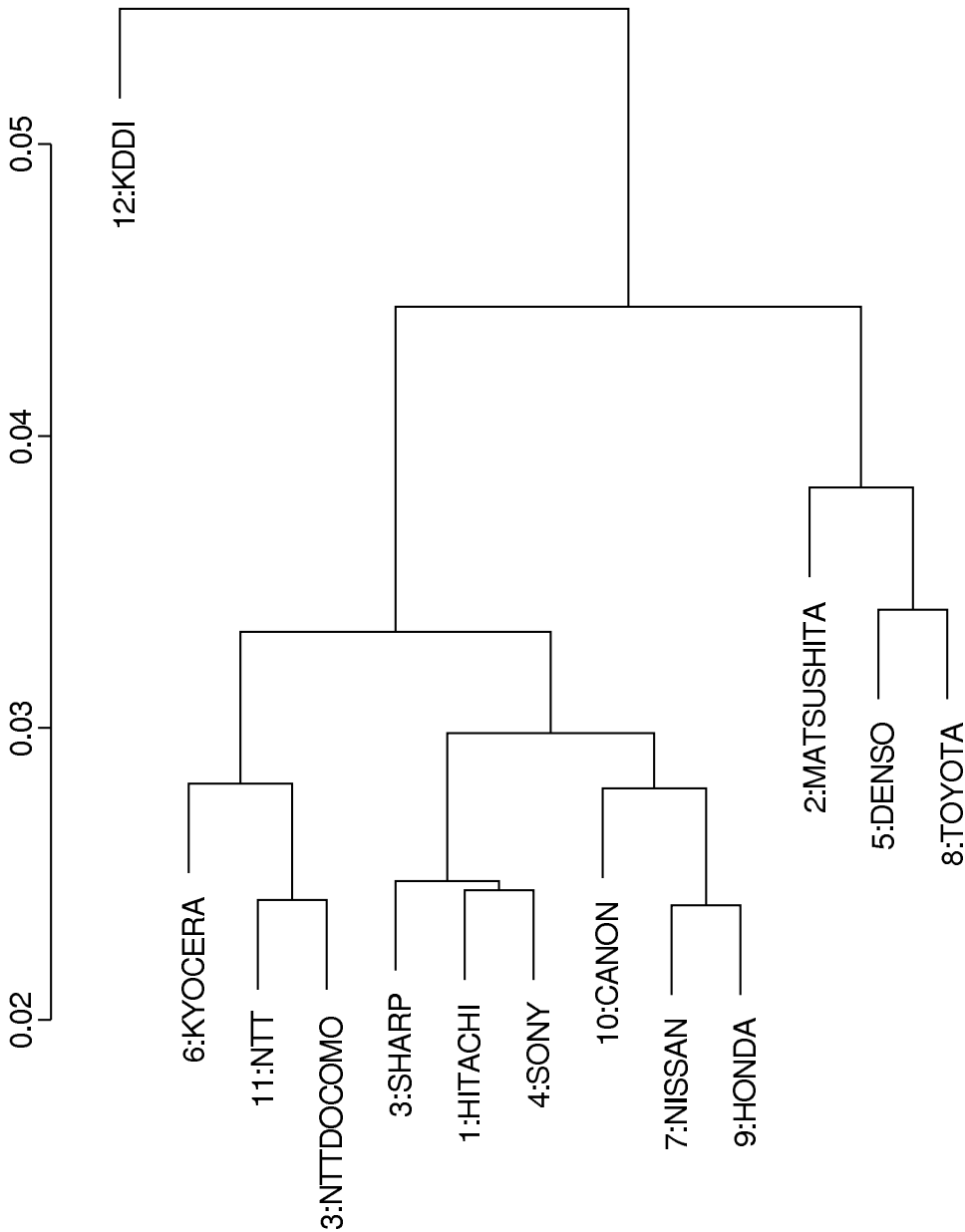


Figure 9.9 *Result of average distance hierarchical clustering using a Chernoff disparity measure with* $\alpha = 0.1$ *(MATSUSHITA is contaminated by a sharp peak).*

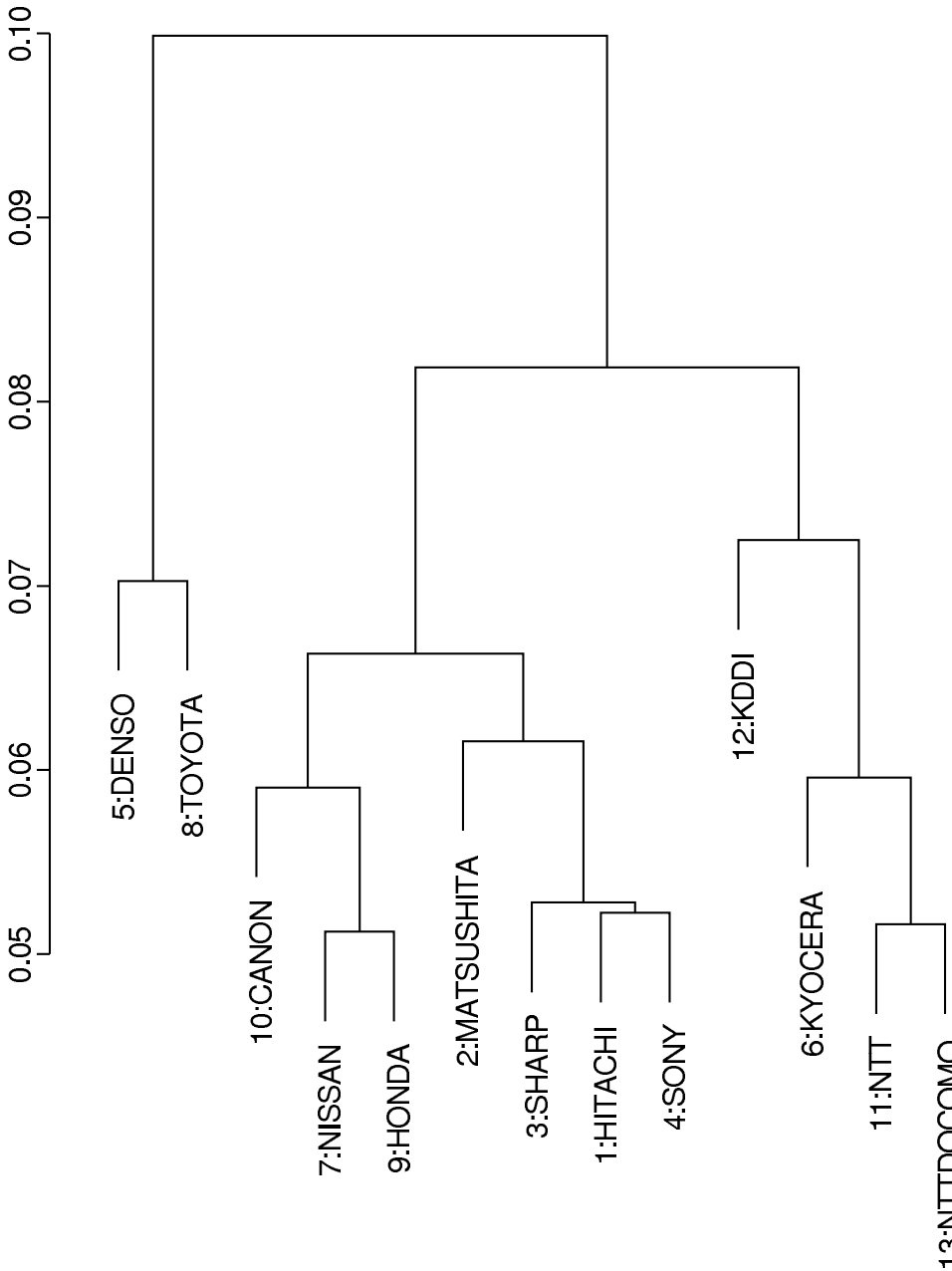


Figure 9.10 *Result of average distance hierarchical clustering using a Chernoff disparity measure with* $\alpha = 0.3$ *(MATSUSHITA is contaminated by a sharp peak).*

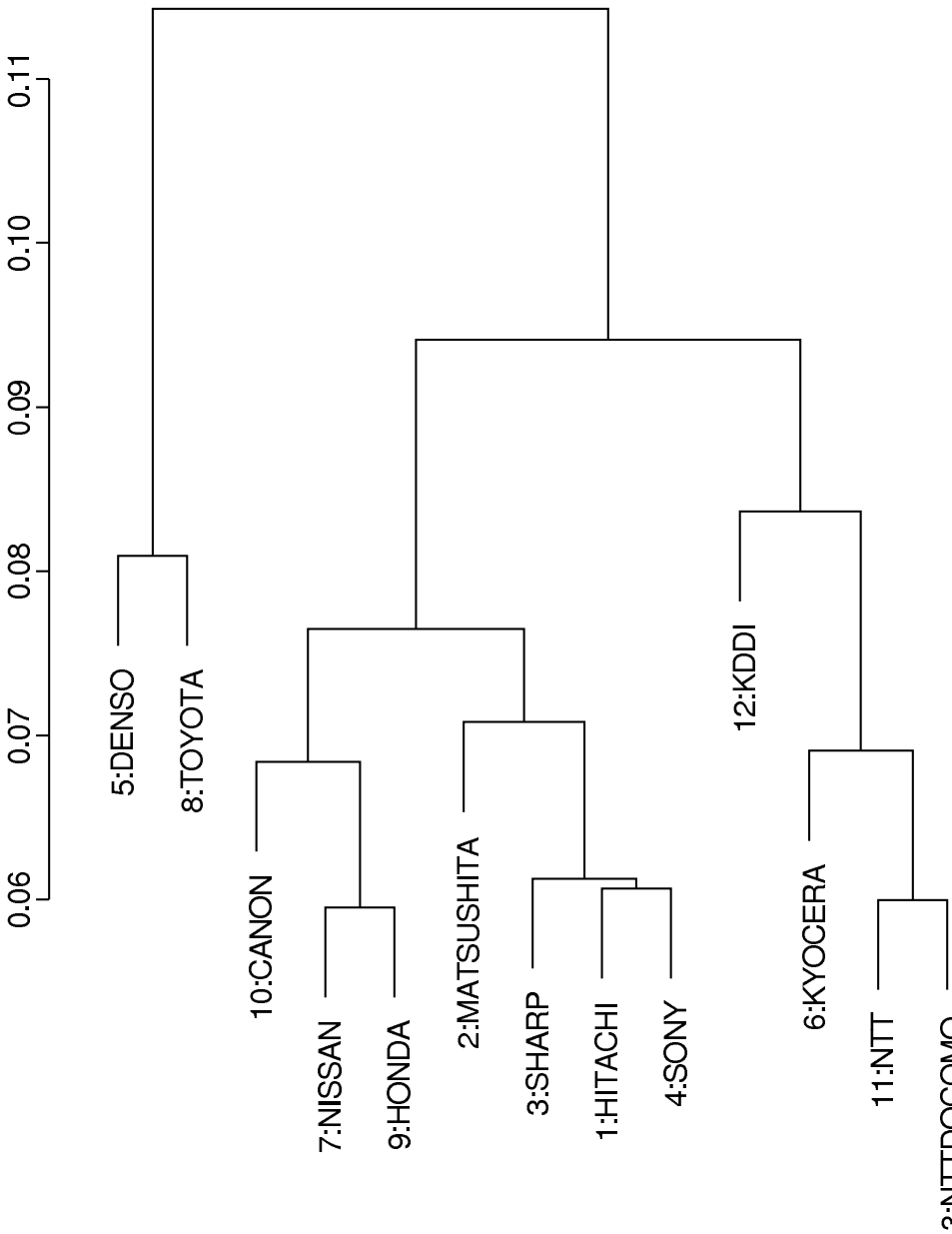


Figure 9.11 *Result of average distance hierarchical clustering using a Chernoff disparity measure with* $\alpha = 0.5$ *(MATSUSHITA is contaminated by a sharp peak).*

9.3 Credit Rating Based on Financial Time Series

So far, we discussed credit rating based on only covariance (or spectral) structures of financial time series and assumed mean vectors are zero. This restriction was approximately ensured from taking log-returns of data. However, in the actual financial time series, we observe that the mean of data smoothly changes even after taking log-returns. (See Figure 1.1.) In what follows, therefore, we consider a credit rating based on taking into account not only covariance structures but also mean structures.

Suppose that a Gaussian locally stationary process $\{X_{t,T}\}$ belongs to one of two categories, Π_1 or Π_2. The category Π_j specifies that the $T \times 1$ time series

$$\mathbf{X}_T = (X_{1,T}, \ldots, X_{T,T})' \tag{9.41}$$

is a Gaussian locally stationary with mean vector

$$\boldsymbol{\mu}_T^{(j)} = \left\{ \mu^{(j)}\left(\frac{1}{T}\right), \ldots, \mu^{(j)}\left(\frac{T}{T}\right) \right\}' \tag{9.42}$$

and covariance matrix

$$\boldsymbol{\Sigma}_T^{(j)} = \left[\int_{-\pi}^{\pi} A_{s,T}^{\circ}{}^{(j)}(\lambda) A_{t,T}^{\circ}{}^{(j)}(-\lambda) \exp\{i(s-t)\lambda\} d\lambda \right]_{s,t=1,\ldots,T}. \tag{9.43}$$

The phrase "$\mathbf{X}_T$ is locally stationary with mean vector $\boldsymbol{\mu}_T$ and covariance matrix $\boldsymbol{\Sigma}_T$" means "$\mathbf{X}_T - \boldsymbol{\mu}_T$ is locally stationary with mean zero and covariance matrix $\boldsymbol{\Sigma}_T$", whereas $E(X_{t,T}) = \mu\left(\frac{t}{T}\right)$ may depend on time t. The probability density of $\mathbf{X}_T$ under Π_j is

$$p_j(\mathbf{x}) = (2\pi)^{-T/2} \left|\boldsymbol{\Sigma}_T^{(j)}\right|^{-1/2} \exp\left\{ -\frac{1}{2} \left(\mathbf{x} - \boldsymbol{\mu}_T^{(j)}\right)' \boldsymbol{\Sigma}_T^{(j)^{-1}} \left(\mathbf{x} - \boldsymbol{\mu}_T^{(j)}\right) \right\}, \tag{9.44}$$

so the likelihood-based rule implies that we should assign $\mathbf{X}_T$ to Π_1 if

$$-\frac{1}{2}\left\{ \log \frac{\left|\boldsymbol{\Sigma}_T^{(1)}\right|}{\left|\boldsymbol{\Sigma}_T^{(2)}\right|} + \left(\mathbf{X}_T - \boldsymbol{\mu}_T^{(1)}\right)' \boldsymbol{\Sigma}_T^{(1)^{-1}} \left(\mathbf{X}_T - \boldsymbol{\mu}_T^{(1)}\right) - \left(\mathbf{X}_T - \boldsymbol{\mu}_T^{(2)}\right)' \boldsymbol{\Sigma}_T^{(2)^{-1}} \left(\mathbf{X}_T - \boldsymbol{\mu}_T^{(2)}\right) \right\} > 0. \tag{9.45}$$

We begin with the standard time domain method for the equal covariance matrices case ($\boldsymbol{\Sigma}_T^{(1)} = \boldsymbol{\Sigma}_T^{(2)}$). In the case of $\boldsymbol{\Sigma}_T^{(1)} = \boldsymbol{\Sigma}_T^{(2)} = \boldsymbol{\Sigma}_T$, the problem is to identify mean functions. This setting includes the signal detection problem which corresponds to $\boldsymbol{\mu}_T^{(1)} = \boldsymbol{\mu}_T$ (a deterministic signal) and $\boldsymbol{\mu}_T^{(2)} = \mathbf{0}$. Then the criterion (9.45) can be expressed in terms of the *linear discriminant*

function

$$D_L(\mathbf{X}_T) = \left(\boldsymbol{\mu}_T^{(1)} - \boldsymbol{\mu}_T^{(2)}\right)' {\boldsymbol{\Sigma}_T}^{-1} \mathbf{X}_T - \frac{1}{2}\boldsymbol{\mu}_T^{(1)\prime} {\boldsymbol{\Sigma}_T}^{-1} \boldsymbol{\mu}_T^{(1)} + \frac{1}{2}\boldsymbol{\mu}_T^{(2)\prime} {\boldsymbol{\Sigma}_T}^{-1} \boldsymbol{\mu}_T^{(2)} \tag{9.46}$$

and $\mathbf{X}_T$ is classified into Π_1 or Π_2 according to $D_L(\mathbf{X}_T) > 0$ or $D_L(\mathbf{X}_T) \leq 0$. It is easy to show that the discriminant function $D_L(\mathbf{X}_T)$ is normally distributed with mean $(-1)^{j+1} d_T^2/2$ and variance d_T^2 under Π_j, where

$$d_T^2 = \left(\boldsymbol{\mu}_T^{(1)} - \boldsymbol{\mu}_T^{(2)}\right)' {\boldsymbol{\Sigma}_T}^{-1} \left(\boldsymbol{\mu}_T^{(1)} - \boldsymbol{\mu}_T^{(2)}\right) \tag{9.47}$$

is the Mahalanobis distance between the two populations Π_1 and Π_2. The two misclassification probabilities are

$$\begin{aligned} P(2|1) &= \Pr\{D_L(\mathbf{X}_T) \leq 0 | \Pi_1\} \\ =P(1|2) &= \Pr\{D_L(\mathbf{X}_T) > 0 | \Pi_2\} = \Phi\left(-\frac{d_T}{2}\right), \end{aligned} \tag{9.48}$$

where $\Phi(\cdot)$ denotes the distribution function of the standard normal distribution. We see that the performance of these probabilities is strictly decreasing in d_T. The asymptotic behavior of these error rates depends on that of d_T^2, more precisely, on the sequence of mean differences

$$\boldsymbol{\delta}_T = \left\{\delta\left(\frac{1}{T}\right), \ldots, \delta\left(\frac{T}{T}\right)\right\}' \equiv \boldsymbol{\mu}_T^{(1)} - \boldsymbol{\mu}_T^{(2)}. \tag{9.49}$$

By Lemma A.2, under Assumption A.2, we have

$$\lim_{T\to\infty} T^{-1} d_T^2 = \frac{1}{2\pi} \int_0^1 \frac{\delta(u)^2}{f(u,0)} du, \tag{9.50}$$

that is,

$$d_T \approx \sqrt{\frac{T}{2\pi} \int_0^1 \frac{\delta(u)^2}{f(u,0)} du} \tag{9.51}$$

and the misclassification probability $\Phi(-d_T/2)$ tends to zero as $T \to \infty$. For this property, we say the discriminant function is *consistent*.

Next, we consider a case that the mean functions are equal. In this case, the problem is to identify the covariance matrix that specifies the structure of the second-order locally stationary process. This problem is reduced to the identification of time varying spectral density when it exists. It is convenient to let $\boldsymbol{\mu}_T^{(1)} = \boldsymbol{\mu}_T^{(2)} = \mathbf{0}$. Then criterion (9.45) becomes the *quadratic discriminant function*

$$D_Q(\mathbf{X}_T) = -\frac{1}{2}\left\{\log\frac{|\boldsymbol{\Sigma}_T^{(1)}|}{|\boldsymbol{\Sigma}_T^{(2)}|} + \mathbf{X}_T' \left({\boldsymbol{\Sigma}_T^{(1)}}^{-1} - {\boldsymbol{\Sigma}_T^{(2)}}^{-1}\right) \mathbf{X}_T\right\} \tag{9.52}$$

with the rule being to classify $\mathbf{X}_T$ into Π_1 if $D_Q(\mathbf{X}_T) > 0$. Contrary to the

linear discriminant function $D_L(\mathbf{X}_T)$, the distribution of $D_Q(\mathbf{X}_T)$ under Π_j is a linear combination of central χ^2-variables where the coefficients are the eigenvalues of a $T \times T$ matrix $\mathbf{\Sigma}_T^{(j)} \left(\mathbf{\Sigma}_T^{(1)^{-1}} - \mathbf{\Sigma}_T^{(2)^{-1}} \right)$ for $j = 1, 2$. Thus it is difficult to determine the theoretical misclassification probabilities in this case. For the situation where the dimensionality T is moderately large, as is often the case of time series analysis, the normal approximation enables us to examine the performance of two misclassification probabilities. It is easy to see that the sth cumulant of

$$\frac{1}{\sqrt{T}} \left[\mathbf{X}_T' \left(\mathbf{\Sigma}_T^{(1)^{-1}} - \mathbf{\Sigma}_T^{(2)^{-1}} \right) \mathbf{X}_T - \mathrm{tr} \left\{ \mathbf{\Sigma}_T^{(j)} \left(\mathbf{\Sigma}_T^{(1)^{-1}} - \mathbf{\Sigma}_T^{(2)^{-1}} \right) \right\} \right] \tag{9.53}$$

under Π_j is given by

$$T^{-s/2} 2^{s-1} (s-1)! \mathrm{tr} \left\{ \mathbf{\Sigma}_T^{(j)} \left(\mathbf{\Sigma}_T^{(1)^{-1}} - \mathbf{\Sigma}_T^{(2)^{-1}} \right) \right\}^s, \tag{9.54}$$

which implies the following.

Proposition 9.1 *Assume $f^{(1)}(u, \lambda) \not\equiv f^{(2)}(u, \lambda)$ on a set of positive Lebesgue measures. Under Π_j,*

$$\frac{1}{\sqrt{T}} \left[D_Q(\mathbf{X}_T) - E_j \left\{ D_Q(\mathbf{X}_T) \right\} \right] \xrightarrow{d} N(0, v_j^2), \tag{9.55}$$

where

$$\begin{aligned} v_j^2 &= \lim_{T \to \infty} \frac{1}{2T} \mathrm{tr} \left\{ \mathbf{\Sigma}_T^{(j)} \left(\mathbf{\Sigma}_T^{(1)^{-1}} - \mathbf{\Sigma}_T^{(2)^{-1}} \right) \right\}^2 \\ &= \frac{1}{4\pi} \int_0^1 \int_{-\pi}^{\pi} f^{(j)}(u, \lambda)^2 \left\{ \frac{1}{f^{(1)}(u, \lambda)} - \frac{1}{f^{(2)}(u, \lambda)} \right\}^2 d\lambda du \end{aligned} \tag{9.56}$$

(see, Lemma A.2).

Combining (9.55) with

$$\begin{aligned} &\lim_{T \to \infty} T^{-1} E_j \left\{ D_Q(\mathbf{X}_T) \right\} \\ &= \lim_{T \to \infty} -\frac{1}{2T} \left[\log \frac{|\mathbf{\Sigma}_T^{(1)}|}{|\mathbf{\Sigma}_T^{(2)}|} + \mathrm{tr} \left\{ \mathbf{\Sigma}_T^{(j)} \left(\mathbf{\Sigma}_T^{(1)^{-1}} - \mathbf{\Sigma}_T^{(2)^{-1}} \right) \right\} \right] \\ &= -\frac{1}{4\pi} \int_0^1 \int_{-\pi}^{\pi} \left[\log \frac{f^{(1)}(u, \lambda)}{f^{(2)}(u, \lambda)} + f^{(j)}(u, \lambda) \left\{ \frac{1}{f^{(1)}(u, \lambda)} - \frac{1}{f^{(2)}(u, \lambda)} \right\} \right] d\lambda du \\ &\equiv m_j \end{aligned} \tag{9.57}$$

(see Lemmas A.2 and A.3), one may approximate the two misclassification probabilities of the rule using $D_Q(\mathbf{X}_T)$ as

$$P(2|1) = \Pr \left\{ D_Q(\mathbf{X}_T) \leq 0 | \Pi_1 \right\} \approx \Phi \left(-\frac{\sqrt{T} m_1}{v_1} \right) \tag{9.58}$$

and

$$P(1|2) = \Pr\{D_Q(\mathbf{X}_T) > 0|\Pi_2\} \approx 1 - \Phi\left(-\frac{\sqrt{T}m_2}{v_2}\right) \tag{9.59}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. Since $m_1 > 0$ and $m_2 < 0$, these probabilities ((9.58) and (9.59)) tend to zero as $T \to \infty$, hence the discrimination function $D_Q(\mathbf{X}_T)$ is consistent.

We may also consider the problem of classification under the assumption that both mean vectors and covariance matrices of the p_j normal populations are unequal. Then the criterion (9.45) can be expressed as

$$\begin{aligned} D_{QL}(\mathbf{X}_T) = &-\frac{1}{2}\log\frac{|\mathbf{\Sigma}_T^{(1)}|}{|\mathbf{\Sigma}_T^{(2)}|} - \frac{1}{2}\mathbf{X}_T'\left(\mathbf{\Sigma}_T^{(1)^{-1}} - \mathbf{\Sigma}_T^{(2)^{-1}}\right)\mathbf{X}_T \\ &+ \left(\boldsymbol{\mu}_T^{(1)'}\mathbf{\Sigma}_T^{(1)^{-1}} - \boldsymbol{\mu}_T^{(2)'}\mathbf{\Sigma}_T^{(2)^{-1}}\right)\mathbf{X}_T \\ &- \frac{1}{2}\boldsymbol{\mu}_T^{(1)'}\mathbf{\Sigma}_T^{(1)^{-1}}\boldsymbol{\mu}_T^{(1)} + \frac{1}{2}\boldsymbol{\mu}_T^{(2)'}\mathbf{\Sigma}_T^{(2)^{-1}}\boldsymbol{\mu}_T^{(2)}, \end{aligned} \tag{9.60}$$

which shows that it depends on a second term which is quadratic with respect to $\mathbf{X}_T$ and a third term which is linear in $\mathbf{X}_T$. The rule is to classify $\mathbf{X}_T$ into Π_1 when $D_{QL}(\mathbf{X}_T) > 0$ and into Π_2 otherwise. The following lemma is useful, which is due to Theorem 3.3.2 of Mathai and Provost (1992).

Lemma 9.1 *When $\mathbf{X}_T \sim N(\boldsymbol{\mu}_T, \mathbf{\Sigma}_T)$, $\mathbf{\Sigma}_T > 0$, the sth cumulant K_s of $\mathbf{Q}_T$ is given as follows:*

For $\mathbf{Q}_T = \mathbf{X}_T'\mathbf{A}_T\mathbf{X}_T + \mathbf{a}_T'\mathbf{X}_T + \mathbf{d}_T$,

$$K_1 = \mathrm{tr}\,(\mathbf{A}_\mathrm{T}\Sigma_\mathrm{T}) + \boldsymbol{\mu}_\mathrm{T}'\mathbf{A}_\mathrm{T}\boldsymbol{\mu}_\mathrm{T} + \mathbf{a}_\mathrm{T}'\boldsymbol{\mu}_\mathrm{T} + \mathbf{d}_\mathrm{t}, \tag{9.61}$$

and for $s \geq 2$

$$\begin{aligned} K_s = 2^{s-1}s!\Bigg\{&\frac{\mathrm{tr}\,(\mathbf{A}_\mathrm{T}\Sigma_\mathrm{T})^\mathrm{s}}{s} + \frac{1}{4}\mathbf{a}_T'\left(\mathbf{\Sigma}_T\mathbf{A}_T\right)^{s-2}\mathbf{\Sigma}_T\mathbf{a}_T \\ &+ \boldsymbol{\mu}_T'\left(\mathbf{A}_T\mathbf{\Sigma}_T\right)^{s-1}\mathbf{A}_T\boldsymbol{\mu}_T + \mathbf{a}_T'\left(\mathbf{\Sigma}_T\mathbf{A}_T\right)^{s-1}\boldsymbol{\mu}_T\Bigg\}. \end{aligned} \tag{9.62}$$

Therefore we can see that

$$\begin{aligned} T^{-1}E_j\{D_{QL}(\mathbf{X}_T)\} = -\frac{1}{2T}\Bigg[&\mathrm{tr}\left(\mathbf{A}_\mathrm{T}\Sigma_\mathrm{T}^{(\mathrm{j})}\right) + \boldsymbol{\mu}_\mathrm{T}^{(\mathrm{j})'}\mathbf{A}_\mathrm{T}\boldsymbol{\mu}_\mathrm{T}^{(\mathrm{j})} + \mathbf{a}_\mathrm{T}'\boldsymbol{\mu}_\mathrm{T}^{(\mathrm{j})} \\ &+ \log\frac{|\mathbf{\Sigma}_T^{(1)}|}{|\mathbf{\Sigma}_T^{(2)}|} + \boldsymbol{\mu}_T^{(1)'}\mathbf{\Sigma}_T^{(1)^{-1}}\boldsymbol{\mu}_T^{(1)} - \boldsymbol{\mu}_T^{(2)'}\mathbf{\Sigma}_T^{(2)^{-1}}\boldsymbol{\mu}_T^{(2)}\Bigg], \end{aligned} \tag{9.63}$$

and for $s \geq 2$, sth cumulant of

$$\frac{-2}{\sqrt{T}} \left[D_{QL}(\mathbf{X}_T) - E_j \left\{ D_{QL}(\mathbf{X}_T) \right\} \right] \tag{9.64}$$

under Π is given by

$$T^{-s/2} 2^{s-1} s! \left\{ \frac{\mathrm{tr}\left(\mathbf{A}_\mathrm{T} \Sigma_\mathrm{T}^{(\mathrm{j})} \right)^\mathrm{s}}{s} + \frac{1}{4} \mathbf{a}_T' \left(\boldsymbol{\Sigma}_T^{(j)} \mathbf{A}_T \right)^{s-2} \boldsymbol{\Sigma}_T^{(j)} \mathbf{a}_T \right.$$
$$\left. + \boldsymbol{\mu}_T^{(j)\prime} \left(\mathbf{A}_T \boldsymbol{\Sigma}_T^{(j)} \right)^{s-1} \mathbf{A}_T \boldsymbol{\mu}_T^{(j)} + \mathbf{a}_T' \left(\boldsymbol{\Sigma}_T^{(j)} \mathbf{A}_T \right)^{s-1} \boldsymbol{\mu}_T^{(j)} \right\} \tag{9.65}$$

with

$$\mathbf{A}_T = \boldsymbol{\Sigma}_T^{(1)^{-1}} - \boldsymbol{\Sigma}_T^{(2)^{-1}} \tag{9.66}$$

and

$$\mathbf{a}_T = -2 \left(\boldsymbol{\mu}_T^{(1)\prime} \boldsymbol{\Sigma}_T^{(1)^{-1}} - \boldsymbol{\mu}_T^{(2)\prime} \boldsymbol{\Sigma}_T^{(2)^{-1}} \right), \tag{9.67}$$

which enables us to examine the performance of two misclassification probability of the rule using $D_{QL}(\mathbf{X}_T)$ by the normal approximation argument as in $D_Q(\mathbf{X}_T)$.

A difficulty in this particular case is that the distribution of $D_{QL}(\mathbf{X}_T)$ is intractable under either Π_1 or Π_2, and the theoretical error rates in this case are difficult to determine. Because of the distributional complexity of such a statistic, one may prefer to use a linear discriminant function of the form $\boldsymbol{b}_T' \mathbf{X}_T$, and then the procedure is as follows: an observed stretch $\mathbf{X}_T$ is classified into Π_1 or Π_2, respectively, according to

$$\boldsymbol{b}_T' \mathbf{X}_T \leq c \quad \text{or} \quad \mathrm{b}_\mathrm{T}' \mathbf{X}_\mathrm{T} > \mathrm{c}, \tag{9.68}$$

where $\boldsymbol{b}_T$ is a $T \times 1$ vector and c is a scalar. Assuming the Gaussianity of the process, $\boldsymbol{b}_T' \mathbf{X}_T$ has a univariate normal distribution with mean $\boldsymbol{b}_T' \boldsymbol{\mu}_T^{(j)}$ and variance $\boldsymbol{b}_T' \boldsymbol{\Sigma}_T^{(j)} \boldsymbol{b}_T$ under Π_j. The misclassification probabilities by this procedure are then given by

$$P(2|1, \{\boldsymbol{b}_T, c\}) \equiv \Pr_1 \left(\boldsymbol{b}_T' \mathbf{X}_T > c \right) = 1 - \Phi \left\{ \frac{c - \boldsymbol{b}_T' \boldsymbol{\mu}_T^{(1)}}{\left(\boldsymbol{b}_T' \boldsymbol{\Sigma}_T^{(1)} \boldsymbol{b}_T \right)^{1/2}} \right\} \tag{9.69}$$

and

$$P(1|2,\{\boldsymbol{b}_T,c\}) \equiv \Pr_2\left(\boldsymbol{b}_T'\mathbf{X}_T \le c\right) = \Phi\left\{\frac{c-\boldsymbol{b}_T'\boldsymbol{\mu}_T^{(2)}}{\left(\boldsymbol{b}_T'\boldsymbol{\Sigma}_T^{(2)}\boldsymbol{b}_T\right)^{1/2}}\right\}$$
$$= 1-\Phi\left\{\frac{\boldsymbol{b}_T'\boldsymbol{\mu}_T^{(2)}-c}{\left(\boldsymbol{b}_T'\boldsymbol{\Sigma}_T^{(2)}\boldsymbol{b}_T\right)^{1/2}}\right\} \qquad (9.70)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution $N(0,1)$. It is desired to make these two probabilities small. The related topics for stationary time series can be found in Taniguchi and Kakizawa (2000), Shumway and Unger (1974) and Shumway (1982).

Exercises

9.1 Show that (9.39) converges to zero, as $\epsilon \to 0$.

9.2 Show that (9.40) diverges, as $\epsilon \to 0$.

Appendix

In Chapter 2 we provided a concise description of the probability measure. More generally we, here, give a brief explanation of the foundation of measure theory and Lebesgue integral. The readers who are interested in details may refer to, e.g., Ash and Doléans-Dade (2000).

If $\mathcal{A}$ is a σ-field of subsets of Ω, $(\Omega, \mathcal{A})$ is called a *measurable space*, and the sets in $\mathcal{A}$ are called *measurable sets*.

Definition A.1 *A* measure *on a σ-field $\mathcal{A}$ is an extended real-valued function μ on $\mathcal{A}$ which satisfies the following:*

(M1) *For every $A \in \mathcal{A}$, $\mu(A) \geq 0$.*

(M2) $\mu(\emptyset) = 0$.

(M3) *Whenever $A_1, A_2, \ldots$ form a finite or countably infinite collection of disjoint sets in $\mathcal{A}$, we have*

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i). \tag{A.71}$$

A *measure space* is a triple $(\Omega, \mathcal{A}, \mu)$ where Ω is a set, $\mathcal{A}$ is a σ-field of subsets of Ω, and μ is a measure on $\mathcal{A}$. If $\mu(\Omega) < \infty$, μ is called a *finite measure*. In particular, if $\mu(\Omega) = 1$, μ becomes a *probability measure*, discussed in Chapter 2. If Ω can be written as $\bigcup_{n=1}^{\infty} A_n$ where the A_n's belong to $\mathcal{A}$ and $\mu(A_n) < \infty$ for all n (the A_n may be assumed disjoint, since $\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} \left(A_1^c \cap \cdots \cap A_{n-1}^c \cap A_n\right)$), then μ is said to be *σ-finite* on $\mathcal{A}$.

We can find that there exists a unique measure μ on a measurable space $(\mathbf{R}, \mathcal{B})$ which satisfies $\mu\{(a, b]\} = b - a$ for any interval $(a, b]$. Henceforth, this μ is called the *Lebesgue measure* on $(\mathbf{R}, \mathcal{B})$ and is denoted by μ_L. Similarly, for a right-continuous and monotone increasing function F on $\mathbf{R}$, we can show that there exists a unique measure μ on $(\mathbf{R}, \mathcal{B})$ which satisfies

$$\mu\{(a, b]\} = F(b) - F(a), \quad (\forall a, b \in \mathbf{R}), \tag{A.72}$$

and we call this μ the *Lebesgue-Stieltjes measure* with respect to F and, henceforth, denote by μ_{LS}.

Write the indicator function of $A \subset \Omega$ as

$$\chi_A(x) = \begin{cases} 1, & (x \in A) \\ 0, & (x \notin A). \end{cases} \tag{A.73}$$

If f is a real-valued function on $(\Omega, \mathcal{A})$, f is said to be a *simple* iff it can be written as a finite sum

$$f(x) = \sum_{i=1}^{k} a_i \chi_{A_i}(x) \tag{A.74}$$

where the A_i are disjoint sets in $\mathcal{A}$ and the $a_i \in \mathbf{R}$. Let f be an arbitrary non-negative $\mathcal{A}$-measurable function on $(\Omega, \mathcal{A})$ (see (2.18) in Chapter 2). If we define f_n as

$$f_n(x) = \begin{cases} \frac{k-1}{2^n} & x \in f^{-1}\left\{\left[\frac{k-1}{2^n}, \frac{k}{2^n}\right)\right\}, \quad (k = 1, 2, \ldots, 2^{2n}), \\ 2^n & x \in f^{-1}\left\{[2^n, \infty)\right\}, \end{cases} \tag{A.75}$$

then f_n becomes a sequence of monotone increasing measurable simple functions and we have

$$f(x) = \lim_{n \to \infty} f_n(x). \tag{A.76}$$

We first define the integral of the simple function f in (A.74) with respect to a measure μ on $(\Omega, \mathcal{A})$ as

$$\int_\Omega f d\mu \equiv \sum_{i=1}^{k} a_i \mu(A_i). \tag{A.77}$$

This does not depend on the representation of $f(x)$, that is, if $f(x)$ has another representation, say, $f(x) = \sum_{i=1}^{l} b_i \chi_{B_i}(x)$, then $\sum_{i=1}^{k} a_i \mu(A_i) = \sum_{i=1}^{l} b_i \mu(B_i)$. Next, if f is a non-negative measurable function, then we can take the sequence of simple functions f_n in (A.75) and define the integral of f_n as (A.74), therefore, we define the integral of f with respect to μ by

$$\int_\Omega f d\mu \equiv \lim_{n \to \infty} \int_\Omega f_n d\mu. \tag{A.78}$$

This integral does not depend on the choice of sequence of simple functions $\{f_n\}$. Finally, for an arbitrary real-valued measurable function f which is not necessarily non-negative, if we take

$$f^+(x) = \max\{f(x), 0\}, \quad f^-(x) = \max\{-f(x), 0\}, \tag{A.79}$$

then we have $f(x) = f^+(x) - f^-(x)$, $f^+(x) \geq 0$ and $f^-(x) \geq 0$. Since for f^+ and f^- we can define the integrals by (A.78), we define the integral of f with respect to μ by

$$\int_\Omega f d\mu \equiv \int_\Omega f^+ d\mu - \int_\Omega f^- d\mu. \tag{A.80}$$

The function f is said to be *integrable* iff $\int_\Omega f^+ d\mu$ and $\int_\Omega f^- d\mu$ are both finite.

The fact that f is integrable is equivalent to that $|f|$ is integrable. If f is a real-valued measurable function on $(\mathbf{R}, \mathcal{B})$ and μ is the Lebesgue measure μ_L on $(\mathbf{R}, \mathcal{B})$, the integral defined by (A.80) is called the *Lebesgue integral* of f and $\int_{\mathbf{R}} f d\mu_L$ is simply denoted by $\int_{\mathbf{R}} f(x)dx$. If μ is the Lebesgue-Stieltjes measure with respect to a right-continuous and monotone increasing function F, then we write $\int_{\mathbf{R}} f d\mu_{LS}$ as $\int_{\mathbf{R}} f(x)dF(x)$ and call this the *Lebesgue-Stieltjes integral* of f with respect to F. In these integrals the fundamental properties hold. For example, if f and g are integrable, then for any $a, b \in \mathbf{R}$, $(af + bg)$ is also integrable and

$$\int_{\mathbf{R}} \{af(x) + bg(x)\}\, dx = a \int_{\mathbf{R}} f(x)dx + b \int_{\mathbf{R}} g(x)dx. \tag{A.81}$$

In relation with the usual Riemann integral, we see that if f is Riemann integrable on $[a, b]$, then f is integrable with respect to the Lebesgue measure on $[a, b]$, and the two integrals are equal.

In a measure space $(\Omega, \mathcal{A}, \mu)$, a proposition $S = S(\omega)$ on Ω is said to hold *almost everywhere* with respect to the measure μ iff

$$\mu\{\omega : S(\omega) \text{ is not true}\} = 0 \tag{A.82}$$

and is written

$$S \quad \mu - a.e. \tag{A.83}$$

or simply S, *a.e.* if μ is understood from the sentence.

We now give fundamental theorems of measure theory in terms of random variables.

Theorem A.1 (Lebesgue's convergence theorem) *For a sequence of random variables $\{X_n, n \in \mathbf{N}\}$, if there exists an integrable random variable Y which satisfies*

$$|X_n| \le Y \ a.e., \quad (n \in \mathbf{N}) \tag{A.84}$$

and $X_n \xrightarrow{P} X$, then it follows that

$$\lim_{n\to\infty} E(X_n) = E(X). \tag{A.85}$$

Theorem A.2 *If a sequence of random variables $\{X_n, n \in \mathbf{N}\}$ satisfies*

$$\sum_{n=1}^{\infty} E(|X_n|) < \infty, \tag{A.86}$$

then $\sum_{n=1}^{\infty} X_n$ almost surely converges to a random variable X and satisfies

$$E\left(\sum_{n=1}^{\infty} X_n\right) = \sum_{n=1}^{\infty}(X_n). \tag{A.87}$$

Let $(\Omega_i, \mathcal{A}_i, \mu_i)$ $i = 1, 2$ be two measure spaces and $(\Omega, \mathcal{A}, \mu)$ be their product measure space $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \equiv \sigma[A_1 \times A_2 : A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2]$ and $\mu = \mu_1 \times \mu_2$. For a function $X = X(\omega_1, \omega_2)$ on Ω, the function on Ω_2 defined by $X_{\omega_1}(\omega_2) = X(\omega_1, \omega_2)$ is called the ω_1-segment of X, and the ω_2-segment of X, X_{ω_2}, is defined similarly.

Theorem A.3 (Fubini's theorem) *Let $(\Omega_1, \mathcal{A}_1, \mu_1)$ and $(\Omega_2, \mathcal{A}_2, \mu_2)$ be σ-finite measure spaces. If $\mathcal{A}_1 \times \mathcal{A}_2$-measurable function X on $\Omega_1 \times \Omega_2$ is non-negative or $\mu_1 \times \mu_2$-integrable, then it follows that*

$$\int_{\Omega_1 \times \Omega_2} X d(\mu_1 \times \mu_2) = \int_{\Omega_1} d\mu_1 \int_{\Omega_2} X_{\omega_1} d\mu_2 = \int_{\Omega_2} d\mu_2 \int_{\Omega_1} X_{\omega_2} d\mu_1. \quad \text{(A.88)}$$

Moreover, in the case that X is $\mu_1 \times \mu_2$-integrable, the ω_i-segment of X, X_{ω_i} $(i = 1, 2)$ is integrable $w_i - a.e.$, respectively.

Let μ and ν be σ-finite measures on $(\Omega, \mathcal{A})$. Then ν is said to be *absolutely continuous* with respect to μ and denoted by $\nu \ll \mu$ iff it follows that if $\mu(A) = 0$, then $\nu(A) = 0$ $(A \in \mathcal{A})$.

Theorem A.4 (Radon-Nikodym theorem) *Let μ and ν be σ-finite measures on $(\Omega, \mathcal{A})$ with $\nu \ll \mu$. Then there exists $\mathcal{A}$-measurable function g such that*

$$\nu(A) = \int_A g d\mu, \quad (\forall A \in \mathcal{A}), \quad \text{(A.89)}$$

where g is unique $\mu - a.e.$ We write this g as $d\nu/d\mu$ and call it the Radon-Nikodym density function *of ν with respect to μ.*

The following theorem is often used for proving the convergence in distribution of multidimensional random variables.

Theorem A.5 (Cramér-Wold device) *Let $\{\mathbf{X}_n\}$ be a sequence of m-dimensional random variables. Then $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ if and only if $\mathbf{a}'\mathbf{X}_n \xrightarrow{d} \mathbf{a}'\mathbf{X}$ for every $\mathbf{a} \in \mathbf{R}^m$.*

In statistical theory, we need the distribution of various statistics. The following theorem is one of the tools for this purpose.

Theorem A.6 (Change of variables) *Let a random vector $\mathbf{X} = (X_1, \ldots, X_n)'$ have the joint probability density function $f_{\mathbf{X}}(\mathbf{x})$, $\mathbf{x} = (x_1, \ldots, x_n)' \in \mathbf{R}^n$. Assume a function ϕ is one to one and random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is defined by $\mathbf{Y} = \phi(\mathbf{X})$. If $\psi = (\psi_1(\mathbf{y}), \ldots, \psi_n(\mathbf{y}))'$, $\mathbf{y} = (y_1, \ldots, y_n)' \in \mathbf{R}^n$ is the inverse transformation of ϕ $(\psi = \phi^{-1})$ and ψ is continuously differentiable, then $\mathbf{Y}$ has the joint probability density function*

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}\{\psi(\mathbf{y})\}|J|, \quad \text{(A.90)}$$

where J is the Jacobian defined by

$$J = \begin{vmatrix} \frac{\partial \psi_1}{\partial y_1} & \cdots & \frac{\partial \psi_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \psi_n}{\partial y_1} & \cdots & \frac{\partial \psi_n}{\partial y_n} \end{vmatrix}. \tag{A.91}$$

In this book we mainly dealt with discrete time stochastic processes from the standpoint of statistical analysis. However, in probabilistic financial engineering, the models are often described in terms of continuous time stochastic processes. The following is the most fundamental example of continuous time stochastic process.

Definition A.2 *A continuous time stochastic process satisfying the following (W1) to (W3) is called the* Wiener process.

(W1) $W_0 = 0$.

(W2) *For any $t_1 < \ldots < t_n$, $(t_1, \ldots, t_n \in [0, \infty))$, the increments*

$$W_{t_2} - W_{t_1}, W_{t_3} - W_{t_2}, \ldots, W_{t_n} - W_{t_{n-1}} \tag{A.92}$$

are mutually independent.

(W3) *For any $t > s$,*

$$W_t - W_s \sim N(0, \sigma^2(t-s)), \quad (\sigma^2 > 0). \tag{A.93}$$

We recommend Tanaka (1996) for descriptions of the stochastic integral of $\{W_t\}$ and the stochastic differential equation based on it from the statistical analytic point of view.

The foundation of spectral analysis for time series is Fourier analysis. The following two theorems are frequently used. Henceforth, we denote the Fourier coefficient of a function $f(\lambda)$ by

$$\hat{f}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\lambda) e^{-in\lambda} d\lambda. \tag{A.94}$$

Theorem A.7 (Parseval's identity) *Let $f(\lambda)$ and $g(\lambda)$ be squared integrable on $[-\pi, \pi]$. Then it follows that*

$$\sum_{-\infty}^{\infty} \hat{f}(n)\overline{\hat{g}(n)} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\lambda)\overline{g(\lambda)} d\lambda. \tag{A.95}$$

Theorem A.8 (Riemann-Lebesgue theorem) *If $f(\lambda)$ is integrable on $[-\pi, \pi]$, then it follows that*

$$\hat{f}(n) \to 0, \quad (|n| \to \infty). \tag{A.96}$$

Related to the convergence in distribution, we state the following two theorems. For proofs, see Billingsley (1995, Chapter 5).

Theorem A.9 *Let $F_0, F_1, F_2, \ldots$ be a sequence of distribution functions on $\mathbf{R}^m$ with corresponding characteristic functions*

$$\phi_n(\mathbf{t}) = \int_{\mathbf{R}^m} \exp(i\mathbf{t}'\mathbf{x}) dF_n(\mathbf{x}), \quad n = 0, 1, 2, \ldots \tag{A.97}$$

Then the following three statements are equivalent:

(i) $F_n \xrightarrow{d} F_0$.

(ii) $\int_{\mathbf{R}^m} g(\mathbf{x}) dF_n(\mathbf{x}) \to \int_{\mathbf{R}^m} g(\mathbf{x}) dF_0(\mathbf{x})$ *for every bounded, continuous function g.*

(iii) $\lim_{n\to\infty} \phi_n(\mathbf{t}) = \phi_0(\mathbf{t})$ *for every* $\mathbf{t} = (t_1, \ldots, t_m)' \in \mathbf{R}^m$.

Theorem A.10 (Helly's theorem) *For every sequence $\{F_n : n \in \mathbf{N}\}$ of distribution functions there exists a sub-sequence $\{F_{n_k}\}$ and a nondecreasing, right continuous function F such that*

$$\lim_{k\to\infty} F_{n_k}(x) = F(x) \tag{A.98}$$

at continuity point x of F.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $T = [0, \infty)$. Suppose that random variables $X_t = X_t(\omega)$ are defined on $(\Omega, \mathcal{F}, P)$ and for all $t \in T$. We say that the stochastic process $X = \{X_t : t \in T\}$ is *measurable* if, for all Borel sets $B \in \mathcal{B}$ of $\mathbf{R}$,

$$\{(\omega, t) : X_t(\omega) \in B\} \in \mathcal{F} \times \mathcal{B}(T), \tag{A.99}$$

where $\mathcal{B}(T)$ is a σ-algebra of Borel sets on T.

Let $\mathcal{F} = \{\mathcal{F}_t : t \in T\}$ be a nondecreasing family of σ-algebras satisfying $\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F}$, $s \le t$. A measurable stochastic process $X = \{X_t : t \in T\}$ is said to be *adapted to a family of σ-algebras* $\mathcal{F} = \{\mathcal{F}_t : t \in T\}$ if for any $t \in T$ the random variables X_t are $\mathcal{F}_t$-measurable. Such a stochastic process is denoted $X = \{X_t, \mathcal{F}_t\}$ and called *$\mathcal{F}$-adapted*. The stochastic process $X = \{X_t, \mathcal{F}_t\}$ is said to be *progressively measurable* if, for any $t \in T$,

$$\{(\omega, s \le t) : X_s(\omega) \in B\} \in \mathcal{F}_t \times \mathcal{B}([0, t]), \tag{A.100}$$

where $B \in \mathcal{B}$ and $\mathcal{B}([0, t])$ is the σ-algebra of Borel sets on $[0, t]$. Evidently any progressively measurable process $X = \{X_t, \mathcal{F}_t\}$ is measurable and adapted to $\mathcal{F} = \{\mathcal{F}_t : t \in T\}$.

Let L_2^τ be the class of progressively measurable processes $\{f_t(\omega), \mathcal{F}_t\}$, $0 \le t \le \tau$, satisfying

$$P\left\{\int_0^\tau f_t(\omega)^2 dt < \infty\right\} = 1. \tag{A.101}$$

For $f = \{f_t, \mathcal{F}_t\} \in L_2^\tau$, we shall define a stochastic integral $I_\tau(f) = \int_0^\tau f_t(\omega) dW_t$, where $\{W_t, \mathcal{F}_t\}$ is a Wiener process. First, we shall determine the stochastic

integral for a certain set of elementary functions. The function $e = e(t,\omega)$, $0 \le t \le \tau$, is said to be *simple* if there exists a subdivision $0 = t_0 < t_1 < \cdots < t_n = \tau$ of $[0,\tau]$ such that $e(t,\omega) = \alpha_i$ if $t \in (t_i, t_{i+1}]$ and α_i is $\mathcal{F}_{t_i}$-measurable. For the simple function $e = e(t,\omega)$ the stochastic integral is defined as

$$I_\tau(e) = \int_0^\tau e(t,\omega)dW_t \equiv \sum_{i=0}^{n-1} \alpha_i \left(W_{t_{i+1}} - W_{t_i}\right). \tag{A.102}$$

For $f = \{f_t, \mathcal{F}_t\} \in L_2^\tau$ it is possible to find a sequence of simple functions $e^{(n)} = e^{(n)}(t,\omega)$, $0 \le t \le \tau$, such that

$$\|f - e^{(n)}\| \xrightarrow{P} 0 \tag{A.103}$$

where $\|\cdot\| = \left\{\int_0^\tau (\cdot)^2 dt\right\}^{1/2}$ (see Liptser and Shiryayev (1977)). We also denote the limit in probability by $P - \lim_{n\to\infty}$, hence, (A.103) is written as $P - \lim_{n\to\infty} \left\|f - e^{(n)}\right\| = 0$.

The stochastic *Itô integral* of $f = \{f_t, \mathcal{F}_t\} \in L_2^\tau$ is defined as the limit

$$I_\tau(f) = \int_0^\tau f_t(\omega)dW_t \equiv P - \lim_{n\to\infty} I_\tau\left(e^{(n)}\right). \tag{A.104}$$

Let M_2^τ be the class of stochastic processes $f = f_t(\omega) \in L_2^\tau$ satisfying the condition

$$E\left\{\int_0^\tau f_t(\omega)^2 dt\right\} < \infty. \tag{A.105}$$

It is known that the stochastic integral $I_\tau(f)$ for $f \in M_2^\tau$ has the following properties:

$$\begin{aligned}
&I_t(af + bg) = aI_t(f) + bI_t(g) \quad (P - a.s.),\ a, b \in \mathbf{R},\\
&E\left\{\int_0^t f_u(\omega)dW_u\right\} = 0,\\
&E\left\{\int_0^t f_u(\omega)dW_u \int_0^s g_u(\omega)dW_u\right\} = E\left\{\int_0^{t\wedge s} f_u(\omega)g_u(\omega)du\right\},
\end{aligned} \tag{A.106}$$

where $f, g \in M_2^\tau$, $0 \le u \le t \le \tau$.

The process $X = \{X_t, \mathcal{F}_t, 0 \le t \le \tau\}$ is called an *Itô process relative to the Wiener process* $W = \{W_t, \mathcal{F}_t, 0 \le t \le \tau\}$ if there exist two adapted processes $A = \{A_t(\omega), \mathcal{F}_t, 0 \le t \le \tau\}$ and $B = \{B_t(\omega), \mathcal{F}_t, 0 \le t \le \tau\}$ such that

$$\begin{aligned}
&P\left\{\int_0^\tau |A_t(\omega)|\, dt < \infty\right\} = 1,\\
&P\left\{\int_0^\tau B_t(\omega)^2 dt < \infty\right\} = 1
\end{aligned} \tag{A.107}$$

and, with probability one for $0 \le t \le \tau$,

$$X_t = X_0 + \int_0^t A_s(\omega)ds + \int_0^t B_s(\omega)dW_s. \tag{A.108}$$

For brevity it is said that the process X_t has the *stochastic differential*

$$dX_t = A_t(\omega)dt + B_t(\omega)dW_t,\ X_0,\ 0 \le t \le \tau, \tag{A.109}$$

which is understood as an abbreviated form of (A.108). The Itô process $X = \{X_t, \mathcal{F}_t, 0 \le t \le \tau\}$ is called a *diffusion-type process* if $A_t(\omega)$ and $B_t(\omega)$ are measurable with respect to $\mathcal{F}_t^X \equiv \sigma\{X_s : 0 \le s \le t\}$.

For a cluster problem of locally stationary processes, the following lemmas are useful, which are due to Dahlhaus (1996a, 1996b). Lemma A.2 is Lemma A.5 of Dahlhaus (1996b) and Lemma A.3 is Theorem 3.2 (ii) of Dahlhaus (1996a).

Set $\boldsymbol{\mu}_T^{(j)} = \left\{\mu^{(j)}\left(\frac{1}{T}\right), \ldots, \mu^{(j)}\left(\frac{T}{T}\right)\right\}'$ and $\boldsymbol{\Sigma}_T^{(j)} = \boldsymbol{\Sigma}_T(A^{(j)}, A^{(j)})$ where

$$\boldsymbol{\Sigma}_T(A, B) = \left\{\int_{-\pi}^{\pi} A_{s,T}^{\circ}(\lambda) B_{t,T}^{\circ}(-\lambda) \exp\left(i\lambda(s-t)\right) d\lambda\right\}_{s,t=1,\ldots,T}.$$

First, we summarize the assumptions used in the following.

Assumption A.2 (i) *Suppose $A : [0,1] \times \mathbf{R} \to \mathbf{C}$ is a 2π-periodic function with $A(u,\lambda) = \overline{A(u,-\lambda)}$ which is differentiable in u and λ with uniformly bounded derivative $(\partial/\partial u)(\partial/\partial\lambda)A$. $f_A(u,\lambda) \equiv |A(u,\lambda)|^2$ denotes the time varying spectral density. $A_{t,T}^{\circ} : \mathbf{R} \to \mathbf{C}$ are 2π-periodic functions with*

$$\sup_{t,\lambda}\left|A_{t,T}^{\circ}(\lambda) - A\left(\frac{t}{T}, \lambda\right)\right| \le KT^{-1}.$$

(ii) *Suppose $\mu : [0,1] \to \mathbf{R}$ is differentiable with uniformly bounded derivative.*

We introduce the following matrices (see Dahlhaus (1996a) for the detailed definition);

$$\mathbf{W}_T(\phi) = \frac{S}{N}\sum_{j=1}^{M} \mathbf{K}_T^{(j)\prime} \mathbf{W}_T^{(j)}(\phi) \mathbf{K}_T^{(j)},$$

where

$$\mathbf{W}_T^{(j)}(\phi) = \left\{\int_{-\pi}^{\pi} \phi(u_j, \lambda) \exp(i\lambda(k-l)) d\lambda\right\}_{k,l=1,\ldots,L_j}$$

and $\mathbf{K}_T^{(j)} = \left(\mathbf{0}_{j1}, \mathbf{I}_{L_j}, \mathbf{0}_{j2}\right)$. According to Lemmas 4.4 and 4.7 of Dahlhaus (1996a), we can see that

$$\|\boldsymbol{\Sigma}_T(A,A)\| \le C + o(1), \quad \left\|\boldsymbol{\Sigma}_T(A,A)^{-1}\right\| \le C + o(1),$$

and $\boldsymbol{W}_T(f_A)$ and $\boldsymbol{W}_T\left(\left\{4\pi^2 f_A\right\}^{-1}\right)$ are the approximations of $\boldsymbol{\Sigma}_T(A,A)$ and $\boldsymbol{\Sigma}_T(A,A)^{-1}$, respectively.

Lemma A.2 *Let $k \in \mathbf{N}$, A_l, B_l fulfill Assumption A.2 (i) and μ_1, μ_2 fulfill Assumption A.2 (ii). Let $\boldsymbol{\Sigma}_l = \boldsymbol{\Sigma}_T(A_l, A_l)$ or $\mathbf{W}_T(f_{A_l})$. Furthermore, let*

$\boldsymbol{\Gamma}_l = \boldsymbol{\Sigma}_T(B_l, B_l)$, $\mathbf{W}_T\left(\left\{4\pi^2\right\}^{-1} f_{B_l}\right)$ *or* $\boldsymbol{\Gamma}_l^{-1} = \mathbf{W}_T\left(\left\{4\pi^2 f_{B_l}\right\}^{-1}\right)$. *Then we have*

$$T^{-1}\mathrm{tr}\left\{\prod_{l=1}^{k}\boldsymbol{\Gamma}_l^{-1}\boldsymbol{\Sigma}_l\right\} = \frac{1}{2\pi}\int_0^1\int_{-\pi}^{\pi}\left\{\prod_{l=1}^{k}\frac{f_{A_l}(u,\lambda)}{f_{B_l}(u,\lambda)}\right\}d\lambda du + O\left(T^{-1/2}\log^{2k+2}T\right)$$

and

$$T^{-1}\boldsymbol{\mu}_{1,T}'\left\{\prod_{l=1}^{k-1}\boldsymbol{\Gamma}_l^{-1}\boldsymbol{\Sigma}_l\right\}\boldsymbol{\Gamma}_k^{-1}\boldsymbol{\mu}_{2,T} = \frac{1}{2\pi}\int_0^1\left\{\prod_{l=1}^{k-1}\frac{f_{A_l}(u,0)}{f_{B_l}(u,0)}\right\}f_{B_k}(u,0)^{-1}\mu_1(u)\mu_2(u)du + O\left(T^{-1/2}\log^{2k+2}T\right).$$

Lemma A.3 *Let D° be the transfer function of a locally stationary process $\{Z_{t,T}\}$, where the corresponding D is bounded from below and has uniformly bounded derivative $\frac{\partial}{\partial u}\frac{\partial}{\partial\lambda}D$. $f_D(u,\lambda) \equiv |D(u,\lambda)|^2$ denotes the time varying spectral density of $Z_{t,T}$. Then, for $\boldsymbol{\Sigma}_T(d) \equiv \boldsymbol{\Sigma}_T(D,D)$, we have*

$$\lim_{T\to\infty} T^{-1}\log|\boldsymbol{\Sigma}_T(d)| = \frac{1}{2\pi}\int_0^1\int_{-\pi}^{\pi}\log 2\pi f_D(u,\lambda)d\lambda du.$$

References

Akahira, M. and Takeuchi, K. 1981. *Asymptotic efficiency of statistical estimators: concepts and higher order asymptotic efficiency.* New York: Springer-Verlag.

Akaike, H. 1970. Statistical predictor identification. *Ann. Inst. Statist. Math.* 22:203–217.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory,* ed. B. N. Petrov and F. Csadki, 267–281. Budapest: Akademiai Kiado.

Akaike, H. 1977. On entropy maximization principle. In *Applications of statistics,* ed. P. R. Krishnaiah, 27–41. Amsterdam: North-Holland.

Akaike, H. 1978. A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.* 30:9–14.

Anderson, T. W. 1971. *The statistical analysis of time series.* New York: John Wiley & Sons.

Anderson, T. W. 1984. *An introduction to multivariate statistical analysis.* 2nd ed. New York: John Wiley & Sons

Ash, R. B. 1972. *Real analysis and probability.* New York: Academic Press.

Ash, R. B. and Doléans-Dade, C. A. 2000. *Probability and measure theory.* 2nd ed. Burlington: Harcourt/Academic Press.

Bai, J. 1994. Weak convergence of the sequential empirical processes of residuals in ARMA models. *Ann. Statist.* 22:2051–2061.

Basak, G., Jagannathan, R. and Sun, G. 2002. A direct test for the mean variance efficiency of a portfolio. *J. Econom. Dynam. Control* 26:1195–1215.

Bellman, R. 1960. *Introduction to matrix analysis.* New York: McGraw-Hill.

Benghabrit, Y. and Hallin, M. 1992. Optimal rank-based tests against first-order superdiagonal bilinear dependence. *J. Statist. Plann. Inference* 32:45–61.

Benghabrit, Y. and Hallin, M. 1996. Locally asymptotically optimal tests for autoregressive against bilinear serial dependence. *Statist. Sinica* 6:147–169.

Bhattacharya, R. N. and Ranga Rao, R. 1976. *Normal approximation and asymptotic expansions.* New York: John Wiley & Sons.

Bickel, P. J. 1982. On adaptive estimation. *Ann. Statist.* 10:647–671.

Bickel, P. J. and Doksum, K. A. 2001. *Mathematical statistics: basic ideas and selected topics.* Vol. 1. 2nd ed. Englewood Cliffs NJ: Prentice Hall.

Billingsley, P. 1995. *Probability and measure.* 3rd ed. New York: John Wiley & Sons.

Black, F. and Scholes, M. 1973. The pricing of options and corporate liabilities. *J. Political Econom.* 81:637–654.

Boldin, M. V. 1982. Estimation of the distribution of noise in an autoregressive scheme. *Theory Probab. Appl.* 27:866–871.

Boldin, M. V. 1998. On residual empirical distribution functions in ARCH models with applications to testing and estimation. *Mitt. Math. Sem. Giessen* 235:49–66.

Boldin, M. V. 2000. On empirical processes in heteroscedastic time series and their use for hypothesis testing and estimation. *Math. Methods Statist.* 9:65–89.

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* 31:307–327.

Brillinger, D. R. 1981. *Time series: data analysis and theory.* Expanded ed. San Francisco: Holden-Day.

Brillinger, D. R. and Rosenblatt, M. 1967a. Asymptotic theory of estimates of k-th order spectra. In *Spectral Analysis of Time Series,* ed. B. Harris, 153–188. New York: John Wiley.

Brillinger, D. R. and Rosenblatt, M. 1967b. Computation and interpretation of k-th order spectra. In *Spectral Analysis of Time Series,* ed. B. Harris, 189–232. New York: John Wiley.

Brockwell, P. J. and Davis, R. A. 1991. *Time series: theory and methods.* 2nd ed. New York: Springer-Verlag.

Brown, B. M. 1971. Martingale central limit theorems. *Ann. Math. Statist.* 42:59–66.

Carmona, R. A. 2004. *Statistical analysis of financial data in S-Plus.* New York: Springer-Verlag.

Chen, M. and An, H. Z. 1998. A note on the stationarity and the existence of moments of the GARCH model. *Statist. Sinica* 8:505–510.

Choi, I. B. and Taniguchi, M. 2001. Misspecified prediction for time series. *J. Forecast.* 20:543–564.

Choi, I. B. and Taniguchi, M. 2003. Prediction problems for square-transformed stationary processes. *Stat. Inference Stoch. Process.* 6:43–64.

Chung, K. L. 2001. *A course in probability theory.* 3rd ed. San Diego: Academic Press.

Corrado, C. J. and Su, T. 1996a. S&P 500 index option tests of Jarrow and Rudd's approximate option valuation formula. *Journal of Futures Markets* 16:611–629.

Corrado, C. J. and Su, T. 1996b. Skewness and kurtosis in S&P 500 index returns implied by option prices. *Journal of Financial Research* 19:175–192.

Corrado, C. J. and Su, T. 1997. Implied volatility skews and stock return skewness and kurtosis implied by stock option prices. *European Journal of Finance* 3:73–85.

Cox, J. C. and Ross, S. A. 1976. The valuation of options for alternative stochastic processes. *Journal of Financial Economics* 3:145–166.

Cox, J. C., Ingersoll, J. E. and Ross, S. A. 1985. A theory of the term structure of interest rates. *Econometrica* 53:385–407.

Dahlhaus, R. 1989. Efficient parameter estimation for self-similar processes. *Ann. Statist.* 17:1749–1766.

Dahlhaus, R. 1996a. On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Process. Appl.* 62:139–168.

Dahlhaus, R. 1996b. Maximum likelihood estimation and model selection for locally stationary processes. *J. Nonparametr. Statist.* 6:171–191.

Dahlhaus, R. 1996c. Asymptotic statistical inference for nonstationary processes with evolutionary spectra. In *Athens conference on applied probability and time series analysis, Vol. 2,* ed. P. M. Robinson and M. Rose, 145–159. New York: Springer.

Dahlhaus, R. 1997. Fitting time series models to nonstationary processes. *Ann. Statist.* 25:1–37.

Dahlhaus, R. 2000. A likelihood approximation for locally stationary processes. *Ann. Statist.* 28:1762–1794.

Drost, F. C., Klaassen, C. A. J. and Werker, B. J. M. 1997. Adaptive estimation in time-series models. *Ann. Statist.* 25:786–817.

Duan, J. C. 1997. Augmented GARCH(p, q) process and its diffusion limit. *J. Econometrics* 79:97–127.

Dunsmuir, W. 1979. A central limit theorem for parameter estimation in stationary vector time series and its application to models for a signal observed with noise. *Ann. Statist.* 7:490–506.

Dzhaparidze, K. 1986. *Parameter estimation and hypothesis testing in spectral analysis of stationary time series.* New York: Springer-Verlag.

Engle, R. F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50:987–1007.

Fan, J. and Kreutzberger, E. 1998. Automatic local smoothing for spectral density estimation. *Scand. J. Statist.* 25:359–369.

Fan, J. and Yao, Q. 2003. *Nonlinear time series: nonparametric and parametric methods.* New York: Springer-Verlag.

Fan, J. and Zhang, W. 2004. Generalised likelihood ratio tests for spectral density. *Biometrika* 91:195–209.

Fuller, W. A. 1996. *Introduction to statistical time series.* 2nd ed. New York: John Wiley & Sons.

Garel, B. and Hallin, M. 1995. Local asymptotic normality of multivariate ARMA processes with a linear trend. *Ann. Inst. Statist. Math.* 47:551–579.

Gill, R. D. 1989. Non- and semi-parametric maximum likelihood estimators and the von Mises method. I. *Scand. J. Statist.* 16:97–128.

Giraitis, L. and Robinson, P. M. 2001. Whittle estimation of ARCH models. *Econometric Theory* 17:608–631.

Giraitis, L. and Surgailis, D. 1990. A central limit theorem for quadratic forms in strongly dependent linear variables and its application to asymptotical normality of Whittle's estimate. *Probab. Theory Related Fields* 86:87–104.

Giraitis, L., Kokoszka, P. and Leipus, R. 2000. Stationary ARCH models: dependence structure and central limit theorem. *Econometric Theory* 16:3–22.

Gouriéroux, C. 1997. *ARCH models and financial applications.* New York: Springer-Verlag.

Granger, C. W. J. and Andersen, A. 1978. Non-linear time series modelling. In *Applied time series analysis*, ed. D. F. Findley, 25–38. New York: Academic Press.

Grenander, U. and Rosenblatt, M. 1957. *Statistical analysis of stationary time series.* New York: John Wiley & Sons.

Guttman, I. 1970. *Statistical tolerance regions, classical and bayesian.* Darien, Conn.: Hafner Publishing Co.

Hafner, C. M. 1998. *Nonlinear time series analysis with applications to foreign exchange rate volatility.* Heidelberg: Physica-Verlag.

Haggan, V. and Ozaki, T. 1981. Modelling nonlinear random vibrations using an amplitude-dependent autoregressive time series model. *Biometrika* 68:189–196.

Hall, P. and Heyde, C. C. 1980. *Martingale limit theory and its application.* New York: Academic Press.

Hallin, M. and Puri, M. L. 1994. Aligned rank tests for linear models with autocorrelated error terms. *J. Multivariate Anal.* 50:175–237.

Hallin, M., Ingenbleek, J. F. and Puri, M. L. 1985. Linear serial rank tests for randomness against ARMA alternatives. *Ann. Statist.* 13:1156–1181.

Hallin, M., Taniguchi, M., Serroukh, A. and Choy, K. 1999. Local asymptotic normal-

ity for regression models with long-memory disturbance. *Ann. Statist.* 27:2054–2080.

Hannan, E. J. 1963. Regression for time series. In *Proc. Sympos. Time Series Analysis,* ed. M. Rosenblatt, 17–37. New York: Wiley.

Hannan, E. J. 1970. *Multiple time series.* New York: John Wiley & Sons.

Hannan, E. J. 1980. The estimation of the order of an ARMA process. *Ann. Statist.* 8:1071–1081.

Hannan, E. J. and Quinn, B. G. 1979. The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* 41:190–195.

Hannan, E. J. and Robinson, P. M. 1973. Lagged regression with unknown lags. *J. Roy. Statist. Soc. Ser. B* 35:252–267.

Härdle, W., Tsybakov, A. and Yang, L. 1998. Nonparametric vector autoregression. *J. Statist. Plann. Inference* 68:221–245.

Hirukawa, J. and Taniguchi, M. 2006. LAN theorem for non-Gaussian locally stationary processes and its applications. *J. Statist. Plann. Inference* 136:640–688.

Hjort, N. L. and Jones, M. C. 1996. Locally parametric nonparametric density estimation. *Ann. Statist.* 24:1619–1647.

Horváth, L., Kokoszka, P. and Teyssière, G. 2001. Empirical process of the squared residuals of an ARCH sequence. *Ann. Statist.* 29:445–469.

Hosoya, Y. 1997. A limit theory for long-range dependence and statistical inference on related models. *Ann. Statist.* 25:105–137.

Hosoya, Y. and Taniguchi, M. 1982. A central limit theorem for stationary processes and the parameter estimation of linear processes. *Ann. Statist.* 10:132–153.

Houglet, M. X. 1980. *Estimating the term structure of interest rates for non-homogeneous bonds.* Dissertation, Graduate School of Business. University of California, Berkeley.

Hull, J. and White, A. 1990. Pricing interest-rate-derivative securities. *Review of Financial Studies* 3:573–592.

Hurst, H. E. 1951. Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civil Engineers* 116:770–799.

Huschens, S. 1998. Confidence intervals for the Value-at-Risk. In *Risk measurement, econometrics and neural networks,* ed. G. Bol, G. Nakhaeizadeh and K. H. Vollmer, 233–244. Heidelberg: Physica-Verlag.

Ibragimov, I. A. 1963. A central limit theorem for a class of dependent random variables. *Theory Probab. Appl.* 8:83–89.

Ibragimov, I. A. and Linnik, Yu. V. 1971. *Independent and stationary sequences of random variables.* Groningen: Wolters-Noordhoff Publishing.

Jarrow, R. and Rudd, A. 1982. Approximate option valuation for arbitrary stochastic processes. *Journal of Financial Economics* 10:347–369.

Jeganathan, P. 1995. Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* 11:818–887.

Jobson, J. D. and Korkie, B. 1980. Estimation for Markowitz efficient portfolios. *J. Amer. Statist. Assoc.* 75:544–554.

Jobson, J. D. and Korkie, B. 1989. A performance interpretation of multivariate tests of asset set intersection, spanning, and mean-variance efficiency. *The Journal of Financial and Quantitative Analysis* 24:185–204.

J. P. Morgan. 1996. Riskmetrics technical document (3rd ed.).

Jurczenko, E., Maillet, B. and Negrea, B. 2002. Multi-moment approximate option

pricing models: a general comparison (Part 1). University of Paris I Panthéon-Sorbonne.

Kakizawa, Y. 1999. Note on the asymptotic efficiency of sample covariances in Gaussian vector stationary processes. *J. Time Ser. Anal.* 20:551–558.

Kakizawa, Y., Shumway, R. H. and Taniguchi, M. 1998. Discrimination and clustering for multivariate time series. *J. Amer. Statist. Assoc.* 93:328–340.

Kariya, T. 1993. *Quantitative methods for portfolio analysis.* Dordrecht: Kluwer Academic Publishers.

Kariya, T. and Liu, R. Y. 2003. *Asset pricing: discrete time approach.* Boston: Kluwer Academic Publishers.

Kato, H., Taniguchi, M. and Honda, M. 2006. Statistical analysis for multiplicatively modulated nonlinear autoregressive model and its applications to electrophysiological signal analysis in humans. *IEEE Trans. Signal Process.* 54:3414–3425.

Kholevo, A. S. 1969. On estimates of regression coefficients. *Theory Probab. Appl.* 14:79–104.

Kreiss, J. P. 1987. On adaptive estimation in stationary ARMA processes. *Ann. Statist.* 15:112–133.

Kreiss, J. P. 1990. Local asymptotic normality for autoregression with infinite order. *J. Statist. Plann. Inference* 26:185–219.

Lauprete, G. J., Samarov, A. M. and Welsch, R. E. 2002. Robust portfolio optimization. *Metrika* 55:139–149.

Lee, S. and Taniguchi, M. 2005. Asymptotic theory for ARCH-SM models: LAN and residual empirical processes. *Statist. Sinica* 15:215–234.

Lehmann, E. L. 1975. *Nonparametrics: statistical methods based on ranks.* San Francisco: Holden-Day.

Lehmann, E. L. 1986. *Testing statistical hypotheses.* 2nd ed. New York: John Wiley & Sons.

Linton, O. 1993. Adaptive estimation in ARCH models. *Econometric Theory* 9:539–569.

Linton, O. 1995. Second order approximation in the partially linear regression model. *Econometrica* 63:1079–1112.

Linton, O. and Xiao, Z. 2001. Second-order approximation for adaptive regression estimators. *Econometric Theory* 17:984–1024.

Liptser, R. S. and Shiryayev, A. N. 1977. *Statistics of random processes I, general theory.* New York: Springer-Verlag.

Longstaff, F. A. 1995. Option pricing and the martingale restriction. *Review of Financial Studies* 8:1091–1124.

Lu, Z. and Jiang, Z. 2001. L_1 geometric ergodicity of a multivariate nonlinear AR model with an ARCH term. *Statist. Probab. Lett.* 51:121–130.

Magnus, J. R. and Neudecker, H. 1988. *Matrix differential calculus with applications in statistics and econometrics.* Chichester: John Wiley & Sons.

Mathai, A. M. and Provost, S. B. 1992. *Quadratic forms in random variables: theory and applications.* New York: Marcel Dekker.

McCulloch, J. H. 1971. Measuring the term structure of interest rates. *Journal of Business* 44:19–31.

McCulloch, J. H. 1975. The tax-adjusted yield curve. *Journal of Finance* 30:811–830.

Naito, T., Asai, K. and Taniguchi, M. 2006. Local Whittle likelihood estimators and tests for non-Gaussian linear processes. *Waseda University Time Series Discussion Paper* No. 24.

Nelson, C. R. and Siegel, A. F. 1987. Parsimonious modeling of yield curves. *Journal of Business* 60:473–489.

Nelson, D. B. 1990. ARCH models as diffusion approximations. *J. Econometrics* 45:7–38.

Nelson, D. B. 1991. Conditional heteroskedasticity in asset returns: a new approach. *Econometrica* 59:347–370.

Parzen, E. 1992. Time series, statistics, and information. In *New directions in time series analysis, Part 1,* ed. D. R. Brillinger, 265–286. New York: Springer-Verlag.

Phillips, P. C. B. 1989. Partially identified econometric models. *Econometric Theory* 5:181–240.

Resnick, S. I. 1987. *Extreme values, regular variation, and point processes.* New York: Springer-Verlag.

Robinson, P. M. 1991. Automatic frequency domain inference on semiparametric and nonparametric models. *Econometrica* 59:1329–1363.

Rothenberg, T. J. 1984. Approximate normality of generalized least squares estimates. *Econometrica* 52:811–825.

Roussas, G. G. 1972. *Contiguity of probability measures: some applications in statistics.* London: Cambridge University Press.

Roussas, G. G. 1979. Asymptotic distribution of the log-likelihood function for stochastic processes. *Z. Wahrsch. Verw. Gebiete* 47:31–46.

Rubinstein, M. 1998. Edgeworth binomial trees. *Journal of Derivatives* 5:20–27.

Sakiyama, K. 2002. Some statistical applications for locally stationary processes. *Sci. Math. Jpn.* 56:231–250.

Sakiyama, K. and Taniguchi, M. 2003. Testing composite hypotheses for locally stationary processes. *J. Time Ser. Anal.* 24:483–504.

Sakiyama, K. and Taniguchi, M. 2004. Discriminant analysis for locally stationary processes. *J. Multivariate Anal.* 90:282–300.

Schwarz, G. 1978. Estimating the dimension of a model. *Ann. Statist.* 6:461–464.

Shibata, R. 1976. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* 63:117–126.

Shibata, R. 1980. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* 8:147–164.

Shiraishi, H. and Taniguchi, M. 2005. Statistical estimation of optimal portfolios for non-Gaussian dependent returns of assets. *Waseda University Time Series Discussion Paper* No. 15.

Shumway, R. H. 1982. Discriminant analysis for time series. In *Handbook of statistics*, Vol. 2, ed. P. R. Krishnaiah and L. N. Kanal, 1–46. Amsterdam: North-Holland.

Shumway, R. H. 2003. Time-frequency clustering and discriminant analysis. *Statist. Probab. Lett.* 63:307–314.

Shumway, R. H. and Unger, A. N. 1974. Linear discriminant functions for stationary time series. *J. Amer. Statist. Assoc.* 69:948–956.

Stout, W. F. 1974. *Almost sure convergence.* New York: Academic Press.

Strasser, H. 1985. *Mathematical theory of statistics.* New York: Walter de Gruyter.

Subba Rao, T. 1981. On the theory of bilinear time series models. *J. Roy. Statist. Soc. B* 43:244–255.

Svensson, L. E. O. 1994. Estimating and interpreting forward interest rates: Sweden 1992-1994, *NBER Working Paper* No. 4871.

Swensen, A. R. 1985. The asymptotic distribution of the likelihood ratio for autoregressive time series with a regression trend. *J. Multivariate Anal.* 16:54–70.

Takeuchi, K. 1976. Distribution of information statistics and criteria for adequacy of models. *Suri-Kagaku (Math. Sci.)* 153:12–18 (in Japanese).

Tamaki, K. 2007. Second order optimality for estimators in time series regression models. *J. Multivariate Anal.* 98:638–659.

Tanaka, K. 1996. *Time series analysis: nonstationary and noninvertible distribution theory.* New York: John Wiley & Sons.

Taniai, H. and Taniguchi, M. 2007. Statistical estimation errors of VaR under ARCH returns. To appear in *J. Statist. Plann. Inference (Ogawa Volume).*

Taniguchi, M. 1980. On estimation of the integrals of certain functions of spectral density. *J. Appl. Probab.* 17:73–83.

Taniguchi, M. 1987. Minimum contrast estimation for spectral densities of stationary processes. *J. Roy. Statist. Soc. Ser. B* 49:315–325.

Taniguchi, M. 1994. Higher order asymptotic theory for discriminant analysis in exponential families of distributions. *J. Multivariate Anal.* 48:169–187.

Taniguchi, M. and Kakizawa, Y. 2000. *Asymptotic theory of statistical inference for time series.* New York: Springer-Verlag.

Taniguchi, M., Puri, M. L. and Kondo, M. 1996. Nonparametric approach for non-Gaussian vector stationary processes. *J. Multivariate Anal.* 56:259–283.

Taniguchi, M., van Garderen, K. J. and Puri, M. L. 2003. Higher order asymptotic theory for minimum contrast estimators of spectral parameters of stationary processes. *Econometric Theory* 19:984–1007.

Tjøstheim, D. 1986. Estimation in nonlinear time series models. *Stochastic Process. Appl.* 21:251–273.

Tong, H. 1990. *Nonlinear time series: a dynamical system approach.* Oxford: Oxford University Press.

Toyooka, Y. 1985. Second-order risk comparison of SLSE with GLSE and MLE in a regression with serial correlation. *J. Multivariate Anal.* 17:107–126.

Toyooka, Y. 1986. Second-order risk structure of GLSE and MLE in a regression with a linear process. *Ann. Statist.* 14:1214–1225.

Tsay, R. S. 2002. *Analysis of financial time series.* New York: Wiley.

van der Vaart, A. W. 1998. *Asymptotic statistics.* Cambridge: Cambridge University Press.

Vasicek, O. A. 1977. An equilibrium characterization of the term structure. *J. Financial Economics* 5:177–188.

Vasicek, O. A. and Fong, H. G. 1982. Term structure modeling using exponential splines. *Journal of Finance* 37:339–348.

Velasco, C. and Robinson, P. M. 2001. Edgeworth expansions for spectral density estimates and Studentized sample mean. *Econometric Theory* 17:497–539.

White, H. 2000. *Asymptotic theory for econometricians.* Revised ed. San Diego: Academic Press.

Wiener, N. 1933. *The Fourier integral and certain of its application.* Cambridge: Cambridge University Press.

Xiao, Z. and Phillips, P. C. B. 1998. Higher-order approximations for frequency domain time series regression. *J. Econometrics* 86:297–336.

Xiao, Z. and Phillips, P. C. B. 2002. Higher order approximations for Wald statistics in time series regressions with integrated processes. *J. Econometrics* 108:157–198.

Zhang, G. and Taniguchi, M. 1994. Discriminant analysis for stationary vector time series. *J. Time Ser. Anal.* 15:117–126.

Zivot, E. and Wang, J. 2006. *Modeling financial time series with S-Plus.* New York:

Springer-Verlag.
Zygmund, A. 1959. *Trigonometric series.* London: Cambridge University Press.