# Analysis of Financial Time Series

# Analysis of Financial Time Series

## Financial Econometrics

RUEY S. TSAY
University of Chicago

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.

This book is printed on acid-free paper. ♾



Published simultaneously in Canada.



For ordering and customer service, call 1-800-CALL-WILEY.

***Library of Congress Cataloging-in-Publication Data***

Tsay, Ruey S., 1951–
Analysis of financial time series / Ruey S. Tsay.
p. cm. — (Wiley series in probability and statistics. Financial engineering section)
"A Wiley-Interscience publication."
Includes bibliographical references and index.
ISBN 0-471-41544-8 (cloth : alk. paper)
1. Time-series analysis. 2. Econometrics. 3. Risk management. I. Title. II. Series.
HA30.3 T76 2001
332′.01′5195—dc21 2001026944

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To my parents and Teresa

# Contents

# Preface

This book grew out of an MBA course in analysis of financial time series that I have been teaching at the University of Chicago since 1999. It also covers materials of Ph.D. courses in time series analysis that I taught over the years. It is an introductory book intended to provide a comprehensive and systematic account of financial econometric models and their application to modeling and prediction of financial time series data. The goals are to learn basic characteristics of financial data, understand the application of financial econometric models, and gain experience in analyzing financial time series.

The book will be useful as a text of time series analysis for MBA students with finance concentration or senior undergraduate and graduate students in business, economics, mathematics, and statistics who are interested in financial econometrics. The book is also a useful reference for researchers and practitioners in business, finance, and insurance facing Value at Risk calculation, volatility modeling, and analysis of serially correlated data.

The distinctive features of this book include the combination of recent developments in financial econometrics in the econometric and statistical literature. The developments discussed include the timely topics of Value at Risk (VaR), high-frequency data analysis, and Markov Chain Monte Carlo (MCMC) methods. In particular, the book covers some recent results that are yet to appear in academic journals; see Chapter 6 on derivative pricing using jump diffusion with closed-form formulas, Chapter 7 on Value at Risk calculation using extreme value theory based on a nonhomogeneous two-dimensional Poisson process, and Chapter 9 on multivariate volatility models with time-varying correlations. MCMC methods are introduced because they are powerful and widely applicable in financial econometrics. These methods will be used extensively in the future.

Another distinctive feature of this book is the emphasis on real examples and data analysis. Real financial data are used throughout the book to demonstrate applications of the models and methods discussed. The analysis is carried out by using several computer packages; the SCA (the Scientific Computing Associates) for building linear time series models, the RATS (Regression Analysis for Time Series) for estimating volatility models, and the S-Plus for implementing neural networks and obtaining postscript plots. Some commands required to run these packages are given

in appendixes of appropriate chapters. In particular, complicated RATS programs used to estimate multivariate volatility models are shown in Appendix A of Chapter 9. Some fortran programs written by myself and others are used to price simple options, estimate extreme value models, calculate VaR, and to carry out Bayesian analysis. Some data sets and programs are accessible from the World Wide Web at http://www.gsb.uchicago.edu/fac/ruey.tsay/teaching/fts.

The book begins with some basic characteristics of financial time series data in Chapter 1. The other chapters are divided into three parts. The first part, consisting of Chapters 2 to 7, focuses on analysis and application of univariate financial time series. The second part of the book covers Chapters 8 and 9 and is concerned with the return series of multiple assets. The final part of the book is Chapter 10, which introduces Bayesian inference in finance via MCMC methods.

A knowledge of basic statistical concepts is needed to fully understand the book. Throughout the chapters, I have provided a brief review of the necessary statistical concepts when they first appear. Even so, a prerequisite in statistics or business statistics that includes probability distributions and linear regression analysis is highly recommended. A knowledge in finance will be helpful in understanding the applications discussed throughout the book. However, readers with advanced background in econometrics and statistics can find interesting and challenging topics in many areas of the book.

An MBA course may consist of Chapters 2 and 3 as a core component, followed by some nonlinear methods (e.g., the neural network of Chapter 4 and the applications discussed in Chapters 5-7 and 10). Readers who are interested in Bayesian inference may start with the first five sections of Chapter 10.

Research in financial time series evolves rapidly and new results continue to appear regularly. Although I have attempted to provide broad coverage, there are many subjects that I do not cover or can only mention in passing.

I sincerely thank my teacher and dear friend, George C. Tiao, for his guidance, encouragement and deep conviction regarding statistical applications over the years. I am grateful to Steve Quigley, Heather Haselkorn, Leslie Galen, Danielle LaCourciere, and Amy Hendrickson for making the publication of this book possible, to Richard Smith for sending me the estimation program of extreme value theory, to Bonnie K. Ray for helpful comments on several chapters, to Steve Kou for sending me his preprint on jump diffusion models, to Robert E. McCulloch for many years of collaboration on MCMC methods, to many students of my courses in analysis of financial time series for their feedback and inputs, and to Jeffrey Russell and Michael Zhang for insightful discussions concerning analysis of high-frequency financial data. To all these wonderful people I owe a deep sense of gratitude. I am also grateful to the support of the Graduate School of Business, University of Chicago and the National Science Foundation. Finally, my heart goes to my wife, Teresa, for her continuous support, encouragement, and understanding, to Julie, Richard, and Vicki for bringing me joys and inspirations; and to my parents for their love and care.

R. S. T.
*Chicago, Illinois*

*Analysis of Financial Time Series.* Ruey S. Tsay

ISBN: 0-471-41544-8

CHAPTER 1

# Financial Time Series and Their Characteristics

Financial time series analysis is concerned with theory and practice of asset valuation over time. It is a highly empirical discipline, but like other scientific fields theory forms the foundation for making inference. There is, however, a key feature that distinguishes financial time series analysis from other time series analysis. Both financial theory and its empirical time series contain an element of uncertainty. For example, there are various definitions of asset volatility, and for a stock return series, the volatility is not directly observable. As a result of the added uncertainty, statistical theory and methods play an important role in financial time series analysis.

The objective of this book is to provide some knowledge of financial time series, introduce some statistical tools useful for analyzing these series, and gain experience in financial applications of various econometric methods. We begin with the basic concepts of asset returns and a brief introduction to the processes to be discussed throughout the book. Chapter 2 reviews basic concepts of linear time series analysis such as stationarity and autocorrelation function, introduces simple linear models for handling serial dependence of the series, and discusses regression models with time series errors, seasonality, unit-root nonstationarity, and long memory processes. Chapter 3 focuses on modeling conditional heteroscedasticity (i.e., the conditional variance of an asset return). It discusses various econometric models developed recently to describe the evolution of volatility of an asset return over time. In Chapter 4, we address nonlinearity in financial time series, introduce test statistics that can discriminate nonlinear series from linear ones, and discuss several nonlinear models. The chapter also introduces nonparametric estimation methods and neural networks and shows various applications of nonlinear models in finance. Chapter 5 is concerned with analysis of high-frequency financial data and its application to market microstructure. It shows that nonsynchronous trading and bid-ask bounce can introduce serial correlations in a stock return. It also studies the dynamic of time duration between trades and some econometric models for analyzing transactions data. In Chapter 6, we introduce continuous-time diffusion models and Ito's lemma. Black-Scholes option pricing formulas are derived and a simple jump diffusion model is used to capture some characteristics commonly observed in options markets. Chapter 7 discusses extreme value theory, heavy-tailed distributions, and their application

to financial risk management. In particular, it discusses various methods for calculating Value at Risk of a financial position. Chapter 8 focuses on multivariate time series analysis and simple multivariate models. It studies the lead-lag relationship between time series and discusses ways to simplify the dynamic structure of a multivariate series and methods to reduce the dimension. Co-integration and threshold co-integration are introduced and used to investigate arbitrage opportunity in financial markets. In Chapter 9, we introduce multivariate volatility models, including those with time-varying correlations, and discuss methods that can be used to reparameterize a conditional covariance matrix to satisfy the positiveness constraint and reduce the complexity in volatility modeling. Finally, in Chapter 10, we introduce some newly developed Monte Carlo Markov Chain (MCMC) methods in the statistical literature and apply the methods to various financial research problems, such as the estimation of stochastic volatility and Markov switching models.

The book places great emphasis on application and empirical data analysis. Every chapter contains real examples, and, in many occasions, empirical characteristics of financial time series are used to motivate the development of econometric models. Computer programs and commands used in data analysis are provided when needed. In some cases, the programs are given in an appendix. Many real data sets are also used in the exercises of each chapter.

## 1.1 ASSET RETURNS

Most financial studies involve returns, instead of prices, of assets. Campbell, Lo, and MacKinlay (1997) give two main reasons for using returns. First, for average investors, return of an asset is a complete and scale-free summary of the investment opportunity. Second, return series are easier to handle than price series because the former have more attractive statistical properties. There are, however, several definitions of an asset return.

Let $P_t$ be the price of an asset at time index $t$. We discuss some definitions of returns that are used throughout the book. Assume for the moment that the asset pays no dividends.

***One-Period Simple Return***

Holding the asset for one period from date $t-1$ to date $t$ would result in a *simple gross return*

$$1 + R_t = \frac{P_t}{P_{t-1}} \quad \text{or} \quad P_t = P_{t-1}(1 + R_t) \tag{1.1}$$

The corresponding one-period *simple net return* or *simple return* is

$$R_t = \frac{P_t}{P_{t-1}} - 1 = \frac{P_t - P_{t-1}}{P_{t-1}}. \tag{1.2}$$

***Multiperiod Simple Return***

Holding the asset for $k$ periods between dates $t - k$ and $t$ gives a $k$-period simple gross return

$$\begin{aligned} 1 + R_t[k] = \frac{P_t}{P_{t-k}} &= \frac{P_t}{P_{t-1}} \times \frac{P_{t-1}}{P_{t-2}} \times \cdots \times \frac{P_{t-k+1}}{P_{t-k}} \\ &= (1 + R_t)(1 + R_{t-1}) \cdots (1 + R_{t-k+1}) \\ &= \prod_{j=0}^{k-1} (1 + R_{t-j}). \end{aligned}$$

Thus, the $k$-period simple gross return is just the product of the $k$ one-period simple gross returns involved. This is called a compound return. The $k$-period simple net return is $R_t[k] = (P_t - P_{t-k})/P_{t-k}$.

In practice, the actual time interval is important in discussing and comparing returns (e.g., monthly return or annual return). If the time interval is not given, then it is implicitly assumed to be one year. If the asset was held for $k$ years, then the annualized (average) return is defined as

$$\text{Annualized } \{R_t[k]\} = \left[ \prod_{j=0}^{k-1} (1 + R_{t-j}) \right]^{1/k} - 1.$$

This is a geometric mean of the $k$ one-period simple gross returns involved and can be computed by

$$\text{Annualized } \{R_t[k]\} = \exp\left[ \frac{1}{k} \sum_{j=0}^{k-1} \ln(1 + R_{t-j}) \right] - 1,$$

where $\exp(x)$ denotes the exponential function and $\ln(x)$ is the natural logarithm of the positive number $x$. Because it is easier to compute arithmetic average than geometric mean and the one-period returns tend to be small, one can use a first-order Taylor expansion to approximate the annualized return and obtain

$$\text{Annualized } \{R_t[k]\} \approx \frac{1}{k} \sum_{j=0}^{k-1} R_{t-j}. \tag{1.3}$$

Accuracy of the approximation in Eq. (1.3) may not be sufficient in some applications, however.

***Continuous Compounding***

Before introducing continuously compounded return, we discuss the effect of compounding. Assume that the interest rate of a bank deposit is 10% per annum and the initial deposit is \$1.00. If the bank pays interest once a year, then the net value

**Table 1.1. Illustration of the Effects of Compounding: The Time Interval Is 1 Year and the Interest Rate is 10% per Annum.**

| Type | Number of payments | Interest rate per period | Net Value |
|---|---|---|---|
| Annual | 1 | 0.1 | \$1.10000 |
| Semiannual | 2 | 0.05 | \$1.10250 |
| Quarterly | 4 | 0.025 | \$1.10381 |
| Monthly | 12 | 0.0083 | \$1.10471 |
| Weekly | 52 | $\frac{0.1}{52}$ | \$1.10506 |
| Daily | 365 | $\frac{0.1}{365}$ | \$1.10516 |
| Continuously | $\infty$ | | \$1.10517 |

of the deposit becomes \$1(1+0.1) = \$1.1 one year later. If the bank pays interest semi-annually, the 6-month interest rate is 10%/2 = 5% and the net value is $\$1(1 + 0.1/2)^2 = \$1.1025$ after the first year. In general, if the bank pays interest $m$ times a year, then the interest rate for each payment is $10\%/m$ and the net value of the deposit becomes $\$1(1 + 0.1/m)^m$ one year later. Table 1.1 gives the results for some commonly used time intervals on a deposit of \$1.00 with interest rate 10% per annum. In particular, the net value approaches \$1.1052, which is obtained by exp(0.1) and referred to as the result of continuous compounding. The effect of compounding is clearly seen.

In general, the net asset value $A$ of continuous compounding is

$$A = C \exp(r \times n), \tag{1.4}$$

where $r$ is the interest rate per annum, $C$ is the initial capital, and $n$ is the number of years. From Eq. (1.4), we have

$$C = A \exp(-r \times n), \tag{1.5}$$

which is referred to as the *present value* of an asset that is worth $A$ dollars $n$ years from now, assuming that the continuously compounded interest rate is $r$ per annum.

***Continuously Compounded Return***

The natural logarithm of the simple gross return of an asset is called the continuously compounded return or *log return*:

$$r_t = \ln(1 + R_t) = \ln \frac{P_t}{P_{t-1}} = p_t - p_{t-1}, \tag{1.6}$$

where $p_t = \ln(P_t)$. Continuously compounded returns $r_t$ enjoy some advantages over the simple net returns $R_t$. First, consider multiperiod returns. We have

$$\begin{aligned} r_t[k] &= \ln(1 + R_t[k]) = \ln[(1 + R_t)(1 + R_{t-1}) \cdots (1 + R_{t-k+1})] \\ &= \ln(1 + R_t) + \ln(1 + R_{t-1}) + \cdots + \ln(1 + R_{t-k+1}) \\ &= r_t + r_{t-1} + \cdots + r_{t-k+1}. \end{aligned}$$

Thus, the continuously compounded multiperiod return is simply the sum of continuously compounded one-period returns involved. Second, statistical properties of log returns are more tractable.

***Portfolio Return***

The simple net return of a portfolio consisting of $N$ assets is a weighted average of the simple net returns of the assets involved, where the weight on each asset is the percentage of the portfolio's value invested in that asset. Let $p$ be a portfolio that places weight $w_i$ on asset $i$, then the simple return of $p$ at time $t$ is $R_{p,t} = \sum_{i=1}^{N} w_i R_{it}$, where $R_{it}$ is the simple return of asset $i$.

The continuously compounded returns of a portfolio, however, do not have the above convenient property. If the simple returns $R_{it}$ are all small in magnitude, then we have $r_{p,t} \approx \sum_{i=1}^{N} w_i r_{it}$, where $r_{p,t}$ is the continuously compounded return of the portfolio at time $t$. This approximation is often used to study portfolio returns.

***Dividend Payment***

If an asset pays dividends periodically, we must modify the definitions of asset returns. Let $D_t$ be the dividend payment of an asset between dates $t-1$ and $t$ and $P_t$ be the price of the asset at the end of period $t$. Thus, dividend is not included in $P_t$. Then the simple net return and continuously compounded return at time $t$ become

$$R_t = \frac{P_t + D_t}{P_{t-1}} - 1, \quad r_t = \ln(P_t + D_t) - \ln(P_{t-1}).$$

***Excess Return***

Excess return of an asset at time $t$ is the difference between the asset's return and the return on some reference asset. The reference asset is often taken to be riskless, such as a short-term U.S. Treasury bill return. The simple excess return and log excess return of an asset are then defined as

$$Z_t = R_t - R_{0t}, \quad z_t = r_t - r_{0t}, \tag{1.7}$$

where $R_{0t}$ and $r_{0t}$ are the simple and log returns of the reference asset, respectively. In the finance literature, the excess return is thought of as the payoff on an arbitrage portfolio that goes long in an asset and short in the reference asset with no net initial investment.

***Remark:*** A long financial position means owning the asset. A short position involves selling asset one does not own. This is accomplished by borrowing the asset from an investor who has purchased. At some subsequent date, the short seller is obligated to buy exactly the same number of shares borrowed to pay back the lender.

Because the repayment requires equal shares rather than equal dollars, the short seller benefits from a decline in the price of the asset. If cash dividends are paid on the asset while a short position is maintained, these are paid to the buyer of the short sale. The short seller must also compensate the lender by matching the cash dividends from his own resources. In other words, the short seller is also obligated to pay cash dividends on the borrowed asset to the lender; see Cox and Rubinstein (1985).

***Summary of Relationship***

The relationships between simple return $R_t$ and continuously compounded (or log) return $r_t$ are

$$r_t = \ln(1 + R_t), \quad R_t = e^{r_t} - 1.$$

Temporal aggregation of the returns produces

$$1 + R_t[k] = (1 + R_t)(1 + R_{t-1}) \cdots (1 + R_{t-k+1}),$$
$$r_t[k] = r_t + r_{t-1} + \cdots + r_{t-k+1}.$$

If the continuously compounded interest rate is $r$ per annum, then the relationship between present and future values of an asset is

$$A = C \exp(r \times n), \quad C = A \exp(-r \times n).$$

## 1.2 DISTRIBUTIONAL PROPERTIES OF RETURNS

To study asset returns, it is best to begin with their distributional properties. The objective here is to understand the behavior of the returns across assets and over time. Consider a collection of $N$ assets held for $T$ time periods, say $t = 1, \ldots, T$. For each asset $i$, let $r_{it}$ be its log return at time $t$. The log returns under study are $\{r_{it}; i = 1, \ldots, N; t = 1, \ldots, T\}$. One can also consider the simple returns $\{R_{it}; i = 1, \ldots, N; t = 1, \ldots, T\}$ and the log excess returns $\{z_{it}; i = 1, \ldots, N; t = 1, \ldots, T\}$.

### 1.2.1 Review of Statistical Distributions and Their Moments

We briefly review some basic properties of statistical distributions and the moment equations of a random variable. Let $R^k$ be the $k$-dimensional Euclidean space. A point in $R^k$ is denoted by $\boldsymbol{x} \in R^k$. Consider two random vectors $\boldsymbol{X} = (X_1, \ldots, X_k)'$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_q)'$. Let $P(\boldsymbol{X} \in A, \boldsymbol{Y} \in B)$ be the probability that $\boldsymbol{X}$ is in the subspace $A \subset R^k$ and $\boldsymbol{Y}$ is in the subspace $B \subset R^q$. For most of the cases considered in this book, both random vectors are assumed to be continuous.

***Joint Distribution***
The function

$$F_{X,Y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = P(\boldsymbol{X} \leq \boldsymbol{x}, \boldsymbol{Y} \leq \boldsymbol{y}),$$

where $\boldsymbol{x} \in R^p$, $\boldsymbol{y} \in R^q$, and the inequality "$\leq$" is a component-by-component operation, is a joint distribution function of $\boldsymbol{X}$ and $\boldsymbol{Y}$ with parameter $\boldsymbol{\theta}$. Behavior of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is characterized by $F_{X,Y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$. If the joint probability density function $f_{x,y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ of $\boldsymbol{X}$ and $\boldsymbol{Y}$ exists, then

$$F_{X,Y}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = \int_{-\infty}^{\boldsymbol{x}} \int_{-\infty}^{\boldsymbol{y}} f_{x,y}(\boldsymbol{w}, \boldsymbol{z}; \boldsymbol{\theta}) d\boldsymbol{z} d\boldsymbol{w}.$$

In this case, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are continuous random vectors.

***Marginal Distribution***
The marginal distribution of $\boldsymbol{X}$ is given by

$$F_X(\boldsymbol{x}; \boldsymbol{\theta}) = F_{X,Y}(\boldsymbol{x}, \infty, \ldots, \infty; \boldsymbol{\theta}).$$

Thus, the marginal distribution of $\boldsymbol{X}$ is obtained by integrating out $\boldsymbol{Y}$. A similar definition applies to the marginal distribution of $\boldsymbol{Y}$.

If $k = 1$, $X$ is a scalar random variable and the distribution function becomes

$$F_X(x) = P(X \leq x; \boldsymbol{\theta}),$$

which is known as the cumulative distribution function (CDF) of $X$. The CDF of a random variable is nondecreasing [i.e., $F_X(x_1) \leq F_X(x_2)$ if $x_1 \leq x_2$, and satisfies $F_X(-\infty) = 0$ and $F_X(\infty) = 1$]. For a given probability $p$, the smallest real number $x_p$ such that $p \leq F_X(x_p)$ is called the $p$th quantile of the random variable $X$. More specifically,

$$x_p = \inf_x \{x \mid p \leq F_X(x)\}.$$

We use CDF to compute the $p$ value of a test statistic in the book.

***Conditional Distribution***
The conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} \leq \boldsymbol{y}$ is given by

$$F_{X|Y \leq y}(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{P(\boldsymbol{X} \leq \boldsymbol{x}, \boldsymbol{Y} \leq \boldsymbol{y})}{P(\boldsymbol{Y} \leq \boldsymbol{y})}.$$

If the probability density functions involved exist, then the conditional density of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$ is

$$f_{x|y}(\boldsymbol{x};\boldsymbol{\theta}) = \frac{f_{x,y}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})}{f_y(\boldsymbol{y};\boldsymbol{\theta})}, \tag{1.8}$$

where the marginal density function $f_y(\boldsymbol{y};\boldsymbol{\theta})$ is obtained by

$$f_y(\boldsymbol{y};\boldsymbol{\theta}) = \int_{-\infty}^{\infty} f_{x,y}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{x}.$$

From Eq. (1.8), the relation among joint, marginal, and conditional distributions is

$$f_{x,y}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}) = f_{x|y}(\boldsymbol{x};\boldsymbol{\theta}) \times f_y(\boldsymbol{y};\boldsymbol{\theta}). \tag{1.9}$$

This identity is used extensively in time series analysis (e.g., in maximum likelihood estimation). Finally, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent random vectors if and only if $f_{x|y}(\boldsymbol{x};\boldsymbol{\theta}) = f_x(\boldsymbol{x};\boldsymbol{\theta})$. In this case, $f_{x,y}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta}) = f_x(\boldsymbol{x};\boldsymbol{\theta})f_y(\boldsymbol{y};\boldsymbol{\theta})$.

***Moments of a Random Variable***

The $\ell$-th moment of a continuous random variable $X$ is defined as

$$m'_\ell = E(X^\ell) = \int_{-\infty}^{\infty} x^\ell f(x)\,dx,$$

where "$E$" stands for expectation and $f(x)$ is the probability density function of $X$. The first moment is called the *mean* or *expectation* of $X$. It measures the central location of the distribution. We denote the mean of $X$ by $\mu_x$. The $\ell$-th central moment of $X$ is defined as

$$m_\ell = E[(X-\mu_x)^\ell] = \int_{-\infty}^{\infty} (x-\mu_x)^\ell f(x)\,dx$$

provided that the integral exists. The second central moment, denoted by $\sigma_x^2$, measures the variability of $X$ and is called the *variance* of $X$. The positive square root, $\sigma_x$, of variance is the *standard deviation* of $X$. The first two moments of a random variable uniquely determine a normal distribution. For other distributions, higher order moments are also of interest.

The third central moment measures the symmetry of $X$ with respect to its mean, whereas the 4th central moment measures the tail behavior of $X$. In statistics, *skewness* and *kurtosis*, which are normalized 3rd and 4th central moments of $X$, are often used to summarize the extent of asymmetry and tail thickness. Specifically, the skewness and kurtosis of $X$ are defined as

$$S(x) = E\left[\frac{(X-\mu_x)^3}{\sigma_x^3}\right], \quad K(x) = E\left[\frac{(X-\mu_x)^4}{\sigma_x^4}\right].$$

The quantity $K(x) - 3$ is called the *excess kurtosis* because $K(x) = 3$ for a normal distribution. Thus, the excess kurtosis of a normal random variable is zero. A distribution with positive excess kurtosis is said to have heavy tails, implying that the distribution puts more mass on the tails of its support than a normal distribution does. In practice, this means that a random sample from such a distribution tends to contain more extreme values.

In application, skewness and kurtosis can be estimated by their sample counterparts. Let $\{x_1, \ldots, x_T\}$ be a random sample of $X$ with $T$ observations. The sample mean is

$$\hat{\mu}_x = \frac{1}{T}\sum_{t=1}^{T} x_t, \tag{1.10}$$

the sample variance is

$$\hat{\sigma}_x^2 = \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \hat{\mu}_x)^2, \tag{1.11}$$

the sample skewness is

$$\hat{S}(x) = \frac{1}{(T-1)\hat{\sigma}_x^3}\sum_{t=1}^{T}(x_t - \hat{\mu}_x)^3, \tag{1.12}$$

and the sample kurtosis is

$$\hat{K}(x) = \frac{1}{(T-1)\hat{\sigma}_x^4}\sum_{t=1}^{T}(x_t - \hat{\mu}_x)^4. \tag{1.13}$$

Under normality assumption, $\hat{S}(x)$ and $\hat{K}(x)$ are distributed asymptotically as normal with zero mean and variances $6/T$ and $24/T$, respectively; see Snedecor and Cochran (1980, p. 78).

### 1.2.2 Distributions of Returns

The most general model for the log returns $\{r_{it}; i = 1, \ldots, N; t = 1, \ldots, T\}$ is its joint distribution function:

$$F_r(r_{11}, \ldots, r_{N1}; r_{12}, \ldots, r_{N2}; \ldots; r_{1T}, \ldots, r_{NT}; \boldsymbol{Y}; \boldsymbol{\theta}), \tag{1.14}$$

where $\boldsymbol{Y}$ is a state vector consisting of variables that summarize the environment in which asset returns are determined and $\boldsymbol{\theta}$ is a vector of parameters that uniquely determine the distribution function $F_r(.)$. The probability distribution $F_r(.)$ governs the stochastic behavior of the returns $r_{it}$ and $\boldsymbol{Y}$. In many financial studies, the state

vector $\boldsymbol{Y}$ is treated as given and the main concern is the conditional distribution of $\{r_{it}\}$ given $\boldsymbol{Y}$. Empirical analysis of asset returns is then to estimate the unknown parameter $\boldsymbol{\theta}$ and to draw statistical inference about behavior of $\{r_{it}\}$ given some past log returns.

The model in Eq. (1.14) is too general to be of practical value. However, it provides a general framework with respect to which an econometric model for asset returns $r_{it}$ can be put in a proper perspective.

Some financial theories such as the Capital Asset Pricing Model (CAPM) of Sharpe (1964) focus on the joint distribution of $N$ returns at a single time index $t$ (i.e., the distribution of $\{r_{1t}, \ldots, r_{Nt}\}$). Other theories emphasize the dynamic structure of individual asset returns (i.e., the distribution of $\{r_{i1}, \ldots, r_{iT}\}$ for a given asset $i$). In this book, we focus on both. In the univariate analysis of Chapters 2 to 7, our main concern is the joint distribution of $\{r_{it}\}_{t=1}^{T}$ for asset $i$. To this end, it is useful to partition the joint distribution as

$$\begin{aligned} F(r_{i1}, \ldots, r_{iT}; \boldsymbol{\theta}) &= F(r_{i1}) F(r_{i2} \mid r_{1t}) \cdots F(r_{iT} \mid r_{i,T-1}, \ldots, r_{i1}) \\ &= F(r_{i1}) \prod_{t=2}^{T} F(r_{it} \mid r_{i,t-1}, \ldots, r_{i1}). \end{aligned} \tag{1.15}$$

This partition highlights the temporal dependencies of the log return $r_{it}$. The main issue then is the specification of the conditional distribution $F(r_{it} \mid r_{i,t-1}, \cdot)$—in particular, how the conditional distribution evolves over time. In finance, different distributional specifications lead to different theories. For instance, one version of the random-walk hypothesis is that the conditional distribution $F(r_{it} \mid r_{i,t-1}, \ldots, r_{i1})$ is equal to the marginal distribution $F(r_{it})$. In this case, returns are temporally independent and, hence, not predictable.

It is customary to treat asset returns as continuous random variables, especially for index returns or stock returns calculated at a low frequency, and use their probability density functions. In this case, using the identity in Eq. (1.9), we can write the partition in Eq. (1.15) as

$$f(r_{i1}, \ldots, r_{iT}; \boldsymbol{\theta}) = f(r_{i1}; \boldsymbol{\theta}) \prod_{t=2}^{T} f(r_{it} \mid r_{i,t-1}, \ldots, r_{i1}, \boldsymbol{\theta}). \tag{1.16}$$

For high-frequency asset returns, discreteness becomes an issue. For example, stock prices change in multiples of a tick size in the New York Stock Exchange (NYSE). The tick size was one eighth of a dollar before July 1997 and was one sixteenth of a dollar from July 1997 to January 2001. Therefore, the tick-by-tick return of an individual stock listed on NYSE is not continuous. We discuss high-frequency stock price changes and time durations between price changes later in Chapter 5.

***Remark:*** On August 28, 2000, the NYSE began a pilot program with seven stocks priced in decimals and the American Stock Exchange (AMEX) began a pilot

program with six stocks and two options classes. The NYSE added 57 stocks and 94 stocks to the program on September 25 and December 4, 2000, respectively. All NYSE and AMEX stocks started trading in decimals on January 29, 2001.

Equation (1.16) suggests that conditional distributions are more relevant than marginal distributions in studying asset returns. However, the marginal distributions may still be of some interest. In particular, it is easier to estimate marginal distributions than conditional distributions using past returns. In addition, in some cases, asset returns have weak empirical serial correlations, and, hence, their marginal distributions are close to their conditional distributions.

Several statistical distributions have been proposed in the literature for the marginal distributions of asset returns, including normal distribution, lognormal distribution, stable distribution, and scale-mixture of normal distributions. We briefly discuss these distributions.

***Normal Distribution***

A traditional assumption made in financial study is that the simple returns $\{R_{it} \mid t = 1, \ldots, T\}$ are independently and identically distributed as normal with fixed mean and variance. This assumption makes statistical properties of asset returns tractable. But it encounters several difficulties. First, the lower bound of a simple return is $-1$. Yet normal distribution may assume any value in the real line and, hence, has no lower bound. Second, if $R_{it}$ is normally distributed, then the multiperiod simple return $R_{it}[k]$ is not normally distributed because it is a product of one-period returns. Third, the normality assumption is not supported by many empirical asset returns, which tend to have a positive excess kurtosis.

***Lognormal Distribution***

Another commonly used assumption is that the log returns $r_t$ of an asset is independent and identically distributed (iid) as normal with mean $\mu$ and variance $\sigma^2$. The simple returns are then iid lognormal random variables with mean and variance given by

$$E(R_t) = \exp\left(\mu + \frac{\sigma^2}{2}\right) - 1, \quad \text{Var}(R_t) = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]. \quad (1.17)$$

These two equations are useful in studying asset returns (e.g., in forecasting using models built for log returns). Alternatively, let $m_1$ and $m_2$ be the mean and variance of the simple return $R_t$, which is lognormally distributed. Then the mean and variance of the corresponding log return $r_t$ are

$$E(r_t) = \ln\left[\frac{m_1 + 1}{\sqrt{1 + \frac{m_2}{(1+m_1)^2}}}\right], \quad \text{Var}(r_t) = \ln\left[1 + \frac{m_2}{(1+m_1)^2}\right].$$

Because the sum of a finite number of iid normal random variables is normal, $r_t[k]$ is also normally distributed under the normal assumption for $\{r_t\}$. In addition,

there is no lower bound for $r_t$, and the lower bound for $R_t$ is satisfied using $1 + R_t = \exp(r_t)$. However, the lognormal assumption is not consistent with all the properties of historical stock returns. In particular, many stock returns exhibit a positive excess kurtosis.

***Stable Distribution***

The stable distributions are a natural generalization of normal in that they are stable under addition, which meets the need of continuously compounded returns $r_t$. Furthermore, stable distributions are capable of capturing excess kurtosis shown by historical stock returns. However, non-normal stable distributions do not have a finite variance, which is in conflict with most finance theories. In addition, statistical modeling using non-normal stable distributions is difficult. An example of non-normal stable distributions is the Cauchy distribution, which is symmetric with respect to its median, but has infinite variance.

***Scale Mixture of Normal Distributions***

Recent studies of stock returns tend to use scale mixture or finite mixture of normal distributions. Under the assumption of scale mixture of normal distributions, the log return $r_t$ is normally distributed with mean $\mu$ and variance $\sigma^2$ [i.e., $r_t \sim N(\mu, \sigma^2)$]. However, $\sigma^2$ is a random variable that follows a positive distribution (e.g., $\sigma^{-2}$ follows a Gamma distribution). An example of finite mixture of normal distributions is

$$r_t \sim (1 - X)N(\mu, \sigma_1^2) + XN(\mu, \sigma_2^2),$$

where $0 \leq \alpha \leq 1$, $\sigma_1^2$ is small and $\sigma_2^2$ is relatively large. For instance, with $\alpha = 0.05$, the finite mixture says that 95% of the returns follow $N(\mu, \sigma_1^2)$ and 5% follow $N(\mu, \sigma_2^2)$. The large value of $\sigma_2^2$ enables the mixture to put more mass at the tails of its distribution. The low percentage of returns that are from $N(\mu, \sigma_2^2)$ says that the majority of the returns follow a simple normal distribution. Advantages of mixtures of normal include that they maintain the tractability of normal, have finite higher order moments, and can capture the excess kurtosis. Yet it is hard to estimate the mixture parameters (e.g., the $\alpha$ in the finite-mixture case).

Figure 1.1 shows the probability density functions of a finite mixture of normal, Cauchy, and standard normal random variable. The finite mixture of normal is $0.95N(0, 1) + 0.05N(0, 16)$ and the density function of Cauchy is

$$f(x) = \frac{1}{\pi(1 + x^2)}, \qquad -\infty < x < \infty.$$

It is seen that Cauchy distribution has fatter tails than the finite mixture of normal, which in turn has fatter tails than the standard normal.

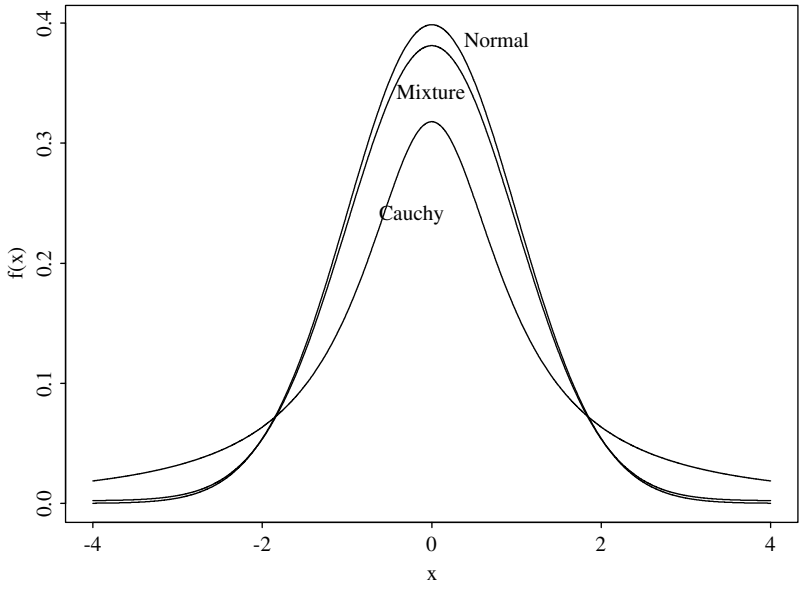


**Figure 1.1.** Comparison of finite-mixture, stable, and standard normal density functions.

### 1.2.3 Multivariate Returns

Let $\boldsymbol{r}_t = (r_{1t}, \ldots, r_{Nt})'$ be the log returns of $N$ assets at time $t$. The multivariate analyses of Chapters 8 and 9 are concerned with the joint distribution of $\{\boldsymbol{r}_t\}_{t=1}^{T}$. This joint distribution can be partitioned in the same way as that of Eq. (1.15). The analysis is then focused on the specification of the conditional distribution function $F(\boldsymbol{r}_t \mid \boldsymbol{r}_{t-1}, \ldots, \boldsymbol{r}_1, \boldsymbol{\theta})$. In particular, how the conditional expectation and conditional covariance matrix of $\boldsymbol{r}_t$ evolve over time constitute the main subjects of Chapters 8 and 9.

The mean vector and covariance matrix of a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ are defined as

$$E(\boldsymbol{X}) = \boldsymbol{\mu}_x = [E(X_1), \ldots, E(X_p)]',$$
$$\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}_x = E[(\boldsymbol{X} - \boldsymbol{\mu}_x)(\boldsymbol{X} - \boldsymbol{\mu}_x)']$$

provided that the expectations involved exist. When the data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ of $\boldsymbol{X}$ are available, the sample mean and covariance matrix are defined as

$$\widehat{\boldsymbol{\mu}}_x = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t, \quad \widehat{\boldsymbol{\Sigma}}_x = \frac{1}{T}\sum_{t=1}^{T} (\boldsymbol{x}_t - \widehat{\boldsymbol{\mu}}_x)(\boldsymbol{x}_t - \widehat{\boldsymbol{\mu}}_x)'.$$

These sample statistics are consistent estimates of their theoretical counterparts provided that the covariance matrix of $\boldsymbol{X}$ exists. In the finance literature, multivariate normal distribution is often used for the log return $\boldsymbol{r}_t$.

### 1.2.4 Likelihood Function of Returns

The partition of Eq. (1.15) can be used to obtain the likelihood function of the log returns $\{r_1, \ldots, r_T\}$ of an asset, where for ease in notation the subscript $i$ is omitted from the log return. If the conditional distribution $f(r_t \mid r_{t-1}, \ldots, r_1, \boldsymbol{\theta})$ is normal with mean $\mu_t$ and variance $\sigma_t^2$, then $\boldsymbol{\theta}$ consists of the parameters in $\mu_t$ and $\sigma_t^2$ and the likelihood function of the data is

$$f(r_1, \ldots, r_T; \boldsymbol{\theta}) = f(r_1; \boldsymbol{\theta}) \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left[\frac{-(r_t - \mu_t)^2}{2\sigma_t^2}\right], \tag{1.18}$$

where $f(r_1; \boldsymbol{\theta})$ is the marginal density function of the first observation $r_1$. The value of $\boldsymbol{\theta}$ that maximizes this likelihood function is the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$. Since log function is monotone, the MLE can be obtained by maximizing the log likelihood function,

$$\ln f(r_1, \ldots, r_T; \boldsymbol{\theta}) = \ln f(r_1; \boldsymbol{\theta}) - \frac{1}{2} \sum_{t=2}^{T} \left[\ln(2\pi) + \ln(\sigma_t^2) + \frac{(r_t - \mu_t)^2}{\sigma_t^2}\right],$$

which is easier to handle in practice. Log likelihood function of the data can be obtained in a similar manner if the conditional distribution $f(r_t \mid r_{t-1}, \ldots, r_1; \boldsymbol{\theta})$ is not normal.

### 1.2.5 Empirical Properties of Returns

The data used in this section are obtained from the Center for Research in Security Prices (CRSP) of the University of Chicago. Dividend payments, if any, are included in the returns. Figure 1.2 shows the time plots of monthly simple returns and log returns of International Business Machines (IBM) stock from January 1926 to December 1997. A *time plot* shows the data against the time index. The upper plot is for the simple returns. Figure 1.3 shows the same plots for the monthly returns of value-weighted market index. As expected, the plots show that the basic patterns of simple and log returns are similar.

Table 1.2 provides some descriptive statistics of simple and log returns for selected U.S. market indexes and individual stocks. The returns are for daily and monthly sample intervals and are in percentages. The data spans and sample sizes are also given in the table. From the table, we make the following observations. (a) Daily returns of the market indexes and individual stocks tend to have high excess kurtoses. For monthly series, the returns of market indexes have higher excess kurtoses than individual stocks. (b) The mean of a daily return series is close to zero, whereas that

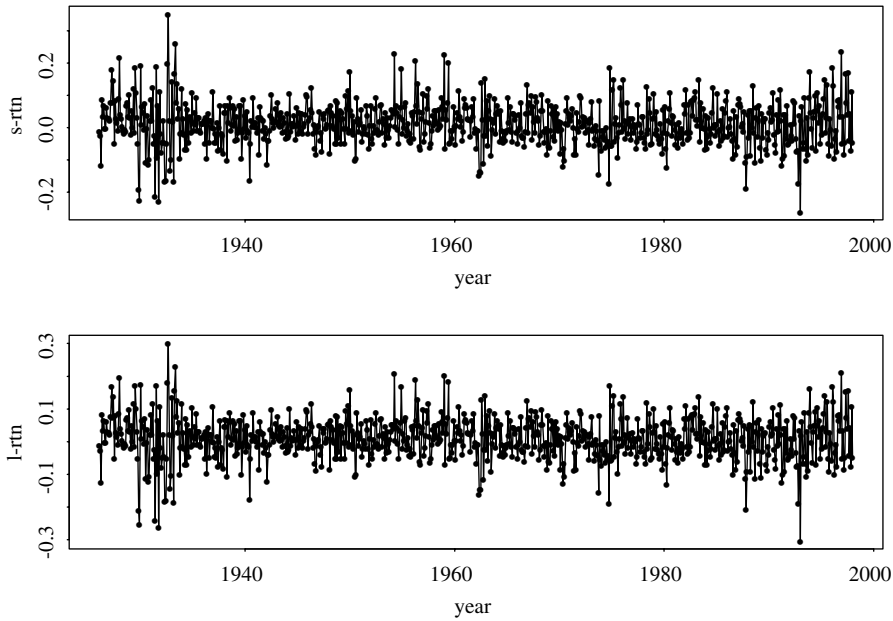


**Figure 1.2.** Time plots of monthly returns of IBM stock from January 1926 to December 1997. The upper panel is for simple net returns, and the lower panel is for log returns.

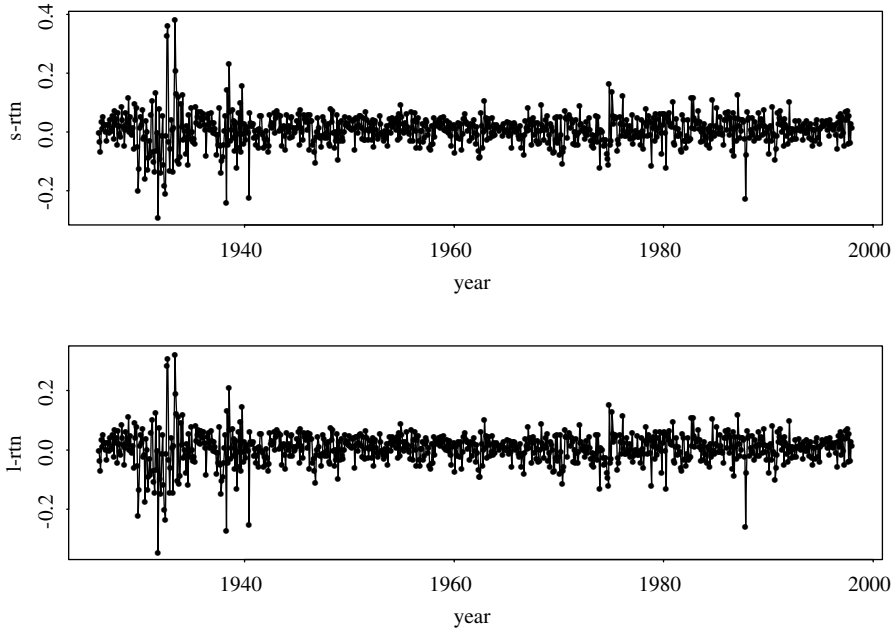


**Figure 1.3.** Time plots of monthly returns of the value-weighted index from January 1926 to December 1997. The upper panel is for simple net returns, and the lower panel is for log returns.

**Table 1.2. Descriptive Statistics for Daily and Monthly Simple and Log Returns of Selected Indexes and Stocks. Returns Are in Percentages, and the Sample Period Ends on December 31, 1997. The Statistics Are Defined in Equations (1.10) to (1.13), and VW and EW Denote Value-Weighted and Equal-Weighted Indexes.**

| Security | Start | Size | Mean | Stan. Dev. | Skew. | Excess Kurt. | Min. | Max. |
|---|---|---|---|---|---|---|---|---|
| | | | | (a) Daily simple returns (%) | | | | |
| VW | 62/7/3 | 8938 | 0.049 | 0.798 | −1.23 | 30.06 | −17.18 | 8.67 |
| EW | 62/7/3 | 8938 | 0.083 | 0.674 | −1.09 | 18.09 | −10.48 | 6.95 |
| I.B.M. | 62/7/3 | 8938 | 0.050 | 1.479 | 0.01 | 11.34 | −22.96 | 12.94 |
| Intel | 72/12/15 | 6329 | 0.138 | 2.880 | −0.17 | 6.76 | −29.57 | 26.38 |
| 3M | 62/7/3 | 8938 | 0.051 | 1.395 | −0.55 | 16.92 | −25.98 | 11.54 |
| Microsoft | 86/3/14 | 2985 | 0.201 | 2.422 | −0.47 | 12.08 | −30.13 | 17.97 |
| Citi-Grp | 86/10/30 | 2825 | 0.125 | 2.124 | −0.06 | 9.16 | −21.74 | 20.75 |
| | | | | (b) Daily log returns (%) | | | | |
| VW | 62/7/3 | 8938 | 0.046 | 0.803 | −1.66 | 40.06 | −18.84 | 8.31 |
| EW | 62/7/3 | 8938 | 0.080 | 0.676 | −1.29 | 19.98 | −11.08 | 6.72 |
| I.B.M. | 62/7/3 | 8938 | 0.039 | 1.481 | −0.33 | 15.21 | −26.09 | 12.17 |
| Intel | 72/12/15 | 6329 | 0.096 | 2.894 | −0.59 | 8.81 | −35.06 | 23.41 |
| 3M | 62/7/3 | 8938 | 0.041 | 1.403 | −1.05 | 27.03 | −30.08 | 10.92 |
| Microsoft | 86/3/14 | 2985 | 0.171 | 2.443 | −1.10 | 19.65 | −35.83 | 16.53 |
| Citi-Grp | 86/10/30 | 2825 | 0.102 | 2.128 | −0.44 | 10.68 | −24.51 | 18.86 |
| | | | | (c) Monthly simple returns (%) | | | | |
| VW | 26/1 | 864 | 0.99 | 5.49 | 0.23 | 8.13 | −29.00 | 38.28 |
| EW | 26/1 | 864 | 1.32 | 7.54 | 1.65 | 15.24 | −31.23 | 65.51 |
| I.B.M. | 26/1 | 864 | 1.42 | 6.70 | 0.17 | 1.94 | −26.19 | 35.12 |
| Intel | 72/12 | 300 | 2.86 | 12.95 | 0.59 | 3.29 | −44.87 | 62.50 |
| 3M | 46/2 | 623 | 1.36 | 6.46 | 0.16 | 0.89 | −27.83 | 25.77 |
| Microsoft | 86/4 | 141 | 4.26 | 10.96 | 0.81 | 2.32 | −24.91 | 51.55 |
| Citi-Grp | 86/11 | 134 | 2.55 | 9.17 | −0.14 | 0.47 | −26.46 | 26.08 |
| | | | | (d) Monthly log returns (%) | | | | |
| VW | 26/1 | 864 | 0.83 | 5.48 | −0.53 | 7.31 | −34.25 | 32.41 |
| EW | 26/1 | 864 | 1.04 | 7.24 | 0.34 | 8.91 | −37.44 | 50.38 |
| I.B.M. | 26/1 | 864 | 1.19 | 6.63 | −0.22 | 2.05 | −30.37 | 30.10 |
| Intel | 72/12 | 300 | 2.03 | 12.63 | −0.32 | 3.20 | −59.54 | 48.55 |
| 3M | 46/2 | 623 | 1.15 | 6.39 | −0.14 | 1.32 | −32.61 | 22.92 |
| Microsoft | 86/4 | 141 | 3.64 | 10.29 | 0.29 | 1.32 | −28.64 | 41.58 |
| Citi-Grp | 86/11 | 134 | 2.11 | 9.11 | −0.50 | 1.14 | −30.73 | 23.18 |

of a monthly return series is slightly larger. (c) Monthly returns have higher standard deviations than daily returns. (d) Among the daily returns, market indexes have smaller standard deviations than individual stocks. This is in agreement with common sense. (e) The skewness is not a serious problem for both daily and monthly

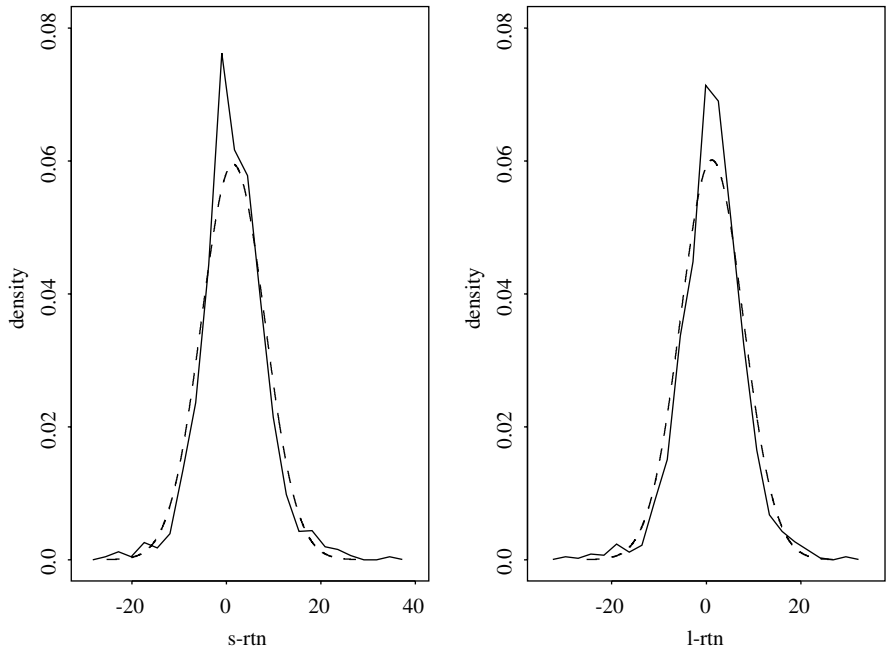


**Figure 1.4.** Comparison of empirical and normal densities for the monthly simple and log returns of IBM stock. The sample period is from January 1926 to December 1997. The left plot is for simple returns and the right plot for log returns. The normal density, shown by the dashed line, uses the sample mean and standard deviation given in Table 1.2.

returns. (f) The descriptive statistics show that the difference between simple and log returns is not substantial.

Figure 1.4 shows the empirical density functions of monthly simple and log returns of IBM stock. Also shown, by a dashed line, in each graph is the normal probability density function evaluated by using the sample mean and standard deviation of IBM returns given in Table 1.2. The plots indicate that the normality assumption is questionable for monthly IBM stock returns. The empirical density function has a higher peak around its mean, but fatter tails than that of the corresponding normal distribution. In other words, the empirical density function is taller, skinnier, but with a wider support than the corresponding normal density.

## 1.3 PROCESSES CONSIDERED

Besides the return series, we also consider the volatility process and the behavior of extreme returns of an asset. The volatility process is concerned with the evolution of conditional variance of the return over time. This is a topic of interest because, as shown in Figures 1.2 and 1.3, the variabilities of returns vary over time and appear in

clusters. In application, volatility plays an important role in pricing stock options. By extremes of a return series, we mean the large positive or negative returns. Table 1.2 shows that the minimum and maximum of a return series can be substantial. The negative extreme returns are important in risk management, whereas positive extreme returns are critical to holding a short position. We study properties and applications of extreme returns, such as the frequency of occurrence, the size of an extreme, and the impacts of economic variables on the extremes, in Chapter 7.

Other financial time series considered in the book include interest rates, exchange rates, bond yields, and quarterly earning per share of a company. Figure 1.5 shows the time plots of two U.S. monthly interest rates. They are the 10-year and 1-year Treasury constant maturity rates from April 1954 to January 2001. As expected, the two interest rates moved in unison, but the 1-year rates appear to be more volatile. Table 1.3 provides some descriptive statistics for selected U.S. financial time series. The monthly bond returns obtained from CRSP are from January 1942 to December 1999. The interest rates are obtained from the Federal Reserve Bank of St Louis. The weekly 3-month Treasury Bill rate started on January 8, 1954, and the 6-month rate started on December 12, 1958. Both series ended on February 16, 2001. For the interest rate series, the sample means are proportional to the time to maturity, but the sample standard deviations are inversely proportional to the time to maturity. For

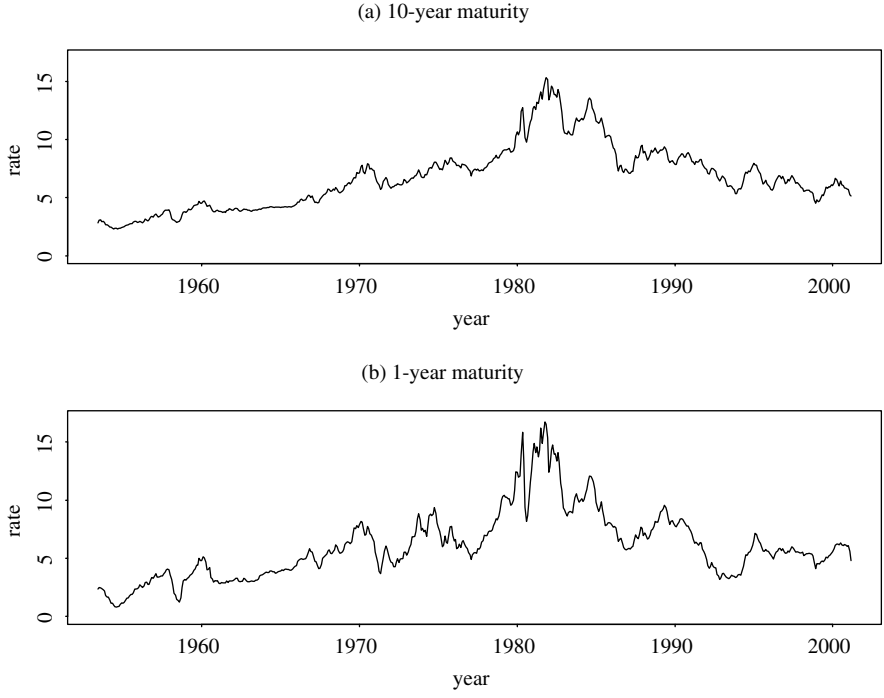


**Figure 1.5.** Time plots of monthly U.S. interest rates from April 1954 to January 2001: (a) the 10-year Treasury constant maturity rate, and (b) the 1-year maturity rate.

**Table 1.3. Descriptive Statistics of Selected U.S. Financial Time Series. The Data Are in Percentages. The Weekly 3-Month Treasury Bill Rate Started from January 8, 1954 and the 6-Month Rate Started from December 12, 1958.**

| Maturity | Mean | Stan. Dev. | Skew. | Excess Kurt. | Min. | Max. |
|---|---|---|---|---|---|---|
| (a) Monthly bond returns: Jan. 1942 to Dec. 1999, $T = 696$ | | | | | | |
| 30 years | 0.43 | 2.53 | 0.66 | 2.77 | −7.73 | 13.31 |
| 20 years | 0.45 | 2.43 | 0.79 | 4.08 | −8.41 | 15.24 |
| 10 years | 0.45 | 1.97 | 0.71 | 2.72 | −6.67 | 10.00 |
| 5 years | 0.46 | 1.39 | 0.87 | 6.61 | −5.80 | 10.61 |
| 1 year | 0.44 | 0.53 | 2.50 | 16.72 | −1.72 | 5.61 |
| (b) Monthly Treasury rates: Apr. 1953 to Jan. 2001, $T = 574$ | | | | | | |
| 10 years | 6.74 | 2.75 | 0.74 | 0.30 | 2.29 | 15.32 |
| 5 years | 6.59 | 2.78 | 0.83 | 0.63 | 1.85 | 15.93 |
| 3 years | 6.43 | 2.82 | 0.89 | 0.85 | 1.47 | 16.22 |
| 1 year | 6.05 | 2.93 | 1.01 | 1.29 | 0.82 | 16.72 |
| (c) Weekly Treasury Bill rates: end on February 16, 2001 | | | | | | |
| 6 months | 6.08 | 2.56 | 1.26 | 1.82 | 2.35 | 15.76 |
| 3 months | 5.51 | 2.76 | 1.14 | 1.88 | 0.58 | 16.76 |

the bond returns, the sample standard deviations are positively related to the time to maturity, whereas the sample means remain stable for all maturities. Most of the series considered have positive excess kurtoses.

With respect to the empirical characteristics of returns shown in Table 1.2, Chapters 2 to 4 focus on the first four moments of a return series and Chapter 7 on the behavior of minimum and maximum returns. Chapters 8 and 9 are concerned with moments of and the relationships between multiple asset returns, and Chapter 5 addresses properties of asset returns when the time interval is small. An introduction to mathematical finance is given in Chapter 6.

## EXERCISES

**1.** Consider the daily stock returns of Alcoa (aa), American Express (axp), Walt Disney (dis), Chicago Tribune (trb), and Tyco International (tyc) from January 1990 to December 1999 for 2528 observations. You may obtain the data directly from CRSP or from files on the Web. The original data are the holding period returns from CRSP. Those on files have been transformed into log returns and are in percentages. Stock tick symbols are used to create file names (e.g., "d-aa9099.dat" contains the daily log returns of Alcoa stock from 1990 to 1999).

- Compute the sample mean, variance, skewness, excess kurtosis, minimum, and maximum of the daily log returns.
- Transform the log returns into simple returns. Compute the sample mean, variance, skewness, excess kurtosis, and minimum and maximum of the daily simple returns.
- Are the sample means of log returns statistically different from zero? Use the 5% significance level to draw your conclusion and discuss their practical implications.

**2.** Consider the monthly stock returns of Alcoa (aa), General Motors (gm), Walt Disney (dis), and Hershey Foods (hsy) from January 1962 to December 1999 for 456 observations and those of American Express (axp) and Mellon Financial Corporation (mel) from January 1973 to December 1999 for 324 observations. Again, you may obtain the data directly from CRSP or from the files on the Web. Tick symbols and years involved are used to create file names (e.g., "m-mel7399.dat" contains the monthly log returns, in percentage, of Mellon Financial Corporation stock from January 1973 to December 1999).

- Compute the sample mean, variance, skewness, excess kurtosis, and minimum and maximum of the monthly log returns.
- Transform the log returns into simple returns. Compute the sample mean, variance, skewness, excess kurtosis, and minimum and maximum of the monthly simple returns.
- Are the sample means of log returns statistically different from zero? Use the 5% significance level to draw your conclusion and discuss their practical implications.

**3.** Focus on the monthly stock returns of Alcoa from 1962 to 1999.

- What is the average annual log return over the data span?
- What is the annualized (average) simple return over the data span?
- Consider an investment that invested one dollar on the Alcoa stock at the beginning of 1962. What was the value of the investment at the end of 1999? Assume that there were no transaction costs.

**4.** Repeat the same analysis as the prior problem for the monthly stock returns of American Express.

**5.** Obtain the histograms of daily simple and log returns of American Express stock from January 1990 to December 1999. Compare them with normal distributions that have the same mean and standard deviation.

**6.** Daily foreign exchange rates can be obtained from the Federal Reserve Bank of Chicago. The data are the noon buying rates in New York City certified by the

Federal Reserve Bank of New York. Consider the exchange rates of Canadian Dollar, German Mark, United Kingdom Pound, Japanese Yen, and French Franc versus the U.S. Dollar from January 1994 to February 2001. The exchange values are payable in foreign currencies, except for U.K. Pound which is in U.S. Dollars. The data are also available in the file "forex-c.dat."

- Compute the daily log returns of the five exchange rate series.
- Compute the sample mean, variance, skewness, excess kurtosis, and minimum and maximum of the five log return series.
- Discuss the empirical characteristics of these exchange rate series.

## REFERENCES

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997), *The Econometrics of Financial Markets*, Princeton University Press: New Jersey.

Cox, J. C., and Rubinstein, M. (1985), *Optoins Markets*, Prentice-Hall: Englewood Cliffs, New Jersey.

Sharpe, W. (1964), "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance*, 19, 425–442.

Snedecor, G. W., and Cochran, W. G. (1980), *Statistical Methods*, 7th edition, Iowa State University Press: Ames, Iowa.

*Analysis of Financial Time Series.* Ruey S. Tsay

ISBN: 0-471-41544-8

CHAPTER 2

# Linear Time Series Analysis and Its Applications

In this chapter, we discuss basic theories of linear time series analysis, introduce some simple econometric models useful for analyzing financial time series, and apply the models to asset returns. Discussions of the concepts are brief with emphasis on those relevant to financial applications. Understanding the simple time series models introduced here will go a long way to better appreciate the more sophisticated financial econometric models of the later chapters. There are many time series textbooks available. For basic concepts of linear time series analysis, see Box, Jenkins, and Reinsel (1994, Chapters 2 and 3) and Brockwell and Davis (1996, Chapters 1–3).

Treating an asset return (e.g., log return $r_t$ of a stock) as a collection of random variables over time, we have a time series $\{r_t\}$. Linear time series analysis provides a natural framework to study the dynamic structure of such a series. The theories of linear time series discussed include stationarity, dynamic dependence, autocorrelation function, modeling, and forecasting. The econometric models introduced include (a) simple autoregressive (AR) models, (b) simple moving-average (MA) models, (c) mixed autoregressive moving-average (ARMA) models, (d) seasonal models, (e) regression models with time series errors, and (f) fractionally differenced models for long-range dependence. For an asset return $r_t$, simple models attempt to capture the linear relationship between $r_t$ and information available prior to time $t$. The information may contain the historical values of $r_t$ and the random vector $\boldsymbol{Y}$ in Eq. (1.14) that describes the economic environment under which the asset price is determined. As such, correlation plays an important role in understanding these models. In particular, correlations between the variable of interest and its past values become the focus of linear time series analysis. These correlations are referred to as *serial correlations* or *autocorrelations*. They are the basic tool for studying a stationary time series.

## 2.1 STATIONARITY

The foundation of time series analysis is stationarity. A time series $\{r_t\}$ is said to be *strictly stationary* if the joint distribution of $(r_{t_1}, \ldots, r_{t_k})$ is identical to that of $(r_{t_1+t}, \ldots, r_{t_k+t})$ for all $t$, where $k$ is an arbitrary positive integer and $(t_1, \ldots, t_k)$ is a collection of $k$ positive integers. In other words, strict stationarity requires that the joint distribution of $(r_{t_1}, \ldots, r_{t_k})$ is invariant under time shift. This is a very strong condition that is hard to verify empirically. A weaker version of stationarity is often assumed. A time series $\{r_t\}$ is *weakly stationary* if both the mean of $r_t$ and the covariance between $r_t$ and $r_{t-\ell}$ are time-invariant, where $\ell$ is an arbitrary integer. More specifically, $\{r_t\}$ is weakly stationary if (a) $E(r_t) = \mu$, which is a constant, and (b) $\text{Cov}(r_t, r_{t-\ell}) = \gamma_\ell$, which only depends on $\ell$. In practice, suppose that we have observed $T$ data points $\{r_t \mid t = 1, \ldots, T\}$. The weak stationarity implies that the time plot of the data would show that the $T$ values fluctuate with constant variation around a constant level.

Implicitly in the condition of weak stationarity, we assume that the first two moments of $r_t$ are finite. From the definitions, if $r_t$ is strictly stationary and its first two moments are finite, then $r_t$ is also weakly stationary. The converse is not true in general. However, if the time series $r_t$ is normally distributed, then weak stationarity is equivalent to strict stationarity. In this book, we are mainly concerned with weakly stationary series.

The covariance $\gamma_\ell = \text{Cov}(r_t, r_{t-\ell})$ is called the lag-$\ell$ autocovariance of $r_t$. It has two important properties: (a) $\gamma_0 = \text{Var}(r_t)$ and (b) $\gamma_{-\ell} = \gamma_\ell$. The second property holds because $\text{Cov}(r_t, r_{t-(-\ell)}) = \text{Cov}(r_{t-(-\ell)}, r_t) = \text{Cov}(r_{t+\ell}, r_t) = \text{Cov}(r_{t_1}, r_{t_1-\ell})$, where $t_1 = t + \ell$.

In the finance literature, it is common to assume that an asset return series is weakly stationary. This assumption can be checked empirically provided that a sufficient number of historical returns are available. For example, one can divide the data into subsamples and check the consistency of the results obtained.

## 2.2 CORRELATION AND AUTOCORRELATION FUNCTION

The correlation coefficient between two random variables $X$ and $Y$ is defined as

$$\rho_{x,y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}} = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sqrt{E(X - \mu_x)^2 E(Y - \mu_y)^2}},$$

where $\mu_x$ and $\mu_y$ are the mean of $X$ and $Y$, respectively, and it is assumed that the variances exist. This coefficient measures the strength of linear dependence between $X$ and $Y$, and it can be shown that $-1 \le \rho_{x,y} \le 1$ and $\rho_{x,y} = \rho_{y,x}$. The two random variables are uncorrelated if $\rho_{x,y} = 0$. In addition, if both $X$ and $Y$ are normal random variables, then $\rho_{x,y} = 0$ if and only if $X$ and $Y$ are independent. When the sample $\{(x_t, y_t)\}_{t=1}^T$ is available, the correlation can be consistently estimated by its

sample counterpart

$$\hat{\rho}_{x,y} = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{T}(x_t - \bar{x})^2 \sum_{t=1}^{T}(y_t - \bar{y})^2}},$$

where $\bar{x} = \sum_{t=1}^{T} x_t/T$ and $\bar{y} = \sum_{t=1}^{T} y_t/T$ are the sample mean of $X$ and $Y$, respectively.

***Autocorrelation Function (ACF)***

Consider a weakly stationary return series $r_t$. When the linear dependence between $r_t$ and its past values $r_{t-i}$ is of interest, the concept of correlation is generalized to autocorrelation. The correlation coefficient between $r_t$ and $r_{t-\ell}$ is called the lag-$\ell$ *autocorrelation* of $r_t$ and is commonly denoted by $\rho_\ell$, which under the weak stationarity assumption is a function of $\ell$ only. Specifically, we define

$$\rho_\ell = \frac{\text{Cov}(r_t, r_{t-\ell})}{\sqrt{\text{Var}(r_t)\,\text{Var}(r_{t-\ell})}} = \frac{\text{Cov}(r_t, r_{t-\ell})}{\text{Var}(r_t)} = \frac{\gamma_\ell}{\gamma_0}, \tag{2.1}$$

where the property $\text{Var}(r_t) = \text{Var}(r_{t-\ell})$ for a weakly stationary series is used. From the definition, we have $\rho_0 = 1$, $\rho_\ell = \rho_{-\ell}$, and $-1 \le \rho_\ell \le 1$. In addition, a weakly stationary series $r_t$ is not serially correlated if and only if $\rho_\ell = 0$ for all $\ell > 0$.

For a given sample of returns $\{r_t\}_{t=1}^{T}$, let $\bar{r}$ be the sample mean (i.e., $\bar{r} = \sum_{t=1}^{T} r_t/T$). Then the lag-1 sample autocorrelation of $r_t$ is

$$\hat{\rho}_1 = \frac{\sum_{t=2}^{T}(r_t - \bar{r})(r_{t-1} - \bar{r})}{\sum_{t=1}^{T}(r_t - \bar{r})^2}.$$

Under some general conditions, $\hat{\rho}_1$ is a consistent estimate of $\rho_1$. For example, if $\{r_t\}$ is an independent and identically distributed (iid) sequence and $E(r_t^2) < \infty$, then $\hat{\rho}_1$ is asymptotically normal with mean zero and variance $1/T$; see Brockwell and Davis (1991, Theorem 7.2.2). This result can be used in practice to test the null hypothesis $H_o : \rho_1 = 0$ versus the alternative hypothesis $H_a : \rho_1 \neq 0$. The test statistic is the usual $t$ ratio, which is $\sqrt{T}\hat{\rho}_1$ and follows asymptotically the standard normal distribution. In general, the lag-$\ell$ sample autocorrelation of $r_t$ is defined as

$$\hat{\rho}_\ell = \frac{\sum_{t=\ell+1}^{T}(r_t - \bar{r})(r_{t-\ell} - \bar{r})}{\sum_{t=1}^{T}(r_t - \bar{r})^2}, \quad 0 \le \ell < T - 1. \tag{2.2}$$

If $\{r_t\}$ is an iid sequence satisfying $E(r_t^2) < \infty$, then $\hat{\rho}_\ell$ is asymptotically normal with mean zero and variance $1/T$ for any fixed positive integer $\ell$. More generally, if $r_t$ is a weakly stationary time series satisfying $r_t = \mu + \sum_{i=0}^{q} \psi_i a_{t-i}$, where $\psi_0 = 1$ and $\{a_j\}$ is a Gaussian white noise series, then $\hat{\rho}_\ell$ is asymptotically normal with mean zero and variance $(1 + 2\sum_{i=1}^{q} \rho_i^2)/T$ for $\ell > q$. This is referred to as Bartlett's formula in the time series literature; see Box, Jenkins, and Reinsel (1994). The previous

result can be used to perform the hypothesis testing of $H_o : \rho_\ell = 0$ vs $H_a : \rho_\ell \neq 0$. For more information about the asymptotic distribution of sample autocorrelations, see Fuller (1976, Chapter 6) and Brockwell and Davis (1991, Chapter 7).

In finite samples, $\hat{\rho}_\ell$ is a biased estimator of $\rho_\ell$. The bias is in the order of $1/T$, which can be substantial when the sample size $T$ is small. In most financial applications, $T$ is relatively large so that the bias is not serious.

***Portmanteau Test***

Financial applications often require to test jointly that several autocorrelations of $r_t$ are zero. Box and Pierce (1970) propose the Portmanteau statistic

$$Q^*(m) = T \sum_{\ell=1}^{m} \hat{\rho}_\ell^2$$

as a test statistic for the null hypothesis $H_o : \rho_1 = \cdots = \rho_m = 0$ against the alternative hypothesis $H_a : \rho_i \neq 0$ for some $i \in \{1, \ldots, m\}$. Under the assumption that $\{r_t\}$ is an iid sequence with certain moment conditions, $Q^*(m)$ is asymptotically a chi-squared random variable with $m$ degrees of freedom.

Ljung and Box (1978) modify the $Q^*(m)$ statistic as below to increase the power of the test in finite samples,

$$Q(m) = T(T+2) \sum_{\ell=1}^{m} \frac{\hat{\rho}_\ell^2}{T-\ell}. \tag{2.3}$$

In practice, the selection of $m$ may affect the performance of the $Q(m)$ statistic. Several values of $m$ are often used. Simulation studies suggest that the choice of $m \approx \ln(T)$ provides better power performance.

The function $\hat{\rho}_1, \hat{\rho}_2, \ldots$ is called the *sample autocorrelation function* (ACF) of $r_t$. It plays an important role in linear time series analysis. As a matter of fact, a linear time series model can be characterized by its ACF, and linear time series modeling makes use of the sample ACF to capture the linear dynamic of the data. Figure 2.1 shows the sample autocorrelation functions of monthly simple and log returns of IBM stock from January 1926 to December 1997. The two sample ACFs are very close to each other, and they suggest that the serial correlations of monthly IBM stock returns are very small, if any. The sample ACFs are all within their two standard-error limits, indicating that they are not significant at the 5% level. In addition, for the simple returns, the Ljung–Box statistics give $Q(5) = 5.4$ and $Q(10) = 14.1$, which correspond to $p$ value of 0.37 and 0.17, respectively, based on chi-squared distributions with 5 and 10 degrees of freedom. For the log returns, we have $Q(5) = 5.8$ and $Q(10) = 13.7$ with $p$ value 0.33 and 0.19, respectively. The joint tests confirm that monthly IBM stock returns have no significant serial correlations. Figure 2.2 shows the same for the monthly returns of the value-weighted index from the Center for Research in Security Prices (CRSP), University of Chicago. There are some significant serial correlations at the 5% level for both return series. The Ljung–Box statistics give $Q(5) = 27.8$ and $Q(10) = 36.0$ for the simple returns and $Q(5) = 26.9$

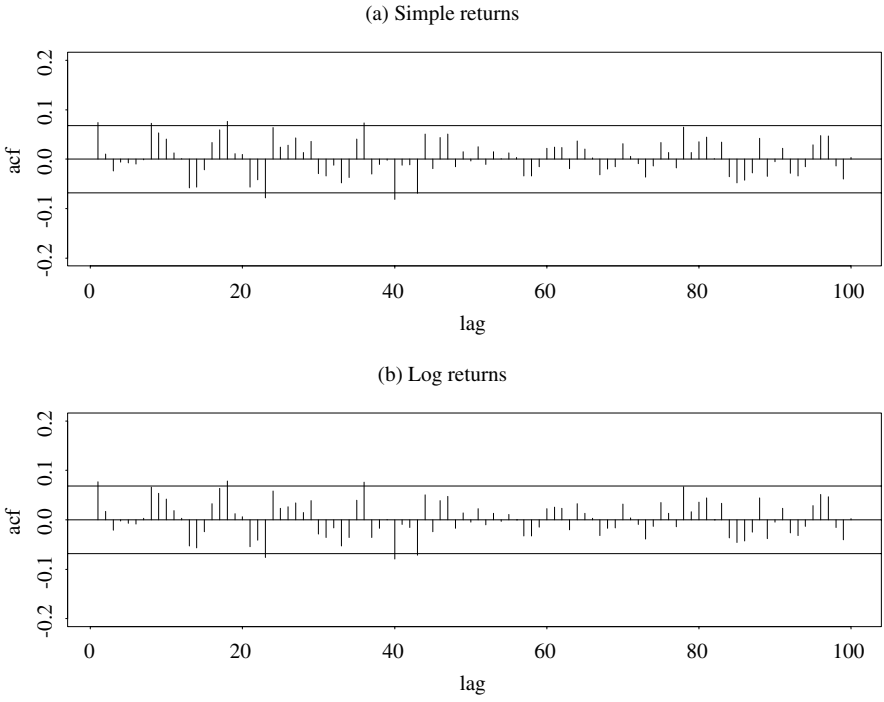


**Figure 2.1.** Sample autocorrelation functions of monthly simple and log returns of IBM stock from January 1926 to December 1997. In each plot, the two horizontal lines denote two standard-error limits of the sample ACF.

and $Q(10) = 32.7$ for the log returns. The $p$ values of these four test statistics are all less than 0.0003, suggesting that monthly returns of the value-weighted index are serially correlated. Thus, the monthly market index return seems to have stronger serial dependence than individual stock returns.

In the finance literature, a version of the Capital Asset Pricing Model (CAPM) theory is that the return $\{r_t\}$ of an asset is not predictable and should have no autocorrelations. Testing for zero autocorrelations has been used as a tool to check the efficient market assumption. However, the way by which stock prices are determined and index returns are calculated might introduce autocorrelations in the observed return series. This is particularly so in analysis of high-frequency financial data. We discuss some of these issues in Chapter 5.

## 2.3 WHITE NOISE AND LINEAR TIME SERIES

### *White Noise*

A time series $r_t$ is called a white noise if $\{r_t\}$ is a sequence of independent and identically distributed random variables with finite mean and variance. In particular,

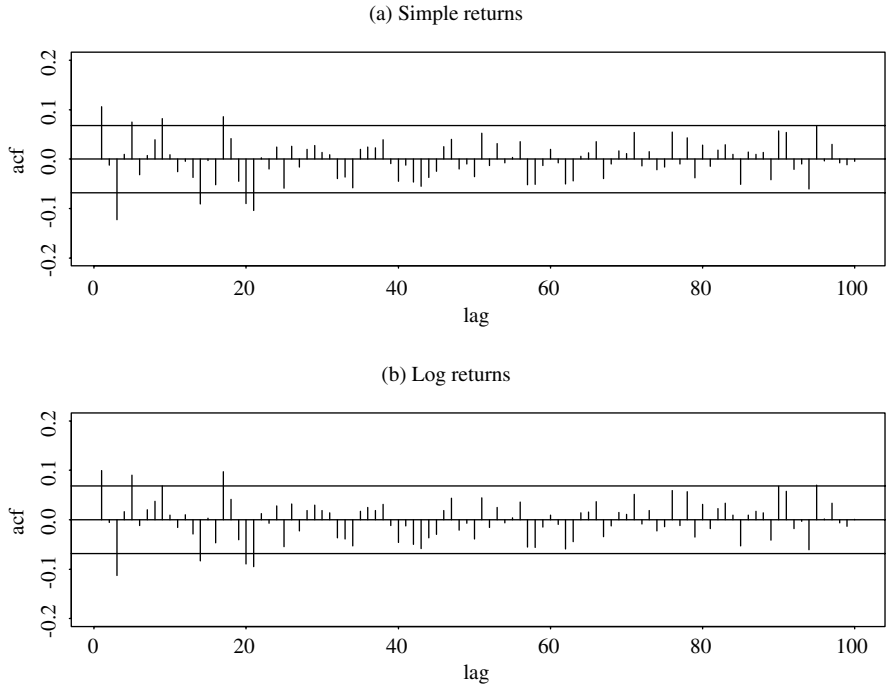


**Figure 2.2.** Sample autocorrelation functions of monthly simple and log returns of the value-weighted index of U.S. Markets from January 1926 to December 1997. In each plot, the two horizontal lines denote two standard-error limits of the sample ACF.

if $r_t$ is normally distributed with mean zero and variance $\sigma^2$, the series is called a Gaussian white noise. For a white noise series, all the ACFs are zero. In practice, if all sample ACFs are close to zero, then the series is a white noise series. Based on Figures 2.1 and 2.2, the monthly returns of IBM stock are close to white noise, whereas those of the value-weighted index are not.

The behavior of sample autocorrelations of the value-weighted index returns indicates that for some asset returns it is necessary to model the serial dependence before further analysis can be made. In what follows, we discuss some simple time series models that are useful in modeling the dynamic structure of a time series. The concepts presented are also useful later in modeling volatility of asset returns.

### *Linear Time Series*

A time series $r_t$ is said to be linear if it can be written as

$$r_t = \mu + \sum_{i=0}^{\infty} \psi_i a_{t-i}, \tag{2.4}$$

where $\mu$ is the mean of $r_t$, $\psi_0 = 1$ and $\{a_t\}$ is a sequence of independent and identically distributed random variables with mean zero and a well-defined distribution (i.e., $\{a_t\}$ is a white noise series). In this book, we are mainly concerned with the case where $a_t$ is a continuous random variable. Not all financial time series are linear, however. We study nonlinearity and nonlinear models in Chapter 4.

For a linear time series in Eq. (2.4), the dynamic structure of $r_t$ is governed by the coefficients $\psi_i$, which are called the *$\psi$-weights* of $r_t$ in the time series literature. If $r_t$ is weakly stationary, we can obtain its mean and variance easily by using the independence of $\{a_t\}$ as

$$E(r_t) = \mu, \qquad \text{Var}(r_t) = \sigma_a^2 \sum_{i=0}^{\infty} \psi_i^2,$$

where $\sigma_a^2$ is the variance of $a_t$. Furthermore, the lag-$\ell$ autocovariance of $r_t$ is

$$\begin{aligned}\gamma_\ell &= \text{Cov}(r_t, r_{t-\ell}) = E\left[\left(\sum_{i=0}^{\infty} \psi_i a_{t-i}\right)\left(\sum_{j=0}^{\infty} \psi_j a_{t-\ell-j}\right)\right] \\ &= E\left(\sum_{i,j=0}^{\infty} \psi_i \psi_j a_{t-i} a_{t-\ell-j}\right) \\ &= \sum_{j=0}^{\infty} \psi_{j+\ell} \psi_j E(a_{t-\ell-j}^2) = \sigma_a^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+\ell}.\end{aligned}$$

Consequently, the $\psi$-weights are related to the autocorrelations of $r_t$ as follows:

$$\rho_\ell = \frac{\gamma_\ell}{\gamma_0} = \frac{\sum_{i=0}^{\infty} \psi_i \psi_{i+\ell}}{1 + \sum_{i=1}^{\infty} \psi_i^2}, \quad \ell \geq 0, \tag{2.5}$$

where $\psi_0 = 1$. Linear time series models are econometric and statistical models used to describe the pattern of the $\psi$-weights of $r_t$.

## 2.4 SIMPLE AUTOREGRESSIVE MODELS

The fact that the monthly return $r_t$ of CRSP value-weighted index has a statistically significant lag-1 autocorrelation indicates that the lagged return $r_{t-1}$ might be useful in predicting $r_t$. A simple model that makes use of such predictive power is

$$r_t = \phi_0 + \phi_1 r_{t-1} + a_t, \tag{2.6}$$

where $\{a_t\}$ is assumed to be a white noise series with mean zero and variance $\sigma_a^2$. This model is in the same form as the well-known simple linear regression model in

which $r_t$ is the dependent variable and $r_{t-1}$ is the explanatory variable. In the time series literature, Model (2.6) is referred to as a simple autoregressive (AR) model of order 1 or simply an AR(1) model. This simple model is also widely used in stochastic volatility modeling when $r_t$ is replaced by its log volatility; see Chapters 3 and 10.

The AR(1) model in Eq. (2.6) has several properties similar to those of the simple linear regression model. However, there are some significant differences between the two models, which we discuss later. Here it suffices to note that an AR(1) model implies that, conditional on the past return $r_{t-1}$, we have

$$E(r_t \mid r_{t-1}) = \phi_0 + \phi_1 r_{t-1}, \qquad \text{Var}(r_t \mid r_{t-1}) = \text{Var}(a_t) = \sigma_a^2.$$

That is, given the past return $r_{t-1}$, the current return is centered around $\phi_0 + \phi_1 r_{t-1}$ with variability $\sigma_a^2$. This is a Markov property such that conditional on $r_{t-1}$, the return $r_t$ is not correlated with $r_{t-i}$ for $i > 1$. Obviously, there are situations in which $r_{t-1}$ alone cannot determine the conditional expectation of $r_t$ and a more flexible model must be sought. A straightforward generalization of the AR(1) model is the AR($p$) model

$$r_t = \phi_0 + \phi_1 r_{t-1} + \cdots + \phi_p r_{t-p} + a_t, \tag{2.7}$$

where $p$ is a non-negative integer and $\{a_t\}$ is defined in Eq. (2.6). This model says that the past $p$ values $r_{t-i}$ ($i = 1, \ldots, p$) jointly determine the conditional expectation of $r_t$ given the past data. The AR($p$) model is in the same form as a multiple linear regression model with lagged values serving as the explanatory variables.

### 2.4.1 Properties of AR models

For effective use of AR models, it pays to study their basic properties. We discuss properties of AR(1) and AR(2) models in detail and give the results for the general AR($p$) model.

***AR(1) Model***

We begin with the sufficient and necessary condition for weak stationarity of the AR(1) model in Eq. (2.6). Assuming that the series is weakly stationary, we have $E(r_t) = \mu$, $\text{Var}(r_t) = \gamma_0$, and $\text{Cov}(r_t, r_{t-j}) = \gamma_j$, where $\mu$ and $\gamma_0$ are constant and $\gamma_j$ is a function of $j$, not $t$. We can easily obtain the mean, variance, and autocorrelations of the series as follows. Taking the expectation of Eq. (2.6) and because $E(a_t) = 0$, we obtain

$$E(r_t) = \phi_0 + \phi_1 E(r_{t-1}).$$

Under the stationarity condition, $E(r_t) = E(r_{t-1}) = \mu$ and hence

$$\mu = \phi_0 + \phi_1 \mu \quad \text{or} \quad E(r_t) = \mu = \frac{\phi_0}{1 - \phi_1}.$$

This result has two implications for $r_t$. First, the mean of $r_t$ exists if $\phi_1 \neq 1$. Second, the mean of $r_t$ is zero if and only if $\phi_0 = 0$. Thus, for a stationary AR(1) process, the constant term $\phi_0$ is related to the mean of $r_t$ and $\phi_0 = 0$ implies that $E(r_t) = 0$.

Next, using $\phi_0 = (1 - \phi_1)\mu$, the AR(1) model can be rewritten as

$$r_t - \mu = \phi_1(r_{t-1} - \mu) + a_t. \tag{2.8}$$

By repeated substitutions, the prior equation implies that

$$\begin{aligned} r_t - \mu &= a_t + \phi_1 a_{t-1} + \phi_1^2 a_{t-2} + \cdots \\ &= \sum_{i=0}^{\infty} \phi_1^i a_{t-i}. \end{aligned} \tag{2.9}$$

Thus, $r_t - \mu$ is a linear function of $a_{t-i}$ for $i \geq 0$. Using this property and the independence of the series $\{a_t\}$, we obtain $E[(r_t - \mu)a_{t+1}] = 0$. By the stationarity assumption, we have $\text{Cov}(r_{t-1}, a_t) = E[(r_{t-1} - \mu)a_t] = 0$. This latter result can also be seen from the fact that $r_{t-1}$ occurred before time $t$ and $a_t$ does not depend on any past information. Taking the square, then the expectation of Eq. (2.8), we obtain

$$\text{Var}(r_t) = \phi_1^2 \,\text{Var}(r_{t-1}) + \sigma_a^2,$$

where $\sigma_a^2$ is the variance of $a_t$ and we make use of the fact that the covariance between $r_{t-1}$ and $a_t$ is zero. Under the stationarity assumption, $\text{Var}(r_t) = \text{Var}(r_{t-1})$, so that

$$\text{Var}(r_t) = \frac{\sigma_a^2}{1 - \phi_1^2}$$

provided that $\phi_1^2 < 1$. The requirement of $\phi_1^2 < 1$ results from the fact that the variance of a random variable is bounded and non-negative. Consequently, the weak stationarity of an AR(1) model implies that $-1 < \phi_1 < 1$. Yet if $-1 < \phi_1 < 1$, then by Eq. (2.9) and the independence of the $\{a_t\}$ series, we can show that the mean and variance of $r_t$ are finite. In addition, by the Cauchy–Schwartz inequality, all the autocovariances of $r_t$ are finite. Therefore, the AR(1) model is weakly stationary. In summary, the necessary and sufficient condition for the AR(1) model in Eq. (2.6) to be weakly stationary is $|\phi_1| < 1$.

***Autocorrelation Function of an AR(1) Model***

Multiplying Eq. (2.8) by $a_t$, using the independence between $a_t$ and $r_{t-1}$, and taking expectation, we obtain

$$E[a_t(r_t - \mu)] = E[a_t(r_{t-1} - \mu)] + E(a_t^2) = E(a_t^2) = \sigma_a^2,$$

where $\sigma_a^2$ is the variance of $a_t$. Multiplying Eq. (2.8) by $(r_{t-\ell} - \mu)$, taking expectation, and using the prior result, we have

$$\gamma_\ell = \begin{cases} \phi_1 \gamma_1 + \sigma_a^2 & \text{if } \ell = 0 \\ \phi_1 \gamma_{\ell-1} & \text{if } \ell > 0, \end{cases}$$

where we use $\gamma_\ell = \gamma_{-\ell}$. Consequently, for a weakly stationary AR(1) model in Eq. (2.6), we have

$$\text{Var}(r_t) = \gamma_0 = \frac{\sigma^2}{1 - \phi_1^2}, \quad \text{and} \quad \gamma_\ell = \phi_1 \gamma_{\ell-1}, \quad \text{for} \quad \ell > 0.$$

From the latter equation, the ACF of $r_t$ satisfies

$$\rho_\ell = \phi_1 \rho_{\ell-1}, \quad \text{for} \quad \ell \geq 0.$$

Because $\rho_0 = 1$, we have $\rho_\ell = \phi_1^\ell$. This result says that the ACF of a weakly stationary AR(1) series decays exponentially with rate $\phi_1$ and starting value $\rho_0 = 1$. For a positive $\phi_1$, the plot of ACF of an AR(1) model shows a nice exponential decay.

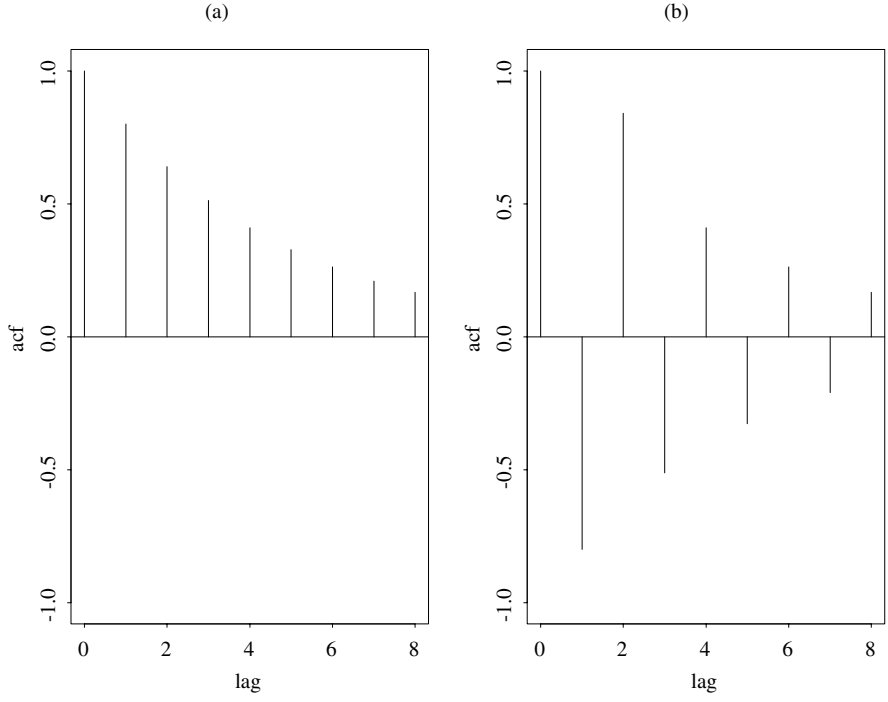


**Figure 2.3.** The autocorrelation function of an AR(1) model: (a) for $\phi_1 = 0.8$, and (b) for $\phi_1 = -0.8$.

For a negative $\phi_1$, the plot consists of two alternating exponential decays with rate $\phi_1^2$. Figure 2.3 shows the ACF of two AR(1) models with $\phi_1 = 0.8$ and $\phi_1 = -0.8$.

***AR(2) Model***

An AR(2) model assumes the form

$$r_t = \phi_0 + \phi_1 r_{t-1} + \phi_2 r_{t-2} + a_t. \tag{2.10}$$

Using the same technique as that of the AR(1) case, we obtain

$$E(r_t) = \mu = \frac{\phi_0}{1 - \phi_1 - \phi_2}$$

provided that $\phi_1 + \phi_2 \neq 1$. Using $\phi_0 = (1 - \phi_1 - \phi_2)\mu$, we can rewrite the AR(2) model as

$$(r_t - \mu) = \phi_1(r_{t-1} - \mu) + \phi_2(r_{t-2} - \mu) + a_t.$$

Multiplying the prior equation by $(r_{t-\ell} - \mu)$, we have

$$(r_{t-\ell} - \mu)(r_t - \mu) = \phi_1(r_{t-\ell} - \mu)(r_{t-1} - \mu) + \phi_2(r_{t-\ell} - \mu)(r_{t-2} - \mu) + (r_{t-\ell} - \mu)a_t.$$

Taking expectation and using $E[(r_{t-\ell} - \mu)a_t] = 0$ for $\ell > 0$, we obtain

$$\gamma_\ell = \phi_1 \gamma_{\ell-1} + \phi_2 \gamma_{\ell-2}, \quad \text{for} \quad \ell > 0.$$

This result is referred to as the *moment equation* of a stationary AR(2) model. Dividing the previous equation by $\gamma_0$, we have the property

$$\rho_\ell = \phi_1 \rho_{\ell-1} + \phi_2 \rho_{\ell-2}, \quad \text{for} \quad \ell > 0, \tag{2.11}$$

for the ACF of $r_t$. In particular, the lag-1 ACF satisfies

$$\rho_1 = \phi_1 \rho_0 + \phi_2 \rho_{-1} = \phi_1 + \phi_2 \rho_1.$$

Therefore, for a stationary AR(2) series $r_t$, we have $\rho_0 = 1$,

$$\rho_1 = \frac{\phi_1}{1 - \phi_2}$$
$$\rho_\ell = \phi_1 \rho_{\ell-1} + \phi_2 \rho_{\ell-2}, \quad \ell \geq 2.$$

The result of Eq. (2.11) says that the ACF of a stationary AR(2) series satisfies the second order difference equation

$$(1 - \phi_1 B - \phi_2 B^2)\rho_\ell = 0,$$

where $B$ is called the *back-shift* operator such that $B\rho_\ell = \rho_{\ell-1}$. This difference equation determines the properties of the ACF of a stationary AR(2) time series. It also determines the behavior of the forecasts of $r_t$. In the time series literature, some people use the notation $L$ instead of $B$ for the back-shift operator. Here $L$ stands for *lag* operator. For instance, $Lr_t = r_{t-1}$ and $L\psi_k = \psi_{k-1}$.

Corresponding to the prior difference equation, there is a second order polynomial equation

$$x^2 - \phi_1 x - \phi_2 = 0.$$

Solutions of this equation are the *characteristic roots* of the AR(2) model, and they are

$$x = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2}.$$

Denote the two characteristic roots by $\omega_1$ and $\omega_2$. If both $\omega_i$ are real valued, then the second order difference equation of the model can be factored as $(1-w_1B)(1-w_2B)$ and the AR(2) model can be regarded as an AR(1) model operates on top of another AR(1) model. The ACF of $r_t$ is then a mixture of two exponential decays. Yet if $\phi_1^2 + 4\phi_2 < 0$, then $\omega_1$ and $\omega_2$ are complex numbers (called a *complex conjugate pair*), and the plot of ACF of $r_t$ would show a picture of damping sine and cosine waves. In business and economic applications, complex characteristic roots are important. They give rise to the behavior of business cycles. It is then common for economic time series models to have complex-valued characteristic roots. For an AR(2) model in Eq. (2.10) with a pair of complex characteristic roots, the *average* length of the stochastic cycles is

$$k = \frac{360^\circ}{\cos^{-1}[\phi_1/(2\sqrt{-\phi_2})]},$$

where the cosine inverse is stated in degrees.

Figure 2.4 shows the ACF of four stationary AR(2) models. Part (b) is the ACF of the AR(2) model $(1-0.6B+0.4B^2)r_t = a_t$. Because $\phi_1^2+4\phi_2 = 0.36+4\times(-0.4) = -1.24 < 0$, this particular AR(2) model contains two complex characteristic roots, and hence its ACF exhibits damping sine and cosine waves. The other three AR(2) models have real-valued characteristic roots. Their ACFs decay exponentially.

**Example 2.1.** As an illustration, consider the quarterly growth rate of U.S. real gross national product (GNP), seasonally adjusted, from the second quarter of 1947 to the first quarter of 1991. This series is used in Chapter 4 as an example of nonlinear economic time series. Here we simply employ an AR(3) model for the data. Denoting the growth rate by $r_t$, we can use the model building procedure of the next subsection to estimate the model. The fitted model is

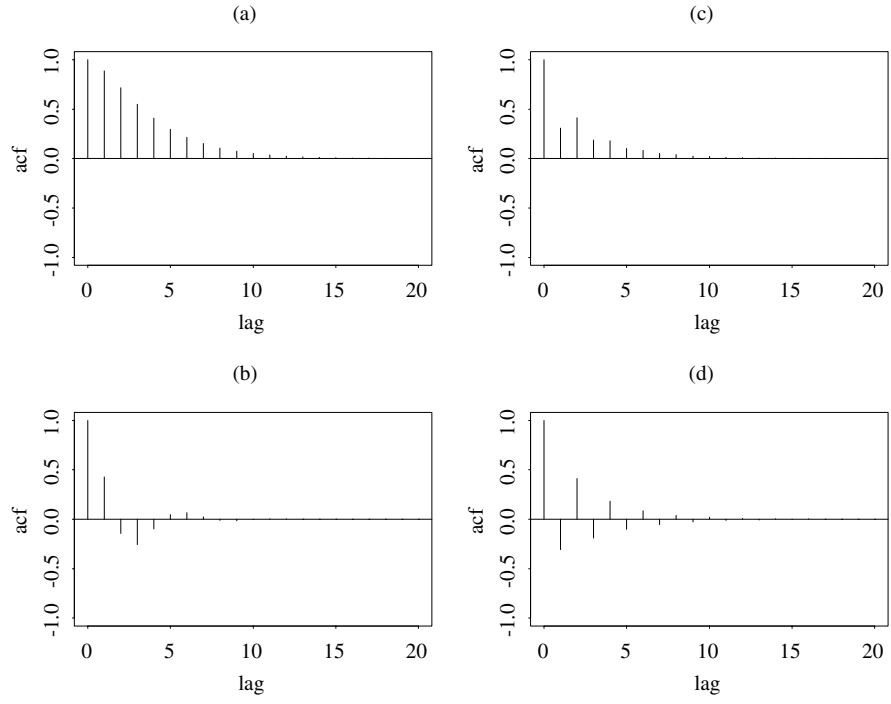


**Figure 2.4.** The autocorrelation function of an AR(2) model: (a) $\phi_1 = 1.2$ and $\phi_2 = -0.35$, (b) $\phi_1 = 0.6$ and $\phi_2 = -0.4$, (c) $\phi_1 = 0.2$ and $\phi_2 = 0.35$, (d) $\phi_1 = -0.2$ and $\phi_2 = 0.35$.

$$r_t = 0.0047 + 0.35r_{t-1} + 0.18r_{t-2} - 0.14r_{t-3} + a_t, \quad \hat{\sigma}_a = 0.0098. \tag{2.12}$$

Rewriting the model as

$$r_t - 0.35r_{t-1} - 0.18r_{t-2} + 0.14r_{t-3} = 0.0047 + a_t,$$

we obtain a corresponding third-order difference equation

$$(1 - 0.35B - 0.18B^2 + 0.14B^3) = 0,$$

which can be factored as

$$(1 + 0.52B)(1 - 0.87B + 0.27B^2) = 0.$$

The first factor $(1 + 0.52B)$ shows an exponentially decaying feature of the GNP growth rate. Focusing on the second-order factor $1 - 0.87B - (-0.27)B^2 = 0$, we have $\phi_1^2 + 4\phi_2 = 0.87^2 + 4(-0.27) = -0.3231 < 0$. Therefore, the second factor of the AR(3) model confirms the existence of stochastic business cycles in the quarterly growth rate of U.S. real GNP. This is reasonable as the U.S. economy went through

expansion and contraction periods. The average length of the stochastic cycles is approximately

$$k = \frac{360^\circ}{\cos^{-1}[\phi_1/(2\sqrt{-\phi_2})]} = 10.83 \text{ quarters},$$

which is about 3 years. If one uses a nonlinear model to separate U.S. economy into "expansion" and "contraction" periods, the data show that the average duration of contraction periods is about three quarters and that of expansion periods is about 3 years; see the analysis in Chapter 4. The average duration of 10.83 quarters is a compromise between the two separate durations. The periodic feature obtained here is common among growth rates of national economies. For example, similar features can be found for OECD countries.

***Stationarity***

The stationarity condition of an AR(2) time series is that the absolute values of its two characteristic roots are less one or, equivalently, its two characteristic roots are less than one in modulus. Under the prior condition, the recursive equation in (2.11) ensures that the ACF of the model converges to zero as the lag $\ell$ increases. This convergence property is a necessary condition for a stationary time series. In fact, the condition also applies to the AR(1) model where the polynomial equation is $x - \phi_1 = 0$. The characteristic root is $x = \phi_1$, which must be less than 1 in modulus for $r_t$ to be stationary. As shown before, $\rho_\ell = \phi_1^\ell$ for a stationary AR(1) model. The condition ensures that $\rho_\ell \to 0$ as $\ell \to \infty$.

**AR($p$) *Model***

The results of AR(1) and AR(2) models can readily be generalized to the general AR($p$) model in Eq. (2.7). The mean of a stationary series is

$$E(r_t) = \frac{\phi_0}{1 - \phi_1 - \cdots - \phi_p}$$

provided that the denominator is not zero. The associated polynomial equation of the model is

$$x^p - \phi_1 x^{p-1} - \phi_2 x^{p-2} - \cdots - \phi_p = 0,$$

which is referred to as the *characteristic equation* of the model. If all the *characteristic roots* of this equation are less than one in modulus, then the series $r_t$ is stationary. For a stationary AR($p$) series, the ACF satisfies the difference equation

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)\rho_\ell = 0, \quad \text{for} \quad \ell > 0.$$

The plot of ACF of a stationary AR($p$) model would then show a mixture of damping sine and cosine patterns and exponential decays depending on the nature of its characteristic roots.

### 2.4.2 Identifying AR Models in Practice

In application, the order $p$ of an AR time series is unknown. It must be specified empirically. This is referred to as the *order determination* of AR models, and it has been extensively studied in the time series literature. Two general approaches are available for determining the value of $p$. The first approach is to use the partial autocorrelation function, and the second approach uses some information criterion function.

***Partial Autocorrelation Function (PACF)***

The PACF of a time series is a function of its ACF and is a useful tool for determining the order $p$ of an AR model. A simple, yet effective way to introduce PACF is to consider the following AR models in consecutive orders:

$$\begin{aligned}
r_t &= \phi_{0,1} + \phi_{1,1} r_{t-1} + e_{1t}, \\
r_t &= \phi_{0,2} + \phi_{1,2} r_{t-1} + \phi_{2,2} r_{t-2} + e_{2t}, \\
r_t &= \phi_{0,3} + \phi_{1,3} r_{t-1} + \phi_{2,3} r_{t-2} + \phi_{3,3} r_{t-3} + e_{3t}, \\
r_t &= \phi_{0,4} + \phi_{1,4} r_{t-1} + \phi_{2,4} r_{t-2} + \phi_{3,4} r_{t-3} + \phi_{4,4} r_{t-4} + e_{4t}, \\
&\vdots \quad \vdots
\end{aligned}$$

where $\phi_{0,j}$, $\phi_{i,j}$, and $\{e_{jt}\}$ are, respectively, the constant term, the coefficient of $r_{t-i}$, and the error term of an AR($j$) model. These models are in the form of a multiple linear regression and can be estimated by the least squares method. As a matter of fact, they are arranged in a sequential order that enables us to apply the idea of partial $F$ test in multiple linear regression analysis. The estimate $\hat{\phi}_{1,1}$ of the first equation is called the lag-1 sample PACF of $r_t$. The estimate $\hat{\phi}_{2,2}$ of the second equation is the lag-2 sample PACF of $r_t$. The estimate $\hat{\phi}_{3,3}$ of the third equation is the lag-3 sample PACF of $r_t$, and so on.

From the definition, the lag-2 PACF $\hat{\phi}_{2,2}$ shows the added contribution of $r_{t-2}$ to $r_t$ over the AR(1) model $r_t = \phi_0 + \phi_1 r_{t-1} + e_{1t}$. The lag-3 PACF shows the added contribution of $r_{t-3}$ to $r_t$ over an AR(2) model, and so on. Therefore, for an AR($p$) model, the lag-$p$ sample PACF should not be zero, but $\hat{\phi}_{j,j}$ should be close to zero for all $j > p$. We make use of this property to determine the order $p$. Indeed, under some regularity conditions, it can be shown that the sample PACF of an AR($p$) process has the following properties:

- $\hat{\phi}_{p,p}$ converges to $\phi_p$ as the sample size $T$ goes to infinity.
- $\hat{\phi}_{\ell,\ell}$ converges to zero for all $\ell > p$.
- The asymptotic variance of $\hat{\phi}_{\ell,\ell}$ is $1/T$ for $\ell > p$.

These results say that, for an AR($p$) series, the sample PACF cuts off at lag $p$.

**Table 2.1. Sample Partial Autocorrelation Function and Akaike Information Criterion for the Monthly Simple Returns of CRSP Value-Weighted Index from January 1926 to December 1997.**

| $p$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| PACF | 0.11 | -0.02 | −0.12 | 0.04 | 0.07 |
| AIC | −5.807 | −5.805 | −5.817 | −5.816 | −5.819 |
| $p$ | 6 | 7 | 8 | 9 | 10 |
| PACF | −0.06 | 0.02 | 0.06 | 0.06 | −0.01 |
| AIC | −5.821 | −5.819 | −5.820 | −5.821 | −5.818 |

As an example, consider the monthly simple returns of CRSP value-weighted index from January 1926 to December 1997. Table 2.1 gives the first 10 lags of sample PACF of the series. With $T = 864$, the asymptotic standard error of the sample PACF is approximately 0.03. Therefore, using the 5% significant level, we identify an AR(3) or AR(5) model for the data (i.e., $p = 3$ or 5).

***Information Criteria***

There are several information criteria available to determine the order $p$ of an AR process. All of them are likelihood based. For example, the well-known *Akaike Information Criterion* (Akaike, 1973) is defined as

$$AIC = \frac{-2}{T} \ln(\text{likelihood}) + \frac{2}{T} \times (\text{number of parameters}), \tag{2.13}$$

where the likelihood function is evaluated at the maximum likelihood estimates and $T$ is the sample size. For a Gaussian AR($\ell$) model, AIC reduces to

$$\text{AIC}(\ell) = \ln(\hat{\sigma}_\ell^2) + \frac{2\ell}{T},$$

where $\hat{\sigma}_\ell^2$ is the maximum likelihood estimate of $\sigma_a^2$, which is the variance of $a_t$, and $T$ is the sample size; see Eq. (1.18). In practice, one computes AIC($\ell$) for $\ell = 0, \ldots, P$, where $P$ is a prespecified positive integer and selects the order $k$ that has the minimum AIC value. The second term of the AIC in Eq. (2.13) is called the *penalty function* of the criterion because it penalizes a candidate model by the number of parameters used. Different penalty functions result in different information criteria.

Table 2.1 also gives the AIC for $p = 1, \ldots, 10$. The AIC values are close to each other with minimum −5.821 occurring at $p = 6$ and 9, suggesting that an AR(6) model is preferred by the criterion. This example shows that different approaches for order determination may result in different choices of $p$. There is no evidence

to suggest that one approach outperforms the other in a real application. Substantive information of the problem under study and simplicity are two factors that also play an important role in choosing an AR model for a given time series.

***Parameter Estimation***

For a specified AR($p$) model in Eq. (2.7), the conditional least squares method, which starts with the $(p + 1)$th observation, is often used to estimate the parameters. Specifically, conditioning on the first $p$ observations, we have

$$r_t = \phi_0 + \phi_1 r_{t-1} + \cdots + \phi_p r_{t-p} + a_t, \quad t = p + 1, \ldots, T,$$

which can be estimated by the least squares method. Denote the estimate of $\phi_i$ by $\hat{\phi}_i$. The *fitted model* is

$$\hat{r}_t = \hat{\phi}_0 + \hat{\phi}_1 r_{t-1} + \cdots + \hat{\phi}_p r_{t-p}$$

and the associated residual is

$$\hat{a}_t = r_t - \hat{r}_t.$$

The series $\{\hat{a}_t\}$ is called the *residual series*, from which we obtain

$$\hat{\sigma}_a^2 = \frac{\sum_{t=p+1}^{T} \hat{a}_t^2}{T - 2p - 1}.$$

For illustration, consider an AR(3) model for the monthly simple returns of the value-weighted index in Table 2.1. The fitted model is

$$r_t = 0.0103 + 0.104 r_{t-1} - 0.010 r_{t-2} - 0.120 r_{t-3} + \hat{a}_t, \quad \hat{\sigma}_a = 0.054.$$

The standard errors of the coefficients are 0.002, 0.034, 0.034, and 0.034, respectively. Except for the lag-2 coefficient, all parameters are statistically significant at the 1% level.

For this example, the AR coefficients of the fitted model are small, indicating that the serial dependence of the series is weak, even though it is statistically significant at the 1% level. The significance of $\hat{\phi}_0$ of the entertained model implies that the expected mean return of the series is positive. In fact, $\hat{\mu} = 0.0103/(1 - 0.104 + 0.010 + 0.120) = 0.01$, which is small, but has an important long-term implication. It implies that the long-term return of the index can be substantial. Using the multi-period simple return defined in Chapter 1, the average annual simple gross return is $[\prod_{t=1}^{864}(1 + R_t)]^{12/864} - 1 \approx 0.1053$. In other words, the monthly simple returns of the CRSP value-weighted index grew about 10.53% per annum from 1926 to 1997, supporting the common belief that equity market performs well in the long term. A one-dollar investment at the beginning of 1926 would be worth about \$1350 at the end of 1997.

***Model Checking***

A fitted model must be examined carefully to check for possible model inadequacy. If the model is adequate, then the residual series should behave as a white noise. The ACF and the Ljung–Box statistics in Eq. (2.3) of the residuals can be used to check the closeness of $\hat{a}_t$ to a white noise. For an AR($p$) model, the Ljung–Box statistic $Q(m)$ follows asymptotically a chi-squared distribution with $m - p$ degrees of freedom. Here the number of degrees of freedom is modified to signify that $p$ AR coefficients are estimated. If a fitted model is found to be inadequate, it must be refined.

Consider the residual series of the fitted AR(3) model for the monthly value-weighted simple returns. We have $Q(10) = 15.8$ with $p$ value 0.027 based on its asymptotic chi-squared distribution with 7 degrees of freedom. Thus, the null hypothesis of no residual serial correlation in the first 10 lags is rejected at the 5% level, but not at the 1% level. If the model is refined to an AR(5) model, then we have

$$r_t = 0.0092 + 0.107r_{t-1} - 0.001r_{t-2} - 0.123r_{t-3} + 0.028r_{t-4} + 0.069r_{t-5} + \hat{a}_t,$$

with $\hat{\sigma}_a = 0.054$. The AR coefficients at lags 1, 3, and 5 are significant at the 5% level. The Ljung–Box statistics give $Q(10) = 11.2$ with $p$ value 0.048. This model shows some improvements and appears to be marginally adequate at the 5% significance level. The mean of $r_t$ based on the refined model is also very close to 0.01, showing that the two models have similar long-term implications.

### 2.4.3 Forecasting

Forecasting is an important application of time series analysis. For the AR($p$) model in Eq. (2.7), suppose that we are at the time index $h$ and are interested in forecasting $r_{h+\ell}$, where $\ell \geq 1$. The time index $h$ is called the *forecast origin* and the positive integer $\ell$ is the *forecast horizon*. Let $\hat{r}_h(\ell)$ be the forecast of $r_{h+\ell}$ using the minimum squared error loss function. In other words, the forecast $\hat{r}_k(\ell)$ is chosen such that

$$E[r_{h+\ell} - \hat{r}_h(\ell)]^2 \leq \min_g E(r_{h+\ell} - g)^2,$$

where $g$ is a function of the information available at time $h$ (inclusive). We referred to $\hat{r}_h(\ell)$ as the $\ell$-step ahead forecast of $r_t$ at the forecast origin $h$.

***1-Step Ahead Forecast***

From the AR($p$) model, we have

$$r_{h+1} = \phi_0 + \phi_1 r_h + \cdots + \phi_p r_{h+1-p} + a_{h+1}.$$

Under the minimum squared error loss function, the point forecast of $r_{h+1}$ given the model and observations up to time $h$ is the conditional expectation

$$\hat{r}_h(1) = E(r_{h+1} \mid r_h, r_{h-1}, \ldots) = \phi_0 + \sum_{i=1}^{p} \phi_i r_{h+1-i}$$

and the associated forecast error is

$$e_h(1) = r_{h+1} - \hat{r}_h(1) = a_{h+1}.$$

Consequently, the variance of the 1-step ahead forecast error is $\text{Var}[e_h(1)] = \text{Var}(a_{h+1}) = \sigma_a^2$. If $a_t$ is normally distributed, then a 95% 1-step ahead interval forecast of $r_{h+1}$ is $\hat{r}_h(1) \pm 1.96 \times \sigma_a$. For the linear model in Eq. (2.4), $a_{t+1}$ is also the 1-step ahead forecast error at the forecast origin $t$. In the econometric literature, $a_{t+1}$ is referred to as the *shock* to the series at time $t + 1$.

In practice, estimated parameters are often used to compute point and interval forecasts. This results in a *conditional forecast* because such a forecast does not take into consideration the uncertainty in the parameter estimates. In theory, one can consider parameter uncertainty in forecasting, but it is much more involved. When the sample size used in estimation is sufficiently large, then the conditional forecast is close to the unconditional one.

***2-Step Ahead Forecast***

Next consider the forecast of $r_{h+2}$ at the forecast origin $h$. From the AR($p$) model, we have

$$r_{h+2} = \phi_0 + \phi_1 r_{h+1} + \cdots + \phi_p r_{h+2-p} + a_{h+2}.$$

Taking conditional expectation, we have

$$\hat{r}_h(2) = E(r_{h+2} \mid r_h, r_{h-1}, \ldots) = \phi_0 + \phi_1 \hat{r}_h(1) + \phi_2 r_h + \cdots + \phi_p r_{h+2-p}$$

and the associated forecast error

$$e_h(2) = r_{h+2} - \hat{r}_h(2) = \phi_1[r_{h+1} - \hat{r}_h(1)] + a_{h+2} = a_{h+2} + \phi_1 a_{h+1}.$$

The variance of the forecast error is $\text{Var}[e_h(2)] = (1 + \phi_1^2)\sigma_a^2$. Interval forecasts of $r_{h+2}$ can be computed in the same way as those for $r_{h+1}$. It is interesting to see that $\text{Var}[e_h(2)] \geq \text{Var}[e_h(1)]$, meaning that as the forecast horizon increases the uncertainty in forecast also increases. This is in agreement with common sense that we are more uncertain about $r_{h+2}$ than $r_{h+1}$ at the time index $h$ for a linear time series.

***Multistep Ahead Forecast***

In general, we have

$$r_{h+\ell} = \phi_0 + \phi_1 r_{h+\ell-1} + \cdots + \phi_p r_{h+\ell-p} + a_{h+\ell}.$$

The $\ell$-step ahead forecast based on the minimum squared error loss function is the conditional expectation of $r_{h+\ell}$ given $\{r_{h-i}\}_{i=0}^{\infty}$, which can be obtained as

$$\hat{r}_h(\ell) = \phi_0 + \sum_{i=1}^{p} \phi_i \hat{r}_h(\ell - i),$$

where it is understood that $\hat{r}_h(i) = r_{h+i}$ if $i \leq 0$. This forecast can be computed recursively using forecasts $\hat{r}_h(i)$ for $i = 1, \ldots, \ell - 1$. The $\ell$-step ahead forecast error is $e_h(\ell) = r_{h+\ell} - \hat{r}_h(\ell)$. It can be shown that for a stationary AR($p$) model, $\hat{r}_h(\ell)$ converges to $E(r_t)$ as $\ell \to \infty$, meaning that for such a series long-term point forecast approaches its unconditional mean. This property is referred to as the *mean reversion* in the finance literature. The variance of the forecast error then approaches the unconditional variance of $r_t$.

Table 2.2 contains the 1-step to 6-step ahead forecasts and the standard errors of the associated forecast errors at the forecast origin 858 for the monthly simple return of the value-weight index using an AR(5) model that was re-estimated using the first 858 observations. The actual returns are also given. Because of the weak serial dependence in the series, the forecasts and standard deviations of forecast errors converge to the sample mean and standard deviation of the data quickly. For the first 858 observations, the sample mean and standard error are 0.0098 and 0.0550, respectively.

Figure 2.5 shows the 1- to 6-step ahead out-of-sample forecasts and their two standard-error limits for the monthly log returns of value-weighted index. As in Table 2.2, the following AR(5) model

$$r_t = 0.0075 + 0.103 r_{t-1} + 0.002 r_{t-2} - 0.114 r_{t-3} + 0.032 r_{t-4} + 0.084 r_{t-5} + a_t,$$

where $\hat{\sigma}_a = 0.054$, is built and used to produce the forecasts. For this example, the forecasts are close to the actual values, and the actual values are all within the 95% interval forecasts. For this particular series, the AR models for monthly simple and log returns are close.

***Remark:*** The prior time series analysis was carried out using the SCA package. The commands used are given in Appendix A.

**Table 2.2. Multistep Ahead Forecasts of an AR(5) Model for the Monthly Simple Returns of CRSP Value-Weighted Index. The Forecast Origin Is 858.**

| Step | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Forecast | 0.0071 | −0.0008 | 0.0086 | 0.0154 | 0.0141 | 0.0100 |
| Std. Error | 0.0541 | 0.0545 | 0.0545 | 0.0549 | 0.0549 | 0.0550 |
| Actual | 0.0762 | −0.0365 | 0.0580 | −0.0341 | 0.0311 | 0.0183 |

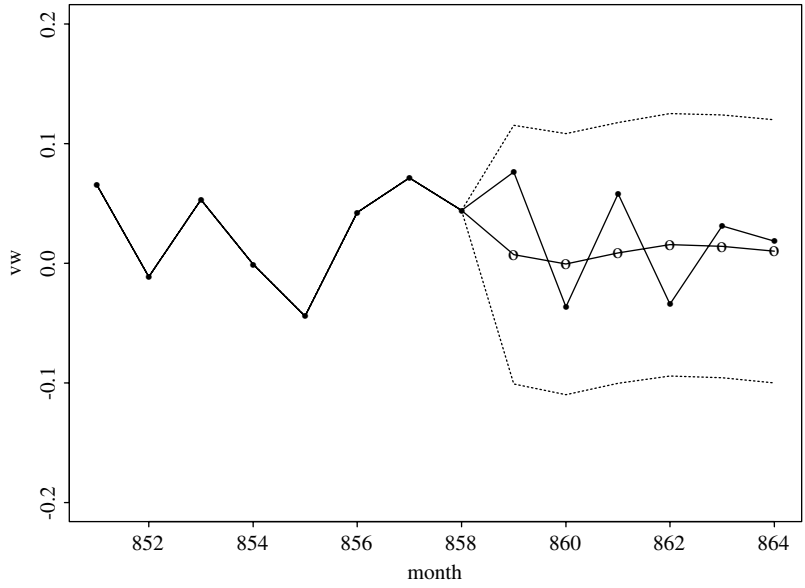


**Figure 2.5.** Plot of 1- to 6-step ahead out-of-sample forecasts for the monthly log returns of the CRSP value-weighted index. The forecast origin is $t = 858$. The forecasts are denoted by "o" and the actual observations by a dot. The two dashed lines denote two standard-error limits of the forecasts.

## 2.5 SIMPLE MOVING-AVERAGE MODELS

We now turn to another class of simple models that are also useful in modeling return series in finance. These models are called moving-average (MA) models. There are several ways to introduce MA models. One approach is to treat the model as a simple extension of white noise series. Another approach is to treat the model as an infinite-order AR model with some parameter constraints. We adopt the second approach. As is shown in Chapter 5, the bid-ask bounce in stock trading may introduce an MA(1) structure in a return series.

There is no particular reason, but simplicity, to assume *a priori* that the order of an AR model is finite. We may entertain, at least in theory, an AR model with infinite order as

$$r_t = \phi_0 + \phi_1 r_{t-1} + \phi_2 r_{t-2} + \cdots + a_t.$$

However, such an AR model is not realistic because it has infinite many parameters. One way to make the model practical is to assume that the coefficients $\phi_i$s satisfy some constraints so that they are determined by a finite number of parameters. A special case of this idea is

$$r_t = \phi_0 - \theta_1 r_{t-1} - \theta_1^2 r_{t-2} - \theta_1^3 r_{t-3} - \cdots + a_t, \tag{2.14}$$

where the coefficients depend on a single parameter $\theta_1$ via $\phi_i = -\theta_1^i$ for $i \geq 1$. For the model in Eq. (2.14) to be stationary, $\theta_1$ must be less than one in absolute value; otherwise, $\theta_1^i$ and the series will explode. Because $|\theta_1| < 1$, we have $\theta_1^i \to 0$ as $i \to \infty$. Thus, the contribution of $r_{t-i}$ to $r_t$ decays exponentially as $i$ increases. This is reasonable as the dependence of a stationary series $r_t$ on its lagged value $r_{t-i}$, if any, should decay over time.

The model in Eq. (2.14) can be rewritten in a rather compact form. To see this, rewrite the model as

$$r_t + \theta_1 r_{t-1} + \theta_1^2 r_{t-2} + \cdots = \phi_0 + a_t. \tag{2.15}$$

The model for $r_{t-1}$ is then

$$r_{t-1} + \theta_1 r_{t-2} + \theta_1^2 r_{t-3} + \cdots = \phi_0 + a_{t-1}. \tag{2.16}$$

Multiplying Eq. (2.16) by $\theta_1$ and subtracting the result from Eq. (2.15), we obtain

$$r_t = \phi_0(1 - \theta_1) + a_t - \theta_1 a_{t-1},$$

which says that except for the constant term $r_t$ is a weighted average of shocks $a_t$ and $a_{t-1}$. Therefore, the model is called an MA model of order 1 or MA(1) model for short. The general form of an MA(1) model is

$$r_t = c_0 + a_t - \theta_1 a_{t-1}, \tag{2.17}$$

where $c_0$ is a constant and $\{a_t\}$ is a white noise series. Similarly, an MA(2) model is in the form

$$r_t = c_0 + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \tag{2.18}$$

and an MA($q$) model is

$$r_t = c_0 + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}, \tag{2.19}$$

where $q > 0$.

### 2.5.1 Properties of MA Models

Again, we focus on the simple MA(1) and MA(2) models. The results of MA($q$) models can easily be obtained by the same techniques.

***Stationarity***

MA models are always weakly stationary because they are finite linear combinations of a white noise sequence for which the first two moments are time-invariant. For example, consider the MA(1) model in Eq. (2.17). Taking expectation of the model,

we have

$$E(r_t) = c_0,$$

which is time-invariant. Taking the variance of Eq. (2.17), we have

$$\text{Var}(r_t) = \sigma_a^2 + \theta_1^2\sigma_a^2 = (1 + \theta_1^2)\sigma_a^2,$$

where we use the fact that $a_t$ and $a_{t-1}$ are uncorrelated. Again, $\text{Var}(r_t)$ is time-invariant. The prior discussion applies to general MA($q$) models, and we obtain two general properties. First, the constant term of an MA model is the mean of the series [i.e., $E(r_t) = c_0$]. Second, the variance of an MA($q$) model is

$$\text{Var}(r_t) = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma_a^2.$$

***Autocorrelation Function***

Assume for simplicity that $c_0 = 0$ for an MA(1) model. Multiplying the model by $r_{t-\ell}$, we have

$$r_{t-\ell}r_t = r_{t-\ell}a_t - \theta_1 r_{t-\ell}a_{t-1}.$$

Taking expectation, we obtain

$$\gamma_1 = -\theta_1\sigma_a^2, \quad \text{and} \quad \gamma_\ell = 0, \quad \text{for} \quad \ell > 1.$$

Using the prior result and the fact that $\text{Var}(r_t) = (1 + \theta_1^2)\sigma_a^2$, we have

$$\rho_0 = 1, \quad \rho_1 = \frac{-\theta_1}{1 + \theta_1^2}, \quad \rho_\ell = 0, \quad \text{for} \quad \ell > 1.$$

Thus, for an MA(1) model, the lag-1 ACF is not zero, but all higher order ACFs are zero. In other words, the ACF of an MA(1) model cuts off at lag 1. For the MA(2) model in Eq. (2.18), the autocorrelation coefficients are

$$\rho_1 = \frac{-\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}, \quad \rho_2 = \frac{-\theta_2}{1 + \theta_1^2 + \theta_2^2}, \quad \rho_\ell = 0, \quad \text{for} \quad \ell > 2.$$

Here the ACF cuts off at lag 2. This property generalizes to other MA models. For an MA($q$) model, the lag-$q$ ACF is not zero, but $\rho_\ell = 0$ for $\ell > q$. Consequently, an MA($q$) series is only linearly related to its first $q$ lagged values and hence is a "finite-memory" model.

### 2.5.2 Identifying MA Order

The ACF is useful in identifying the order of an MA model. For a time series $r_t$ with ACF $\rho_\ell$, if $\rho_q \neq 0$, but $\rho_\ell = 0$ for $\ell > q$, then $r_t$ follows an MA($q$) model.

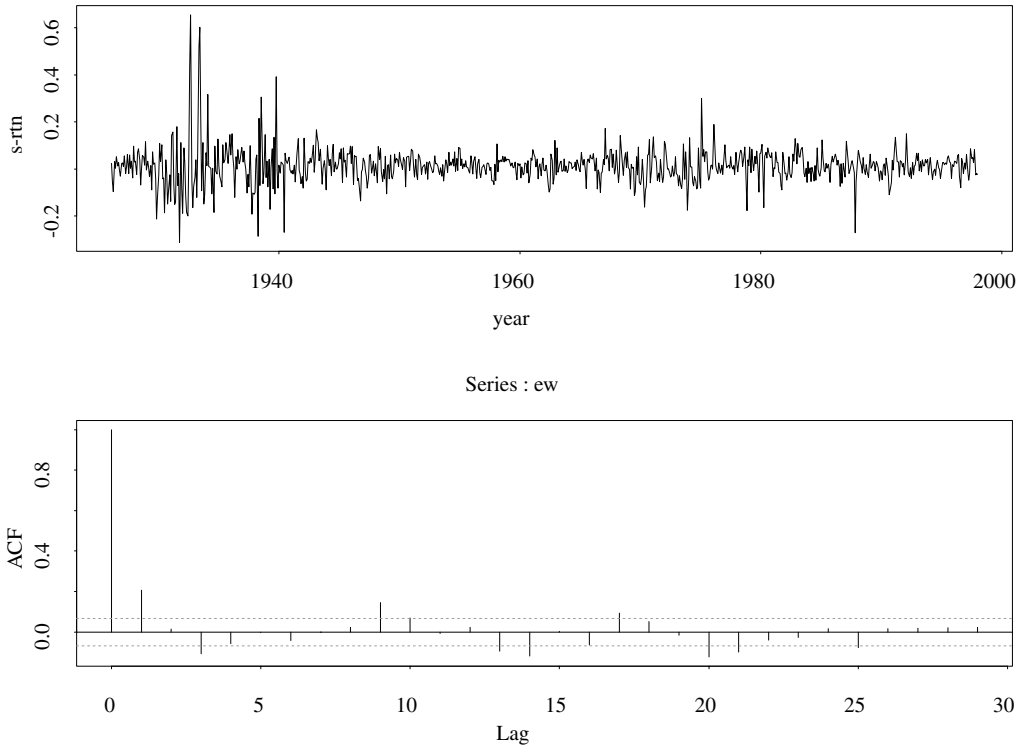


**Figure 2.6.** Time plot and sample autocorrelation function of the monthly simple returns of the CRSP equal-weighted index from January 1926 to December 1997.

Figure 2.6 shows the time plot of monthly simple returns of the CRSP equal-weighted index from January 1926 to December 1997 and the sample ACF of the series. The two dashed lines shown on the ACF plot denote the two standard-error limits. It is seen that the series has significant ACF at lags 1, 3, and 9. There are some marginally significant ACF at higher lags, but we do not consider them here. Based on the sample ACF, the following MA(9) model

$$r_t = c_0 + a_t - \theta_1 a_{t-1} - \theta_3 a_{t-3} - \theta_9 a_{t-9}$$

is identified for the series.

### 2.5.3 Estimation

Maximum likelihood estimation is commonly used to estimate MA models. There are two approaches for evaluating the likelihood function of an MA model. The first approach assumes that the initial shocks (i.e., $a_t$ for $t \leq 0$) are zero. As such, the shocks needed in likelihood function calculation are obtained recursively from the

model, starting with $a_1 = r_1 - c_0$ and $a_2 = r_2 - c_0 + \theta_1 a_1$. This approach is referred to as the *conditional likelihood method* and the resulting estimates are the conditional maximum likelihood estimates. The second approach treats the initial shocks $a_t, t \leq 0$, as additional parameters of the model and estimate them jointly with other parameters. This approach is referred to as the *exact likelihood method*. The exact likelihood estimates are preferred over the conditional ones, but they require more intensive computation. If the sample size is large, then the two types of maximum likelihood estimates are close to each other. For details of conditional and exact likelihood estimates of MA models, readers are referred to Box, Jenkins, and Reinsel (1994) or Chapter 8.

For illustration, consider the monthly simple return series of the CRSP equal-weighted index and the specified MA(9) model. The conditional maximum likelihood method produces the fitted model

$$r_t = 0.0132 + a_t + 0.1775a_{t-1} - 0.1324a_{t-3} + 0.1349a_{t-9}, \quad \hat{\sigma}_a = 0.0727, \quad (2.20)$$

where standard errors of the coefficient estimates are 0.0030, 0.0327, 0.0328, and 0.0328, respectively. The Ljung–Box statistics of the residuals give $Q(10) = 11.4$ with $p$ value 0.122, which is based on an asymptotic chi-squared distribution with 7 degrees of freedom. The model appears to be adequate except for a few marginal residual ACFs at lags 14, 17, and 20. The exact maximum likelihood method produces the fitted model

$$r_t = 0.0132 + a_t + 0.1806a_{t-1} - 0.1315a_{t-3} + 0.1379a_{t-9}, \quad \hat{\sigma}_a = 0.0727, \quad (2.21)$$

where standard errors of the estimates are 0.0029, 0.0329, 0.0330, and 0.0328, respectively. The Ljung–Box statistics of the residuals give $Q(10) = 11.6$ with $p$ value 0.116. This fitted model is also adequate. Comparing models (2.20) and (2.21), we see that for this particular instance, the difference between the conditional and exact likelihood methods is negligible.

### 2.5.4 Forecasting Using MA Models

Forecasts of an MA model can easily be obtained. Because the model has finite memory, its point forecasts go to the mean of the series quickly. To see this, assume that the forecast origin is $h$. For the 1-step ahead forecast of an MA(1) process, the model says

$$r_{h+1} = c_0 + a_{h+1} - \theta_1 a_h.$$

Taking the conditional expectation, we have

$$\hat{r}_h(1) = E(r_{h+1} \mid r_h, r_{h-1}, \ldots) = c_0 - \theta_1 a_h$$
$$e_h(1) = r_{h+1} - \hat{r}_h(1) = a_{h+1}.$$

The variance of the 1-step ahead forecast error is $\text{Var}[e_h(1)] = \sigma_a^2$. In practice, the quantity $a_h$ can be obtained in several ways. For instance, assume that $a_0 = 0$, then $a_1 = r_1 - c_0$, and we can compute $a_t$ for $2 \leq t \leq h$ recursively by using $a_t = r_t - c_0 + \theta_1 a_{t-1}$. Alternatively, it can be computed by using the AR representation of the MA(1) model; see Subsection 2.6.5.

For the 2-step ahead forecast from the equation

$$r_{h+2} = c_0 + a_{h+2} - \theta_1 a_{h+1},$$

we have

$$\hat{r}_h(2) = E(r_{h+2} \mid r_h, r_{h-1}, \ldots) = c_0,$$
$$e_h(2) = r_{h+2} - \hat{r}_h(2) = a_{h+2} - \theta_1 a_{h+1}.$$

The variance of the forecast error is $\text{Var}[e_h(2)] = (1 + \theta_1^2)\sigma_a^2$, which is the variance of the model and is greater than or equal to that of the 1-step ahead forecast error. The prior result shows that for an MA(1) model the 2-step ahead forecast of the series is simply the unconditional mean of the model. This is true for any forecast origin $h$. More generally, $\hat{r}_h(\ell) = c_0$ for $\ell \geq 2$. In summary, for an MA(1) model, the 1-step ahead point forecast at the forecast origin $h$ is $c_0 - \theta_1 a_h$ and the multistep ahead forecasts are $c_0$, which is the unconditional mean of the model. If we plot the forecasts $\hat{r}_h(\ell)$ versus $\ell$, we see that the forecasts form a horizontal line after one step.

Similarly, for an MA(2) model, we have

$$r_{h+\ell} = c_0 + a_{h+\ell} - \theta_1 a_{h+\ell-1} - \theta_2 a_{h+\ell-2},$$

from which we obtain

$$\hat{r}_h(1) = c_0 - \theta_1 a_h - \theta_2 a_{h-1}$$
$$\hat{r}_h(2) = c_0 - \theta_2 a_h$$
$$\hat{r}_h(\ell) = c_0, \quad \text{for} \quad \ell > 2.$$

Thus, the multistep ahead forecasts of an MA(2) model go to the mean of the series after two steps. The variances of forecast errors go to the variance of the series after two steps. In general, for an MA($q$) model, multistep ahead forecasts go to the mean after the first $q$ steps.

Table 2.3 gives some forecasts of the MA(9) model in Eq. (2.20) for the monthly simple returns of the equal-weighted index at the forecast origin $h = 854$. The sample mean and standard error of the first 854 observations of the series are 0.0131 and 0.0757, respectively. As expected, the table shows that (a) the 10-step ahead forecast is the sample mean, and (b) the standard deviations of the forecast errors converge to the standard deviation of the series as the forecast horizon increases.

**Table 2.3. Forecast Performance of an MA(9) Model for the Monthly Simple Returns of the CRSP Equal-Weighted Index. The Forecast Origin is $h = 854$. The Model Is Estimated by the Conditional Maximum Likelihood Method.**

| Step | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Fcst | .0026 | −.0016 | .0239 | .0133 | .0072 |
| S. Er | .0730 | .0741 | .0741 | .0747 | .0747 |
| Actu. | −.0479 | −.0213 | .0851 | .0442 | .0486 |
| Step | 6 | 7 | 8 | 9 | 10 |
| Fcst | .0163 | .0106 | .0186 | .0087 | .0131 |
| S. Er | .0747 | .0747 | .0747 | .0747 | .0754 |
| Actu. | .0271 | .0814 | −.0257 | −.0198 | −.0226 |

***Summary***

A brief summary of AR and MA models is in order. We have discussed the following properties:

- for MA models, ACF is useful in specifying the order because ACF cuts off at lag $q$ for an MA($q$) series;
- for AR models, PACF is useful in order determination because PACF cuts off at lag $p$ for an AR($p$) process;
- an MA series is always stationary, but for an AR series to be stationary, all of its characteristic roots must be less than 1 in modulus;
- for a stationary series, the multistep ahead forecasts converge to the mean of the series and the variances of forecast errors converge to the variance of the series.

## 2.6 SIMPLE ARMA MODELS

In some applications, the AR or MA models discussed in the previous sections become cumbersome because one may need a high-order model with many parameters to adequately describe the dynamic structure of the data. To overcome this difficulty, the autoregressive moving-average (ARMA) models are introduced; see Box, Jenkins, and Reinsel (1994). Basically, an ARMA model combines the ideas of AR and MA models into a compact form so that the number of parameters used is kept small. For the return series in finance, the chance of using ARMA models is low. However, the concept of ARMA models is highly relevant in volatility modeling. As a matter of fact, the generalized autoregressive conditional heteroscedastic (GARCH) model can be regarded as an ARMA model, albeit nonstandard, for the $a_t^2$ series; see Chapter 3 for details. In this section, we study the simplest ARMA(1, 1) model.

A time series $r_t$ follows an ARMA(1, 1) model if it satisfies

$$r_t - \phi_1 r_{t-1} = \phi_0 + a_t - \theta_1 a_{t-1}, \tag{2.22}$$

where $\{a_t\}$ is a white noise series. The left-hand side of Eq. (2.22) is the AR component of the model and the right-hand side gives the MA component. The constant term is $\phi_0$. For this model to be meaningful, we need $\phi_1 \neq \theta_1$; otherwise, there is a cancellation in the equation and the process reduces to a white noise series.

### 2.6.1 Properties of ARMA(1, 1) Models

Properties of ARMA(1, 1) models are generalizations of those of AR(1) models with some minor modifications to handle the impact of the MA(1) component. We start with the stationarity condition. Taking the expectation of Eq. (2.22), we have

$$E(r_t) - \phi_1 E(r_{t-1}) = \phi_0 + E(a_t) - \theta_1 E(a_{t-1}).$$

Because $E(a_i) = 0$ for all $i$, the mean of $r_t$ is

$$E(r_t) = \mu = \frac{\phi_0}{1 - \phi_1}$$

provided that the series is weakly stationary. This result is exactly the same as that of the AR(1) model in Eq. (2.6).

Next, assuming for simplicity that $\phi_0 = 0$, we consider the autocovariance function of $r_t$. First, multiplying the model by $a_t$ and taking expectation, we have

$$E(r_t a_t) = E(a_t^2) - \theta_1 E(a_t a_{t-1}) = E(a_t^2) = \sigma_a^2. \tag{2.23}$$

Rewriting the model as

$$r_t = \phi_1 r_{t-1} + a_t - \theta_1 a_{t-1}$$

and taking the variance of the prior equation, we have

$$\text{Var}(r_t) = \phi_1^2 \text{Var}(r_{t-1}) + \sigma_a^2 + \theta_1^2 \sigma_a^2 - 2\phi_1 \theta_1 E(r_{t-1} a_{t-1}).$$

Here we make use of the fact that $r_{t-1}$ and $a_t$ are uncorrelated. Using Eq. (2.23), we obtain

$$\text{Var}(r_t) - \phi_1^2 \text{Var}(r_{t-1}) = (1 - 2\phi_1 \theta_1 + \theta_1^2)\sigma_a^2.$$

Therefore, if the series $r_t$ is weakly stationary, then $\text{Var}(r_t) = \text{Var}(r_{t-1})$ and we have

$$\text{Var}(r_t) = \frac{(1 - 2\phi_1 \theta_1 + \theta_1^2)\sigma_a^2}{1 - \phi_1^2}.$$

Because the variance is positive, we need $\phi_1^2 < 1$ (i.e., $|\phi_1| < 1$). Again, this is precisely the same stationarity condition as that of the AR(1) model.

To obtain the autocovariance function of $r_t$, we assume $\phi_0 = 0$ and multiply the model in Eq. (2.22) by $r_{t-\ell}$ to obtain

$$r_t r_{t-\ell} - \phi_1 r_{t-1} r_{t-\ell} = a_t r_{t-\ell} - \theta_1 a_{t-1} r_{t-\ell}.$$

For $\ell = 1$, taking expectation and using Eq. (2.23) for $t - 1$, we have

$$\gamma_1 - \phi_1 \gamma_0 = -\theta_1 \sigma_a^2,$$

where $\gamma_\ell = \text{Cov}(r_t, r_{t-\ell})$. This result is different from that of the AR(1) case for which $\gamma_1 - \phi_1 \gamma_0 = 0$. However, for $\ell = 2$ and taking expectation, we have

$$\gamma_2 - \phi_1 \gamma_1 = 0,$$

which is identical to that of the AR(1) case. In fact, the same technique yields

$$\gamma_\ell - \phi_1 \gamma_{\ell-1} = 0, \quad \text{for} \quad \ell > 1. \tag{2.24}$$

In terms of ACF, the previous results show that for a stationary ARMA(1, 1) model

$$\rho_1 = \phi_1 - \frac{\theta_1 \sigma_a^2}{\gamma_0}, \quad \rho_\ell = \phi_1 \rho_{\ell-1}, \quad \text{for} \quad \ell > 1.$$

Thus, the ACF of an ARMA(1, 1) model behaves very much like that of an AR(1) model except that the exponential decay starts with lag 2. Consequently, the ACF of an ARMA(1, 1) model does not cut off at any finite lag.

Turning to PACF, one can show that the PACF of an ARMA(1, 1) model does not cut off at any finite lag either. It behaves very much like that of an MA(1) model except that the exponential decay starts with lag 2 instead of lag 1.

In summary, the stationarity condition of an ARMA(1, 1) model is the same as that of an AR(1) model, and the ACF of an ARMA(1, 1) exhibits a similar pattern like that of an AR(1) model except that the pattern starts at lag 2.

### 2.6.2 General ARMA Models

A general ARMA($p, q$) model is in the form

$$r_t = \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i} + a_t - \sum_{i=1}^{q} \theta_i a_{t-i},$$

where $\{a_t\}$ is a white noise series and $p$ and $q$ are non-negative integers. The AR and MA models are special cases of the ARMA($p, q$) model. Using the back-shift

operator, the model can be written as

$$(1 - \phi_1 B - \cdots - \phi_p B^p) r_t = \phi_0 + (1 - \theta_1 B - \cdots - \theta_q B^q) a_t. \tag{2.25}$$

The polynomial $1 - \phi_1 B - \cdots - \phi_p B^p$ is the AR polynomial of the model. Similarly, $1 - \theta_1 B - \cdots - \theta_q B^q$ is the MA polynomial. We require that there are no common factors between the AR and MA polynomials; otherwise the order $(p, q)$ of the model can be reduced. Like a pure AR model, the AR polynomial introduces the characteristic equation of an ARMA model. If all of the solutions of the characteristic equation are less than 1 in absolute value, then the ARMA model is weakly stationary. In this case, the unconditional mean of the model is $E(r_t) = \phi_0/(1 - \phi_1 - \cdots - \phi_p)$.

### 2.6.3 Identifying ARMA Models

The ACF and PACF are not informative in determining the order of an ARMA model. Tsay and Tiao (1984) propose a new approach that uses the extended autocorrelation function (EACF) to specify the order of an ARMA process. The basic idea of EACF is relatively simple. If we can obtain a consistent estimate of the AR component of an ARMA model, then we can derive the MA component. From the derived MA series, we can use ACF to identify the order of the MA component.

The derivation of EACF is relatively involved; see Tsay and Tiao (1984) for details. Yet the function is easy to use. The output of EACF is a two-way table, where the rows correspond to AR order $p$ and the columns to MA order $q$. The theoretical version of EACF for an ARMA(1, 1) model is given in Table 2.4. The key feature of the table is that it contains a triangle of "O" with the upper left vertex located at the order (1, 1). This is the characteristic we use to identify the order of an ARMA process. In general, for an ARMA$(p, q)$ model, the triangle of "O" will have its upper left vertex at the $(p, q)$ position.

For illustration, consider the monthly log stock returns of the 3M Company from February 1946 to December 1997. There are 623 observations. The return series and its sample ACF are shown in Figure 2.7. The ACF indicates that there are no

**Table 2.4. The Theoretical EACF Table for an ARMA(1, 1) Model, Where "X" Denotes Nonzero, "O" Denotes Zero, and "*" Denotes Either Zero or Nonzero. This Latter Category Does Not Play Any Role in Identifying the Order (1, 1).**

| | MA | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| AR | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 0 | X | X | X | X | X | X | X | X |
| 1 | X | O | O | O | O | O | O | O |
| 2 | * | X | O | O | O | O | O | O |
| 3 | * | * | X | O | O | O | O | O |
| 4 | * | * | * | X | O | O | O | O |
| 5 | * | * | * | * | X | O | O | O |

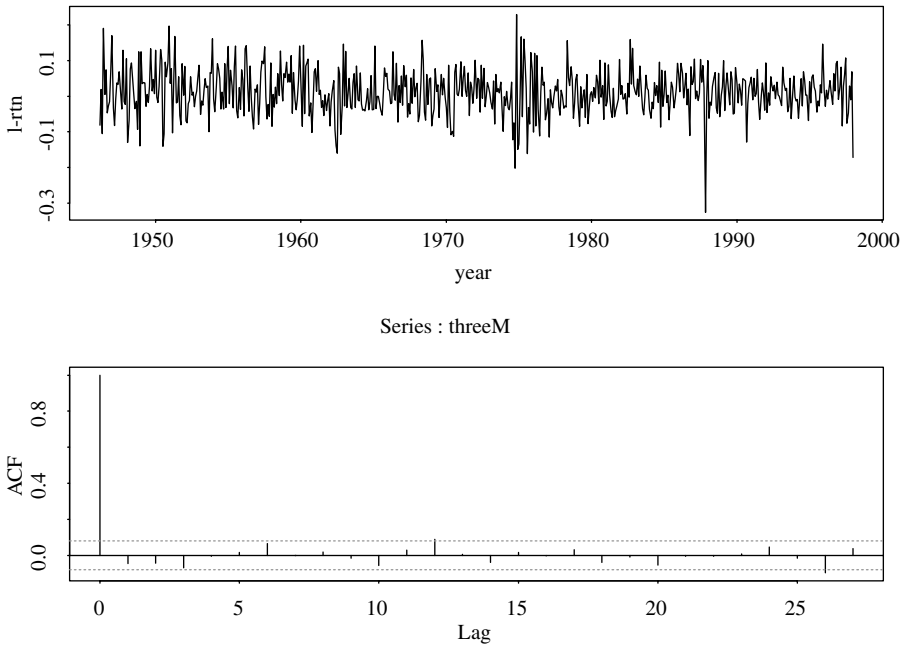


**Figure 2.7.** Time plot and sample autocorrelation function of the monthly log stock returns of 3M Company from February 1946 to December 1997.

significant serial correlations in the data at the 5% level. Table 2.5 shows the sample EACF and a corresponding simplified table for the series. The simplified table is constructed by using the following notation:

1. "X" denotes that the absolute value of the corresponding EACF is greater than or equal to $2/\sqrt{T}$, which is twice of the asymptotic standard error of the EACF;
2. "O" denotes that the corresponding EACF is less than $2/\sqrt{T}$ in modulus.

The simplified table clearly exhibits a triangular pattern of "O" with its upper left vertex at the order $(p, q) = (0, 0)$. The only exception is a single "X" in the first row, which corresponds to a sample EACF of 0.09 that is only slightly greater than $2/\sqrt{623} = 0.08$. Therefore, the EACF suggests that the monthly log returns of 3M stock follow an ARMA(0, 0) model (i.e., a white noise series). This is in agreement with the result suggested by the sample ACF in Figure 2.7.

Once an ARMA$(p, q)$ model is specified, its parameters can be estimated by either the conditional or exact likelihood method. In addition, the Ljung–Box statistics of the residuals can be used to check the adequacy of a fitted model. If the model is correctly specified, then $Q(m)$ follows asymptotically a chi-squared distribution with $m - g$ degrees of freedom, where $g$ denotes the number of parameters used in the model.

**Table 2.5. Sample Extended Autocorrelation Function and a Simplified Table for the Monthly Log Returns of 3M Stock from February 1946 to December 1997.**

(a) Sample extended autocorrelation function

| | MA order: $q$ | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 0 | −.05 | −.04 | −.07 | −.01 | .02 | .06 | −.00 | .02 | −.01 | −.06 | .03 | .09 | .01 |
| 1 | −.49 | .01 | −.06 | −.03 | −.00 | .06 | .01 | .01 | −.01 | −.05 | .02 | .08 | .02 |
| 2 | −.45 | −.18 | −.05 | .01 | −.02 | .06 | .03 | .02 | −.01 | −.00 | .01 | .05 | .05 |
| 3 | −.18 | .15 | .40 | −.01 | −.01 | .05 | −.00 | .03 | −.03 | −.00 | .00 | .02 | .05 |
| 4 | .42 | .04 | .39 | −.08 | −.01 | .01 | −.01 | .04 | .02 | .02 | −.00 | .01 | .03 |
| 5 | −.13 | .24 | .41 | .07 | .23 | .01 | .01 | .05 | −.03 | .02 | −.01 | .00 | .04 |
| 6 | −.07 | −.37 | .06 | .31 | .20 | −.09 | .01 | .06 | −.03 | .02 | −.01 | .00 | .03 |

(b) Simplified EACF table

| | MA order: $q$ | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 0 | O | O | O | O | O | O | O | O | O | O | O | X | O |
| 1 | X | O | O | O | O | O | O | O | O | O | O | O | O |
| 2 | X | X | O | O | O | O | O | O | O | O | O | O | O |
| 3 | X | X | X | O | O | O | O | O | O | O | O | O | O |
| 4 | X | O | X | O | O | O | O | O | O | O | O | O | O |
| 5 | X | X | X | O | X | O | O | O | O | O | O | O | O |
| 6 | O | X | O | X | X | O | O | O | O | O | O | O | O |

### 2.6.4 Forecasting Using an ARMA Model

Like the behavior of ACF, forecasts of an ARMA$(p, q)$ model have similar characteristics as those of an AR$(p)$ model after adjusting for the impacts of the MA component on the lower horizon forecasts. Denote the forecast origin by $h$. The 1-step ahead forecast of $r_{h+1}$ can be easily obtained from the model as

$$\hat{r}_h(1) = E(r_{h+1} \mid r_h, r_{h-1}, \ldots) = \phi_0 + \sum_{i=1}^{p} \phi_i r_{h+1-i} - \sum_{i=1}^{q} \theta_i a_{h+1-i},$$

and the associated forecast error is $e_h(1) = r_{h+1} - \hat{r}_h(1) = a_{h+1}$. The variance of 1-step ahead forecast error is $\text{Var}[e_h(1)] = \sigma_a^2$. For the $\ell$-step ahead forecast, we have

$$\hat{r}_h(\ell) = E(r_{h+\ell} \mid r_h, r_{h-1}, \ldots) = \phi_0 + \sum_{i=1}^{p} \phi_i \hat{r}_h(\ell - i) r_{h+\ell-i} - \sum_{i=1}^{q} \theta_i a_h(\ell - i)$$

where it is understood that $\hat{r}_h(\ell - i) = r_{h+\ell-i}$ if $\ell - i \leq 0$ and $a_h(\ell - i) = 0$ if $\ell - i > 0$ and $a_h(\ell - i) = a_{h+\ell-i}$ if $\ell - i \leq 0$. Thus, the multistep ahead forecasts of an ARMA model can be computed recursively. The associated forecast error is

$$e_h(\ell) = r_{h+\ell} - \hat{r}_h(\ell),$$

which can be computed easily via a formula to be given in the next subsection.

### 2.6.5 Three Model Representations for an ARMA Model

In this subsection, we briefly discuss three model representations for a stationary ARMA$(p, q)$ model. The three representations serve three different purposes. Knowing these representations can lead to a better understanding of the model. The first representation is the ARMA$(p, q)$ model in Eq. (2.25). This representation is compact and useful in parameter estimation. It is also useful in computing recursively multistep ahead forecasts of $r_t$; see the discussion of the last subsection.

For the other two representations, we use long division of two polynomials. Given two polynomials $\phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i$ and $\theta(B) = 1 - \sum_{i=1}^{q} \theta_i B^i$, we can obtain, by long division, that

$$\frac{\theta(B)}{\phi(B)} = 1 + \psi_1 B + \psi_2 B^2 + \cdots \equiv \psi(B) \tag{2.26}$$

and

$$\frac{\phi(B)}{\theta(B)} = 1 - \pi_1 B - \pi_2 B^2 - \cdots \equiv \pi(B). \tag{2.27}$$

For instance, if $\phi(B) = 1 - \phi_1 B$ and $\theta(B) = 1 - \theta_1 B$, then

$$\psi(B) = \frac{1 - \theta_1 B}{1 - \phi_1 B} = 1 + (\phi_1 - \theta_1)B + \phi_1(\phi_1 - \theta_1)B^2 + \phi_1^2(\phi_1 - \theta_1)B^3 + \cdots$$

$$\pi(B) = \frac{1 - \phi_1 B}{1 - \theta_1 B} = 1 - (\phi_1 - \theta_1)B - \theta_1(\phi_1 - \theta_1)B^2 - \theta_1^2(\phi_1 - \theta_1)B^3 - \cdots.$$

From the definition, $\psi(B)\pi(B) = 1$. Making use of the fact that $Bc = c$ for any constant (because the value of a constant is time-invariant), we have

$$\frac{\phi_0}{\theta(1)} = \frac{\phi_0}{1 - \theta_1 - \cdots - \theta_q} \quad \text{and} \quad \frac{\phi_0}{\phi(1)} = \frac{\phi_0}{1 - \phi_1 - \cdots - \phi_p}.$$

***AR Representation***

Using the result of long division in Eq. (2.27), the ARMA$(p, q)$ model can be written as

$$r_t = \frac{\phi_0}{1 - \theta_1 - \cdots - \theta_q} + \pi_1 r_{t-1} + \pi_2 r_{t-2} + \pi_3 r_{t-3} + \cdots + a_t. \tag{2.28}$$

This representation shows the dependence of the current return $r_t$ on the past returns $r_{t-i}$, where $i > 0$. The coefficients $\{\pi_i\}$ are referred to as the $\pi$-*weights* of an ARMA model. To show that the contribution of the lagged value $r_{t-i}$ to $r_t$ is diminishing as $i$ increases, the $\pi_i$ coefficient should decay to zero as $i$ increases. An ARMA$(p, q)$ model that has this property is said to be invertible. For a pure AR model, $\theta(B) = 1$ so that $\pi(B) = \phi(B)$, which is a finite-degree polynomial. Thus, $\pi_i = 0$ for $i > p$, and the model is invertible. For other ARMA models, a sufficient condition for invertibility is that all the zeros of the polynomial $\theta(B)$ are greater than unity in modulus. For example, consider the MA(1) model $r_t = (1 - \theta_1 B)a_t$. The zero of the first order polynomial $1 - \theta_1 B$ is $B = 1/\theta_1$. Therefore, an MA(1) model is invertible if $|1/\theta_1| > 1$. This is equivalent to $|\theta_1| < 1$.

From the AR representation in Eq. (2.28), an invertible ARMA$(p, q)$ series $r_t$ is a linear combination of the current shock $a_t$ and a weighted average of the past values. The weights decay exponentially for more remote past values.

***MA Representation***

Again, using the result of long division in Eq. (2.26), an ARMA$(p, q)$ model can also be written as

$$r_t = \mu + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots = \mu + \psi(B)a_t, \tag{2.29}$$

where $\mu = E(r_t) = \phi_0/(1 - \phi_1 - \cdots - \phi_p)$. This representation shows explicitly the impact of the past shock $a_{t-i}$ $(i > 0)$ on the current return $r_t$. The coefficients $\{\psi_i\}$ are referred to as the *impulse response function* of the ARMA model. For a weakly stationary series, the $\psi_i$ coefficients decay exponentially as $i$ increases. This is understandable as the effect of shock $a_{t-i}$ on the return $r_t$ should diminish over time. Thus, for a stationary ARMA model, the shock $a_{t-i}$ does not have a permanent impact on the series. If $\phi_0 \neq 0$, then the MA representation has a constant term, which is the mean of $r_t$ [i.e., $\phi_0/(1 - \phi_1 - \cdots - \phi_p$].

The MA representation in Eq. (2.29) is also useful in computing the variance of a forecast error. At the forecast origin $h$, we have the shocks $a_h, a_{h-1}, \ldots$. Therefore, the $\ell$-step ahead point forecast is

$$\hat{r}_h(\ell) = \mu + \psi_\ell a_h + \psi_{\ell+1} a_{h-1} + \cdots, \tag{2.30}$$

and the associated forecast error is

$$e_h(\ell) = a_{h+\ell} + \psi_1 a_{h+\ell-1} + \cdots + \psi_{\ell-1} a_{h+1}.$$

Consequently, the variance of $\ell$-step ahead forecast error is

$$\text{Var}[e_h(\ell)] = (1 + \psi_1^2 + \cdots + \psi_{\ell-1}^2)\sigma_a^2, \tag{2.31}$$

which, as expected, is a nondecreasing function of the forecast horizon $\ell$.

Finally, the MA representation in Eq. (2.29) provides a simple proof of mean reversion of a stationary time series. The stationarity implies that $\psi_i$ approaches zero as $i \to \infty$. Therefore, by Eq. (2.30), we have $\hat{r}_h(\ell) \to \mu$ as $\ell \to \infty$. Because $\hat{r}_h(\ell)$ is the conditional expectation of $r_{h+\ell}$ at the forecast origin $h$, the result says that in the long-term the return series is expected to approach its mean, that is, the series is mean reverting. Furthermore, using the MA representation in Eq. (2.29), we have $\text{Var}(r_t) = (1 + \sum_{i=1}^{\infty} \psi_i^2)\sigma_a^2$. Consequently, by Eq. (2.31), we have $\text{Var}[e_h(\ell)] \to \text{Var}(r_t)$ as $\ell \to \infty$. The speed by which $\hat{r}_h(\ell)$ approaches $\mu$ determines the speed of mean reverting.

## 2.7 UNIT-ROOT NONSTATIONARITY

So far we have focused on return series that are stationary. In some studies, interest rates, foreign exchange rates, or the price series of an asset are of interest. These series tend to be nonstationary. For a price series, the nonstationarity is mainly due to the fact that there is no fixed level for the price. In the time series literature, such a nonstationary series is called unit-root nonstationary time series. The best known example of unit-root nonstationary time series is the random-walk model.

### 2.7.1 Random Walk

A time series $\{p_t\}$ is a random walk if it satisfies

$$p_t = p_{t-1} + a_t, \tag{2.32}$$

where $p_0$ is a real number denoting the starting value of the process and $\{a_t\}$ is a white noise series. If $p_t$ is the log price of a particular stock at date $t$, then $p_0$ could be the log price of the stock at its initial public offering (i.e., the logged IPO price). If $a_t$ has a symmetric distribution around zero, then conditional on $p_{t-1}$, $p_t$ has a 50–50 chance to go up or down, implying that $p_t$ would go up or down at random. If we treat the random-walk model as a special AR(1) model, then the coefficient of $p_{t-1}$ is unity, which does not satisfy the weak stationarity condition of an AR(1) model. A random-walk series is, therefore, not weakly stationary, and we call it a unit-root nonstationary time series.

The random-walk model has been widely considered as a statistical model for the movement of logged stock prices. Under such a model, the stock price is not predictable or mean reverting. To see this, the 1-step ahead forecast of model (2.32) at the forecast origin $h$ is

$$\hat{p}_h(1) = E(p_{h+1} \mid p_h, p_{h-1}, \ldots) = p_h,$$

which is the log price of the stock at the forecast origin. Such a forecast has no practical value. The 2-step ahead forecast is

$$\hat{p}_h(2) = E(p_{h+2} \mid p_h, p_{h-1}, \ldots) = E(p_{h+1} + a_{h+2} \mid p_h, p_{h-1}, \ldots)$$
$$= E(p_{h+1} \mid p_h, p_{h-1}, \ldots) = \hat{p}_h(1) = p_h,$$

which again is the log price at the forecast origin. In fact, for any forecast horizon $\ell > 0$, we have

$$\hat{p}_h(\ell) = p_h.$$

Thus, for all forecast horizons, point forecasts of a random-walk model are simply the value of the series at the forecast origin. Therefore, the process is not mean-reverting.

The MA representation of the random-walk model in Eq. (2.32) is

$$p_t = a_t + a_{t-1} + a_{t-2} + \cdots.$$

This representation has several important practical implications. First, the $\ell$-step ahead forecast error is

$$e_h(\ell) = a_{h+\ell} + \cdots + a_{h+1},$$

so that $\text{Var}[e_h(\ell)] = \ell\sigma_a^2$, which diverges to infinity as $\ell \to \infty$. The length of an interval forecast of $p_{h+\ell}$ will approach infinity as the forecast horizon increases. This result says that the usefulness of point forecast $\hat{p}_h(\ell)$ diminishes as $\ell$ increases, which again implies that the model is not predictable. Second, the unconditional variance of $p_t$ is unbounded because $\text{Var}[e_h(\ell)]$ approaches infinity as $\ell$ increases. Theoretically, this means that $p_t$ can assume any real value for a sufficiently large $t$. For the log price $p_t$ of an individual stock, this is plausible. Yet for market indexes, negative log price is very rare if it happens at all. In this sense, the adequacy of a random-walk model for market indexes is questionable. Third, from the representation, $\psi_i = 1$ for all $i$. Thus, the impact of any past shock $a_{t-i}$ on $p_t$ does not decay over time. Consequently, the series has a strong memory as it remembers all of the past shocks. In economics, the shocks are said to have a permanent effect on the series.

### 2.7.2 Random Walk with a Drift

As shown by empirical examples considered so far, the log return series of a market index tends to have a small and positive mean. This implies that the model for the log price is

$$p_t = \mu + p_{t-1} + a_t, \tag{2.33}$$

where $\mu = E(p_t - p_{t-1})$ and $\{a_t\}$ is a white noise series. The constant term $\mu$ of model (2.33) is very important in financial study. It represents the time-trend of the log price $p_t$ and is often referred to as the *drift* of the model. To see this, assume that

the initial log price is $p_0$. Then we have

$$\begin{aligned} p_1 &= \mu + p_0 + a_1 \\ p_2 &= \mu + p_1 + a_2 = 2\mu + p_0 + a_2 + a_1 \\ &\vdots = \vdots \\ p_t &= t\mu + p_0 + a_t + a_{t-1} + \cdots + a_1. \end{aligned}$$

The last equation shows that the log price consists of a time trend $t\mu$ and a pure random-walk process $\sum_{i=1}^{t} a_i$. Because $\text{Var}(\sum_{i=1}^{t} a_i) = t\sigma_a^2$, where $\sigma_a^2$ is the variance of $a_t$, the conditional standard deviation of $p_t$ is $\sqrt{t}\sigma_a$, which grows at a slower rate than the conditional expectation of $p_t$. Therefore, if we graph $p_t$ against the time index $t$, we have a time trend with slope $\mu$. A positive slope $\mu$ implies that the log price eventually goes to infinity. In contrast, a negative $\mu$ implies that the log price would converge to $-\infty$ as $t$ increases. Based on this discussion, it is not surprising to see that the log return series of the CRSP value- and equal-weighted indexes have a small, but statistically significant, positive mean.

To illustrate the effect of the drift parameter on the price series, we consider the monthly log stock returns of the 3M Company from February 1946 to December 1997. As shown by the sample EACF in Table 2.5, the series has no significant serial correlation. The series thus follows the simple model

$$r_t = 0.0115 + a_t, \quad \hat{\sigma}_a = 0.0639, \tag{2.34}$$

where 0.0115 is the sample mean of $r_t$ and has a standard error 0.0026. The mean of the monthly log returns of 3M stock is, therefore, significantly different from zero at the 1% level. We use the log return series to construct two log price series—namely,

$$p_t = \sum_{i=1}^{t} r_i \quad \text{and} \quad p_t^* = \sum_{i=1}^{t} a_i,$$

where $a_i$ is the mean-corrected log return in Eq. (2.34) (i.e., $a_t = r_t - 0.0115$). The $p_t$ is the log price of 3M stock, assuming that the initial price is zero (i.e., the log price of January 1946 was zero). The $p_t^*$ is the corresponding log price if the mean of log returns were zero. Figure 2.8 shows the time plots of $p_t$ and $p_t^*$ as well as a straight line $y_t = 0.0115 \times t$. From the plots, the importance of the constant 0.0115 in Eq. (2.34) is evident. In addition, as expected, the slope of the upward trend of $p_t$ is about 0.0115.

Finally, it is important to understand the meaning of a constant term in a time series model. First, for an MA($q$) model in Eq. (2.19), the constant term is simply the mean of the series. Second, for a stationary AR($p$) model in Eq. (2.7) or ARMA($p, q$) model in Eq. (2.25), the constant term is related to the mean via $\mu = \phi_0/(1 - \phi_1 - \cdots - \phi_p)$. Third, for a random walk with a drift, the constant term becomes the time slope. These different interpretations for the constant term

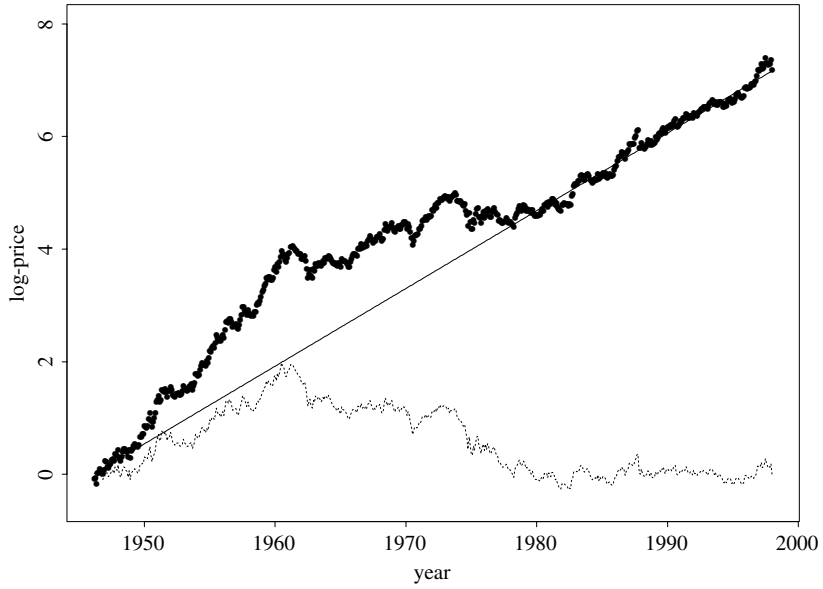


**Figure 2.8.** Time plots of log prices for 3M stock from February 1946 to December 1997, assuming that the log price of January 1946 was zero. The dashed line is for log price without time trend. The straight line is $y_t = 0.0115 \times t$.

in a time series model clearly highlights the difference between dynamic and usual linear regression models.

Another important difference between dynamic and regression models is shown by an AR(1) model and a simple linear regression model,

$$r_t = \phi_0 + \phi_1 r_{t-1} + a_t \quad \text{and} \quad y_t = \beta_0 + \beta_1 x_t + a_t.$$

For the AR(1) model to be meaningful, the coefficient $\phi_1$ must satisfy $|\phi_1| \leq 1$. However, the coefficient $\beta_1$ can assume any fixed real number.

### 2.7.3 General Unit-Root Nonstationary Models

Consider an ARMA model. If one extends the model by allowing the AR polynomial to have 1 as a characteristic root, then the model becomes the well-known autoregressive integrated moving-average (ARIMA) model. An ARIMA model is said to be unit-root nonstationary because its AR polynomial has a unit root. Like a random-walk model, an ARIMA model has strong memory because the $\psi_i$ coefficients in its MA representation do not decay over time to zero, implying that the past shock $a_{t-i}$ of the model has a permanent effect on the series. A conventional approach for handling unit-root nonstationarity is to use *differencing*.

***Differencing***

A time series $y_t$ is said to be an ARIMA$(p, 1, q)$ process if the change series $c_t = y_t - y_{t-1} = (1 - B)y_t$ follows a stationary and invertible ARMA$(p, q)$ model. In finance, price series are commonly believed to be nonstationary, but the log return series, $r_t = \ln(p_t) - \ln(p_{t-1})$, is stationary. In this case, the log price series is unit-root nonstationary and hence can be treated as an ARIMA process. The idea of transforming a nonstationary series into a stationary one by considering its change series is called *differencing* in the time series literature. More formally, $c_t = y_t - y_{t-1}$ is referred to as the first differenced series of $y_t$. In some scientific fields, a time series $y_t$ may contain multiple unit roots and needs to be differenced multiple times to become stationary. For example, if both $y_t$ and its first differenced series $c_t = y_t - y_{t-1}$ are unit-root nonstationary, but $s_t = c_t - c_{t-1} = y_t - 2y_{t-1} + y_{t-2}$ is weakly stationary, then $y_t$ has double unit roots, and $s_t$ is the second differenced series of $y_t$. In addition, if $s_t$ follows an ARMA$(p, q)$ model, then $y_t$ is an ARIMA$(p, 2, q)$ process. For such a time series, if $s_t$ has a nonzero mean, then $y_t$ has a quadratic time function and the quadratic time coefficient is related to the mean of $s_t$. The seasonally adjusted series of U.S. quarterly gross domestic product implicit price deflator might have double unit roots. However, the mean of the second differenced series is not significantly different from zero; see Exercises of the chapter. Box, Jenkins, and Reinsel (1994) discuss many properties of general ARIMA models.

### 2.7.4 Unit-Root Test

To test whether the log price $p_t$ of an asset follows a random walk or a random walk with a drift, we employ the models

$$p_t = \phi_1 p_{t-1} + e_t \tag{2.35}$$

$$p_t = \phi_0 + \phi_1 p_{t-1} + e_t, \tag{2.36}$$

where $e_t$ denotes the error term, and consider the null hypothesis $H_o : \phi_1 = 1$ versus the alternative hypothesis $H_a : \phi_1 < 1$. This is the well-known unit-root testing problem; see Dickey and Fuller (1979). A convenient test statistic is the $t$ ratio of the least squares (LS) estimate of $\phi_1$ under the null hypothesis. For Eq. (2.35), the LS method gives

$$\hat{\phi}_1 = \frac{\sum_{t=1}^{T} p_{t-1} p_t}{\sum_{t=1}^{T} p_{t-1}^2}, \quad \hat{\sigma}_e^2 = \frac{\sum_{t=1}^{T} (p_t - \hat{\phi}_1 p_{t-1})^2}{T - 1},$$

where $p_0 = 0$ and $T$ is the sample size. The $t$ ratio is

$$\text{DF} \equiv t\text{-ratio} = \frac{\hat{\phi}_1 - 1}{\text{std}(\hat{\phi}_1)} = \frac{\sum_{t=1}^{T} p_{t-1} e_t}{\hat{\sigma}_e \sqrt{\sum_{t=1}^{T} p_{t-1}^2}},$$

which is commonly referred to as the Dickey–Fuller test. If $\{e_t\}$ is a white noise series with finite moments of order slightly greater than 2, then the DF-statistic converges to a function of the standard Brownian motion as $T \to \infty$; see Chan and Wei (1988) and Phillips (1987) for more information. If $\phi_0$ is zero but Eq. (2.36) is employed anyway, then the resulting $t$ ratio for testing $\phi_1 = 1$ will converge to another non-standard asymptotic distribution. In either case, simulation is used to obtain critical values of the test statistics; see Fuller (1976, Chapter 8) for selected critical values. Yet if $\phi_0 \neq 0$ and Eq. (2.36) is used, then the $t$ ratio for testing $\phi_1 = 1$ is asymptotically normal. However, large sample sizes are needed for the asymptotic normal distribution to hold. Standard Brownian motion is introduced in Chapter 6.

## 2.8 SEASONAL MODELS

Some financial time series such as quarterly earning per share of a company exhibits certain cyclical or periodic behavior. Such a time series is called a *seasonal time series*. Figure 2.9(a) shows the time plot of quarterly earning per share of Johnson and Johnson from the first quarter of 1960 to the last quarter of 1980. The data obtained from Shumway and Stoffer (2000) possess some special characteristics. In

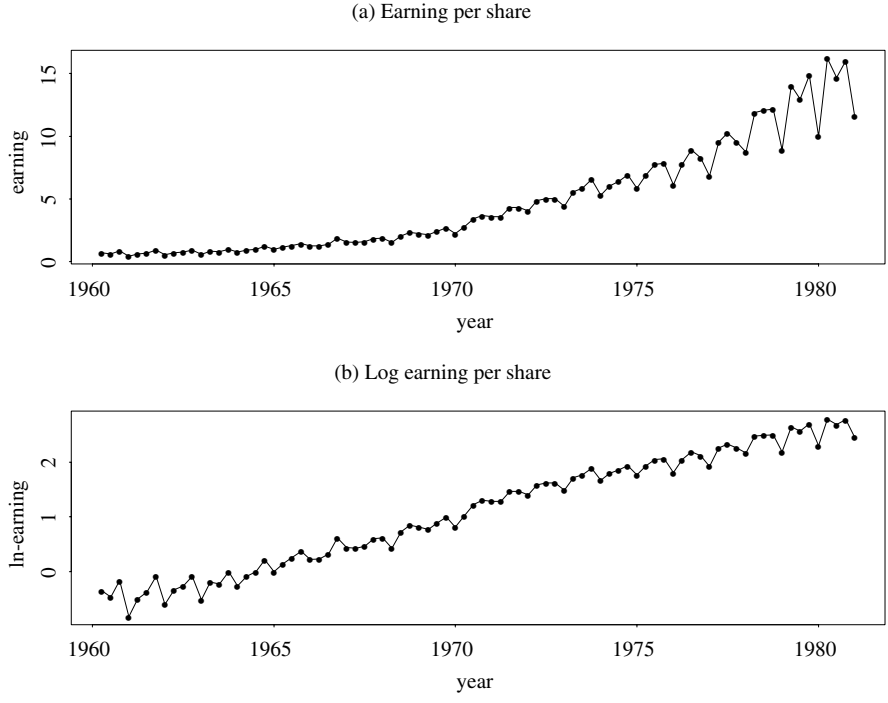


**Figure 2.9.** Time plots of quarterly earning per share of Johnson and Johnson from 1960 to 1980: (a) observed earning, (b) log earning.

particular, the earning grew exponentially during the sample period and had a strong seasonality. Furthermore, the variability of earning increased over time. The cyclical pattern repeats itself every year so that the periodicity of the series is 4. If monthly data are considered (e.g., monthly sales of Wal-Mart Stores), then the periodicity is 12. Seasonal time series models are also useful in pricing weather-related derivatives and energy futures.

Analysis of seasonal time series has a long history. In some applications, seasonality is of secondary importance and is removed from the data, resulting in a seasonally adjusted time series that is then used to make inference. The procedure to remove seasonality from a time series is referred to as *seasonal adjustment*. Most economic data published by the U.S. government are seasonally adjusted (e.g., the growth rate of domestic gross product and the unemployment rate). In other applications such as forecasting, seasonality is as important as other characteristics of the data and must be handled accordingly. Because forecasting is a major objective of financial time series analysis, we focus on the latter approach and discuss some econometric models that are useful in modeling seasonal time series.

### 2.8.1 Seasonal Differencing

Figure 2.9(b) shows the time plot of log earning per share of Johnson and Johnson. We took the log transformation for two reasons. First, it is used to handle the exponential growth of the series. Indeed, the new plot confirms that the growth is linear in the log scale. Second, the transformation is used to stablize the variability of the series. Again, the increasing pattern in variability of Figure 2.9(a) disappears in the new plot. Log transformation is commonly used in analysis of financial and economic time series. In this particular instance, all earnings are positive so that no adjustment is needed before taking the transformation. In some cases, one may need to add a positive constant to every data point before taking the transformation.

Denote the log earning by $x_t$. The upper left panel of Figure 2.10 shows the sample ACF of $x_t$, which indicates that the quarterly log earning per share has strong serial correlations. A conventional method to handle such strong serial correlations is to consider the first differenced series of $x_t$ [i.e., $\Delta x_t = x_t - x_{t-1} = (1 - B)x_t$]. The lower left plot of Figure 2.10 gives the sample ACF of $\Delta x_t$. The ACF is strong when the lag is a multiple of periodicity 4. This is a well-documented behavior of sample ACF of a seasonal time series. Following the procedure of Box, Jenkins, and Reinsel (1994, Chapter 9), we take another difference of the data—that is,

$$\Delta_4(\Delta x_t) = (1 - B^4)\Delta x_t = \Delta x_t - \Delta x_{t-4} = x_t - x_{t-1} - x_{t-4} + x_{t-5}.$$

The operation $\Delta_4 = (1 - B^4)$ is called a *seasonal differencing*. In general, for a seasonal time series $y_t$ with periodicity $s$, seasonal differencing means

$$\Delta_s y_t = y_t - y_{t-s} = (1 - B^s)y_t.$$

The conventional difference $\Delta y_t = y_t - y_{t-1} = (1-B)y_t$ is referred to as the *regular differencing*. The lower right plot of Figure 2.10 shows the sample ACF of $\Delta_4 \Delta x_t$,

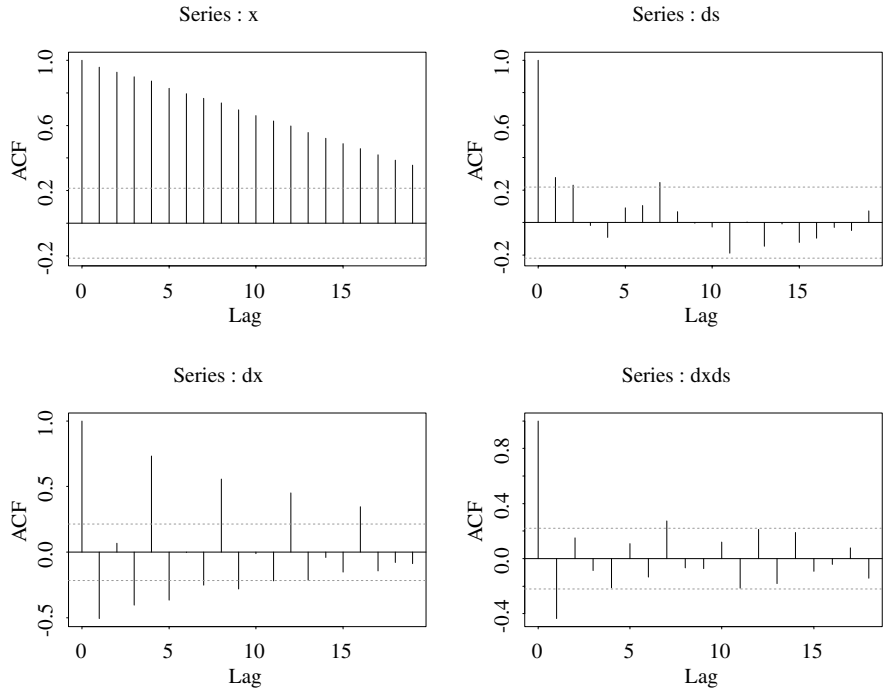


**Figure 2.10.** Sample ACF of the log series of quarterly earning per share of Johnson and Johnson from 1960 to 1980, where $x_t$ is the log earning, "dx" is the first differenced series, "ds" is the seasonally differenced series, and "dxds" denotes series with regular and seasonal differencing.

which has a significant negative ACF at lag 1 and a marginal negative correlation at lag 4. For completeness, Figure 2.10 also gives the sample ACF of the seasonally differenced series $\Delta_4 x_t$.

### 2.8.2 Multiplicative Seasonal Models

The behavior of the sample ACF of $(1 - B^4)(1 - B)x_t$ in Figure 2.10 is common among seasonal time series. It led to the development of the following special seasonal time series model

$$(1 - B^s)(1 - B)x_t = (1 - \theta B)(1 - \Theta B^s)a_t, \tag{2.37}$$

where $s$ is the periodicity of the series, $a_t$ is a white noise series, $|\theta| < 1$, and $|\Theta| < 1$. This model is referred to as the *airline model* in the literature; see Box, Jenkins, and Reinsel (1994, Chapter 9). It has been found to be widely applicable in modeling seasonal time series. The AR part of the model simply consists of the regular and seasonal differences, whereas the MA part involves two parameters. Focusing on the

MA part (i.e., on the model),

$$w_t = (1 - \theta B)(1 - \Theta B^s)a_t = a_t - \theta a_{t-1} - \Theta a_{t-s} + \theta\Theta a_{t-s-1},$$

where $w_t = (1 - B^s)(1 - B)x_t$ and $s > 1$. It is easy to obtain that $E(w_t) = 0$ and

$$\begin{aligned}
\text{Var}(w_t) &= (1 + \theta^2)(1 + \Theta^2)\sigma_a^2 \\
\text{Cov}(w_t, w_{t-1}) &= -\theta(1 + \Theta^2)\sigma_a^2 \\
\text{Cov}(w_t, w_{t-s+1}) &= \theta\Theta\sigma_a^2 \\
\text{Cov}(w_t, w_{t-s}) &= -\Theta(1 + \theta^2)\sigma_a^2 \\
\text{Cov}(w_t, w_{t-s-1}) &= \theta\Theta\sigma_a^2 \\
\text{Cov}(w_t, w_{t-\ell}) &= 0, \quad \text{for} \quad \ell \neq 0, 1, s-1, s, s+1.
\end{aligned}$$

Consequently, the ACF of the $w_t$ series is given by

$$\rho_1 = \frac{-\theta}{1 + \theta^2}, \quad \rho_s = \frac{-\Theta}{1 + \Theta^2}, \quad \rho_{s-1} = \rho_{s+1} = \rho_1\rho_s = \frac{\theta\Theta}{(1 + \theta^2)(1 + \Theta^2)},$$

and $\rho_\ell = 0$ for $\ell > 0$ and $\ell \neq 1, s-1, s, s+1$. For example, if $w_t$ is a quarterly time series, then $s = 4$ and the ACF is nonzero at lags 1, 3, 4, and 5 only.

It is interesting to compare the prior ACF with those of the MA(1) model $y_t = (1 - \theta B)a_t$ and the MA($s$) model $z_t = (1 - \Theta B^s)a_t$. The ACF of $y_t$ and $z_t$ series are

$$\begin{aligned}
\rho_1(y) &= \frac{-\theta}{1 + \theta^2}, \text{ and} \quad \rho_\ell(y) = 0, \quad \ell > 1, \\
\rho_s(z) &= \frac{-\Theta}{1 + \Theta^2}, \text{and} \quad \rho_\ell(z) = 0, \quad \ell > 0, \neq s.
\end{aligned}$$

We see that (i) $\rho_1 = \rho_1(y)$, (ii) $\rho_s = \rho_s(z)$, and (iii) $\rho_{s-1} = \rho_{s+1} = \rho_1(y) \times \rho_s(z)$. Therefore, the ACF of $w_t$ at lags $(s - 1)$ and $(s + 1)$ can be regarded as the *interaction* between lag-1 and lag-$s$ serial dependence, and the model of $w_t$ is called a *multiplicative* seasonal MA model. In practice, a multiplicative seasonal model says that the dynamics of the regular and seasonal components of the series are approximately orthogonal.

The model

$$w_t = (1 - \theta B - \Theta B^s)a_t, \tag{2.38}$$

where $|\theta| < 1$ and $|\Theta| < 1$, is a nonmultiplicative seasonal MA model. It is easy to see that for the model in Eq. (2.38), $\rho_{s+1} = 0$. A multiplicative model is more parsimonious than the corresponding nonmultiplicative model because both

models use the same number of parameters, but the multiplicative model has more nonzero ACFs.

**Example 2.2.** In this example, we apply the airline model to the log series of quarterly earning per share of Johnson and Johnson from 1960 to 1980. Based on the exact likelihood method, the fitted model is

$$(1 - B)(1 - B^4)x_t = (1 - 0.678B)(1 - 0.314B^4)a_t, \quad \hat{\sigma}_a = 0.089,$$

where standard errors of the two MA parameters are 0.080 and 0.101, respectively. The Ljung-Box statistics of the residuals show $Q(12) = 10.0$ with $p$ value 0.44. The model appears to be adequate.

To illustrate the forecasting performance of the prior seasonal model, we reestimate the model using the first 76 observations and reserve the last eight data points for forecasting evaluation. We compute 1-step to 8-step ahead forecasts and their standard errors of the fitted model at the forecast origin $h = 76$. An antilog transformation is taken to obtain forecasts of earning per share using the relationship between normal and log-normal distributions given in Chapter 1. Figure 2.11 shows the forecast performance of the model, where the observed data are in solid line,

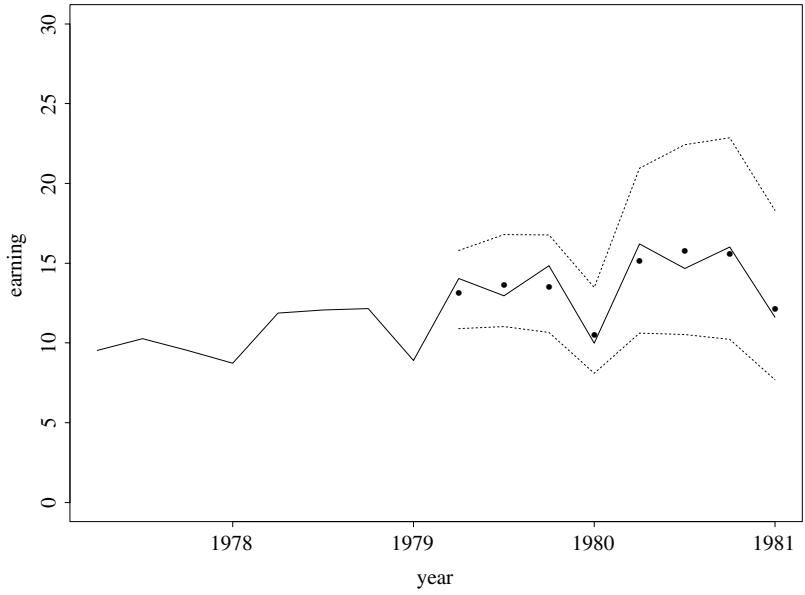


**Figure 2.11.** Out-of-sample point and interval forecasts for the quarterly earning of Johnson and Johnson. The forecast origin is the fourth quarter of 1978. In the plot, solid line shows the actual observations, dots represent point forecasts, and dashed lines show a 95% interval forecasts.

point forecasts are shown by dots, and the dashed lines show 95% interval forecasts. The forecasts show a strong seasonal pattern and are close to the observed data.

When the seasonal pattern of a time series is stable over time (e.g., close to a deterministic function), dummy variables may be used to handle the seasonality. This approach is taken by some analysts. However, deterministic seasonality is a special case of the multiplicative seasonal model discussed before. Specifically, if $\Theta = 1$, then model (2.37) contains a deterministic seasonal component. Consequently, the same forecasts are obtained by using either dummy variables or a multiplicative seasonal model when the seasonal pattern is deterministic. Yet use of dummy variables can lead to inferior forecasts if the seasonal pattern is not deterministic. In practice, we recommend that the exact likelihood method should be used to estimate a multiplicative seasonal model, especially when the sample size is small or when there is the possibility of having a deterministic seasonal component.

## 2.9 REGRESSION MODELS WITH TIME SERIES ERRORS

In many applications, the relationship between two time series is of major interest. The Market Model in finance is an example that relates the return of an individual stock to the return of a market index. The term structure of interest rates is another example in which the time evolution of the relationship between interest rates with different maturities is investigated. These examples lead to the consideration of a linear regression in the form

$$r_{1t} = \alpha + \beta r_{2t} + e_t, \tag{2.39}$$

where $r_{1t}$ and $r_{2t}$ are two time series and $e_t$ denotes the error term. The least squares (LS) method is often used to estimate model (2.39). If $\{e_t\}$ is a white noise series, then the LS method produces consistent estimates. In practice, however, it is common to see that the error term $e_t$ is serially correlated. In this case, we have a regression model with time series errors, and the LS estimates of $\alpha$ and $\beta$ may not be consistent.

Regression model with time series errors is widely applicable in economics and finance, but it is one of the most commonly misused econometric models because the serial dependence in $e_t$ is often overlooked. It pays to study the model carefully.

We introduce the model by considering the relationship between two U.S. weekly interest rate series:

1. $r_{1t}$: The 1-year Treasury constant maturity rate.
2. $r_{3t}$: The 3-year Treasury constant maturity rate.

Both series have 1967 observations from January 5, 1962 to September 10, 1999 and are measured in percentages. The series are obtained from the Federal Reserve Bank of St Louis. Figure 2.12 shows the time plots of the two interest rates with solid line denoting the 1-year rate and dashed line the 3-year rate. Figure 2.13(a) plots $r_{1t}$ versus $r_{3t}$, indicating that, as expected, the two interest rates are highly correlated.

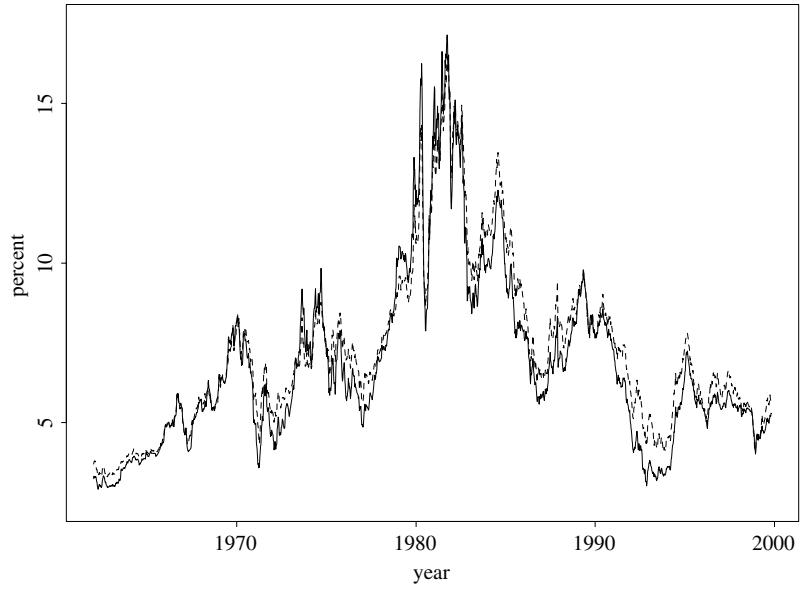


**Figure 2.12.** Time plots of U.S. weekly interest rates (in percentages) from January 5, 1962 to September 10, 1999. The solid line is the Treasury 1-year constant maturity rate and the dashed line the Treasury 3-year constant maturity rate.

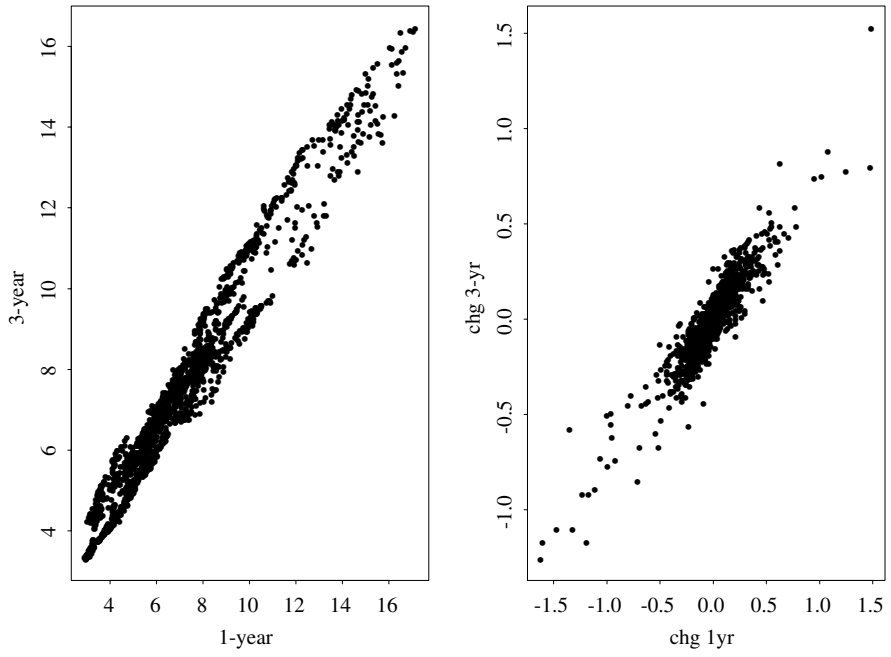


**Figure 2.13.** Scatterplots of U.S. weekly interest rates from January 5, 1962 to September 10, 1999: (a) 3-year rate versus 1-year rate, and (b) changes in 3-year rate versus changes in 1-year rate.

A naive way to describe the relationship between the two interest rates is to use the simple model $r_{3t} = \alpha + \beta r_{1t} + e_t$. This results in a fitted model

$$r_{3t} = 0.911 + 0.924 r_{1t} + e_t, \quad \hat{\sigma}_e = 0.538 \tag{2.40}$$

with $R^2 = 95.8\%$, where the standard errors of the two coefficients are 0.032 and 0.004, respectively. Model (2.40) confirms the high correlation between the two interest rates. However, the model is seriously inadequate as shown by Figure 2.14, which gives the time plot and ACF of its residuals. In particular, the sample ACF of the residuals is highly significant and decays slowly, showing the pattern of a unit-root nonstationary time series. The behavior of the residuals suggests that marked differences exist between the two interest rates. Using the modern econometric terminology, if one assumes that the two interest rate series are unit-root nonstationary, then the behavior of the residuals of Eq. (2.40) indicates that the two interest rates are not *co-integrated*; see Chapter 8 for discussion of co-integration. In other words, the data fail to support the hypothesis that there exists a long-term equilibrium between the two interest rates. In some sense, this is not surprising because the pattern of "inverted yield curve" did occur during the data span. By inverted yield curve, we mean the situation under which interest rates are inversely related to their time to maturities.

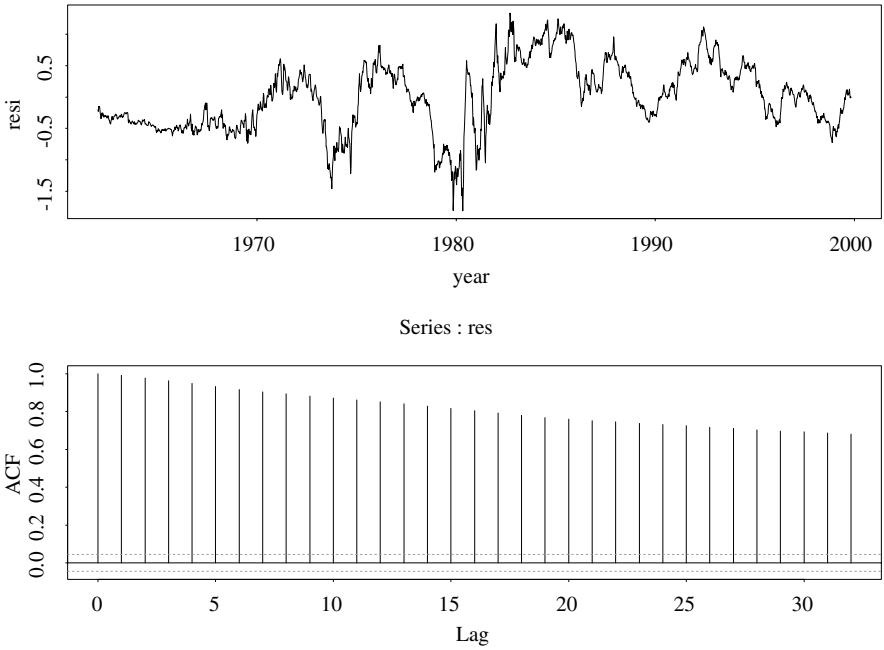


**Figure 2.14.** Residual series of linear regression (2.40) for two U.S. weekly interest rates: (a) time plot, and (b) sample ACF.

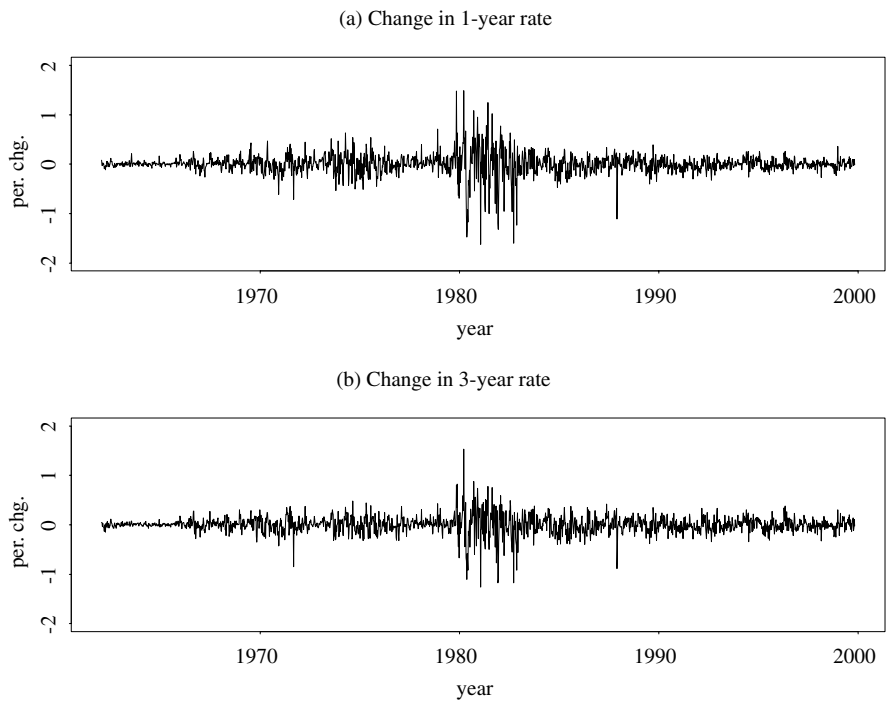


**Figure 2.15.** Time plots of the change series of U.S. weekly interest rates from January 12, 1962 to September 10, 1999: (a) changes in the Treasury 1-year constant maturity rate, and (b) changes in the Treasury 3-year constant maturity rate.

The unit-root behavior of both interest rates and the residuals of Eq. (2.40) leads to the consideration of the change series of interest rates. Let

**1.** $c_{1t} = r_{1t} - r_{1,t-1} = (1 - B)r_{1t}$ for $t \geq 2$: Changes in the 1-year interest rate;
**2.** $c_{3t} = r_{3t} - r_{3,t-1} = (1 - B)r_{3t}$ for $t \geq 2$: Changes in the 3-year interest rate,

and consider the linear regression $c_{3t} = \alpha + \beta c_{1t} + e_t$. Figure 2.15 shows time plots of the two change series, whereas Figure 2.13(b) provides a scatterplot between them. The change series remain highly correlated with a fitted linear regression model given by

$$c_{3t} = 0.0002 + 0.7811c_{1t} + e_t, \quad \hat{\sigma}_e = 0.0682, \tag{2.41}$$

with $R^2 = 84.8\%$. The standard errors of the two coefficients are 0.0015 and 0.0075, respectively. This model further confirms the strong linear dependence between interest rates. Figure 2.16 shows the time plot and sample ACF of the residuals of Eq. (2.41). Once again, the ACF shows some significant serial correlation in the residuals, but the magnitude of the correlation is much smaller. This weak serial

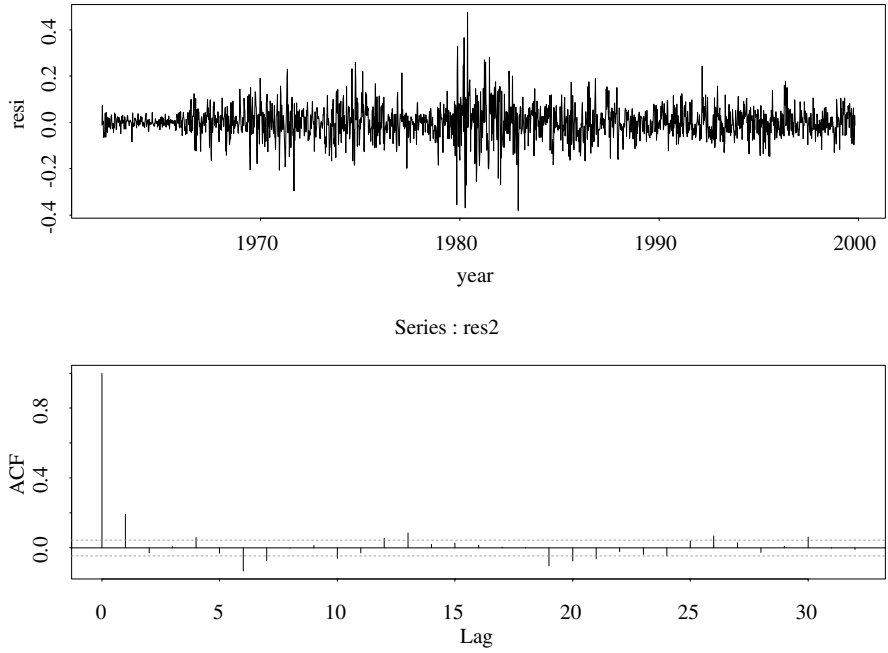


**Figure 2.16.** Residual series of the linear regression (2.41) for two change series of U.S. weekly interest rates: (a) time plot, and (b) sample ACF.

dependence in the residuals can be modeled by using the simple time series models discussed in the previous sections, and we have a linear regression with time series errors.

The main objective of this section is to discuss a simple approach for building a linear regression model with time series errors. The approach is straightforward. We employ a simple time series model discussed in this chapter for the residual series and estimate the whole model jointly. For illustration, consider the simple linear regression in Eq. (2.41). Because residuals of the model are serially correlated, we identify a simple ARMA model for the residuals. From the sample ACF of the residuals shown in Figure 2.16, we specify an MA(1) model for the residuals and modify the linear regression model to

$$c_{3t} = \alpha + \beta c_{1t} + e_t, \quad e_t = a_t - \theta_1 a_{t-1}, \tag{2.42}$$

where $\{a_t\}$ is assumed to be a white noise series. In other words, we simply use an MA(1) model, without the constant term, to capture the serial dependence in the error term of Eq. (2.41). The resulting model is a simple example of linear regression with time series errors. In practice, more elaborated time series models can be added to a linear regression equation to form a general regression model with time series errors.

Estimating a regression model with time series errors was not easy before the advent of modern computers. Special methods such as the Cochrane–Orcutt estimator have been proposed to handle the serial dependence in the residuals; see Greene (2000, p. 546). By now, the estimation is as easy as that of other time series models. If the time series model used is stationary and invertible, then one can estimate the model jointly via the maximum likelihood method. This is the approach we take by using the SCA package. For the U.S. weekly interest rate data, the fitted version of model (2.42) is

$$c_{3t} = 0.0002 + 0.7824c_{1t} + e_t, \quad e_t = a_t + 0.2115a_{t-1}, \quad \hat{\sigma}_a = 0.0668, \quad (2.43)$$

with $R^2 = 85.4\%$. The standard errors of the parameters are 0.0018, 0.0077, and 0.0221, respectively. The model no longer has a significant lag-1 residual ACF, even though some minor residual serial correlations remain at lags 4 and 6. The incremental improvement of adding additional MA parameters at lags 4 and 6 to the residual equation is small and the result is not reported here.

Comparing the models in Eqs. (2.40), (2.41), and (2.43), we make the following observations. First, the high $R^2$ and coefficient 0.924 of model (2.40) are misleading because the residuals of the model show strong serial correlations. Second, for the change series, $R^2$ and the coefficient of $c_{1t}$ of models (2.41) and (2.43) are close. In this particular instance, adding the MA(1) model to the change series only provides a marginal improvement. This is not surprising because the estimated MA coefficient is small numerically, even though it is statistically highly significant. Third, the analysis demonstrates that it is important to check residual serial dependence in linear regression analysis.

Because the constant term of Eq. (2.43) is insignificant, the model shows that the two weekly interest rate series are related as

$$r_{3t} = r_{3,t-1} + 0.782(r_{1t} - r_{1,t-1}) + a_t + 0.212a_{t-1}.$$

The interest rates are concurrently and serially correlated.

***Summary***

We outline a general procedure for analyzing linear regression models with time series errors:

1. Fit the linear regression model and check serial correlations of the residuals.
2. If the residual series is unit-root nonstationary, take the first difference of both the dependent and explanatory variables. Go to step 1. If the residual series appears to be stationary, identify an ARMA model for the residuals and modify the linear regression model accordingly.
3. Perform a joint estimation via the maximum likelihood method and check the fitted model for further improvement.

To check the serial correlations of residuals, we recommend that the Ljung–Box statistics be used instead of the Durbin–Watson (DW) statistic because the latter only considers the lag-1 serial correlation. There are cases in which residual serial dependence appears at higher order lags. This is particularly so when the time series involved exhibits some seasonal behavior.

***Remark:*** For a residual series $e_t$ with $T$ observations, the Durbin–Watson statistic is

$$\text{DW} = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}.$$

Straightforward calculation shows that $\text{DW} \approx 2(1 - \hat{\rho}_1)$, where $\hat{\rho}_1$ is the lag-1 ACF of $\{e_t\}$.

## 2.10 LONG-MEMORY MODELS

We have discussed that for a stationary time series the ACF decays exponentially to zero as lag increases. Yet for a unit-root nonstationary time series, it can be shown that the sample ACF converges to 1 for all fixed lags as the sample size increases; see Chan and Wei (1988) and Tiao and Tsay (1983). There exist some time series whose ACF decays slowly to zero at a polynomial rate as the lag increases. These processes are referred to as long-memory time series. One such an example is the fractionally differenced process defined by

$$(1 - B)^d x_t = a_t, \quad -0.5 < d < 0.5, \tag{2.44}$$

where $\{a_t\}$ is a white noise series. Properties of model (2.44) have been widely studied in the literature (e.g., Hosking, 1981). We summarize some of these properties below.

1. If $d < 0.5$, then $x_t$ is a weakly stationary process and has the infinite MA representation

$$x_t = a_t + \sum_{i=1}^{\infty} \psi_i a_{t-i}, \quad \text{with} \quad \psi_k = \frac{d(1+d)\cdots(k-1+d)}{k!}$$
$$= \frac{(k+d-1)!}{k!(d-1)!}.$$

2. If $d > -0.5$, then $x_t$ is invertible and has the infinite AR representation

$$x_t = \sum_{i=1}^{\infty} \pi_i x_{t-i} + a_t, \quad \text{with} \quad \pi_k = \frac{-d(1-d)\cdots(k-1-d)}{k!}$$

$$= \frac{(k-d-1)!}{k!(-d-1)!}.$$

**3.** For $-0.5 < d < 0.5$, the ACF of $x_t$ is

$$\rho_k = \frac{d(1+d)\cdots(k-1+d)}{(1-d)(2-d)\cdots(k-d)}, \quad k = 1, 2, \ldots.$$

In particular, $\rho_1 = d/(1-d)$ and

$$\rho_k \approx \frac{(-d)!}{(d-1)!} k^{2d-1}, \quad \text{as} \quad k \to \infty.$$

**4.** For $-0.5 < d < 0.5$, the PACF of $x_t$ is $\phi_{k,k} = d/(k-d)$ for $k = 1, 2, \ldots$.

**5.** For $-0.5 < d < 0.5$, the spectral density function $f(\omega)$ of $x_t$, which is the Fourier transform of the ACF of $x_t$, satisfies

$$f(\omega) \sim \omega^{-2d}, \quad \text{as} \quad \omega \to 0, \tag{2.45}$$

where $\omega \in [0, 2\pi]$ denotes the frequency.

Of particular interest here is the behavior of ACF of $x_t$ when $d < 0.5$. The property says that $\rho_k \sim k^{2d-1}$, which decays at a polynomial, instead of exponential, rate. For this reason, such an $x_t$ process is called a long-memory time series. A special characteristic of the spectral density function in Eq. (2.45) is that the spectrum diverges to infinity as $\omega \to 0$. However, the spectral density function of a stationary ARMA process is bounded for all $\omega \in [0, 2\pi]$.

Earlier we used the binomial theorem for noninteger powers

$$(1-B)^d = \sum_{k=0}^{\infty} (-1)^k \binom{d}{k} B^k, \quad \binom{d}{k} = \frac{d(d-1)\cdots(d-k+1)}{k!}.$$

If the fractionally differenced series $(1-B)^d x_t$ follows an ARMA$(p, q)$ model, then $x_t$ is called an ARFIMA$(p, d, q)$ process, which is a generalized ARIMA model by allowing for noninteger $d$.

In practice, if the sample ACF of a time series is not large in magnitude, but decays slowly, then the series may have long memory. As an illustration, Figure 2.17 shows the sample ACFs of the absolute series of daily simple returns for the CRSP value- and equal-weighted indexes from July 3, 1962 to December 31, 1997. The ACFs are relatively small in magnitude, but decay very slowly; they appear to be significant at the 5% level even after 300 lags. For more information about the behavior of sample ACF of absolute return series, see Ding, Granger, and Engle (1993). For the pure fractionally differenced model in Eq. (2.44), one can estimate $d$ using either a maximum likelihood method or a regression method with logged periodogram at the

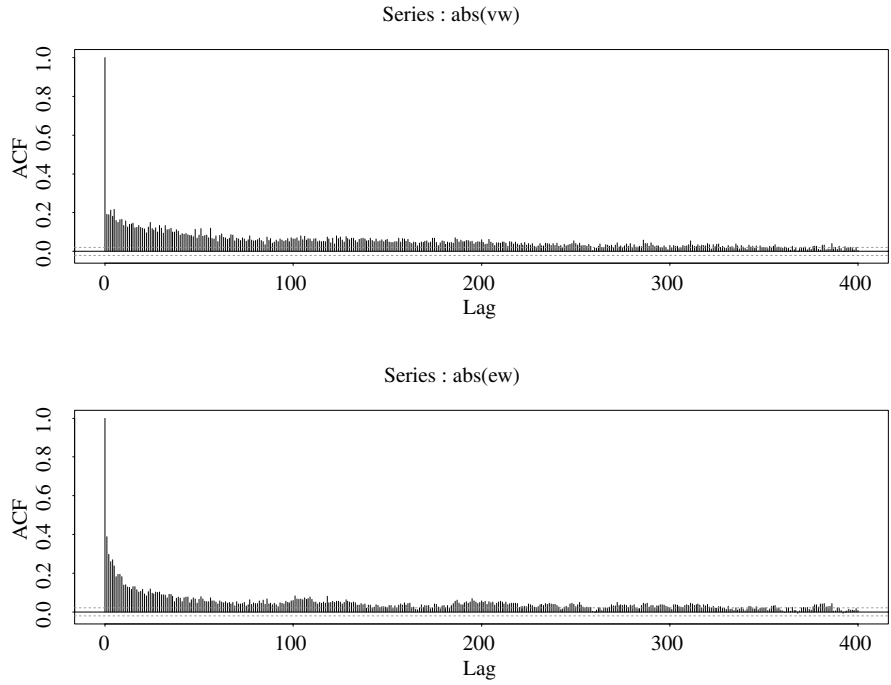


**Figure 2.17.** Sample autocorrelation function of the absolute series of daily simple returns for the CRSP value- and equal-weighted indexes: (a) the value-weighted index return, and (b) the equal-weighted index return.

lower frequencies. Finally, long-memory models have attracted some attention in the finance literature in part because of the work on fractional Brownian motion in the continuous-time models.

## APPENDIX A: SOME SCA COMMANDS

### A. Commands Used in Section 2.4

The data file is m-vw.dat and comments start with "–". These comments explain the function of each command.

```
-- load data into SCA and denote the series by vw.
input vw. file 'm-vw.dat'
-- compute 10 lags of PACF.
pacf vw. maxl 10.
-- compute AIC for AR(1) to AR(10).
miden vw. no ccm. arfits 1 to 10.
-- specify an AR(3) model and denote the model by m1.
tsm m1. model (1,2,3)vw=c0+noise.
-- estimate the model and store the residuals in r1.
```

```
estim m1. hold resi(r1)
-- compute ACF of the residuals, including Q statistics.
acf r1.
-- refine the model to an AR(5).
tsm m1. model (1,2,3,4,5)vw=c0+noise.
-- estimate the model and store the residuals in r1.
estim m1. hold resi(r1)
-- compute ACF of the residuals.
acf r1. maxl 10.
-- compute p-value of the Q(5) statistic.
p=1.0-cdfc(11.2,5)
-- print p-value.
print p
-- re-estimate the model using the first 858 observations.
estim m1. span 1,858.
-- compute 1-step to 6-step ahead forecasts at origin 858.
ufore m1. orig 858. nofs 6.
-- quit SCA.
stop
```

## B. Commands used in Section 2.9

The 1-year maturity interest rates are in the file "wgs1yr.dat" and the 3-year rates are in the file "wgs3yr.dat."

```
-- load the data into SCA, denote the data by rate1 and rate3.
input date, rate1. file 'wgs1yr.dat'
--
input date,rate3. file 'wgs3yr.dat'
-- specify a simple linear regression model.
tsm m1. model rate3=b0+(b1)rate1+noise.
-- estimate the specified model and store residual in r1.
estim m1. hold resi(r1).
-- compute 10 lags of residual acf.
acf r1. maxl 10.
-- difference the two series, denote the new series by c1t and c3t
diff old rate1,rate3. new c1t, c3t. compress.
-- specify a linear regression model for the differenced data
tsm m2. model c3t=h0+(h1)c1t+noise.
-- estimation
estim m2. hold resi(r2).
-- compute residual acf.
acf r2. maxl 10.
-- specify a regression model with time series errors.
tsm m3. model c3t=g0+(g1)c1t+(1)noise.
-- estimate the model using the exact likelihood method.
estim m3. method exact. hold resi(r3).
-- compute residual acf.
acf r3. maxl 10.
-- refine the model to include more MA lags.
tsm m4. model c3t=g0+(g1)c1t+(1,4,6)noise.
-- estimation
estim m4. method exact. hold resi(r4).
```

```
-- compute residual acf.
acf r4. maxl 10.
-- exit SCA
stop
```

## EXERCISES

**1.** Suppose that the simple return of a monthly bond index follows the MA(1) model

$$R_t = a_t + 0.2a_{t-1}, \quad \sigma_a = 0.025.$$

Assume that $a_{100} = 0.01$. Compute the 1-step and 2-step ahead forecasts of the return at the forecast origin $t = 100$. What are the standard deviations of the associated forecast errors? Also compute the lag-1 and lag-2 autocorrelations of the return series.

**2.** Suppose that the daily log return of a security follows the model

$$r_t = 0.01 + 0.2r_{t-2} + a_t,$$

where $\{a_t\}$ is a Gaussian white noise series with mean zero and variance 0.02. What are the mean and variance of the return series $r_t$? Compute the lag-1 and lag-2 autocorrelations of $r_t$. Assume that $r_{100} = -0.01$, and $r_{99} = 0.02$. Compute the 1- and 2-step ahead forecasts of the return series at the forecast origin $t = 100$. What are the associated standard deviations of the forecast errors?

**3.** The file "bnd.dat" contains simple returns of monthly indexes of U.S. government bonds with maturities in 30 years, 20 years, 10 years, 5 years, and 1 year (in column order). The data are obtained from CRSP, and the sample period is from January 1942 to December 1999. Build an AR or MA model for the simple return of bond index with maturity 5 years. Is the fitted model adequate?

**4.** Consider the sampling period from January 1990 to December 1999. Are the daily log returns of ALCOA stock predictable? You may test the hypothesis using (a) the first 5 lags of the autocorrelation function, and (b) the first 10 lags of the autocorrelation function. Draw your conclusion by using the 5% significance level. The data are available from CRSP.

**5.** Consider the daily log returns of Hewlett-Packard stock, value-weighted index, equal-weighted index, and S&P 500 index from January 1980 to December 1999 for 5056 observations. The returns include all distributions and are in percentages. The data can be obtained from CRSP or from the file "d-hwp3dx8099.dat," which has four columns with the same ordering as stated before. For each return series, test the hypothesis $H_o : \rho_1 = \cdots = \rho_{10} = 0$ versus the alternative hypothesis $H_a : \rho_i \neq 0$ for some $i \in \{1, \ldots, 10\}$, where $\rho_i$ is the lag-$i$ autocor-

relation. Draw your conclusion based on the 5% significance level. Compare the results between returns of individual stocks and market indexes.

**6.** Consider the monthly log returns of CRSP equal-weighted index from January 1962 to December 1999 for 456 observations. You may obtain the data from CRSP directly or from the file "m-ew6299.dat" on the Web.

- Build an AR model for the series and check the fitted model.
- Build an MA model for the series and check the fitted model.
- Compute 1- and 2-step ahead forecasts of the AR and MA models built in the previous two questions.
- Compare the fitted AR and MA models.

**7.** Column 3 of the file "d-hwp3dx8099.dat" contains the daily log returns of the CRSP equal-weighted index from January 1980 to December 1999.

- Build an AR model for the series and check the fitted model.
- Build an ARMA model for the series and check the fitted model.
- Use the fitted AR model to compute 1-step to 7-step ahead forecasts at the forecast origin December 27, 1999 (i.e., $h = 5052$). Note that for this particular instance the lag-5 coefficient is statistically significant. This might be due to the weekend effects.

**8.** Again, consider the daily log return of CRSP equal-weighted index from January 1980 to December 1999. Create indicator variables for Mondays, Tuesdays, Wednesdays, and Thursdays and use a regression model, possibly with time series errors, to study the effects of trading days on the index return. What is the fitted model? Are there serial correlations in the residuals?

**9.** This problem is concerned with the dynamic relationship between the spot and futures prices of the S&P500 index. The data file "sp5may.dat" has three columns: log(futures price), log(spot price), and cost-of-carry ($\times 100$). The data were obtained from the Chicago Mercantile Exchange for the S&P 500 stock index in May 1993 and its June futures contract. The time interval is 1 minute (intraday). Several authors used the data to study index futures arbitrage. Here we focus on the first two columns. Let $f_t$ and $s_t$ be the log prices of futures and spot, respectively. Consider $y_t = f_t - f_{t-1}$ and $x_t = s_t - s_{t-1}$. Build a regression model with time series errors between $\{y_t\}$ and $\{x_t\}$, with $y_t$ being the dependent variable.

**10.** The data file "qunemrate.dat" contains the U.S. quarterly unemployment rate, seasonally adjusted, from 1948 to the second quarter of 1991. Consider the change series $y_t = x_t - x_{t-1}$, where $x_t$ is the quarterly unemployment rate.

Build an AR model for the $y_t$ series. Does the fitted model suggest the existence of business cycles?

**11.** The quarterly gross domestic product implicit price deflator is often used as a measure of inflation. The file "gdpipd.dat" contains the data for U.S. from the first quarter of 1947 to the last quarter of 2000. The data are seasonally adjusted and equal to 100 for year 1996. Build an ARIMA model for the series and check the validity of the fitted model. The data are obtained from the Federal Reserve Bank of St Louis.

## REFERENCES


Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in B.N. Petrov and F. Csaki, ed. *2nd International Symposium on Information Theory*, 267–281. Akademia Kiado: Budapest.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994), *Time Series Analysis: Forecasting and Control*, 3rd edition, Prentice Hall: Englewood Cliffs, New Jersey.

Box, G. E. P., and Pierce, D. (1970), "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models," *Journal of the American Statistical Association*, 65, 1509–1526.

Brockwell, P. J., and Davis, R. A. (1996), *Introduction to Time Series and Forecasting*, Springer: New York.

Brockwell, P. J., and Davis, R. A. (1991), *Time Series: Theory and Methods*, 2nd edition, Springer-Verlag: New York.

Chan, N. H., and Wei, C. Z. (1988), "Limiting Distributions of Least Squares Estimates of Unstable Autoregressive Processes," *Annals of Statistics*, 16, 367–401.

Dickey, D. A., and Fuller, W. A. (1979), "Distribution of the Estimates for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, 427–431.

Ding, Z., Granger, C. W. J., and Engle, R. F. (1993), "A Long Memory Property of Stock Returns and a New Model," *Journal of Empirical Finance*, 1, 83–106.

Fuller, W. A. (1976), *Introduction to Statistical Time Series*, Wiley: New York.

Greene, W. H. (2000), *Econometric Analysis*, 4th edition, Prentice-Hall: Upper Saddle River, New Jersey.

Hosking, J. R. M. (1981), "Fractional Differencing," *Biometrika*, 68, 165–176.

Ljung, G., and Box, G. E. P. (1978), "On a Measure of Lack of Fit in Time Series Models," *Biometrika*, 66, 67–72.

Phillips, P. C. B. (1987), "Time Series Regression with a Unit Root," *Econometrica*, 55, 277–301.

Shumway, R. H., and Stoffer, D. S. (2000), *Time Series Analysis and its Applications*, Springer-Velag: New York.

Tiao, G. C., and Tsay, R. S. (1983), "Consistency Properties of Least Squares Estimates of Autoregressive Parameters in ARMA Models," *Annals of Statistics*, 11, 856–871.

Tsay, R. S., and Tiao, G. C. (1984), "Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Function for Stationary and Nonstationary ARMA Models," *Journal of the American Statistical Association*, 79, 84–96.

*Analysis of Financial Time Series.* Ruey S. Tsay

ISBN: 0-471-41544-8

CHAPTER 3

# Conditional Heteroscedastic Models

The objective of this chapter is to study some econometric models available in the literature for modeling the volatility of an asset return. These models are referred to as conditional heteroscedastic models.

Volatility is an important factor in options trading. Here volatility means the conditional variance of the underlying asset return. Consider, for example, the price of a European *call option*, which is a contract giving its holder the right, but not the obligation, to buy a fixed number of shares of a specified common stock at a fixed price on a given date. The fixed price is called the *strike price* and is commonly denoted by $K$. The given date is called the expiration date. The important time duration here is the time to expiration, and we denote it by $\ell$. If the holder can exercise her right any time on or before the expiration date, then the option is called an *American call option*. The well-known Black–Scholes option pricing formula states that the price of a European call option is

$$c_t = P_t\Phi(x) - Kr^{-\ell}\Phi(x - \sigma_t\sqrt{\ell}), \quad \text{and} \quad x = \frac{\ln(P_t/Kr^{-\ell})}{\sigma_t\sqrt{\ell}} + \frac{1}{2}\sigma_t\sqrt{\ell}, \qquad (3.1)$$

where $P_t$ is the current price of the underlying stock, $r$ is the risk-free interest rate, $\sigma_t$ is the conditional standard deviation of the log return of the specified stock, and $\Phi(x)$ is the cumulative distribution function of the standard normal random variable evaluated at $x$. A derivation of the formula is given later in Chapter 6. The formula has several nice interpretations, but it suffices to say here that the conditional variance of the log return of the underlying stock plays an important role. This volatility evolves over time.

Volatility is also important in risk management. As discussed in Chapter 7, volatility modeling provides a simple approach to calculating value at risk of a financial position. Finally, modeling the volatility of a time series can improve the efficiency in parameter estimation and the accuracy in interval forecast.

The univariate volatility models discussed in this chapter include the autoregressive conditional heteroscedastic (ARCH) model of Engle (1982), the generalized ARCH (GARCH) model of Bollerslev (1986), the exponential GARCH (EGARCH)

model of Nelson (1991), the conditional heteroscedastic autoregressive moving-average (CHARMA) model of Tsay (1987), the random coefficient autoregressive (RCA) model of Nicholls and Quinn (1982), and the stochastic volatility (SV) models of Melino and Turnbull (1990), Harvey, Ruiz, and Shephard (1994), and Jacquier, Polson, and Rossi (1994). We also discuss advantages and weaknesses of each volatility model and show some applications of the models. Multivariate volatility models, including those with time-varying correlations, are discussed in Chapter 9.

## 3.1 CHARACTERISTICS OF VOLATILITY

A special feature of stock volatility is that it is not directly observable. For example, consider the daily log returns of IBM stock. The daily volatility is not directly observable from the returns because there is only one observation in a trading day. If intraday data of the stock, such as 5-minute returns, are available, then one can estimate the daily volatility. The accuracy of such an estimate deserves a careful study, however. Furthermore, stock volatility consists of intraday volatility and variation between trading days. The unobservability of volatility makes it difficult to evaluate the forecasting performance of conditional heteroscedastic models. We discuss this issue later.

In options markets, if one accepts the idea that the prices are governed by an econometric model such as the Black–Scholes formula, then one can use the price to obtain the "implied" volatility. Yet this approach is often criticized for using a specific model, which is based on some assumptions that might not hold in practice. For instance, from the observed prices of a European call option, one can use the Black–Scholes formula in Eq. (3.1) to deduce the conditional standard deviation $\sigma_t$. The resulting value of $\sigma_t^2$ is called the *implied volatility* of the underlying stock. However, this implied volatility is derived under the log normal assumption for the return series. It might be very different from the actual volatility. Experience shows that implied volatility of an asset return tends to be larger than that obtained by using a GARCH type of volatility model.

Although volatility is not directly observable, it has some characteristics that are commonly seen in asset returns. First, there exist volatility clusters (i.e., volatility may be high for certain time periods and low for other periods). Second, volatility evolves over time in a continuous manner—that is, volatility jumps are rare. Third, volatility does not diverge to infinity—that is, volatility varies within some fixed range. Statistically speaking, this means that volatility is often stationary. Fourth, volatility seems to react differently to a big price increase or a big price drop. These properties play an important role in the development of volatility models. Some volatility models were proposed specifically to correct the weaknesses of the existing ones for their inability to capture the characteristics mentioned earlier. For example, the EGARCH model was developed to capture the asymmetry in volatility induced by big "positive" and "negative" asset returns.

## 3.2 STRUCTURE OF A MODEL

Let $r_t$ be the log return of an asset at time index $t$. The basic idea behind volatility study is that the series $\{r_t\}$ is either serially uncorrelated or with minor lower order serial correlations, but it is dependent. For illustration, Figure 3.1 shows the ACF and PACF of some functions of the monthly log stock returns of Intel Corporation from January 1973 to December 1997. The upper left panel shows the sample ACF of the return, which suggests no significant serial correlations except for a minor one at lag 7. The upper right panel shows the sample ACF of the absolute log returns (i.e., $|r_t|$), whereas the lower left panel shows the sample ACF of the squared returns $r_t^2$. These two plots clearly suggest that the monthly returns are not independent. Combining the three plots, it seems that the returns are indeed serially uncorrelated, but dependent. Volatility models attempt to capture such dependence in the return series.

To put the volatility models in a proper perspective, it is informative to consider the conditional mean and conditional variance of $r_t$ given $F_{t-1}$—that is,

$$\mu_t = E(r_t \mid F_{t-1}), \qquad \sigma_t^2 = \text{Var}(r_t \mid F_{t-1}) = E[(r_t - \mu_t)^2 \mid F_{t-1}], \tag{3.2}$$

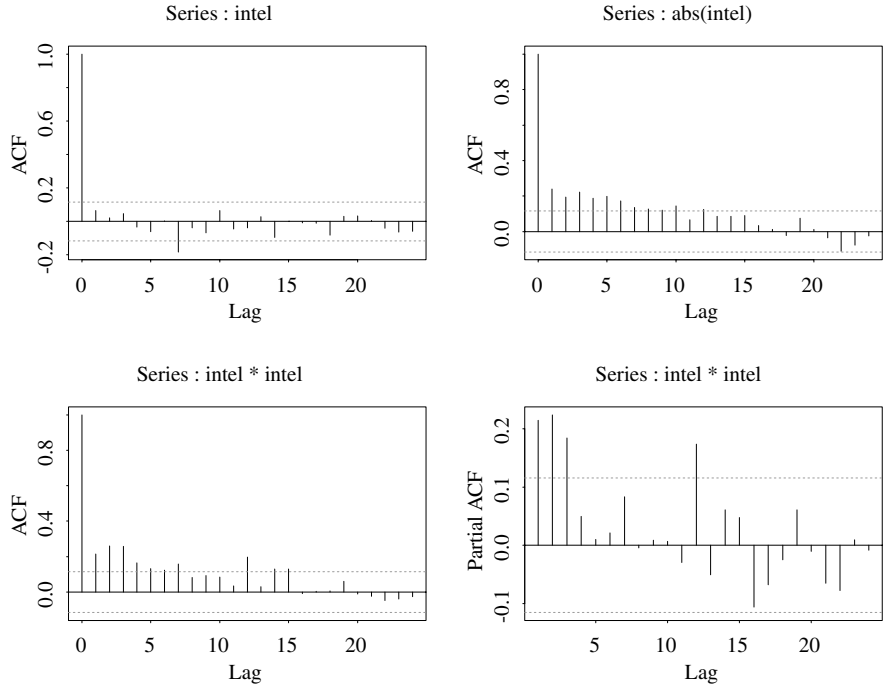


**Figure 3.1.** Sample ACF and PACF of various functions of monthly log stock returns of Intel Corporation from January 1973 to December 1997: (a) ACF of the log returns, (b) ACF of the squared returns (lower left), (c) ACF of the absolute returns (upper right), and (d) PACF of the squared returns.

where $F_{t-1}$ denotes the information set available at time $t-1$. Typically, $F_{t-1}$ consists of all linear functions of the past returns. As shown by the empirical examples of Chapter 2 and Figure 3.1, serial dependence of a stock return series $r_t$ is weak if it exists at all. Therefore, the equation for $\mu_t$ in (3.2) should be simple, and we assume that $r_t$ follows a simple time series model such as a stationary ARMA$(p, q)$ model. In other words, we entertain the model

$$r_t = \mu_t + a_t, \quad \mu_t = \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i} - \sum_{i=1}^{q} \theta_i a_{t-i}, \tag{3.3}$$

for $r_t$, where $p$ and $q$ are non-negative integers.

Model (3.3) illustrates a possible financial application of the linear time series models of Chapter 2. The order $(p, q)$ of an ARMA model may depend on the frequency of the return series. For example, daily returns of a market index often show some minor serial correlations, but monthly returns of the index may not contain any significant serial correlation. One may include some explanatory variables to the conditional mean equation and use a linear regression model with time series errors to capture the behavior of $\mu_t$. For example, a dummy variable can be used for the Mondays to study the effect of weekend on daily stock returns.

Combining Eqs. (3.2) and (3.3), we have

$$\sigma_t^2 = \text{Var}(r_t \mid F_{t-1}) = \text{Var}(a_t \mid F_{t-1}). \tag{3.4}$$

The conditional heteroscedastic models of this chapter are concerned with the evolution of $\sigma_t^2$. The manner under which $\sigma_t^2$ evolves over time distinguishes one volatility model from another.

Conditional heteroscedastic models can be classified into two general categories. Those in the first category use an exact function to govern the evolution of $\sigma_t^2$, whereas those in the second category use a stochastic equation to describe $\sigma_t^2$. The GARCH model belongs to the first category, and the stochastic volatility model is in the second category.

For simplicity in introducing volatility models, we assume that the model for the conditional mean is given. However, we estimate the conditional mean and variance equations jointly in empirical studies. Throughout the book, $a_t$ is referred to as the *shock* or *mean-corrected return* of an asset return at time $t$ and $\sigma_t$ is the positive square-root of $\sigma_t^2$. The model for $\mu_t$ in Eq. (3.3) is referred to as the *mean* equation for $r_t$ and the model for $\sigma_t^2$ is the *volatility* equation for $r_t$. Therefore, modeling conditional heteroscedasticity amounts to augmenting a dynamic equation to a time series model to govern the time evolution of the conditional variance of the shock.

## 3.3 THE ARCH MODEL

The first model that provides a systematic framework for volatility modeling is the ARCH model of Engle (1982). The basic idea of ARCH models is that (a) the mean-

corrected asset return $a_t$ is serially uncorrelated, but dependent, and (b) the dependence of $a_t$ can be described by a simple quadratic function of its lagged values. Specifically, an ARCH($m$) model assumes that

$$a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_m a_{t-m}^2, \tag{3.5}$$

where $\{\epsilon_t\}$ is a sequence of independent and identically distributed (iid) random variables with mean zero and variance 1, $\alpha_0 > 0$, and $\alpha_i \geq 0$ for $i > 0$. The coefficients $\alpha_i$ must satisfy some regularity conditions to ensure that the unconditional variance of $a_t$ is finite. In practice, $\epsilon_t$ is often assumed to follow the standard normal or a standardized Student-$t$ distribution.

From the structure of the model, it is seen that large past squared shocks $\{a_{t-i}^2\}_{i=1}^m$ imply a large conditional variance $\sigma_t^2$ for the mean-corrected return $a_t$. Consequently, $a_t$ tends to assume a large value (in modulus). This means that, under the ARCH framework, large shocks tend to be followed by another large shock. Here I use the word *tend* because a large variance does not necessarily produce a large variate. It only says that the probability of obtaining a large variate is greater than that of a smaller variance. This feature is similar to the volatility clusterings observed in asset returns.

The ARCH effect also occurs in other financial time series. Figure 3.2 shows the time plots of (a) the percentage changes in Deutsche Mark/U.S. Dollar exchange rate measured in 10-minute intervals from June 5, 1989 to June 19, 1989 for 2488 observations, and (b) the squared series of the percentage changes. Big percentage changes occurred occasionally, but there exist certain stable periods. Figure 3.3(a) shows the sample ACF of the percentage change series. Clearly, the series has no serial correlation. Figure 3.3(b) shows the sample PACF of the squared series of percentage changes. It is seen that there are some big spikes in the PACF. Such spikes suggest that the percentage changes are not independent and have some ARCH effects.

***Remark:*** Some authors use $h_t$ to denote the conditional variance in Eq. (3.5). In this case, the shock becomes $a_t = \sqrt{h_t}\epsilon_t$.

### 3.3.1 Properties of ARCH Models

To understand the ARCH models, it pays to carefully study the ARCH(1) model

$$a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2,$$

where $\alpha_0 > 0$ and $\alpha_1 \geq 0$. First, the unconditional mean of $a_t$ remains zero because

$$E(a_t) = E[E(a_t \mid F_{t-1})] = E[\sigma_t E(\epsilon_t)] = 0.$$

Second, the unconditional variance of $a_t$ can be obtained as

$$\text{Var}(a_t) = E(a_t^2) = E[E(a_t^2 \mid F_{t-1})] = E(\alpha_0 + \alpha_1 a_{t-1}^2) = \alpha_0 + \alpha_1 E(a_{t-1}^2).$$

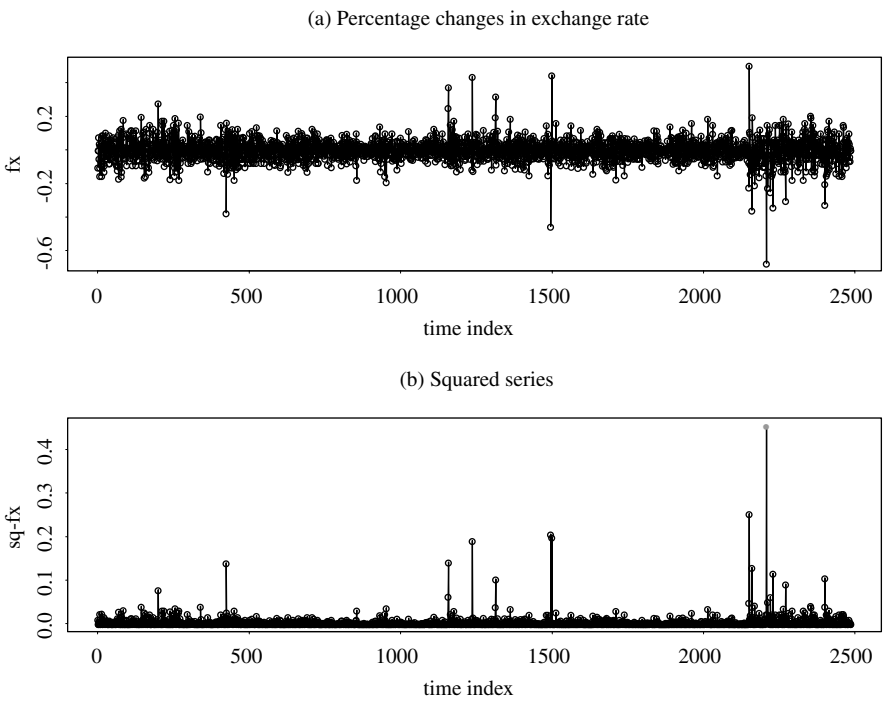


**Figure 3.2.** (a) Time plot of 10-minute returns of the exchange rate between Deutsche Mark and Dollar, and (b) the squared returns.

Because $a_t$ is a stationary process with $E(a_t) = 0$, $\text{Var}(a_t) = \text{Var}(a_{t-1}) = E(a_{t-1}^2)$. Therefore, we have $\text{Var}(a_t) = \alpha_0 + \alpha_1 \text{Var}(a_t)$ and $\text{Var}(a_t) = \alpha_0/(1 - \alpha_1)$. Because the variance of $a_t$ must be positive, we need $0 \leq \alpha_1 < 1$. Third, in some applications, we need higher order moments of $a_t$ to exist and, hence, $\alpha_1$ must also satisfy some additional constraints. For instance, to study its tail behavior, we require that the fourth moment of $a_t$ is finite. Under the normality assumption of $\epsilon_t$ in Eq. (3.5), we have

$$E(a_t^4 \mid F_{t-1}) = 3[E(a_t^2 \mid F_{t-1})]^2 = 3(\alpha_0 + \alpha_1 a_{t-1}^2)^2.$$

Therefore,

$$E(a_t^4) = E[E(a_t^4 \mid F_{t-1})] = 3E(\alpha_0 + \alpha_1 a_{t-1}^2)^2 = 3E[\alpha_0^2 + 2\alpha_0\alpha_1 a_{t-1}^2 + \alpha_1^2 a_{t-1}^4].$$

If $a_t$ is fourth-order stationary with $m_4 = E(a_t^4)$, then we have

$$\begin{aligned} m_4 &= 3[\alpha_0^2 + 2\alpha_0\alpha_1 \text{Var}(a_t) + \alpha_1^2 m_4] \\ &= 3\alpha_0^2 \left(1 + 2\frac{\alpha_1}{1 - \alpha_1}\right) + 3\alpha_1^2 m_4. \end{aligned}$$

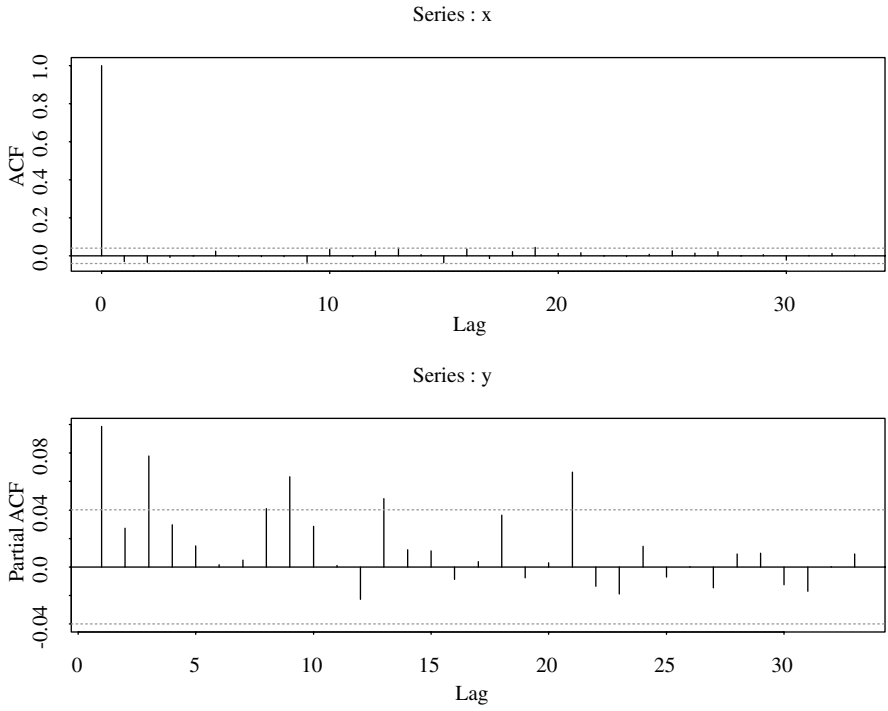


**Figure 3.3.** (a) Sample autocorrelation function of the return series of Mark/Dollar exchange rate, and (b) sample partial autocorrelation function of the squared returns.

Consequently,

$$m_4 = \frac{3\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)}.$$

This result has two important implications: (a) since the fourth moment of $a_t$ is positive, we see that $\alpha_1$ must also satisfy the condition $1-3\alpha_1^2 > 0$; that is, $0 \leq \alpha_1^2 < 1/3$; and (b) the unconditional kurtosis of $a_t$ is

$$\frac{E(a_t^4)}{[\text{Var}(a_t)]^2} = 3\frac{\alpha_0^2(1+\alpha_1)}{(1-\alpha_1)(1-3\alpha_1^2)} \times \frac{(1-\alpha_1)^2}{\alpha_0^2} = 3\frac{1-\alpha_1^2}{1-3\alpha_1^2} > 3.$$

Thus, the excess kurtosis of $a_t$ is positive and the tail distribution of $a_t$ is heavier than that of a normal distribution. In other words, the shock $a_t$ of a conditional Gaussian ARCH(1) model is more likely than a Gaussian white noise series to produce "outliers." This is in agreement with the empirical finding that "outliers" appear more often in asset returns than that implied by an iid sequence of normal random variates.

These properties continue to hold for general ARCH models, but the formulas become more complicated for higher order ARCH models. The condition $\alpha_i \geq 0$ in Eq. (3.5) can be relaxed. It is a condition to ensure that the conditional variance $\sigma_t^2$ is positive for all $t$. In fact, a natural way to achieve positiveness of the conditional variance is to rewrite an ARCH($m$) model as

$$a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + A'_{m,t-1} \Omega A_{m,t-1}, \tag{3.6}$$

where $A_{m,t-1} = (a_{t-1}, \ldots, a_{t-m})'$ and $\Omega$ is a $m \times m$ non-negative definite matrix. The ARCH($m$) model in Eq. (3.5) requires $\Omega$ to be diagonal. Thus, Engle's model uses a parsimonious approach to approximate a quadratic function. A simple way to achieve Eq. (3.6) is to employ a random-coefficient model for $a_t$; see the CHARMA and RCA models discussed later.

### 3.3.2 Weaknesses of ARCH Models

The advantages of ARCH models include properties discussed in the previous subsection. The model also has some weaknesses:

1. The model assumes that positive and negative shocks have the same effects on volatility because it depends on the square of the previous shocks. In practice, it is well known that price of a financial asset responds differently to positive and negative shocks.
2. The ARCH model is rather restrictive. For instance, $\alpha_1^2$ of an ARCH(1) model must be in the interval $[0, \frac{1}{3}]$ if the series is to have a finite fourth moment. The constraint becomes complicated for higher order ARCH models.
3. The ARCH model does not provide any new insight for understanding the source of variations of a financial time series. They only provide a mechanical way to describe the behavior of the conditional variance. It gives no indication about what causes such behavior to occur.
4. ARCH models are likely to overpredict the volatility because they respond slowly to large isolated shocks to the return series.

### 3.3.3 Building an ARCH Model

A simple way to build an ARCH model consists of three steps: (1) build an econometric model (e.g., an ARMA model) for the return series to remove any linear dependence in the data, and use the residual series of the model to test for ARCH effects; (2) specify the ARCH order and perform estimation; and (3) check the fitted ARCH model carefully and refine it if necessary. More details are given later.

***Modeling the Mean Effect and Testing***

An ARMA model is built for the observed time series to remove any serial correlations in the data. For most asset return series, this step amounts to removing the

sample mean from the data if the sample mean is significantly different from zero. For some daily return series, a simple AR model might be needed. The squared series $a_t^2$ is used to check for conditional heteroscedasticity, where $a_t = r_t - \mu_t$ is the residual of the ARMA model. Two tests are available here. The first test is to check the usual Ljung–Box statistics of $a_t^2$; see McLeod and Li (1983). The second test for conditional heteroscedasticity is the Lagrange multiplier test of Engle (1982). This test is equivalent to the usual $F$ statistic for testing $\alpha_i = 0$ $(i = 1, \ldots, m)$ in the linear regression

$$a_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_m a_{t-m}^2 + e_t, \quad t = m+1, \ldots, T,$$

where $e_t$ denotes the error term, $m$ is a prespecified positive integer, and $T$ is the sample size. Let $SSR_0 = \sum_{t=m+1}^{T} (a_t^2 - \bar{\omega})^2$, where $\bar{\omega}$ is the sample mean of $a_t^2$, and $SSR_1 = \sum_{t=m+1}^{T} \hat{e}_t^2$, where $\hat{e}_t$ is the least squares residual of the prior linear regression. Then we have

$$F = \frac{(SSR_0 - SSR_1)/m}{SSR_1/(T - 2m - 1)},$$

which is asymptotically distributed as a chi-squared distribution with $m$ degrees of freedom under the null hypothesis.

***Order Determination***

If the test statistic $F$ is significant, then conditional heteroscedasticity of $a_t$ is detected, and we use the PACF of $a_t^2$ to determine the ARCH order. Using PACF of $a_t^2$ to select the ARCH order can be justified as follows. From the model in Eq. (3.5), we have

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_m a_{t-m}^2.$$

For a given sample, $a_t^2$ is an unbiased estimate of $\sigma_t^2$. Therefore, we expect that $a_t^2$ is linearly related to $a_{t-1}^2, \ldots, a_{t-m}^2$ in a manner similar to that of an autoregressive model of order $m$. Note that a single $a_t^2$ is generally not an efficient estimate of $\sigma_t^2$, but it can serve as an approximation that could be informative in specifying the order $m$.

Alternatively, define $\eta_t = a_t^2 - \sigma_t^2$. It can be shown that $\{\eta_t\}$ is an un-correlated series with mean 0. The ARCH model then becomes

$$a_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_m a_{t-m}^2 + \eta_t,$$

which is in the form of an AR($m$) model for $a_t^2$, except that $\{\eta_t\}$ is not an iid series. From Chapter 2, PACF of $a_t^2$ is a useful tool to determine the order $m$. Because $\{\eta_t\}$ are not identically distributed, the least squares estimates of the prior model are consistent, but not efficient. The PACF of $a_t^2$ may not be effective when the sample size is small.

***Estimation***

Two likelihood functions are commonly used in ARCH estimation. Under the normality assumption, the likelihood function of an ARCH($m$) model is

$$\begin{aligned} & f(a_1, \ldots, a_T \mid \boldsymbol{\alpha}) \\ &\quad = f(a_T \mid F_{T-1}) f(a_{T-1} \mid F_{T-2}) \cdots f(a_{m+1} \mid F_m) f(a_1, \ldots, a_m \mid \boldsymbol{\alpha}) \\ &\quad = \prod_{t=m+1}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left[-\frac{a_t^2}{2\sigma_t^2}\right] \times f(a_1, \ldots, a_m \mid \boldsymbol{\alpha}), \end{aligned}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_m)'$ and $f(a_1, \ldots, a_m \mid \boldsymbol{\alpha})$ is the joint probability density function of $a_1, \ldots, a_m$. Since the exact form of $f(a_1, \ldots, a_m \mid \boldsymbol{\alpha})$ is complicated, it is commonly dropped from the prior likelihood function, especially when the sample size is sufficiently large. This results in using the conditional likelihood function

$$f(a_{m+1}, \ldots, a_T \mid \boldsymbol{\alpha}, a_1, \ldots, a_m) = \prod_{t=m+1}^{T} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left[-\frac{a_t^2}{2\sigma_t^2}\right],$$

where $\sigma_t^2$ can be evaluated recursively. We refer to estimates obtained by maximizing the prior likelihood function as the conditional maximum likelihood estimates (MLE) under normality.

Maximizing the conditional likelihood function is equivalent to maximizing its logarithm, which is easier to handle. The conditional log likelihood function is

$$\ell(a_{m+1}, \ldots, a_T \mid \boldsymbol{\alpha}, a_1, \ldots, a_m) = \sum_{t=m+1}^{T} -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma_t^2) - \frac{1}{2}\frac{a_t^2}{\sigma_t^2}.$$

Since the first term $\ln(2\pi)$ does not involve any parameters, the log likelihood function becomes

$$\ell(a_{m+1}, \ldots, a_T \mid \boldsymbol{\alpha}, a_1, \ldots, a_m) = -\sum_{t=m+1}^{T} \left[\frac{1}{2}\ln(\sigma_t^2) + \frac{1}{2}\frac{a_t^2}{\sigma_t^2}\right],$$

where $\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \cdots + \alpha_m a_{t-m}^2$ can be evaluated recursively.

In some applications, it is more appropriate to assume that $\epsilon_t$ follows a heavy-tailed distribution such as a standardized Student-$t$ distribution. Let $x_v$ be a Student-$t$ distribution with $v$ degrees of freedom. Then $\mathrm{Var}(x_v) = v/(v-2)$ for $v > 2$, and we use $\epsilon_t = x_v/\sqrt{v/(v-2)}$. The probability density function of $\epsilon_t$ is

$$f(\epsilon_t \mid v) = \frac{\Gamma((v+1)/2)}{\Gamma(v/2)\sqrt{(v-2)\pi}} \left(1 + \frac{\epsilon_t^2}{v-2}\right)^{-(v+1)/2}, \quad v > 2, \tag{3.7}$$

where $\Gamma(x)$ is the usual Gamma function [i.e., $\Gamma(x) = \int_0^\infty y^{x-1}e^{-y}dy$]. Using $a_t = \sigma_t\epsilon_t$, we obtain the conditional likelihood function of $a_t$s as

$$f(a_{m+1}, \ldots, a_T \mid \boldsymbol{\alpha}, A_m) = \prod_{t=m+1}^{T} \frac{\Gamma((v+1)/2)}{\Gamma(v/2)\sqrt{(v-2)\pi}} \frac{1}{\sigma_t} \left[1 + \frac{a_t^2}{(v-2)\sigma_t^2}\right]^{-(v+1)/2},$$

where $v > 2$ and $A_m = (a_1, a_2, \ldots, a_m)$. We refer to the estimates that maximize the prior likelihood function as the conditional MLE under $t$-distribution. The degrees of freedom of the $t$-distribution can be specified a priori or estimated jointly with other parameters. A value between 3 and 6 is often used if it is prespecified.

If the degrees of freedom $v$ of Student-$t$ distribution is prespecified, then the conditional log likelihood function is

$$\ell(a_{m+1}, \ldots, a_T \mid \boldsymbol{\alpha}, A_m) = -\sum_{t=m+1}^{T} \left[\frac{v+1}{2} \ln\left(1 + \frac{a_t^2}{(v-2)\sigma_t^2}\right) + \frac{1}{2}\ln(\sigma_t^2)\right]. \tag{3.8}$$

If one wishes to estimate $v$ jointly with other parameters, then the log likelihood function involving degrees of freedom

$$\begin{aligned}\ell(a_{m+1}, \ldots, a_T \mid \boldsymbol{\alpha}, v, A_m) &= (T-m)[\ln(\Gamma((v+1)/2)) - \ln(\Gamma(v/2)) - 0.5\ln((v-2)\pi)] \\ &\quad + \ell(a_{m+1}, \ldots, a_T \mid \boldsymbol{\alpha}, A_m),\end{aligned}$$

where the second term is given in Eq. (3.8).

***Model Checking***

For an ARCH model, the standardized shocks

$$\tilde{a}_t = \frac{a_t}{\sigma_t}$$

are iid random variates following either a standard normal or standardized Student-$t$ distribution. Therefore, one can check the adequacy of a fitted ARCH model by examining the series $\{\tilde{a}_t\}$. In particular, the Ljung–Box statistics of $\tilde{a}_t$ can be used to check the adequacy of the mean equation and that of $\tilde{a}_t^2$ can be used to test the validity of the volatility equation. The skewness, kurtosis, and quantile-to-quantile plot (i.e., QQ-plot) of $\{\tilde{a}_t\}$ can be used to check the validity of the distribution assumption.

***Forecasting***

Forecasts of the ARCH model in Eq. (3.5) can be obtained recursively as those of an AR model. Consider an ARCH($m$) model. At the forecast origin $h$, the 1-step ahead

forecast of $\sigma_{h+1}^2$ is

$$\sigma_h^2(1) = \alpha_0 + \alpha_1 a_h^2 + \cdots + \alpha_m a_{h+1-m}^2.$$

The 2-step ahead forecast is

$$\sigma_h^2(2) = \alpha_0 + \alpha_1 \sigma_h^2(1) + \alpha_2 a_h^2 + \cdots + \alpha_m a_{h+2-m}^2,$$

and the $\ell$-step ahead forecast for $\sigma_{h+\ell}^2$ is

$$\sigma_h^2(\ell) = \alpha_0 + \sum_{i=1}^{m} \alpha_i \sigma_h^2(\ell - i), \tag{3.9}$$

where $\sigma_h^2(\ell - i) = a_{h+\ell-i}^2$ if $\ell - i \leq 0$.

### 3.3.4 Examples

In this subsection, we illustrate ARCH modeling by considering two examples.

**Example 3.1.** We first apply the modeling procedure to build a simple ARCH model for the monthly log stock returns of Intel Corporation. The sample ACF and PACF of the squared returns in Figure 3.1 clearly show the existence of conditional heteroscedasticity. Thus, it is unnecessary to perform any statistical tests to confirm the need of ARCH modeling, and we proceed to identify the order of an ARCH model. The sample PACF in the lower right panel of Figure 3.1 indicates that an ARCH(3) model might be appropriate. Consequently, we specify the model

$$r_t = \mu + a_t, \quad a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \alpha_2 a_{t-2}^2 + \alpha_3 a_{t-3}^2$$

for the monthly log returns of Intel stock. Assuming that $\epsilon_t$ are iid standard normal, we obtain the fitted model

$$r_t = 0.0196 + a_t, \quad \sigma_t^2 = 0.0090 + 0.2973 a_{t-1}^2 + 0.0900 a_{t-2}^2 + 0.0626 a_{t-3}^2,$$

where the standard errors of the parameters are 0.0062, 0.0013, 0.0887, 0.0645, and 0.0777, respectively. While the estimates meet the general requirement of an ARCH(3) model, the estimates of $\alpha_2$ and $\alpha_3$ appear to be statistically nonsignificant at the 5% level. Therefore, the model can be simplified.

Dropping the two nonsignificant parameters, we obtain the model

$$r_t = 0.0213 + a_t, \quad \sigma_t^2 = 0.00998 + 0.4437 a_{t-1}^2, \tag{3.10}$$

where the standard errors of the parameters are 0.0062, 0.00124, and 0.0938, respectively. All the estimates are highly significant. Figure 3.4 shows the standardized

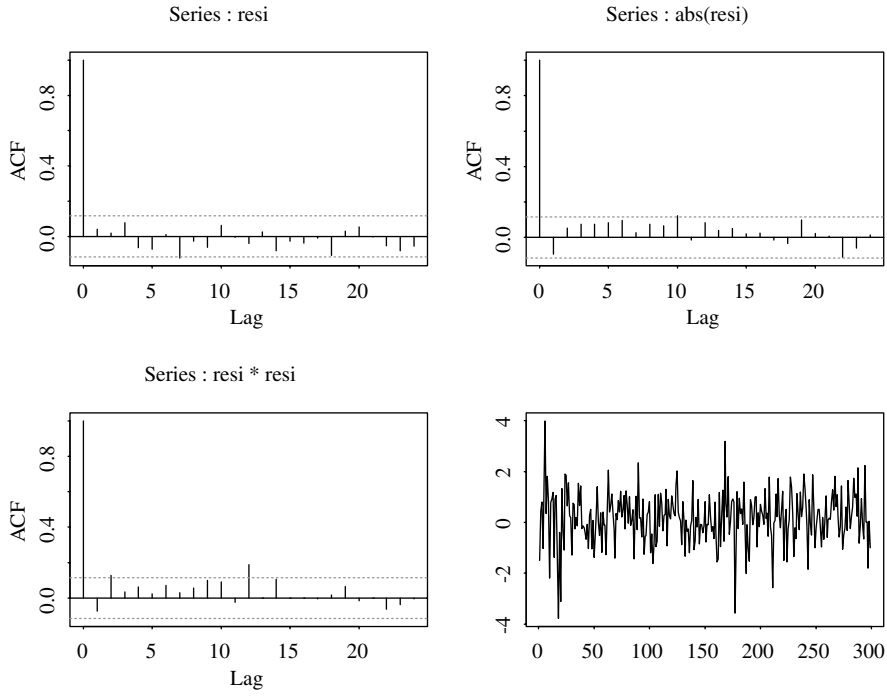


**Figure 3.4.** Model checking statistics of the Gaussian ARCH(1) model in Eq. (3.10) for the monthly log stock returns of Intel from January 1973 to December 1997: parts (a), (b), and (c) show the sample ACF of the standardized shocks, their squared series, and absolute series, respectively, and (d) is the time plot of standardized shocks.

shocks and the sample ACF of some functions of the standardized shocks. The Ljung–Box statistics of the standardized shocks $\{\tilde{a}_t\}$ give $Q(10) = 12.53$ with $p$ value 0.25 and those of $\{\tilde{a}_t^2\}$ give $Q(10) = 17.23$ with $p$ value 0.07. Consequently, the ARCH(1) model in Eq. (3.10) is adequate for the data at the 5% significance level.

The ARCH(1) model in Eq. (3.10) has some interesting properties. First, the expected monthly log return for Intel stock is about 2.1%, which is remarkable. Second, $\hat{\alpha}_1^2 = 0.444^2 < 1/3$ so that the unconditional fourth moment of the monthly log return of Intel stock exists. Third, the unconditional variance of $r_t$ is $0.00998/(1 - 0.4437) = 0.0179$. Finally, the ARCH(1) model can be used to predict the monthly volatility of Intel stock returns.

***t* Innovation**

For comparison, we also fit an ARCH(1) model to the series, assuming that $\epsilon_t$ follows a standardized Student-$t$ distribution with 5 degrees of freedom. The resulting model

is

$$r_t = 0.0222 + a_t, \quad \sigma_t^2 = 0.0121 + 0.3029a_{t-1}^2, \tag{3.11}$$

where the standard errors of the parameters are 0.0019, 0.1443, and 0.0061, respectively. All the estimates are significant at the 5% level, but the $t$ ratio of $\hat{\alpha}_1$ is only 2.10. The unconditional variance of $a_t$ is $0.0121/(1 - 0.3029) = 0.0174$, which is close to that obtained under normality. The Ljung–Box statistics of the standardized shocks give $Q(10) = 13.66$ with $p$-value 0.19, confirming that the mean equation is adequate. However, the Ljung–Box statistics for the squared standardized shocks show $Q(10) = 23.83$ with $p$ value 0.008. The volatility equation is inadequate at the 5% level. We refine the model by considering an ARCH(2) model and obtain

$$r_t = 0.0225 + a_t. \quad \sigma_t^2 = 0.0113 + 0.226a_{t-1}^2 + 0.108a_{t-2}^2, \tag{3.12}$$

where the standard errors of the parameters are 0.006, 0.002, 0.135, and 0.094, respectively. The coefficient of $a_{t-1}^2$ is marginally significant at the 10% level, but that of $a_{t-2}^2$ is only slightly greater than its standard error. The Ljung–Box statistics for the squared standardized shocks give $Q(10) = 8.82$ with $p$ value 0.55. Consequently, the fitted ARCH(2) model appears to be adequate.

Comparing models (3.10), (3.11), and (3.12), we see that (a) using a heavy-tailed distribution for $\epsilon_t$ reduces the ARCH effect, and (b) the difference among the three models is small for this particular instance. Finally, a more appropriate conditional heteroscedastic model for this data set is a GARCH(1, 1) model, which is discussed in the next section.

**Example 3.2.** Consider the percentage changes of the exchange rate between Mark and Dollar in 10-minute intervals. The data are shown in Figure 3.2(a). As shown in Figure 3.3(a), the series has no serial correlations. However, the sample PACF of the squared series $a_t^2$ shows some big spikes, especially at lags 1 and 3. There are some large PACF at higher lags, but the lower order lags tend to be more important. Following the procedure discussed in the previous subsection, we specify an ARCH(3) model for the series. Using the conditional Gaussian likelihood function, we obtain the fitted model

$$\sigma_t^2 = 0.22 \times 10^{-6} + 0.328a_{t-1}^2 + 0.073a_{t-2}^2 + 0.103a_{t-3}^2,$$

where all the estimates are statistically significant at the 5% significant level, and the standard errors of the parameters are $0.46 \times 10^{-8}$, 0.0162, 0.0160, and 0.0147, respectively. Model checking, using the standardized shock $\tilde{a}_t$, indicates that the model is adequate.

***Remark:*** The estimation of conditional heteroscedastic models of this chapter is carried out by the Regression Analysis of Time Series (RATS) package. There are

other softwares available, including Eviews, Scientific Computing Associates (SCA), and S-Plus.

## 3.4 THE GARCH MODEL

Although the ARCH model is simple, it often requires many parameters to adequately describe the volatility process of an asset return. For instance, consider the monthly excess returns of S&P 500 index. An ARCH(9) model is needed for the volatility process. Some alternative model must be sought. Bollerslev (1986) proposes a useful extension known as the generalized ARCH (GARCH) model. For a log return series $r_t$, we assume that the mean equation of the process can be adequatedly described by an ARMA model. Let $a_t = r_t - \mu_t$ be the mean-corrected log return. Then $a_t$ follows a GARCH$(m, s)$ model if

$$a_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{m} \alpha_i a_{t-i}^2 + \sum_{j=1}^{s} \beta_j \sigma_{t-j}^2, \tag{3.13}$$

where again $\{\epsilon_t\}$ is a sequence of iid random variables with mean 0 and variance 1.0, $\alpha_0 > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, and $\sum_{i=1}^{\max(m,s)} (\alpha_i + \beta_i) < 1$. Here it is understood that $\alpha_i = 0$ for $i > m$ and $\beta_j = 0$ for $j > s$. The latter constraint on $\alpha_i + \beta_i$ implies that the unconditional variance of $a_t$ is finite, whereas its conditional variance $\sigma_t^2$ evolves over time. As before, $\epsilon_t$ is often assumed to be a standard normal or standardized Student-$t$ distribution. Equation (3.13) reduces to a pure ARCH$(m)$ model if $s = 0$.

To understand properties of GARCH models, it is informative to use the following representation. Let $\eta_t = a_t^2 - \sigma_t^2$ so that $\sigma_t^2 = a_t^2 - \eta_t$. By plugging $\sigma_{t-i}^2 = a_{t-i}^2 - \eta_{t-i} (i = 0, \ldots, s)$ into Eq. (3.13), we can rewrite the GARCH model as

$$a_t^2 = \alpha_0 + \sum_{i=1}^{\max(m,s)} (\alpha_i + \beta_i) a_{t-i}^2 + \eta_t - \sum_{j=1}^{s} \beta_j \eta_{t-j}. \tag{3.14}$$

It is easy to check that $\{\eta_t\}$ is a martingale difference series [i.e., $E(\eta_t) = 0$ and $\text{cov}(\eta_t, \eta_{t-j}) = 0$ for $j \geq 1$]. However, $\{\eta_t\}$ in general is not an iid sequence. Equation (3.14) is an ARMA form for the squared series $a_t^2$. Thus, a GARCH model can be regarded as an application of the ARMA idea to the squared series $a_t^2$. Using the unconditional mean of an ARMA model, we have

$$E(a_t^2) = \frac{\alpha_0}{1 - \sum_{i=1}^{\max(m,s)} (\alpha_i + \beta_i)}$$

provided that the denominator of the prior fraction is positive.

The strengths and weaknesses of GARCH models can easily be seen by focusing on the simplest GARCH(1, 1) model with

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \quad 0 \le \alpha_1, \beta_1 \le 1, (\alpha_1 + \beta_1) < 1. \tag{3.15}$$

First, a large $a_{t-1}^2$ or $\sigma_{t-1}^2$ gives rise to a large $\sigma_t^2$. This means that a large $a_{t-1}^2$ tends to be followed by another large $a_t^2$, generating, again, the well-known behavior of volatility clustering in financial time series. Second, it can be shown that if $1 - 2\alpha_1^2 - (\alpha_1 + \beta_1)^2 > 0$, then

$$\frac{E(a_t^4)}{[E(a_t^2)]^2} = \frac{3[1 - (\alpha_1 + \beta_1)^2]}{1 - (\alpha_1 + \beta_1)^2 - 2\alpha_1^2} > 3.$$

Consequently, similar to ARCH models, the tail distribution of a GARCH(1, 1) process is heavier than that of a normal distribution. Third, the model provides a simple parametric function that can be used to describe the volatility evolution.

Forecasts of a GARCH model can be obtained using methods similar to those of an ARMA model. Consider the GARCH(1, 1) model in Eq. (3.15) and assume that the forecast origin is $h$. For 1-step ahead forecast, we have

$$\sigma_{h+1}^2 = \alpha_0 + \alpha_1 a_h^2 + \beta_1 \sigma_h^2,$$

where $a_h$ and $\sigma_h^2$ are known at the time index $h$. Therefore, the 1-step ahead forecast is

$$\sigma_h^2(1) = \alpha_0 + \alpha_1 a_h^2 + \beta_1 \sigma_h^2.$$

For multistep ahead forecasts, we use $a_t^2 = \sigma_t^2 \epsilon_t^2$ and rewrite the volatility equation in Eq. (3.15) as

$$\sigma_{t+1}^2 = \alpha_0 + (\alpha_1 + \beta_1)\sigma_t^2 + \alpha_1 \sigma_t^2(\epsilon_t^2 - 1).$$

When $t = h + 1$, the equation becomes

$$\sigma_{h+2}^2 = \alpha_0 + (\alpha_1 + \beta_1)\sigma_{h+1}^2 + \alpha_1 \sigma_{h+1}^2(\epsilon_{h+1}^2 - 1).$$

Since $E(\epsilon_{h+1}^2 - 1 \mid F_h) = 0$, the 2-step ahead volatility forecast at the forecast origin $h$ satisfies the equation

$$\sigma_h^2(2) = \alpha_0 + (\alpha_1 + \beta_1)\sigma_h^2(1).$$

In general, we have

$$\sigma_h^2(\ell) = \alpha_0 + (\alpha_1 + \beta_1)\sigma_h^2(\ell - 1), \quad \ell > 1. \tag{3.16}$$

This result is exactly the same as that of an ARMA(1, 1) model with AR polynomial $1 - (\alpha_1 + \beta_1)B$. By repeated substitutions in Eq. (3.16), we obtain that the $\ell$-step

ahead forecast can be written as

$$\sigma_h^2(\ell) = \frac{\alpha_0[1-(\alpha_1+\beta_1)^{\ell-1}]}{1-\alpha_1-\beta_1} + (\alpha_1+\beta_1)^{\ell-1}\sigma_h^2(1).$$

Therefore,

$$\sigma_h^2(\ell) \to \frac{\alpha_0}{1-\alpha_1-\beta_1}, \quad \text{as} \quad \ell \to \infty$$

provided that $\alpha_1 + \beta_1 < 1$. Consequently, the multistep ahead volatility forecasts of a GARCH(1, 1) model converge to the unconditional variance of $a_t$ as the forecast horizon increases to infinity provided that $\text{Var}(a_t)$ exists.

The literature on GARCH models is enormous; see Bollerslev, Chou, and Kroner (1992), Bollerslev, Engle, and Nelson (1994), and the references therein. The model encounters the same weaknesses as the ARCH model. For instance, it responds equally to positive and negative shocks. In addition, recent empirical studies of high-frequency financial time series indicate that the tail behavior of GARCH models remains too short even with standardized Student-$t$ innovations.

### 3.4.1 An Illustrative Example

The modeling procedure of ARCH models can also be used to build a GARCH model. However, specifying the order of an GARCH model is not easy. Only lower order GARCH models are used in most applications, say GARCH(1, 1), GARCH(2, 1), and GARCH(1, 2) models. The conditional maximum likelihood method continues to apply provided that the starting values of the volatility $\{\sigma_t^2\}$ are assumed to be known. Consider, for instance, a GARCH(1, 1) model. If $\sigma_1^2$ is treated as fixed, then $\sigma_t^2$ can be computed recursively for a GARCH(1, 1) model. In some applications, the sample variance of $a_t$ serves as a good starting value of $\sigma_1^2$. The fitted model can be checked by using the standardized residual $\tilde{a}_t = a_t/\sigma_t$ and its squared process.

**Example 3.3.** In this example, we consider the monthly excess returns of S&P 500 index starting from 1926 for 792 observations. The series is shown in Figure 3.5. Denote the excess return series by $r_t$. Figure 3.6 shows the sample ACF of $r_t$ and the sample PACF of $r_t^2$. The $r_t$ series has some serial correlations at lags 1 and 3, but the key feature is that the PACF of $r_t^2$ shows strong linear dependence. If an MA(3) model is entertained, we obtain

$$r_t = 0.0062 + a_t + 0.0944a_{t-1} - 0.1407a_{t-3}, \quad \hat{\sigma}_a = 0.0576$$

for the series, where all of the coefficients are significant at the 5% level. However, for simplicity, we use instead an AR(3) model

$$r_t = \phi_1 r_{t-1} + \phi_2 r_{t-2} + \phi_3 r_{t-3} + \beta_0 + a_t.$$

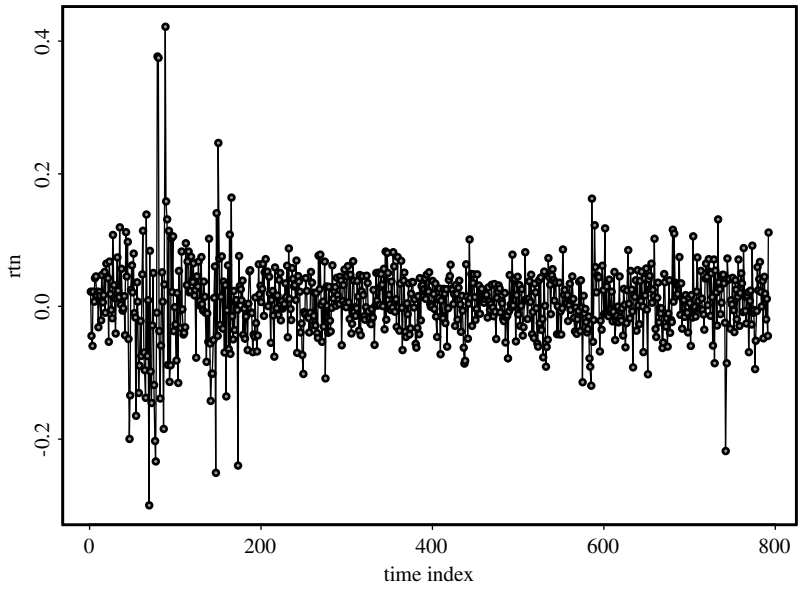


**Figure 3.5.** Time series plot of the monthly excess returns of S&P 500 index.

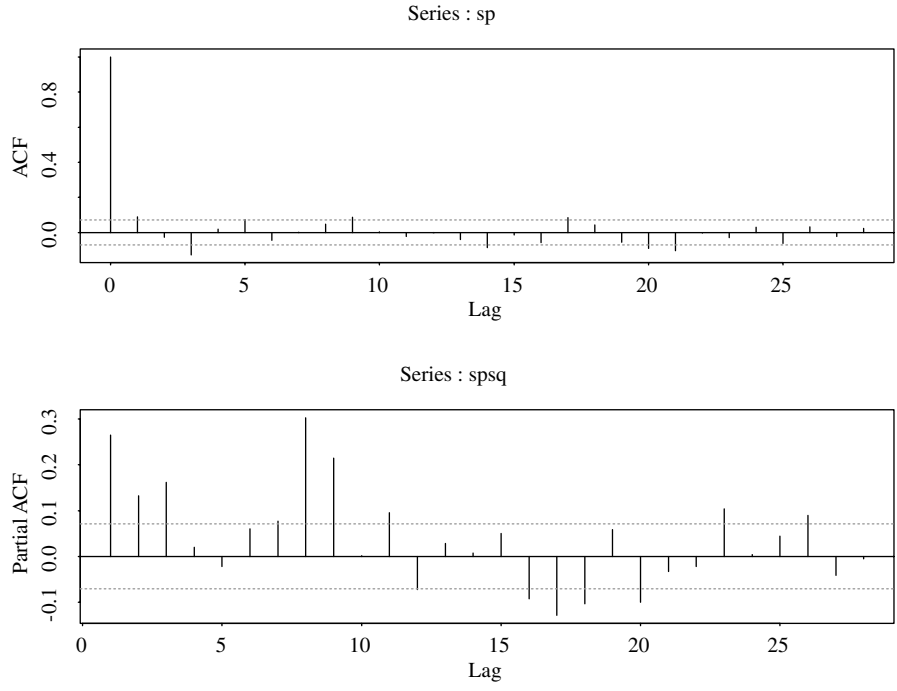


**Figure 3.6.** (a) Sample ACF of the monthly excess returns of S&P 500 index, and (b) sample PACF of the squared monthly excess returns.

The fitted AR(3) model, under the normality assumption, is

$$r_t = 0.088r_{t-1} - 0.023r_{t-2} - 0.123r_{t-3} + 0.0066 + a_t, \quad \hat{\sigma}_a^2 = 0.00333. \quad (3.17)$$

For the GARCH effects, we use the GARCH(1, 1) model

$$a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + \beta_1 \sigma_{t-1}^2 + \alpha_1 a_{t-1}^2.$$

A joint estimation of the AR(3)-GARCH(1, 1) model gives

$$\begin{aligned} r_t &= 0.021r_{t-1} - 0.034r_{t-2} - 0.013r_{t-3} + 0.0085 + a_t \\ \sigma_t^2 &= 0.000099 + 0.8476\sigma_{t-1}^2 + 0.1219a_{t-1}^2. \end{aligned}$$

From the volatility equation, the implied unconditional variance of $a_t$ is

$$\frac{0.000099}{1 - 0.8476 - 0.1219} = 0.00325,$$

which is very close to that of Eq. (3.17). However, $t$ ratios of the parameters in the mean equation suggest that all AR coefficients are insignificant at the 5% level. Therefore, we refine the model by dropping all AR coefficients. The refined model is

$$r_t = 0.0065 + a_t, \quad \sigma_t^2 = 0.00014 + 0.8220\sigma_{t-1}^2 + 0.1352a_{t-1}^2. \quad (3.18)$$

The standard error of the parameter in the mean equation is 0.0015, whereas those of the parameters in the volatility equation are 0.00002, 0.0208, and 0.0166, respectively. The unconditional variance of $a_t$ is $0.0001/(1 - 0.822 - 0.1352) = 0.00324$. This is a simple stationary GARCH(1, 1) model. Figure 3.7 shows the estimated volatility process and the standardized shocks $\tilde{a}_t = a_t/\sigma_t$ for the GARCH(1, 1) model in Eq. (3.18). The $\tilde{a}_t$ series looks like a white noise process. Figure 3.8 provides the sample ACF of the standardized shocks $\tilde{a}_t$ and the squared process $\tilde{a}_t^2$. These ACFs fail to suggest any significant serial correlations in the two processes. More specifically, we have $Q(10) = 10.32(0.41)$ and $Q(20) = 22.66(0.31)$ for $\tilde{a}_t$, and $Q(10) = 8.83(0.55)$ and $Q(20) = 15.82(0.73)$ for $\tilde{a}_t^2$, where the number in parentheses is the $p$ value of the test statistic. Thus, the model appears to be adequate. Note that the fitted model shows $\hat{\alpha}_1 + \hat{\beta}_1 = 0.9572$, which is close to 1. This phenomenon is commonly observed in practice and it leads to imposing the constraint $\alpha_1 + \beta_1 = 1$ in a GARCH(1, 1) model, resulting in an integrated GARCH (or IGARCH) model; see Section 3.5.

Finally, to forecast the volatility of monthly excess returns of S&P500 index, we can use the volatility equation in Eq. (3.18). For instance, at the forecast origin $h$, we have $\sigma_{h+1}^2 = 0.00014 + 0.822\sigma_h^2 + 0.1352a_h^2$. The 1-step ahead forecast is then

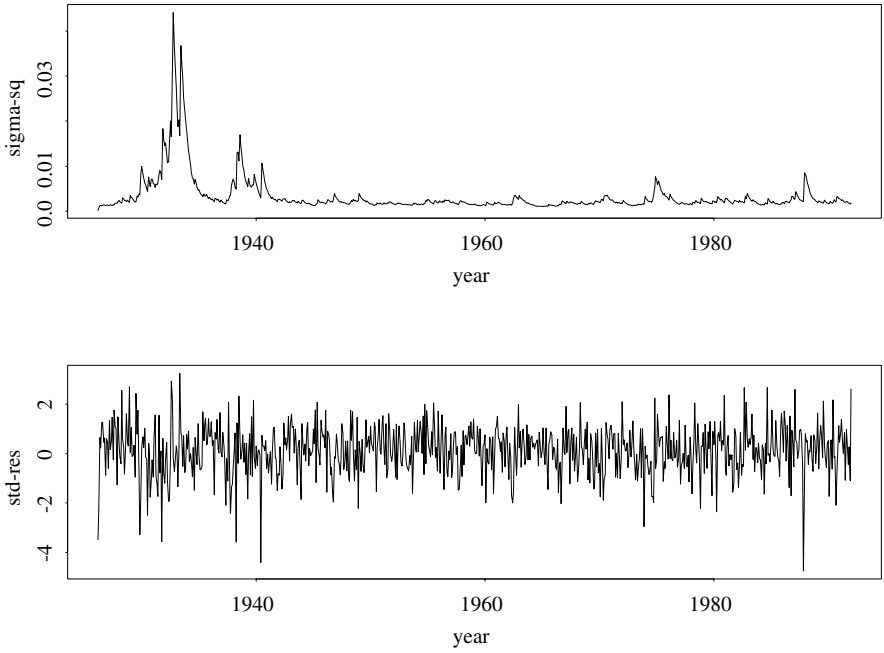


**Figure 3.7.** (a) Time series plot of estimated volatility for the monthly excess returns of S&P 500 index, and (b) the standardized shocks of the monthly excess returns of S&P 500 index. Both plots are based on the GARCH(1, 1) model in Eq. (3.18).

$$\sigma_h^2(1) = 0.00014 + 0.822\sigma_h^2 + 0.1352a_h^2,$$

where $a_h$ is the residual of the mean equation at time $h$ and $\sigma_h^2$ is obtained from the volatility equation. The starting value $\sigma_0^2$ is fixed at either zero or the unconditional variance of $a_t$. For multistep ahead forecasts, we use the recursive formula in Eq. (3.16). Table 3.1 shows some mean and volatility forecasts for the monthly excess return of S&P500 index with forecast origin $h = 792$ based on the GARCH(1, 1) model in Eq. (3.18).

***t* Innovation**

Assuming that $\epsilon_t$ follows a standardized Student-$t$ distribution with 5 degrees of freedom, we reestimate the GARCH(1, 1) model and obtain

$$r_t = 0.0085 + a_t, \quad \sigma_t^2 = 0.00018 + 0.1272a_{t-1}^2 + 0.8217\sigma_{t-1}^2, \tag{3.19}$$

where the standard errors of the parameters are 0.0014, $0.55\times10^{-4}$, 0.0349, and 0.0382, respectively. This model is essentially an IGARCH(1, 1) model as $\hat{\alpha}_1 + \hat{\beta}_1 \approx 0.96$, which is close to 1. The Ljung–Box statistics of the standardized residuals give $Q(10) = 10.45$ with $p$ value 0.40 and those of the $\{\tilde{a}_t^2\}$ series give $Q(10) = 9.33$

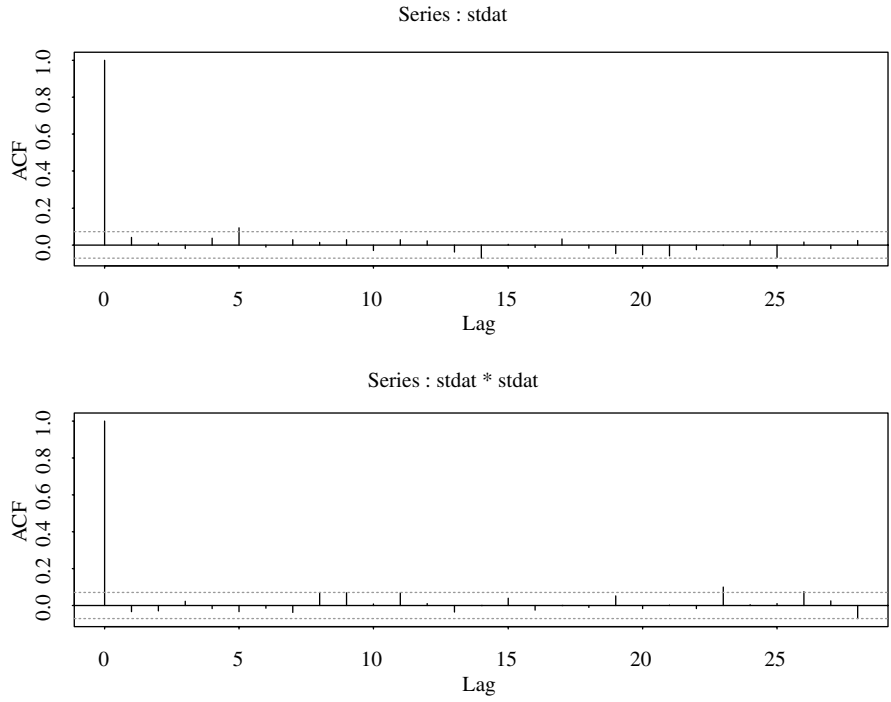


**Figure 3.8.** Model checking of the GARCH(1, 1) model in Eq. (3.18) for monthly excess returns of S&P 500 index: (a) Sample ACF of standardized shocks, and (b) sample ACF of the squared standardized shocks.

with $p$ value 0.50. Thus, the fitted GARCH(1, 1) model with Student-$t$ distribution is adequate.

***Estimation of Degrees of Freedom***

If we further extend the GARCH(1, 1) model by estimating the degrees of freedom of the Student-$t$ distribution used, we obtain the model

$$r_t = 0.0083 + a_t, \quad \sigma_t^2 = 0.00017 + 0.1227a_{t-1}^2 + 0.8193\sigma_{t-1}^2, \tag{3.20}$$

**Table 3.1. Volatility Forecasts for the Monthly Excess Returns of S&P500 Index. The Forecast Origin Is $h = 792$, Which Corresponds to December, 1991. Here Volatility Denotes Conditional Variance.**

| Horizon | 1 | 2 | 3 | 4 | 5 | $\infty$ |
|---|---|---|---|---|---|---|
| Return | 0.0065 | 0.0065 | 0.0065 | 0.0065 | 0.0065 | 0.0065 |
| Volatility | 0.00311 | 0.00312 | 0.00312 | 0.00313 | 0.00314 | 0.00324 |

where the estimated degrees of freedom is 6.51. Standard errors of the estimates in Eq. (3.20) are close to those in Eq. (3.19). The standard error of the estimated degrees of freedom is 1.49. Consequently, we cannot reject the hypothesis of using a standardized Student-$t$ distribution with 5 degrees of freedom at the 5% significance level.

### 3.4.2 Forecasting Evaluation

Since the volatility of an asset return is not directly observable, comparing the forecasting performance of different volatility models is a challenge to data analysts. In the literature, some researchers use out-of-sample forecasts and compare the volatility forecasts $\sigma_h^2(\ell)$ with the shock $a_{h+\ell}^2$ in the forecasting sample to assess the forecasting performance of a volatility model. This approach often finds a low correlation coefficient between $a_{h+\ell}^2$ and $\sigma_h^2(\ell)$. However, such a finding is not surprising because $a_{h+\ell}^2$ alone is not an adequate measure of the volatility at time index $h + \ell$. Consider the 1-step ahead forecasts. From a statistical point of view, $E(a_{h+1}^2 \mid F_h) = \sigma_{h+1}^2$ so that $a_{h+1}^2$ is a consistent estimate of $\sigma_{h+1}^2$. But it is not an accurate estimate of $\sigma_{h+1}^2$ because a single observation of a random variable with a known mean value cannot provide an accurate estimate of its variance. Consequently, such an approach to evaluate forecasting performance of volatility models is strictly speaking not proper. For more information concerning forecasting evaluation of GARCH models, readers are referred to Andersen and Bollerslev (1998).

## 3.5 THE INTEGRATED GARCH MODEL

If the AR polynomial of the GARCH representation in Eq. (3.14) has a unit root, then we have an IGARCH model. Thus, IGARCH models are unit-root GARCH models. Similar to ARIMA models, a key feature of IGARCH models is that the impact of past squared shocks $\eta_{t-i} = a_{t-i}^2 - \sigma_{t-i}^2$ for $i > 0$ on $a_t^2$ is persistent.

An IGARCH(1, 1) model can be written as

$$a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + \beta_1 \sigma_{t-1}^2 + (1 - \beta_1) a_{t-1}^2,$$

where $\{\epsilon_t\}$ is defined as before and $1 > \beta_1 > 0$. For the monthly excess returns of S&P 500 index, an estimated IGARCH(1, 1) model is

$$\begin{aligned} r_t &= 0.0067 + a_t, \quad a_t = \sigma_t \epsilon_t \\ \sigma_t^2 &= 0.000119 + 0.8059 \sigma_{t-1}^2 + 0.1941 a_{t-1}^2, \end{aligned}$$

where the standard errors of the estimates in the volatility equation are 0.0017, 0.000013, and 0.0144, respectively. The parameter estimates are close to those of the GARCH(1, 1) model shown before, but there is a major difference between the

two models. The unconditional variance of $a_t$, hence that of $r_t$, is not defined under the above IGARCH(1, 1) model. This seems hard to justify for an excess return series. From a theoretical point of view, the IGARCH phenomenon might be caused by occasional level shifts in volatility. The actual cause of persistence in volatility deserves a careful investigation.

When $\alpha_1 + \beta_1 = 1$, repeated substitutions in Eq. (3.16) give

$$\sigma_h^2(\ell) = \sigma_h^2(1) + (\ell - 1)\alpha_0, \quad \ell \geq 1,$$

where $h$ is the forecast origin. Consequently, the effect of $\sigma_h^2(1)$ on future volatilities is also persistent, and the volatility forecasts form a straight line with slope $\alpha_0$. Nelson (1990) studies some probability properties of the volatility process $\sigma_t^2$ under an IGARCH model. The process $\sigma_t^2$ is a martingale for which some nice results are available in the literature. Under certain conditions, the volatility process is strictly stationary, but not weakly stationary because it does not have the first two moments.

The case of $\alpha_0 = 0$ is of particular interest in studying the IGARCH(1, 1) model. In this case, the volatility forecasts are simply $\sigma_h^2(1)$ for all forecast horizons. This special IGARCH(1, 1) model is the volatility model used in RiskMetrics, which is an approach for calculating Value at Risk; see Chapter 7.

## 3.6 THE GARCH-M MODEL

In finance, the return of a security may depend on its volatility. To model such a phenomenon, one may consider the GARCH-M model, where "M" stands for GARCH *in mean*. A simple GARCH(1, 1)-M model can be written as

$$\begin{aligned} r_t &= \mu + c\sigma_t^2 + a_t, \quad a_t = \sigma_t \epsilon_t, \\ \sigma_t^2 &= \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \end{aligned} \tag{3.21}$$

where $\mu$ and $c$ are constant. The parameter $c$ is called the risk premium parameter. A positive $c$ indicates that the return is positively related to its past volatility. Other specifications of risk premium have also been used in the literature, including $r_t = \mu + c\sigma_t + a_t$.

The formulation of the GARCH-M model in Eq. (3.21) implies that there are serial correlations in the return series $r_t$. These serial correlations are introduced by those in the volatility process $\{\sigma_t^2\}$. The existence of risk premium is, therefore, another reason that some historical stock returns have serial correlations.

For illustration, we consider a GARCH(1, 1)-M model for the monthly excess returns of S&P 500 index from January 1926 to December 1991. The fitted model is

$$r_t = 0.0028 + 1.99\sigma_t^2 + a_t, \quad \sigma_t^2 = 0.00016 + 0.1328a_{t-1}^2 + 0.8137\sigma_{t-1}^2,$$

where the standard errors for the two parameters in the mean equation are 0.0022 and 0.7425, respectively, and those for the parameters in the volatility equation are

0.00002, 0.0220, and 0.0192, respectively. The estimated risk premium for the index return is positive and significant at the 5% level. The idea of risk premium applies to other GARCH models.

## 3.7 THE EXPONENTIAL GARCH MODEL

To overcome some weaknesses of the GARCH model in handling financial time series, Nelson (1991) proposes the exponential GARCH (EGARCH) model. In particular, to allow for asymmetric effects between positive and negative asset returns, he considers the weighted innovation

$$g(\epsilon_t) = \theta\epsilon_t + \gamma[|\,\epsilon_t\,| - E(|\,\epsilon_t\,|)], \tag{3.22}$$

where $\theta$ and $\gamma$ are real constants. Both $\epsilon_t$ and $|\,\epsilon_t\,| - E(|\,\epsilon_t\,|)$ are zero-mean iid sequences with continuous distributions. Therefore, $E[g(\epsilon_t)] = 0$. The asymmetry of $g(\epsilon_t)$ can easily be seen by rewriting it as

$$g(\epsilon_t) = \begin{cases} (\theta + \gamma)\epsilon_t - \gamma E(|\,\epsilon_t\,|) & \text{if } \epsilon_t \geq 0, \\ (\theta - \gamma)\epsilon_t - \gamma E(|\,\epsilon_t\,|) & \text{if } \epsilon_t < 0. \end{cases}$$

***Remark:*** For the standard Gaussian random variable $\epsilon_t$, $E(|\,\epsilon_t\,|) = \sqrt{2/\pi}$. For the standardized Student-$t$ distribution in Eq. (3.7), we have

$$E(|\,\epsilon_t\,|) = \frac{2\sqrt{v-2}\Gamma((v+1)/2)}{(v-1)\Gamma(v/2)\sqrt{\pi}}.$$

An EGARCH($m$, $s$) model can be written as

$$a_t = \sigma_t\epsilon_t, \quad \ln(\sigma_t^2) = \alpha_0 + \frac{1 + \beta_1 B + \cdots + \beta_s B^s}{1 - \alpha_1 B - \cdots - \alpha_m B^m} g(\epsilon_{t-1}), \tag{3.23}$$

where $\alpha_0$ is a constant, $B$ is the back-shift (or lag) operator such that $Bg(\epsilon_t) = g(\epsilon_{t-1})$, and $1 + \beta_1 B + \cdots + \beta_s B^s$ and $1 - \alpha_1 B - \cdots - \alpha_m B^m$ are polynomials with zeros outside the unit circle and have no common factors. By outside the unit circle, we mean that absolute values of the zeros are greater than 1. Again, Eq. (3.23) uses the usual ARMA parameterization to describe the evolution of the conditional variance of $a_t$. Based on this representation, some properties of the EGARCH model can be obtained in a similar manner as those of the GARCH model. For instance, the unconditional mean of $\ln(\sigma_t^2)$ is $\alpha_0$. However, the model differs from the GARCH model in several ways. First, it uses logged conditional variance to relax the positiveness constraint of model coefficients. Second, the use of $g(\epsilon_t)$ enables the model to respond asymmetrically to positive and negative lagged values of $a_t$. Some additional properties of the EGARCH model can be found in Nelson (1991).

To better understand the EGARCH model, let us consider the simple model with order $(1, 0)$

$$a_t = \sigma_t \epsilon_t, \qquad (1 - \alpha B) \ln(\sigma_t^2) = (1 - \alpha)\alpha_0 + g(\epsilon_{t-1}), \tag{3.24}$$

where $\epsilon_t$s are iid standard normal and the subscript of $\alpha_1$ is omitted. In this case, $E(|\,\epsilon_t\,|) = \sqrt{2/\pi}$ and the model for $\ln(\sigma_t^2)$ becomes

$$(1 - \alpha B) \ln(\sigma_t^2) = \begin{cases} \alpha_* + (\theta + \gamma)\epsilon_{t-1} & \text{if } \epsilon_{t-1} \geq 0, \\ \alpha_* + (\theta - \gamma)\epsilon_{t-1} & \text{if } \epsilon_{t-1} < 0, \end{cases} \tag{3.25}$$

where $\alpha_* = (1 - \alpha)\alpha_0 - \sqrt{2/\pi}\gamma$. This is a nonlinear function similar to that of the threshold autoregressive model (TAR) of Tong (1978, 1990). It suffices to say that for this simple EGARCH model the conditional variance evolves in a nonlinear manner depending on the sign of $a_{t-1}$. Specifically, we have

$$\sigma_t^2 = \sigma_{t-1}^{2\alpha} \exp(\alpha_*) \begin{cases} \exp\left[(\theta + \gamma)\frac{a_{t-1}}{\sqrt{\sigma_{t-1}^2}}\right] & \text{if } a_{t-1} \geq 0, \\ \exp\left[(\theta - \gamma)\frac{a_{t-1}}{\sqrt{\sigma_{t-1}^2}}\right] & \text{if } a_{t-1} < 0. \end{cases}$$

The coefficients $(\theta + \gamma)$ and $(\theta - \gamma)$ show the asymmetry in response to positive and negative $a_{t-1}$. The model is, therefore, nonlinear if $\gamma \neq 0$. For higher order EGARCH models, the nonlinearity becomes much more complicated. Cao and Tsay (1992) use nonlinear models, including EGARCH models, to obtain multistep ahead volatility forecasts. We discuss nonlinearity in financial time series in Chapter 4.

### 3.7.1 An Illustrative Example

Nelson (1991) applies an EGARCH model to the daily excess returns of the value-weighted market index from the Center for Research in Security Prices from July 1962 to December 1987. The excess returns are obtained by removing monthly Treasury bill returns from the value-weighted index returns, assuming that the Treasury bill return was constant for each calendar day within a given month. There are 6408 observations. Denote the excess return by $r_t$. The model used is as follows:

$$\begin{aligned} r_t &= \phi_0 + \phi_1 r_{t-1} + c\sigma_t^2 + a_t \\ \ln(\sigma_t^2) &= \alpha_0 + \ln(1 + wN_t) + \frac{1 + \beta B}{1 - \alpha_1 B - \alpha_2 B^2} g(\epsilon_{t-1}), \end{aligned} \tag{3.26}$$

where $\sigma_t^2$ is the conditional variance of $a_t$ given $F_{t-1}$, $N_t$ is the number of nontrading days between trading days $t - 1$ and $t$, $\alpha_0$ and $w$ are real parameters, $g(\epsilon_t)$ is defined in Eq. (3.22), and $\epsilon_t$ follows a generalized error distribution with probability density

**Table 3.2. An Estimated AR(1)-EGARCH(2, 1) Model for the Daily Excess Returns of the Value-Weighted CRSP Market Index: July 1962 to December 1987.**

| Par. | $\alpha_0$ | $w$ | $\gamma$ | $\alpha_1$ | $\alpha_2$ | $\beta$ |
|---|---|---|---|---|---|---|
| Est. | −10.06 | .183 | .156 | 1.929 | −.929 | −.978 |
| Err. | .346 | .028 | .013 | .015 | .015 | .006 |
| Par. | $\theta$ | $\phi_0$ | $\phi_1$ | $c$ | $v$ | |
| Est. | −.118 | $3.5 \cdot 10^{-4}$ | .205 | −3.361 | 1.576 | |
| Err. | .009 | $9.9 \cdot 10^{-5}$ | .012 | 2.026 | .032 | |

function

$$f(x) = \frac{v \exp[-(1/2)|x/\lambda|^v]}{\lambda 2^{(1+1/v)}\Gamma(1/v)}, \qquad -\infty < x < \infty, \quad 0 < v \leq \infty,$$

where again $\Gamma(.)$ is the gamma function and

$$\lambda = [2^{(-2/v)}\Gamma(1/v)/\Gamma(3/v)]^{1/2}.$$

Similar to a GARCH-M model, the parameter $c$ in Eq. (3.26) is the risk premium parameter. Table 3.2 gives the parameter estimates and their standard errors of the model. The mean equation of model (3.26) has two features that are of interest. First, it uses an AR(1) model to take care of possible serial correlation in the excess returns. Second, it uses the volatility $\sigma_t^2$ as a regressor to account for risk premium. The estimated risk premium is negative, but statistically insignificant.

### 3.7.2 Another Example

As another illustration, we consider the monthly log returns of IBM stock from January 1926 to December 1997 for 864 observations. An AR(1)-EGARCH(1, 0) model is entertained and the fitted model is

$$r_t = 0.0105 + 0.092 r_{t-1} + a_t, \quad a_t = \sigma_t \epsilon_t \tag{3.27}$$

$$\ln(\sigma_t^2) = -5.496 + \frac{g(\epsilon_{t-1})}{1 - 0.856B},$$

$$g(\epsilon_{t-1}) = -0.0795\epsilon_{t-1} + 0.2647\left[|\epsilon_{t-1}| - \sqrt{2/\pi}\right], \tag{3.28}$$

where $\{\epsilon_t\}$ is a sequence of independent standard Gaussian random variates. All parameter estimates are statistically significant at the 5% level. For model checking, the Ljung–Box statistics give $Q(10) = 6.31(0.71)$ and $Q(20) = 21.4(0.32)$ for the standardized residual process $\tilde{a}_t = a_t/\sigma_t$ and $Q(10) = 4.13(0.90)$ and $Q(20) = 15.93(0.66)$ for the squared process $\tilde{a}_t^2$, where again the number in parentheses denotes $p$ value. Therefore, there is no serial correlation or conditional het-

eroscedasticity in the standardized residuals of the fitted model. The prior AR(1)-EGARCH(1, 0) model is adequate.

From the estimated volatility equation in (3.28) and using $\sqrt{2/\pi} \approx 0.7979$, we obtain the volatility equation as

$$\ln(\sigma_t^2) = -1.001 + 0.856 \ln(\sigma_{t-1}^2) + \begin{cases} 0.1852\epsilon_{t-1} & \text{if } \epsilon_{t-1} \geq 0 \\ -0.3442\epsilon_{t-1} & \text{if } \epsilon_{t-1} < 0. \end{cases}$$

Taking antilog transformation, we have

$$\sigma_t^2 = \sigma_{t-1}^{2\times 0.856} e^{-1.001} \times \begin{cases} e^{0.1852\epsilon_{t-1}} & \text{if } \epsilon_{t-1} \geq 0 \\ e^{-0.3442\epsilon_{t-1}} & \text{if } \epsilon_{t-1} < 0. \end{cases}$$

This equation highlights the asymmetric responses in volatility to the past positive and negative shocks under an EGARCH model. For example, for a standardized shock with magnitude 2 (i.e., two standard deviations), we have

$$\frac{\sigma_t^2(\epsilon_{t-1} = -2)}{\sigma_t^2(\epsilon_{t-1} = 2)} = \frac{\exp[-0.3442 \times (-2)]}{\exp(0.1852 \times 2)} = e^{0.318} = 1.374.$$

Therefore, the impact of a negative shock of size two standard deviations is about 37.4% higher than that of a positive shock of the same size. This example clearly demonstrates the asymmetric feature of EGARCH models. In general, the bigger the shock, the larger the difference in volatility impact.

### 3.7.3 Forecasting Using an EGARCH Model

We use the EGARCH(1, 0) model to illustrate multistep ahead forecasts of EGARCH models, assuming that the model parameters are known and the innovations are standard Gaussian. For such a model, we have

$$\ln(\sigma_t^2) = (1 - \alpha_1)\alpha_0 + \alpha_1 \ln(\sigma_{t-1}^2) + g(\epsilon_{t-1}),$$
$$g(\epsilon_{t-1}) = \theta\epsilon_{t-1} + \gamma(|\epsilon_{t-1}| - \sqrt{2/\pi}).$$

Taking exponentials, the model becomes

$$\sigma_t^2 = \sigma_{t-1}^{2\alpha_1} \exp[(1 - \alpha_1)\alpha_0] \exp[g(\epsilon_{t-1})],$$
$$g(\epsilon_{t-1}) = \theta\epsilon_{t-1} + \gamma(|\epsilon_{t-1}| - \sqrt{2/\pi}). \tag{3.29}$$

Let $h$ be the forecast origin. For the 1-step ahead forecast, we have

$$\sigma_{h+1}^2 = \sigma_h^{2\alpha_1} \exp[(1 - \alpha_1)\alpha_0] \exp[g(\epsilon_h)],$$

where all of the quantities on the right-hand side are known. Thus, the 1-step ahead volatility forecast at the forecast origin $h$ is simply $\hat{\sigma}_h^2(1) = \sigma_{h+1}^2$ given earlier. For

the 2-step ahead forecast, Eq. (3.29) gives

$$\sigma_{h+2}^2 = \sigma_{h+1}^{2\alpha_1} \exp[(1-\alpha_1)\alpha_0] \exp[g(\epsilon_{h+1})].$$

Taking conditional expectation at time $h$, we have

$$\hat{\sigma}_h^2(2) = \hat{\sigma}_h^{2\alpha_1}(1) \exp[(1-\alpha_1)\alpha_0] E_h\{\exp[g(\epsilon_{h+1})]\},$$

where $E_h$ denotes a conditional expectation taking at the time origin $h$. The prior expectation can be obtained as follows:

$$\begin{aligned} E\{\exp[g(\epsilon)]\} &= \int_{-\infty}^{\infty} \exp[\theta\epsilon + \gamma(|\epsilon| - \sqrt{2/\pi})] f(\epsilon) d\epsilon \\ &= \exp\left(-\gamma\sqrt{2/\pi}\right) \left[ \int_0^{\infty} e^{(\theta+\gamma)\epsilon} \frac{1}{\sqrt{2\pi}} e^{-\epsilon^2/2}\, d\epsilon \right. \\ &\quad \left. + \int_{-\infty}^{0} e^{(\theta-\gamma)\epsilon} \frac{1}{\sqrt{2\pi}} e^{-\epsilon^2/2}\, d\epsilon \right] \\ &= \exp\left(-\gamma\sqrt{2/\pi}\right) \left[ e^{(\theta+\gamma)^2/2} \Phi(\theta+\gamma) + e^{(\theta-\gamma)^2/2} \Phi(\gamma-\theta) \right], \end{aligned}$$

where $f(\epsilon)$ and $\Phi(x)$ are the probability density function and CDF of the standard normal distribution, respectively. Consequently, the 2-step ahead volatility forecast is

$$\begin{aligned} \hat{\sigma}_h^2(2) &= \hat{\sigma}_h^{2\alpha_1}(1) \exp\left[(1-\alpha_1)\alpha_0 - \gamma\sqrt{2/\pi}\right] \\ &\quad \times \left[ \exp\{(\theta+\gamma)^2/2\} \Phi(\theta+\gamma) + \exp\{(\theta-\gamma)^2/2\} \Phi(\gamma-\theta) \right]. \end{aligned}$$

Repeating the previous procedure, we obtain a recursive formula for $j$-step ahead forecast

$$\begin{aligned} \hat{\sigma}_h^2(j) &= \hat{\sigma}_h^{2\alpha_1}(j-1) \exp(\omega) \\ &\quad \times \left[ \exp\{(\theta+\gamma)^2/2\} \Phi(\theta+\gamma) + \exp\{(\theta-\gamma)^2/2\} \Phi(\gamma-\theta) \right], \end{aligned}$$

where $\omega = (1-\alpha_1)\alpha_0 - \gamma\sqrt{2/\pi}$. The values of $\Phi(\theta+\gamma)$ and $\Phi(\theta-\gamma)$ can be obtained from most statistical packages. Alternatively, accurate approximations to these values can be obtained by using the method in Appendix B of Chapter 6.

For illustration, consider the AR(1)-EGARCH(1, 0) model of the previous subsection for the monthly log returns of IBM stock. Using the fitted EGARCH(1, 0) model, we can compute the volatility forecasts for the series. At the forecast origin $t = 864$, the forecasts are $\hat{\sigma}_{864}^2(1) = 6.05 \times 10^{-3}$, $\hat{\sigma}_{864}^2(2) = 5.82 \times 10^{-3}$,

$\hat{\sigma}^2_{864}(3) = 5.63 \times 10^{-3}$, and $\hat{\sigma}^2_{864}(10) = 4.94 \times 10^{-3}$. These forecasts converge gradually to the sample variance $4.37 \times 10^{-3}$ of the shock process $a_t$ of Eq. (3.27).

## 3.8 THE CHARMA MODEL

Many other econometric models have been proposed in the literature to describe the evolution of the conditional variance $\sigma_t^2$ in Eq. (3.2). We mention the conditional heteroscedastic ARMA (CHARMA) model that uses random coefficients to produce conditional heteroscedasticity; see Tsay (1987). The CHARMA model is not the same as the ARCH model, but the two models have similar second-order conditional properties. A CHARMA model is defined as

$$r_t = \mu_t + a_t, \quad a_t = \delta_{1t} a_{t-1} + \delta_{2t} a_{t-2} + \cdots + \delta_{mt} a_{t-m} + \eta_t, \tag{3.30}$$

where $\{\eta_t\}$ is a Gaussian white noise series with mean zero and variance $\sigma_\eta^2$, $\{\boldsymbol{\delta}_t\} = \{(\delta_{1t}, \ldots, \delta_{mt})'\}$ is a sequence of iid random vectors with mean zero and non-negative definite covariance matrix $\boldsymbol{\Omega}$, and $\{\boldsymbol{\delta}_t\}$ is independent of $\{\eta_t\}$. In this section, we use some basic properties of vector and matrix operations to simplify the presentation. Readers may consult Appendix A of Chapter 8 for a brief review of these properties. For $m > 0$, the model can be written as

$$a_t = \boldsymbol{a}'_{t-1} \boldsymbol{\delta}_t + \eta_t,$$

where $\boldsymbol{a}_t = (a_{t-1}, \ldots, a_{t-m})'$ is a vector of lagged values of $a_t$ and is available at time $t-1$. The conditional variance of $a_t$ of the CHARMA model in Eq. (3.30) is then

$$\begin{aligned} \sigma_t^2 &= \sigma_\eta^2 + \boldsymbol{a}'_{t-1} \operatorname{Cov}(\boldsymbol{\delta}_t) \boldsymbol{a}_{t-1} \\ &= \sigma_\eta^2 + (a_{t-1}, \ldots, a_{t-m}) \boldsymbol{\Omega} (a_{t-1}, \ldots, a_{t-m})'. \end{aligned} \tag{3.31}$$

Denote the $(i, j)$th element of $\boldsymbol{\Omega}$ by $\omega_{ij}$. Because the matrix is symmetric, we have $\omega_{ij} = \omega_{ji}$. If $m = 1$, then Eq. (3.31) reduces to $\sigma_t^2 = \sigma_\eta^2 + \omega_{11} a_{t-1}^2$, which is an ARCH(1) model. If $m = 2$, then Eq. (3.31) reduces to

$$\sigma_t^2 = \sigma_\eta^2 + \omega_{11} a_{t-1}^2 + 2\omega_{12} a_{t-1} a_{t-2} + \omega_{22} a_{t-2}^2,$$

which differs from an ARCH(2) model by the cross-product term $a_{t-1} a_{t-2}$. In general, the conditional variance of a CHARMA($m$) model is equivalent to that of an ARCH($m$) model if $\boldsymbol{\Omega}$ is a diagonal matrix. Because $\boldsymbol{\Omega}$ is a covariance matrix, which is non-negative definite, and $\sigma_\eta^2$ is a variance, which is positive, we have $\sigma_t^2 \geq \sigma_\eta^2 > 0$ for all $t$. In other words, the positiveness of $\sigma_t^2$ is automatically satisfied under a CHARMA model.

An obvious difference between ARCH and CHARMA models is that the latter use cross-products of the lagged values of $a_t$ in the volatility equation. The cross-product terms might be useful in some applications. For example, in modeling an asset return series, cross-product terms denote interactions between previous returns. It is conceivable that stock volatility may depend on such interactions. However, the number of cross-product terms increases rapidly with the order $m$, and some constraints are needed to keep the model simple. A possible constraint is to use a small number of cross-product terms in a CHARMA model. Another difference between the two models is that higher order properties of CHARMA models are harder to obtain than those of ARCH models because it is harder to handle random coefficients than constant coefficients.

For illustration, we employ the CHARMA model

$$r_t = \phi_0 + a_t, \qquad a_t = \delta_{1t} a_{t-1} + \delta_{2t} a_{t-2} + \eta_t$$

for the monthly excess returns of S&P 500 index used before in GARCH modeling. The fitted model is

$$r_t = 0.00635 + a_t, \quad \sigma_t^2 = 0.00179 + (a_{t-1}, a_{t-2})\widehat{\mathbf{\Omega}}(a_{t-1}, a_{t-2})',$$

where

$$\widehat{\mathbf{\Omega}} = \begin{bmatrix} 0.1417(0.0333) & -0.0594(0.0365) \\ -0.0594(0.0365) & 0.3081(0.0340) \end{bmatrix}$$

and the numbers in parentheses are standard errors. The cross-product term of $\widehat{\mathbf{\Omega}}$ has a $t$ ratio of $-1.63$, which is marginally significant at the 10% level. If we refine the model to

$$r_t = \phi_0 + a_t, \qquad a_t = \delta_{1t} a_{t-1} + \delta_{2t} a_{t-2} + \delta_{3t} a_{t-3} + \eta_t,$$

but assume that $\delta_{3t}$ is uncorrelated with $(\delta_{1t}, \delta_{2t})$, then we obtain the fitted model

$$r_t = 0.0068 + a_t, \quad \sigma_t^2 = .00136 + (a_{t-1}, a_{t-2}, a_{t-3})\widehat{\mathbf{\Omega}}(a_{t-1}, a_{t-2}, a_{t-3})',$$

where the elements of $\widehat{\mathbf{\Omega}}$ and their standard errors, shown in parentheses, are

$$\hat{\mathbf{\Sigma}} = \begin{bmatrix} 0.1212(.0355) & -0.0622(.0283) & 0 \\ -0.0622(.0283) & 0.1913(.0254) & 0 \\ 0 & 0 & 0.2988(0.0420) \end{bmatrix}.$$

All of the estimates are now statistically significant at the 5% level. From the model, $a_t = r_t - 0.0068$ is the deviation of the monthly excess return from its average. The fitted CHARMA model shows that there is some interaction effect between the first

two lagged deviations. Indeed, the volatility equation can be written approximately as

$$\sigma_t^2 = 0.00136 + 0.12a_{t-1}^2 - 0.12a_{t-1}a_{t-2} + 0.19a_{t-2}^2 + 0.30a_{t-3}^2.$$

The conditional variance is slightly larger when $a_{t-1}a_{t-2}$ is negative.

### 3.8.1 Effects of Explanatory Variables

The CHARMA model can easily be generalized so that the volatility of $r_t$ may depend on some explanatory variables. Let $\{x_{it}\}_{i=1}^m$ be $m$ explanatory variables available at time $t$. Consider the model

$$r_t = \mu_t + a_t, \qquad a_t = \sum_{i=1}^{m} \delta_{it} x_{i,t-1} + \eta_t, \tag{3.32}$$

where $\boldsymbol{\delta}_t = (\delta_{1t}, \ldots, \delta_{mt})'$ and $\eta_t$ are random vector and variable defined in Eq. (3.30). Then the conditional variance of $a_t$ is

$$\sigma_t^2 = \sigma_\eta^2 + (x_{1,t-1}, \ldots, x_{m,t-1})\boldsymbol{\Omega}(x_{1,t-1}, \ldots, x_{m,t-1})'.$$

In application, the explanatory variables may include some lagged values of $a_t$.

## 3.9 RANDOM COEFFICIENT AUTOREGRESSIVE MODELS

In the literature, the random coefficient autoregressive (RCA) model is introduced to account for variability among different subjects under study, similar to the panel data analysis in econometrics and the hierarchical model in statistics. We classify the RCA model as a conditional heteroscedastic model, but historically it is used to obtain a better description of the conditional mean equation of the process by allowing for the parameters to evolve over time. A time series $r_t$ is said to follow an RCA($p$) model if it satisfies

$$r_t = \phi_0 + \sum_{i=1}^{p} (\phi_i + \delta_{it}) r_{t-i} + a_t, \tag{3.33}$$

where $p$ is a positive integer, $\{\boldsymbol{\delta}_t\} = \{(\delta_{1t}, \ldots, \delta_{pt})'\}$ is a sequence of independent random vectors with mean zero and covariance matrix $\boldsymbol{\Omega}_\delta$, and $\{\boldsymbol{\delta}_t\}$ is independent of $\{a_t\}$; see Nicholls and Quinn (1982) for further discussions of the model. The conditional mean and variance of the RCA model in Eq. (3.33) are

$$\mu_t = E(a_t \mid F_{t-1}) = \sum_{i=1}^{p} \phi_i a_{t-i}, \quad \sigma_t^2 = \sigma_a^2 + (r_{t-1}, \ldots, r_{t-p})\boldsymbol{\Omega}_\delta (r_{t-1}, \ldots, r_{t-p})',$$

which is in the same form as that of a CHARMA model. However, there is a subtle difference between RCA and CHARMA models. For the RCA model, the volatility is a quadratic function of the observed lagged values $r_{t-i}$. Yet the volatility is a quadratic function of the lagged innovations $a_{t-i}$ in a CHARMA model.

## 3.10 THE STOCHASTIC VOLATILITY MODEL

An alternative approach to describe the volatility evolution of a financial time series is to introduce an innovation to the conditional variance equation of $a_t$; see Melino and Turnbull (1990), Harvey, Ruiz, and Shephard (1994) and Jacquier, Polson, and Rossi (1994). The resulting model is referred to as a stochastic volatility (SV) model. Similar to EGARCH models, to ensure positiveness of the conditional variance, SV models use $\ln(\sigma_t^2)$ instead of $\sigma_t^2$. A SV model is defined as

$$a_t = \sigma_t \epsilon_t, \qquad (1 - \alpha_1 B - \cdots - \alpha_m B^m) \ln(\sigma_t^2) = \alpha_0 + v_t, \tag{3.34}$$

where $\epsilon_t$s are iid $N(0, 1)$, $v_t$s are iid $N(0, \sigma_v^2)$, $\{\epsilon_t\}$ and $\{v_t\}$ are independent, $\alpha_0$ is a constant, and all zeros of the polynomial $1 - \sum_{i=1}^{m} \alpha_i B^i$ are greater 1 in modulus. Introducing the innovation $v_t$ substantially increases the flexibility of the model in describing the evolution of $\sigma_t^2$, but it also increases the difficulty in parameter estimation. To estimate a SV model, we need a quasi-likelihood method via Kalman filtering or a Monte Carlo method. Jacquier, Polson, and Rossi (1994) provide some comparison of estimation results between quasi-likelihood and Monte Carlo Markov Chain (MCMC) methods. The difficulty in estimating a SV model is understandable because for each shock $a_t$ the model uses two innovations $\epsilon_t$ and $v_t$. We discuss a MCMC method to estimate SV models in Chapter 10. For more discussions on stochastic volatility models, see Taylor (1994).

The appendixes of Jacquier, Polson, and Rossi (1994) provide some properties of the SV model when $m = 1$. For instance, with $m = 1$, we have

$$\ln(\sigma_t^2) \sim N\left(\frac{\alpha_0}{1-\alpha_1}, \frac{\sigma_v^2}{1-\alpha_1^2}\right) \equiv N(\mu_h, \sigma_h^2),$$

and $E(a_t^2) = \exp[\mu_h + 1/(2\sigma_h^2)]$, $E(a_t^4) = 3\exp[2\mu_h^2 + 2\sigma_h^2]$, and $\text{corr}(a_t^2, a_{t-i}^2) = [\exp(\sigma_h^2 \alpha_1^i) - 1]/[3\exp(\sigma_h^2) - 1]$. Limited experience shows that SV models often provided improvements in model fitting, but their contributions to out-of-sample volatility forecasts received mixed results.

## 3.11 THE LONG-MEMORY STOCHASTIC VOLATILITY MODEL

More recently, the SV model is further extended to allow for long memory in volatility, using the idea of fractional difference. As stated in Chapter 2, a time series is a

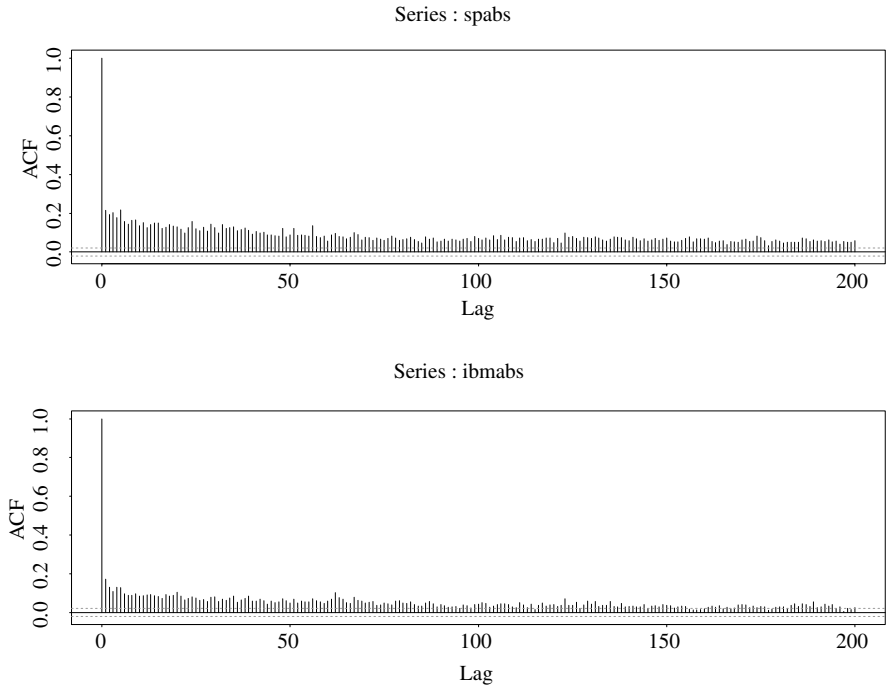


**Figure 3.9.** The sample ACF of daily absolute log returns for the S&P 500 index, and the IBM stock for the period from July 3, 1962 to December 31, 1997. The two horizontal lines denote the asymptotic 5% limits.

long-memory process if its autocorrelation function decays at a hyperbolic, instead of an exponential, rate as the lag increases. The extension to long-memory models in volatility study is motivated by the fact that autocorrelation function of the squared or absolute-valued series of an asset return often decays slowly, even though the return series has no serial correlation; see Ding, Granger, and Engle (1993). Figure 3.9 shows the sample ACF of the daily absolute returns for IBM stock and the S&P 500 index from July 3, 1962 to December 31, 1997. These sample ACFs are positive and of moderate magnitude, but decay slowly.

A simple long-memory stochastic volatility (LMSV) model can be written as

$$a_t = \sigma_t \epsilon_t, \qquad \sigma_t = \sigma \exp(u_t/2), \qquad (1-B)^d u_t = \eta_t, \tag{3.35}$$

where $\sigma > 0$, $\epsilon_t$s are iid $N(0,1)$, $\eta_t$s are iid $N(0, \sigma_\eta^2)$ and independent of $\epsilon_t$, and $0 < d < 0.5$. The feature of long memory stems from the fractional difference $(1-B)^d$, which implies that the ACF of $u_t$ decays slowly at a hyperbolic, instead of an exponential, rate as the lag increases. For model (3.35), we have

$$\begin{aligned}\ln(a_t^2) &= \ln(\sigma^2) + u_t + \ln(\epsilon_t^2) \\ &= [\ln(\sigma^2) + E(\ln \epsilon_t^2)] + u_t + [\ln(\epsilon_t^2) - E(\ln \epsilon_t^2)] \\ &\equiv \mu + u_t + e_t.\end{aligned}$$

Thus, the $\ln(a_t^2)$ series is a Gaussian long-memory signal plus a non-Gaussian white noise; see Breidt, Crato, and de Lima (1998). Estimation of the long-memory stochastic volatility model is complicated, but the fractional difference parameter $d$ can be estimated by using either a quasi-maximum likelihood method or a regression method. Using the log series of squared daily returns for companies in S&P 500 index, Bollerslev and Jubinski (1999) and Ray and Tsay (2000) found that the median estimate of $d$ is about 0.38. For applications, Ray and Tsay (2000) study common long-memory components in daily stock volatilities of groups of companies classified by various characteristics. They found that companies in the same industrial or business sector tend to have more common long-memory components (e.g., big U.S. national banks and financial institutions).

## 3.12 AN ALTERNATIVE APPROACH

French, Schwert, and Stambaugh (1987) consider an alternative approach for volatility estimation that uses high-frequency data to calculate volatility of low-frequency returns. In recent years, this approach has attracted some interest due in part to the availability of high-frequency financial data, especially in the foreign exchange markets (e.g., Andersen, Bollerslev, Diebold, and Labys, 1999).

Suppose that we are interested in the monthly volatility of an asset for which daily returns are available. Let $r_t^m$ be the monthly log return of the asset at month $t$. Assume that there are $n$ trading days in month $t$ and the daily log returns of the asset in the month are $\{r_{t,i}\}_{i=1}^n$. Using properties of log returns, we have

$$r_t^m = \sum_{i=1}^{n} r_{t,i}.$$

Assuming that the conditional variance and covariance exist, we have

$$\text{Var}(r_t^m \mid F_{t-1}) = \sum_{i=1}^{n} \text{Var}(r_{t,i} \mid F_{t-1}) + 2 \sum_{i<j} \text{Cov}[(r_{t,i}, r_{t,j}) \mid F_{t-1}], \qquad (3.36)$$

where $F_{t-1}$ denotes the information available at month $t-1$ (inclusive). The prior equation can be simplified if additional assumptions are made. For example, if we assume that $\{r_{t,i}\}$ is a white noise series, then

$$\text{Var}(r_t^m \mid F_{t-1}) = n \,\text{Var}(r_{t,1}),$$

where $\text{Var}(r_{t,1})$ can be estimated from the daily returns $\{r_{t,i}\}_{i=1}^{n}$ by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(r_{t,i} - \bar{r}_t)^n}{n-1},$$

where $\bar{r}_t$ is the sample mean of the daily log returns in month $t$ (i.e., $\bar{r}_t = \sum_{i=1}^{n} r_{t,i}/n$). The estimated monthly volatility is then

$$\hat{\sigma}_m^2 = \frac{n}{n-1}\sum_{i=1}^{n}(r_{t,i} - \bar{r}_t)^2. \tag{3.37}$$

If $\{r_{t,i}\}$ follows an MA(1) model, then

$$\text{Var}(r_t^m \mid F_{t-1}) = n\,\text{Var}(r_{t,1}) + 2(n-1)\,\text{Cov}(r_{t,1}, r_{t,2}),$$

which can be estimated by

$$\hat{\sigma}_m^2 = \frac{n}{n-1}\sum_{i=1}^{n}(r_{t,i} - \bar{r}_t)^2 + 2\sum_{i=1}^{n-1}(r_{t,i} - \bar{r}_t)(r_{t,i+1} - \bar{r}_t). \tag{3.38}$$

The previous approach for volatility estimation is simple, but it encounters several difficulties in practice. First, the model for daily returns $\{r_{t,i}\}$ is unknown. This complicates the estimation of covariances in Eq. (3.36). Second, there are roughly 21 trading days in a month, resulting in a small sample size. The accuracy of the estimates of variance and covariance in Eq. (3.36) might be questionable. The accuracy depends on the dynamic structure of $\{r_{t,i}\}$ and their distribution. If the daily log returns have high excess kurtosis and serial correlations, then the sample estimates $\hat{\sigma}_m^2$ in Eqs. (3.37) and (3.38) may not even be consistent; see Bai, Russell, and Tiao (2000). Further research is needed to make this approach valuable.

**Example 3.4.** Consider the monthly volatility of the log returns of S&P 500 index from January 1980 to December 1999. We calculate the volatility by three methods. In the first method, we use daily log returns and Eq. (3.37) (i.e., assuming that the daily log returns form a white noise series). The second method also uses daily returns but assumes an MA(1) model [i.e., using Eq. (3.38)]. The third method applies a GARCH(1, 1) model to the monthly returns from January 1962 to December 1999. We use a longer data span to obtain a more accurate estimate of the monthly volatility. The GARCH(1, 1) model used is

$$r_t^m = 0.658 + a_t, \quad a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = 3.349 + 0.086 a_{t-1}^2 + 0.735\sigma_{t-1}^2,$$

where $\epsilon_t$ is a standard Gaussian white noise series. Figure 3.10 shows the time plots of the estimated monthly volatility. Clearly the estimated volatilities based on daily returns are much higher than those based on monthly returns and a GARCH(1, 1)

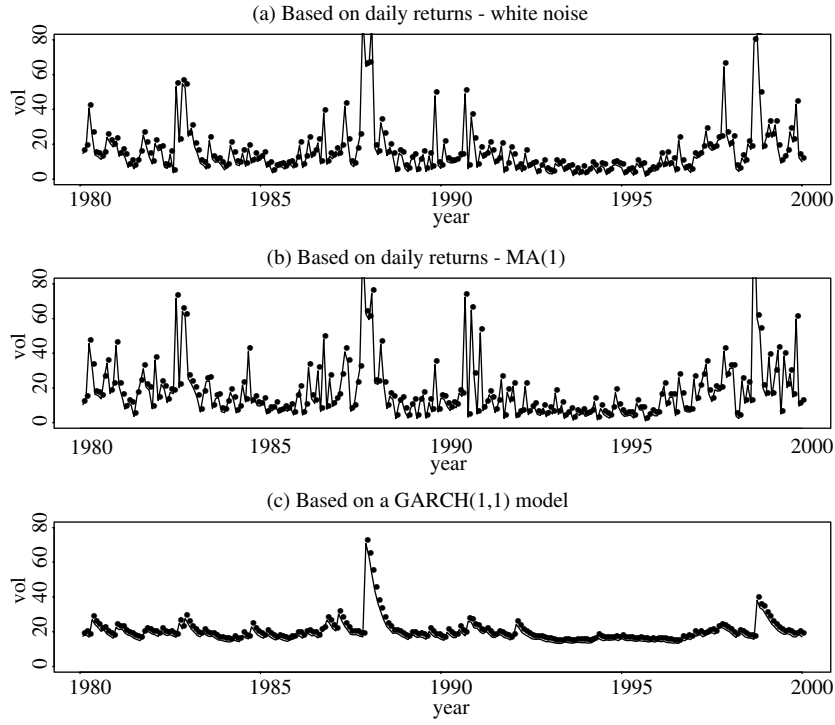


**Figure 3.10.** Time plots of estimated monthly volatility for the log returns of S&P 500 index from January 1980 to December 1999: (a) assumes that the daily log returns form a white noise series, (b) assumes that the daily log returns follow an MA(1) model, and (c) uses monthly returns from January 1962 to December 1999 and a GARCH(1, 1) model.

model. In particular, the estimated volatility for October 1987 was about 680 when daily returns are used. The plots shown were truncated to have the same scale.

***Remark:*** In Eq. (3.37), if we further assume that the sample mean $\bar{r}_t$ is zero, then we have $\hat{\sigma}_m^2 \approx \sum_{i=1}^{n} r_{t,i}^2$. In this case, the cumulative sum of squares of daily log returns in a month can be used as an estimate of monthly volatility.

## 3.13 APPLICATION

In this section, we apply the volatility models discussed in this chapter to investigate some problems of practical importance. The data used are the monthly log returns of IBM stock and S&P 500 index from January 1926 to December 1999. There are 888 observations, and the returns are in percentages and include dividends. Figure 3.11 shows the time plots of the two return series.

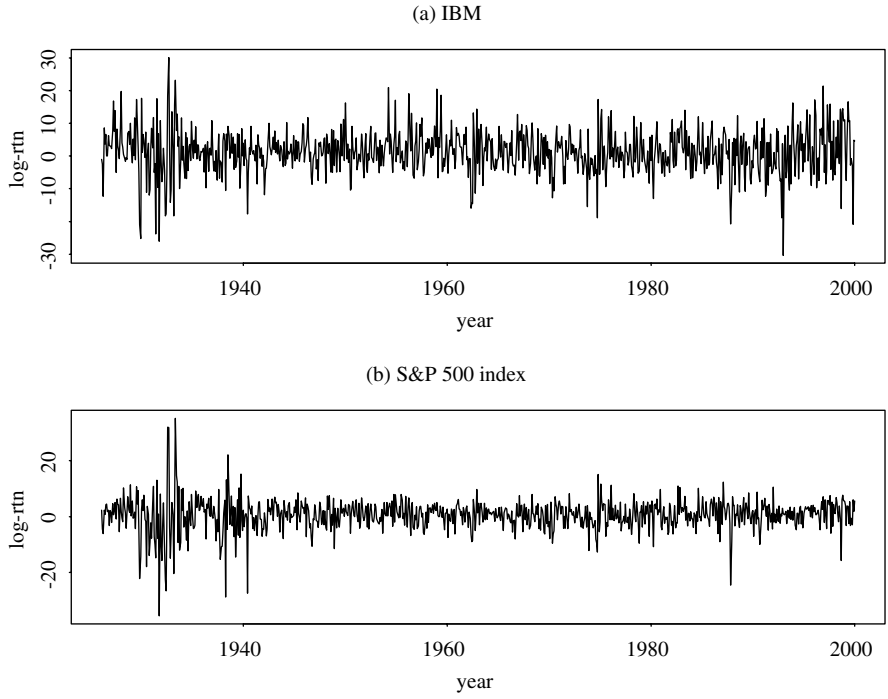


**Figure 3.11.** Time plots of monthly log returns for IBM stock and S&P 500 index. The sample period is from January 1926 to December 1999. The returns are in percentages and include dividends.

**Example 3.5.** The questions we address here are whether the daily volatility of a stock is lower in the Summer and, if so, by how much. Affirmative answers to these two questions have practical implications in stock option pricing. We use the monthly log returns of IBM stock shown in Figure 3.11(a) as an illustrative example.

Denote the monthly log return series by $r_t$. If Gaussian GARCH models are entertained, we obtain the GARCH(1, 1) model

$$\begin{aligned} r_t &= 1.23 + 0.099 r_{t-1} + a_t, \quad a_t = \sigma_t \epsilon_t \\ \sigma_t^2 &= 3.206 + 0.103 a_{t-1}^2 + 0.825 \sigma_{t-1}^2, \end{aligned} \tag{3.39}$$

for the series. The standard errors of the two parameters in the mean equation are 0.222 and 0.037, respectively, whereas those of the parameters in the volatility equation are 0.947, 0.021, and 0.037, respectively. Using the standardized residuals $\tilde{a}_t = a_t/\sigma_t$, we obtain $Q(10) = 7.82(0.553)$ and $Q(20) = 21.22(0.325)$, where $p$ value is in parentheses. Therefore, there are no serial correlations in the residuals of the mean equation. The Ljung–Box statistics of the $\tilde{a}_t^2$ series show $Q(10) = 2.89(0.98)$ and $Q(20) = 7.26(0.99)$, indicating that the standardized residuals have no conditional

heteroscedasticity. The fitted model seems adequate. This model serves as a starting point for further study.

To study the Summer effect on stock volatility of an asset, we define an indicator variable

$$u_t = \begin{cases} 1 & \text{if } t \text{ is June, July, or August} \\ 0 & \text{otherwise} \end{cases} \tag{3.40}$$

and modify the volatility equation to

$$\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + u_t(\alpha_{00} + \alpha_{10} a_{t-1}^2 + \beta_{10}\sigma_{t-1}^2).$$

This equation uses two GARCH(1, 1) models to describe the volatility of a stock return; one model for the Summer months and the other for the remaining months. For the monthly log returns of IBM stock, estimation results show that the estimates of $\alpha_{10}$ and $\beta_{10}$ are statistically nonsignificance at the 10% level. Therefore, we refine the equation and obtain the model

$$\begin{aligned} r_t &= 1.21 + 0.099 r_{t-1} + a_t, \quad a_t = \sigma_t \epsilon_t \\ \sigma_t^2 &= 4.539 + 0.113 a_{t-1}^2 + 0.816\sigma_{t-1}^2 - 5.154 u_t. \end{aligned} \tag{3.41}$$

The standard errors of the parameters in the mean equation are 0.218 and 0.037, respectively, and those of the parameters in the volatility equation are 1.071, 0.022, 0.037, and 1.900, respectively. The Ljung–Box statistics for the standardized residuals $\tilde{a}_t = a_t/\sigma_t$ show $Q(10) = 7.66(0.569)$ and $Q(20) = 21.64(0.302)$. Therefore, there are no serial correlations in the standardized residuals. The Ljung–Box statistics for $\tilde{a}_t^2$ give $Q(10) = 3.38(0.97)$ and $Q(20) = 6.82(0.99)$, indicating no conditional heteroscedasticity in the standardized residuals, either. The refined model seems adequate.

Comparing the volatility models in Eqs. (3.39) and (3.41), we obtain the following conclusions. First, because the coefficient $-5.514$ is significantly different from zero with $p$ value 0.0067, the Summer effect on stock volatility is statistically significant at the 1% level. Furthermore, the negative sign of the estimate confirms that the volatility of IBM monthly log stock returns is indeed lower during the Summer. Second, rewrite the volatility model in Eq. (3.41) as

$$\sigma_t^2 = \begin{cases} -0.615 + 0.113 a_{t-1}^2 + 0.816\sigma_{t-1}^2 & \text{if } t \text{ is June, July, or August} \\ 4.539 + 0.113 a_{t-1}^2 + 0.816\sigma_{t-1}^2, & \text{otherwise.} \end{cases}$$

The negative constant term $-0.615 = 4.539 - 5.514$ is counterintuitive. However, since the standard errors of 4.539 and 5.514 are relatively large, the estimated difference $-0.615$ might not be significantly different from zero. To verify the assertion, we refit the model by imposing the constraint that the constant term of the volatility equation is zero for the Summer months. This can easily be done by using the

equation

$$\sigma_t^2 = \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \gamma(1 - u_t).$$

The fitted model is

$$\begin{aligned} r_t &= 1.21 + 0.099 r_{t-1} + a_t, \quad a_t = \sigma_t \epsilon_t \\ \sigma_t^2 &= 0.114 a_{t-1}^2 + 0.811 \sigma_{t-1}^2 + 4.552(1 - u_t). \end{aligned} \tag{3.42}$$

The standard errors of the parameters in the mean equation are 0.219 and 0.038, respectively, and those of the parameters in the volatility equation are 0.022, 0.034, and 1.094, respectively. The Ljung–Box statistics of the standardized residuals show $Q(10) = 7.68$ and $Q(20) = 21.67$ and those of the $\tilde{a}_t^2$ series give $Q(10) = 3.17$ and $Q(20) = 6.85$. These test statistics are close to what we had before and are not significant at the 5% level.

The volatility Eq. (3.42) can readily be used to assess the Summer effect on the IBM stock volatility. For illustration, based on the model in Eq. (3.42), the medians of $a_t^2$ and $\sigma_t^2$ are 29.4 and 75.1, respectively, for the IBM monthly log returns in 1999. Using these values, we have $\sigma_t^2 = 0.114 \times 29.4 + 0.811 \times 75.1 = 64.3$ for the Summer months and $\sigma_t^2 = 68.8$ for the other months. Ratio of the two volatilities is $64.3/68.8 \approx 93\%$. Thus, there is a 7% reduction in the volatility of the monthly log return of IBM stock in the Summer months.

**Example 3.6.** The S&P 500 index is widely used in the derivative markets. As such, modeling its volatility is a subject of intensive study. The question we ask in this example is whether the past returns of individual components of the index contribute to the modeling of the S&P 500 index volatility in the presence of its own returns. A thorough investigation on this topic is beyond the scope of this chapter, but we use the past returns of IBM stock as explanatory variables to address the question.

The data used are shown in Figure 3.11. Denote by $r_t$ the monthly log return series of S&P 500 index. Using the $r_t$ series and Gaussian GARCH models, we obtain the following special GARCH(2, 1) model

$$r_t = 0.609 + a_t, \quad a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = 0.717 + 0.147 a_{t-2}^2 + 0.839 \sigma_{t-1}^2. \tag{3.43}$$

The standard error of the constant term in the mean equation is 0.138 and those of the parameters in the volatility equation are 0.214, 0.021, and 0.017, respectively. Based on the standardized residuals $\tilde{a}_t = a_t/\sigma_t$, we have $Q(10) = 11.51(0.32)$ and $Q(20) = 23.71(0.26)$, where the number in parentheses denotes $p$ value. For the $\tilde{a}_t^2$ series, we have $Q(10) = 9.42(0.49)$ and $Q(20) = 13.01(0.88)$. Therefore, the model seems adequate at the 5% significance level.

Next we evaluate the contributions, if any, of using the past returns of IBM stock, which is a component of the S&P 500 index, in modeling the index volatility. As a simple illustration, we modify the volatility equation as

$$\sigma_t^2 = \alpha_0 + \alpha_2 a_{t-2}^2 + \beta_1 \sigma_{t-1}^2 + \gamma(x_{t-1} - 1.24)^2,$$

**Table 3.3. Fitted Volatilities for the Monthly Log Returns of the S&P 500 Index from July to December 1999 Using Models with and without the Past Log Return of IBM Stock.**

| Month | 7/99 | 8/99 | 9/99 | 10/99 | 11/99 | 12/99 |
|---|---|---|---|---|---|---|
| Model (3.43) | 26.30 | 26.01 | 24.73 | 21.69 | 20.71 | 22.46 |
| Model (3.44) | 23.32 | 23.13 | 22.46 | 20.00 | 19.45 | 18.27 |

where $x_t$ is the monthly log return of IBM stock and 1.24 is the sample mean of $x_t$. The fitted model for $r_t$ becomes

$$\begin{aligned} r_t &= 0.616 + a_t, \quad a_t = \sigma_t \epsilon_t, \\ \sigma_t^2 &= 1.069 + 0.148 a_{t-2}^2 + 0.834 \sigma_{t-1}^2 - 0.007 (x_{t-1} - 1.24)^2. \end{aligned} \tag{3.44}$$

The standard error of the parameter in the mean equation is 0.139 and those of the parameters in the volatility equation are 0.271, 0.020, 0.018, and 0.002, respectively. For model checking, we have $Q(10) = 11.39(0.33)$ and $Q(20) = 23.63(0.26)$ for the standardized residuals $\tilde{a}_t = a_t/\sigma_t$ and $Q(10) = 9.35(0.50)$ and $Q(20) = 13.51(0.85)$ for the $\tilde{a}_t^2$ series. Therefore, the model is adequate.

Since the $p$ value for testing $\gamma = 0$ is 0.0039, the contribution of the lag-1 IBM stock return to the S&P 500 index volatility is statistically significant at the 1% level. The negative sign is understandable because it implies that using the lag-1 past return of IBM stock reduces the volatility of the S&P 500 index return. Table 3.3 gives the fitted volatility of S&P 500 index from July to December of 1999 using models (3.43) and (3.44). From the table, the past value of IBM log stock return indeed contributes to the modeling of the S&P 500 index volatility.

## 3.14 KURTOSIS OF GARCH MODELS

Uncertainty in volatility estimation is an important issue, but it is often overlooked. To assess the variability of an estimated volatility, one must consider the kurtosis of a volatility model. In this section, we derive the excess kurtosis of a GARCH(1, 1) model. The same idea applies to other GARCH models, however. The model considered is

$$a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

where $\alpha_0 > 0$, $\alpha_1 \geq 0$, $\beta_1 \geq 0$, $\alpha_1 + \beta_1 < 1$, and $\{\epsilon_t\}$ is an iid sequence satisfying

$$E(\epsilon_t) = 0, \quad \text{Var}(\epsilon_t) = 1, \quad E(\epsilon_t^4) = K_\epsilon + 3,$$

where $K_\epsilon$ is the excess kurtosis of the innovation $\epsilon_t$. Based on the assumption, we have

- $\text{Var}(a_t) = E(\sigma_t^2) = \alpha_0/[1 - (\alpha_1 + \beta_1)]$.
- $E(a_t^4) = (K_\epsilon + 3)E(\sigma_t^4)$ provided that $E(\sigma_t^4)$ exists.

Taking the square of the volatility model, we have

$$\sigma_t^4 = \alpha_0^2 + \alpha_1^2 a_{t-1}^4 + \beta_1^2 \sigma_{t-1}^4 + 2\alpha_0\alpha_1 a_{t-1}^2 + 2\alpha_0\beta_1\sigma_{t-1}^2 + 2\alpha_1\beta_1\sigma_{t-1}^2 a_{t-1}^2.$$

Taking expectation of the equation and using the two properties mentioned earlier, we obtain

$$E(\sigma_1^4) = \frac{\alpha_0^2(1 + \alpha_1 + \beta_1)}{[1 - (\alpha_1 + \beta_1)][1 - \alpha_1^2(K_\epsilon + 2) - (\alpha_1 + \beta_1)^2]}$$

provided that $1 > \alpha_1 + \beta_1 \geq 0$ and $1 - \alpha_1^2(K_\epsilon + 2) - (\alpha_1 + \beta_1)^2 > 0$. The excess kurtosis of $a_t$, if it exists, is then

$$\begin{aligned} K_a &= \frac{E(a_t^4)}{[E(a_t^2)]^2} - 3 \\ &= \frac{(K_\epsilon + 3)[1 - (\alpha_1 + \beta_1)^2]}{1 - 2\alpha_1^2 - (\alpha_1 + \beta_1)^2 - K_\epsilon\alpha_1^2} - 3. \end{aligned}$$

This excess kurtosis can be written in an informative expression. First, consider the case that $\epsilon_t$ is normally distributed. In this case, $K_\epsilon = 0$, and some algebra shows that

$$K_a^{(g)} = \frac{6\alpha_1^2}{1 - 2\alpha_1^2 - (\alpha_1 + \beta_1)^2},$$

where the superscript $(g)$ is used to denote Gaussian distribution. This result has two important implications: (a) the kurtosis of $a_t$ exists if $1 - 2\alpha_1^2 - (\alpha_1 + \beta_1)^2 > 0$, and (b) if $\alpha_1 = 0$, then $K_a^{(g)} = 0$, meaning that the corresponding GARCH(1, 1) model does not have heavy tails.

Second, consider the case that $\epsilon_t$ is not Gaussian. Using the prior result, we have

$$\begin{aligned} K_a &= \frac{K_\epsilon - K_\epsilon(\alpha_1 + \beta_1) + 6\alpha_1^2 + 3K_\epsilon\alpha_1^2}{1 - 2\alpha_1^2 - (\alpha_1 + \beta_1)^2 - K_\epsilon\alpha_1^2} \\ &= \frac{K_\epsilon[1 - 2\alpha_1^2 - (\alpha_1 + \beta_1)^2] + 6\alpha_1^2 + 5K_\epsilon\alpha_1^2}{1 - 2\alpha_1^2 - (\alpha_1 + \beta_1)^2 - K_\epsilon\alpha_1^2} \end{aligned}$$

$$= \frac{K_\epsilon + K_a^{(g)} + \frac{5}{6}K_\epsilon K_a^{(g)}}{1 - \frac{1}{6}K_\epsilon K_a^{(g)}}.$$

This result was obtained originally by George C. Tiao; see Bai, Russell, and Tiao (2001). It holds for all GARCH models provided that the kurtosis exists. For instance, if $\beta_1 = 0$, then the model reduces to an ARCH(1) model. In this case, it is easy to verify that $K_a^{(g)} = 6\alpha_1^2/(1 - 3\alpha_1^2)$ provided that $1 > 3\alpha_1^2$ and the excess kurtosis of $a_t$ is

$$\begin{aligned} K_a &= \frac{(K_\epsilon + 3)(1 - \alpha_1^2)}{1 - (K_\epsilon + 3)\alpha_1^2} - 3 = \frac{K_\epsilon + 2K_\epsilon\alpha_1^2 + 6\alpha_1^2}{1 - 3\alpha_1^2 - K_\epsilon\alpha_1^2} \\ &= \frac{K_\epsilon(1 - 3\alpha_1^2) + 6\alpha_1^2 + 5K_\epsilon\alpha_1^2}{1 - 3\alpha_1^2 - K_\epsilon\alpha_1^2} \\ &= \frac{K_\epsilon + K_a^{(g)} + \frac{5}{6}K_\epsilon K_a^{(g)}}{1 - \frac{1}{6}K_\epsilon K_a^{(g)}}. \end{aligned}$$

The prior result shows that for a GARCH(1, 1) model the coefficient $\alpha_1$ plays a critical role in determining the tail behavior of $a_t$. If $\alpha_1 = 0$, then $K_a^{(g)} = 0$ and $K_a = K_\epsilon$. In this case, the tail behavior of $a_t$ is similar to that of the standardized noise $\epsilon_t$. Yet if $\alpha_1 > 0$, then $K_a^{(g)} > 0$ and the $a_t$ process has heavy tails.

For a (standardized) Student-$t$ distribution with $v$ degrees of freedom, we have $E(\epsilon_t^4) = 6/(v-4)+3$ if $v > 4$. Therefore, the excess kurtosis of $\epsilon_t$ is $K_\epsilon = 6/(v-4)$ for $v > 4$. This is part of the reason that we used $t_5$ in the chapter when the degrees of freedom of $t$ distribution are prespecified. The excess kurtosis of $a_t$ becomes $K_a = [6+(v+1)K_a^{(g)}]/[v-4-K_a^{(g)}]$ provided that $1-2\alpha_1^2(v-1)/(v-4)-(\alpha_1+\beta_1)^2 > 0$.

## APPENDIX A. SOME RATS PROGRAMS FOR ESTIMATING VOLATILITY MODELS

The data file used in the illustration is "sp500.dat," which contains the monthly excess returns of S&P 500 index with 792 observations. Comments in a RATS program start with "*".

### A. A Gaussian GARCH(1, 1) Model with a Constant Mean Equation

```
all 0  792:1
open data sp500.dat
data(org=obs) / rt
*** initialize the conditional variance function
set h = 0.0
*** specify the parameters of the model
nonlin mu a0 a1 b1
```

```
*** specify the mean equation
frml at = rt(t)-mu
*** specify the volatility equation
frml gvar = a0+a1*at(t-1)**2+b1*h(t-1)
*** specify the log likelihood function
frml garchln = -0.5*log(h(t)=gvar(t))-0.5*at(t)**2/h(t)
*** sample period used in estimation
smpl 2 792
*** initial estimates
compute a0 = 0.01, a1 = 0.1, b1 = 0.5, mu = 0.1
maximize(method=bhhh,recursive,iterations=150) garchln
set fv = gvar(t)
set resid = at(t)/sqrt(fv(t))
set residsq = resid(t)*resid(t)
*** Checking standardized residuals
cor(qstats,number=20,span=10) resid
*** Checking squared standardized residuals
cor(qstats,number=20,span=10) residsq
```

### B. A GARCH(1, 1) Model with Student-*t* Innovation

```
all 0  792:1
open data sp500.dat
data(org=obs) / rt
set h = 0.0
nonlin mu a0 a1 b1 v
frml at = rt(t)-mu
frml gvar = a0+a1*at(t-1)**2+b1*h(t-1)
frml tt = at(t)**2/(h(t)=gvar(t))
frml tln = %LNGAMMA((v+1)/2.)-%LNGAMMA(v/2.)-0.5*log(v-2.)
frml gln = tln-((v+1)/2.)*log(1.0+tt(t)/(v-2.0))-0.5*log(h(t))
smpl  2 792
compute a0 = 0.01, a1 = 0.1, b1 = 0.5, mu = 0.1, v = 10
maximize(method=bhhh,recursive,iterations=150) gln
set fv = gvar(t)
set resid = at(t)/sqrt(fv(t))
set residsq = resid(t)*resid(t)
cor(qstats,number=20,span=10) resid
cor(qstats,number=20,span=10) residsq
```

### C. An AR(1)-EGARCH(1,0) Model for Monthly Log Returns of IBM Stock

```
all 0  864:1
open data m-ibm.dat
data(org=obs) / rt
set h = 0.0
nonlin c0 p1 th ga a0 a1
frml at = rt(t)-c0-p1*rt(t-1)
frml epsi = at(t)/(sqrt(exp(h(t))))
frml g = th*epsi(t)+ga*(abs(epsi(t))-sqrt(2./%PI))
frml gvar = a1*h(t-1)+(1-a1)*a0+g(t-1)
frml garchln = -0.5*(h(t)=gvar(t))-0.5*epsi(t)**2
smpl 3  864
```

```
compute c0 = 0.01, p1 = 0.01, th = 0.1, ga = 0.1
compute a0 = 0.01, a1 = 0.5
maximize(method=bhhh,recursive,iterations=150) garchln
set fv = gvar(t)
set resid = epsi(t)
set residsq = resid(t)*resid(t)
cor(qstats,number=20,span=10) resid
cor(qstats,number=20,span=10) residsq
```

## EXERCISES

**1.** Derive multistep ahead forecasts for a GARCH(1, 2) model at the forecast origin $h$.

**2.** Derive multistep ahead forecasts for a GARCH(2, 1) model at the forecast origin $h$.

**3.** Suppose that $r_1, \ldots, r_n$ are observations of a return series that follows the AR(1)-GARCH(1, 1) model

$$r_t = \mu + \phi_1 r_{t-1} + a_t, \quad a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2,$$

where $\epsilon_t$ is a standard Gaussian white noise series. Derive the conditional log likelihood function of the data.

**4.** In the previous equation, assume that $\epsilon_t$ follows a standardized Student-$t$ distribution with $v$ degrees of freedom. Derive the conditional log likelihood function of the data.

**5.** Consider the monthly log return of Intel stock from 1973 to 1997. Build a GARCH model for the series and compute 1 to 5-step ahead volatility forecasts at the forecast origin December 1997.

**6.** The file "m-mrk.dat" contains monthly simple returns of Merck stock. There are three columns—namely, monthly simple returns, years, and months. Transform the simple returns to log returns.

- Is there evidence of ARCH effects in the log returns? Use Ljung–Box statistics for the squared returns with 5 and 10 lags of autocorrelation and 5% significance level to answer the question.
- Use the PACF of the squared log returns to identify an ARCH model for the data and fit the identified model. Write down the fitted model.

**7.** The file "m-mmm.dat" contains two columns. They are monthly simple return and date for 3M stock. Transform the returns to log returns.

- Is there evidence of ARCH effects in the log returns? Use Ljung–Box statistics with 5 and 10 lags of autocorrelation and 5% significance level to answer the question.
- Use the PACF of the squared returns to identify an ARCH model. What is the fitted model?
- There are 623 data points. Use the fitted model earlier to predict the volatilities for $t = 624$ and $t = 625$ (the forecast origin is 623).
- Build a ARCH-M model for the log return series of 3M stock. Test the hypothesis that the risk premium is zero at the 5% significance level. Draw your conclusion.
- Build an EGARCH model for the log return series of 3M stock. Use the fitted model to compute 1- and 2-step ahead volatility forecasts at the forecast origin $h = 623$.

**8.** The file "m-gmsp5099.dat" contains the monthly log returns, in percentages, of General Motors stock and S&P 500 index from 1950 to 1999. The GM stock returns are in column 1.

- Build a GARCH model with Gaussian innovations for the log returns of GM stock. Check the model and write down the fitted model.
- Build a GARCH-M model with Gaussian innovations for the log returns of GM stock. What is the fitted model?
- Build a GARCH model with Student-$t$ distribution with 6 degrees of freedom for the GM log returns. Check the model and write down the fitted model.
- Build a GARCH model with Student-$t$ distribution for the log returns of GM stock, including estimation of the degrees of freedom. Write down the fitted model. Let $v$ be the degrees of freedom of the Student-$t$ distribution. Test the hypothesis $H_o : v = 6$ versus $H_a : v \neq 6$, using the 5% significance level.
- Build an EGARCH model for the log returns of GM stock. What is the fitted model?
- Compare all the volatility models obtained for the log returns of GM stock. Is there any significant difference? Why?

**9.** Again, consider the file "m-gmsp5099.dat."

- Build a Gaussian GARCH model for the monthly log returns of S&P 500 index. Check the model carefully.
- Is there a Summer effect on the volatility of the index return? Use the GARCH model built in part (a) to answer this question.
- Are lagged returns of GM stock useful in modeling the index volatility? Again, use the GARCH model of part (a) as a baseline model for comparison.

**10.** The file "d-ibmln.dat" contains the daily log returns, in percentages, of IBM stock from July 1962 to December 1997 with 8938 observations. The file has only one column. Fit a GARCH(1, 1) model to the series. What is the fitted model?

## REFERENCES

Andersen, T. G., and Bollerslev, T. (1998), "Answering the skeptics: Yes, standard volatility models do provide accurate forecasts," *International Economic Review*, 39, 885–905.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (1999), "The distribution of exchange rate volatility," Working paper, Economics Department, University of Pennsylvania.

Bai, X., Russell, J. R., and Tiao, G. C. (2000), "Beyond Metron's Utopia: effects of dependence and non-normality on variance estimates using high-frequency data," Working paper, Graduate School of Business, University of Chicago.

Bai, X., Russell, J. R., and Tiao, G. C. (2001), "Kurtosis of GARCH and stochastic volatility models," Working paper, Graduate School of Business, University of Chicago.

Bollerslev, T. (1986), "Generalized autoregressive conditional heteroskedasticity." *Journal of Econometrics*, 31, 307–327.

Bollerslev, T. (1990), "Modeling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH approach," *Review of Economics and Statistics*, 72, 498–505.

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992), "ARCH modeling in finance," *Journal of Econometrics*, 52, 5–59.

Bollerslev, T., Engle, R. F., and Nelson, D. B. (1994), "ARCH model" in Handbook of Econometrics IV, 2959–3038, ed. Engle, R. F., and McFadden, D. C. Amsterdam: Elsrier Science.

Bollerslev, T., and Jubinski, D. (1999), "Equality trading volume and volatility: Latent information arrivals and common long-run dependencies," *Journal of Business & Economic Statistics*, 17, 9–21.

Breidt, F. J., Crato, N., and de Lima, P. (1998), "On the detection and estimation of long memory in stochastic volatility," *Journal of Econometrics*, 83, 325–348.

Cao, C., and Tsay, R. S. (1992), "Nonlinear time series analysis of stock volatilities," *Journal of Applied Econometrics*, 7, s165–s185.

Ding, Z., Granger, C. W. J., and Engle, R. F. (1993), "A long memory property of stock returns and a new model," *Journal of Empirical Finance*, 1, 83–106.

Engle, R. F. (1982), "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflations," *Econometrica*, 50, 987–1007.

French, K. R., Schwert, G. W., and Stambaugh, R. F. (1987), "Expected stock returns and volatility," *Journal of Financial Economics*, 19, 3–29.

Harvey, A. C., Ruiz, E., and Shephard, N. (1994), "Multivariate stochastic variance models," *Review of Economic Studies*, 61, 247–264.

Jacquier, E., Polson, N. G., and Rossi, P. (1994), "Bayesian analysis of stochastic volatility models" (with discussion), *Journal of Business & Economic Statistics*, 12, 371–417.

McLeod, A. I., and Li, W. K. (1983), "Diagnostic checking ARMA time series models using squared-residual autocorrelations," *Journal of Time Series Analysis*, 4, 269–273.

Melino, A., and Turnbull, S. M. (1990), "Pricing foreign currency options with stochastic volatility," *Journal of Econometrics*, 45, 239–265.

Nelson, D. B. (1990), "Stationarity and persistence in the GARCH(1, 1) model," *Econometric Theory*, 6, 318–334.

Nelson, D. B. (1991), "Conditional heteroskedasticity in asset returns: A new approach," *Econometrica*, 59, 347–370.

Nicholls, D. F., and Quinn, B. G. (1982), *Random Coefficient Autoregressive Models: An Introduction*, Lecture Notes in Statistics, 11. Springer-Verlag: New York.

Ray, B. K., and Tsay, R. S. (2000), "Long-range dependence in daily stock volatilities," *Journal of Business & Economic Statistics*, 18, 254–262.

Taylor, S. J. (1994), "Modeling stochastic volatility," *Mathematical Finance*, 4, 183–204.

Tong, H. (1978), "On a threshold model," in *Pattern Recognition and Signal Processing*, ed. C.H. Chen, Sijhoff & Noordhoff: Amsterdam.

Tong, H. (1990), *Non-Linear Time Series: A Dynamical System Approach*, Oxford University Press: Oxford.

Tsay, R. S. (1987), "Conditional heteroscedastic time series models," *Journal of the American Statistical Association*, 82, 590–604.

*Analysis of Financial Time Series.* Ruey S. Tsay

ISBN: 0-471-41544-8

CHAPTER 4

# Nonlinear Models and Their Applications

This chapter focuses on nonlinearity in financial data and nonlinear econometric models useful in analysis of financial time series. Consider a univariate time series $x_t$, which, for simplicity, is observed at equally spaced time intervals. We denote the observations by $\{x_t \mid t = 1, \ldots, T\}$, where $T$ is the sample size. As stated in Chapter 2, a purely stochastic time series $x_t$ is said to be linear if it can be written as

$$x_t = \mu + \sum_{i=0}^{\infty} \psi_i a_{t-i}, \tag{4.1}$$

where $\mu$ is a constant, $\psi_i$ are real numbers with $\psi_0 = 1$, and $\{a_t\}$ is a sequence of independent and identically distributed (iid) random variables with a well-defined distribution function. We assume that the distribution of $a_t$ is continuous and $E(a_t) = 0$. In many cases, we further assume that $\text{Var}(a_t) = \sigma_a^2$ or, even stronger, that $a_t$ is Gaussian. If $\sigma_a^2 \sum_{i=1}^{\infty} \psi_i^2 < \infty$, then $x_t$ is weakly stationary (i.e., the first two moments of $x_t$ are time-invariant). The ARMA process of Chapter 2 is linear because it has an MA representation in Eq. (4.1). Any stochastic process that does not satisfy the condition of Eq. (4.1) is said to be nonlinear. The prior definition of nonlinearity is for purely stochastic time series. One may extend the definition by allowing the mean of $x_t$ to be a linear function of some exogenous variables, including the time index and some periodic functions. But such a mean function can be handled easily by methods discussed in Chapter 2, and we do not discuss it here. Mathematically, a purely stochastic time series model for $x_t$ is a function of an iid sequence consisting of the current and past shocks—that is,

$$x_t = f(a_t, a_{t-1}, \ldots). \tag{4.2}$$

The linear model in Eq. (4.1) says that $f(.)$ is a linear function of its arguments. Any nonlinearity in $f(.)$ results in a nonlinear model. The general nonlinear model in Eq. (4.2) is not directly applicable because it contains too many parameters.

To put nonlinear models available in the literature in a proper perspective, we write the model of $x_t$ in terms of its conditional moments. Let $F_{t-1}$ be the $\sigma$-field generated by available information at time $t-1$ (inclusive). Typically, $F_{t-1}$ denotes the collection of linear combinations of elements in $\{x_{t-1}, x_{t-2}, \ldots\}$ and $\{a_{t-1}, a_{t-2}, \ldots\}$. The conditional mean and variance of $x_t$ given $F_{t-1}$ are

$$\mu_t = E(x_t \mid F_{t-1}) \equiv g(F_{t-1}), \quad \sigma_t^2 = \text{Var}(x_t \mid F_{t-1}) \equiv h(F_{t-1}), \tag{4.3}$$

where $g(.)$ and $h(.)$ are well-defined functions with $h(.) > 0$. Thus, we restrict the model to

$$x_t = g(F_{t-1}) + \sqrt{h(F_{t-1})}\epsilon_t,$$

where $\epsilon_t = a_t/\sigma_t$ is a standardized shock. For the linear series $x_t$ in Eq. (4.1), $g(.)$ is a linear function of elements of $F_{t-1}$ and $h(.) = \sigma_a^2$. The development of nonlinear models involves making extensions of the two equations in Eq. (4.3). If $g(.)$ is nonlinear, $x_t$ is said to be *nonlinear in mean*. If $h(.)$ is time-variant, then $x_t$ is *nonlinear in variance*. The conditional heteroscedastic models of Chapter 3 are nonlinear in variance because their conditional variances $\sigma_t^2$ evolve over time. In fact, except for the GARCH-M models, in which $\mu_t$ depends on $\sigma_t^2$ and hence also evolves over time, all of the volatility models of Chapter 3 focus on modifications or extensions of the conditional variance equation in Eq. (4.3). Based on the well-known Wold Decomposition, a weakly stationary and purely stochastic time series can be expressed as a linear function of uncorrelated shocks. For stationary volatility series, these shocks are uncorrelated, but dependent. The models discussed in this chapter represent another extension to nonlinearity derived from modifying the conditional mean equation in Eq. (4.3).

Many nonlinear time series models have been proposed in the statistical literature, such as the bilinear models of Granger and Andersen (1978), the threshold autoregressive (TAR) model of Tong (1978), the state-dependent model of Priestley (1980), and the Markov switching model of Hamilton (1989). The basic idea underlying these nonlinear models is to let the conditional mean $\mu_t$ evolve over time according to some simple parametric nonlinear function. Recently, a number of nonlinear models have been proposed by making use of advances in computing facilities and computational methods. Examples of such extensions include the nonlinear state-space modeling of Carlin, Polson, and Stoffer (1992), the functional-coefficient autoregressive model of Chen and Tsay (1993a), the nonlinear additive autoregressive model of Chen and Tsay (1993b), and the multivariate adaptive regression spline of Lewis and Stevens (1991). The basic idea of these extensions is either using simulation methods to describe the evolution of the conditional distribution of $x_t$ or using data-driven methods to explore the nonlinear characteristics of a series. Finally, nonparametric and semiparametric methods such as kernel regression and artificial neural networks have also been applied to explore the nonlinearity in a time series. We discuss some nonlinear models in Section 4.1 that are applicable to financial time series. The discussion includes some nonparametric and semiparametric methods.

Apart from the development of various nonlinear models, there is substantial interest in studying test statistics that can discriminate linear series from nonlinear ones. Both parametric and nonparametric tests are available. Most parametric tests employ either the Lagrange multiplier or likelihood ratio statistics. Nonparametric tests depend on either higher order spectra of $x_t$ or the concept of dimension correlation developed for chaotic time series. We review some nonlinearity tests in Section 4.2. Sections 4.3 and 4.4 discuss modeling and forecasting of nonlinear models. Finally, an application of nonlinear models is given in Section 4.5.

## 4.1 NONLINEAR MODELS

Most nonlinear models developed in the statistical literature focus on the conditional mean equation in Eq. (4.3); see Priestley (1988) and Tong (1990) for summaries of nonlinear models. Our goal here is to introduce some nonlinear models that are applicable to financial time series.

### 4.1.1 Bilinear Model

The linear model in Eq. (4.1) is simply the first-order Taylor series expansion of the $f(.)$ function in Eq. (4.2). As such, a natural extension to nonlinearity is to employ the second-order terms in the expansion to improve the approximation. This is the basic idea of bilinear models, which can be defined as

$$x_t = c + \sum_{i=1}^{p} \phi_i x_{t-i} - \sum_{j=1}^{q} \theta_j a_{t-j} + \sum_{i=1}^{m} \sum_{j=1}^{s} \beta_{ij} x_{t-i} a_{t-j} + a_t, \tag{4.4}$$

where $p$, $q$, $m$, and $s$ are non-negative integers. This model was introduced by Granger and Andersen (1978) and has been widely investigated. Subba Rao and Gabr (1984) discuss some properties and applications of the model, and Liu and Brockwell (1988) study general bilinear models. Properties of bilinear models such as stationarity conditions are often derived by (a) putting the model in a state-space form, and (b) using the state transition equation to express the state as a product of past innovations and random coefficient vectors. A special generalization of the bilinear model in Eq. (4.4) has conditional heteroscedasticity. For example, consider the model

$$x_t = \mu + \sum_{i=1}^{s} \beta_i a_{t-i} a_t + a_t, \tag{4.5}$$

where $\{a_t\}$ is a white noise series. The first two conditional moments of $x_t$ are

$$E(x_t \mid F_{t-1}) = \mu, \quad \text{Var}(x_t \mid F_{t-1}) = \left(1 + \sum_{i=1}^{s} \beta_i a_{t-i}\right)^2 \sigma_a^2,$$

which are similar to that of the RCA or CHARMA model of Chapter 3.

**Example 4.1.** Consider the monthly simple returns of CRSP equal-weighted index from January 1926 to December 1997 for 864 observations. Denote the series by $R_t$. The sample PACF of $R_t$ shows significant partial autocorrelations at lags 1 and 3, whereas that of $R_t^2$ suggests that the conditional heteroscedasticity might depend on the past three innovations. Therefore, we employ the special bilinear model

$$R_t = \mu + \phi_1 R_{t-1} + \phi_3 R_{t-3} + (1 + \beta_1 a_{t-1} + \beta_2 a_{t-2} + \beta_3 a_{t-3})a_t$$

for the series. Assuming that the conditional distribution of $a_t$ is normal, we use the conditional maximum likelihood method and obtain the fitted model

$$\begin{aligned} R_t &= 0.014 + 0.160R_{t-1} - 0.104R_{t-3} \\ &\quad + (1 + 0.337a_{t-1} - 0.022a_{t-2} - 0.601a_{t-3})a_t, \end{aligned} \tag{4.6}$$

where $\hat{\sigma}_a^2 = .0052$ and the standard errors of the parameters are, in the order of appearance, 0.003, 0.026, 0.018, 0.083, 0.084, and 0.079. The only insignificant estimate is the coefficient of $a_{t-2}$. Define

$$\hat{a}_t = \frac{R_t - 0.014 - 0.160R_{t-1} + 0.014R_{t-3}}{1 + 0.337\hat{a}_{t-1} - 0.022\hat{a}_{t-2} - 0.601\hat{a}_{t-3}},$$

where $\hat{a}_t = 0$ for $t \leq 3$ as the residual series of the model. The sample ACF of $\hat{a}_t$ shows no significant serial correlations, but the series is not independent because the squared series $\hat{a}_t^2$ has significant serial correlations. The validity of model (4.6) deserves further investigation. For comparison, we also consider an ARCH(3) model for the series and obtain

$$\begin{aligned} R_t &= 0.013 + 0.222R_{t-1} - 0.140R_{t-3} + a_t, \\ \sigma_t^2 &= 0.002 + 0.168a_{t-1}^2 + 0.00001a_{t-2}^2 + 0.274a_{t-3}^2, \end{aligned} \tag{4.7}$$

where all estimates but the coefficient of $a_{t-2}^2$ are highly significant. The standardized residual series and its squared series show no serial correlations, indicating that the ARCH(3) model is adequate for the data. Models (4.6) and (4.7) appear to be similar, but the latter seems to fit the data better.

### 4.1.2 Threshold Autoregressive (TAR) Model

This model is motivated by several nonlinear characteristics commonly observed in practice such as asymmetry in declining and rising patterns of a process. It uses piecewise linear models to obtain a better approximation of the conditional mean equation. However, in contrast to the traditional piecewise linear model that allows

for model changes to occur in the "time" space, the TAR model uses threshold space to improve linear approximation. Let us start with a simple two-regime AR(1) model

$$x_t = \begin{cases} -1.5x_{t-1} + a_t & \text{if } x_{t-1} < 0, \\ 0.5x_{t-1} + a_t & \text{if } x_{t-1} \geq 0, \end{cases} \tag{4.8}$$

where $a_t$s are iid $N(0, 1)$. Here the delay is 1 and the threshold is 0. Figure 4.1 shows the time plot of a simulated series of $x_t$ with 200 observations. A horizontal line of zero is added to the plot, which illustrates several characteristics of TAR models. First, despite of the coefficient $-1.5$ in the first regime, the process $x_t$ is geometrically ergodic and stationary. In fact, the necessary and sufficient condition for model (4.8) to be geometrically ergodic is $\phi_1^{(1)} < 1$, $\phi_1^{(2)} < 1$, and $\phi_1^{(1)}\phi_1^{(2)} < 1$, where $\phi^{(i)}$ is the AR coefficient of regime $i$; see Petruccelli and Woolford (1984) and Chen and Tsay (1991). Ergodicity is an important concept in time series analysis. For example, the statistical theory showing that the sample mean $\bar{x} = \sum_{t=1}^{T} x_t/T$ of $x_t$ converges to the mean of $x_t$ is referred to as the *ergodic theorem*, which can be regarded as the counterpart of the central limit theory for the iid case. Second, the series exhibits an asymmetric increasing and decreasing pattern. If $x_{t-1}$ is negative, then $x_t$ tends to switch to a positive value due to the negative and explosive coefficient $-1.5$. Yet when $x_{t-1}$ is positive, it tends to take multiple time indexes for $x_t$ to reduce to a negative value. Consequently, the time plot of $x_t$ shows that regime 2 has more observations than regime 1, and the series contains large upward

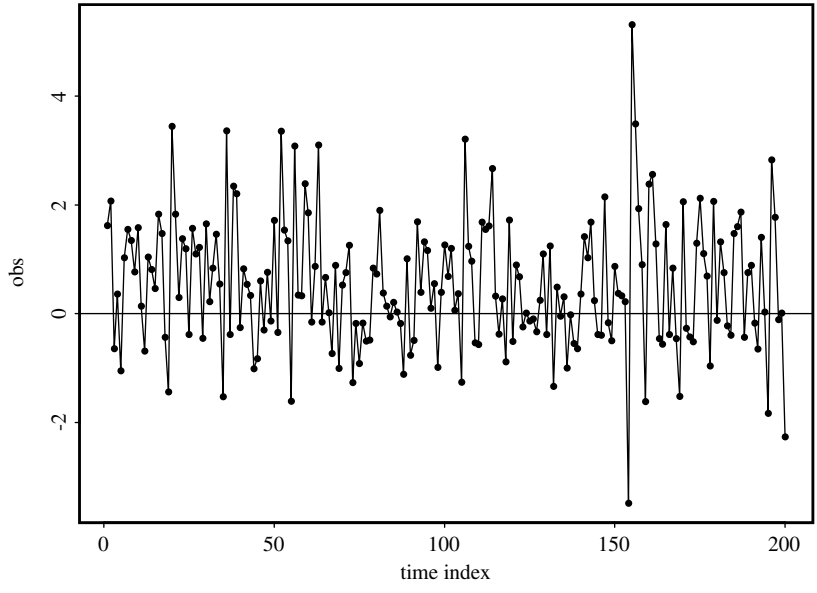


**Figure 4.1.** Time plot of a simulated two-regime TAR(1) series.

jumps when it becomes negative. The series is therefore not time-reversible. Third, the model contains no constant terms, but $E(x_t)$ is not zero. The sample mean of the particular realization is 0.61 with a standard deviation 0.07. In general, $E(x_t)$ is a weighted average of the conditional means of the two regimes, which are nonzero. The weight for each regime is simply the probability that $x_t$ is in that regime under its stationary distribution. It is also clear from the discussion that, for a TAR model to have zero mean, nonzero constant terms in some of the regimes are needed. This is very different from a stationary linear model for which a nonzero constant implies that the mean of $x_t$ is not zero.

A time series $x_t$ is said to follow a $k$-regime self-exciting TAR (SETAR) model with threshold variable $x_{t-d}$ if it satisfies

$$x_t = \phi_0^{(j)} + \phi_1^{(j)} x_{t-1} - \cdots - \phi_p^{(j)} x_{t-p} + a_t^{(j)}, \qquad \text{if} \quad \gamma_{j-1} \le x_{t-d} < \gamma_j, \quad (4.9)$$

where $k$ and $d$ are positive integers, $j = 1, \ldots, k$, $\gamma_i$s are real numbers such that $-\infty = \gamma_0 < \gamma_1 < \cdots < \gamma_{k-1} < \gamma_k = \infty$, the superscript $(j)$ is used to signify the regime, $\{a_t^{(j)}\}$ are iid sequences with mean 0 and variance $\sigma_j^2$ and are mutually independent for different $j$. The parameter $d$ is referred to as the *delay parameter* and $\gamma_j$s as the *thresholds*. Here it is understood that the AR models are different for different regimes; otherwise, the number of regimes can be reduced. Equation (4.9) says that a SETAR model is a piecewise linear AR model in the threshold space. It is similar in spirit to the usual piecewise linear models in regression analysis, where model changes occur in the order by which observations are taken. The SETAR model is nonlinear provided that $k > 1$.

Properties of general SETAR models are hard to obtain, but some of them can be found in Tong (1990), Chan (1993), Chan and Tsay (1998), and the references therein. In recent years, there is increasing interest in TAR models and their applications; see, for instance, Hansen (1997), Tsay (1998), and Montgomery et al. (1998). A testing and modeling procedure for univariate SETAR models is proposed in Tsay (1989). The SETAR model in Eq. (4.9) can be generalized by using a threshold variable $z_t$ that is measurable with respect to $F_{t-1}$ (i.e., a function of elements of $F_{t-1}$). The main requirements are that $z_t$ is stationary with a continuous distribution function over a compact subset of the real line and that $z_{t-d}$ is known at time $t$. Such a generalized model is referred to as an *open-loop TAR model*.

An important application of SETAR models in finance is to handle the asymmetric responses in volatility between positive and negative returns. The models can also be used to study arbitrage tradings in index futures and cash prices; see Chapter 8 on multivariate time series analysis. Here we focus on volatility modeling.

**Example 4.2.** To illustrate the application of TAR models in finance, we consider the daily log returns, in percentages and including dividends, of IBM stock from July 3, 1962 to December 31, 1999 for 9442 observations. Figure 4.2 shows the time plot of the series. The volatility seems to be larger in recent years. If GARCH

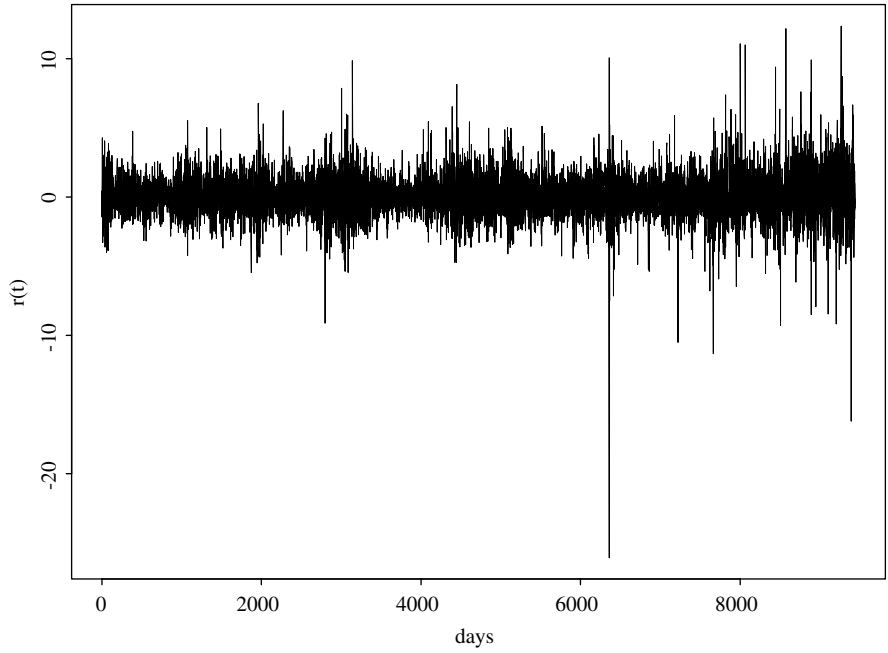


**Figure 4.2.** Time plot of the daily log returns for IBM stock from July 3, 1962 to December 31, 1999.

models of Chapter 3 are entertained, we obtain the following AR(2)-GARCH(1, 1) model for the series

$$r_t = 0.067 - 0.023r_{t-2} + a_t, \quad a_t = \sigma_t \epsilon_t$$
$$\sigma_t^2 = 0.031 + 0.076a_{t-1}^2 + 0.915\sigma_{t-1}^2, \tag{4.10}$$

where $r_t$ is the log return, $\{\epsilon_t\}$ is a Gaussian white noise sequence with mean zero and variance 1.0, the standard errors of the parameters in the mean equation are 0.013 and 0.011, respectively, and those of the volatility equation are 0.003, 0.002, and 0.003, respectively. All estimates but the coefficient of $r_{t-2}$ are highly significant. The Ljung–Box statistics of the standardized residuals give $Q(10) = 11.31(0.33)$ and $Q(20) = 27.00(0.14)$, where the number in parentheses denotes $p$ value. For the squared standardized residuals, we obtain $Q(10) = 11.86(0.29)$ and $Q(20) = 19.19(0.51)$. The model is adequate in modeling the serial dependence and conditional heteroscedasticity of the data. But the unconditional mean for $r_t$ of model (4.10) is 0.065, which is substantially larger than the sample mean 0.045, indicating that the model might be misspecified. The TAR model can be used to refine the model by allowing for asymmetric response in volatility to the sign of shock $a_{t-1}$. More specifically, we consider an AR(2)-TAR-GARCH(1, 1) model for

the series and obtain

$$\begin{aligned} r_t &= 0.043 - 0.022 r_{t-2} + a_t, \quad a_t = \sigma_t \epsilon_t \\ \sigma_t^2 &= 0.002 + 0.097 a_{t-1}^2 + 0.954 \sigma_{t-1}^2 \\ &\quad + (0.056 - 0.051 a_{t-1}^2 - 0.067 \sigma_{t-1}^2) I(a_{t-1} > 0), \end{aligned}$$

where $I(a_{t-1}) = 1$ if $a_{t-1} > 0$ and it is zero otherwise. Because the estimate 0.002 of the volatility equation is insignificant at the 5% level, we further refine the model to

$$\begin{aligned} r_t &= 0.043 - 0.022 r_{t-2} + a_t, \quad a_t = \sigma_t \epsilon_t \\ \sigma_t^2 &= 0.098 a_{t-1}^2 + 0.954 \sigma_{t-1}^2 \\ &\quad + (0.060 - 0.052 a_{t-1}^2 - 0.069 \sigma_{t-1}^2) I(a_{t-1} > 0), \end{aligned} \tag{4.11}$$

where the standard errors of the two parameters in the mean equation are 0.013 and 0.010, respectively, and those of the TAR-GARCH(1, 1) model are 0.003, 0.004, 0.005, 0.004, and 0.009. All of the estimates are statistically significant at the 5% level. The unconditional mean for $r_t$ of model (4.11) is 0.042, which is very close to the sample mean of $r_t$. Residual analysis based on the Ljung–Box statistics finds no significant serial correlations or conditional heteroscedasticity in the standardized residuals. The AR coefficient in the mean equation is small, indicating that, as expected, the daily log returns of IBM stock are essentially serially uncorrelated. However, the volatility model of the returns shows strong dependence in the innovational process $\{a_t\}$ and evidence of asymmetry in the conditional variance. Rewriting the TAR-GARCH(1, 1) equation as

$$\sigma_t^2 = \begin{cases} 0.098 a_{t-1}^2 + 0.954 \sigma_{t-1}^2 & \text{if } a_{t-1} \le 0 \\ 0.060 + 0.046 a_{t-1}^2 + 0.885 \sigma_{t-1}^2 & \text{if } a_{t-1} > 0, \end{cases} \tag{4.12}$$

we obtain some interesting implications. First, if we interpret $a_{t-1}$ as the deviation of IBM daily log return from its conditional expectation, then volatility follows essentially an IGARCH(1, 1) model without a drift when the deviation is nonpositive. Second, when the deviation is positive, the volatility has a persistent parameter $0.046 + 0.885 = 0.931$, which is close to, but less than, 1. Therefore, the volatility follows a GARCH(1, 1) model when the deviation is positive. Consequently, the volatility responds differently to positive and negative shocks. Finally, other threshold volatility models have also been proposed in the literature (e.g., Rabemananjara and Zakoian, 1993; Zakoian, 1994).

***Remark:*** The RATS program used to estimate the AR(2)-Tar-GARCH(1, 1) model is in the appendix.

### 4.1.3 Smooth Transition AR (STAR) Model

A criticism of the SETAR model is that its conditional mean equation is not continuous. The thresholds $\{\gamma_j\}$ are the discontinuity points of the conditional mean function $\mu_t$. In response to this criticism, smooth TAR models have been proposed; see Chan and Tong (1986) and Teräsvirta (1994) and the references therein. A time series $x_t$ is said to follow a two-regime STAR($p$) model if it satisfies

$$x_t = c_0 + \sum_{i=1}^{p} \phi_{0,i} x_{t-i} + F\left(\frac{x_{t-d} - \Delta}{s}\right)\left(c_1 + \sum_{i=1}^{p} \phi_{1,i} x_{t-i}\right) + a_t, \qquad (4.13)$$

where $d$ is the delay parameter, $\Delta$ and $s$ are parameters representing the location and scale of model transition, and $F(.)$ is a smooth transition function. In practice, $F(.)$ often assumes one of three forms—namely, logistic, exponential, or a cumulative distribution function. From Eq. (4.13), the conditional mean of a STAR model is a weighted linear combination between the following two equations:

$$\mu_{1t} = c_0 + \sum_{i=1}^{p} \phi_{0,i} x_{t-i},$$

$$\mu_{2t} = (c_0 + c_1) + \sum_{i=1}^{p} (\phi_{0,i} + \phi_{1,i}) x_{t-i}.$$

The weights are determined in a continuous manner by $F(\frac{x_{t-d}-\Delta}{s})$. The prior two equations also determine properties of a STAR model. For instance, a prerequisite for the stationarity of a STAR model is that all zeros of both AR polynomials are outside the unit circle. An advantage of the STAR model over the TAR model is that the conditional mean function is differentiable. However, experience shows that the transition parameters $\Delta$ and $s$ of a STAR model are hard to estimate. In particular, most empirical studies show that standard errors of the estimates of $\Delta$ and $s$ are often quite large resulting in $t$ ratios about 1.0; see Teräsvirta (1994). This uncertainty leads to various complications in interpreting an estimated STAR model.

**Example 4.3.** To illustrate the application of STAR models in financial time series analysis, we consider the monthly simple stock returns for Minnesota Mining and Manufacturing (3M) Company from February 1946 to December 1997. If ARCH models are entertained, we obtain the following ARCH(2) model

$$R_t = 0.014 + a_t, \quad a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = 0.003 + 0.108 a_{t-1}^2 + 0.151 a_{t-2}^2, \quad (4.14)$$

where standard errors of the estimates are 0.002, 0.0003, 0.045, and 0.058, respectively. As discussed before, such an ARCH model fails to show the asymmetric responses of stock volatility to positive and negative prior shocks. The STAR model provides a simple alternative that may overcome this difficulty. Applying STAR mod-

els to the monthly returns of 3M stock, we obtain the model

$$R_t = 0.017 + a_t, \quad a_t = \sigma_t \epsilon_t,$$
$$\sigma_t^2 = (0.002 + 0.256a_{t-1}^2 + 0.141a_{t-2}^2) + \frac{0.002 - 0.314a_{t-1}^2}{1 + \exp(-1000a_{t-1})}, \quad (4.15)$$

where the standard error of the constant term in the mean equation is 0.002 and those of the estimates in the volatility equation are 0.0003, 0.092, 0.056, 0.001, and 0.102, respectively. The scale parameter 1000 of the logistic transition function is fixed *a priori* to simplify the estimation. This STAR model provides some support for asymmetric responses to positive and negative prior shocks. For a large negative $a_{t-1}$, the volatility model approaches the ARCH(2) model

$$\sigma_t^2 = 0.002 + 0.256a_{t-1}^2 + 0.141a_{t-2}^2.$$

Yet for a large positive $a_{t-1}$, the volatility process behaves like the ARCH(2) model

$$\sigma_t^2 = 0.005 - 0.058a_{t-1}^2 + 0.141a_{t-2}^2.$$

The negative coefficient of $a_{t-1}^2$ in the prior model is counterintuitive, but the magnitude is small. As a matter of fact, for a large positive shock $a_{t-1}$, the ARCH effects appear to be weak even though the parameter estimates remain statistically significant. The RATS program used is given in the appendix.

### 4.1.4 Markov Switching Model

The idea of using probability switching in nonlinear time series analysis is discussed in Tong (1983). Using a similar idea, but emphasizing aperiodic transition between various states of an economy, Hamilton (1989) considers the Markov switching autoregressive (MSA) model. Here the transition is driven by a hidden two-state Markov chain. A time series $x_t$ follows an MSA model if it satisfies

$$x_t = \begin{cases} c_1 + \sum_{i=1}^{p} \phi_{1,i} x_{t-i} + a_{1t} & \text{if } s_t = 1, \\ c_2 + \sum_{i=1}^{p} \phi_{2,i} x_{t-i} + a_{2t} & \text{if } s_t = 2, \end{cases} \quad (4.16)$$

where $s_t$ assumes values in $\{1, 2\}$ and is a first-order Markov chain with transition probabilities

$$P(s_t = 2 \mid s_{t-1} = 1) = w_1, \quad P(s_t = 1 \mid s_{t-1} = 2) = w_2.$$

The innovational series $\{a_{1t}\}$ and $\{a_{2t}\}$ are sequences of iid random variables with mean zero and finite variance and are independent of each other. A small $w_i$ means that the model tends to stay longer in state $i$. In fact, $1/w_i$ is the expected duration of the process to stay in State $i$. From the definition, an MSA model uses a hidden

Markov chain to govern the transition from one conditional mean function to another. This is different from that of a SETAR model for which the transition is determined by a particular lagged variable. Consequently, a SETAR model uses a deterministic scheme to govern the model transition, whereas an MSA model uses a stochastic scheme. In practice, the stochastic nature of the states implies that one is never certain about which state $x_t$ belongs to in an MSA model. When the sample size is large, one can use some filtering techniques to draw inference on the state of $x_t$. Yet as long as $x_{t-d}$ is observed, the regime of $x_t$ is known in a SETAR model. This difference has important practical implications in forecasting. For instance, forecasts of an MSA model are always a linear combination of forecasts produced by submodels of individual states. But those of a SETAR model only come from a single regime provided that $x_{t-d}$ is observed. Forecasts of a SETAR model also become a linear combination of those produced by models of individual regimes when the forecast horizon exceeds the delay $d$. It is much harder to estimate an MSA model than other models because the states are not directly observable. Hamilton (1990) uses the EM algorithm, which is a statistical method iterating between taking expectation and maximization. McCulloch and Tsay (1994) consider a Markov Chain Monte Carlo (MCMC) method to estimate a general MSA model. We discuss MCMC methods in Chapter 10.

McCulloch and Tsay (1993) generalize the MSA model in Eq. (4.16) by letting the transition probabilities $w_1$ and $w_2$ be logistic, or probit, functions of some explanatory variables available at time $t-1$. Chen, McCulloch, and Tsay (1997) use the idea of Markov switching as a tool to perform model comparison and selection between non-nested nonlinear time series models (e.g., comparing bilinear and SETAR models). Each competing model is represented by a state. This approach to select a model is a generalization of the odds ratio commonly used in Bayesian analysis. Finally, the MSA model can easily be generalized to the case of more than two states. The computational intensity involved increases rapidly, however. For more discussions of Markov switching models in econometrics, see Hamilton (1994, Chapter 22).

**Example 4.4.** Consider the growth rate, in percentage, of U.S. quarterly real gross national product (GNP) from the second quarter of 1947 to the first quarter of 1991. The data are seasonally adjusted and shown in Figure 4.3, where a horizontal line of zero growth is also given. It is reassuring to see that a majority of the growth rates are positive. This series has been widely used in nonlinear analysis of economic time series. Tiao and Tsay (1994) and Potter (1995) use TAR models, whereas Hamilton (1989) and McCulloch and Tsay (1994) employ Markov switching models.

Employing the MSA model in Eq. (4.16) with $p = 4$ and using a Markov Chain Monte Carlo method, which is discussed in Chapter 10, McCulloch and Tsay (1994) obtain the estimates shown in Table 4.1. The results have several interesting findings. First, the mean growth rate of the marginal model for State 1 is $0.909/(1 - 0.265 - 0.029 + 0.126 + 0.11) = 0.965$ and that of State 2 is $-0.42/(1 - 0.216 - 0.628 + 0.073 + 0.097) = -1.288$. Thus, State 1 corresponds to quarters with positive growth, or expansion periods, whereas State 2 consists of quarters with negative

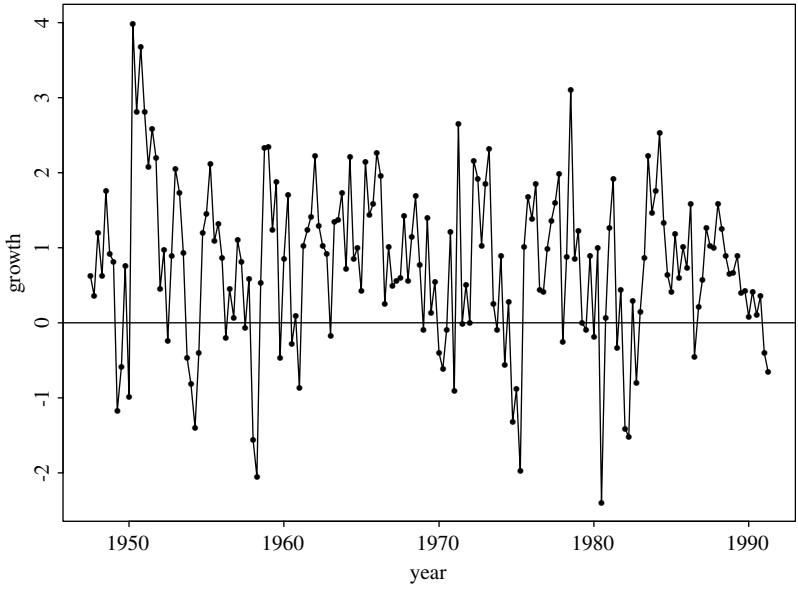


**Figure 4.3.** Time plot of the growth rate of U.S. quarterly real GNP from 1947.II to 1991.I. The data are seasonally adjusted and in percentages.

growth, or a contraction period. Second, the relatively large posterior standard deviations of the parameters in State 2 reflect that there are few observations in that state. This is expected as Figure 4.3 shows few quarters with negative growth. Third, the transition probabilities appear to be different for different states. The estimates indicate that it is more likely for the U.S. GNP to get out of a contraction period than to jump into one—0.286 versus 0.118. Fourth, treating $1/w_i$ as the expected duration for the process to stay in State $i$, we see that the expected durations for a contraction

**Table 4.1. Estimation Results of a Markov Switching Model with $p = 4$ for the Growth Rate of U.S. Quarterly Real GNP, Seasonally Adjusted. The Estimates and Their Standard Errors Are Posterior Means and Standard Errors of a Gibbs Sampling with 5000 Iterations.**

| | State 1 | | | | | | |
|---|---|---|---|---|---|---|---|
| Parameter | $c_i$ | $\phi_1$ | $\phi_2$ | $\phi_3$ | $\phi_4$ | $\sigma_i$ | $w_i$ |
| Estimate | 0.909 | 0.265 | 0.029 | −0.126 | −0.110 | 0.816 | 0.118 |
| Std. Error | 0.202 | 0.113 | 0.126 | 0.103 | 0.109 | 0.125 | 0.053 |
| | State 2 | | | | | | |
| Estimate | −0.420 | 0.216 | 0.628 | −0.073 | −0.097 | 1.017 | 0.286 |
| Std. Error | 0.324 | 0.347 | 0.377 | 0.364 | 0.404 | 0.293 | 0.064 |

period and an expansion period are approximately 3.69 and 11.31 quarters. Thus, on average, a contraction in the U.S. economy lasts about a year, whereas an expansion can last for 3 years. Finally, the estimated AR coefficients of $x_{t-2}$ differ substantially between the two states, indicating that the dynamics of the U.S. economy are different between expansion and contraction periods.

### 4.1.5 Nonparametric Methods

In some financial applications, we may not have sufficient knowledge to pre-specify the nonlinear structure between two variables $Y$ and $X$. In other applications, we may wish to take advantage of the advances in computing facilities and computational methods to explore the functional relationship between $Y$ and $X$. These considerations lead to the use of nonparametric methods and techniques. Nonparametric methods, however, are not without any cost. They are highly data-dependent and can easily result in overfitting. Our goal here is to introduce some nonparametric methods for financial applications and some nonlinear models that make use of nonparametric methods and techniques. The nonparametric methods discussed include kernel regression, local least squares estimation, and neural network.

The essence of nonparametric methods is *smoothing*. Consider two financial variables $Y$ and $X$, which are related by

$$Y_t = m(X_t) + a_t, \tag{4.17}$$

where $m(.)$ is an arbitrary, smooth, but unknown function and $\{a_t\}$ is a white noise sequence. We wish to estimate the nonlinear function $m(.)$ from the data. For simplicity, consider the problem of estimating $m(.)$ at a particular date for which $X = x$. That is, we are interested in estimating $m(x)$. Suppose that at $X = x$ we have repeated independent observations $y_1, \ldots, y_T$. Then the data become

$$y_t = m(x) + a_t, \quad t = 1, \ldots, T.$$

Taking the average of the data, we have

$$\frac{\sum_{t=1}^{T} y_t}{T} = m(x) + \frac{\sum_{t=1}^{T} a_t}{T}.$$

By the Law of Large Number, the average of the shocks converges to zero as $T$ increases. Therefore, the average $\bar{y} = \sum_{t=1}^{T} y_t/T$ is a consistent estimate of $m(x)$. That the average $\bar{y}$ provides a consistent estimate of $m(x)$ or, alternatively, that the average of shocks converges to zero shows the power of smoothing.

In financial time series, we do not have repeated observations available at $X = x$. What we observed are $\{(y_t, x_t)\}$ for $t = 1, \ldots, T$. But if the function $m(.)$ is sufficiently smooth, then the value of $Y_t$ for which $X_t \approx x$ continues to provide accurate approximation of $m(x)$. The value of $Y_t$ for which $X_t$ is far away from $x$ provides less accurate approximation for $m(x)$. As a compromise, one can use a weighted

average of $y_t$ instead of the simple average to estimate $m(x)$. The weight should be larger for those $Y_t$ with $X_t$ close to $x$ and smaller for those $Y_t$ with $X_t$ far away from $x$. Mathematically, the estimate of $m(x)$ for a given $x$ can be written as

$$\hat{m}(x) = \frac{1}{T}\sum_{t=1}^{T} w_t(x) y_t, \tag{4.18}$$

where the weights $w_t(x)$ are larger for those $y_t$ with $x_t$ close to $x$ and smaller for those $y_t$ with $x_t$ far away from $x$.

From Eq. (4.18), the estimate $\hat{m}(x)$ is simply a *local weighted average* with weights determined by two factors. The first factor is the distance measure (i.e., the distance between $x_t$ and $x$). The second factor is the assignment of weight for a given distance. Different ways to determine the distance between $x_t$ and $x$ and to assign the weight using the distance give rise to different nonparametric methods. In what follows, we discuss the commonly used kernel regression and local linear regression methods.

#### *4.1.5.1 Kernel Regression*

Kernel regression is perhaps the most commonly used nonparametric method in smoothing. The weights here are determined by a *kernel*, which is typically a probability density function, is denoted by $K(x)$, and satisfies

$$K(x) \geq 0, \quad \int K(z)\, dz = 1.$$

However, to increase the flexibility in distance measure, one often rescales the kernel using a variable $h > 0$, which is referred to as the *bandwidth*. The rescaled kernel becomes

$$K_h(x) = \frac{1}{h} K(x/h), \quad \int K_h(z)\, dz = 1. \tag{4.19}$$

The weight function can now be defined as

$$w_t(x) = \frac{K_h(x - x_t)}{\frac{1}{T}\sum_{t=1}^{T} K_h(x - x_t)}, \tag{4.20}$$

where the denominator is a normalization constant that makes the smoother adaptive to the local intensity of the $X$ variable and ensures the weights sum to $T$. Plugging Eq. (4.20) into the smoothing formula (4.18), we have the well-known Nadaraya–Watson kernel estimator

$$\hat{m}(x) = \frac{1}{T}\sum_{t=1}^{T} w_t(x) y_t = \frac{\sum_{t=1}^{T} K_h(x - x_t) y_t}{\sum_{t=1}^{T} K_h(x - x_t)}; \tag{4.21}$$

see Nadaraya (1964) and Watson (1964). In practice, many choices are available for the kernel $K(x)$. However, theoretical and practical considerations lead to a few choices, including the Gaussian kernel

$$K_h(x) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{x^2}{2h^2}\right)$$

and the *Epanechnikov* kernel (Epanechnikov, 1969)

$$K_h(x) = \frac{0.75}{h}\left(1 - \frac{x^2}{h^2}\right) I\left(|\frac{x}{h}| \leq 1\right),$$

where $I(A)$ is an indicator such that $I(A) = 1$ if $A$ holds and $I(A) = 0$ otherwise. Figure 4.4 shows the Gaussian and Epanechnikov kernels for $h = 1$.

To understand the role played by the bandwidth $h$, we evaluate the Nadaraya–Watson estimator with the Epanechnikov kernel at the observed values $\{x_t\}$ and consider two extremes. First, if $h \to 0$, then

$$\hat{m}(x_t) \to \frac{K_h(0)y_t}{K_h(0)} = y_t,$$

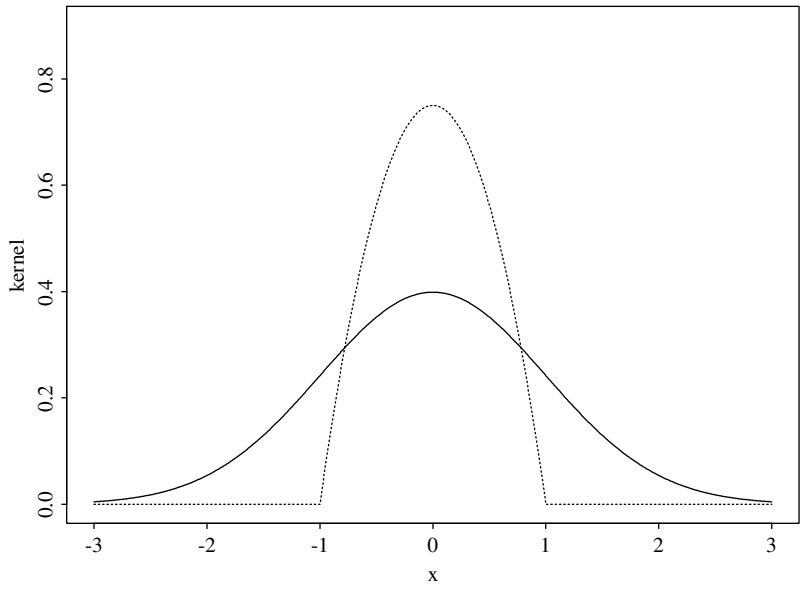


**Figure 4.4.** Standard normal kernel (solid line) and Epanechnikov kernel (dashed line) with bandwidth $h = 1$.

indicating that small bandwidths reproduce the data. Second, if $h \to \infty$, then

$$\hat{m}(x_t) \to \frac{\sum_{t=1}^{T} K_h(0) y_t}{\sum_{t=1}^{T} K_h(0)} = \frac{1}{T} \sum_{t=1}^{T} y_t = \bar{y},$$

suggesting that large bandwidths lead to an oversmoothed curve—the sample mean. In general, the bandwidth function $h$ acts as follows. If $h$ is very small, then the weights focus on a few observations that are in the neighborhood around each $x_t$. If $h$ is very large, then the weights will spread over larger neighborhoods of $x_t$. Consequently, the choice of $h$ plays an important role in kernel regression. This is the well-known problem of bandwidth selection in kernel regression.

#### *4.1.5.2 Bandwidth Selection*

There are several approaches for bandwidth selection; see Härdle (1990). The first approach is the plug-in method, which is based on the asymptotic expansion of the mean squared error (MSE) for kernel smoothers

$$E[\hat{m}(x) - m(x)]^2,$$

where $m(.)$ is the true function. Under some regularity conditions, one can derive the *optimal bandwidth* that minimizes the MSE. The optimal bandwidth typically depends on several unknown quantities that must be estimated from the data with some preliminary smoothing. Several iterations are often needed to obtain a reasonable estimate of the optimal bandwidth. In practice, the choice of preliminary smoothing can become a problem.

The second approach to bandwidth selection is the leave-one-out *cross-validation*. First, one observation $(x_j, y_j)$ is left out. The remaining $T - 1$ data points are used to obtain the following smoother at $x_j$

$$\hat{m}_{h,j}(x_j) = \frac{1}{T-1} \sum_{t \neq j} w_t(x_j) y_t,$$

which is an estimate of $y_j$. Second, perform Step-1 for $j = 1, \ldots, T$ and define the function

$$CV(h) = \frac{1}{T} \sum_{j=1}^{T} [y_j - \hat{m}_{h,j}(x_j)]^2 W(x_j),$$

where $W(.)$ is a non-negative weight function that can be used to down-weight the boundary points if necessary. Decreasing the weights assigned to data points close to the boundary is needed because those points often have fewer neighboring observations. The function $CV(h)$ is called the cross-validation function because it validates the ability of the smoother to predict $\{y_t\}_{t=1}^{T}$. One chooses the bandwidth $h$ that minimizes the $CV(.)$ function.

#### *4.1.5.3 Local Linear Regression Method*

Assume that the second derivative of $m(.)$ in model (4.17) exists and is continuous at $x$, where $x$ is a given point in the support of $m(.)$. Denote the data available by $\{(y_t, x_t)\}_{t=1}^{T}$. The local linear regression method to nonparametric regression is to find $a$ and $b$ that minimize

$$L(a, b) = \sum_{t=1}^{T} [y_t - a - b(x - x_t)]^2 K_h(x - x_t), \tag{4.22}$$

where $K_h(.)$ is a kernel function defined in Eq. (4.19) and $h$ is a bandwidth. Denote the resulting value of $a$ by $\hat{a}$. The estimate of $m(x)$ is then defined as $\hat{a}$.

Under the least squares theory, Eq. (4.22) is a weighted least squares problem and one can derive a closed-form solution for $a$. Specifically, taking the partial derivatives of $L(a, b)$ with respect to both $a$ and $b$ and equating the derivatives to zero, we have a system of two equations with two unknowns:

$$\begin{aligned}
\sum_{t=1}^{T} K_h(x - x_t) y_t &= a \sum_{t=1}^{T} K_h(x - x_t) + b \sum_{t=1}^{T} (x - x_t) K_h(x - x_t) \\
\sum_{t=1}^{T} y_t (x - x_t) K_h(x - x_t) &= a \sum_{t=1}^{T} (x - x_t) K_h(x - x_t) \\
&\quad + b \sum_{t=1}^{T} (x - x_t)^2 K_h(x - x_t).
\end{aligned}$$

Define

$$s_{T,\ell} = \sum_{t=1}^{T} K_h(x - x_t)(x - x_t)^{\ell}, \quad \ell = 0, 1, 2.$$

The prior system of equations becomes

$$\begin{bmatrix} s_{T,0} & s_{T,1} \\ s_{T,1} & s_{T,2} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{T} K_h(x - x_t) y_t \\ \sum_{t=1}^{T} (x - x_t) K_h(x - x_t) y_t \end{bmatrix}.$$

Consequently, we have

$$\hat{a} = \frac{s_{T,2} \sum_{t=1}^{T} K_h(x - x_t) y_t - s_{T,1} \sum_{t=1}^{T} (x - x_t) K_h(x - x_t) y_t}{s_{T,0} s_{T,2} - s_{T,1}^2}.$$

The numerator and denominator of the prior fraction can be further simplified as

$$s_{T,2}\sum_{t=1}^{T}K_h(x-x_t)y_t - s_{T,1}\sum_{t=1}^{T}(x-x_t)K_h(x-x_t)y_t$$
$$= \sum_{t=1}^{T}[K_h(x-x_t)(s_{T,2}-(x-x_t)s_{T,1})]y_t.$$

$$s_{T,0}s_{T,2} - s_{T,1}^2 = \sum_{t=1}^{T}K_h(x-x_t)s_{T,2} - \sum_{t=1}^{T}(x-x_t)K_h(x-x_t)s_{T,1}$$
$$= \sum_{t=1}^{T}K_h(x-x_t)[s_{T,2}-(x-x_t)s_{T,1}].$$

In summary, we have

$$\hat{a} = \frac{\sum_{t=1}^{T} w_t y_t}{\sum_{t=1}^{T} w_t}, \tag{4.23}$$

where $w_t$ is defined as

$$w_t = K_h(x-x_t)[s_{T,2}-(x-x_t)s_{T,1}].$$

In practice, to avoid possible zero in the denominator, we use $\hat{m}(x)$ next to estimate $m(x)$

$$\hat{m}(x) = \frac{\sum_{t=1}^{T} w_t y_t}{\sum_{t=1}^{T} w_t + \frac{1}{T^2}}. \tag{4.24}$$

Notice that a nice feature of Eq. (4.24) is that the weight $w_t$ satisfies

$$\sum_{t=1}^{T}(x-x_t)w_t = 0.$$

Also, if one assumes that $m(.)$ of Eq. (4.17) has the first derivative and finds the minimizer of

$$\sum_{t=1}^{T}(y_t-a)^2K_h(x-x_t),$$

then the resulting estimator is the Nadaraya–Watson estimator mentioned earlier. In general, if one assumes that $m(x)$ has a bounded $k$-th derivative, then one can replace the linear polynomial in Eq. (4.22) by a $(k-1)$-order polynomial. We refer to the estimator in Eq. (4.24) as the local linear regression smoother. Fan (1993) shows

that, under some regularity conditions, the local linear regression estimator has some important sampling properties. The selection of bandwidth can be carried out via the same methods as before.

#### *4.1.5.4 Time Series Application*

In time series analysis, the explanatory variables are often the lagged values of the series. Consider the simple case of a single explanatory variable. Here model (4.17) becomes

$$x_t = m(x_{t-1}) + a_t,$$

and the kernel regression and local linear regression method discussed before are directly applicable. When multiple explanatory variables exist, some modifications are needed to implement the nonparametric methods. For the kernel regression, one can use a multivariate kernel such as a multivariate normal density function with a prespecified covariance matrix:

$$K_h(\boldsymbol{x}) = \frac{1}{(h\sqrt{2\pi})^p |\,\boldsymbol{\Sigma}\,|^{1/2}} \exp\left(-\frac{1}{2h^2}\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right),$$

where $p$ is the number of explanatory variables and $\boldsymbol{\Sigma}$ is a prespecified positive-definite matrix. Alternatively, one can use the product of univariate kernel functions as a multivariate kernel—for example,

$$K_h(\boldsymbol{x}) = \prod_{i=1}^{p} \frac{0.75}{h_i}\left(1 - \frac{x_i^2}{h_i^2}\right) I\left(\left|\frac{x_i}{h_i}\right| < 1\right).$$

This latter approach is simple, but it overlooks the relationship between the explanatory variables.

The following nonlinear models are derived with the help of nonparametric methods.

### 4.1.6 Functional Coefficient AR Model

Recent advances in nonparametric techniques enable researchers to relax parametric constraints in proposing nonlinear models. In some cases, nonparametric methods are used in a preliminary study to help select a parametric nonlinear model. This is the approach taken by Chen and Tsay (1993a) in proposing the functional-coefficient autoregressive (FAR) model that can be written as

$$x_t = f_1(X_{t-1})x_{t-1} + \cdots + f_p(X_{t-1})x_{t-p} + a_t, \tag{4.25}$$

where $X_{t-1} = (x_{t-1}, \ldots, x_{t-k})'$ is a vector of lagged values of $x_t$. If necessary, $X_{t-1}$ may also include other explanatory variables available at time $t-1$. The functions

$f_i(.)$ of Eq. (4.25) are assumed to be continuous, even twice differentiable, almost surely with respect to their arguments. Most of the nonlinear models discussed before are special cases of the FAR model. In application, one can use nonparametric methods such as kernel regression or local linear regression to estimate the functional coefficients $f_i(.)$, especially when the dimension of $X_{t-1}$ is low (e.g., $X_{t-1}$ is a scalar). Recently, Cai, Fan, and Yao (1999) apply the local linear regression method to estimate $f_i(.)$ and show that substantial improvements in 1-step ahead forecasts can be achieved by using FAR models.

### 4.1.7 Nonlinear Additive AR Model

A major difficulty in applying nonparametric methods to nonlinear time series analysis is the "curse of dimensionality." Consider a general nonlinear AR($p$) process $x_t = f(x_{t-1}, \ldots, x_{t-p}) + a_t$. A direct application of nonparametric methods to estimate $f(.)$ would require $p$-dimensional smoothing, which is hard to do when $p$ is large, especially if the number of data points is not large. A simple, yet effective way to overcome this difficulty is to entertain an additive model that only requires lower dimensional smoothing. A time series $x_t$ follows a nonlinear additive AR (NAAR) model if

$$x_t = f_0(t) + \sum_{i=1}^{p} f_i(x_{t-i}) + a_t, \tag{4.26}$$

where $f_i(.)$s are continuous functions almost surely. Because each function $f_i(.)$ has a single argument, it can be estimated nonparametrically using one-dimensional smoothing techniques and hence avoids the curse of dimensionality. In application, an iterative estimation method that estimates $f_i(.)$ nonparametrically conditioned on estimates of $f_j(.)$ for all $j \neq i$ is used to estimate a NAAR model; see Chen and Tsay (1993b) for further details and examples of NAAR models.

The additivity assumption is rather restrictive and needs to be examined carefully in application. Chen, Liu, and Tsay (1995) consider test statistics for checking the additivity assumption.

### 4.1.8 Nonlinear State-Space Model

Making using of recent advances in MCMC methods (Gelfand and Smith, 1990), Carlin, Polson, and Stoffer (1992) propose a Monte Carlo approach for nonlinear state-space modeling. The model considered is

$$S_t = f_t(S_{t-1}) + u_t, \quad x_t = g_t(S_t) + v_t, \tag{4.27}$$

where $S_t$ is the state vector, $f_t(.)$ and $g_t(.)$ are known functions depending on some unknown parameters, $\{u_t\}$ is a sequence of iid multivariate random vectors with zero mean and non-negative definite covariance matrix $\boldsymbol{\Sigma}_u$, $\{v_t\}$ is a sequence of iid ran-

dom variables with mean zero and variance $\sigma_v^2$, and $\{u_t\}$ is independent of $\{v_t\}$. Monte Carlo techniques are employed to handle the nonlinear evolution of the state transition equation because the whole conditional distribution function of $S_t$ given $S_{t-1}$ is needed for a nonlinear system. Other numerical smoothing methods for nonlinear time series analysis have been considered by Kitagawa (1998) and the references therein. MCMC methods (or computing-intensive numerical methods) are powerful tools for nonlinear time series analysis. Their potential has not been fully explored. However, the assumption of knowing $f_t(.)$ and $g_t(.)$ in model (4.27) may hinder practical use of the proposed method. A possible solution to overcome this limitation is to use nonparametric methods such as the analyses considered in FAR and NAAR models to specify $f_t(.)$ and $g_t(.)$ before using nonlinear state-space models.

### 4.1.9 Neural Networks

A popular topic in modern data analysis is neural network, which can be classified as a semiparametric method. The literature on neural network is enormous, and its application spreads over many scientific areas with varying degrees of success; see section 2 of Ripley (1993) for a list of applications and section 10 for remarks concerning its application in finance. Cheng and Titterington (1994) provide information on neural networks from a statistical viewpoint. In this subsection, we focus solely on the *feed-forward* neural networks in which inputs are connected to one or more *neurons*, or *nodes*, in the input layer, and these nodes are connected forward to further layers until they reach the output layer. Figure 4.5 shows an example of a simple feed-forward network for univariate time series analysis with one hidden layer. The input layer has two nodes, and the hidden layer has three. The input nodes are connected forward to each and every node in the hidden layer, and these hidden nodes are connected to the single node in the output layer. We call the network a 2-3-1 feed-forward network. More complicated neural networks, including those with feedback connections, have been proposed in the literature, but the feed-forward networks are most relevant to our study.

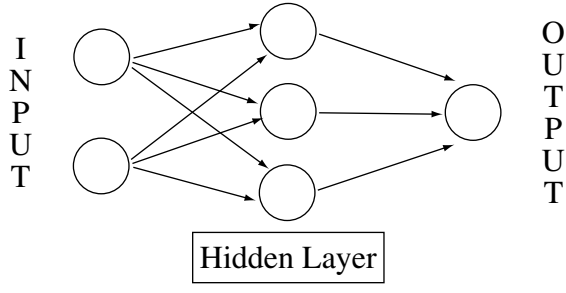


**Figure 4.5.** A feed-forward neural network with one hidden layer for univariate time series analysis.

#### *4.1.9.1 Feed-Forward Neural Networks*

A neural network processes information from one layer to the next by an "activation function." Consider a feed-forward network with one hidden layer. The $j$th node in the hidden layer is defined as

$$h_j = f_j\left(\alpha_{0j} + \sum_{i \to j} w_{ij} x_i\right) \tag{4.28}$$

where $x_i$ is the value of the $i$th input node, $f_j(.)$ is an activation function typically taken to be the logistic function

$$f_j(z) = \frac{\exp(z)}{1 + \exp(z)},$$

$\alpha_{0j}$ is called the bias, the summation $i \to j$ means summing over all input nodes feeding to $j$, and $w_{ij}$ are the weights. For illustration, the $j$th node of the hidden layer of the 2-3-1 feed-forward network in Figure 4.5 is

$$h_j = \frac{\exp(\alpha_{0j} + w_{1j} x_1 + w_{2j} x_2)}{1 + \exp(\alpha_{0j} + w_{1j} x_1 + w_{2j} x_2)}, \quad j = 1, 2, 3. \tag{4.29}$$

For the output layer, the node is defined as

$$o = f_o\left(\alpha_{0o} + \sum_{j \to o} w_{jo} h_j\right), \tag{4.30}$$

where the activation function $f_o(.)$ is either linear or a Heaviside function. If $f_o(.)$ is linear, then

$$o = \alpha_{0o} + \sum_{j=1}^{k} w_{jo} h_j,$$

where $k$ is the number of nodes in the hidden layer. By a Heaviside function, we mean $f_o(z) = 1$ if $z > 0$ and $f_o(z) = 0$ otherwise. A neuron with a Heaviside function is called a *threshold neuron*, with "1" denoting that the neuron fires its message. For example, the output of the 2-3-1 network in Figure 4.5 is

$$o = \alpha_{0o} + w_{1o} h_1 + w_{2o} h_2 + w_{3o} h_3$$

if the activation function is linear; it is

$$o = \begin{cases} 1 & \text{if } \alpha_{0o} + w_{1o} h_1 + w_{2o} h_2 + w_{3o} h_3 > 0 \\ 0 & \text{if } \alpha_{0o} + w_{1o} h_1 + w_{2o} h_2 + w_{3o} h_3 \leq 0 \end{cases}$$

if $f_o(.)$ is a Heaviside function.

Combining the layers, the output of a feed-forward neural network can be written as

$$o = f_o\left[\alpha_{0o} + \sum_{j\to o} w_{jo} f_j\left(\alpha_{0j} + \sum_{i\to j} w_{ij} x_i\right)\right]. \tag{4.31}$$

If one also allows for direct connections from the input layer to the output layer, then the network becomes

$$o = f_o\left[\alpha_{0o} + \sum_{i\to o} \alpha_{io} x_i + \sum_{j\to o} w_{jo} f_j\left(\alpha_{0j} + \sum_{i\to j} w_{ij} x_i\right)\right], \tag{4.32}$$

where the first summation is summing over the input nodes. When the activation function of the output layer is linear, the direct connections from the input nodes to the output node represent a linear function between the inputs and output. Consequently, in this particular case model (4.32) is a generalization of linear models. For the 2-3-1 network in Figure 4.5, if the output activation function is linear, then Eq. (4.31) becomes

$$o = \alpha_{0o} + \sum_{j=1}^{3} w_{jo} h_j,$$

where $h_j$ is given in Eq. (4.29). The network thus has 13 parameters. If Eq. (4.32) is used, then the network becomes

$$o = \alpha_{0o} + \sum_{i=1}^{2} \alpha_{io} x_i + \sum_{j=1}^{3} w_{jo} h_j,$$

where again $h_j$ is given in Eq. (4.29). The number of parameters of the network increases to 15.

We refer to the function in Eq. (4.31) or (4.32) as a semiparametric function because its functional form is known, but the number of nodes and their biases and weights are unknown. The direct connections from the input layer to the output layer in Eq. (4.32) mean that the network can skip the hidden layer. We refer to such a network as a *skip-layer* feed-forward network.

Feed-forward networks are known as *multilayer percetrons* in the neural network literature. They can approximate any continuous function uniformly on compact sets by increasing the number of nodes in the hidden layer; see Hornik, Stinchcombe, and White (1989), Hornik (1993), and Chen and Chen (1995). This property of neural networks is the universal approximation property of the multilayer percetrons. In short, feed-forward neural networks with a hidden layer can be seen as a way to parametrize a general continuous nonlinear function.

#### *4.1.9.2 Training and Forecasting*

Application of neural networks involves two steps. The first step is to *train* the network (i.e., to build a network, including determining the number of nodes and estimating their biases and weights). The second step is inference, especially forecasting. The data are often divided into two nonoverlapping subsamples in the training stage. The first subsample is used to estimate the parameters of a given feed-forward neural network. The network so built is then used in the second subsample to perform forecasting and compute its forecasting accuracy. By comparing the forecasting performance, one selects the network that outperforms the others as the "best" network for making inference. This is the idea of cross-validation widely used in statistical model selection. Other model selection methods are also available.

In a time series application, let $\{(r_t, \boldsymbol{x}_t) \mid t = 1, \ldots, T\}$ be the available data for network training, where $\boldsymbol{x}_t$ denotes the vector of inputs and $r_t$ is the series of interest (e.g., log returns of an asset). For a given network, let $o_t$ be the output of the network with input $\boldsymbol{x}_t$; see Eq. (4.32). Training a neural network amounts to choosing its biases and weights to minimize some fitting criterion—for example, the least squares

$$S^2 = \sum_{t=1}^{T} (r_t - o_t)^2.$$

This is a nonlinear estimation problem that can be solved by several iterative methods. To ensure the smoothness of the fitted function, some additional constraints can be added to the prior minimization problem. In the neural network literature, *Back Propagation* (BP) learning algorithm is a popular method for network training. The BP method, introduced by Bryson and Ho (1969), works backward starting with the output layer and uses a gradient rule to modify the biases and weights iteratively. Appendix 2A of Ripley (1993) provides a derivation of Back Propagation. Once a feed-forward neural network is built, it can be used to compute forecasts in the forecasting subsample.

**Example 4.5.** To illustrate applications of neural network in finance, we consider the monthly log returns, in percentages and including dividends, for IBM stock from January 1926 to December 1999. We divide the data into two subsamples. The first subsample consisting of returns from January 1926 to December 1997 for 864 observations is used for modeling. Using model (4.32) with three inputs and two nodes in the hidden layer, we obtain a 3-2-1 network for the series. The three inputs are $r_{t-1}$, $r_{t-2}$, and $r_{t-3}$, and the biases and weights are given next:

$$\hat{r}_t = 3.22 - 1.81 f_1(\boldsymbol{r}_{t-1}) - 2.28 f_2(\boldsymbol{r}_{t-1}) - 0.09 r_{t-1} - 0.05 r_{t-2} - 0.12 r_{t-3}, \quad (4.33)$$

where $\boldsymbol{r}_{t-1} = (r_{t-1}, r_{t-2}, r_{t-3})$ and the two logistic functions are

$$f_1(\boldsymbol{r}_{t-1}) = \frac{\exp(-8.34 - 18.97 r_{t-1} + 2.17 r_{t-2} - 19.17 r_{t-3})}{1 + \exp(-8.34 - 18.97 r_{t-1} + 2.17 r_{t-2} - 19.17 r_{t-3})}$$

$$f_2(\boldsymbol{r}_{t-1}) = \frac{\exp(39.25 - 22.17r_{t-1} - 17.34r_{t-2} - 5.98r_{t-3})}{1 + \exp(39.25 - 22.17r_{t-1} - 17.34r_{t-2} - 5.98r_{t-3})}.$$

The standard error of the residuals for the prior model is 6.56. For comparison, we also built an AR model for the data and obtained

$$r_t = 1.101 + 0.077r_{t-1} + a_t, \quad \sigma_a = 6.61. \tag{4.34}$$

The residual standard error is slightly greater than that of the feed-forward model in Eq. (4.33).

***Forecast Comparison***

The monthly returns of IBM stock in 1998 and 1999 form the second subsample and are used to evaluate the out-of-sample forecasting performance of neural networks. As a benchmark for comparison, we use the sample mean of $r_t$ in the first subsample as the 1-step ahead forecast for all the monthly returns in the second subsample. This corresponds to assuming that the log monthly price of IBM stock follows a random walk with a drift. The mean squared forecast error (MSE) of this benchmark model is 91.85. For the AR(1) model in Eq. (4.34), the MSE of 1-step ahead forecasts is 91.70. Thus, the AR(1) model outperforms slightly the benchmark. For the 3-2-1 feed-forward network in Eq. (4.33), the MSE is 91.74, which is essentially the same as that of the AR(1) model.

***Remark:*** The estimation of feed-forward networks is done by using the S-Plus program with default starting weights; see Venables and Ripley (1999) for more information. Our limited experience shows that the estimation results vary. For the IBM stock returns used in Example 4.5, the out-of-sample MSE for a 3-2-1 network can be as low as 89.46 and as high as 93.65. If we change the number of nodes in the hidden layer, the range for the MSE becomes even wider. The S-Plus commands used in Example 4.5 are given in Appendix B.

**Example 4.6.** Nice features of the feed-forward networks include its flexibility and wide applicability. For illustration, we use the network with a Heaviside activation function for the output layer to forecast the direction of price movement for IBM stock considered in Example 4.5. Define a direction variable as

$$d_t = \begin{cases} 1 & \text{if } r_t \geq 0 \\ 0 & \text{if } r_t < 0. \end{cases}$$

We use eight input nodes consisting of the first four lagged values of both $r_t$ and $d_t$ and four nodes in the hidden layer to build an 8-4-1 feed-forward network for $d_t$ in the first subsample. The resulting network is then used to compute the 1-step ahead probability of an "upward movement" (i.e., a positive return) for the following month in the second subsample. Figure 4.6 shows a typical output of probability

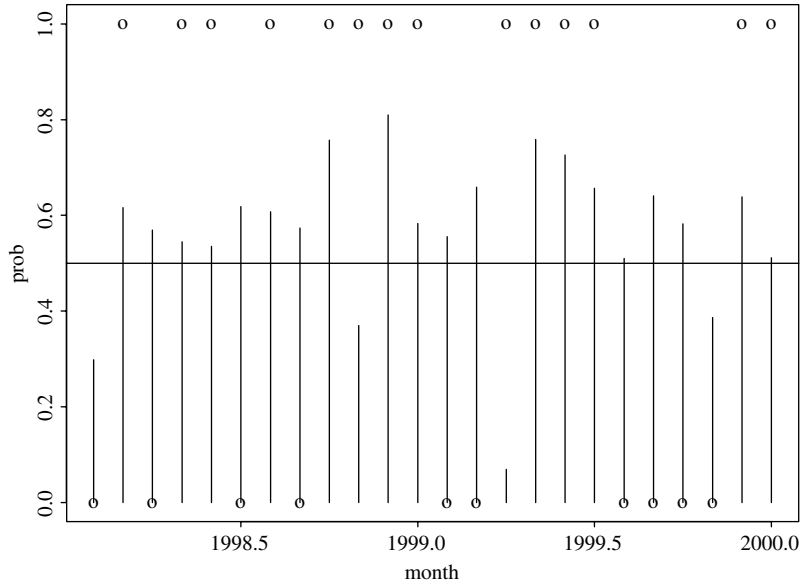


**Figure 4.6.** One-step ahead probability forecasts for a positive monthly return for IBM stock using an 8-4-1 feed-forward neural network. The forecasting period is from January 1998 to December 1999.

forecasts and the actual directions in the second subsample with the latter denoted by "o." A horizontal line of 0.5 is added to the plot. If we take a rigid approach by letting $\hat{d}_t = 1$ if the probability forecast is greater than or equal to 0.5 and $\hat{d}_t = 0$ otherwise, then the neural network has a successful rate of 0.58. The success rate of the network varies substantially from one estimation to another, and the network uses 49 parameters. To gain more insight, we did a simulation study of running the 8-4-1 feed-forward network 500 times and computed the number of errors in predicting the upward and downward movement using the same method as before. The mean and median of errors over the 500 runs are 11.28 and 11, respectively, whereas the maximum and minimum numbers of errors are 18 and 4. For comparison, we also did a simulation with 500 runs using a random walk with a drift—that is,

$$\hat{d}_t = \begin{cases} 1 & \text{if } \hat{r}_t = 1.19 + \epsilon_t \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where 1.19 is the average monthly log return for IBM stock from January 1926 to December 1997 and $\{\epsilon_t\}$ is a sequence of iid $N(0, 1)$ random variables. The mean and median of the number of forecast errors become 10.53 and 11, whereas the maximum and minimum numbers of errors are 17 and 5, respectively. Figure 4.7 shows the histograms of the number of forecast errors for the two simulations. The results show

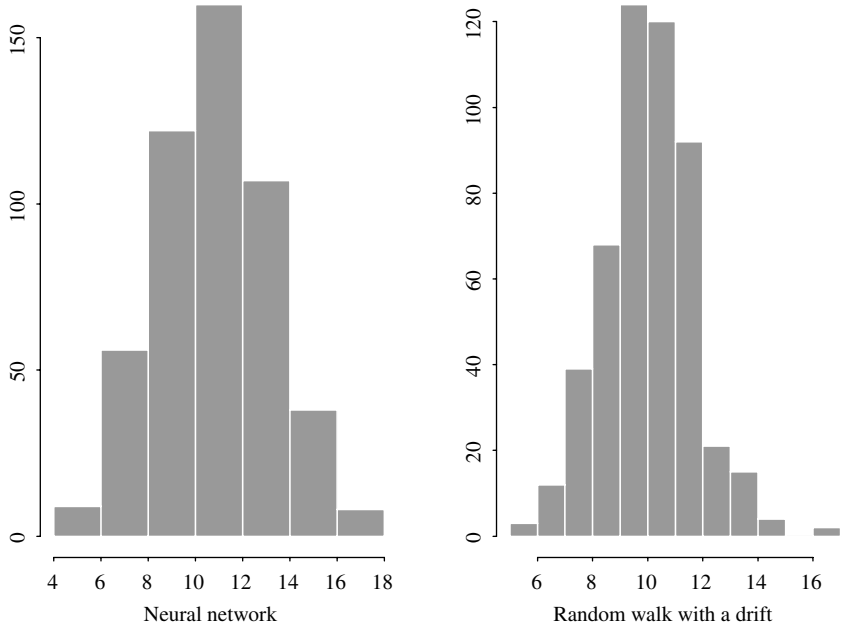


**Figure 4.7.** Histograms of the number of forecasting errors for the directional movements of monthly log returns of IBM stock. The forecasting period is from January 1998 to December 1999.

that the 8-4-1 feed-forward neural network does not outperform the simple model that assumes a random walk with a drift for the monthly log price of IBM stock.

## 4.2 NONLINEARITY TESTS

In this section, we discuss some nonlinearity tests available in the literature that have decent power against the nonlinear models considered in Section 4.1. The tests discussed include both parametric and nonparametric statistics. The Ljung–Box statistics of squared residuals, the bispectral test, and the BDS test are nonparametric methods. The RESET test (Ramsey, 1969), the $F$ tests of Tsay (1986, 1989), and other Lagrange multiplier and likelihood ratio tests depend on specific parametric functions. Because nonlinearity may occur in many ways, there exists no single test that dominates the others in detecting nonlinearity.

### 4.2.1 Nonparametric Tests

Under the null hypothesis of linearity, residuals of a properly specified linear model should be independent. Any violation of independence in the residuals indicates inadequacy of the entertained model, including the linearity assumption. This is the basic

idea behind various nonlinearity tests. In particular, some of the nonlinearity tests are designed to check for possible violation in quadratic forms of the underlying time series.

#### 4.2.1.1 Q-Statistic of Squared Residuals

McLeod and Li (1983) apply the Ljung–Box statistics to the squared residuals of an ARMA$(p, q)$ model to check for model inadequacy. The test statistic is

$$Q(m) = T(T+2)\sum_{i=1}^{m} \frac{\hat{\rho}_i^2(a_t^2)}{T-i},$$

where $T$ is the sample size, $m$ is a properly chosen number of autocorrelations used in the test, $a_t$ denotes the residual series, and $\hat{\rho}_i(a_t^2)$ is the lag-$i$ ACF of $a_t^2$. If the entertained linear model is adequate, $Q(m)$ is asymptotically a chi-squared random variable with $m - p - q$ degrees of freedom. As mentioned in Chapter 3, the prior Q-statistic is useful in detecting conditional heteroscedasticity of $a_t$ and is asymptotically equivalent to the Lagrange multiplier test statistic of Engle (1982) for ARCH models; see subsection 3.3.3. The null hypothesis of the statistics is $H_o : \beta_1 = \cdots = \beta_m = 0$, where $\beta_i$ is the coefficient of $a_{t-i}^2$ in the linear regression

$$a_t^2 = \beta_0 + \beta_1 a_{t-1}^2 + \cdots + \beta_m a_{t-m}^2 + e_t$$

for $t = m+1, \ldots, T$. Because the statistic is computed from residuals (not directly from the observed returns), the number of degrees of freedom is $m - p - q$.

#### 4.2.1.2 Bispectral Test

This test can be used to test for linearity and Gaussianity. It depends on the result that a properly normalized bispectrum of a linear time series is constant over all frequencies and that the constant is zero under normality. The bispectrum of a time series is the Fourier transform of its third-order moments. For a stationary time series $x_t$ in Eq. (4.1), the third-order moment is defined as

$$c(u, v) = g\sum_{k=-\infty}^{\infty} \psi_k \psi_{k+u} \psi_{k+v}, \tag{4.35}$$

where $u$ and $v$ are integers, $g = E(a_t^3)$, $\psi_0 = 1$, and $\psi_k = 0$ for $k < 0$. Taking Fourier transforms of Eq. (4.35), we have

$$b_3(w_1, w_2) = \frac{g}{4\pi^2}\Gamma[-(w_1 + w_2)]\Gamma(w_1)\Gamma(w_2), \tag{4.36}$$

where $\Gamma(w) = \sum_{u=0}^{\infty} \psi_u \exp(-iwu)$ with $i = \sqrt{-1}$, and $w_i$ are frequencies. Yet the spectral density function of $x_t$ is given by

$$p(w) = \frac{\sigma_a^2}{2\pi} |\, \Gamma(w) \,|^2,$$

where $w$ denotes the frequency. Consequently, the function

$$b(w_1, w_2) = \frac{|\, b_3(w_1, w_2) \,|^2}{p(w_1)p(w_2)p(w_1 + w_2)} = \text{constant for all } (w_1, w_2). \tag{4.37}$$

The bispectrum test makes use of the property in Eq. (4.37). Basically, it estimates the function $b(w_1, w_2)$ in Eq. (4.37) over a suitably chosen grid of points and applies a test statistic similar to Hotelling's $T^2$ statistic to check the constancy of $b(w_1, w_2)$. For a linear Gaussian series, $E(a_t^3) = g = 0$ so that the bispectrum is zero for all frequencies $(w_1, w_2)$. For further details of the bispectral test, see Priestley (1988), Subba Rao and Gabr (1984), and Hinich (1982). Limited experience shows that the test has decent power when the sample size is large.

#### *4.2.1.3 BDS Statistic*

Brock, Dechert, and Scheinkman (1987) propose a test statistic, commonly referred to as the *BDS test*, to detect the iid assumption of a time series. The statistic is, therefore, different from other test statistics discussed because the latter mainly focus on either the second- or third-order properties of $x_t$. The basic idea of the BDS test is to make use of "correlation integral" popular in chaotic time series analysis. Given a $k$-dimensional time series $X_t$ and observations $\{X_t\}_{t=1}^{T_k}$, define the correlation integral as

$$C_k(\delta) = \lim_{T_k \to \infty} \frac{2}{T_k(T_k - 1)} \sum_{i<j} I_\delta(X_i, X_j), \tag{4.38}$$

where $I_\delta(u, v)$ is an indicator variable that equals one if $\| u - v \| < \delta$, and zero otherwise, where $\| \, . \, \|$ is the supnorm. The correlation integral measures the fraction of data pairs of $\{X_t\}$ that are within a distance of $\delta$ from each other. Consider next a time series $x_t$. Construct $k$-dimensional vectors $X_t^k = (x_t, x_{t+1}, \ldots, x_{t+k-1})'$, which are called *k-histories*. The idea of the BDS test is as follows. Treat a $k$-history as a point in the $k$-dimensional space. If $\{x_t\}_{t=1}^{T}$ are indeed iid random variables, then the $k$-histories $\{X_t\}_{t=1}^{T_k}$ should show no pattern in the $k$-dimensional space. Consequently, the correlation integrals should satisfy the relation $C_k(\delta) = [C_1(\delta)]^k$. Any departure from the prior relation suggests that $x_t$ are not iid. As a simple, but informative example, consider a sequence of iid random variables from the uniform distribution over [0, 1]. Let $[a, b]$ be a subinterval of [0, 1] and consider the "2-history" $(x_t, x_{t+1})$, which represents a point in the two-dimensional space. Under the iid assumption, the expected number of 2-histories in the subspace $[a, b] \times [a, b]$ should equal the square of the expected number of $x_t$ in $[a, b]$. This idea can be formally examined by using sample counterparts of correlation integrals. Define

$$C_\ell(\delta, T) = \frac{2}{T_k(T_k - 1)} \sum_{i<j} I_\delta(X_i^*, X_j^*), \quad \ell = 1, k,$$

where $T_\ell = T - \ell + 1$ and $X_i^* = x_i$ if $\ell = 1$ and $X_i^* = X_i^k$ if $\ell = k$. Under the null hypothesis that $\{x_t\}$ are iid with a nondegenerated distribution function $F(.)$, Brock, Dechert, and Scheinkman (1987) show that

$$C_k(\delta, T) \to [C_1(\delta)]^k \quad \text{with probability 1,} \quad \text{as} \quad T \to \infty$$

for any fixed $k$ and $\delta$. Furthermore, the statistic $\sqrt{T}\{C_k(\delta, T) - [C_1(\delta, T)]^k\}$ is asymptotically distributed as normal with mean zero and variance

$$\sigma_k^2(\delta) = 4\left[N^k + 2\sum_{j=1}^{k-1} N^{k-j}C^{2j} + (k-1)^2C^{2k} - k^2NC^{2k-2}\right],$$

where $C = \int [F(z+\delta) - F(z-\delta)]dF(z)$ and $N = \int [F(z+\delta) - F(z-\delta)]^2 dF(z)$. Note that $C_1(\delta, T)$ is a consistent estimate of $C$, and $N$ can be consistently estimated by

$$N(\delta, T) = \frac{6}{T_k(T_k - 1)(T_k - 2)} \sum_{t<s<u} I_\delta(x_t, x_s) I_\delta(x_s, x_u).$$

The BDS test statistic is then defined as

$$D_k(\delta, T) = \sqrt{T}\{C_k(\delta, T) - [C_1(\delta, T)]^k\}/\sigma_k(\delta, T), \tag{4.39}$$

where $\sigma_k(\delta, T)$ is obtained from $\sigma_k(\delta)$ when $C$ and $N$ are replaced by $C_1(\delta, T)$ and $N(\delta, T)$, respectively. This test statistic has a standard normal limiting distribution. For further discussion and examples of applying the BDS test, see Hsieh (1989) and Brock, Hsieh, and LeBaron (1991). In application, one should remove linear dependence, if any, from the data before applying the BDS test. The test may be sensitive to the choices of $\delta$ and $k$, especially when $k$ is large.

### 4.2.2 Parametric Tests

Turning to parametric tests, we consider the RESET test of Ramsey (1969) and its generalizations. We also discuss some test statistics for detecting threshold nonlinearity. To simplify the notation, we use vectors and matrices in the discussion. If necessary, readers may consult Appendix A of Chapter 8 for a brief review on vectors and matrices.

#### *4.2.2.1 The RESET Test*

Ramsey (1969) proposes a specification test for linear least squares regression analysis. The test is referred to as a RESET test and is readily applicable to linear AR

models. Consider the linear AR($p$) model

$$x_t = \boldsymbol{X}_{t-1}'\boldsymbol{\phi} + a_t, \tag{4.40}$$

where $\boldsymbol{X}_{t-1} = (1, x_{t-1}, \ldots, x_{t-p})'$ and $\boldsymbol{\phi} = (\phi_0, \phi_1, \ldots, \phi_p)'$. The first step of the RESET test is to obtain the least squares estimate $\hat{\boldsymbol{\phi}}$ of Eq. (4.40) and compute the fit $\hat{x}_t = \boldsymbol{X}_{t-1}'\widehat{\boldsymbol{\phi}}$, the residual $\hat{a}_t = x_t - \hat{x}_t$, and the sum of squared residuals $SSR_0 = \sum_{t=p+1}^{T} \hat{a}_t^2$, where $T$ is the sample size. In the second step, consider the linear regression

$$\hat{a}_t = \boldsymbol{X}_{t-1}'\boldsymbol{\alpha}_1 + \boldsymbol{M}_{t-1}'\boldsymbol{\alpha}_2 + v_t, \tag{4.41}$$

where $\boldsymbol{M}_{t-1} = (\hat{x}_t^2, \ldots, \hat{x}_t^{s+1})'$ for some $s \geq 1$, and compute the least squares residuals

$$\hat{v}_t = \hat{a}_t - \boldsymbol{X}_{t-1}'\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{M}_{t-1}'\widehat{\boldsymbol{\alpha}}_2$$

and the sum of squared residuals $SSR_1 = \sum_{t=p+1}^{T} \hat{v}_t^2$ of the regression. The basic idea of the RESET test is that if the linear AR($p$) model in Eq. (4.40) is adequate, then $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ of Eq. (4.41) should be zero. This can be tested by the usual $F$ statistic of Eq. (4.41) given by

$$F = \frac{(SSR_0 - SSR_1)/g}{SSR_1/(T - p - g)} \qquad \text{with} \quad g = s + p + 1 \tag{4.42}$$

which, under the linearity and normality assumption, has an $F$ distribution with degrees of freedom $g$ and $T - p - g$.

***Remark:*** Because $\hat{x}_t^k$ for $k = 2, \ldots, s + 1$ tend to be highly correlated with $\boldsymbol{X}_{t-1}$ and among themselves, principal components of $\boldsymbol{M}_{t-1}$ that are not co-linear with $\boldsymbol{X}_{t-1}$ are often used in fitting Eq. (4.41). Principal component analysis is a statistical tool for dimension reduction; see Chapter 8 for more information.

Kennan (1985) proposes a nonlinearity test for time series that uses $\hat{x}_t^2$ only and modifies the second step of the RESET test to avoid multicollinearity between $\hat{x}_t^2$ and $\boldsymbol{X}_{t-1}$. Specifically, the linear regression (4.41) is divided into two steps. In step 2(a), one removes linear dependence of $\hat{x}_t^2$ on $\boldsymbol{X}_{t-1}$ by fitting the regression

$$\hat{x}_t^2 = \boldsymbol{X}_{t-1}'\boldsymbol{\beta} + u_t$$

and obtaining the residual $\hat{u}_t = \hat{x}_t^2 - \boldsymbol{X}_{t-1}\widehat{\boldsymbol{\beta}}$. In step 2(b), consider the linear regression

$$\hat{a}_t = \hat{u}_t\alpha + v_t,$$

and obtain the sum of squared residuals $SSR_1 = \sum_{t=p+1}^{T}(\hat{a}_t - \hat{u}_t\hat{\alpha})^2 = \sum_{t=p+1}^{T}\hat{v}_t^2$ to test the null hypothesis $\alpha = 0$.

#### *4.2.2.2 The F Test*

To improve the power of Kennan's and RESET tests, Tsay (1986) uses a different choice of the regressor $\boldsymbol{M}_{t-1}$. Specifically, he suggests using $\boldsymbol{M}_{t-1} = \text{vech}(\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}')$, where vech($\boldsymbol{A}$) denotes the half-stacking vector of the matrix $\boldsymbol{A}$ using elements on and below the diagonal only; see Appendix B of Chapter 8 for more information about the operator. For example, if $p = 2$, then $\boldsymbol{M}_{t-1} = (x_{t-1}^2, x_{t-1}x_{t-2}, x_{t-2}^2)'$. The dimension of $\boldsymbol{M}_{t-1}$ is $p(p+1)/2$ for an AR($p$) model. In practice, the test is simply the usual partial $F$ statistic for testing $\boldsymbol{\alpha} = 0$ in the linear least squares regression

$$x_t = \boldsymbol{X}_{t-1}'\boldsymbol{\phi} + \boldsymbol{M}_{t-1}'\boldsymbol{\alpha} + e_t,$$

where $e_t$ denotes the error term. Under the assumption that $x_t$ is a linear AR($p$) process, the partial $F$ statistic follows an $F$ distribution with degrees of freedom $g$ and $T - p - g - 1$, where $g = p(p+1)/2$. We refer to this $F$ test as the *Ori-F test*. Luukkonen, Saikkonen, and Teräsvirta (1988) further extend the test by augmenting $\boldsymbol{M}_{t-1}$ with cubic terms $x_{t-i}^3$ for $i = 1, \ldots, p$.

#### *4.2.2.3 Threshold Test*

When the alternative model under study is a SETAR model, one can derive specific test statistics to increase the power of the test. One of the specific tests is the likelihood ratio statistic. This test, however, encounters the difficulty of undefined parameters under the null hypothesis of linearity because the threshold is undefined for a linear AR process. Another specific test seeks to transform testing threshold nonlinearity into detecting model changes. It is then interesting to discuss the differences between these two specific tests for threshold nonlinearity.

To simplify the discussion, let us consider the simple case that the alternative model is a two-regime SETAR model with threshold variable $x_{t-d}$. The null hypothesis is $H_o$: $x_t$ follows the linear AR($p$) model

$$x_t = \phi_0 + \sum_{i=1}^{p}\phi_i x_{t-i} + a_t, \tag{4.43}$$

whereas the alternative hypothesis is $H_a$: $x_t$ follows the SETAR model

$$x_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{p}\phi_i^{(1)}x_{t-i} + a_{1t} & \text{if } x_{t-d} < r_1 \\ \phi_0^{(2)} + \sum_{i=1}^{p}\phi_i^{(2)}x_{t-i} + a_{2t} & \text{if } x_{t-d} \geq r_1, \end{cases} \tag{4.44}$$

where $r_1$ is the threshold. For a given realization $\{x_t\}_{t=1}^{T}$ and assuming normality, let $l_0(\hat{\phi}, \hat{\sigma}_a^2)$ be the log likelihood function evaluated at the maximum likelihood estimates of $\phi = (\phi_0, \ldots, \phi_p)'$ and $\sigma_a^2$. This is easy to compute. The likelihood

function under the alternative is also easy to compute if the threshold $r_1$ is given. Let $l_1(r_1; \hat{\phi}_1, \hat{\sigma}_1^2; \hat{\phi}_2, \hat{\sigma}_2^2)$ be the log likelihood function evaluated at the maximum likelihood estimates of $\phi_i = (\phi_0^{(i)}, \ldots, \phi_p^{(i)})'$ and $\sigma_i^2$ conditioned on knowing the threshold $r_1$. The log likelihood ratio $l(r_1)$ defined as

$$l(r_1) = l_1(r_1; \hat{\phi}_1, \hat{\sigma}_1^2; \hat{\phi}_2, \hat{\sigma}_2^2) - l_0(\hat{\phi}, \hat{\sigma}_a^2),$$

is then a function of the threshold $r_1$, which is unknown. Yet under the null hypothesis, there is no threshold and $r_1$ is not defined. The parameter $r_1$ is referred to as a *nuisance parameter* under the null hypothesis. Consequently, the asymptotic distribution of the likelihood ratio is very different from that of the conventional likelihood ratio statistics. See Chan (1991) for further details and critical values of the test. A common approach is to use $l_{\max} = \sup_{v<r_1<u} l(r_1)$ as the test statistic, where $v$ and $u$ are prespecified lower and upper bounds of the threshold. Davis (1987) and Andrews and Ploberger (1994) provide further discussion on hypothesis testing involving nuisance parameters under the null hypothesis. Simulation is often used to obtain empirical critical values of the test statistic $l_{\max}$, which depends on the choices of $v$ and $u$. The average of $l(r_1)$ over $r_1 \in [v, u]$ is also considered by Andrews and Ploberger as a test statistic.

Tsay (1989) makes use of arranged autoregression and recursive estimation to derive an alternative test for threshold nonlinearity. The arranged autoregression seeks to transfer the SETAR model under the alternative hypothesis $H_a$ into a model change problem with the threshold $r_1$ serving as the change point. To see this, the SETAR model in Eq. (4.44) says that $x_t$ follows essentially two linear models depending on whether $x_{t-d} < r_1$ or $x_{t-d} \geq r_1$. For a realization $\{x_t\}_{t=1}^T$, $x_{t-d}$ can assume values $\{x_1, \ldots, x_{T-d}\}$. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(T-d)}$ be the ordered statistics of $\{x_t\}_{t=1}^{T-d}$ (i.e., arranging the observations in increasing order). The SETAR model can then be written as

$$x_{(j)+d} = \beta_0 + \sum_{i=1}^{p} \beta_i x_{(j)+d-i} + a_{(j)+d}, \quad j = 1, \ldots, T-d, \tag{4.45}$$

where $\beta_i = \phi_i^{(1)}$ if $x_{(j)} < r_1$ and $\beta_i = \phi_i^{(2)}$ if $x_{(j)} \geq r_1$. Consequently, the threshold $r_1$ is a change point for the linear regression in Eq. (4.45), and we refer to Eq. (4.45) as an arranged autoregression (in increasing order of the threshold $x_{t-d}$). Note that the arranged autoregression in Eq. (4.45) does not alter the dynamic dependence of $x_t$ on $x_{t-i}$ for $i = 1, \ldots, p$ because $x_{(j)+d}$ still depends on $x_{(j)+d-i}$ for $i = 1, \ldots, p$. What is done is simply to present the SETAR model in the threshold space instead of in the time space. That is, the equation with a smaller $x_{t-d}$ appears before that with a larger $x_{t-d}$. The threshold test of Tsay (1989) is obtained as follows.

- Step 1: Fit the Eq. (4.45) using $j = 1, \ldots, m$, where $m$ is a pre-specified positive integer (e.g., 30). Denote the least squares estimates of $\beta_i$ by $\hat{\beta}_{i,m}$, where $m$ denotes the number of data points used in estimation.

- Step 2: Compute the predictive residual

$$\hat{a}_{(m+1)+d} = x_{(m+1)+d} - \hat{\beta}_{0,m} - \sum_{i=1}^{p} \hat{\beta}_{i,m} x_{(m+1)+d-i}$$

  and its standard error. Let $\hat{e}_{(m+1)+d}$ be the standardized predictive residual.

- Step 3: Use the recursive least squares method to update the least squares estimates to $\hat{\beta}_{i,m+1}$ by incorporating the new data point $x_{(m+1)+d}$.

- Step 4: Repeat Steps 2 and 3 until all data points are processed.

- Step 5: Consider the linear regression of the standardized predictive residual

$$\hat{e}_{(m+j)+d} = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_{(m+j)+d-i} + v_t, \quad j = 1, \ldots, T - d - m \tag{4.46}$$

  and compute the usual $F$ statistic for testing $\alpha_i = 0$ in Eq. (4.46) for $i = 0, \ldots, p$. Under the null hypothesis that $x_t$ follows a linear AR($p$) model, the $F$ ratio has a limiting $F$ distribution with degrees of freedom $p + 1$ and $T - d - m - p$.

We refer to the earlier $F$ test as a *Tar-F test*. The idea behind the test is that under the null hypothesis there is no model change in the arranged autoregression in Eq. (4.45) so that the standardized predictive residuals should be close to iid with mean zero and variance 1. In this case, they should have no correlations with the regressors $x_{(m+j)+d-i}$. For further details including formulas for a recursive least squares method and some simulation study on performance of the Tar-F test, see Tsay (1989). The Tar-F test avoids the problem of nuisance parameters encountered by the likelihood ratio test. It does not require knowing the threshold $r_1$. It simply tests that the predictive residuals have no correlations with regressors if the null hypothesis holds. Therefore, the test does not depend on knowing the number of regimes in the alternative model. Yet the Tar-F test is not as powerful as the likelihood ratio test if the true model is indeed a two-regime SETAR model with a known innovational distribution.

### 4.2.3 Applications

In this subsection, we apply some of the nonlinearity tests discussed previously to five time series. For a real financial time series, an AR model is used to remove any serial correlation in the data, and the tests apply to the residual series of the model. The five series employed are as follows:

1. $r_{1t}$: A simulated series of iid $N(0, 1)$ with 500 observations.
2. $r_{2t}$: A simulated series of iid Student-$t$ distribution with 6 degrees of freedom. The sample size is 500.

**3.** $a_{3t}$: The residual series of monthly log returns of CRSP equal-weighted index from 1926 to 1997 with 864 observations. The linear AR model used is

$$(1 - 0.180B + 0.099B^3 - 0.105B^9)r_{3t} = 0.0086 + a_{3t}.$$

**4.** $a_{4t}$: The residual series of monthly log returns of CRSP value-weighted index from 1926 to 1997 with 864 observations. The linear AR model used is

$$(1 - 0.098B + 0.111B^3 - 0.088B^5)r_{4t} = 0.0078 + a_{4t}.$$

**5.** $a_{5t}$: The residual series of monthly log returns of IBM stock from 1926 to 1997 with 864 observations. The linear AR model used is

$$(1 - 0.077B)r_{5t} = 0.011 + a_{5t}.$$

Table 4.2 shows the results of nonlinearity test. For the simulated series and IBM returns, the $F$ tests are based on an AR(6) model. For the index returns, the AR order is the same as the model given earlier. For the BDS test, we chose $\delta = \hat{\sigma}_a$ and $\delta = 1.5\hat{\sigma}_a$ with $k = 2, \ldots, 5$. Also given in the table are the Ljung–Box statistics that confirm no serial correlation in the residual series before applying nonlinearity tests. Compared with their asymptotic critical values, the BDS test and the $F$ tests are insignificant at the 5% level for the simulated series. However, the BDS tests are highly significant for the real financial time series. The $F$ tests also show significant results for the index returns, but they fail to suggest nonlinearity in the IBM log

**Table 4.2. Nonlinearity Tests for Simulated Series and Some Log Stock Returns. The Sample Size of Simulated Series is 500 and That of Stock Returns is 864. The BDS Test Uses $k = 2, \ldots, 5$.**

| | Q | Q | BDS($\delta = 1.5\hat{\sigma}_a$) | | | |
|---|---|---|---|---|---|---|
| Data | (5) | (10) | 2 | 3 | 4 | 5 |
| N(0,1) | 3.2 | 6.5 | −0.32 | −0.14 | −0.15 | −0.33 |
| $t_6$ | 0.9 | 1.7 | −0.87 | −1.18 | −1.56 | −1.71 |
| ln(ew) | 2.9 | 4.9 | 9.94 | 11.72 | 12.83 | 13.65 |
| ln(vw) | 1.0 | 9.8 | 8.61 | 9.88 | 10.70 | 11.29 |
| ln(ibm) | 0.6 | 7.1 | 4.96 | 6.09 | 6.68 | 6.82 |
| | | $d = 1$ | BDS($\delta = \hat{\sigma}_a$) | | | |
| Data | Ori-F | Tar-F | 2 | 3 | 4 | 5 |
| N(0,1) | 1.13 | 0.87 | −0.77 | −0.71 | −1.04 | −1.27 |
| $t_6$ | 0.69 | 0.81 | −0.35 | −0.76 | −1.25 | −1.49 |
| ln(ew) | 5.05 | 6.77 | 10.01 | 11.85 | 13.14 | 14.45 |
| ln(vw) | 4.95 | 6.85 | 7.01 | 7.83 | 8.64 | 9.53 |
| ln(ibm) | 1.32 | 1.51 | 3.82 | 4.70 | 5.45 | 5.72 |

returns. In summary, the tests confirm that the simulated series are linear and suggest that the stock returns are nonlinear.

## 4.3 MODELING

Nonlinear time series modeling necessarily involves subjective judgment. However, there are some general guidelines to follow. It starts with building an adequate linear model on which nonlinearity tests are based. For financial time series, the Ljung–Box statistics and Engle's test are commonly used to detect conditional heteroscedasticity. For general series, other tests of Section 4.2 apply. If nonlinearity is statistically significant, then one chooses a class of nonlinear models to entertain. The selection here may depend on the experience of the analyst and the substantive matter of the problem under study. For volatility models, the order of an ARCH process can often be determined by checking the partial autocorrelation function of the squared series. For GARCH and EGARCH models, only lower orders such as (1, 1), (1, 2), and (2, 1) are considered in most applications. Higher order models are hard to estimate and understand. For TAR models, one may use the procedures given in Tong (1990) and Tsay (1989, 1998) to build an adequate model. When the sample size is sufficiently large, one may apply nonparametric techniques to explore the nonlinear feature of the data and choose a proper nonlinear model accordingly; see Chen and Tsay (1993a) and Cai, Fan, and Yao (1999). The MARS procedure of Lewis and Stevens (1991) can also be used to explore the dynamic structure of the data. Finally, information criteria such as Akaike information criterion (Akaike, 1974) and the generalized odd ratios in Chen, McCulloch, and Tsay (1997) can be used to discriminate between competing nonlinear models. The chosen model should be carefully checked before it is used for prediction.

## 4.4 FORECASTING

Unlike the linear model, there exist no closed-form formulas to compute forecasts of most nonlinear models when the forecast horizon is greater than 1. We use parametric bootstraps to compute nonlinear forecasts. It is understood that the model used in forecasting has been rigorously checked and is judged to be adequate for the series under study. By a model, we mean the dynamic structure and innovational distributions. In some cases, we may treat the estimated parameters as given.

### 4.4.1 Parametric Bootstrap

Let $T$ be the forecast origin and $\ell$ be the forecast horizon ($\ell > 0$). That is, we are at time index $T$ and interested in forecasting $x_{T+\ell}$. The parametric bootstrap considered computes realizations $x_{T+1}, \ldots, X_{T+\ell}$ sequentially by (a) drawing a new innovation from the specified innovational distribution of the model, and (b) computing $x_{T+i}$ using the model, data, and previous forecasts $x_{T+1}, \ldots, x_{T+i-1}$. This results in a

realization for $x_{T+\ell}$. The procedure is repeated $M$ times to obtain $M$ realizations of $x_{T+\ell}$ denoted by $\{x_{T+\ell}^{(j)}\}_{j=1}^{M}$. The point forecast of $x_{T+\ell}$ is then the sample average of $x_{T+\ell}^{(j)}$. Let the forecast be $x_T(\ell)$. We used $M = 3000$ in some applications and the results seem fine. The realizations $\{x_{T+\ell}^{(j)}\}_{j=1}^{M}$ can also be used to obtain an empirical distribution of $x_{T+\ell}$. We make use of this empirical distribution later to evaluate forecasting performance.

### 4.4.2 Forecasting Evaluation

There are many ways to evaluate the forecasting performance of a model, ranging from directional measures to magnitude measures to distributional measures. A directional measure considers the future direction (up or down) implied by the model. Predicting that tomorrow's S&P 500 index will go up or down is an example of directional forecasts that are of practical interest. Predicting the year-end value of the daily S&P 500 index belongs to the case of magnitude measure. Finally, assessing the likelihood that the daily S&P 500 index will go up 10% or more between now and the year end requires knowing the future conditional probability distribution of the index. Evaluating the accuracy of such an assessment needs a distributional measure.

In practice, the available data set is divided into two subsamples. The first subsample of the data is used to build a nonlinear model, and the second subsample is used to evaluate the forecasting performance of the model. We refer to the two subsamples of data as *estimation* and *forecasting subsamples*. In some studies, a rolling forecasting procedure is used in which a new data point is moved from the forecasting subsample into the estimation subsample as the forecast origin advances. In what follows, we briefly discuss some measures of forecasting performance that are commonly used in the literature. Keep in mind, however, that there exists no widely accepted single measure to compare models. A utility function based on the objective of the forecast might be needed to better understand the comparison.

#### *4.4.2.1 Directional Measure*

A typical measure here is to use a $2 \times 2$ contingency table that summarizes the numbers of "hits" and "misses" of the model in predicting ups and downs of $x_{T+\ell}$ in the forecasting subsample. Specifically, the contingency table is given as

| | Predicted | | |
|---|---|---|---|
| Actual | up | down | |
| up | $m_{11}$ | $m_{12}$ | $m_{10}$ |
| down | $m_{21}$ | $m_{22}$ | $m_{20}$ |
| | $m_{01}$ | $m_{02}$ | $m$ |

where $m$ is the total number of $\ell$-step ahead forecasts in the forecasting subsample, $m_{11}$ is the number of "hits" in predicting upward movements, $m_{21}$ is the number of "misses" in predicting downward movements of the market, and so on. Larger values

in $m_{11}$ and $m_{22}$ indicate better forecasts. The test statistic

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(m_{ij} - \frac{m_{i0}m_{0j}}{m})^2}{\frac{m_{i0}m_{0j}}{m}}$$

can then be used to evaluate the performance of the model. A large $\chi^2$ signifies that the model outperforms the chance of random choice. Under some mild conditions, $\chi^2$ has an asymptotic chi-squared distribution with 1 degree of freedom. For further discussion of this measure, see Dahl and Hylleberg (1999).

For illustration of the directional measure, consider the 1-step ahead probability forecasts of the 8-4-1 feed-forward neural network shown in Figure 4.6. The $2 \times 2$ table of "hits" and "misses" of the network is

| | Predicted | | |
|---|---|---|---|
| Actual | up | down | |
| up | 12 | 2 | 14 |
| down | 8 | 2 | 10 |
| | 20 | 4 | 24 |

The table shows that the network predicts the upward movement well, but fares poorly in forecasting the downward movement of the stock. The chi-squared statistic of the table is 0.137 with $p$ value 0.71. Consequently, the network does not significantly outperform a random walk model with equal probabilities for "upward" and "downward" movements.

#### *4.4.2.2 Magnitude Measure*

Three statistics are commonly used to measure performance of point forecasts. They are the mean squared error (MSE), mean absolute deviation (MAD), and mean absolute percentage error (MAPE). For $\ell$-step ahead forecasts, these measures are defined as

$$MSE(\ell) = \frac{1}{m}\sum_{j=0}^{m-1}[x_{T+\ell+j} - x_{T+j}(\ell)]^2 \tag{4.47}$$

$$MAD(\ell) = \frac{1}{m}\sum_{j=0}^{m-1} | x_{T+\ell+j} - x_{T+j}(\ell) | \tag{4.48}$$

$$MAPE(\ell) = \frac{1}{m}\sum_{j=0}^{m-1} \left| \frac{x_{T+j}(\ell)}{x_{T+j+\ell}} - 1 \right|, \tag{4.49}$$

where $m$ is the number of $\ell$-step ahead forecasts available in the forecasting subsample. In application, one often chooses one of the above three measures, and the

model with the smallest magnitude on that measure is regarded as the best $\ell$-step ahead forecasting model. It is possible that different $\ell$ may result in selecting different models. The measures also have other limitations in model comparison; see, for instance, Clements and Hendry (1993).

#### *4.4.2.3 Distributional Measure*

Practitioners recently began to assess forecasting performance of a model using its predictive distributions. Strictly speaking, a predictive distribution incorporates parameter uncertainty in forecasts. We call it *conditional predictive distribution* if the parameters are treated as fixed. The empirical distribution of $x_{T+\ell}$ obtained by the parametric bootstrap is a conditional predictive distribution. This empirical distribution is often used to compute a distributional measure. Let $u_T(\ell)$ be the percentile of the observed $x_{T+\ell}$ in the prior empirical distribution. We then have a set of $m$ percentiles $\{u_{T+j}(\ell)\}_{j=0}^{m-1}$, where again $m$ is the number of $\ell$-step ahead forecasts in the forecasting subsample. If the model entertained is adequate, $\{u_{T+j}(\ell)\}$ should be a random sample from the uniform distribution on [0, 1]. For a sufficiently large $m$, one can compute the Kolmogorov–Smirnov statistic of $\{u_{T+j}(\ell)\}$ with respect to uniform [0, 1]. The statistic can be used for both model checking and forecasting comparison.

## 4.5 APPLICATION

In this section, we illustrate nonlinear time series models by analyzing the quarterly U.S. civilian unemployment rate, seasonally adjusted, from 1948 to 1993. This series was analyzed in detail by Montgomery, Zarnowitz, Tsay, and Tiao (1998). We repeat some of the analyses here using nonlinear models. Figure 4.8 shows the time plot of the data. Well-known characteristics of the series include that (a) it tends to move countercyclically with U.S. business cycles, and (b) the rate rises quickly, but decays slowly. The latter characteristic suggests that the dynamic structure of the series is nonlinear.

Denote the series by $x_t$ and let $\Delta x_t = x_t - x_{t-1}$ be the change in unemployment rate. The linear model

$$(1 - 0.31B^4)(1 - 0.65B)\Delta x_t = (1 - 0.78B^4)a_t, \quad \hat{\sigma}_a^2 = 0.090 \tag{4.50}$$

was built by Montgomery et al. (1998), where the standard errors of the three coefficients are 0.11, 0.06, and 0.07, respectively. This is a seasonal model even though the data were seasonally adjusted. It indicates that the seasonal adjustment procedure used did not successfully remove the seasonality. This model is used as a benchmark model for forecasting comparison.

To test for nonlinearity, we apply some of the nonlinearity tests of Section 4.2 with an AR(5) model for the differenced series $\Delta x_t$. The results are given in Table 4.3. All of the tests reject the linearity assumption. In fact, the linearity assumption is rejected for all AR($p$) models we applied, where $p = 2, \ldots, 10$.

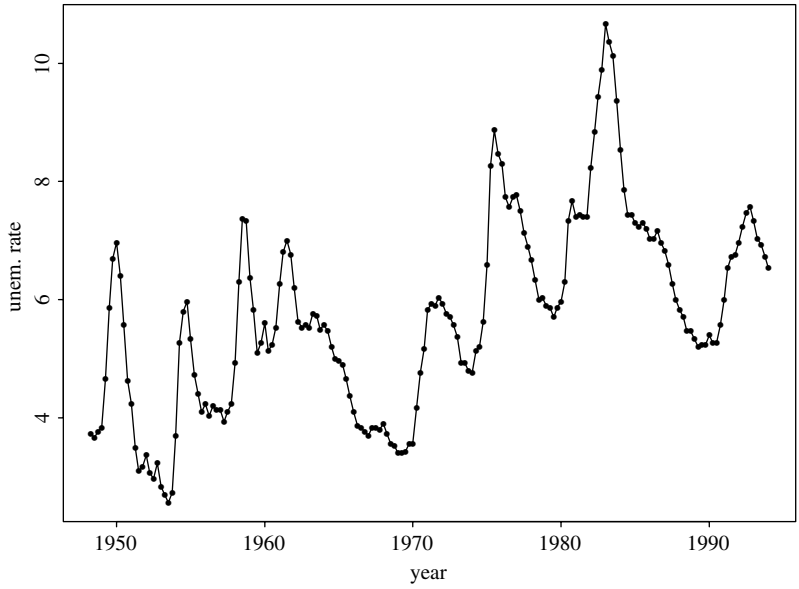


**Figure 4.8.** Time plot of the U.S. quarterly unemployment rate, seasonally adjusted, from 1948 to 1993.

Using a modeling procedure similar to that of Tsay (1989), Montgomery et al. (1998) build the following TAR model for the $\Delta x_t$ series:

$$\Delta x_t = \begin{cases} 0.01 + 0.73\Delta x_{t-1} + 0.10\Delta x_{t-2} + a_{1t} & \text{if } \Delta x_{t-2} \leq 0.1, \\ 0.18 + 0.80\Delta x_{t-1} - 0.56\Delta x_{t-2} + a_{2t} & \text{otherwise.} \end{cases} \tag{4.51}$$

The sample variances of $a_{1t}$ and $a_{2t}$ are 0.76 and 0.165, respectively, the standard errors of the three coefficients of Regime 1 are 0.03, 0.10, and 0.12, respectively, and those of Regime 2 are 0.09, 0.1, and 0.16. This model says that the change in the U.S. quarterly unemployment rate, $\Delta x_t$, behaves like a piecewise linear model in the reference space of $x_{t-2} - x_{t-3}$ with threshold 0.1. Intuitively, the model implies that the dynamics of unemployment act differently depending on the recent change in the unemployment rate. In the first regime, the unemployment rate has had either

**Table 4.3. Nonlinearity Test for Changes in the U.S. Quarterly Unemployment Rate: 1948.II-1993.IV. An AR(5) Model Was Used in the Tests, Where LST Denotes the Test of Luukkonen et al. (1988) and TAR($d$) Means Threshold Test with Delay $d$.**

| Type | Ori-F | LST | TAR(1) | TAR(2) | TAR(3) | TAR(4) |
|---|---|---|---|---|---|---|
| Test | 2.80 | 2.83 | 2.41 | 2.16 | 2.84 | 2.98 |
| $p$ value | .0007 | .0002 | .0298 | .0500 | .0121 | .0088 |

a decrease or a minor increase. Here the economy should be stable, and essentially the change in the rate follows a simple AR(1) model because the lag-2 coefficient is insignificant. In the second regime, there is a substantial jump in the unemployment rate (0.1 or larger). This typically corresponds to the contraction phase in the business cycle. It is also the period during which government interventions and industrial restructuring are likely to occur. Here $\Delta x_t$ follows an AR(2) model with a positive constant, indicating an upward trend in $x_t$. The AR(2) polynomial contains two complex characteristic roots, which indicate possible cyclical behavior in $\Delta x_t$. Consequently, the chance of having a turning point in $x_t$ increases, suggesting that the period of large increases in $x_t$ should be short. This implies that the contraction phases in the U.S. economy tend to be shorter than the expansion phases.

Applying a Markov Chain Monte Carlo method, Montgomery et al. (1998) obtain the following Markov switching model for $\Delta x_t$:

$$\Delta x_t = \begin{cases} -0.07 + 0.38\Delta x_{t-1} - 0.05\Delta x_{t-2} + \epsilon_{1t} & \text{if } s_t = 1 \\ 0.16 + 0.86\Delta x_{t-1} - 0.38\Delta x_{t-2} + \epsilon_{2t} & \text{if } s_t = 2. \end{cases} \tag{4.52}$$

The conditional means of $\Delta x_t$ are $-0.10$ for $s_t = 1$ and 0.31 for $s_t = 2$. Thus, the first state represents the expansionary periods in the economy, and the second state represents the contractions. The sample variances of $\epsilon_{1t}$ and $\epsilon_{2t}$ are 0.031 and 0.192, respectively. The standard errors of the three parameters in state $s_t = 1$ are 0.03, 0.14, and 0.11, and those of state $s_t = 2$ are 0.04, 0.13, and 0.14, respectively. The state transition probabilities are $P(s_t = 2 \mid s_{t-1} = 1) = 0.084(0.060)$ and $P(s_t = 1 \mid s_{t-1} = 2) = 0.126(0.053)$, where the number in parentheses is the corresponding standard error. This model implies that in the second state the unemployment rate $x_t$ has an upward trend with an AR(2) polynomial possessing complex characteristic roots. This feature of the model is similar to the second regime of the TAR model in Eq. (4.51). In the first state, the unemployment rate $x_t$ has a slightly decreasing trend with a much weaker autoregressive structure.

***Forecasting Performance***

A rolling procedure was used by Montgomery et al. (1998) to forecast the unemployment rate $x_t$. The procedure works as follows:

1. Begin with forecast origin $T = 83$, corresponding to 1968.II which was used in the literature to monitor performance of various econometric models in forecasting unemployment rate. Estimate the linear, TAR, and MSA models using the data from 1948.I to the forecast origin (inclusive).
2. Perform 1-quarter to 5-quarter ahead forecasts and compute the forecast errors of each model. Forecasts of nonlinear models used are computed by using the parametric bootstrap method of Section 4.4.
3. Advance the forecast origin by 1 and repeat the estimation and forecasting processes until all data are employed.
4. Use MSE and mean forecast error to compare performance of the models.

**Table 4.4. Out-of-Sample Forecast Comparison Among Linear, TAR, and MSA Models for the U.S. Quarterly Unemployment Rate. The Starting Forecast Origin is 1968.II, Where the Row Marked by "MSE" Shows the MSE of the Benchmark Linear Model.**

| (A) | Relative MSE of forecast | | | | |
|---|---|---|---|---|---|
| Model | 1-step | 2-step | 3-step | 4-step | 5-step |
| | (a) Overall comparison | | | | |
| Linear | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| TAR | 1.00 | 1.04 | 0.99 | 0.98 | 1.03 |
| MSA | 1.19 | 1.39 | 1.40 | 1.45 | 1.61 |
| MSE | 0.08 | 0.31 | 0.67 | 1.13 | 1.54 |
| | (b) Forecast origins in economic contractions | | | | |
| Linear | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| TAR | 0.85 | 0.91 | 0.83 | 0.72 | 0.72 |
| MSA | 0.97 | 1.03 | 0.96 | 0.86 | 1.02 |
| MSE | 0.22 | 0.97 | 2.14 | 3.38 | 3.46 |
| | (c) Forecast origins in economic expansions | | | | |
| Linear | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| TAR | 1.06 | 1.13 | 1.10 | 1.15 | 1.17 |
| MSA | 1.31 | 1.64 | 1.73 | 1.84 | 1.87 |
| MSE | 0.06 | 0.21 | 0.45 | 0.78 | 1.24 |
| (B) | Mean of forecast errors | | | | |
| Model | 1-step | 2-step | 3-step | 4-step | 5-step |
| | (a) Overall comparison | | | | |
| Linear | .03 | .09 | .17 | .25 | .33 |
| TAR | −.10 | −.02 | −.03 | −.03 | −.01 |
| MSA | .00 | −.02 | −.04 | −.07 | −.12 |
| | (b) Forecast origins in economic contractions | | | | |
| Linear | 0.31 | 0.68 | 1.08 | 1.41 | 1.38 |
| TAR | 0.24 | 0.56 | 0.87 | 1.01 | 0.86 |
| MSA | 0.20 | 0.41 | 0.57 | 0.52 | 0.14 |
| | (c) Forecast origins in economic expansions | | | | |
| Linear | −.01 | .00 | .03 | .08 | .17 |
| TAR | −.05 | −.11 | −.17 | −.19 | −.14 |
| MSA | −.03 | −.08 | −.13 | −.17 | −.16 |

Table 4.4 shows the relative MSE of forecasts and mean forecast errors for the linear model in Eq. (4.50), the TAR model in Eq. (4.51), and the MSA model in Eq. (4.52), using the linear model as a benchmark. The comparisons are based on overall performance as well as the status of the U.S. economy at the forecast origin. From the table, we make the following observations:

1. For the overall comparison, TAR model and the linear model are very close in MSE, but the TAR model has smaller biases. Yet the MSA model has the highest MSE, but smallest biases.
2. For forecast origins in economic contractions, the TAR model shows improvements over the linear model both in MSE and bias. The MSA model also shows some improvement over the linear model, but the improvement is not as large as that of the TAR model.
3. For forecast origins in economic expansions, the linear model outperforms both nonlinear models.

The results suggest that the contributions of nonlinear models over linear ones in forecasting the U.S. quarterly unemployment rate are mainly in the periods when the U.S. economy is in contractions. This is not surprising because, as mentioned before, it is during the economic contractions that government interventions and industrial restructuring are most likely to occur. These external events could introduce nonlinearity in the U.S. unemployment rate. Intuitively, such improvements are important because it is during the contractions that people pay more attention to economic forecasts.

## APPENDIX A. SOME RATS PROGRAMS FOR NONLINEAR VOLATILITY MODELS

### A. This program was used to estimate an AR(2)-TAR-GARCH(1, 1) model for daily log returns of IBM stock. The data file is "d-ibmln99.dat."

```
all 0 9442:1
open data d-ibmln99.dat
data(org=obs) / rt
set h = 0.0
*nonlin mu p1 p2 a0 a1 a2 b0 b1 b2
nonlin mu p2 a1 a2 b0 b1 b2
*frml at = rt(t)-mu-p1*rt(t-1)-p2*rt(t-2)
frml at = rt(t)-mu-p2*rt(t-2)
frml u = (at(t-1)/abs(at(t-1))+1.0)/2.0
frml gvar1 = a1*at(t-1)**2+a2*h(t-1)
frml gvar = gvar1(t)+u(t)*(b0+b1*at(t-1)**2+b2*h(t-1))
frml garchln = -0.5*log(h(t)=gvar(t))-0.5*at(t)**2/h(t)
smpl 4 9442
compute mu = 0.03, p1 = 0.1, p2 = -0.03
```

```
compute a0 = 0.1, a1 = 0.1, a2 = 0.6, b0 = 0.1, b1 = 0.05
compute b2 = 0.1, a3 = 0.1, b3 = 0.1
maximize(method=simplex,iterations=10) garchln
smpl 4 9442
maximize(method=bhhh,recursive,iterations=150) garchln
set fv = gvar(t)
set resid = at(t)/sqrt(fv(t))
set residsq = resid(t)*resid(t)
cor(qstats,number=20,span=10) resid
cor(qstats,number=20,span=10) residsq
```

**B. This program was used to estimate a smooth TAR model for the monthly simple returns of 3M stock. The data file is "m-mmm.dat."**

```
all 0 623:1
open data m-mmm.dat
data(org=obs) / mmm
set h = 0.0
nonlin a0 a1 a2 a00 a11 mu
frml at = mmm(t) - mu
frml var1 = a0+a1*at(t-1)**2+a2*at(t-2)**2
frml var2 = a00+a11*at(t-1)**2
frml gvar = var1(t)+var2(t)/(1.0+exp(-at(t-1)*1000.0))
frml garchlog = -0.5*log(h(t)=gvar(t))-0.5*at(t)**2/h(t)
smpl 3 623
compute a0 = .01, a1 = 0.2, a2 = 0.1
compute a00 = .01, a11 = -.2, mu = 0.02
maximize(method=bhhh,recursive,iterations=150) garchlog
set fv = gvar(t)
set resid = at(t)/sqrt(fv(t))
set residsq = resid(t)*resid(t)
cor(qstats,number=20,span=10) resid
cor(qstats,number=20,span=10) residsq
```

## APPENDIX B. S-PLUS COMMANDS FOR NEURAL NETWORK

The following commands are used in S-Plus to build the 3-2-1 skip-layer feed-forward network of Example 4.5. A line starting with "#" denotes comment. The data file is "m-ibmln.dat."

```
# load the data into S-Plus workspace.
x_scan(file='m-ibmln.dat')
# select the output: r(t)
y_x[4:864]
# obtain the input variables: r(t-1), r(t-2), and r(t-3)
ibm.x_cbind(x[3:863],x[2:862],x[1:861])
# build a 3-2-1 network with skip layer connections
# and linear output.
ibm.nn_nnet(ibm.x,y,size=2,linout=T,skip=T,maxit=10000,
decay=1e-2,reltol=1e-7,abstol=1e-7,range=1.0)
# print the summary results of the network
```

```
summary(ibm.nn)
# compute \& print the residual sum of squares.
sse_sum((y-predict(ibm.nn,ibm.x))^2)
print(sse)
#eigen(nnet.Hess(ibm.nn,ibm.x,y),T)$values
# setup the input variables in the forecasting subsample
ibm.p_cbind(x[864:887],x[863:886],x[862:885])
# compute the forecasts
yh_predict(ibm.nn,ibm.p)
# The observed returns in the forecasting subsample
yo_x[865:888]
# compute \& print the sum of squares of forecast errors
ssfe_sum((yo-yh)^2)
print(ssfe)
# quit S-Plus
q()
```

## EXERCISES

**1.** Consider the monthly log returns of General Electric (GE) stock from January 1926 to December 1999. You may download the data from CRSP or use the file "m-ge2699.dat" on the Web. The log returns in the file are in percentages. Build a threshold GARCH model for the series using $a_{t-1}$ as the threshold variable with zero threshold, where $a_{t-1}$ is the shock at time $t-1$. Check the fitted model.

**2.** Suppose that the monthly log returns of GE stock, measured in percentages, follows a smooth threshold GARCH(1, 1) model. For the sampling period from January 1926 to December 1999, the fitted model is

$$r_t = 1.06 + a_t, \quad a_t = \sigma_t \epsilon_t$$
$$\sigma_t^2 = 0.103 a_{t-1}^2 + 0.952\sigma_{t-1}^2 + \frac{1}{1+\exp(-10a_{t-1})}(4.490 - 0.193\sigma_{t-1}^2),$$

where all of the estimates are highly significant, the coefficient 10 in the exponent is fixed *a priori* to simplify the estimation, and $\{\epsilon_t\}$ are iid $N(0, 1)$. Assume that $a_{888} = 16.0$ and $\sigma_{888}^2 = 50.2$, what is the 1-step ahead volatility forecast $\widehat{\sigma}_{888}^2(1)$? Suppose instead that $a_{888} = -16.0$, what is the 1-step ahead volatility forecast $\widehat{\sigma}_{888}^2(1)$?

**3.** Suppose that the monthly log returns, in percentages, of a stock follow the following Markov switching model

$$r_t = 1.25 + a_t, \quad a_t = \sigma_t \epsilon_t$$
$$\sigma_t^2 = \begin{cases} 0.10 a_{t-1}^2 + 0.93\sigma_{t-1}^2 & \text{if } s_t = 1 \\ 4.24 + 0.10 a_{t-1}^2 + 0.78\sigma_{t-1}^2 & \text{if } s_t = 2, \end{cases}$$

where the transition probabilities are

$$P(s_t = 2 \mid s_{t-1} = 1) = 0.15, \qquad P(s_t = 1 \mid s_{t-1} = 2) = 0.05.$$

Suppose that $a_{100} = 6.0$, $\sigma^2_{100} = 50.0$ and $s_{100} = 2$ with probability 1.0; what is the 1-step ahead volatility forecast at the forecast origin $t = 100$? Also, if the probability of $s_{100} = 2$ is reduced to 0.8, what is the 1-step ahead volatility forecast at the forecast origin $t = 100$?

**4.** Again, consider the monthly log returns of GE stock from January 1926 to December 1999. Reserve the returns in 1998 and 1999 for forecasting evaluation.

- Fit a 3-2-1 feed-forward neural network to the return series and calculate the mean squared error of the 1-step ahead forecasts in the forecasting subsample. Write down the biases and weights of the network in the estimation subsample.
- Suppose that we are interested in forecasting the direction of the 1-month ahead stock movement. Fit a 6-5-1 feed-forward neural network to the return series using a Heaviside function for the output node. Compute the 1-step ahead forecasts in the forecasting subsample and compare them with the actual movements.

**5.** Because of the existence of inverted yield curves in the term structure of interest rates, the spread of interest rates should be nonlinear. To verify this, consider the weekly U.S. interest rates of (a) Treasury 1-year constant maturity rate, and (b) Treasury 3-year constant maturity rate. As in Chapter 2, denote the two interest rates by $r_{1t}$ and $r_{3t}$, respectively, and the data span is from January 5, 1962 to September 10, 1999. The data are in files "wgs3yr.dat" and "wgs1yr.dat" on the Web.

- Let $s_t = r_{3t} - r_{1t}$ be the spread in log interest rates. Is $\{s_t\}$ linear? Perform some nonlinearity tests and draw the conclusion using the 5% significance level.
- Let $s_t^* = (r_{3t} - r_{3,t-1}) - (r_{1t} - r_{1,t-1}) = s_t - s_{t-1}$ be the change in interest rate spread. Is $\{s_t^*\}$ linear? Perform some nonlinearity tests and draw the conclusion using the 5% significance level.
- Build a threshold model for the $s_t$ series and check the fitted model.
- Build a threshold model for the $s_t^*$ series and check the fitted model.

## REFERENCES

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Andrews, D. W. K., and Ploberger, W. (1994), "Optimal tests when a nuisance parameter is present only under the alternative," *Econometrica*, 62, 1383–1414.

Brock, W., Dechert, W. D., and Scheinkman, J. (1987), "A test for independence based on the correlation dimension," Working paper, Department of Economics, University of Wisconsin, Madison.

Brock, W., Hsieh, D. A., and LeBaron, B. (1991), *Nonlinear Dynamics, Chaos and Instability: Statistical Theory and Economic Evidence*, MIT Press: Cambridge.

Bryson, A. E., and Ho, Y. C. (1969), *Applied Optimal Control*. Blaisdell: New York.

Cai, Z., Fan, J., and Yao, Q. (1999), "Functional-coefficient regression models for nonlinear time series," working paper, University of North Carolina, Charlotte.

Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992), "A Monte Carlo Approach to Nonnormal and Nonlinear State Space Modeling," *Journal of the American Statistical Association*, 87, 493–500.

Chan, K. S. (1991), "Percentage points of likelihood ratio tests for threshold autoregression," *Journal of the Royal Statistical Society, Series B*, 53, 691–696.

Chan, K. S. (1993), "Consistency and limiting distribution of the least squares estimator of a continuous autoregressive model," *The Annals of Statistics*, 21, 520–533.

Chan, K. S., and Tong H. (1986), "On estimating thresholds in autoregressive models," *Journal of Time Series Analysis*, 7, 179–190.

Chan, K. S., and Tsay, R. S. (1998), "Limiting properties of the conditional least squares estimator of a continuous TAR model," *Biometrika*, 85, 413–426.

Chen, C., McCulloch, R. E., and Tsay, R. S. (1997), "A unified approach to estimating and modeling univariate linear and nonlinear time series," *Statistica Sinica*, 7, 451–472.

Chen, R., and Tsay, R. S. (1991), "On the Ergodicity of TAR(1) processes," *Annals of Applied Probability*, 1, 613–634.

Chen, R., and Tsay, R. S. (1993a), "Functional-coefficient autoregressive models," *Journal of the American Statistical Association*, 88, 298–308.

Chen, R., and Tsay, R. S. (1993b), "Nonlinear additive ARX models," *Journal of the American Statistical Association*, 88, 955–967.

Chen, R., Liu, J., and Tsay, R. S. (1995), "Additivity tests for nonlinear autoregressive models," *Biometrika* (1995), 82, 369–383.

Chen, T., and Chen, H. (1995), "Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems," *IEEE Transactions on Neural Networks* 6, 911–917.

Cheng, B., and Titterington, D. M. (1994), "Neural networks: a review from a statistical perspective," *Statistical Science* 9, 2–54.

Clements, M. P., and Hendry, D. F. (1993), "On the limitations of comparing mean square forecast errors," *Journal of Forecasting*, 12, 617–637.

Dahl, C. M., and Hylleberg, S. (1999), "Specifying nonlinear econometric models by flexible regression models and relative forecast performance," working paper, Department of Economics, University of Aarhus, Denmark.

Davis, R. B. (1987), "Hypothesis testing when a nuisance parameter is present only under the alternative," *Biometrika*, 74, 33–43.

Engle, R. F. (1982), "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflations," *Econometrica*, 50, 987–1007.

Epanechnikov, V. (1969), "Nonparametric estimates of a multivariate probability density," *Theory of Probability and its applications*, 14, 153–158.

Fan, J. (1993), "Local linear regression smoother and their minimax efficiencies," *The Annals of Statistics*, 21, 196–216.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Granger, C. W. J., and Andersen, A. P. (1978), *An Introduction to Bilinear Time Series Models*. Vandenhoek and Ruprecht: Gottingen.

Hamilton, J. D. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57, 357–384.

Hamilton, J. D. (1990), "Analysis of time series subject to changes in regime," *J. Econometrics*, 45, 39–70.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton University Press: New Jersey.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press: New York.

Hansen, B. E. (1997), "Inference in TAR models," *Studies in Nonlinear Dynamics and Econometrics*, 1, 119–131.

Hinich, M. (1982), "Testing for Gaussianity and linearity of a stationary time series," *Journal of Time Series Analysis*, 3, 169–176,

Hornik, K. (1993), "Some new results on neural network approximation," *Neural Networks* 6, 1069–1072.

Hornik, K., Stinchcombe, M., and White, H. (1989), "Multilayer feedforward networks are universal approximators," *Neural Networks* 2, 359–366.

Hsieh, D. A. (1989), "Testing for nonlinear dependence in daily foreign exchange rates," *Journal of Business*, 62, 339–368.

Kennan, D. M. (1985), "A Tukey non-additivity-type test for time series nonlinearity," *Biometrika*, 72, 39–44.

Kitagawa, G. (1998), "A self-organizing state space model," *Journal of the American Statistical Association*, 93, 1203–1215.

Lewis, P. A. W., and Stevens, J. G. (1991), "Nonlinear Modeling of Time Series Using Multivariate Adaptive Regression Spline (MARS)," *Journal of the American Statistical Association*, 86, 864–877.

Liu, J., and Brockwell,P. J. (1988), "On the general bilinear time-series model," *Journal of Applied Probability*, 25, 553–564.

Luukkonen, R., Saikkonen, P. and Teräsvirta, T. (1988), "Testing linearity against smooth transition autoregressive models," *Biometrika*, 75, 491–499.

McCulloch, R. E., and Tsay, R. S. (1993), "Bayesian Inference and Prediction for Mean and Variance Shifts in Autoregressive Time Series," *Journal of the American Statistical Association*, 88, 968–978.

McCulloch, R. E., and Tsay, R. S. (1994), "Statistical inference of macroeconomic time series via Markov switching models,"
*Journal of Time Series Analysis*, 15, 523–539.

McLeod, A. I., and Li, W. K. (1983), "Diagnostic checking ARMA time series models using squared-residual autocorrelations," *Journal of Time Series Analysis*, 4, 269–273.

Montgomery, A. L., Zarnowitz, V., Tsay, R. S., and Tiao, G. C. (1998), "Forecasting the U.S. Unemployment Rate," *Journal of the American Statistical Association*, 93, 478–493.

Nadaraya, E. A. (1964), "On estimating regression," *Theory and Probability Application*, 10, 186–190.

Petruccelli, J., and Woolford, S. W. (1984), "A threshold AR(1) model," *Journal of Applied Probability*, 21, 270–286.

Potter, S. M. (1995), "A nonlinear approach to U.S. GNP," *Journal of Applied Econometrics*, 10, 109–125.

Priestley, M. B. (1980), "State-dependent models: a general approach to nonlinear time series analysis," *Journal of Time Series Analysis*, 1, 47–71.

Priestley, M. B. (1988), *Non-linear and Non-stationary Time Series Analysis*, Academic Press: London.

Rabemananjara, R., and Zakoian, J. M. (1993), "Threshold ARCH models and asymmetries in volatility," *Journal of Applied Econometrics*, 8, 31–49.

Ramsey, J. B. (1969), "Tests for specification errors in classical linear least squares regression analysis," *Journal of the Royal Statistical Society, Series B*, 31, 350–371.

Ripley, B. D. (1993), "Statistical aspects of neural networks," in *Networks and Chaos—Statistical and Probablistic Aspects*, eds O. E. Barndorff-Nielsen, J. L. Jensen, and W. S. Kendall, pp. 40–123. Chapman and Hall: London.

Subba Rao, T., and Gabr, M. M. (1984), *An Introduction to Bispectral Analysis and Bilinear Time Series Models*, Lecture Notes in Statistics, 24. Springer-Verlag: New York.

Teräsvirta, T. (1994), "Specification, estimation, and evaluation of smooth transition autoregressive models," *Journal of the American Statistical Association*, 89, 208–218.

Tiao, G. C., and Tsay, R. S. (1994), "Some Advances in Nonlinear and Adaptive Modeling in Time Series," *Journal of Forecasting*, 13, 109–131.

Tong, H. (1978), "On a threshold model," in *Pattern Recognition and Signal Processing*, ed. C.H. Chen, Sijhoff & Noordhoff: Amsterdam.

Tong, H. (1983), *Threshold Models in Nonlinear Time Series Analysis*, Lecture Notes in Statistics, Springer-Verlag: New York.

Tong, H. (1990), *Non-Linear Time Series: A Dynamical System Approach*, Oxford University Press: Oxford.

Tsay, R. S. (1986), "Nonlinearity tests for time series," *Biometrika*, 73, 461–466.

Tsay, R. S. (1989),"Testing and modeling threshold autoregressive processes," *Journal of the American Statistical Association*, 84, 231–240.

Tsay, R. S. (1998), "Testing and modeling multivariate threshold models," *Journal of the American Statistical Association*, 93, 1188–1202.

Venables, W. N., and Ripley, B. D. (1999), *Modern Applied Statistics with S-Plus*, 3rd ed. Springer-Verlag: New York.

Watson, G. S. (1964), "Smooth regression analysis," *Sabkhya, Series A*, 26, 359–372.

Zakoian, J. M. (1994), "Threshold heteroscedastic models," *Journal of Economic Dynamics and Control*, 18, 931–955.

*Analysis of Financial Time Series.* Ruey S. Tsay

ISBN: 0-471-41544-8

CHAPTER 5

# High-Frequency Data Analysis and Market Microstructure

High-frequency data are observations taken at fine time intervals. In finance, they often mean observations taken daily or at a finer time scale. These data have become available primarily due to advances in data acquisition and processing techniques, and they have attracted much attention because they are important in empirical study of market microstructure. The ultimate high-frequency data in finance are the transaction-by-transaction or trade-by-trade data in security markets. Here time is often measured in seconds. The Trades and Quotes (TAQ) database of the New York Stock Exchange (NYSE) contains all equity transactions reported on the *Consolidated Tape* from 1992 to present, which includes transactions on NYSE, AMEX, NASDAQ, and the regional exchanges. The Berkeley Options Data Base provides similar data for options transactions from August 1976 to December 1996. Transactions data for many other securities and markets, both domestic and foreign, are continuously collected and processed. Wood (2000) provides some historical perspective of high-frequency financial study.

High-frequency financial data are important in studying a variety of issues related to trading process and market microstructure. They can be used to compare the efficiency of different trading systems in price discovery (e.g., the open out-cry system of NYSE and the computer trading system of NASDAQ). They can also be used to study the dynamics of bid and ask quotes of a particular stock (e.g., Hasbrouck, 1999; Zhang, Russell, and Tsay, 2001b). In an order-driven stock market (e.g., the Taiwan Stock Exchange), high-frequency data can be used to study the order dynamic and, more interesting, to investigate the question "who provides the market liquidity." Cho, Russell, Tiao, and Tsay (2000) use intraday 5-minute returns of more than 340 stocks traded in the Taiwan Stock Exchange to study the impact of daily stock price limits and find significant evidence of magnet effects toward the price ceiling.

However, high-frequency data have some unique characteristics that do not appear in lower frequencies. Analysis of these data thus introduces new challenges to financial economists and statisticians. In this chapter, we study these special characteristics, consider methods for analyzing high-frequency data, and discuss implications

of the results obtained. In particular, we discuss nonsynchronous trading, bid-ask spread, duration models, price movements that are in multiples of tick size, and bivariate models for price changes and time durations between transactions associated with price changes. The models discussed are also applicable to other scientific areas such as telecommunications and environmental studies.

## 5.1 NONSYNCHRONOUS TRADING

We begin with nonsynchronous trading. Stock tradings such as those on the NYSE do not occur in a synchronous manner; different stocks have different trading frequencies, and even for a single stock the trading intensity varies from hour to hour and from day to day. Yet we often analyze a return series in a fixed time interval such as daily, weekly, or monthly. For daily series, price of a stock is its *closing* price, which is the last transaction price of the stock in a trading day. The actual time of the last transaction of the stock varies from day to day. As such we incorrectly assume daily returns as an equally-spaced time series with a 24-hour interval. It turns out that such an assumption can lead to erroneous conclusions about the predictability of stock returns even if the true return series are serially independent.

For daily stock returns, nonsynchronous trading can introduce (a) lag-1 cross-correlation between stock returns, (b) lag-1 serial correlation in a portfolio return, and (c) in some situations negative serial correlations of the return series of a single stock. Consider stocks A and B. Assume that the two stocks are independent and stock A is traded more frequently than stock B. For special news affecting the market that arrives near the closing hour on one day, stock A is more likely than B to show the effect of the news on the same day simply because A is traded more frequently. The effect of the news on B will eventually appear, but it may be delayed until the following trading day. If this situation indeed happens, return of stock A appears to lead that of stock B. Consequently, the return series may show a significant lag-1 cross-correlation from A to B even though the two stocks are independent. For a portfolio that holds stocks A and B, the prior cross-correlation would become a significant lag-1 serial correlation.

In a more complicated manner, nonsynchronous trading can also induce erroneous negative serial correlations for a single stock. There are several models available in the literature to study this phenomenon; see Campbell, Lo, and MacKinlay (1997) and the references therein. Here we adopt a simplified version of the model proposed in Lo and MacKinlay (1990). Let $r_t$ be the continuously compounded return of a security at the time index $t$. For simplicity, assume that $\{r_t\}$ is a sequence of independent and identically distributed random variables with mean $E(r_t) = \mu$ and variance $\text{Var}(r_t) = \sigma^2$. For each time period, the probability that the security is not traded is $\pi$, which is time-invariant and independent of $r_t$. Let $r_t^o$ be the observed return. When there is no trade at time index $t$, we have $r_t^o = 0$ because there is no information available. Yet when there is a trade at time index $t$, we define $r_t^o$ as the cumulative return from the previous trade (i.e., $r_t^o = r_t + r_{t-1} + \cdots + r_{t-k_t}$, where $k_t$ is the largest non-negative integer such that no trade occurred in the periods

$t - k_t, t - k_t + 1, \ldots, t - 1$). Mathematically, the relationship between $r_t$ and $r_t^o$ is

$$r_t^o = \begin{cases} 0 & \text{with probability } \pi \\ r_t & \text{with probability } (1-\pi)^2 \\ r_t + r_{t-1} & \text{with probability } (1-\pi)^2\pi \\ r_t + r_{t-1} + r_{t-2} & \text{with probability } (1-\pi)^2\pi^2 \\ \vdots & \vdots \\ \sum_{i=0}^{k} r_{t-i} & \text{with probability } (1-\pi)^2\pi^{k-1} \\ \vdots & \vdots \end{cases} \tag{5.1}$$

These probabilities are easy to understand. For example, $r_t^o = r_t$ if and only if there are trades at both $t$ and $t-1$, $r_t^o = r_t + r_{t-1}$ if and only if there are trades at $t$ and $t-2$, but no trade at $t-1$, and $r_t^o = r_t + r_{t-1} + r_{t-2}$ if and only if there are trades at $t$ and $t-3$, but no trades at $t-1$ and $t-2$, and so on. As expected, the total probability is 1 given by

$$\pi + (1-\pi)^2[1 + \pi + \pi^2 + \cdots] = \pi + (1-\pi)^2 \frac{1}{1-\pi} = \pi + 1 - \pi = 1.$$

We are ready to consider the moment equations of the observed return series $\{r_t^o\}$. First, the expectation of $r_t^o$ is

$$\begin{aligned} E(r_t^o) &= (1-\pi)^2 E(r_t) + (1-\pi)^2\pi E(r_t + r_{t-1}) + \cdots \\ &= (1-\pi)^2\mu + (1-\pi)^2\pi 2\mu + (1-\pi)^2\pi^2 3\mu + \cdots \\ &= (1-\pi)^2\mu[1 + 2\pi + 3\pi^2 + 4\pi^3 + \cdots] \\ &= (1-\pi)^2\mu \frac{1}{(1-\pi)^2} = \mu. \end{aligned} \tag{5.2}$$

In the prior derivation, we use the result $1 + 2\pi + 3\pi^2 + 4\pi^3 + \cdots = \frac{1}{(1-\pi)^2}$. Next, for the variance of $r_t^o$, we use $\text{Var}(r_t^o) = E[(r_t^o)^2] - [E(r_t^o)]^2$ and

$$\begin{aligned} E(r_t^o)^2 &= (1-\pi)^2 E[(r_t)^2] + (1-\pi)^2\pi E[(r_t + r_{t-1})^2] + \cdots \\ &= (1-\pi)^2[(\sigma^2 + \mu^2) + \pi(2\sigma^2 + 4\mu^2) + \pi^2(3\sigma^2 + 9\mu^2) + \cdot] \qquad (5.3) \\ &= (1-\pi)^2\{\sigma^2[1 + 2\pi + 3\pi^2 + \cdots] + \mu^2[1 + 4\pi + 9\pi^2 + \cdots]\} \qquad (5.4) \\ &= \sigma^2 + \mu^2\left[\frac{2}{1-\pi} - 1\right]. \qquad (5.5) \end{aligned}$$

In Eq. (5.3), we use

$$E\left(\sum_{i=0}^{k} r_{t-i}\right)^2 = \text{Var}\left(\sum_{i=0}^{k} r_{t-i}\right) + \left[E\left(\sum_{i=0}^{k} r_{t-i}\right)\right]^2 = (k+1)\sigma^2 + [(k+1)\mu]^2$$

under the serial independence assumption of $r_t$. Using techniques similar to that of Eq. (5.2), we can show that the first term of Eq. (5.4) reduces to $\sigma^2$. For the second term of Eq. (5.4), we use the identity

$$1 + 4\pi + 9\pi^2 + 16\pi^3 + \cdots = \frac{2}{(1-\pi)^3} - \frac{1}{(1-\pi)^2},$$

which can be obtained as follows: Let

$$H = 1 + 4\pi + 9\pi^2 + 16\pi^3 + \cdots \quad \text{and} \quad G = 1 + 3\pi + 5\pi^2 + 7\pi^3 + \cdots.$$

Then $(1-\pi)H = G$ and

$$\begin{aligned}(1-\pi)G &= 1 + 2\pi + 2\pi^2 + 2\pi^3 + \cdots \\ &= 2(1 + \pi + \pi^2 + \cdots) - 1 = \frac{2}{(1-\pi)} - 1.\end{aligned}$$

Consequently, from Eqs. (5.2) and (5.5), we have

$$\text{Var}(r_t^o) = \sigma^2 + \mu^2\left[\frac{2}{1-\pi} - 1\right] - \mu^2 = \sigma^2 + \frac{2\pi\mu^2}{1-\pi}. \tag{5.6}$$

Consider next the lag-1 autocovariance of $\{r_t^o\}$. Here we use $\text{Cov}(r_t^o, r_{t-1}^o) = E(r_t^o r_{t-1}^o) - E(r_t^0)E(r_{t-1}^o) = E(r_t^o r_{t-1}^o) - \mu^2$. The question then reduces to finding $E(r_t^o r_{t-1}^o)$. Notice that $r_t^o r_{t-1}^o$ is zero if there is no trade at $t$, no trade at $t-1$, or no trade at both $t$ and $t-1$. Therefore, we have

$$r_t^o r_{t-1}^o = \begin{cases} 0 & \text{with probability } 2\pi - \pi^2 \\ r_t r_{t-1} & \text{with probability } (1-\pi)^3 \\ r_t(r_{t-1} + r_{t-2}) & \text{with probability } (1-\pi)^3\pi \\ r_t(r_{t-1} + r_{t-2} + r_{t-3}) & \text{with probability } (1-\pi)^3\pi^2 \\ \vdots & \vdots \\ r_t(\sum_{i=1}^{k} r_{t-i}) & \text{with probability } (1-\pi)^3\pi^{k-1} \\ \vdots & \vdots \end{cases} \tag{5.7}$$

Again the total probability is unity. To understand the prior result, notice that $r_t^o r_{t-1}^o = r_t r_{t-1}$ if and only if there are three consecutive trades at $t-2$, $t-1$, and $t$. Using Eq. (5.7) and the fact that $E(r_t r_{t-j}) = E(r_t)E(r_{t-j}) = \mu^2$ for $j > 0$, we have

$$\begin{aligned}E(r_t^o r_{t-1}^o) &= (1-\pi)^3\{E(r_t r_{t-1}) + \pi E[r_t(r_{t-1} + r_{t-2})] \\ &\quad + \pi^2 E\left[r_t\left(\sum_{i=1}^{3} r_{t-i}\right)\right] + \cdots\} \\ &= (1-\pi)^3\mu^2[1 + 2\pi + 3\pi^2 + \cdots] = (1-\pi)\mu^2.\end{aligned}$$

The lag-1 autocovariance of $\{r_t^o\}$ is then

$$\text{Cov}(r_t^o, r_{t-1}^o) = -\pi\mu^2. \tag{5.8}$$

Provided that $\mu$ is not zero, the nonsynchronous trading induces a *negative* lag-1 autocorrelation in $r_t^o$ given by

$$\rho_1(r_t^o) = \frac{-(1-\pi)\pi\mu^2}{(1-\pi)\sigma^2 + 2\pi\mu^2}.$$

In general, we can extend the prior result and show that

$$\text{Cov}(r_t^o, r_{t-j}^o) = -\mu^2\pi^j, \quad j \geq 1.$$

The magnitude of the lag-1 ACF depends on the choices of $\mu$, $\pi$, and $\sigma$ and can be substantial. Thus, when $\mu \neq 0$, the nonsynchronous trading induces negative autocorrelations in an observed security return series.

The previous discussion can be generalized to the return series of a portfolio that consists of $N$ securities; see Campbell, Lo, and MacKinlay (1997, Chapter 3). In the time series literature, effects of nonsynchronous trading on the return of a single security are equivalent to that of random temporal aggregation on a time series, with the trading probability $\pi$ governing the mechanism of aggregation.

## 5.2 BID-ASK SPREAD

In some stock exchanges (e.g., NYSE) market makers play an important role in facilitating trades. They provide market liquidity by standing ready to buy or sell whenever the public wishes to sell or buy. By market liquidity, we mean the ability to buy or sell significant quantities of a security quickly, anonymously, and with little price impact. In return for providing liquidity, market makers are granted monopoly rights by the exchange to post different prices for purchases and sales of a security. They buy at the *bid* price $P_b$ and sell at a higher ask price $P_a$. (For the public, $P_b$ is the sale price and $P_a$ is the purchase price.) The difference $P_a - P_b$ is call the *bid-ask spread*, which is the primary source of compensation for market makers. Typically, the bid-ask spread is small—namely, one or two ticks.

The existence of bid-ask spread, although small in magnitude, has several important consequences in time series properties of asset returns. We briefly discuss the bid-ask bounce—namely, the bid-ask spread introduces *negative* lag-1 serial correlation in an asset return. Consider the simple model of Roll (1984). The observed market price $P_t$ of an asset is assumed to satisfy

$$P_t = P_t^* + I_t\frac{S}{2}, \tag{5.9}$$

where $S = P_a - P_b$ is the bid-ask spread, $P_t^*$ is the time-$t$ fundamental value of the asset in a frictionless market, and $\{I_t\}$ is a sequence of independent binary random variables with equal probabilities (i.e., $I_t = 1$ with probability 0.5 and $= -1$ with probability 0.5). The $I_t$ can be interpreted as an order-type indicator, with 1 signifying buyer-initiated transaction and $-1$ seller-initiated transaction. Alternatively, the model can be written as

$$P_t = P_t^* + \begin{cases} +S/2 & \text{with probability 0.5,} \\ -S/2 & \text{with probability 0.5.} \end{cases}$$

If there is no change in $P_t^*$, then the observed process of price changes is

$$\Delta P_t = (I_t - I_{t-1})\frac{S}{2}. \tag{5.10}$$

Under the assumption of $I_t$ in Eq. (5.9), $E(I_t) = 0$ and $\text{Var}(I_t) = 1$, and we have $E(\Delta P_t) = 0$ and

$$\text{Var}(\Delta P_t) = S^2/2 \tag{5.11}$$

$$\text{Cov}(\Delta P_t, \Delta P_{t-1}) = -S^2/4 \tag{5.12}$$

$$\text{Cov}(\Delta P_t, \Delta P_{t-j}) = 0, \quad j > 1. \tag{5.13}$$

Therefore, the autocorrelation function of $\Delta P_t$ is

$$\rho_j(\Delta P_t) = \begin{cases} -0.5 & \text{if } j = 1, \\ 0 & \text{if } j > 1. \end{cases} \tag{5.14}$$

The bid-ask spread thus introduces a negative lag-1 serial correlation in the series of observed price changes. This is referred to as the *bid-ask bounce* in the finance literature. Intuitively, the bounce can be seen as follows. Assume that the fundamental price $P_t^*$ is equal to $(P_a + P_b)/2$. Then $P_t$ assumes the value $P_a$ or $P_b$. If the previously observed price is $P_a$ (the higher value), then the current observed price is either unchanged or lower at $P_b$. Thus, $\Delta P_t$ is either 0 or $-S$. However, if the previous observed price is $P_b$ (the lower value), then $\Delta P_t$ is either 0 or $S$. The negative lag-1 correlation in $\Delta P_t$ becomes apparent. The bid-ask spread does not introduce any serial correlation beyond lag 1, however.

A more realistic formulation is to assume that $P_t^*$ follows a random walk so that $\Delta P_t^* = P_t^* - P_{t-1}^* = \epsilon_t$, which forms a sequence of independent and identically distributed random variables with mean zero and variance $\sigma^2$. In addition, $\{\epsilon_t\}$ is independent of $\{I_t\}$. In this case, $\text{Var}(\Delta P_t) = \sigma^2 + S^2/2$, but $\text{Cov}(\Delta P_t, \Delta P_{t-j})$ remains unchanged. Therefore,

$$\rho_1(\Delta P_t) = \frac{-S^2/4}{S^2/2 + \sigma^2} \leq 0.$$

The magnitude of lag-1 autocorrelation of $\Delta P_t$ is reduced, but the negative effect remains when $S = P_a - P_b > 0$. In finance, it might be of interest to study the components of the bid-ask spread. Interested readers are referred to Campbell, Lo, and MacKinlay (1997) and the references therein.

The effect of bid-ask spread continues to exist in portfolio returns and in multivariate financial time series. Consider the bivariate case. Denote the bivariate order-type indicator by $I_t = (I_{1t}, I_{2t})'$, where $I_{1t}$ is for the first security and $I_{2t}$ for the second security. If $I_{1t}$ and $I_{2t}$ are contemporaneously correlated, then the bid-ask spreads can introduce negative lag-1 cross-correlations.

## 5.3 EMPIRICAL CHARACTERISTICS OF TRANSACTIONS DATA

Let $t_i$ be the calendar time, measured in seconds from midnight, at which the $i$-th transaction of an asset takes place. Associated with the transaction are several variables such as the transaction price, the transaction volume, the prevailing bid and ask quotes, and so on. The collection of $t_i$ and the associated measurements are referred to as the *transactions data*. These data have several important characteristics that do not exist when the observations are aggregated over time. Some of the characteristics are given next.

1. Unequally spaced time intervals: Transactions such as stock tradings on an exchange do not occur at equally spaced time intervals. As such the observed transaction prices of an asset do not form an equally spaced time series. The time duration between trades becomes important and might contain useful information about market microstructure (e.g., trading intensity).
2. Discrete-valued prices: The price change of an asset from one transaction to the next only occurs in multiples of tick size. In the NYSE, the tick size was one eighth of a dollar before June 24, 1997, and was one sixteenth of a dollar before January 29, 2001. All NYSE and AMEX stocks started to trade in decimals on January 29, 2001. Therefore, the price is a discrete-valued variable in transactions data. In some markets, price change may also be subject to limit constraints set by regulators.
3. Existence of a daily periodic or diurnal pattern: Under the normal trading conditions, transaction activity can exhibit periodic pattern. For instance, in the NYSE, transactions are heavier at the beginning and closing of the trading hours and thinner during the lunch hours, resulting in a "U-shape" transaction intensity. Consequently, time durations between transactions also exhibit a daily cyclical pattern.
4. Multiple transactions within a single second: It is possible that multiple transactions, even with different prices, occur at the same time. This is partly due to the fact that time is measured in seconds that may be too long a time scale in periods of heavy tradings.

**Table 5.1. Frequencies of Price Change in Multiples of Tick Size for IBM Stock from November 1, 1990 to January 31, 1991.**

| Number (tick) | $\leq -3$ | $-2$ | $-1$ | 0 | 1 | 2 | $\geq 3$ |
|---|---|---|---|---|---|---|---|
| Percentage | 0.66 | 1.33 | 14.53 | 67.06 | 14.53 | 1.27 | 0.63 |

To demonstrate these characteristics, we consider first the IBM transactions data from November 1, 1990 to January 31, 1991. These data are from the Trades, Orders Reports, and Quotes (TORQ) dataset; see Hasbrouck (1992). There are 63 trading days and 60,328 transactions. To simplify the discussion, we ignore the price changes between trading days and focus on the transactions that occurred in the normal trading hours from 9:30 am to 4:00 pm Eastern Time. It is well known that overnight stock returns differ substantially from intraday returns; see Stoll and Whaley (1990) and the references therein. Table 5.1 gives the frequencies in percentages of price change measured in the tick size of $\$1/8 = \$0.125$. From the table, we make the following observations:

1. About two-thirds of the intraday transactions were without price change.
2. The price changed in one tick approximately 29% of the intraday transactions.
3. Only 2.6% of the transactions were associated with two-tick price changes.
4. Only about 1.3% of the transactions resulted in price changes of three ticks or more.
5. The distribution of positive and negative price changes was approximately symmetric.

Consider next the number of transactions in a 5-minute time interval. Denote the series by $x_t$. That is, $x_1$ is the number of IBM transactions from 9:30 am to 9:35 am on November 1, 1990 Eastern time, $x_2$ is the number of transactions from 9:35 am to 9:40 am, and so on. The time gaps between trading days are ignored. Figure 5.1(a) shows the time plot of $x_t$, and Figure 5.1(b) the sample ACF of $x_t$ for lags 1 to 260. Of particular interest is the cyclical pattern of the ACF with a periodicity of 78, which is the number of 5-minute intervals in a trading day. The number of transactions thus exhibits a daily pattern. To further illustrate the daily trading pattern, Figure 5.2 shows the average number of transactions within 5-minute time intervals over the 63 days. There are 78 such averages. The plot exhibits a "smiling" or "U" shape, indicating heavier tradings at the opening and closing of the market and thinner tradings during the lunch hours.

Since we focus on transactions that occurred in the normal trading hours of a trading day, there are 59,838 time intervals in the data. These intervals are called the intraday *durations* between trades. For IBM stock, there were 6531 zero time intervals. That is, during the normal trading hours of the 63 trading days from November 1, 1990 to January 31, 1991, multiple transactions in a second occurred 6531 times, which is about 10.91%. Among these multiple transactions, 1002 of them had

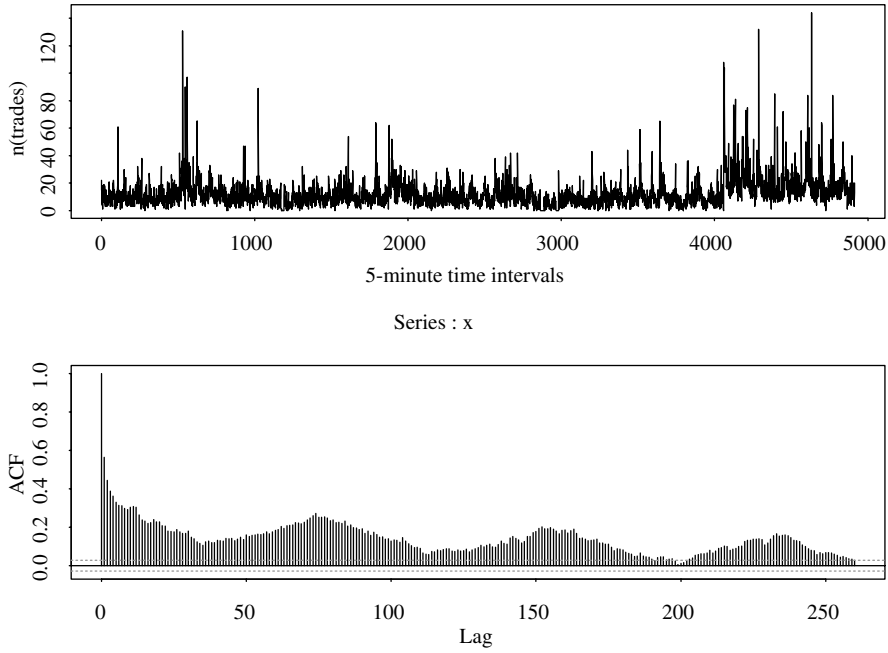


**Figure 5.1.** IBM intraday transactions data from 11/01/90 to 1/31/91: (a) the number of transactions in 5-minute time intervals, and (b) the sample ACF of the series in part(a).

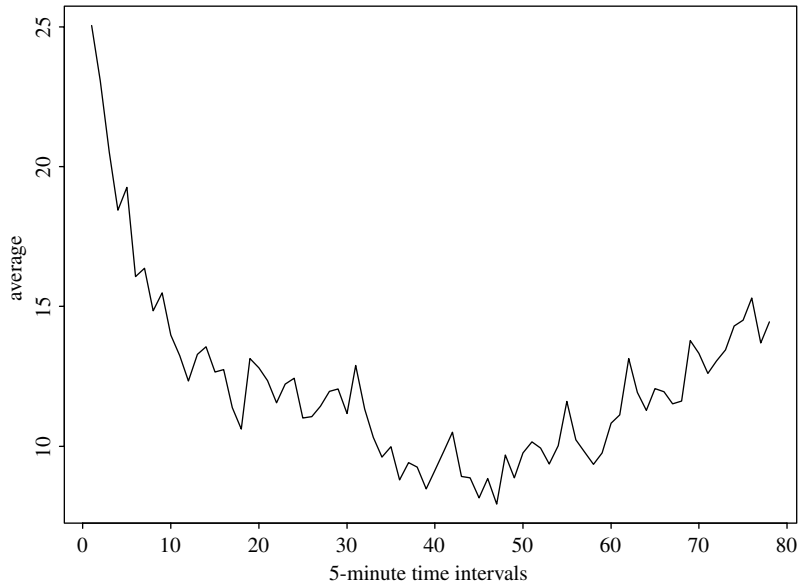


**Figure 5.2.** Time plot of the average number of transactions in 5-minute time intervals. There are 78 observations, averaging over the 63 trading days from 11/01/90 to 1/31/91 for IBM stock.

**Table 5.2. Two-Way Classification of Price Movements in Consecutive Intraday Trades for IBM Stock. The Price Movements Are Classified Into "Up," "Unchanged," and "Down." The Data Span is From 11/01/90 to 1/31/91.**

| | $i$th trade | | | |
|---|---|---|---|---|
| $(i-1)$th trade | "+" | "0" | "−" | Margin |
| "+" | 441 | 5498 | 3948 | 9887 |
| "0" | 4867 | 29779 | 5473 | 40119 |
| "−" | 4580 | 4841 | 410 | 9831 |
| Margin | 9888 | 40118 | 9831 | 59837 |

different prices, which is about 1.67% of the total number of intraday transactions. Therefore, multiple transactions (i.e., zero durations) may become an issue in statistical modeling of the time durations between trades.

Table 5.2 provides a two-way classification of price movements. Here price movements are classified into "up," "unchanged," and "down." We denote them by "+," "0," and "−," respectively. The table shows the price movements between two consecutive trades (i.e., from the $[i-1]$th to the $i$th transaction) in the sample. From the table, trade-by-trade data show that

1. consecutive price increases or decreases are relatively rare, which are about $441/59837 = 0.74\%$ and $410/59837 = 0.69\%$, respectively;
2. there is a slight edge to move from "up" to "unchanged" than to "down"; see row 1 of the table;
3. there is a high tendency for price to remain "unchanged";
4. the probabilities of moving from "down" to "up" or "unchanged" are about the same. See row 3.

The first observation mentioned before is a clear demonstration of bid-ask bounce, showing *price reversals* in intraday transactions data. To confirm this phenomenon, we consider a directional series $D_i$ for price movements, where $D_i$ assumes the value $+1$, 0, $-1$ for "up," "unchanged," and "down" price movement, respectively, for the $i$th transaction. The ACF of $\{D_i\}$ has a single spike at lag 1 with value $-0.389$, which is highly significant for a sample size of 59,837 and confirms the price reversal in consecutive trades.

As a second illustration, we consider the transactions data of IBM stock in December 1999 obtained from the TAQ database. The normal trading hours are from 9:30 am to 4:00 pm Eastern time, except for December 31 when the market closed at 13:00 pm. Comparing with the 1990–1991 data, two important changes have occurred. First, the number of intraday tradings has increased sixfold. There were 134,120 intraday tradings in December 1999 alone. The increased trading intensity also increased the chance of multiple transactions within a second. The percentage of trades with zero time duration doubled to 22.98%. At the extreme, there were

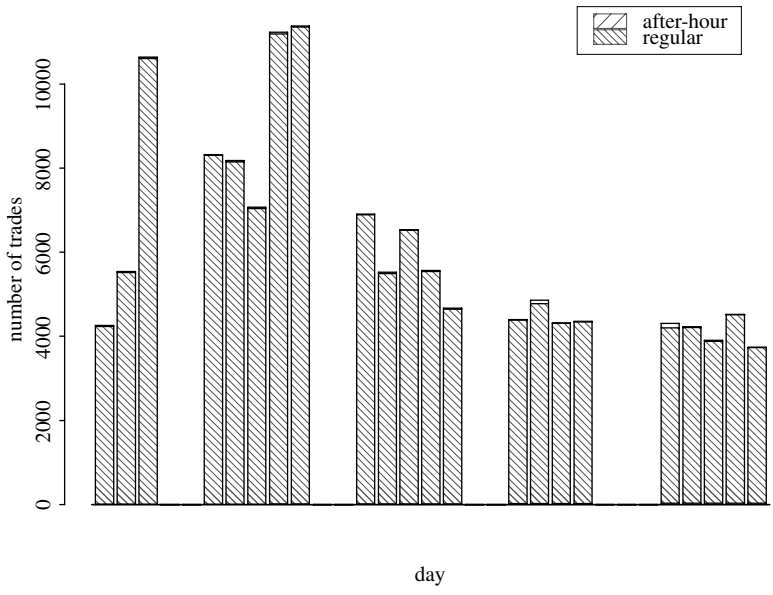


**Figure 5.3.** IBM transactions data for December 1999. The plot shows the number of transactions in each trading day with the after-hours portion denoting the number of trades with time stamp after 4:00 pm.

42 transactions within a given second that happened twice on December 3, 1999. Second, the tick size of price movement was $\$1/16 = \$0.0625$ instead of $\$1/8$. The change in tick size should reduce the bid-ask spread. Figure 5.3 shows the daily number of transactions in the new sample. Figure 5.4(a) shows the time plot of time durations between trades, measured in seconds, and Figure 5.4(b) is the time plot of price changes in consecutive intraday trades, measured in multiples of the tick size of $\$1/16$. As expected, Figures 5.3 and 5.4(a) show clearly the inverse relationship between the daily number of transactions and the time interval between trades. Figure 5.4(b) shows two unusual price movements for IBM stock on December 3, 1999. They were a drop of 63 ticks followed by an immediate jump of 64 ticks and a drop of 68 ticks followed immediately by a jump of 68 ticks. Unusual price movements like these occurred infrequently in intraday transactions.

Focusing on trades recorded within the regular trading hours, we have 61,149 trades out of 133,475 with no price change. This is about 45.8% and substantially lower than that between November 1990 and January 1991. It seems that reducing the tick size increased the chance of a price change. Table 5.3 gives the percentages of trades associated with a price change. The price movements remain approximately symmetric with respect to zero. Large price movements in intraday tradings are still relatively rare.

***Remark:*** The record keeping of high-frequency data is often not as good as that of observations taken at lower frequencies. Data cleaning becomes a necessity in

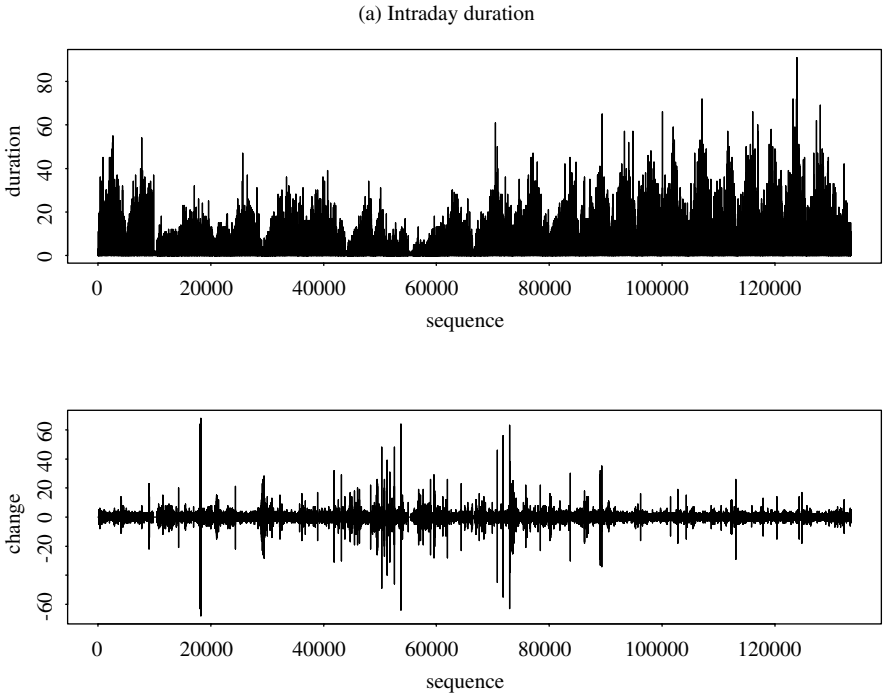


**Figure 5.4.** IBM transactions data for December 1999. Part (a) is the time plot of time durations between trades and part (b) is the time plot of price changes in consecutive trades measured in multiples of the tick size of \$1/16. Only data in the normal trading hours are included.

high-frequency data analysis. For transactions data, missing observations may happen in many ways, and the accuracy of the exact transaction time might be questionable for some trades. For example, recorded trading times may be beyond 4:00 pm Eastern time even before the opening of after-hours tradings. How to handle these observations deserves a careful study. A proper method of data cleaning requires a

**Table 5.3. Percentages of Intraday Transactions Associated with a Price Change for IBM Stock Traded in December 1999. The Percentage of Transactions without Price Change Is 45.8% and the Total Number of Transactions Recorded within the Regular Trading Hours Is 133,475. The Size Is Measured in Multiples of Tick Size \$1/16.**

| | (a) Upward movements | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | $> 7$ |
| percentage | 18.03 | 5.80 | 1.79 | 0.66 | 0.25 | 0.15 | 0.09 | 0.32 |
| | (b) Downward movements | | | | | | | |
| percentage | 18.24 | 5.57 | 1.79 | 0.71 | 0.24 | 0.17 | 0.10 | 0.31 |

deep understanding of the way by which the market operates. As such, it is important to specify clearly and precisely the methods used in data cleaning. These methods must be taken into consideration in making inference.

Again, let $t_i$ be the calendar time, measured in seconds from the midnight, when the $i$th transaction took place. Let $P_{t_i}$ be the transaction price. The price change from the $(i-1)$th to the $i$th trade is $y_i \equiv \Delta P_{t_i} = P_{t_i} - P_{t_{i-1}}$ and the time duration is $\Delta t_i = t_i - t_{i-1}$. Here it is understood that the subscript $i$ in $\Delta t_i$ and $y_i$ denotes the time sequence of transactions, not the calendar time. In what follows, we consider models for $y_i$ and $\Delta t_i$ both individually and jointly.

## 5.4 MODELS FOR PRICE CHANGES

The discreteness and concentration on "no change" make it difficult to model the intraday price changes. Campbell, Lo, and MacKinlay (1997) discuss several econometric models that have been proposed in the literature. Here we mention two models that have the advantage of employing explanatory variables to study the intraday price movements. The first model is the ordered probit model used by Hauseman, Lo, and MacKinlay (1992) to study the price movements in transactions data. The second model has been considered recently by McCulloch and Tsay (2000) and is a simplified version of the model proposed by Rydberg and Shephard (1998); see also Ghysels (2000).

### 5.4.1 Ordered Probit Model

Let $y_i^*$ be the unobservable price change of the asset under study (i.e., $y_i^* = P_{t_i}^* - P_{t_{i-1}}^*$), where $P_t^*$ is the *virtual* price of the asset at time $t$. The ordered probit model assumes that $y_i^*$ is a continuous random variable and follows the model

$$y_i^* = \boldsymbol{x}_i \boldsymbol{\beta} + \epsilon_i, \tag{5.15}$$

where $\boldsymbol{x}_i$ is a $p$-dimensional row vector of explanatory variables available at time $t_{i-1}$, $\boldsymbol{\beta}$ is a $k \times 1$ parameter vector, $E(\epsilon_i \mid \boldsymbol{x}_i) = 0$, $\text{Var}(\epsilon_i \mid \boldsymbol{x}_i) = \sigma_i^2$, and $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. The conditional variance $\sigma_i^2$ is assumed to be a positive function of the explanatory variable $\boldsymbol{w}_i$—that is,

$$\sigma_i^2 = g(\boldsymbol{w}_i), \tag{5.16}$$

where $g(.)$ is a positive function. For financial transactions data, $\boldsymbol{w}_i$ may contain the time interval $t_i - t_{i-1}$ and some conditional heteroscedastic variables. Typically, one also assumes that the conditional distribution of $\epsilon_i$ given $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$ is Gaussian.

Suppose that the observed price change $y_i$ may assume $k$ possible values. In theory, $k$ can be infinity, but countable. In practice, $k$ is finite and may involve combin-

ing several categories into a single value. For example, we have $k = 7$ in Table 5.1, where the first value "$-3$ ticks" means that the price change is $-3$ ticks or lower. We denote the $k$ possible values as $\{s_1, \ldots, s_k\}$. The ordered probit model postulates the relationship between $y_i$ and $y_i^*$ as

$$y_i = s_j \quad \text{if} \quad \alpha_{j-1} < y_i^* \le \alpha_j, \quad j = 1, \ldots, k, \tag{5.17}$$

where $\alpha_j$s are real numbers satisfying $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{k-1} < \alpha_k = \infty$. Under the assumption of conditional Gaussian distribution, we have

$$\begin{aligned} P(y_i = s_j \mid \boldsymbol{x}_i, \boldsymbol{w}_i) &= P(\alpha_{j-1} < \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \le \alpha_j \mid \boldsymbol{x}_i, \boldsymbol{w}_i) \\ &= \begin{cases} P(\boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \le \alpha_1 \mid \boldsymbol{x}_i, \boldsymbol{w}_i) & \text{if } j = 1 \\ P(\alpha_{j-1} < \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \le \alpha_j \mid \boldsymbol{x}_i, \boldsymbol{w}_i) & \text{if } j = 2, \ldots, k-1 \\ P(\alpha_{k-1} < \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \mid \boldsymbol{x}_i, \boldsymbol{w}_i) & \text{if } j = k \end{cases} \\ &= \begin{cases} \Phi\left[\frac{\alpha_1 - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i(\boldsymbol{w}_i)}\right] & \text{if } j = 1 \\ \Phi\left[\frac{\alpha_j - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i(\boldsymbol{w}_i)}\right] - \Phi\left[\frac{\alpha_{j-1} - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i(\boldsymbol{w}_i)}\right] & \text{if } j = 2, \ldots, k-1 \\ 1 - \Phi\left[\frac{\alpha_{k-1} - \boldsymbol{x}_i\boldsymbol{\beta}}{\sigma_i(\boldsymbol{w}_i)}\right] & \text{if } j = k, \end{cases} \end{aligned} \tag{5.18}$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal random variable evaluated at $x$, and we write $\sigma_i(\boldsymbol{w}_i)$ to denote that $\sigma_i^2$ is a positive function of $\boldsymbol{w}_i$. From the definition, an ordered probit model is driven by an unobservable continuous random variable. The observed values, which have a natural ordering, can be regarded as categories representing the underlying process.

The ordered probit model contains parameters $\boldsymbol{\beta}$, $\alpha_i$ $(i = 1, \ldots, k-1)$, and those in the conditional variance function $\sigma_i(\boldsymbol{w}_i)$ in Eq. (5.16). These parameters can be estimated by the maximum likelihood or Markov Chain Monte Carlo methods.

**Example 5.1.** Hauseman, Lo, and MacKinlay (1992) apply the ordered probit model to the 1988 transactions data of more than 100 stocks. Here we only report their result for IBM. There are 206,794 trades. The sample mean (standard deviation) of price change $y_i$, time duration $\Delta t_i$, and bid-ask spread are $-0.0010(0.753)$, $27.21(34.13)$, and $1.9470(1.4625)$, respectively. The bid-ask spread is measured in ticks. The model used has nine categories for price movement, and the functional specifications are

$$\begin{aligned} \boldsymbol{x}_i\boldsymbol{\beta} = \beta_1 \Delta t_i^* &+ \sum_{v=1}^{3} \beta_{v+1} y_{i-v} + \sum_{v=1}^{3} \beta_{v+4} \mathrm{SP5}_{i-v} + \sum_{v=1}^{3} \beta_{v+7} \mathrm{IBS}_{i-v} \\ &+ \sum_{v=1}^{3} \beta_{v+10} [T_\lambda(V_{i-v}) \times \mathrm{IBS}_{i-v}] \end{aligned} \tag{5.19}$$

$$\sigma_i^2(\boldsymbol{w}_i) = 1.0 + \gamma_1^2 \Delta t_i^* + \gamma_2^2 \mathrm{AB}_{i-1}, \tag{5.20}$$

where $T_\lambda(V) = (V^\lambda - 1)/\lambda$ is the Box-Cox (1964) transformation of $V$ with $\lambda \in [0, 1]$ and the explanatory variables are defined by the following:

- $\Delta t_i^* = (t_i - t_{i-1})/100$ is a rescaled time duration between the $(i-1)$th and $i$th trades with time measured in seconds.
- $\mathrm{AB}_{i-1}$ is the bid-ask spread prevailing at time $t_{i-1}$ in ticks.
- $y_{i-v}(v = 1, 2, 3)$ is the lagged value of price change at $t_{i-v}$ in ticks. With $k = 9$, the possible values of price changes are $\{-4, -3, -2, -1, 0, 1, 2, 3, 4\}$ in ticks.
- $V_{i-v}(v = 1, 2, 3)$ is the lagged value of dollar volume at the $(i - v)$th transaction, defined as the price of the $(i - v)$th transaction in dollars times the number of shares traded (denominated in hundreds of shares). That is, the dollar volume is in hundreds of dollars.
- $\mathrm{SP5}_{i-v}(v = 1, 2, 3)$ is the 5-minute continuously compounded returns of the Standard and Poor's 500 index futures price for the contract maturing in the closest month beyond the month in which transaction $(i - v)$ occurred, where the return is computed with the futures price recorded one minute before the nearest round minute *prior* to $t_{i-v}$ and the price recorded 5 minutes before this.
- $\mathrm{IBS}_{i-v}(v = 1, 2, 3)$ is an indicator variable defined by

$$\mathrm{IBS}_{i-v} = \begin{cases} 1 & \text{if } P_{i-v} > (P_{i-v}^a + P_{i-v}^b)/2 \\ 0 & \text{if } P_{i-v} = (P_{i-v}^a + P_{i-v}^b)/2 \\ -1 & \text{if } P_{i-v} < (P_{i-v}^a + P_{i-v}^b)/2, \end{cases}$$

  where $P_j^a$ and $P_j^b$ are the ask and bid price at time $t_j$.

The parameter estimates and their $t$ ratios are given in Table 5.4. All the $t$ ratios are large except one, indicating that the estimates are highly significant. Such high $t$ ratios are not surprising as the sample size is large. For the heavily traded IBM stock, the estimation results suggest the following conclusions:

1. The boundary partitions are not equally spaced, but are almost symmetric with respect to zero.
2. The transaction duration $\Delta t_i$ affects both the conditional mean and conditional variance of $y_i$ in Eqs. (5.19) and (5.20).
3. The coefficients of lagged price changes are negative and highly significant, indicating *price reversals*.
4. As expected, the bid-ask spread at time $t_{i-1}$ significantly affects the conditional variance.

**Table 5.4. Parameter Estimates of the Ordered-Probit Model in Eq. (5.19) and Eq. (5.20) for the 1988 Transaction Data of IBM, Where *t* Denotes the *t* Ratio.**

| | (a) Boundary partitions of the probit model | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Par. | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ | $\alpha_8$ |
| Est. | −4.67 | −4.16 | −3.11 | −1.34 | 1.33 | 3.13 | 4.21 | 4.73 |
| *t* | −145.7 | −157.8 | −171.6 | −155.5 | 154.9 | 167.8 | 152.2 | 138.9 |
| | (b) Equation parameters of the probit model | | | | | | | |
| Par. | $\gamma_1$ | $\gamma_2$ | $\beta_1$: $\Delta t_i^*$ | $\beta_2$: $y_{-1}$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ |
| Est. | 0.40 | 0.52 | −0.12 | −1.01 | −0.53 | −0.21 | 1.12 | −0.26 |
| *t* | 15.6 | 71.1 | −11.4 | −135.6 | −85.0 | −47.2 | 54.2 | −12.1 |
| Par. | $\beta_7$ | $\beta_8$ | $\beta_9$: | $\beta_{10}$ | $\beta_{11}$ | $\beta_{12}$ | $\beta_{13}$ | |
| Est. | 0.01 | −1.14 | −0.37 | −0.17 | 0.12 | 0.05 | 0.02 | |
| *t* | 0.26 | −63.6 | −21.6 | −10.3 | 47.4 | 18.6 | 7.7 | |

### 5.4.2 A Decomposition Model

An alternative approach to modeling price change is to decompose it into three components and use conditional specifications for the components; see Rydberg and Shephard (1998). The three components are an indicator for price change, the direction of price movement if there is a change, and the size of price change if a change occurs. Specifically, the price change at the $i$th transaction can be written as

$$y_i \equiv P_{t_i} - P_{t_{i-1}} = A_i D_i S_i, \tag{5.21}$$

where $A_i$ is a binary variable defined as

$$A_i = \begin{cases} 1 & \text{if there is a price change at the } i\text{th trade} \\ 0 & \text{if price remains the same at the } i\text{th trade.} \end{cases} \tag{5.22}$$

$D_i$ is also a discrete variable signifying the *direction* of the price change if a change occurs—that is,

$$D_i \mid (A_i = 1) = \begin{cases} 1 & \text{if price increases at the } i\text{th trade} \\ -1 & \text{if price drops at the } i\text{th trade,} \end{cases} \tag{5.23}$$

where $D_i \mid (A_i = 1)$ means that $D_i$ is defined under the condition of $A_i = 1$, and $S_i$ is size of the price change in ticks if there is a change at the $i$th trade and $S_i = 0$ if there is no price change at the $i$th trade. When there is a price change, $S_i$ is a positive integer-valued random variable.

Note that $D_i$ is not needed when $A_i = 0$, and there is a natural ordering in the decomposition. $D_i$ is well defined only when $A_i = 1$ and $S_i$ is meaningful when

$A_i = 1$ and $D_i$ is given. Model specification under the decomposition makes use of the ordering.

Let $F_i$ be the information set available at the $i$th transaction. Examples of elements in $F_i$ are $\Delta t_{i-j}$, $A_{i-j}$, $D_{i-j}$, and $S_{i-j}$ for $j \geq 0$. The evolution of price change under model (5.21) can then be partitioned as

$$\begin{aligned} P(y_i \mid F_{i-1}) &= P(A_i D_i S_i \mid F_{i-1}) \\ &= P(S_i \mid D_i, A_i, F_{i-1}) P(D_i \mid A_i, F_{i-1}) P(A_i \mid F_{i-1}). \end{aligned} \tag{5.24}$$

Since $A_i$ is a binary variable, it suffices to consider the evolution of the probability $p_i = P(A_i = 1)$ over time. We assume that

$$\ln\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_i\boldsymbol{\beta} \quad \text{or} \quad p_i = \frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}}, \tag{5.25}$$

where $\boldsymbol{x}_i$ is a finite-dimensional vector consisting of elements of $F_{i-1}$ and $\boldsymbol{\beta}$ is a parameter vector. Conditioned on $A_i = 1$, $D_i$ is also a binary variable, and we use the following model for $\delta_i = P(D_i = 1 \mid A_i = 1)$,

$$\ln\left(\frac{\delta_i}{1-\delta_i}\right) = \boldsymbol{z}_i\boldsymbol{\gamma} \quad \text{or} \quad \delta_i = \frac{e^{\boldsymbol{z}_i\boldsymbol{\gamma}}}{1+e^{\boldsymbol{z}_i\boldsymbol{\gamma}}}, \tag{5.26}$$

where $\boldsymbol{z}_i$ is a finite-dimensional vector consisting of elements of $F_{i-1}$ and $\boldsymbol{\gamma}$ is a parameter vector. To allow for asymmetry between positive and negative price changes, we assume that

$$S_i \mid (D_i, A_i = 1) \sim 1 + \begin{cases} g(\lambda_{u,i}) & \text{if } D_i = 1, A_i = 1 \\ g(\lambda_{d,i}) & \text{if } D_i = -1, A_i = 1, \end{cases} \tag{5.27}$$

where $g(\lambda)$ is a geometric distribution with parameter $\lambda$ and the parameters $\lambda_{j,i}$ evolve over time as

$$\ln\left(\frac{\lambda_{j,i}}{1-\lambda_{j,i}}\right) = \boldsymbol{w}_i\boldsymbol{\theta}_j \quad \text{or} \quad \lambda_{j,i} = \frac{e^{\boldsymbol{w}_i\boldsymbol{\theta}_j}}{1+e^{\boldsymbol{w}_i\boldsymbol{\theta}_j}}, \quad j = u, d, \tag{5.28}$$

where $\boldsymbol{w}_i$ is again a finite-dimensional explanatory variables in $F_{i-1}$ and $\boldsymbol{\theta}_j$ is a parameter vector.

In Eq. (5.27), the probability mass function of a random variable $x$, which follows the geometric distribution $g(\lambda)$, is

$$p(x = m) = \lambda(1-\lambda)^m, \quad m = 0, 1, 2, \ldots.$$

We added 1 to the geometric distribution so that the price change, if it occurs, is at least 1 tick. In Eq. (5.28), we take the logistic transformation to ensure that $\lambda_{j,i} \in [0, 1]$.

The previous specification classifies the $i$th trade, or transaction, into one of three categories:

1. no price change: $A_i = 0$ and the associated probability is $(1 - p_i)$;
2. a price increase: $A_i = 1$, $D_i = 1$, and the associated probability is $p_i\delta_i$. The size of the price increase is governed by $1 + g(\lambda_{u,i})$.
3. a price drop: $A_i = 1$, $D_i = -1$, and the associated probability is $p_i(1 - \delta_i)$. The size of the price drop is governed by $1 + g(\lambda_{d,i})$.

Let $I_i(j)$ for $j = 1, 2, 3$ be the indicator variables of the prior three categories. That is, $I_i(j) = 1$ if the $j$th category occurs and $I_i(j) = 0$ otherwise. The log likelihood function of Eq. (5.24) becomes

$$\begin{aligned}\ln[P(y_i \mid F_{i-1})] = {} & I_i(1)\ln[(1 - p_i)] + I_i(2)[\ln(p_i) + \ln(\delta_i) \\ & + \ln(\lambda_{u,i}) + (S_i - 1)\ln(1 - \lambda_{u,i})] \\ & + I_i(3)[\ln(p_i) + \ln(1 - \delta_i) + \ln(\lambda_{d,i}) + (S_i - 1)\ln(1 - \lambda_{d,i})],\end{aligned}$$

and the overall log likelihood function is

$$\ln[P(y_1, \ldots, y_n \mid F_0)] = \sum_{i=1}^{n} \ln P(y_i \mid F_{i-1})], \tag{5.29}$$

which is a function of parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\theta}_u$, and $\boldsymbol{\theta}_d$.

**Example 5.2.** We illustrate the decomposition model by analyzing the intraday transactions of IBM stock from November 1, 1990 to January 31, 1991. There were 63 trading days and 59,838 intraday transactions in the normal trading hours. The explanatory variables used are

1. $A_{i-1}$: The action indicator of the previous trade (i.e., the $[i - 1]$th trade within a trading day).
2. $D_{i-1}$: The direction indicator of the previous trade.
3. $S_{i-1}$: The size of the previous trade.
4. $V_{i-1}$: The volume of the previous trade, divided by 1000.
5. $\Delta t_{i-1}$: Time duration from the $(i - 2)$th to $(i - 1)$th trade.
6. $BA_i$: The bid-ask spread prevailing at the time of transaction.

Because we use lag-1 explanatory variables, the actual sample size is 59,775. It turns out that $V_{i-1}$, $\Delta t_{i-1}$ and $BA_i$ are not statistically significant for the model entertained. Thus, only the first three explanatory variables are used. The model employed is

$$\begin{aligned}
\ln\left(\frac{p_i}{1-p_i}\right) &= \beta_0 + \beta_1 A_{i-1} \\
\ln\left(\frac{\delta_i}{1-\delta_i}\right) &= \gamma_0 + \gamma_1 D_{i-1} \\
\ln\left(\frac{\lambda_{u,i}}{1-\lambda_{u,i}}\right) &= \theta_{u,0} + \theta_{u,1} S_{i-1} \\
\ln\left(\frac{\lambda_{d,i}}{1-\lambda_{d,i}}\right) &= \theta_{d,0} + \theta_{d,1} S_{i-1}.
\end{aligned} \tag{5.30}$$

The parameter estimates, using the log-likelihood function in Eq. (5.29), are given in Table 5.5. The estimated simple model shows some dynamic dependence in the price change. In particular, the trade-by-trade price changes of IBM stock exhibit some appealing features:

**1.** The probability of a price change depends on the previous price change. Specifically, we have

$$P(A_i = 1 \mid A_{i-1} = 0) = 0.258, \quad P(A_i = 1 \mid A_{i-1} = 1) = 0.476.$$

The result indicates that a price change may occur in clusters and, as expected, most transactions are without price change. When no price change occurred at the $(i-1)$th trade, then only about one out of four trades in the subsequent transaction has a price change. When there is a price change at the $(i-1)$th transaction, the probability of a price change in the $i$th trade increases to about 0.5.

**2.** The direction of price change is governed by

$$P(D_i = 1 \mid F_{i-1}, A_i) = \begin{cases} 0.483 & \text{if } D_{i-1} = 0 \text{ (i.e., } A_{i-1} = 0) \\ 0.085 & \text{if } D_{i-1} = 1,\ A_i = 1 \\ 0.904 & \text{if } D_{i-1} = -1,\ A_i = 1. \end{cases}$$

This result says that (a) if no price change occurred at the $(i-1)$th trade, then the chances for a price increase or decrease at the $i$th trade are about even; and (b) the probabilities of consecutive price increases or decreases are very low. The probability of a price increase at the $i$th trade given that a price change

**Table 5.5. Parameter Estimates of the ADS Model in Eq. (5.30) for IBM Intraday Transactions: 11/01/90 to 1/31/91.**

| Parameter | $\beta_0$ | $\beta_1$ | $\gamma_0$ | $\gamma_1$ | $\theta_{u,0}$ | $\theta_{u,1}$ | $\theta_{d,0}$ | $\theta_{d,1}$ |
|---|---|---|---|---|---|---|---|---|
| Estimate | −1.057 | 0.962 | −0.067 | −2.307 | 2.235 | −0.670 | 2.085 | −0.509 |
| Std.Err. | 0.104 | 0.044 | 0.023 | 0.056 | 0.029 | 0.050 | 0.187 | 0.139 |

occurs at the $i$th trade and there was a price increase at the $(i-1)$th trade is only 8.6%. However, the probability of a price increase is about 90% given that a price change occurs at the $i$th trade and there was a price decrease at the $(i-1)$th trade. Consequently, this result shows the effect of bid-ask bounce and supports price reversals in high-frequency trading.

3. There is weak evidence suggesting that big price changes have a higher probability to be followed by another big price change. Consider the size of a price increase. We have

$$S_i \mid (D_i = 1) \sim 1 + g(\lambda_{u,i}), \quad \lambda_{u,i} = 2.235 - 0.670 S_{i-1}.$$

Using the probability mass function of a geometric distribution, we obtain that the probability of a price increase by one tick is 0.827 at the $i$th trade if the transaction results in a price increase and $S_{i-1} = 1$. The probability reduces to 0.709 if $S_{i-1} = 2$ and to 0.556 if $S_{i-1} = 3$. Consequently, the probability of a large $S_i$ is proportional to $S_{i-1}$ given that there is a price increase at the $i$th trade.

A difference between the ADS and ordered probit models is that the ADS model does not require any truncation or grouping in the size of a price change.

## 5.5 DURATION MODELS

Duration models are concerned with time intervals between trades. Longer durations indicate lack of trading activities, which in turn signify a period of no new information. The dynamic behavior of durations, thus, contains useful information about intraday market activities. Using concepts similar to the ARCH models for volatility, Engle and Russell (1998) propose an autoregressive conditional duration (ACD) model to describe the evolution of time durations for (heavily traded) stocks. Zhang, Russell, and Tsay (2001) extend the ACD model to account for nonlinearity and structural breaks in the data. In this section, we introduce some simple duration models. As mentioned before, intraday transactions exhibit some diurnal pattern. Therefore, we focus on the adjusted time duration

$$\Delta t_i^* = \Delta t_i / f(t_i), \tag{5.31}$$

where $f(t_i)$ is a deterministic function consisting of the cyclical component of $\Delta t_i$. Obviously, $f(t_i)$ depends on the underlying asset and the systematic behavior of the market. In practice, there are many ways to estimate $f(t_i)$, but no single method dominates the others in terms of statistical properties. A common approach is to use smoothing spline. Here we use simple quadratic functions and indicator variables to take care of the deterministic component of daily trading activities.

For the IBM data employed in the illustration of ADS models, we assume

$$f(t_i) = \exp[d(t_i)], \quad d(t_i) = \beta_0 + \sum_{j=1}^{7} \beta_j f_j(t_i), \tag{5.32}$$

where

$$f_1(t_i) = -\left(\frac{t_i - 43200}{14400}\right)^2, \quad f_3(t_i) = \begin{cases} -\left(\frac{t_i - 38700}{7500}\right)^2 & \text{if } t_i < 43200 \\ 0 & \text{otherwise,} \end{cases}$$

$$f_2(t_i) = -\left(\frac{t_i - 48300}{9300}\right)^2, \quad f_4(t_i) = \begin{cases} -\left(\frac{t_i - 48600}{9000}\right)^2 & \text{if } t_i \geq 43200 \\ 0 & \text{otherwise,} \end{cases}$$

$f_5(t_i)$ and $f_6(t_i)$ are indicator variables for the first and second 5 minutes of market opening [i.e., $f_5(.) = 1$ if and only if $t_i$ is between 9:30 am and 9:35 am Eastern

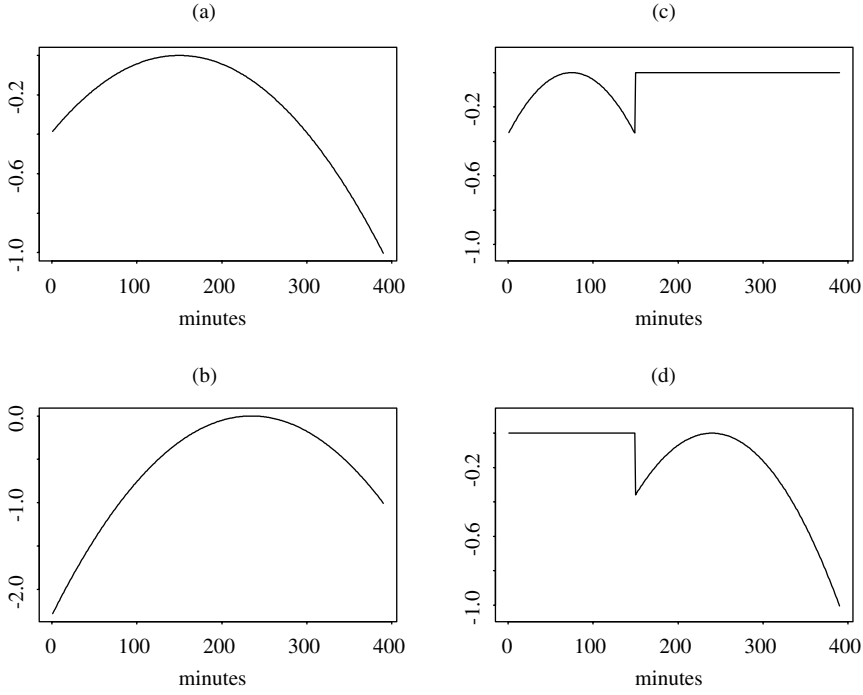


**Figure 5.5.** Quadratic functions used to remove the deterministic component of IBM intraday trading durations: (a)–(d) are the functions $f_1(.)$ to $f_4(.)$ of Eq. (5.32), respectively.

Time], and $f_7(t_i)$ is the indicator for the last 30 minutes of daily trading [i.e., $f_7(t_i) = 1$ if and only if the trade occurred between 3:30 pm and 4:00 pm Eastern Time]. Figure 5.5 shows the plot of $f_i(.)$ for $i = 1, \ldots, 4$, where the time scales in the x-axis is in minutes. Note that $f_3(43{,}200) = f_4(43{,}200)$, where 43,200 corresponds to 12:00 noon.

The coefficients $\beta_j$ of Eq. (5.32) are obtained by the least squares method of the linear regression

$$\ln(\Delta t_i) = \beta_0 + \sum_{j=1}^{7} \beta_j f_j(t_i) + \epsilon_i.$$

The fitted model is

$$\begin{aligned}\widehat{\ln(\Delta t_i)} = {} & 2.555 + 0.159 f_1(t_i) + 0.270 f_2(t_i) + 0.384 f_3(t_i) \\ & + 0.061 f_4(t_i) - 0.611 f_5(t_i) - 0.157 f_6(t_i) + 0.073 f_7(t_i).\end{aligned}$$

Figure 5.6 shows the time plot of average durations in 5-minute time intervals over the 63 trading days before and after adjusting for the deterministic component. Part

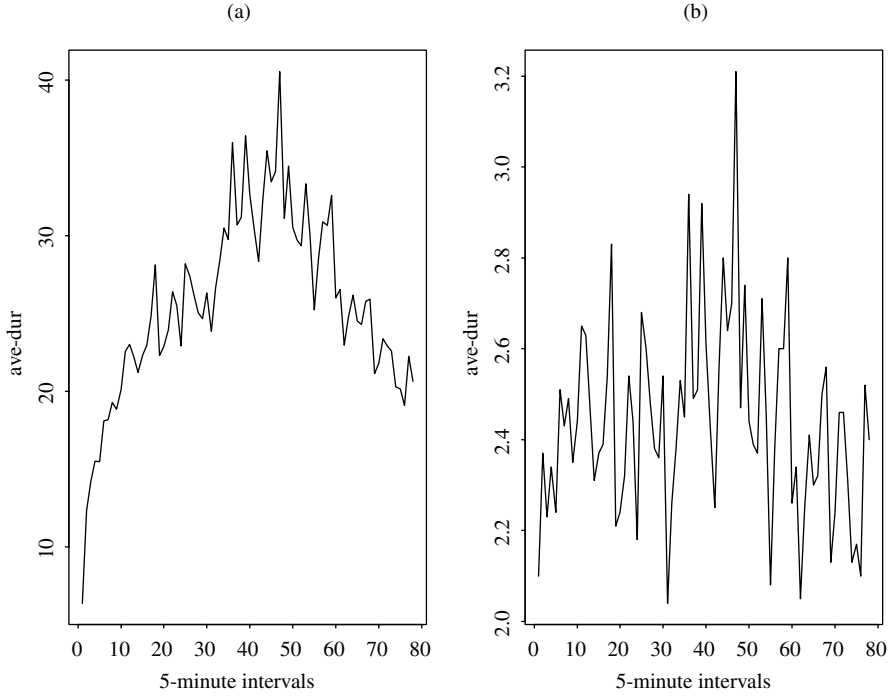


**Figure 5.6.** IBM transactions data from 11/01/90 to 1/31/91: (a) The average durations in 5-minute time intervals, and (b) the average durations in 5-minute time intervals after adjusting for the deterministic component.

(a) is the average durations of $\Delta t_i$ and, as expected, it exhibits a diurnal pattern. Part (b) is the average durations of $\Delta t_i^*$ (i.e., after the adjustment), and the diurnal pattern is largely removed.

### 5.5.1 The ACD Model

The autoregressive conditional duration (ACD) model uses the idea of GARCH models to study the dynamic structure of the adjusted duration $\Delta t_i^*$ of Eq. (5.31). For ease in notation, we define $x_i = \Delta t_i^*$.

Let $\psi_i = E(x_i \mid F_{i-1})$ be the conditional expectation of the adjusted duration between the $(i-1)$th and $i$th trades, where $F_{i-1}$ is the information set available at the $(i-1)$th trade. In other words, $\psi_i$ is the expected adjusted duration given $F_{i-1}$. The basic ACD model is defined as

$$x_i = \psi_i \epsilon_i, \tag{5.33}$$

where $\{\epsilon_i\}$ is a sequence of independent and identically distributed non-negative random variables such that $E(\epsilon_i) = 1$. In Engle and Russell (1998), $\epsilon_i$ follows a standard exponential or a standardized Weibull distribution, and $\psi_i$ assumes the form

$$\psi_i = \omega + \sum_{j=1}^{r} \gamma_j x_{i-j} + \sum_{j=1}^{s} \omega_j \psi_{i-j}. \tag{5.34}$$

Such a model is referred to as an ACD$(r, s)$ model. When the distribution of $\epsilon_i$ is exponential, the resulting model is called an EACD$(r, s)$ model. Similarly, if $\epsilon_i$ follows a Weibull distribution, the model is a WACD$(r, s)$ model. If necessary, readers are referred to Appendix A for a quick review of exponential and Weibull distributions.

Similar to GARCH models, the process $\eta_i = x_i - \psi_i$ is a Martingale difference sequence [i.e., $E(\eta_i \mid F_{i-1}) = 0$], and the ACD$(r, s)$ model can be written as

$$x_i = \omega + \sum_{j=1}^{\max(r,s)} (\gamma_j + \omega_j) x_{i-j} - \sum_{j=1}^{s} \omega_j \eta_{i-j} + \eta_j, \tag{5.35}$$

which is in the form of an ARMA process with non-Gaussian innovations. It is understood here that $\gamma_j = 0$ for $j > r$ and $\omega_j = 0$ for $j > s$. Such a representation can be used to obtain the basic conditions for weak stationarity of the ACD model. For instance, taking expectation on both sides of Eq. (5.35) and assuming weak stationarity, we have

$$E(x_i) = \frac{\omega}{1 - \sum_{j=1}^{\max(r,s)} (\gamma_j + \omega_j)}.$$

Therefore, we assume $\omega > 0$ and $1 > \sum_j (\gamma_j + \omega_j)$ because the expected duration is positive. As another application of Eq. (5.35), we study properties of the EACD(1, 1) model.

**EACD(1, 1) *Model***
An EACD(1, 1) model can be written as

$$x_i = \psi_i \epsilon_i, \quad \psi_i = \omega + \gamma_1 x_{i-1} + \omega_1 \psi_{i-1}, \tag{5.36}$$

where $\epsilon_i$ follows the standard exponential distribution. Using the moments of a standard exponential distribution in Appendix A, we have $E(\epsilon_i) = 1$, $\text{Var}(\epsilon_i) = 1$, and $E(\epsilon_i^2) = \text{Var}(x_i) + [E(x_i)]^2 = 2$. Assuming that $x_i$ is weakly stationary (i.e., the first two moments of $x_i$ are time-invariant), we derive the variance of $x_i$. First, taking expectation of Eq. (5.36), we have

$$E(x_i) = E[E(\psi_i \epsilon_i \mid F_{i-1})] = E(\psi_i), \quad E(\psi_i) = \omega + \gamma_1 E(x_{i-1}) + \omega_1 E(\psi_{i-1}). \tag{5.37}$$

Under weak stationarity, $E(\psi_i) = E(\psi_{i-1})$ so that Eq. (5.37) gives

$$\mu_x \equiv E(x_i) = E(\psi_i) = \frac{\omega}{1 - \gamma_1 - \omega_1}. \tag{5.38}$$

Next, because $E(\epsilon_i^2) = 2$, we have $E(x_i^2) = E[E(\psi_i^2 \epsilon_i^2 \mid F_{i-1})] = 2E(\psi_i^2)$.

Taking square of $\psi_i$ in Eq. (5.36) and expectation and using weak stationarity of $\psi_i$ and $x_i$, we have, after some algebra, that

$$E(\psi_i^2) = \mu_x^2 \times \frac{1 - (\gamma_1 + \omega_1)^2}{1 - 2\gamma_1^2 - \omega_1^2 - 2\gamma_1\omega_1}. \tag{5.39}$$

Finally, using $\text{Var}(x_i) = E(x_i^2) - [E(x_i)]^2$ and $E(x_i^2) = 2E(\psi_i^2)$, we have

$$\text{Var}(x_i) = 2E(\psi_i^2) - \mu_x^2 = \mu_x^2 \times \frac{1 - \omega_1^2 - 2\gamma_1\omega_1}{1 - \omega_1^2 - 2\gamma_1\omega_1 - 2\gamma_1^2},$$

where $\mu_x$ is defined in Eq. (5.38). This result shows that, to have time-invariant unconditional variance, the EACD(1, 1) model in Eq. (5.36) must satisfy $1 > 2\gamma_1^2 + \omega_1^2 + 2\gamma_1\omega_1$. The variance of an WACD(1, 1) model can be obtained by using the same techniques and the first two moments of a standardized Weibull distribution.

***ACD Models with a Generalized Gamma Distribution***
In the statistical literature, intensity function is often expressed in terms of hazard function. As shown in Appendix B, the hazard function of an EACD model is constant over time and that of an WACD model is a monotonous function. These hazard functions are rather restrictive in application as the intensity function of stock trans-

actions might not be constant or monotone over time. To increase the flexibility of the associated hazard function, Zhang, Russell, and Tsay (2001) employ a (standardized) generalized Gamma distribution for $\epsilon_i$. See Appendix A for some basic properties of a generalized Gamma distribution. The resulting hazard function may assume various patterns, including U shape or inverted U shape. We refer to an ACD model with innovations that follow a generalized Gamma distribution as a GACD($r$, $s$) model.

### 5.5.2 Simulation

To illustrate ACD processes, we generated 500 observations from the ACD(1, 1) model

$$x_i = \psi_i \epsilon_i, \quad \psi_i = 0.3 + 0.2x_{i-1} + 0.7\psi_{i-1} \tag{5.40}$$

using two different innovational distributions for $\epsilon_i$. In case 1, $\epsilon_i$ is assumed to follow a standardized Weibull distribution with parameter $\alpha = 1.5$. In case 2, $\epsilon_i$ follows a (standardized) generalized Gamma distribution with parameters $\kappa = 1.5$ and $\alpha = 0.5$.

Figure 5.7(a) shows the time plot of the WACD(1, 1) series, whereas Figure 5.8(a) is the GACD(1, 1) series. Figure 5.9 plots the histograms of both simulated series.

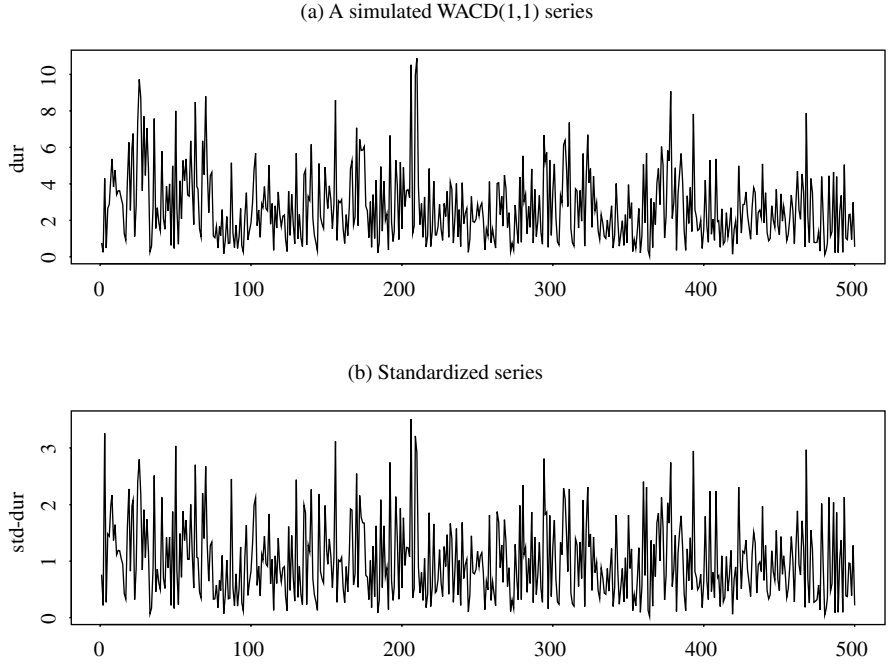


**Figure 5.7.** A simulated WACD(1, 1) series in Eq. (5.40): (a) the original series, and (b) the standardized series after estimation. There are 500 observations.

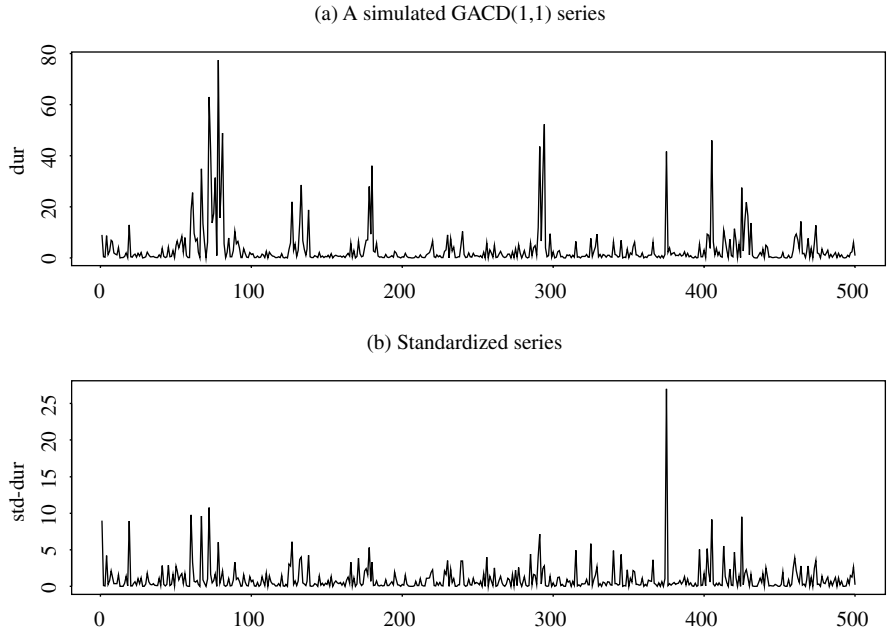


**Figure 5.8.** A simulated GACD(1, 1) series in Eq. (5.40): (a) the original series, and (b) the standardized series after estimation. There are 500 observations.

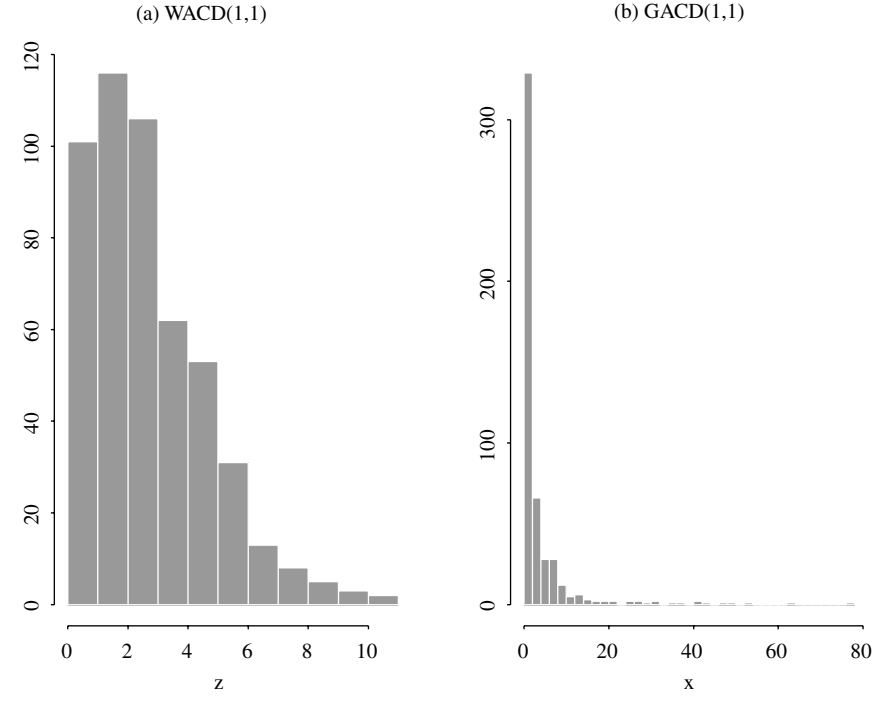


**Figure 5.9.** Histograms of simulated duration processes with 500 observations: (a) WACD(1, 1) model, and (b) GACD(1, 1) model

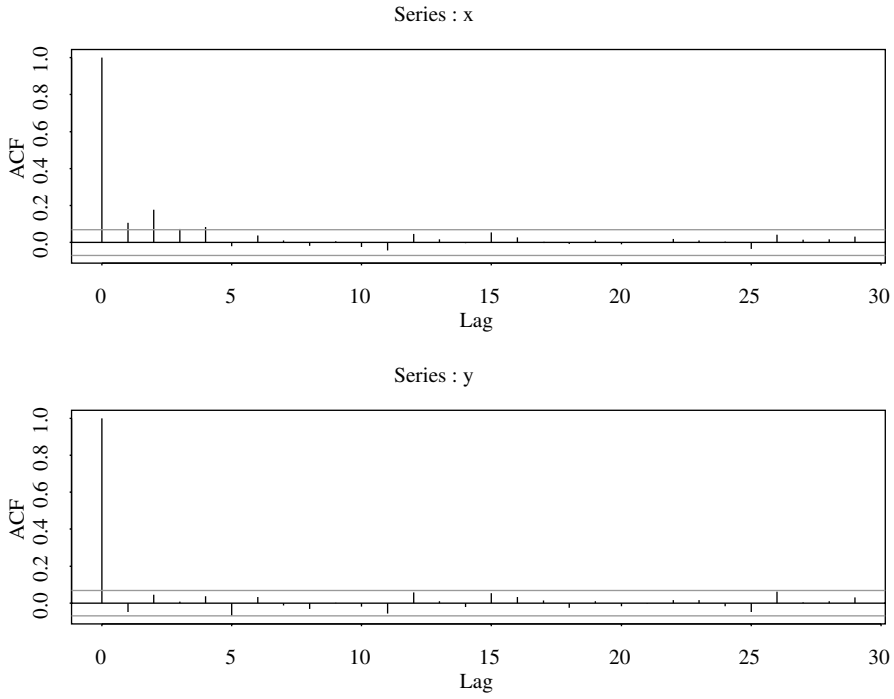


**Figure 5.10.** The sample autocorrelation function of a simulated WACD(1, 1) series with 500 observations: (a) the original series, and (b) the standardized residual series.

The difference between the two models is evident. Finally, the sample ACF of the two simulated series are shown in Figure 5.10(a) and Figure 5.11(b), respectively. The serial dependence of the data is clearly seen.

### 5.5.3 Estimation

For an ACD$(r, s)$ model, let $i_o = \max(r, s)$ and $\boldsymbol{x}_t = (x_1, \ldots, x_t)'$. The likelihood function of the durations $x_1, \ldots, x_T$ is

$$f(\boldsymbol{x}_T \mid \boldsymbol{\theta}) = \left[ \prod_{i=i_o+1}^{T} f(x_i \mid F_{i-1}, \boldsymbol{\theta}) \right] \times f(\boldsymbol{x}_{i_o} \mid \boldsymbol{\theta}),$$

where $\boldsymbol{\theta}$ denotes the vector of model parameters, and $T$ is the sample size. The marginal probability density function $f(\boldsymbol{x}_{i_o} \mid \boldsymbol{\theta})$ of the previous equation is rather complicated for a general ACD model. Because its impact on the likelihood function is diminishing as the sample size $T$ increases, this marginal density is often ignored, resulting in the use of conditional likelihood method. For a WACD model, we use the probability density function (pdf) of Eq. (5.55) and obtain the conditional log

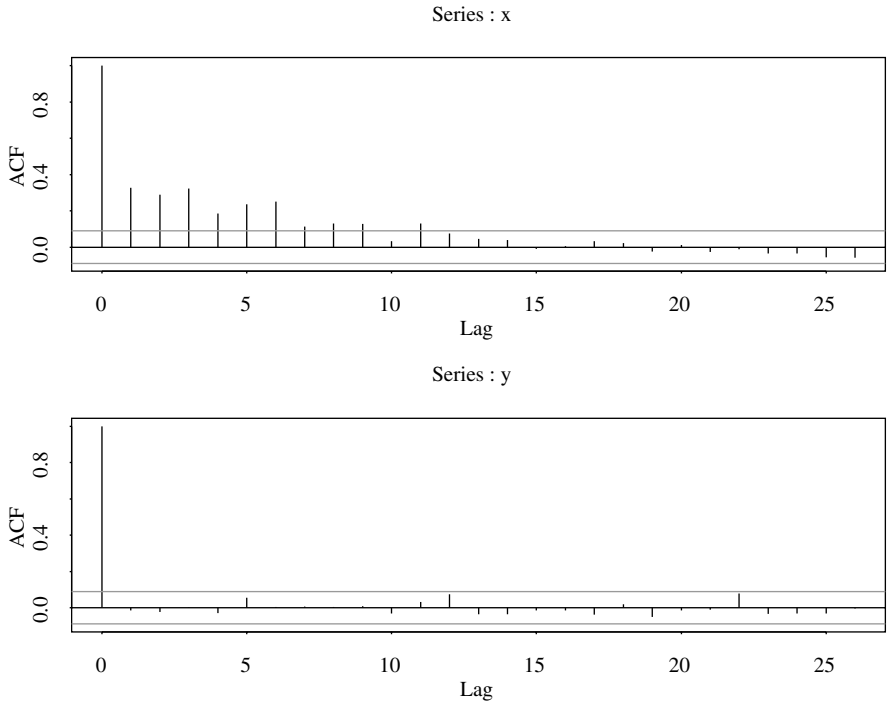


**Figure 5.11.** The sample autocorrelation function of a simulated GACD(1, 1) series with 500 observations: (a) the original series, and (b) the standardized residual series.

likelihood function

$$\ell(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{x}_{i_o}) = \sum_{i=i_0+1}^{T} \alpha \ln\left[\Gamma\left(1+\frac{1}{\alpha}\right)\right] + \ln\left(\frac{\alpha}{x_i}\right) + \alpha \ln\left(\frac{x_i}{\psi_i}\right) - \left[\frac{\Gamma\left(1+\frac{1}{\alpha}\right) x_i}{\psi_i}\right]^{\alpha}, \tag{5.41}$$

where $\psi_i = \omega + \sum_{j=1}^{r} \gamma_j x_{i-j} + \sum_{j=1}^{s} \omega_j \psi_{i-j}$, $\boldsymbol{\theta} = (\omega, \gamma_1, \ldots, \gamma_r, \omega_1, \ldots, \omega_s, \alpha)'$ and $\boldsymbol{x} = (x_{i_o+1}, \ldots, x_T)'$. When $\alpha = 1$, the (conditional) log likelihood function reduces to that of an EACD($r$, $s$) model.

For a GACD($r$, $s$) model, the conditional log likelihood function is

$$\ell(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{x}_{i_o}) = \sum_{i=i_o+1}^{T} \ln\left(\frac{\alpha}{\Gamma(\kappa)}\right) + (\kappa\alpha - 1)\ln(x_i) - \kappa\alpha \ln(\lambda\psi_i) - \left(\frac{x_i}{\lambda\psi_i}\right)^{\alpha}, \tag{5.42}$$

where $\lambda = \Gamma(\kappa)/\Gamma(\kappa + \frac{1}{\alpha})$ and the parameter vector $\boldsymbol{\theta}$ now also includes $\kappa$. As expected, when $\kappa = 1$, $\lambda = 1/\Gamma(1+\frac{1}{\alpha})$ and the log likelihood function in Eq. (5.42)

reduces to that of a WACD($r$, $s$) model in Eq. (5.41). This log likelihood function can be rewritten in many ways to simplify the estimation.

Under some regularity conditions, the conditional maximum likelihood estimates are asymptotically normal; see Engle and Russell (1998) and the references therein. In practice, simulation can be used to obtain finite-sample reference distributions for the problem of interest once a duration model is specified.

**Example 5.3.** (Simulated ACD(1,1) series continued) Consider the simulated WACD(1,1) and GACD(1, 1) series of Eq. (5.40). We apply the conditional likelihood method and obtain the results in Table 5.6. The estimates appear to be reasonable. Let $\hat{\psi}_i$ be the 1-step ahead prediction of $\psi_i$ and $\hat{\epsilon}_i = x_i/\hat{\psi}_i$ be the standardized series, which can be regarded as standardized residuals of the series. If the model is adequately specified, $\{\hat{\epsilon}_i\}$ should behave as a sequence of independent and identically distributed random variables. Figure 5.7(b) and Figure 5.8(b) show the time plot of $\hat{\epsilon}_i$ for both models. The sample ACF of $\hat{\epsilon}_i$ for both fitted models are shown in Figure 5.10(b) and Figure 5.11(b), respectively. It is evident that no significant serial correlations are found in the $\hat{\epsilon}_i$ series.

**Example 5.4.** As an illustration of duration models, we consider the transaction durations of IBM stock on five consecutive trading days from November 1 to November 7, 1990. Focusing on positive transaction durations, we have 3534 observations. In addition, the data have been adjusted by removing the deterministic component in Eq. (5.32). That is, we employ 3534 positive adjusted durations as defined in Eq. (5.31).

Figure 5.12(a) shows the time plot of the adjusted (positive) durations for the first five trading days of November 1990, and Figure 5.13(a) gives the sample ACF of the series. There exist some serial correlations in the adjusted durations. We fit a WACD(1, 1) model to the data and obtain the model

$$x_i = \psi_i \epsilon_i, \quad \psi_i = 0.169 + 0.064 x_{i-1} + 0.885 \psi_{i-1}, \tag{5.43}$$

**Table 5.6. Estimation Results for Simulated ACD(1,1) Series with 500 Observations: (a) for WACD(1,1) Series and (b) for GACD(1,1) Series.**

| | | | (a) WACD(1,1) model | | |
|---|---|---|---|---|---|
| Parameter | $\omega$ | $\gamma_1$ | $\omega_1$ | $\alpha$ | |
| True | 0.3 | 0.2 | 0.7 | 1.5 | |
| Estimate | 0.364 | 0.100 | 0.767 | 1.477 | |
| Std Error | (0.139) | (0.025) | (0.060) | (0.052) | |
| | | | (b) GACD(1,1) model | | |
| Parameter | $\omega$ | $\gamma_1$ | $\omega_1$ | $\alpha$ | $\kappa$ |
| True | 0.3 | 0.2 | 0.7 | 0.5 | 1.5 |
| Estimate | 0.401 | 0.343 | 0.561 | 0.436 | 2.077 |
| Std Error | (0.117) | (0.074) | (0.065) | (0.078) | (0.653) |

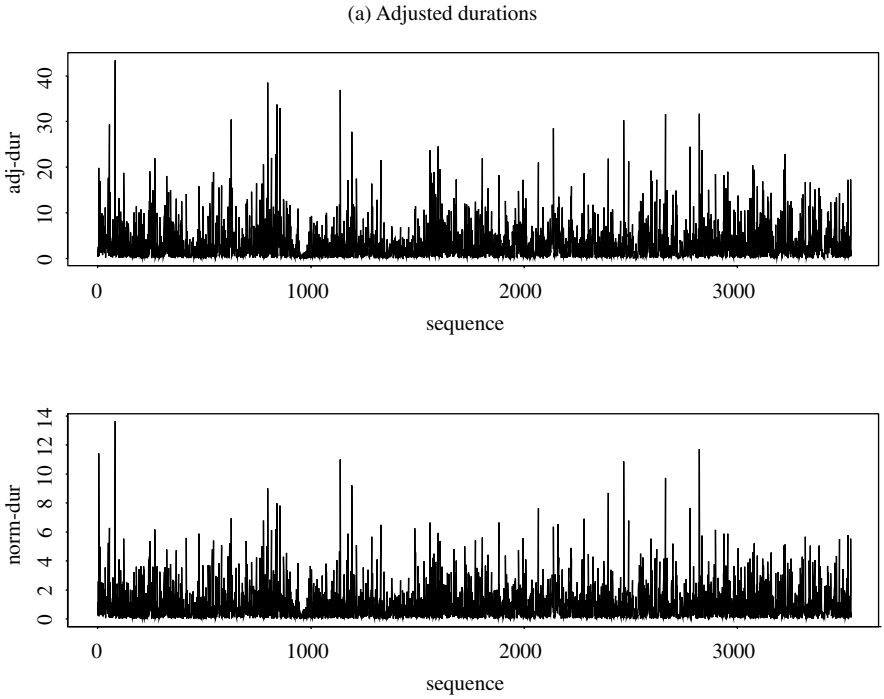


**Figure 5.12.** Time plots of durations for IBM stock traded in the first five trading days of November 1990: (a) the adjusted series, and (b) the normalized innovations of an WACD(1, 1) model. There are 3534 nonzero durations.

where $\{\epsilon_i\}$ is a sequence of independent and identically distributed random variates that follow the standardized Weibull distribution with parameter $\hat{\alpha} = 0.879(0.012)$, where 0.012 is the estimated standard error. Standard errors of the estimates in Eq. (5.43) are 0.039, 0.010, and 0.018, respectively. All $t$ ratios of the estimates are greater than 4.2, indicating that the estimates are significant at the 1% level. Figure 5.12(b) shows the time plot of $\hat{\epsilon}_i = x_i/\hat{\psi}_i$, and Figure 5.13(b) provides the sample ACF of $\hat{\epsilon}_i$. The Ljung–Box statistics show $Q(10) = 4.96$ and $Q(20) = 10.75$ for the $\hat{\epsilon}_i$ series. Clearly, the standardized innovations have no significant serial correlations. In fact, the sample autocorrelations of the squared series $\{\hat{\epsilon}_i^2\}$ are also small with $Q(10) = 6.20$ and $Q(20) = 11.16$, further confirming lack of serial dependence in the normalized innovations. In addition, the mean and standard deviation of a standardized Weibull distribution with $\alpha = 0.879$ are 1.00 and 1.14, respectively. These numbers are close to the sample mean and standard deviation of $\{\hat{\epsilon}_i\}$, which are 1.01 and 1.22, respectively. The fitted model seems adequate.

In model (5.43), the estimated coefficients show $\hat{\gamma}_1 + \hat{\omega}_1 \approx 0.949$, indicating certain persistence in the adjusted durations. The expected adjusted duration is $0.169/(1 - 0.064 - 0.885) = 3.31$ seconds, which is close to the sample mean 3.29 of the adjusted durations. The estimated $\alpha$ of the standardized Weibull distribution

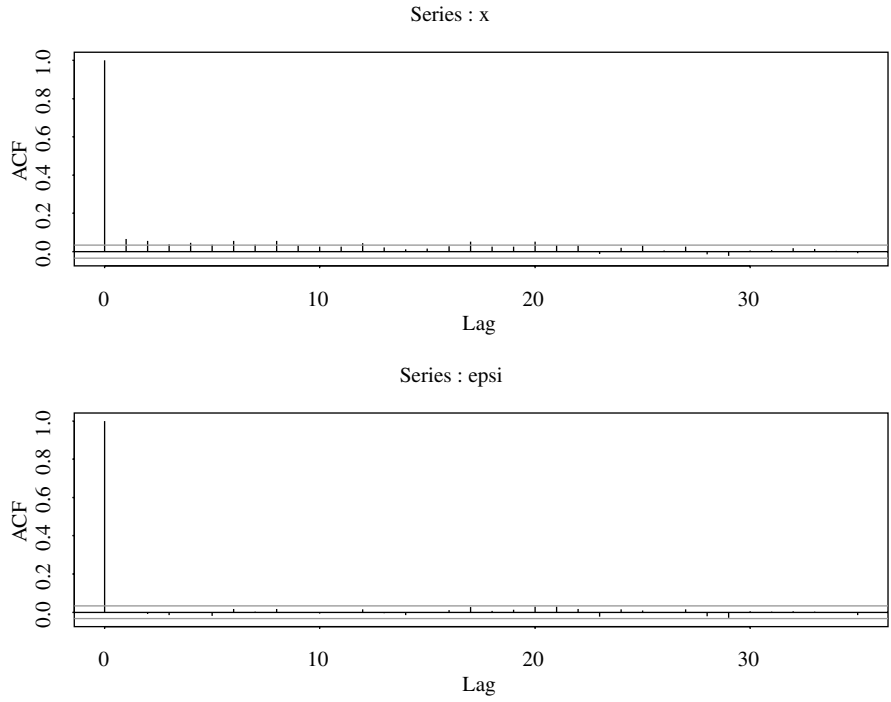


**Figure 5.13.** The sample autocorrelation function of adjusted durations for IBM stock traded in the first five trading days of November 1990: (a) the adjusted series, and (b) the normalized innovations for a WACD(1, 1) model.

is 0.879, which is less than but close to 1. Thus, the conditional hazard function is monotonously decreasing at a slow rate.

If a generalized Gamma distribution function is used for the innovations, then the fitted GACD(1, 1) model is

$$x_i = \psi_i \epsilon_i, \quad \psi_i = 0.141 + 0.063 x_{i-1} + 0.897 \psi_{i-1}, \tag{5.44}$$

where $\{\epsilon_i\}$ follows a standardized, generalized Gamma distribution in Eq. (5.56) with parameters $\kappa = 4.248(1.046)$ and $\alpha = 0.395(0.053)$, where the number in parentheses denotes estimated standard error. Standard errors of the three parameters in Eq. (5.44) are 0.041, 0.010, and 0.019, respectively. All of the estimates are statistically significant at the 1% level. Again, the normalized innovational process $\{\hat{\epsilon}_i\}$ and its squared series have no significant serial correlation, where $\hat{\epsilon}_i = x_i/\hat{\psi}_i$ based on model (5.44). Specifically, for the $\hat{\epsilon}_i$ process, we have $Q(10) = 4.95$ and $Q(20) = 10.28$. For the $\hat{\epsilon}_i^2$ series, we have $Q(10) = 6.36$ and $Q(20) = 10.89$.

The expected duration of model (5.44) is 3.52, which is slightly greater than that of the WACD(1, 1) model in Eq. (5.43). Similarly, the persistence parameter $\hat{\gamma}_1 + \hat{\omega}_1$ of model (5.44) is also slightly higher at 0.96.

***Remark:*** Estimation of EACD models can be carried out by using programs for ARCH models with some minor modification; see Engle and Russell (1998). In this book, we use either the RATS program or some Fortran programs developed by the author to estimate the duration models. Limited experience indicates that it is harder to estimate a GACD model than an EACD or a WACD model. RATS programs used to estimate WACD and GACD models are given in Appendix C.

## 5.6 NONLINEAR DURATION MODELS

Nonlinear features are also commonly found in high-frequency data. As an illustration, we apply some nonlinearity tests discussed in Chapter 4 to the normalized innovations $\hat{\epsilon}_i$ of the WACD(1, 1) model for the IBM transaction durations in Example 5.4; see Eq. (5.43). Based on an AR(4) model, the test results are given in part (a) of Table 5.7. As expected from the model diagnostics of Example 5.4, the Ori-F test indicates no quadratic nonlinearity in the normalized innovations. However, the TAR-F test statistics suggest strong nonlinearity.

Based on the test results in Table 5.7, we entertain a threshold duration model with two regimes for the IBM intraday durations. The threshold variable is $x_{t-1}$ (i.e., lag-1 adjusted duration). The estimated threshold value is 3.79. The fitted threshold WACD(1, 1) model is $x_i = \psi_i \epsilon_i$, where

$$\psi_i = \begin{cases} 0.020 + 0.257x_{i-1} + 0.847\psi_{i-1}, & \epsilon_i \sim w(0.901) \quad \text{if } x_{i-1} \le 3.79 \\ 1.808 + 0.027x_{i-1} + 0.501\psi_{i-1}, & \epsilon_i \sim w(0.845) \quad \text{if } x_{i-1} > 3.79, \end{cases} \tag{5.45}$$

**Table 5.7. Nonlinearity Tests for IBM Transaction Durations from November 1 to November 7, 1990. Only Intraday Durations Are Used. The Number in the Parentheses of Tar-F Tests Denotes Time Delay.**

| | (a) Normalized innovations of a WACD(1,1) model | | | | |
|---|---|---|---|---|---|
| Type | Ori-F | Tar-F(1) | Tar-F(2) | Tar-F(3) | Tar-F(4) |
| Test | 0.343 | 3.288 | 3.142 | 3.128 | 0.297 |
| $p$ value | 0.969 | 0.006 | 0.008 | 0.008 | 0.915 |
| | (b) Normalized innovations of a threshold WACD(1,1) model | | | | |
| Type | Ori-F | Tar-F(1) | Tar-F(2) | Tar-F(3) | Tar-F(4) |
| Test | 0.163 | 0.746 | 1.899 | 1.752 | 0.270 |
| $p$ value | 0.998 | 0.589 | 0.091 | 0.119 | 0.929 |

where $w(\alpha)$ denotes a standardized Weibull distribution with parameter $\alpha$. The number of observations in the two regimes are 2503 and 1030, respectively. In Eq. (5.45), the standard errors of the parameters for the first regime are 0.043, 0.041, 0.024, and 0.014, whereas those for the second regime are 0.526, 0.020, 0.147, and 0.020, respectively.

Consider the normalized innovations $\hat{\epsilon}_i = x_i/\hat{\psi}_i$ of the threshold WACD(1, 1) model in Eq. (5.45). We obtain $Q(12) = 9.8$ and $Q(24) = 23.9$ for $\hat{\epsilon}_i$ and $Q(12) = 8.0$ and $Q(24) = 16.7$ for $\hat{\epsilon}_i^2$. Thus, there are no significant serial correlations in the $\hat{\epsilon}_i$ and $\hat{\epsilon}_i^2$ series. Furthermore, applying the same nonlinearity tests as before to this newly normalized innovational series $\hat{\epsilon}_i$, we detect no nonlinearity; see part (b) of Table 5.7. Consequently, the two-regime threshold WACD(1, 1) model in Eq. (5.45) is adequate.

If we classify the two regimes as heavy and thin trading periods, then the threshold model suggests that the trading dynamics measured by intraday transaction durations are different between heavy and thin trading periods for IBM stock even after the adjustment of diurnal pattern. This is not surprising as market activities are often driven by arrivals of news and other information.

The estimated threshold WACD(1, 1) model in Eq. (5.45) contains some insignificant parameters. We refine the model and obtain the result:

$$\psi_i = \begin{cases} 0.225x_{i-1} + 0.867\psi_{i-1}, & \epsilon_i \sim w(0.902) \quad \text{if } x_{i-1} \le 3.79 \\ 1.618 + 0.614\psi_{i-1}, & \epsilon_i \sim w(0.846) \quad \text{if } x_{i-1} > 3.79. \end{cases}$$

All of the estimates of the refined model are highly significant. The Ljung–Box statistics of the standardized innovations $\hat{\epsilon}_i = x_i/\hat{\psi}_i$ show $Q(10) = 5.91(0.82)$ and $Q(20) = 16.04(0.71)$ and those of $\hat{\epsilon}_i^2$ give $Q(10) = 5.35(0.87)$ and $Q(20) = 15.20(0.76)$, where the number in parentheses is the $p$ value. Therefore, the refined model is adequate. The RATS program used to estimate the prior model is given in Appendix C.

## 5.7 BIVARIATE MODELS FOR PRICE CHANGE AND DURATION

In this section, we introduce a model that considers jointly the process of price change and the associated duration. As mentioned before, many intraday transactions of a stock result in no price change. Those transactions are highly relevant to trading intensity, but they do not contain direct information on price movement. Therefore, to simplify the complexity involved in modeling price change, we focus on transactions that result in a price change and consider a price change and duration (PCD) model to describe the multivariate dynamics of price change and the associated time duration.

We continue to use the same notation as before, but the definition is changed to transactions with a price change. Let $t_i$ be the calendar time of the $i$th price change of an asset. As before, $t_i$ is measured in seconds from midnight of a trading day. Let $P_{t_i}$ be the transaction price when the $i$th price change occurred and $\Delta t_i = t_i - t_{i-1}$ be

the time duration between price changes. In addition, let $N_i$ be the number of trades in the time interval $(t_{i-1}, t_i)$ that result in no price change. This new variable is used to represent trading intensity during a period of no price change. Finally, let $D_i$ be the direction of the $i$th price change with $D_i = 1$ when price goes up and $D_i = -1$ when the price comes down, and let $S_i$ be the size of the $i$th price change measured in ticks. Under the new definitions, the price of a stock evolves over time by

$$P_{t_i} = P_{t_{i-1}} + D_i S_i, \tag{5.46}$$

and the transactions data consist of $\{\Delta t_i, N_i, D_i, S_i\}$ for the $i$th price change. The PCD model is concerned with the joint analysis of $(\Delta t_i, N_i, D_i, S_i)$.

***Remark:*** Focusing on transactions associated with a price change can reduce the sample size dramatically. For example, consider the intraday data of IBM stock from November 1, 1990 to January 31, 1991. There were 60,265 intraday trades, but only 19,022 of them resulted in a price change. In addition, there is no diurnal pattern in time durations between price changes.

To illustrate the relationship among the price movements of all transactions and those of transactions associated with a price change, we consider the intraday tradings of IBM stock on November 21, 1990. There were 726 transactions on that day during the normal trading hours, but only 195 trades resulted in a price change. Figure 5.14 shows the time plot of the price series for both cases. As expected, the price series are the same.

The PCD model decomposes the joint distribution of $(\Delta t_i, N_i, D_i, S_i)$ given $F_{i-1}$ as

$$\begin{aligned} &f(\Delta t_i, N_i, D_i, S_i \mid F_{i-1}) \\ &= f(S_i \mid D_i, N_i, \Delta t_i, F_{i-1}) f(D_i \mid N_i, \Delta t_i, F_{i-1}) f(N_i \mid \Delta t_i, F_{i-1}) f(\Delta t_i \mid F_{i-1}). \end{aligned} \tag{5.47}$$

This partition enables us to specify suitable econometric models for the conditional distributions and, hence, to simplify the modeling task. There are many ways to specify models for the conditional distributions. A proper specification might depend on the asset under study. Here we employ the specifications used by McCulloch and Tsay (2000), who use generalized linear models for the discrete-valued variables and a time series model for the continuous variable $\ln(\Delta t_i)$.

For the time duration between price changes, we use the model

$$\ln(\Delta t_i) = \beta_0 + \beta_1 \ln(\Delta t_{i-1}) + \beta_2 S_{i-1} + \sigma \epsilon_i, \tag{5.48}$$

where $\sigma$ is a positive number and $\{\epsilon_i\}$ is a sequence of iid $N(0, 1)$ random variables. This is a multiple linear regression model with lagged variables. Other explanatory variables can be added if necessary. The log transformation is used to ensure the positiveness of time duration.

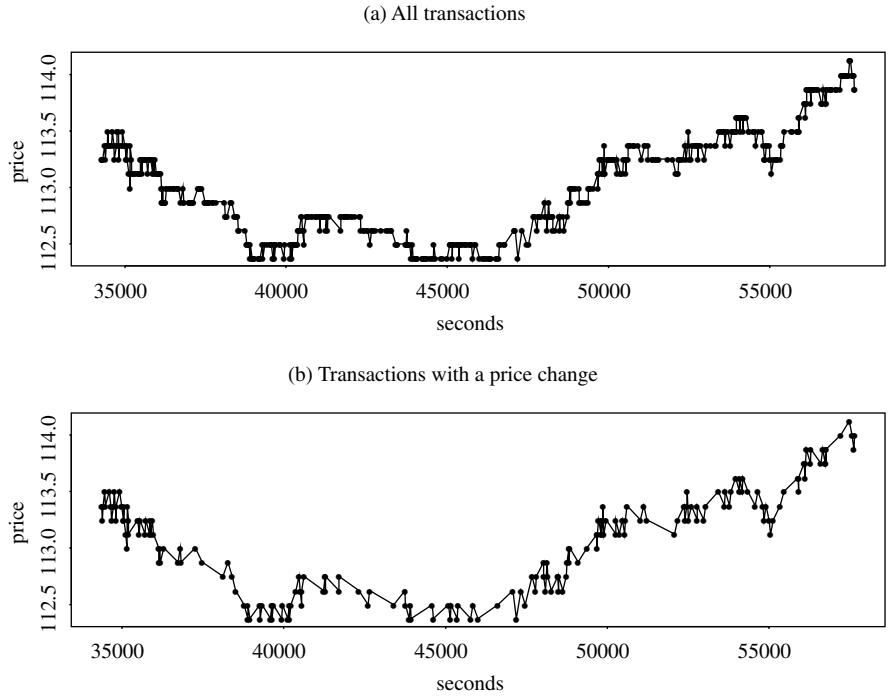


**Figure 5.14.** Time plots of the intraday transaction prices of IBM stock on November 21, 1990: (a) all transactions, and (b) transactions that resulted in a price change.

The conditional model for $N_i$ is further partitioned into two parts because empirical data suggest a concentration of $N_i$ at 0. The first part of the model for $N_i$ is the logit model

$$p(N_i = 0 \mid \Delta t_i, F_{i-1}) = \text{logit}[\alpha_0 + \alpha_1 \ln(\Delta t_i)], \tag{5.49}$$

where $\text{logit}(x) = \exp(x)/[1 + \exp(x)]$, whereas the second part of the model is

$$N_i \mid (N_i > 0, \Delta t_i, F_{i-1}) \sim 1 + g(\lambda_i), \quad \lambda_i = \frac{\exp[\gamma_0 + \gamma_1 \ln(\Delta t_i)]}{1 + \exp[\gamma_0 + \gamma_1 \ln(\Delta t_i)]}, \tag{5.50}$$

where $\sim$ means "is distributed as," and $g(\lambda)$ denotes a geometric distribution with parameter $\lambda$, which is in the interval (0, 1).

The model for direction $D_i$ is

$$D_i \mid (N_i, \Delta t_i, F_{i-1}) = \text{sign}(\mu_i + \sigma_i \epsilon), \tag{5.51}$$

where $\epsilon$ is a $N(0, 1)$ random variable, and

$$\mu_i = \omega_0 + \omega_1 D_{i-1} + \omega_2 \ln(\Delta t_i)$$
$$\ln(\sigma_i) = \beta \left| \sum_{j=1}^{4} D_{i-j} \right| = \beta | D_{i-1} + D_{i-2} + D_{i-3} + D_{i-4} |.$$

In other words, $D_i$ is governed by the sign of a normal random variable with mean $\mu_i$ and variance $\sigma_i^2$. A special characteristic of the prior model is the function for $\ln(\sigma_i)$. For intraday transactions, a key feature is the *price reversal* between consecutive price changes. This feature is modeled by the dependence of $D_i$ on $D_{i-1}$ in the mean equation with a negative $\omega_1$ parameter. However, there exists occasional local trend in the price movement. The previous variance equation allows for such a local trend by increasing the uncertainty in the direction of price movement when the past data showed evidence of a local trend. For a normal distribution with a fixed mean, increasing its variance makes a random draw have the same chance to be positive and negative. This in turn increases the chance for a sequence of all positive or all negative draws. Such a sequence produces a local trend in price movement.

To allow for different dynamics between positive and negative price movements, we use different models for the size of a price change. Specifically, we have

$$S_i \mid (D_i = -1, N_i, \Delta t_i, F_{i-1}) \sim p(\lambda_{d,i}) + 1, \quad \text{with} \tag{5.52}$$
$$\ln(\lambda_{d,i}) = \eta_{d,0} + \eta_{d,1} N_i + \eta_{d,2} \ln(\Delta t_i) + \eta_{d,3} S_{i-1}$$
$$S_i \mid (D_i = 1, N_i, \Delta t_i, F_{i-1}) \sim p(\lambda_{u,i}) + 1, \quad \text{with} \tag{5.53}$$
$$\ln(\lambda_{u,i}) = \eta_{u,0} + \eta_{u,1} N_i + \eta_{u,2} \ln(\Delta t_i) + \eta_{u,3} S_{i-1},$$

where $p(\lambda)$ denotes a Poisson distribution with parameter $\lambda$, and 1 is added to the size because the minimum size is 1 tick when there is a price change.

The specified models in Eqs. (5.48)–(5.53) can be estimated jointly by either the maximum likelihood method or the Markov Chain Monte Carlo methods. Based on Eq. (5.47), the models consist of six conditional models that can be estimated separately.

**Example 5.5.** Consider the intraday transactions of IBM stock on November 21, 1990. There are 194 price changes within the normal trading hours. Figure 5.15 shows the histograms of $\ln(\Delta t_i)$, $N_i$, $D_i$, and $S_i$. The data for $D_i$ are about equally distributed between "upward" and "downward" movements. Only a few transactions resulted in a price change of more than 1 tick; as a matter of fact, there were seven changes with two ticks and one change with three ticks. Using Markov Chain Monte Carlo (MCMC) methods (see Chapter 10), we obtained the following models for the data. The reported estimates and their standard deviations are the posterior means and standard deviations of MCMC draws with 9500 iterations. The model for the time duration between price changes is

$$\ln(\Delta t_i) = 4.023 + 0.032 \ln(\Delta t_{i-1}) - 0.025 S_{i-1} + 1.403 \epsilon_i,$$

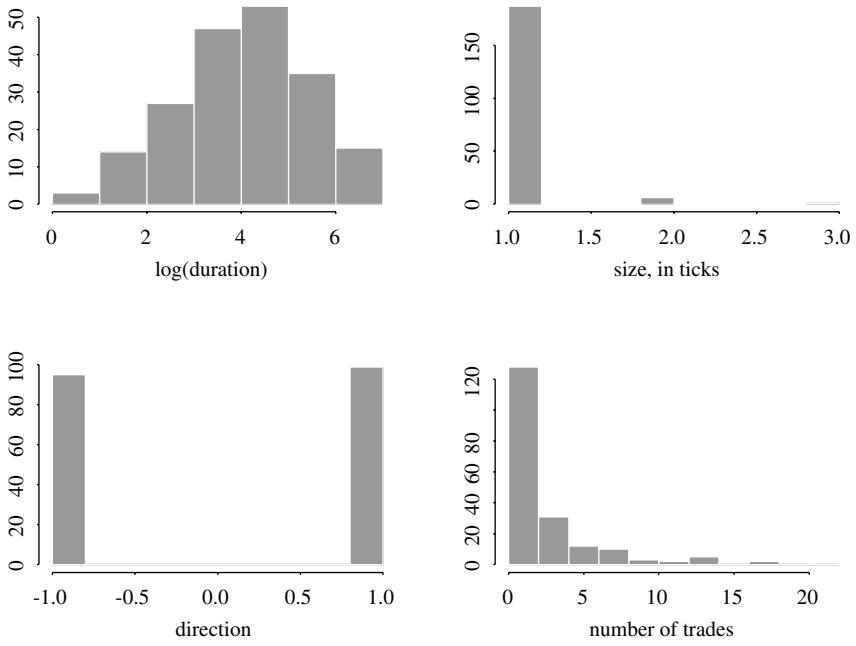


**Figure 5.15.** Histograms of intraday transactions data for IBM stock on November 21, 1990: (a) log durations between price changes, (b) direction of price movement, (c) size of price change measured in ticks, and (d) number of trades without a price change.

where standard deviations of the coefficients are 0.415, 0.073, 0.384, and 0.073, respectively. The fitted model indicates that there was no dynamic dependence in the time duration. For the $N_i$ variable, we have

$$Pr(N_i > 0 \mid \Delta t_i, F_{i-1}) = \text{logit}[-0.637 + 1.740 \ln(\Delta t_i)],$$

where standard deviations of the estimates are 0.238 and 0.248, respectively. Thus, as expected, the number of trades with no price change in the time interval $(t_{i-1}, t_i)$ depends positively on the length of the interval. The magnitude of $N_i$ when it is positive is

$$N_i \mid (N_i > 0, \Delta t_i, F_{i-1}) \sim 1 + g(\lambda_i), \quad \lambda_i = \frac{\exp[0.178 - 0.910 \ln(\Delta t_i)]}{1 + \exp[0.178 - 0.910 \ln(\Delta t_i)]},$$

where standard deviations of the estimates are 0.246 and 0.138, respectively. The negative and significant coefficient of $\ln(\Delta t_i)$ means that $N_i$ is positively related to the length of the duration $\Delta t_i$ because a large $\ln(\Delta t_i)$ implies a small $\lambda_i$, which in turn implies higher probabilities for larger $N_i$; see the geometric distribution in Eq. (5.27).

The fitted model for $D_i$ is

$$\mu_i = 0.049 - 0.840 D_{i-1} - 0.004 \ln(\Delta t_i)$$
$$\ln(\sigma_i) = 0.244 | D_{i-1} + D_{i-2} + D_{i-3} + D_{i-4} |,$$

where standard deviations of the parameters in the mean equation are 0.129, 0.132, and 0.082, respectively, whereas that for the parameter in the variance equation is 0.182. The price reversal is clearly shown by the highly significant negative coefficient of $D_{i-1}$. The marginally significant parameter in the variance equation is exactly as expected. Finally, the fitted models for the size of a price change are

$$\ln(\lambda_{d,i}) = 1.024 - 0.327 N_i + 0.412 \ln(\Delta t_i) - 4.474 S_{i-1}$$
$$\ln(\lambda_{u,i}) = -3.683 - 1.542 N_i + 0.419 \ln(\Delta t_i) + 0.921 S_{i-1},$$

where standard deviations of the parameters for the "down size" are 3.350, 0.319, 0.599, and 3.188, respectively, whereas those for the "up size" are 1.734, 0.976, 0.453, and 1.459. The interesting estimates of the prior two equations are the negative estimates of the coefficient of $N_i$. A large $N_i$ means there were more transactions in the time interval $(t_{i-1}, t_i)$ with no price change. This can be taken as evidence of no new information available in the time interval $(t_{i-1}, t_i)$. Consequently, the size for the price change at $t_i$ should be small. A small $\lambda_{u,i}$ or $\lambda_{d,i}$ for a Poisson distribution gives precisely that.

In summary, granted that a sample of 194 observations in a given day may not contain sufficient information about the trading dynamic of IBM stock, but the fitted models appear to provide some sensible results. McCulloch and Tsay (2000) extend the PCD model to a hierarchical framework to handle all the data of the 63 trading days between November 1, 1990 and January 31, 1991. Many of the parameter estimates become significant in this extended sample, which has more than 19,000 observations. For example, the overall estimate of the coefficient of $\ln(\Delta t_{i-1})$ in the model for time duration ranges from 0.04 to 0.1, which is small, but significant.

Finally, using transactions data to test microstructure theory often requires a careful specification of the variables used. It also requires a deep understanding of the way by which the market operates and the data are collected. However, ideas of the econometric models discussed in this chapter are useful and widely applicable in analysis of high-frequency data.

## APPENDIX A. REVIEW OF SOME PROBABILITY DISTRIBUTIONS

### *Exponential distribution*

A random variable $X$ has an exponential distribution with parameter $\beta > 0$ if its probability density function (pdf) is given by

$$f(x \mid \beta) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Denoting such a distribution by $X \sim \exp(\beta)$, we have $E(X) = \beta$ and $\text{Var}(X) = \beta^2$. The cumulative distribution function (CDF) of $X$ is

$$F(x \mid \beta) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-x/\beta} & \text{if } x \geq 0. \end{cases}$$

When $\beta = 1$, $X$ is said to have a standard exponential distribution.

***Gamma function***

For $\kappa > 0$, the gamma function $\Gamma(\kappa)$ is defined by

$$\Gamma(\kappa) = \int_0^\infty x^{\kappa-1} e^{-x} dx.$$

The most important properties of the gamma function are:

1. For any $\kappa > 1$, $\Gamma(\kappa) = (\kappa - 1)\Gamma(\kappa - 1)$.
2. For any positive integer $m$, $\Gamma(m) = (m - 1)!$.
3. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

The integration

$$\Gamma(y \mid \kappa) = \int_0^y x^{\kappa-1} e^{-x} dx, \quad y > 0$$

is an *incomplete* gamma function. Its values have been tabulated in the literature. Computer programs are now available to evaluate the incomplete gamma function.

***Gamma distribution***

A random variable $X$ has a Gamma distribution with parameter $\kappa$ and $\beta$ ($\kappa > 0$, $\beta > 0$) if its pdf is given by

$$f(x \mid \kappa, \beta) = \begin{cases} \frac{1}{\beta^\kappa \Gamma(\kappa)} x^{\kappa-1} e^{-x/\beta} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

By changing variable $y = x/\beta$, one can easily obtain the moments of $X$:

$$E(X^m) = \int_0^\infty x^m f(x \mid \kappa, \beta) dx = \frac{1}{\beta^\kappa \Gamma(\kappa)} \int_0^\infty x^{\kappa+m-1} e^{-x/\beta} dx$$

$$= \frac{\beta^m}{\Gamma(\kappa)} \int_0^\infty y^{\kappa+m-1} e^{-y} dy = \frac{\beta^m \Gamma(\kappa+m)}{\Gamma(\kappa)}.$$

In particular, the mean and variance of $X$ are $E(X) = \kappa\beta$ and $\text{Var}(X) = \kappa\beta^2$. When $\beta = 1$, the distribution is called a standard Gamma distribution with parameter $\kappa$. We use the notation $G \sim \text{Gamma}(\kappa)$ to denote that $G$ follows a standard Gamma distribution with parameter $\kappa$. The moments of $G$ are

$$E(G^m) = \frac{\Gamma(\kappa+m)}{\Gamma(\kappa)}, \quad m > 0. \tag{5.54}$$

***Weibull distribution***

A random variable $X$ has a Weibull distribution with parameters $\alpha$ and $\beta$ ($\alpha > 0$, $\beta > 0$) if its pdf is given by

$$f(x \mid \alpha, \beta) = \begin{cases} \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0, \end{cases}$$

where $\beta$ and $\alpha$ are the scale and shape parameters of the distribution. The mean and variance of $X$ are

$$E(X) = \beta\Gamma\left(1 + \frac{1}{\alpha}\right), \quad \text{Var}(X) = \beta^2 \left\{\Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^2\right\}$$

and the CDF of $X$ is

$$F(x \mid \alpha, \beta) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-(x/\beta)^\alpha} & \text{if } x \geq 0. \end{cases}$$

When $\alpha = 1$, the Weibull distribution reduces to an exponential distribution.

Define $Y = X/[\beta\Gamma(1 + \frac{1}{\alpha})]$. We have $E(Y) = 1$ and the pdf of $Y$ is

$$f(y \mid \alpha) = \begin{cases} \alpha \left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^\alpha y^{\alpha-1} \exp\left\{-\left[\Gamma\left(1 + \frac{1}{\alpha}\right) y\right]^\alpha\right\} & \text{if } y \geq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5.55}$$

where the scale parameter $\beta$ disappears due to standardization. The CDF of the standardized Weibull distribution is

$$F(y \mid \alpha) = \begin{cases} 0 & \text{if } y < 0 \\ 1 - \exp\left\{-\left[\Gamma\left(1 + \frac{1}{\alpha}\right) y\right]^\alpha\right\} & \text{if } y > 0, \end{cases}$$

and we have $E(Y) = 1$ and $\text{Var}(Y) = \Gamma(1 + \frac{2}{\alpha})/[\Gamma(1 + \frac{1}{\alpha})]^2 - 1$. For a duration model with Weibull innovations, the prior pdf is used in the maximum likelihood estimation.

***Generalized Gamma distribution***

A random variable $X$ has a generalized Gamma distribution with parameter $\alpha, \beta, \kappa$ ($\alpha > 0$, $\beta > 0$, and $\kappa > 0$) if its pdf is given by

$$f(x \mid \alpha, \beta, \kappa) = \begin{cases} \dfrac{\alpha x^{\kappa\alpha-1}}{\beta^{\kappa\alpha}\Gamma(\kappa)} \exp\left[-\left(\dfrac{x}{\beta}\right)^{\alpha}\right] & \text{if } x \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\beta$ is a scale parameter, and $\alpha$ and $\kappa$ are shape parameters. This distribution can be written as

$$G = \left(\frac{X}{\beta}\right)^{\alpha},$$

where $G$ is a standard Gamma random variable with parameter $\kappa$. The pdf of $X$ can be obtained from that of $G$ by the technique of changing variables. Similarly, the moments of $X$ can be obtained from that of $G$ in Eq. (5.54) by

$$E(X^m) = E[(\beta G^{1/\alpha})^m] = \beta^m E(G^{m/\alpha}) = \beta^m \frac{\Gamma(\kappa + \frac{m}{\alpha})}{\Gamma(\kappa)} = \frac{\beta^m \Gamma(\kappa + \frac{m}{\alpha})}{\Gamma(\kappa)}.$$

When $\kappa = 1$, the generalized Gamma distribution reduces to that of a Weibull distribution. Thus, the exponential and Weibull distributions are special cases of the generalized Gamma distribution.

The expectation of a generalized Gamma distribution is $E(X) = \beta\Gamma(\kappa + \frac{1}{\alpha})/\Gamma(\kappa)$. In duration models, we need a distribution with unit expectation. Therefore, defining a random variable $Y = \lambda X/\beta$, where $\lambda = \Gamma(\kappa)/\Gamma(\kappa + \frac{1}{\alpha})$, we have $E(Y) = 1$ and the pdf of $Y$ is

$$f(y \mid \alpha, \kappa) = \begin{cases} \dfrac{\alpha y^{\kappa\alpha-1}}{\lambda^{\kappa\alpha}\Gamma(\kappa)} \exp\left[-\left(\dfrac{y}{\lambda}\right)^{\alpha}\right] & \text{if } y > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{5.56}$$

where again the scale parameter $\beta$ disappears and $\lambda = \Gamma(\kappa)/\Gamma(\kappa + \frac{1}{\alpha})$.

## APPENDIX B. HAZARD FUNCTION

A useful concept in modeling duration is the *Hazard function* implied by a distribution function. For a random variable $X$, the *survival function* is defined as

$$S(x) \equiv P(X > x) = 1 - P(X \leq x) = 1 - \text{CDF}(x), \quad x > 0,$$

which gives the probability that a subject, which follows the distribution of $X$, survives at the time $x$. The hazard function (or intensity function) of $X$ is then defined

by

$$h(x) = \frac{f(x)}{S(x)} \tag{5.57}$$

where $f(.)$ and $S(.)$ are the pdf and survival function of $X$, respectively.

**Example 5.6.** For the Weibull distribution with parameters $\alpha$ and $\beta$, the survival function and hazard function are:

$$S(x \mid \alpha, \beta) = \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right], \quad h(x \mid \alpha, \beta) = \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1}, \quad x > 0.$$

In particular, when $\alpha = 1$, we have $h(x \mid \beta) = 1/\beta$. Therefore, for an exponential distribution, the hazard function is constant. For a Weibull distribution, the hazard is a monotone function. If $\alpha > 1$, then the hazard function is monotonously increasing. If $\alpha < 1$, the hazard function is monotonously decreasing. For the generalized Gamma distribution, the survival function and hence, the hazard function involve the incomplete Gamma function. Yet the hazard function may exhibit various patterns, including U shape or inverted U shape. Thus, the generalized Gamma distribution provides a flexible approach to modeling the duration of stock transactions.

For the standardized Weibull distribution, the survival and hazard functions are

$$S(y \mid \alpha) = \exp\left\{-\left[\Gamma\left(1 + \frac{1}{\alpha}\right) y\right]^{\alpha}\right\},$$

$$h(y \mid \alpha) = \alpha\left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^{\alpha} y^{\alpha-1}, \quad y > 0.$$

## APPENDIX C. SOME RATS PROGRAMS FOR DURATION MODELS

The data used are adjusted time durations of intraday transactions of IBM stock from November 1 to November 9, 1990. The file name is "ibm1to5.dat" and it has 3534 observations.

### A. Program for Estimating a WACD(1, 1) Model

```
all 0 3534:1
open data ibm1to5.dat
data(org=obs) / x r1
set psi = 1.0
nonlin a0 a1 b1 al
frml gvar = a0+a1*x(t-1)+b1*psi(t-1)
frml gma = %LNGAMMA(1.0+1.0/al)
frml gln =al*gma(t)+log(al)-log(x(t)) $
 +al*log(x(t)/(psi(t)=gvar(t)))-(exp(gma(t))*x(t)/psi(t))**al
```

```
smpl 2 3534
compute a0 = 0.2, a1 = 0.1, b1 = 0.1, al = 0.8
maximize(method=bhhh,recursive,iterations=150) gln
set fv = gvar(t)
set resid = x(t)/fv(t)
set residsq = resid(t)*resid(t)
cor(qstats,number=20,span=10) resid
cor(qstats,number=20,span=10) residsq
```

## B. Program for Estimating a GACD(1, 1) Models

```
all 0 3534:1
open data ibm1to5.dat
data(org=obs) / x r1
set psi = 1.0
nonlin a0 a1 b1 al ka
frml cv = a0+a1*x(t-1)+b1*psi(t-1)
frml gma = %LNGAMMA(ka)
frml lam = exp(gma(t))/exp(%LNGAMMA(ka+(1.0/al)))
frml xlam = x(t)/(lam(t)*(psi(t)=cv(t)))
frml gln =-gma(t)+log(al/x(t))+ka*al*log(xlam(t))-(xlam(t))**al
smpl 2 3534
compute a0 = 0.238, a1 = 0.075, b1 = 0.857, al = 0.5, ka = 4.0
nlpar(criterion=value,cvcrit=0.00001)
maximize(method=bhhh,recursive,iterations=150) gln
set fv = cv(t)
set resid = x(t)/fv(t)
set residsq = resid(t)*resid(t)
cor(qstats,number=20,span=10) resid
cor(qstats,number=20,span=10) residsq
```

## C. A program for estimating a Tar-WACD(1, 1) model. The threshold 3.79 is prespecified.

```
all 0 3534:1
open data ibm1to5.dat
data(org=obs) / x rt
set psi = 1.0
nonlin a1 a2 al b0 b2 bl
frml u = ((x(t-1)-3.79)/abs(x(t-1)-3.79)+1.0)/2.0
frml cp1 = a1*x(t-1)+a2*psi(t-1)
frml gma1 = %LNGAMMA(1.0+1.0/al)
frml cp2 = b0+b2*psi(t-1)
frml gma2 = %LNGAMMA(1.0+1.0/bl)
frml cp = cp1(t)*(1-u(t))+cp2(t)*u(t)
frml gln1 =al*gma1(t)+log(al)-log(x(t)) $
 +al*log(x(t)/(psi(t)=cp(t)))-(exp(gma1(t))*x(t)/psi(t))**al
frml gln2 =bl*gma2(t)+log(bl)-log(x(t)) $
 +bl*log(x(t)/(psi(t)=cp(t)))-(exp(gma2(t))*x(t)/psi(t))**bl
frml gln = gln1(t)*(1-u(t))+gln2(t)*u(t)
smpl 2 3534
compute a1 = 0.2, a2 = 0.85, al = 0.9
```

```
compute b0 = 1.8, b2 = 0.5, bl = 0.8
maximize(method=bhhh,recursive,iterations=150) gln
set fv = cp(t)
set resid = x(t)/fv(t)
set residsq = resid(t)*resid(t)
cor(qstats,number=20,span=10) resid
cor(qstats,number=20,span=10) residsq
```

## EXERCISES

**1.** Let $r_t$ be the log return of an asset at time $t$. Assume that $\{r_t\}$ is a Gaussian white noise series with mean 0.05 and variance 1.5. Suppose that the probability of a trade at each time point is 40% and is independent of $r_t$. Denote the observed return by $r_t^o$. Is $r_t^o$ serially correlated? If yes, calculate the first three lags of autocorrelations of $r_t^o$.

**2.** Let $P_t$ be the observed market price of an asset, which is related to the fundamental value of the asset $P_t^*$ via Eq. (5.9). Assume that $\Delta P_t^* = P_t^* - P_{t-1}^*$ forms a Gaussian white noise series with mean zero and variance 1.0. Suppose that the bid-ask spread is two ticks. What is the lag-1 autocorrelation of the price change series $\Delta P_t = P_t - P_{t-1}$ when the tick size is \$1/8? What is the lag-1 autocorrelation of the price change when the tick size is \$1/16?

**3.** The file "ibm-d2-dur.dat" contains the adjusted durations between trades of IBM stock on November 2, 1990. The file has three columns consisting of day, time of trade measured in seconds from midnight, and adjusted durations.

**(a)** Build an EACD model for the adjusted duration and check the fitted model.

**(b)** Build a WACD model for the adjusted duration and check the fitted model.

**(c)** Build a GACD model for the adjusted duration and check the fitted model.

**(d)** Compare the prior three duration models.

**4.** The file "mmm9912-dtp.dat" contains the transactions data of the stock of 3M Company in December 1999. There are three columns: day of the month, time of transaction in seconds from midnight, and transaction price. Transactions that occurred after 4:00 pm Eastern time are excluded.

**(a)** Is there a diurnal pattern in 3M stock trading? You may construct a time series $n_t$, which denotes the number of trades in 5-minute time interval to answer this question.

**(b)** Use the price series to confirm the existence of bid-ask bounce in intraday trading of 3M stock.

**(c)** Tabulate the frequencies of price change in multiples of tick size \$1/16. You may combine changes with 5 ticks or more into a category and those with $-5$ ticks or beyond into another category.

**5.** Consider again the transactions data of 3M stock in December 1999.

**(a)** Use the data to construct an intraday 5-minute log return series. Use the simple average of all transaction prices within a 5-minute interval as the stock price for the interval. Is the series serially correlated? You may use Ljung–Box statistics to test the hypothesis with the first 10 lags of sample autocorrelation function.

**(b)** There are seventy-seven 5-minute returns in a normal trading day. Some researchers suggest that the sum of squares of the intraday 5-minute returns can be used as a measure of daily volatility. Apply this approach and calculate the daily volatility of the log return of 3M stock in December 1999. Discuss the validity of such a procedure to estimate daily volatility.

**6.** The file "mmm9912-adur.dat" contains an adjusted intraday trading duration of 3M stock in December 1999. There are thirty-nine 10-minute time intervals in a trading day. Let $d_i$ be the average of all log durations for the $i$th 10-minute interval across all trading days in December 1999. Define an adjusted duration as $t_j / \exp(d_i)$, where $j$ is in the $i$th 10-minute interval. Note that more sophisticated methods can be used to adjust the diurnal pattern of trading duration. Here we simply use a local average.

**(a)** Is there a diurnal pattern in the adjusted duration series? Why?

**(b)** Build a duration model for the adjusted series using exponential innovations. Check the fitted model.

**(c)** Build a duration model for the adjusted series using Weibull innovations. Check the fitted model.

**(d)** Build a duration model for the adjusted series using generalized Gamma innovations. Check the fitted model.

**(e)** Compare and comment on the three duration models built before.

## REFERENCES

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997), *The Econometrics of Financial Markets*, Princeton University Press: New Jersey.

Cho, D., Russell, J. R., Tiao, G. C., and Tsay, R. S. (2000), "The magnet effect of price limits: Evidence from high frequency data on Taiwan stock exchange," Working paper, Graduate School of Business, University of Chicago.

Engle, R. F., and Russell, J. R. (1998), "Autoregressive conditional duration: A new model for irregularly spaced transaction data," *Econometrica*, 66, 1127–1162.

Ghysels, E. (2000), "Some econometric recipes for high-frequency data cooking," *Journal of Business and Economic Statistics*, 18, 154–163.

Hasbrouck, J. (1992), *Using the TORQ database*, Stern School of Business, New York University.

Hasbrouck, J. (1999), "The dynamics of discrete bid and ask quotes," *Journal of Finance*, 54, 2109–2142.

Hauseman, J., Lo, A., and MacKinlay, C. (1992), "An ordered probit analysis of transaction stock prices," *Journal of Financial Economics*, 31, 319–379.

Lo, A., and MacKinlay, A. C. (1990), "An econometric analysis of nonsynchronous trading," *Journal of Econometrics*, 45, 181–212.

McCulloch, R. E., and Tsay, R. S. (2000), "Nonlinearity in high frequency data and hierarchical models," Working paper, Graduate School of Business, University of Chicago.

Roll, R. (1984), "A simple implicit measure of the effective bid-ask spread in an efficient market," *Journal of Finance*, 39, 1127–1140.

Rydberg, T. H., and Shephard, N. (1998), "Dynamics of trade-by-trade price movements: decomposition and models," Working paper, Nuffield College, Oxford University.

Stoll, H., and Whaley, R. (1990), "Stock market structure and volatility," *Review of Financial Studies*, 3, 37–71.

Wood, R. A. (2000), "Market microstructure research databases: History and projections," *Journal of Business & Economic Statistics*, 18, 140–145.

Zhang, M. Y., Russell, J. R., and Tsay, R. S. (2001), "A nonlinear autoregressive conditional duration model with applications to financial transaction data," *Journal of Econometrics* (to appear).

Zhang, M. Y., Russell, J. R., and Tsay, R. S. (2001b), "Determinants of bid and ask quotes and implications for the cost of trading," Working paper, Graduate School of Business, University of Chicago.

*Analysis of Financial Time Series.* Ruey S. Tsay

ISBN: 0-471-41544-8

CHAPTER 6

# Continuous-Time Models and Their Applications

Price of a financial asset evolves over time and forms a *stochastic process*, which is a statistical term used to describe the evolution of a random variable over time. The observed prices are a realization of the underlying stochastic process. The theory of stochastic process is the basis on which the observed prices are analyzed and statistical inference is made.

There are two types of stochastic process for modeling the price of an asset. The first type is called the *discrete-time stochastic process*, in which the price changes at discrete time points. All the processes discussed in the previous chapters belong to this category. For example, the daily closing price of IBM stock on the New York Stock Exchange forms a discrete-time stochastic process. Here the price changes only at the closing of a trading day. Price movements within a trading day are not necessarily relevant to the observed daily price. The second type of stochastic process is the *continuous-time process*, in which the price changes continuously, even though the price is only observed at discrete time points. One can think of the price as the "true value" of the stock that always exists and is time varying.

For both types of process, the price can be continuous or discrete. A continuous price can assume any positive real number, whereas a discrete price can only assume a countable number of possible values. Assume that the price of an asset is a continuous-time stochastic process. If the price is a continuous random variable, then we have a continuous-time continuous process. If the price itself is discrete, then we have a continuous-time discrete process. Similar classifications apply to discrete-time processes. The series of price change in Chapter 5 is an example of discrete-time discrete process.

In this chapter, we treat the price of an asset as a continuous-time continuous stochastic process. Our goal is to introduce the statistical theory and tools needed to model financial assets and to price options. We begin the chapter with some terminologies of stock options used in the chapter. In Section 6.2, we provide a brief introduction of Brownian motion, which is also known as a Wiener process. We then discuss some diffusion equations and stochastic calculus, including the well-known Ito's lemma. Most option pricing formulas are derived under the assumption that the

price of an asset follows a diffusion equation. We use the Black–Scholes formula to demonstrate the derivation. Finally, to handle the price variations caused by rare events (e.g., a profit warning), we also study some simple diffusion models with jumps.

If the price of an asset follows a diffusion equation, then the price of an option contingent to the asset can be derived by using hedging methods. However, with jumps the market becomes incomplete and there is no perfect hedging of options. The price of an option is then valued either by using diversifiability of jump risk or defining a notion of risk and choosing a price and a hedge that minimize this risk. For basic applications of stochastic processes in derivative pricing, see Cox and Rubinstein (1985) and Hull (1997).

## 6.1 OPTIONS

A stock option is a financial contract that gives the holder the right to trade a certain number of shares of a specified common stock by a certain date for a specified price. There are two types of options. A *call option* gives the holder the right to buy the underlying stock; see Chapter 3 for a formal definition. A *put option* gives the holder the right to sell the underlying stock. The specified price in the contract is called the *strike price* or *exercise price*. The date in the contract is known as the *expiration date* or *maturity*. *American options* can be exercised at any time up to the expiration date. *European options* can be exercised only on the expiration date.

The value of a stock option depends on the value of the underlying stock. Let $K$ be the strike price and $P$ be the stock price. A call option is *in-the-money* when $P > K$, *at-the-money* when $P = K$, and *out-of-the-money* when $P < K$. A put option is in-the-money when $P < K$, at-the-money when $P = K$, and out-of-the-money when $P > K$. In general, an option is in-the-money when it would lead to a positive cash flow to the holder if it were exercised immediately. An option is out-of-the-money when it would lead to a negative cash flow to the holder if it were exercised immediately. Finally, an option is at-the-money when it would lead to zero cash flow if it were exercised immediately. Obviously, only in-the-money options are exercised in practice. For more description on options, see Hull (1997).

## 6.2 SOME CONTINUOUS-TIME STOCHASTIC PROCESSES

In mathematical statistics, a continuous-time continuous stochastic process is defined on a probability space $(\Omega, F, \mathbf{P})$, where $\Omega$ is a nonempty space, $F$ is a $\sigma$-field consisting of subsets of $\Omega$, and $\mathbf{P}$ is a probability measure; see Chapter 1 of Billingsley (1986). The process can be written as $\{x(\eta, t)\}$, where $t$ denotes time and is continuous in $[0, \infty)$. For a given $t$, $x(\eta, t)$ is a real-valued continuous random variable (i.e., a mapping from $\Omega$ to the real line), and $\eta$ is an element of $\Omega$. For the price of an asset at time $t$, the range of $x(\eta, t)$ is the set of non-negative real numbers. For a given $\eta$, $\{x(\eta, t)\}$ is a time series with values depending on the time $t$. For simplicity, we

write a continuous-time stochastic process as $\{x_t\}$ with the understanding that, for a given $t$, $x_t$ is a random variable. In the literature, some authors use $x(t)$ instead of $x_t$ to emphasize that $t$ is continuous. However, we use the same notation $x_t$, but call it a continuous-time stochastic process.

### 6.2.1 The Wiener Process

In a discrete-time econometric model, we assume that the shocks form a white noise process, which is not predictable. What is the counterpart of shocks in a continuous-time model? The answer is the increments of a *Wiener process*, which is also known as a *standard Brownian motion*. There are many ways to define a Wiener process $\{w_t\}$. We use a simple approach that focuses on the small change $\Delta w_t = w_{t+\Delta t} - w_t$ associated with a small increment $\Delta t$ in time. A continuous-time stochastic process $\{w_t\}$ is a Wiener process if it satisfies

1. $\Delta w_t = \epsilon\sqrt{\Delta t}$, where $\epsilon$ is a standard normal random variable, and
2. $\Delta w_t$ is independent of $w_j$ for all $j \leq t$.

The second condition is a Markov property saying that conditional on the present value $w_t$, any past information of the process, $w_j$ with $j < t$, is irrelevant to the future $w_{t+\ell}$ with $\ell > 0$. From this property, it is easily seen that for any two nonoverlapping time intervals $\Delta_1$ and $\Delta_2$, the increments $w_{t_1+\Delta_1} - w_{t_1}$ and $w_{t_2+\Delta_2} - w_{t_2}$ are independent. In finance, this Markov property is related to a weak form of efficient market.

From the first condition, $\Delta w_t$ is normally distributed with mean zero and variance $\Delta t$. That is, $\Delta w_t \sim N(0,\ \Delta t)$, where $\sim$ denotes probability distribution. Consider next the process $w_t$. We assume that the process starts at $t = 0$ with initial value $w_0$, which is fixed and often set to zero. Then $w_t - w_0$ can be treated as a sum of many small increments. More specifically, define $T = \frac{t}{\Delta t}$, where $\Delta t$ is a small positive increment. Then

$$w_t - w_0 = w_{T\,\Delta t} - w_0 = \sum_{i=1}^{T} \Delta w_i = \sum_{i=1}^{T} \epsilon_i \sqrt{\Delta t},$$

where $\Delta w_i = w_{i\,\Delta t} - w_{(i-1)\,\Delta t}$. Because $\epsilon_i$'s are independent, we have

$$E(w_t - w_0) = 0, \quad \text{Var}(w_t - w_0) = \sum_{i=1}^{T} \Delta t = T\,\Delta t = t.$$

Thus, the increment in $w_t$ from time 0 to time $t$ is normally distributed with mean zero and variance $t$. To put it formally, for a Wiener process $w_t$, we have that $w_t - w_0 \sim N(0, t)$. This says that the variance of a Wiener process increases linearly with the length of time interval.

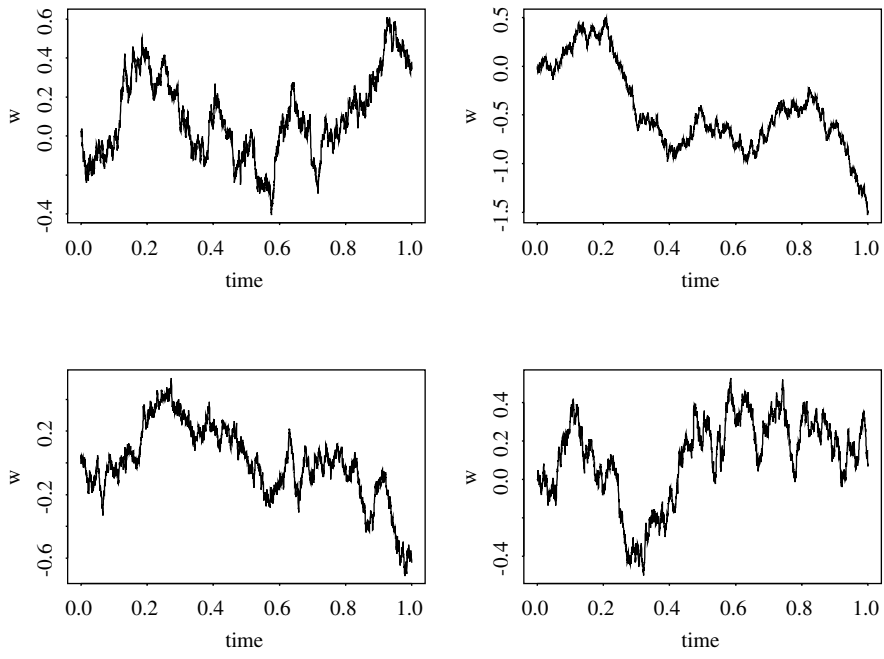


**Figure 6.1.** Four simulated Wiener processes.

Figure 6.1 shows four simulated Wiener processes on the unit time interval [0, 1]. They are obtained by using a simple version of the Donsker's Theorem in the statistical literature with $n = 3000$; see Donsker (1951) or Billingsley (1968).

***Donsker's Theorem***
Assume that $\{z_i\}_{i=1}^n$ is a sequence of independent standard normal random variates. For any $t \in [0, 1]$, let $[nt]$ be the integer part of $nt$. Define $w_{n,t} = \frac{1}{\sqrt{n}} \sum_{i=1}^{[nt]} z_i$. Then $w_{n,t}$ converges in distribution to a Wiener process $w_t$ on $[0, 1]$ as $n$ goes to infinity.

The four plots start with $w_0 = 0$, but drift apart as time increases, illustrating that the variance of a Wiener process increases with time. A simple time transformation from $[0, 1)$ to $[0, \infty)$ can be used to obtain simulated Wiener processes for $t \in [0, \infty)$.

***Remark:*** A formal definition of a Brownian motion $w_t$ on a probability space $(\Omega, F, \mathbf{P})$ is that it is a real-valued, continuous stochastic process for $t \geq 0$ with independent and stationary increments. In other words, $w_t$ satisfies

1. continuity: the map from $t$ to $w_t$ is continuous almost surely with respect to the probability measure $\mathbf{P}$;
2. independent increments: if $s \leq t$, $w_t - w_s$ is independent of $w_v$ for all $v \leq s$;

**3.** stationary increments: if $s \leq t$, $w_t - w_s$ and $w_{t-s} - w_0$ have the same probability distribution.

It can be shown that the probability distribution of the increment $w_t - w_s$ is normal with mean $\mu(t-s)$ and variance $\sigma^2(t-s)$. Furthermore, for any given time indexes $0 \leq t_1 < t_2 < \cdots < t_k$, the random vector $(w_{t_1}, w_{t_2}, \ldots, w_{t_k})$ follows a multivariate normal distribution. Finally, a Brownian motion is *standard* if $w_0 = 0$ almost surely, $\mu = 0$, and $\sigma^2 = 1$.

***Remark:*** An important property of Brownian motions is that their paths are not differentiable almost surely. In other words, for a standard Brownian motion $w_t$, it can be shown that $dw_t/dt$ does not exist for all elements of $\Omega$ except for elements in a subset $\Omega_1 \subset \Omega$ such that $P(\Omega_1) = 0$. As a result, we cannot use the usual intergation in calculus to handle integrals involving a standard Brownian motion when we consider the value of an asset over time. Another approach must be sought. This is the purpose of discussing Ito's calculus in the next section.

### 6.2.2 Generalized Wiener Processes

The Wiener process is a special stochastic process with zero drift and variance proportional to the length of time interval. This means that the rate of change in expectation is zero and the rate of change in variance is 1. In practice, the mean and variance of a stochastic process can evolve over time in a more complicated manner. Hence, further generalization of stochastic process is needed. To this end, we consider the *generalized Wiener process* in which the expectation has a drift rate $\mu$ and the rate of variance change is $\sigma^2$. Denote such a process by $x_t$ and use the notation $dy$ for a small change in the variable $y$. Then the model for $x_t$ is

$$dx_t = \mu\, dt + \sigma\, dw_t, \tag{6.1}$$

where $w_t$ is a Wiener process. If we consider a discretized version of Eq. (6.1), then

$$x_t - x_0 = \mu t + \sigma\epsilon\sqrt{t}$$

for increment from 0 to $t$. Consequently,

$$E(x_t - x_0) = \mu t, \quad \text{Var}(x_t - x_0) = \sigma^2 t.$$

The results say that the increment in $x_t$ has a growth rate of $\mu$ for the expectation and a growth rate of $\sigma^2$ for the variance. In the literature, $\mu$ and $\sigma$ of Eq. (6.1) are referred to as the drift and volatility parameters of the generalized Wiener process $x_t$.

### 6.2.3 Ito's Processes

The drift and volatility parameters of a generalized Wiener process are time-invariant. If one further extends the model by allowing $\mu$ and $\sigma$ to be functions

of the stochastic process $x_t$, then we have an Ito's process. Specifically, a process $x_t$ is an Ito's process if it satisfies

$$dx_t = \mu(x_t, t)\,dt + \sigma(x_t, t)\,dw_t, \tag{6.2}$$

where $w_t$ is a Wiener process. This process plays an important role in mathematical finance and can be written as

$$x_t = x_0 + \int_0^t \mu(x_s, s)\,ds + \int_0^t \sigma(x_s, s)\,dw_s,$$

where $x_0$ denotes the starting value of the process at time 0 and the last term on the right-hand side is a stochastic integral. Equation (6.2) is referred to as a stochastic diffusion equation with $\mu(x_t, t)$ and $\sigma(x_t, t)$ being the drift and diffusion functions, respectively.

The Wiener process is a special Ito's process because it satisfies Eq. (6.2) with $\mu(x_t, t) = 0$ and $\sigma(x_t, t) = 1$.

## 6.3 ITO'S LEMMA

In finance, when using continuous-time models, it is common to assume that the price of an asset is an Ito's process. Therefore, to derive the price of a financial derivative, one needs to use Ito's calculus. In this section, we briefly review Ito's lemma by treating it as a natural extension of the differentiation in calculus. Ito's lemma is the basis of stochastic calculus.

### 6.3.1 Review of Differentiation

Let $G(x)$ be a differentiable function of $x$. Using Taylor expansion, we have

$$\Delta G \equiv G(x + \Delta x) - G(x) = \frac{\partial G}{\partial x}\,\Delta x + \frac{1}{2}\frac{\partial^2 G}{\partial x^2}(\Delta x)^2 + \frac{1}{6}\frac{\partial^3 G}{\partial x^3}(\Delta x)^3 + \cdots.$$

Taking the limit as $\Delta x \to 0$ and ignoring the higher order terms of $\Delta x$, we have

$$dG = \frac{\partial G}{\partial x}\,dx.$$

When $G$ is a function of $x$ and $y$, we have

$$\Delta G = \frac{\partial G}{\partial x}\,\Delta x + \frac{\partial G}{\partial y}\,\Delta y + \frac{1}{2}\frac{\partial^2 G}{\partial x^2}(\Delta x)^2 + \frac{\partial^2 G}{\partial x \partial y}\,\Delta x\,\Delta y + \frac{1}{2}\frac{\partial^2 G}{\partial y^2}(\Delta y)^2 + \cdots.$$

Taking the limit as $\Delta x \to 0$ and $\Delta y \to 0$, we have

$$dG = \frac{\partial G}{\partial x}\,dx + \frac{\partial G}{\partial y}\,dy.$$

### 6.3.2 Stochastic Differentiation

Turn next to the case in which $G$ is a differentiable function of $x_t$ and $t$, and $x_t$ is an Ito's process. The Taylor expansion becomes

$$\Delta G = \frac{\partial G}{\partial x}\,\Delta x + \frac{\partial G}{\partial t}\,\Delta t + \frac{1}{2}\frac{\partial^2 G}{\partial x^2}(\Delta x)^2 + \frac{\partial^2 G}{\partial x \partial t}\,\Delta x\,\Delta t + \frac{1}{2}\frac{\partial^2 G}{\partial t^2}(\Delta t)^2 + \cdots. \quad (6.3)$$

A discretized version of the Ito's process is

$$\Delta x = \mu\,\Delta t + \sigma\epsilon\sqrt{\Delta t}, \quad (6.4)$$

where, for simplicity, we omit the arguments of $\mu$ and $\sigma$, and $\Delta x = x_{t+\Delta t} - x_t$. From Eq. (6.4), we have

$$(\Delta x)^2 = \mu^2(\Delta t)^2 + \sigma^2\epsilon^2\,\Delta t + 2\mu\sigma\epsilon(\Delta t)^{3/2} = \sigma^2\epsilon^2\,\Delta t + H(\Delta t), \quad (6.5)$$

where $H(\Delta t)$ denotes higher order terms of $\Delta t$. This result shows that $(\Delta x)^2$ contains a term of order $\Delta t$, which cannot be ignored when we take the limit as $\Delta t \to 0$. However, the first term in the right-hand side of Eq. (6.5) has some nice properties:

$$\begin{aligned} E(\sigma^2\epsilon^2\,\Delta t) &= \sigma^2\,\Delta t, \\ \mathrm{Var}(\sigma^2\epsilon^2\,\Delta t) &= E[\sigma^4\epsilon^4(\Delta t)^2] - [E(\sigma^2\epsilon^2\,\Delta t)]^2 = 2\sigma^4(\Delta t)^2, \end{aligned}$$

where we use $E(\epsilon^4) = 3$ for a standard normal random variable. These two properties show that $\sigma^2\epsilon^2\,\Delta t$ converges to a nonstochastic quantity $\sigma^2\,\Delta t$ as $\Delta t \to 0$. Consequently, from Eq. (6.5), we have

$$(\Delta x)^2 \to \sigma^2\,dt \quad \text{as} \quad \Delta t \to 0.$$

Plugging the prior result into Eq. (6.3) and using the Ito's equation of $x_t$ in Eq. (6.2), we obtain

$$\begin{aligned} dG &= \frac{\partial G}{\partial x}\,dx + \frac{\partial G}{\partial t}\,dt + \frac{1}{2}\frac{\partial^2 G}{\partial x^2}\sigma^2\,dt \\ &= \left(\frac{\partial G}{\partial x}\mu + \frac{\partial G}{\partial t} + \frac{1}{2}\frac{\partial^2 G}{\partial x^2}\sigma^2\right)dt + \frac{\partial G}{\partial x}\sigma\,dw_t, \end{aligned}$$

which is the well-known Ito's lemma in Stochastic Calculus.

Recall that we suppressed the argument $(x_t, t)$ from the drift and volatility terms $\mu$ and $\sigma$ in the derivation of Ito's lemma. To avoid any possible confusion in the future, we restate the lemma as follows.

***Ito's Lemma***
Assume that $x_t$ is a continuous-time stochastic process satisfying

$$dx_t = \mu(x_t, t)\,dt + \sigma(x_t, t)\,dw_t,$$

where $w_t$ is a Wiener process. Furthermore, $G(x_t, t)$ is a differentiable function of $x_t$ and $t$. Then,

$$dG = \left[\frac{\partial G}{\partial x}\mu(x_t, t) + \frac{\partial G}{\partial t} + \frac{1}{2}\frac{\partial^2 G}{\partial x^2}\sigma^2(x_t, t)\right] dt + \frac{\partial G}{\partial x}\sigma(x_t, t)\,dw_t. \tag{6.6}$$

**Example 6.1.** As a simple illustration, consider the square function $G(w_t, t) = w_t^2$ of the Wiener process. Here we have $\mu(w_t, t) = 0, \sigma(w_t, t) = 1$ and

$$\frac{\partial G}{\partial w_t} = 2w_t, \quad \frac{\partial G}{\partial t} = 0, \quad \frac{\partial^2 G}{\partial w_t^2} = 2.$$

Therefore,

$$dw_t^2 = \left(2w_t \times 0 + 0 + \frac{1}{2} \times 2 \times 1\right) dt + 2w_t\,dw_t = dt + 2w_t\,dw_t. \tag{6.7}$$

### 6.3.3 An Application

Let $P_t$ be the price of a stock at time $t$, which is continuous in $[0, \infty)$. In the literature, it is common to assume that $P_t$ follows the special Ito's process

$$dP_t = \mu P_t dt + \sigma P_t\,dw_t, \tag{6.8}$$

where $\mu$ and $\sigma$ are constant. Using the notation of the general Ito's process in Eq. (6.2), we have $\mu(x_t, t) = \mu x_t$ and $\sigma(x_t, t) = \sigma x_t$, where $x_t = P_t$. Such a special process is referred to as a *geometric Brownian motion*. We now apply the Ito's lemma to obtain a continuous-time model for the logarithm of the stock price $P_t$. Let $G(P_t, t) = \ln(P_t)$ be the log price of the underlying stock. Then we have

$$\frac{\partial G}{\partial P_t} = \frac{1}{P_t}, \quad \frac{\partial G}{\partial t} = 0, \quad \frac{1}{2}\frac{\partial^2 G}{\partial P_t^2} = \frac{1}{2}\frac{-1}{P_t^2}.$$

Consequently, via Ito's lemma, we obtain

$$d\ln(P_t) = \left(\frac{1}{P_t}\mu P_t + \frac{1}{2}\frac{-1}{P_t^2}\sigma^2 P_t^2\right) dt + \frac{1}{P_t}\sigma P_t\,dw_t = \left(\mu - \frac{\sigma^2}{2}\right) dt + \sigma\,dw_t.$$

This result shows that the logarithm of a price follows a generalized Wiener Process with drift rate $\mu - \sigma^2/2$ and variance rate $\sigma^2$ if the price is a geometric Brownian motion. Consequently, the change in logarithm of price (i.e., log return) between current time $t$ and some future time $T$ is normally distributed with mean $(\mu - \sigma^2/2)(T - t)$ and variance $\sigma^2(T - t)$. If the time interval $T - t = \Delta$ is fixed and we are interested in equally spaced increments in log price, then the increment series is a Gaussian process with mean $(\mu - \sigma^2/2)\,\Delta$ and variance $\sigma^2\,\Delta$.

### 6.3.4 Estimation of $\mu$ and $\sigma$

The two unknown parameters $\mu$ and $\sigma$ of the geometric Brownian motion in Eq. (6.8) can be estimated empirically. Assume that we have $n + 1$ observations of stock price $P_t$ at equally spaced time interval $\Delta$ (e.g., daily, weekly, or monthly). We measure $\Delta$ in years. Denote the observed prices as $\{P_0, P_1, \ldots, P_n\}$ and let $r_t = \ln(P_t) - \ln(P_{t-1})$ for $t = 1, \ldots, n$.

Since $P_t = P_{t-1}\exp(r_t)$, $r_t$ is the continuously compounded return in the $t$th time interval. Using the result of the previous subsection and assuming that the stock price $P_t$ follows a geometric Brownian motion, we obtain that $r_t$ is normally distributed with mean $(\mu - \sigma^2/2)\,\Delta$ and variance $\sigma^2\,\Delta$. In addition, $r_t$s are not serially correlated.

For simplicity, define $\mu_r = E(r_t) = (\mu - \sigma^2/2)\,\Delta$ and $\sigma_r^2 = \mathrm{Var}(r_t) = \sigma^2\,\Delta$. Let $\bar{r}$ and $s_r$ be the sample mean and standard deviation of the data—that is,

$$\bar{r} = \frac{\sum_{t=1}^{n} r_t}{n}, \quad s_r = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(r_t - \bar{r})^2}.$$

As mentioned in Chapter 1, $\bar{r}$ and $s_r$ are consistent estimates of the mean and standard deviation of $r_i$, respectively. That is, $\bar{r} \to \mu_r$ and $s_r \to \sigma_r$ as $n \to \infty$. Therefore, we may estimate $\sigma$ by

$$\hat{\sigma} = \frac{s_r}{\sqrt{\Delta}}.$$

Furthermore, it can be shown that the standard error of this estimate is approximately $\hat{\sigma}/\sqrt{2n}$. From $\hat{\mu}_r = \bar{r}$, we can estimate $\mu$ by

$$\hat{\mu} = \frac{\bar{r}}{\Delta} + \frac{\hat{\sigma}^2}{2} = \frac{\bar{r}}{\Delta} + \frac{s_r^2}{2\,\Delta}.$$

When the series $r_t$ is serially correlated or when the price of the asset does not follow the geometric Brownian motion in Eq. (6.8), then other estimation methods must be used to estimate the drift and volatility parameters of the diffusion equation. We return to this issue later.

**Example 6.2.** Consider the daily log returns of IBM stock in 1998. Figure 6.2(a) shows the time plot of the data, which have 252 observations. Figure 6.2(b) shows the sample autocorrelations of the series. It is seen that the log returns are indeed serially uncorrelated. The Ljung–Box statistic gives $Q(10) = 4.9$, which is highly insignificant compared with a chi-squared distribution with 10 degrees of freedom.

If we assume that the price of IBM stock in 1998 follows the geometric Brownian motion in Eq. (6.8), then we can use the daily log returns to estimate the parameters $\mu$ and $\sigma$. From the data, we have $\bar{r} = 0.002276$ and $s_r = 0.01915$. Since 1 trading day is equivalent to $\Delta = 1/252$ year, we obtain that

$$\hat{\sigma} = \frac{s_r}{\sqrt{\Delta}} = 0.3040, \quad \hat{\mu} = \frac{\bar{r}}{\Delta} + \frac{\hat{\sigma}^2}{2} = 0.6198.$$

Thus, the estimated expected return was 61.98% and the standard deviation was 30.4% per annum for IBM stock in 1998.

The normality assumption of the daily log returns may not hold, however. In this particular instance, the skewness $-0.464(0.153)$ and excess kurtosis $2.396(0.306)$ raise some concern, where the number in parentheses denotes asymptotic standard error.

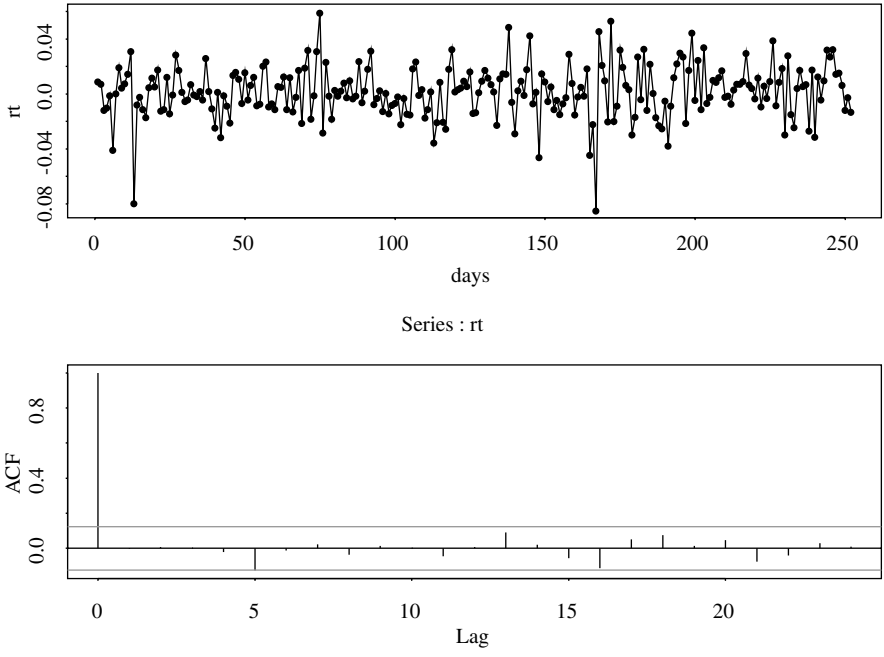


**Figure 6.2.** Daily returns of IBM stock in 1998: (a) log returns, and (b) sample autocorrelations.

**Example 6.3.** Consider the daily log return of the stock of Cisco Systems, Inc. in 1999. There are 252 observations, and the sample mean and standard deviation are 0.00332 and 0.026303, respectively. The log return series also shows no serial correlation with $Q(12) = 10.8$, which is not significant even at the 10% level. Therefore, we have

$$\hat{\sigma} = \frac{s_r}{\sqrt{\Delta}} = \frac{0.00332}{\sqrt{1.0/252.0}} = 0.418, \quad \hat{\mu} = \frac{\bar{r}}{\Delta} + \frac{\hat{\sigma}^2}{2} = 0.924.$$

Consequently, the estimated expected return for Cisco Systems' stock was 92.4% per annum, and the estimated standard deviation was 41.8% per annum in 1999.

## 6.4 DISTRIBUTIONS OF STOCK PRICES AND LOG RETURNS

The result of the previous section shows that if one assumes that price of a stock follows the geometric Brownian motion

$$dP_t = \mu P_t dt + \sigma P_t \, dw_t,$$

then the logarithm of the price follows a generalized Wiener process

$$d\ln(P_t) = \left(\mu - \frac{\sigma^2}{2}\right) dt + \sigma \, dw_t,$$

where $P_t$ is the price of the stock at time $t$ and $w_t$ is a Wiener process. Therefore, the change in log price from time $t$ to $T$ is normally distributed as

$$\ln(P_T) - \ln(P_t) \sim N\left[\left(\mu - \frac{\sigma^2}{2}\right)(T-t), \sigma^2(T-t)\right]. \tag{6.9}$$

Consequently, conditional on the price $P_t$ at time $t$, the log price at time $T > t$ is normally distributed as

$$\ln(P_T) \sim N\left[\ln(P_t) + \left(\mu - \frac{\sigma^2}{2}\right)(T-t), \sigma^2(T-t)\right]. \tag{6.10}$$

Using the result of lognormal distribution discussed in Chapter 1, we obtain the (conditional) mean and variance of $P_T$ as

$$\begin{aligned} E(P_T) &= P_t \exp[\mu(T-t)], \\ \text{Var}(P_T) &= P_t^2 \exp[2\mu(T-t)]\{\exp[\sigma^2(T-t)] - 1\}. \end{aligned}$$

Note that the expectation confirms that $\mu$ is the expected rate of return of the stock.

The prior distribution of stock price can be used to make inference. For example, suppose that the current price of Stock A is \$50, the expected return of the stock is 15% per annum, and the volatility is 40% per annum. Then the expected price of Stock A in 6-month (0.5 year) and the associated variance are given by

$$E(P_T) = 50\exp(0.15 \times 0.5) = 53.89,$$
$$\text{Var}(P_T) = 2500\exp(0.3 \times 0.5)[\exp(0.16 \times 0.5) - 1] = 241.92.$$

The standard deviation of the price 6 months from now is $\sqrt{241.92} = 15.55$.

Next, let $r$ be the continuously compounded rate of return per annum from time $t$ to $T$. Then we have

$$P_T = P_t \exp[r(T-t)],$$

where $T$ and $t$ are measured in years. Therefore,

$$r = \frac{1}{T-t}\ln\left(\frac{P_T}{P_t}\right).$$

By Eq. (6.9), we have

$$\ln\left(\frac{P_T}{P_t}\right) \sim N\left[\left(\mu - \frac{\sigma^2}{2}\right)(T-t), \sigma^2(T-t)\right].$$

Consequently, the distribution of the continuously compounded rate of return per annum is

$$r \sim N\left(\mu - \frac{\sigma^2}{2}, \frac{\sigma^2}{T-t}\right).$$

The continuously compounded rate of return is, therefore, normally distributed with mean $\mu - \sigma^2/2$ and standard deviation $\sigma/\sqrt{T-t}$.

Consider a stock with an expected rate of return of 15% per annum and a volatility of 10% per annum. The distribution of the continuously compounded rate of return of the stock over two years is normal with mean $0.15 - 0.01/2 = 0.145$ or 14.5% per annum and standard deviation $0.1/\sqrt{2} = 0.071$ or 7.1% per annum. These results allow us to construct confidence intervals (C.I.) for $r$. For instance, a 95% C.I. for $r$ is $0.145 \pm 1.96 \times 0.071$ per annum (i.e., 0.6%, 28.4%).

## 6.5 DERIVATION OF BLACK–SCHOLES DIFFERENTIAL EQUATION

In this section, we use Ito's lemma and assume no arbitrage to derive the Black–Scholes differential equation for the price of a derivative contingent to a stock valued at $P_t$. Assume that the price $P_t$ follows the geometric Brownian motion in Eq. (6.8)

and $G_t = G(P_t, t)$ is the price of a derivative (e.g., a call option) contingent on $P_t$. By Ito's lemma,

$$dG_t = \left( \frac{\partial G_t}{\partial P_t} \mu P_t + \frac{\partial G_t}{\partial t} + \frac{1}{2} \frac{\partial^2 G_t}{\partial P_t^2} \sigma^2 P_t^2 \right) dt + \frac{\partial G_t}{\partial P_t} \sigma P_t \, dw_t.$$

The discretized versions of the process and previous result are

$$\Delta P_t = \mu P_t \, \Delta t + \sigma P_t \, \Delta w_t, \tag{6.11}$$

$$\Delta G_t = \left( \frac{\partial G_t}{\partial P_t} \mu P_t + \frac{\partial G_t}{\partial t} + \frac{1}{2} \frac{\partial^2 G_t}{\partial P_t^2} \sigma^2 P_t^2 \right) \Delta t + \frac{\partial G_t}{\partial P_t} \sigma P_t \, \Delta w_t, \tag{6.12}$$

where $\Delta P_t$ and $\Delta G_t$ are changes in $P_t$ and $G_t$ in a small time interval $\Delta t$. Because $\Delta w_t = \epsilon \sqrt{\Delta t}$ for both Eqs. (6.11) and (6.12), one can construct a portfolio of the stock and the derivative that does not involve the Wiener process. The appropriate portfolio is short on derivative and long $\frac{\partial G_t}{\partial P_t}$ shares of the stock. Denote the value of the portfolio by $V_t$. By construction,

$$V_t = -G_t + \frac{\partial G_t}{\partial P_t} P_t. \tag{6.13}$$

The change in $V_t$ is then

$$\Delta V_t = -\Delta G_t + \frac{\partial G_t}{\partial P_t} \, \Delta P_t. \tag{6.14}$$

Substituting Eqs. (6.11) and (6.12) into Eq. (6.14), we have

$$\Delta V_t = \left( -\frac{\partial G_t}{\partial t} - \frac{1}{2} \frac{\partial^2 G_t}{\partial P_t^2} \sigma^2 P_t^2 \right) \Delta t. \tag{6.15}$$

This equation does not involve the stochastic component $\Delta w_t$. Therefore, under the no arbitrage assumption, the portfolio $V_t$ must be riskless during the small time interval $\Delta t$. In other words, the assumptions used imply that the portfolio must instantaneously earn the same rate of return as other short-term, risk-free securities. Otherwise there exists an arbitrage opportunity between the portfolio and the short-term, risk-free securities. Consequently, we have

$$\Delta V_t = r V_t \, \Delta t, \tag{6.16}$$

where $r$ is the risk-free interest rate. By Eqs. (6.13) to (6.16), we have

$$\left( \frac{\partial G_t}{\partial t} + \frac{1}{2} \frac{\partial^2 G_t}{\partial P_t^2} \sigma^2 P_t^2 \right) \Delta t = r \left( G_t - \frac{\partial G_t}{\partial P_t} P_t \right) \Delta t.$$

Therefore,

$$\frac{\partial G_t}{\partial t} + r P_t \frac{\partial G_t}{\partial P_t} + \frac{1}{2}\sigma^2 P_t^2 \frac{\partial^2 G_t}{\partial P_t^2} = r G_t. \tag{6.17}$$

This is the Black–Scholes differential equation for derivative pricing. It can be solved to obtain the price of a derivative with $P_t$ as the underlying variable. The solution so obtained depends on the boundary conditions of the derivative. For a European call option, the boundary condition is

$$G_T = \max(P_T - K, 0),$$

where $T$ is the expiration time and $K$ is the strike price. For a European put option, the boundary condition becomes

$$G_T = \max(K - P_T, 0).$$

**Example 6.4.** As a simple example, consider a forward contract on a stock that pays no dividend. In this case, the value of the contract is given by

$$G_t = P_t - K \exp[-r(T - t)],$$

where $K$ is the delivery price, $r$ is the risk-free interest rate, and $T$ is the expiration time. For such a function, we have

$$\frac{\partial G_t}{\partial t} = -rK \exp[-r(T - t)], \quad \frac{\partial G_t}{\partial P_t} = 1, \quad \frac{\partial^2 G_t}{\partial P_t^2} = 0.$$

Substituting these quantities into the left-hand side of Eq. (6.17) yields

$$-rK \exp[-r(T - t)] + r P_t = r\{P_t - K \exp[-r(T - t)]\},$$

which equals the right-hand side of Eq. (6.17). Thus, the Black–Scholes differential equation is indeed satisfied.

## 6.6 BLACK–SCHOLES PRICING FORMULAS

Black and Scholes (1973) successfully solve their differential equation in Eq. (6.17) to obtain exact formulas for the price of European call and put options. In what follows, we derive these formulas using what is called *Risk-Neutral Valuation* in finance.

### 6.6.1 Risk-Neutral World

The drift parameter $\mu$ drops out from the Black–Scholes differential equation. In finance, this means the equation is independent of risk preferences. In other words,

risk preferences cannot affect the solution of the equation. A nice consequence of this property is that one can assume that investors are risk-neutral. In a risk-neutral world, we have the following results:

- The expected return on all securities is the risk-free interest rate $r$, and
- The present value of any cash flow can be obtained by discounting its expected value at the risk-free rate.

### 6.6.2 Formulas

The expected value of a European call option at maturity in a risk-neutral world is

$$E_*[\max(P_T - K, 0)],$$

where $E_*$ denotes expected value in a risk-neutral world. The price of the call option at time $t$ is

$$c_t = \exp[-r(T-t)]E_*[\max(P_T - K, 0)]. \tag{6.18}$$

Yet in a risk-neutral world, we have $\mu = r$, and by Eq. (6.10), $\ln(P_T)$ is normally distributed as

$$\ln(P_T) \sim N\left[\ln(P_t) + \left(r - \frac{\sigma^2}{2}\right)(T-t), \sigma^2(T-t)\right].$$

Let $g(P_T)$ be the probability density function of $P_T$. Then the price of the call option in Eq. (6.18) is

$$c_t = \exp[-r(T-t)]\int_K^\infty (P_T - K)g(P_T)dP_T.$$

By changing the variable in the integration and some algebraic calculations (details are given in Appendix A), we have

$$c_t = P_t\Phi(h_+) - K\exp[-r(T-t)]\Phi(h_-), \tag{6.19}$$

where $\Phi(x)$ is the cumulative distribution function (CDF) of the standard normal random variable evaluated at $x$,

$$h_+ = \frac{\ln(P_t/K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}$$

$$h_- = \frac{\ln(P_t/K) + (r - \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} = h_+ - \sigma\sqrt{T-t}.$$

In practice, $\Phi(x)$ can easily be obtained from most statistical packages. Alternatively, one can use an approximation given in Appendix B.

The Black–Scholes call formula in Eq. (6.19) has some nice interpretations. First, if we exercise the call option on the expiration date, we receive the stock, but we have to pay the strike price. This exchange will take place only when the call finishes in-the-money (i.e., $P_T > K$). The first term $P_t\Phi(h_+)$ is the present value of receiving the stock if and only if $P_T > K$ and the second term $-K\exp[-r(T-t)]\Phi(h_-)$ is the present value of paying the strike price if and only if $P_T > K$. A second interpretation is particularly useful. As shown in the derivation of Black–Scholes differential equation in Section 6.5, $\Phi(h_+) = \frac{\partial G_t}{\partial P_t}$ is the number of shares in the portfolio that does not involve uncertainty, the Wiener process. We know that $c_t = P_t\Phi(h_+) + B_t$, where $B_t$ is the dollar amount invested in risk-free bonds in the portfolio (or short on the derivative). We can then see that $B_t = -K\exp[-r(T-t)]\Phi(h_-)$ directly from inspection of the Black–Scholes formula. The first term of the formula $P_t\Phi(h_+)$ is the amount invested in the stock, whereas the second term, $K\exp[-r(T-t)]\Phi(h_-)$, is the amount borrowed.

Similarly, we can obtain the price of a European put option as

$$p_t = K\exp[-r(T-t)]\Phi(-h_-) - P_t\Phi(-h_+). \tag{6.20}$$

Since the standard normal distribution is symmetric with respect to its mean 0.0, we have $\Phi(x) = 1-\Phi(-x)$ for all $x$. Using this property, we have $\Phi(-h_i) = 1-\Phi(h_i)$. Thus, the information needed to compute the price of a put option is the same as that of a call option. Alternatively, using the symmetry of normal distribution, it is easy to verify that

$$p_t - c_t = K\exp[-r(T-t)] - P_t,$$

which is referred to as the *put-call parity* and can be used to obtain $p_t$ from $c_t$.

**Example 6.5.** Suppose that the current price of Intel stock is \$80 per share with volatility $\sigma = 20\%$ per annum. Suppose further that the risk-free interest rate is 8% per annum. What is the price of a European call option on Intel with a strike price of \$90 that will expire in 3 months?

From the assumptions, we have $P_t = 80$, $K = 90$, $T - t = 0.25$, $\sigma = 0.2$, and $r = 0.08$. Therefore,

$$h_+ = \frac{\ln(80/90) + (0.08 + 0.04/2) \times 0.25}{0.2\sqrt{0.25}} = -0.9278$$

$$h_- = h_+ - 0.2\sqrt{0.25} = -1.0278.$$

Using any statistical software (e.g., Minitab or SCA), or the approximation in Appendix B, we have

$$\Phi(-0.9278) = 0.1767, \quad \Phi(-1.0278) = 0.1520.$$

Consequently, the price of a European call option is

$$c_t = \$80\Phi(-0.9278) - \$90\Phi(-1.0278)\exp(-0.02) = \$0.73.$$

The stock price has to rise by \$10.73 for the purchaser of the call option to break even.

Under the same assumptions, the price of a European put option is

$$p_t = \$90 \exp(-0.08 \times 0.25)\Phi(1.0278) - \$80\Phi(0.9278) = \$8.95.$$

Thus, the stock price can rise an additional \$1.05 for the purchaser of the put option to break even.

**Example 6.6.** The strike price of the previous example is well beyond the current stock price. A more realistic strike price is \$85. Assume that the other conditions of the previous example continue to hold. We now have $P_t = 80$, $K = 85$, $r = 0.08$, and $T - t = 0.25$, and the $h_i$s become

$$h_+ = \frac{\ln(80/85) + (0.08 + 0.04/2) \times 0.25}{0.2\sqrt{0.25}} = -0.356246$$
$$h_- = h_+ - 0.2\sqrt{0.25} = -0.456246.$$

Using the approximation in Appendix B, we have $\Phi(-0.356246) = 0.3608$ and $\Phi(-0.456246) = 0.3241$. The price of a European call option is then

$$c_t = \$85\Phi(-0.356246) - \$85 \exp(-0.02)\Phi(-0.456246) = \$1.86.$$

The price of the stock has to rise by \$6.86 for the purchaser of the call option to break even. Yet under the same assumptions, the price of a European put option is

$$\begin{aligned} p_t &= \$85 \exp(-0.02)\Phi(0.456246) - \$80\Phi(0.356246) \\ &= \$85 \exp(-0.02) \times 0.6759 - \$80 \times 0.6392 = \$5.18. \end{aligned}$$

The stock price must fall \$0.18 for the purchaser of the put option to break even.

### 6.6.3 Discussion

From the formulas, the price of a call or put option depends on five variables—namely, the current stock price $P_t$, the strike price $K$, the time to expiration $T - t$ measured in years, the volatility $\sigma$ per annum, and the interest rate $r$ per annum. It pays to study the effects of these five variables on the price of an option.

#### *6.6.3.1 Marginal Effects*

Consider first the marginal effects of the five variables on the price of a call option $c_t$. By marginal effects we mean that changing one variable while holding the others fixed. The effects on a call option can be summarized as follows.

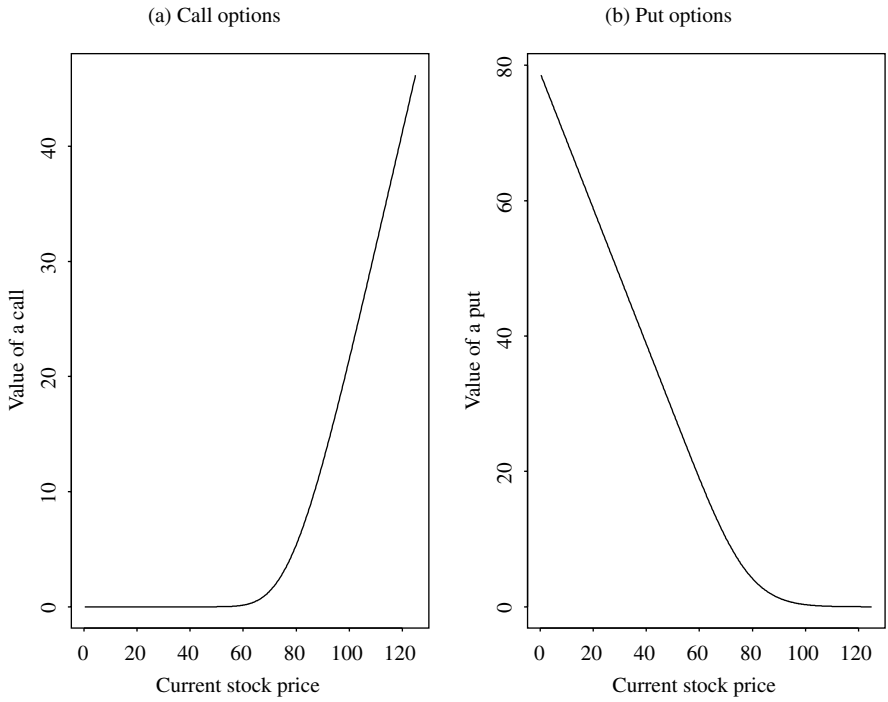


**Figure 6.3.** Marginal effects of the current stock price on the price of an option with $K = 80$, $T - t = 0.25$, $\sigma = 0.3$, and $r = 0.06$: (a) call option, and (b) put option.

1. The current stock price $P_t$: $c_t$ is positively related to $\ln(P_t)$. In particular, $c_t \to 0$ as $P_t \to 0$ and $c_t \to \infty$ as $P_t \to \infty$. Figure 6.3(a) illustrates the effects with $K = 80$, $r = 6\%$ per annum, $T - t = 0.25$ years, and $\sigma = 30\%$ per annum.
2. The strike price $K$: $c_t$ is negatively related to $\ln(K)$. In particular, $c_t \to P_t$ as $K \to 0$ and $c_t \to 0$ as $K \to \infty$.
3. Time to expiration: $c_t$ is related to $T - t$ in a complicated manner, but we can obtain the limiting results by writing $h_+$ and $h_-$ as

$$h_+ = \frac{\ln(P_t/K)}{\sigma\sqrt{T-t}} + \frac{(r+\sigma^2/2)\sqrt{T-t}}{\sigma},$$

$$h_- = \frac{\ln(P_t/K)}{\sigma\sqrt{T-t}} + \frac{(r-\sigma^2/2)\sqrt{T-t}}{\sigma}.$$

If $P_t < K$, then $c_t \to 0$ as $(T - t) \to 0$. If $P_t > K$, then $c_t \to P_t - K$ as $(T - t) \to 0$ and $c_t \to P_t$ as $(T - t) \to \infty$. Figure 6.4(a) shows the marginal effects of $T - t$ on $c_t$ for three different current stock prices. The fixed variables are $K = 80$, $r = 6\%$, and $\sigma = 30\%$. The solid, dotted, and dashed lines of the plot are for $P_t = 70$, 80, and 90, respectively.

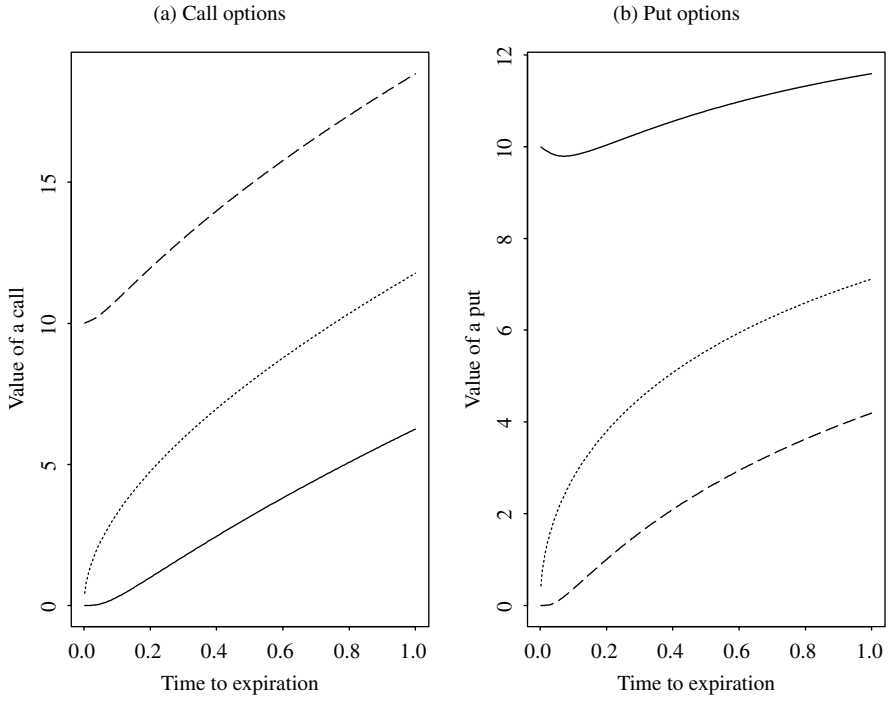


**Figure 6.4.** Marginal effects of the time to expiration on the price of an option with $K = 80$, $\sigma = 0.3$, and $r = 0.06$: (a) call option, and (b) put option. The solid, dotted, and dashed lines are for the current stock price $P_t = 70$, 80, and 90, respectively.

**4.** Volatility $\sigma$: Rewriting $h_+$ and $h_-$ as

$$h_+ = \frac{\ln(P_t/K) + r(T-t)}{\sigma\sqrt{T-t}} + \frac{\sigma}{2}\sqrt{T-t}$$
$$h_- = \frac{\ln(P_t/K) + r(T-t)}{\sigma\sqrt{T-t}} - \frac{\sigma}{2}\sqrt{T-t},$$

we obtain that (a) if $\ln(P_t/K) + r(T-t) < 0$, then $c_t \to 0$ as $\sigma \to 0$, and (b) if $\ln(P_t/K) + r(T-t) \geq 0$, then $c_t \to P_t - Ke^{-r(T-t)}$ as $\sigma \to 0$ and $c_t \to P_t$ as $\sigma \to \infty$. Figure 6.5(a) shows the effects of $\sigma$ on $c_t$ for $K = 80$, $T - t = 0.25$, $r = 0.06$, and three different values of $P_t$. The solid, dotted, and dashed lines are for $P_t = 70$, 80, and 90, respectively.

**5.** Interest rate: $c_t$ is positively related to $r$ such that $c_t \to P_t$ as $r \to \infty$.

The marginal effects of the five variables on a put option can be obtained similarly. Part (b) of Figures 6.3–6.5 illustrates the effects for some selected cases.

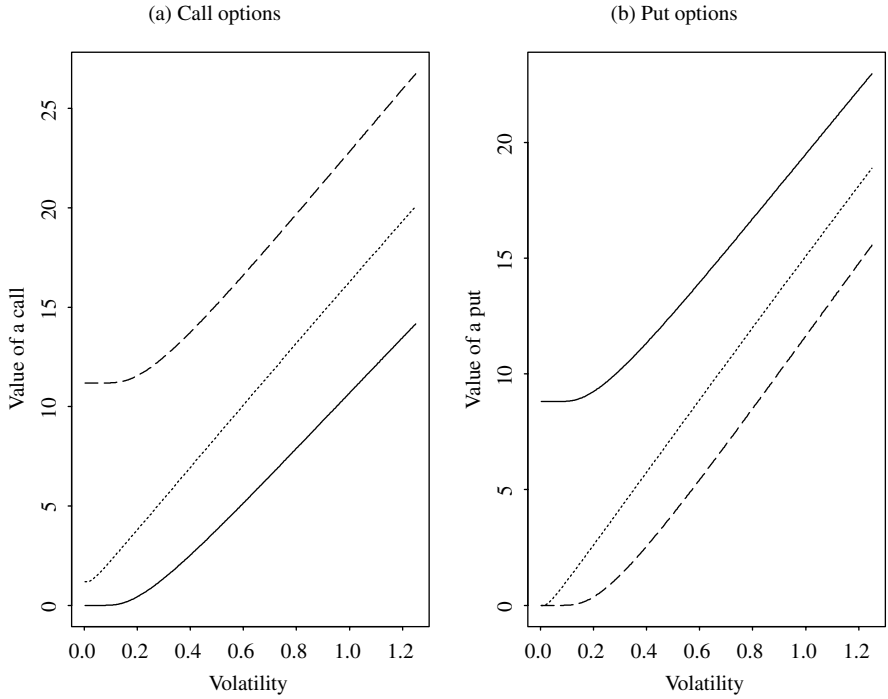


**Figure 6.5.** Marginal effects of stock volatility on the price of an option with $K = 80$, $T - t = 0.25$, and $r = 0.06$: (a) call option, and (b) put option. The solid, dotted, and dashed lines are for the current stock price $P_t = 70$, 80, and 90, respectively.

#### *6.6.3.2 Some Joint Effects*

Figure 6.6 shows the joint effects of volatility and strike price on a call option, where the other variables are fixed at $P_t = 80$, $r = 0.06$, and $T - t = 0.25$. As expected, the price of a call option is higher when the volatility is high and the strike price is well below the current stock price. Figure 6.7 shows the effects on a put option under the same conditions. The price of a put option is higher when the volatility is high and the strike price is well above the current stock price. Furthermore, the plot also shows that the effects of a strike price on the price of a put option becomes more linear as the volatility increases.

## 6.7 AN EXTENSION OF ITO'S LEMMA

In derivative pricing, a derivative may be contingent on multiple securities. When the prices of these securities are driven by multiple factors, the price of the derivative is a function of several stochastic processes. The two-factor model for the term structure of interest rate is an example of two stochastic processes. In this section, we briefly discuss the extension of Ito's lemma to the case of several stochastic processes.

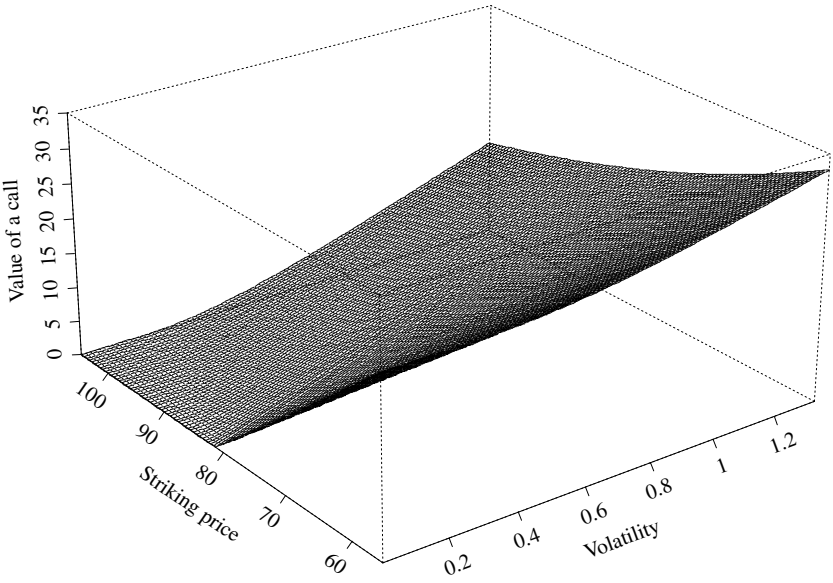


**Figure 6.6.** Joint effects of stock volatility and the strike price on a call option with $P_t = 80$, $r = 0.06$, and $T - t = 0.25$.

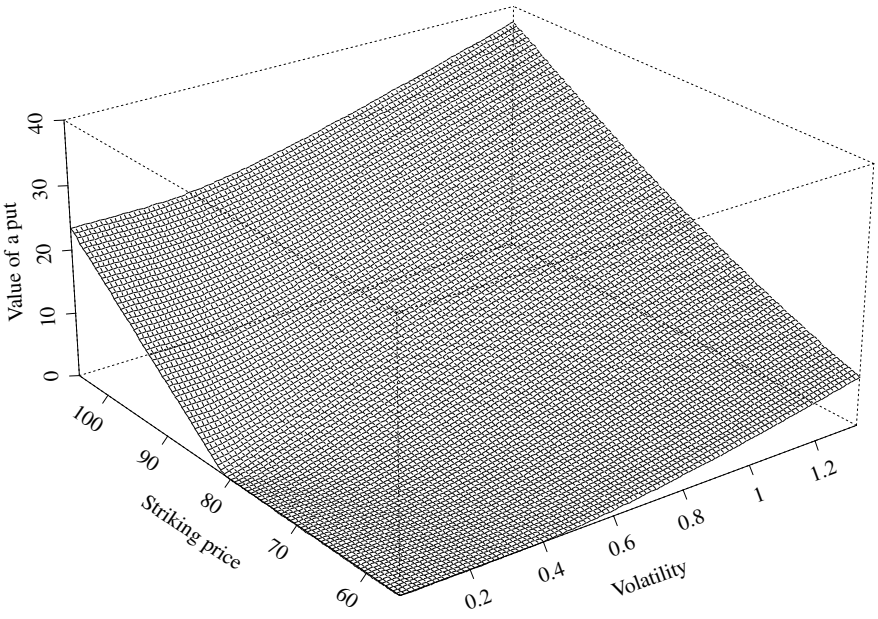


**Figure 6.7.** Joint effects of stock volatility and the strike price on a put option with $K = 80$, $T - t = 0.25$, and $r = 0.06$.

Consider a $k$-dimensional continuous-time process $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{kt})'$, where $k$ is a positive integer and $x_{it}$ is a continuous-time stochastic process satisfying

$$dx_{it} = \mu_i(\boldsymbol{x}_t)dt + \sigma_i(\boldsymbol{x}_t)\, dw_{it}, \quad i = 1, \ldots, k, \tag{6.21}$$

where $w_{it}$ is a Wiener process. It is understood that the drift and volatility functions $\mu_i(x_{it})$ and $\sigma_i(x_{it})$ are functions of time index $t$ as well. We omit $t$ from their arguments to simplify the notation. For $i \neq j$, the Wiener processes $w_{it}$ and $w_{jt}$ are different. We assume that the correlation between $dw_{it}$ and $dw_{jt}$ is $\rho_{ij}$. This means that $\rho_{ij}$ is the correlation between the two standard normal random variables $\epsilon_i$ and $\epsilon_j$ defined by $\Delta w_{it} = \epsilon_i \Delta t$ and $\Delta w_{jt} = \epsilon_j \Delta t$. Assume that $G_t = G(\boldsymbol{x}_t, t)$ be a function of the stochastic processes $x_{it}$ and time $t$. The Taylor expansion gives

$$\Delta G_t = \sum_{i=1}^{k} \frac{\partial G_t}{\partial x_{it}} \Delta x_{it} + \frac{\partial G_t}{\partial t} \Delta t + \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{\partial^2 G_t}{\partial x_{it} \partial x_{jt}} \Delta x_{it} \Delta x_{jt}$$
$$+ \frac{1}{2} \sum_{i=1}^{k} \frac{\partial^2 G_t}{\partial x_{it} \partial t} \Delta x_{it} \Delta t + \cdots.$$

The discretized version of Eq. (6.21) is

$$\Delta w_{it} = \mu_i(\boldsymbol{x}_t)\, \Delta t + \sigma_i(\boldsymbol{x}_t)\, \Delta w_{it}, \quad i = 1, \ldots, k.$$

Using a similar argument as that of Eq. (6.5) in Section 6.3, we can obtain that

$$\lim_{\Delta t \to 0} (\Delta x_{it})^2 \to \sigma_i^2(\boldsymbol{x}_t)dt \tag{6.22}$$

$$\lim_{\Delta t \to 0} \Delta x_{it}\, \Delta x_{jt} \to \sigma_i(\boldsymbol{x}_t)\sigma_j(\boldsymbol{x}_t)\rho_{ij}dt. \tag{6.23}$$

Using Eqs. (6.21)–(6.23), taking the limit as $\Delta t \to 0$, and ignoring higher order terms of $\Delta t$, we have

$$dG_t = \left[ \sum_{i=1}^{k} \frac{\partial G_t}{\partial x_{it}} \mu_i(\boldsymbol{x}_t) + \frac{\partial G_t}{\partial t} + \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{\partial^2 G_t}{\partial x_{it} \partial x_{jt}} \sigma_i(\boldsymbol{x}_t)\sigma_j(\boldsymbol{x}_t)\rho_{ij} \right] dt$$
$$+ \sum_{i=1}^{k} \frac{\partial G_t}{\partial x_{it}} \sigma_i(\boldsymbol{x}_t)\, dw_{it}. \tag{6.24}$$

This is a generalization of Ito's lemma to the case of multiple stochastic processes.

## 6.8 STOCHASTIC INTEGRAL

We briefly discuss stochastic integration so that the price of an asset can be obtained under the assumption that it follows an Ito's process. We deduce the integration result

using the Ito's formula. For a rigorous treatment on the topic, readers may consult textbooks on stochastic calculus. First, like the usual integration of a deterministic function, integration is the opposite side of differentiation so that

$$\int_0^t dx_s = x_t - x_0$$

continues to hold for a stochastic process $x_t$. In particular, for the Wiener process $w_t$, we have $\int_0^t dw_s = w_t$ because $w_0 = 0$. Next, consider the integration $\int_0^t w_s\, dw_s$. Using the prior result and taking integration of Eq. (6.7), we have

$$w_t^2 = t + 2\int_0^t w_s\, dw_s.$$

Therefore,

$$\int_0^t w_s\, dw_s = \frac{1}{2}(w_t^2 - t).$$

This is different from the usual deterministic integration for which $\int_0^t y\, dy = (y_t^2 - y_0^2)/2$.

Turn to the case that $x_t$ is a geometric Brownian motion—that is, $x_t$ satisfies

$$dx_t = \mu x_t dt + \sigma x_t\, dw_t,$$

where $\mu$ and $\sigma$ are constant with $\sigma > 0$; see Eq. (6.8). Applying the Ito's lemma to $G(x_t, t) = \ln(x_t)$, we obtain

$$d\ln(x_t) = \left(\mu - \frac{\sigma^2}{2}\right) dt + \sigma\, dw_t.$$

Taking integration and using the results obtained before, we have

$$\int_0^t d\ln(x_s) = \left(\mu - \frac{\sigma^2}{2}\right)\int_0^t ds + \sigma\int_0^t dw_s.$$

Consequently,

$$\ln(x_t) = \ln(x_0) + (\mu - \sigma^2/2)t + \sigma w_t$$

and

$$x_t = x_0 \exp[(\mu - \sigma^2/2)t + \sigma w_t].$$

Changing the notation $x_t$ to $P_t$ for the price of an asset, we have a solution for the price under the assumption that it is a geometric Brownian motion. The price is

$$P_t = P_0 \exp[(\mu - \sigma^2/2)t + \sigma w_t]. \tag{6.25}$$

## 6.9 JUMP DIFFUSION MODELS

Empirical studies have found that the stochastic diffusion model based on Brownian motion fails to explain some characteristics of asset returns and the prices of their derivatives (e.g., the "volatility smile" of implied volatilities; see Bakshi, Cao, and Chen, 1997, and the references therein). Volatility smile is referred to as the convex function between the implied volatility and strike price of an option. Both out-of-the-money and in-the-money options tend to have higher implied volatilities than at-the-money options especially in the foreign exchange markets. Volatility smile is less pronounced for equity options. The inadequacy of the standard stochastic diffusion model has led to the developments of alternative continuous-time models. For example, jump diffusion and stochastic volatility models have been proposed in the literature to overcome the inadequacy; see Merton (1976) and Duffie (1995).

Jumps in stock prices are often assumed to follow a probability law. For example, the jumps may follow a Poisson process, which is a continuous-time discrete process. For a given time $t$, let $X_t$ be the number of times a special event occurs during the time period $[0, t]$. Then $X_t$ is a Poisson process if

$$Pr(X_t = m) = \frac{\lambda^m t^m}{m!} \exp(-\lambda t), \quad \lambda > 0.$$

That is, $X_t$ follows a Poisson distribution with parameter $\lambda t$. The parameter $\lambda$ governs the occurrence of the special event and is referred to as the *rate* or *intensity* of the process. A formal definition also requires that $X_t$ be a right-continuous homogeneous Markov process with left-hand limit.

In this section, we discuss a simple jump diffusion model proposed by Kou (2000). This simple model enjoys several nice properties. The returns implied by the model are leptokurtic and asymmetric with respect to zero. In addition, the model can reproduce volatility smile and provide analytical formulas for the prices of many options. The model consists of two parts, with the first part being continuous and following a geometric Brownian motion and the second part being a jump process. The occurrences of jump are governed by a Poisson process, and the jump size follows a double exponential distribution. Let $P_t$ be the price of an asset at time $t$. The simple jump diffusion model postulates that the price follows the stochastic differential equation

$$\frac{dP_t}{P_t} = \mu dt + \sigma\, dw_t + d\left(\sum_{i=1}^{n_t}(J_i - 1)\right), \tag{6.26}$$

where $w_t$ is a Wiener process, $n_t$ is a Poisson process with rate $\lambda$, and $\{J_i\}$ is a sequence of independent and identically distributed nonnegative random variables such that $X = \ln(J)$ has a double exponential distribution with probability density function

$$f_X(x) = \frac{1}{2\eta} e^{-|x-\kappa|/\eta}, \quad 0 < \eta < 1. \tag{6.27}$$

In model (6.26), $n_t$, $w_t$, and $J_i$ are independent so that there is no relation between the randomness of the model. Notice that $n_t$ is the number of jumps in the time interval $[0, t]$ and follows a Poisson distribution with parameter $\lambda t$, where $\lambda$ is a constant. At the $i$th jump, the proportion of price jump is $J_i - 1$.

The double exponential distribution can be written as

$$X - \kappa = \begin{cases} \xi & \text{with probability 0.5} \\ -\xi & \text{with probability 0.5,} \end{cases} \tag{6.28}$$

where $\xi$ is an exponential random variable with mean $\eta$ and variance $\eta^2$. The probability density function of $\xi$ is

$$f(x) = \frac{1}{\eta} e^{-x/\eta}, \quad 0 < x < \infty.$$

Some useful properties of the double exponential distribution are

$$E(X) = \kappa, \quad \text{Var}(X) = 2\eta^2, \quad E(e^X) = \frac{e^\kappa}{1 - \eta^2}.$$

For finite samples, it is hard to distinguish a double exponential distribution from a Student-$t$ distribution. However, a double exponential distribution is more tractable analytically and can generate a higher probability concentration (e.g., higher peak) around its mean value. As stated in Chapter 1, histograms of observed asset returns tend to have a higher peak than the normal density. Figure 6.8 shows the probability density function of a double exponential random variable in the solid line and that of a normal random variable in the dotted line. Both variables have mean zero and variance 0.0008. The high peak of the double exponential density is clearly seen.

Solving the stochastic differential equation in Eq. (6.26), we obtain the dynamics of the asset price as

$$P_t = P_0 \exp[(\mu - \sigma^2/2)t + \sigma w_t] \prod_{i=1}^{n_t} J_i, \tag{6.29}$$

where it is understood that $\prod_{i=1}^{0} = 1$. This result is a generalization of Eq. (6.25) by including the stochastic jumps. It can be obtained as follows. Let $t_i$ be the time of the $i$th jump. For $t \in [0, t_1)$, there is no jump and the price is given in Eq. (6.25).

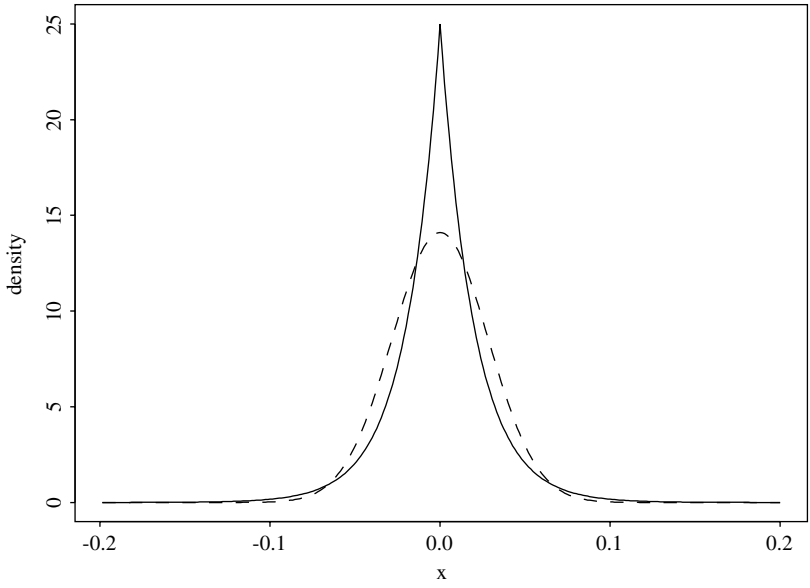


**Figure 6.8.** Probability density functions of a double exponential and a normal random variable with mean zero and variance 0.0008. The solid line denotes the double exponential distribution.

Consequently, the left-hand price limit at time $t_1$ is

$$P_{t_1^-} = P_0 \exp[(\mu - \sigma^2/2)t_1 + \sigma w_{t_1}].$$

At time $t_1$, the proportion of price jump is $J_1 - 1$ so that the price becomes

$$P_{t_1} = (1 + J_1 - 1)P_{t_1^-} = J_1 P_{t_1^-} = P_0 \exp[(\mu - \sigma^2/2)t_1 + \sigma w_{t_1}]J_1.$$

For $t \in (t_1, t_2)$, there is no jump in the interval $(t_1, t]$ so that

$$P_t = P_{t_1} \exp[(\mu - \sigma^2/2)(t - t_1) + \sigma(w_t - w_{t_1})].$$

Plugging in $P_{t_1}$, we have

$$P_t = P_0 \exp[(\mu - \sigma^2/2)t + \sigma w_t]J_1.$$

Repeating the scheme, we obtain Eq. (6.29).

From Eq. (6.29), the simple return of the underlying asset in a small time increment $\Delta t$ becomes

$$\frac{P_{t+\Delta t} - P_t}{P_t} = \exp\left[\left(\mu - \frac{1}{2}\sigma^2\right)\Delta t + \sigma(w_{t+\Delta t} - w_t) + \sum_{i=n_t+1}^{n_{t+\Delta t}} X_i\right] - 1,$$

where it is understood that a summation over an empty set is zero and $X_i = \ln(J_i)$. For a small $\Delta t$, we may use the approximation $e^x \approx 1 + x + x^2/2$ and the result $(\Delta w_t)^2 \approx \Delta t$ discussed in Section 6.3 to obtain

$$\frac{P_{t+\Delta t} - P_t}{P_t} \approx \left(\mu - \frac{1}{2}\sigma^2\right) \Delta t + \sigma \, \Delta w_t + \sum_{i=n_t+1}^{n_{t+\Delta t}} J_i + \frac{1}{2}\sigma^2 (\Delta w_t)^2$$

$$\approx \mu \, \Delta t + \sigma \epsilon \sqrt{\Delta t} + \sum_{i=n_t+1}^{n_{t+\Delta t}} X_i,$$

where $\Delta w_t = w_{t+\Delta t} - w_t$ and $\epsilon$ is a standard normal random variable.

Under the assumption of Poisson process, the probability of having one jump in the time interval $(t, t + \Delta t]$ is $\lambda \, \Delta t$ and that of having more than one jump is $o(\Delta t)$, where the symbol $o(\Delta t)$ means that if we divide this term by $\Delta t$ then its value tends to zero as $\Delta t$ tends to zero. Therefore, for a small $\Delta t$, by ignoring multiple jumps, we have

$$\sum_{i=n_t+1}^{n_{t+\Delta t}} X_i \approx \begin{cases} X_{n_t+1} & \text{with probability } \lambda \, \Delta t \\ 0 & \text{with probability } 1 - \lambda \, \Delta t. \end{cases}$$

Combining the prior results, we see that the simple return of the underlying asset is approximately distributed as

$$\frac{P_{t+\Delta t} - P_t}{P_t} \approx \mu \, \Delta t + \sigma \epsilon \sqrt{\Delta t} + I \times X, \tag{6.30}$$

where $I$ is a Bernoulli random variable with $Pr(I = 1) = \lambda \, \Delta t$ and $Pr(I = 0) = 1 - \lambda \, \Delta t$, and $X$ is a double exponential random variable defined in Eq. (6.28). Equation (6.30) reduces to that of a geometric Brownian motion without jumps.

Let $G = \mu \, \Delta t + \sigma \epsilon \sqrt{\Delta t} + I \times X$ be the random variable in the right-hand side of Eq. (6.30). Using the independence between the exponential and normal distributions used in the model, Kou (2000) obtains the probability density function of $G$ as

$$g(x) = \frac{\lambda \, \Delta t}{2\eta} e^{\sigma^2 \, \Delta t/(2\eta^2)} \left\{ e^{-\omega/\eta} \Phi\left(\frac{\omega\eta - \sigma^2 \, \Delta t}{\sigma\eta\sqrt{\Delta t}}\right) + e^{\omega/\eta} \Phi\left(\frac{\omega\eta + \sigma^2 \, \Delta t}{\sigma\eta\sqrt{\Delta t}}\right) \right\}$$

$$+ (1 - \lambda \, \Delta t) \frac{1}{\sigma\sqrt{\Delta t}} f\left(\frac{x - \mu \, \Delta t}{\sigma\sqrt{\Delta t}}\right), \tag{6.31}$$

where $\omega = x - \mu \, \Delta t - \kappa$, and $f(.)$ and $\Phi(.)$ are, respectively, the probability density and cumulative distribution functions of the standard normal random variable. Furthermore,

$$E(G) = \mu \, \Delta t + \kappa\lambda \, \Delta t, \quad \text{Var}(G) = \sigma^2 \, \Delta t + \lambda \, \Delta t[2\eta^2 + \kappa^2(1 - \lambda \, \Delta t)].$$

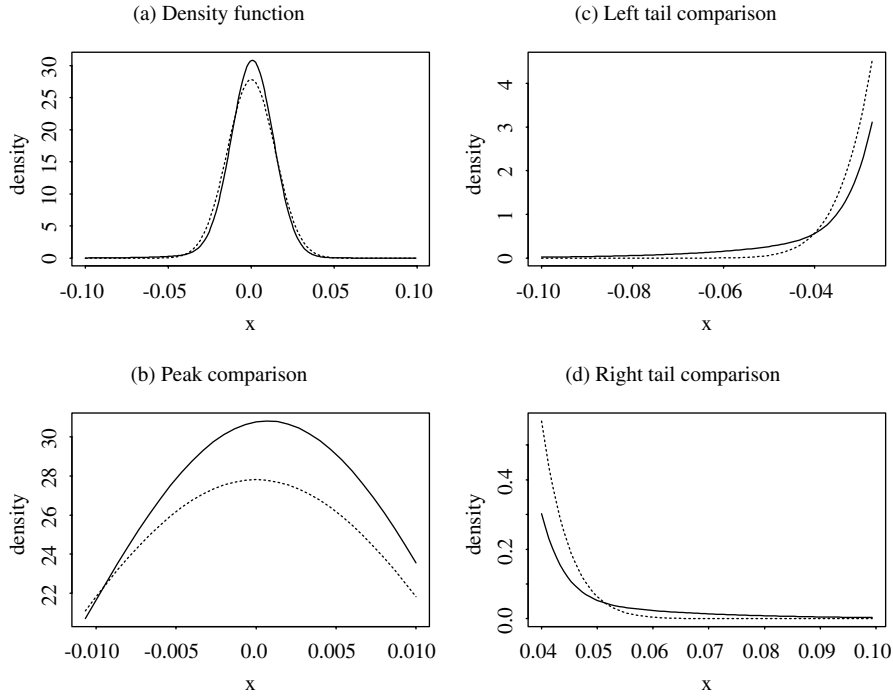


**Figure 6.9.** Density comparisons between a normal distribution and the distribution of Eq. (6.31). The dotted line denotes the normal distribution. Both distributions have mean zero and variance $2.0572 \times 10^{-4}$. (a) Overall comparison, (b) comparison of the peaks, (c) left tails, and (d) right tails.

Figure 6.9 shows some comparisons between probability density functions of a normal distribution and the distribution of Eq. (6.31). Both distributions have mean zero and variance $2.0572 \times 10^{-4}$. The mean and variance are obtained by assuming that the return of the underlying asset satisfies $\mu = 20\%$ per annum, $\sigma = 20\%$ per annum, $\Delta t = 1$ day $= 1/252$ years, $\lambda = 10$, $\kappa = -0.02$, and $\eta = 0.02$. In other words, we assume that there are about 10 daily jumps per year with average jump size $-2\%$, and the jump volatility is 2%. These values are reasonable for a U.S. stock. From the plots, the leptokurtic feature of the distribution derived from the jump diffusion process in Eq. (6.26) is clearly shown. The distribution has a higher peak and fatter tails than the corresponding normal distribution.

### 6.9.1 Option Pricing under Jump Diffusion

In the presence of random jumps, the market becomes incomplete. In this case, the standard hedging arguments are not applicable to price an option. But we can still derive an option pricing formula that does not depend on attitudes toward risk by assuming that the number of securities available is very large so that the risk of the

sudden jumps is diversifiable and the market will therefore pay no risk premium over the risk-free rate for bearing this risk. Alternatively, for a given set of risk premiums, one can consider a risk-neutral measure $P^*$ such that

$$\begin{aligned}\frac{dP_t}{P_t} &= [r - \lambda E(J-1)]dt + \sigma\, dw_t + d\left[\sum_{i=1}^{n_t}(J_i - 1)\right] \\ &= (r - \lambda\psi)dt + \sigma\, dw_t + d\left[\sum_{i=1}^{n_t}(J_i - 1)\right],\end{aligned}$$

where $r$ is the risk-free interest rate, $J = \exp(X)$ such that $X$ follows the double exponential distribution of Eq. (6.27), $\psi = e^{\kappa}/(1-\eta^2) - 1$, $0 < \eta < 1$, and the parameters $\kappa$, $\eta$, $\psi$, and $\sigma$ become risk-neutral parameters taking consideration of the risk premiums; see Kou (2000) for more details. The unique solution of the prior equation is given by

$$P_t = P_0 \exp\left[\left(r - \frac{\sigma^2}{2} - \lambda\psi\right)t + \sigma w_t\right]\prod_{i=1}^{n_t} J_i.$$

To price a European option in the jump diffusion model, it remains to compute the expectation, under the measure $P^*$, of the discounted final payoff of the option. In particular, the price of a European call option at time $t$ is given by

$$\begin{aligned}c_t &= E_*[e^{-r(T-t)}(P_T - K)_+] \\ &= E_*\left[e^{-r(T-t)}\left(P_t \exp\left[\left(\frac{r-\sigma^2}{2-\lambda\psi}\right)(T-t) + \sigma\sqrt{T-t}\epsilon\right]\prod_{i=1}^{n_T} J_i - K\right)_+\right],\end{aligned} \tag{6.32}$$

where $T$ is the expiration time, $(T-t)$ is the time to expiration measured in years, $K$ is the strike price, $(y)_+ = \max(0, y)$, and $\epsilon$ is a standard normal random variable. Kou (2000) shows that $c_t$ is analytically tractable as

$$\begin{aligned}c_t = &\sum_{n=1}^{\infty}\sum_{j=1}^{n} e^{-\lambda(T-t)}\frac{\lambda^n(T-t)^n}{n!}\frac{2^j}{2^{2n-1}}\binom{2n-j-1}{n-1} \\ &\times (A_{1,n,j} + A_{2,n,j} + A_{3,n,j}) \\ &+ e^{-\lambda(T-t)}\left[P_t e^{-\lambda\psi(T-t)}\Phi(h_+) - K e^{-r(T-t)}\Phi(h_-)\right],\end{aligned} \tag{6.33}$$

where $\Phi(.)$ is the CDF of the standard normal random variable,

$$A_{1,n,j} = P_t e^{-\lambda\psi(T-t)+n\kappa}\frac{1}{2}\left[\frac{1}{(1-\eta)^j} + \frac{1}{(1+\eta)^j}\right]\Phi(b_+) - e^{-r(T-t)}K\Phi(b_-)$$

$$A_{2,n,j} = \frac{1}{2} e^{-r(T-t)-\omega/\eta+\sigma^2(T-t)/(2\eta^2)} K \times \sum_{i=0}^{j-1} \left[ \frac{1}{(1-\eta)^{j-i}} - 1 \right] \left( \frac{\sigma\sqrt{T-t}}{\eta} \right)^i \frac{1}{\sqrt{2\pi}} Hh_i(c_-)$$

$$A_{3,n,j} = \frac{1}{2} e^{-r(T-t)+\omega/\eta+\sigma^2(T-t)/(2\eta^2)} K \times \sum_{i=0}^{j-1} \left[ 1 - \frac{1}{(1+\eta)^{j-i}} \right] \left( \frac{\sigma\sqrt{T-t}}{\eta} \right)^i \frac{1}{\sqrt{2\pi}} Hh_i(c_+)$$

$$b_\pm = \frac{\ln(P_t/K) + (r \pm \sigma^2/2 - \lambda\psi)(T-t) + n\kappa}{\sigma\sqrt{T-t}}$$

$$h_\pm = \frac{\ln(P_t/K) + (r \pm \sigma^2/2 - \lambda\psi)(T-t)}{\sigma\sqrt{T-t}}$$

$$c_\pm = \frac{\sigma\sqrt{T-t}}{\eta} \pm \frac{\omega}{\sigma\sqrt{T-t}}$$

$$\omega = \ln(K/P_t) + \lambda\psi(T-t) - (r - \sigma^2/2)(T-t) - n\kappa$$

$$\psi = \frac{e^\kappa}{1-\eta^2} - 1,$$

and the $Hh_i(.)$ functions are defined as

$$Hh_n(x) = \frac{1}{n!} \int_x^\infty (s-x)^n e^{-s^2/2} ds, \quad n = 0, 1, \ldots \tag{6.34}$$

and $Hh_{-1}(x) = \exp(-x^2/2)$, which is $\sqrt{2\pi} f(x)$ with $f(x)$ being the probability density function of a standard normal random variable; see Abramowitz and Stegun (1972). The $Hh_n(x)$ functions satisfy the recursion

$$nHh_n(x) = Hh_{n-2}(x) - xHh_{n-1}(x), \quad n \geq 1, \tag{6.35}$$

with starting values $Hh_{-1}(x) = e^{-x^2/2}$ and $Hh_0(x) = \sqrt{2\pi}\Phi(-x)$.

The pricing formula involves an infinite series, but its numerical value can be approximated quickly and accurately through truncation (e.g., the first 10 terms). Also, if $\lambda = 0$ (i.e., there are no jumps), then it is easily seen that $c_t$ reduces to the Black–Scholes formula for a call option discussed before.

Finally, the price of a European put option under the jump diffusion model considered can be obtained by using the put-call parity—that is,

$$p_t = c_t + Ke^{-r(T-t)} - P_t.$$

Pricing formulas for other options under the jump diffusion model in Eq. (6.26) can be found in Kou (2000).

**Example 6.7.** Consider the stock of Example 6.6, which has a current price of \$80. As before, assume that the strike price of a European option is $K = \$85$ and other parameters are $r = 0.08$ and $T - t = 0.25$. In addition, assume that the price of the stock follows the jump diffusion model in Eq. (6.26) with parameters $\lambda = 10$, $\kappa = -0.02$, and $\eta = 0.02$. In other words, there are about 10 jumps per year with average jump size $-2\%$ and jump volatility 2%. Using the formula in Eq. (6.33), we obtain $c_t = \$2.25$, which is higher than \$1.86 of Example 6.6 when there are no jumps. The corresponding put option assumes the value $p_t = \$5.57$, which is also higher than what we had before. As expected, adding the jumps while keeping the other parameters fixed increases the prices of both European options. Keep in mind, however, that adding the jump process to the stock price in a real application often leads to different estimates for the stock volatility $\sigma$.

## 6.10 ESTIMATION OF CONTINUOUS-TIME MODELS

Next we consider the problem of estimating directly the diffusion equation (i.e., the Ito's process) from discretely-sampled data. Here the drift and volatility functions $\mu(x_t, t)$ and $\sigma(x_t, t)$ are time-varying and may not follow a specific parametric form. This is a topic of considerable interest in recent years. Details of the available methods are beyond the scope of this chapter. Hence, we only outline the approaches proposed in the literature. Interested readers can consult the corresponding references and Lo (1988).

There are several approaches available for estimating a diffusion equation. The first approach is the quasi-maximum likelihood approach, which makes use of the fact that for a small time interval $dw_t$ is normally distributed; see Kessler (1997) and the references therein. The second approach uses methods of moments; see Conley, Hansen, Luttmer, and Scheinkman (1997) and the references therein. The third approach uses nonparametric methods; see Ait-Sahalia (1996, 1997). The fourth approach uses semiparametric and reprojection methods; see Gallant and Long (1997) and Gallant and Tauchen (1997). Recently, many researchers have applied Markov Chain Monte Carlo methods to estimate the diffusion equation; see Eraker (2001) and Elerian, Chib, and Shephard (2001).

## APPENDIX A. INTEGRATION OF BLACK–SCHOLES FORMULA

In this appendix, we derive the price of a European call option given in Eq. (6.19). Let $x = \ln(P_T)$. By changing variable and using $g(P_T)dP_T = f(x)\,dx$, where $f(x)$ is the probability density function of $x$, we have

$$
\begin{aligned}
c_t &= \exp[-r(T-t)] \int_K^\infty (P_T - K) g(P_T) dP_t \\
&= e^{-r(T-t)} \int_{\ln(K)}^\infty (e^x - K) f(x)\, dx \\
&= e^{-r(T-t)} \left[ \int_{\ln(K)}^\infty e^x f(x)\, dx - K \int_{\ln(K)}^\infty f(x)\, dx \right]. \qquad (6.36)
\end{aligned}
$$

Because $x = \ln(P_T) \sim N[\ln(P_t) + (r - \sigma^2/2)(T-t), \sigma^2(T-t)]$, the integration of the second term of Eq. (6.36) reduces to

$$
\begin{aligned}
\int_{\ln(K)}^\infty f(x)\, dx &= 1 - \int_{-\infty}^{\ln(K)} f(x)\, dx \\
&= 1 - \text{CDF}(\ln(K)) \\
&= 1 - \Phi(-h_-) = \Phi(h_-),
\end{aligned}
$$

where CDF $\ln(K))$ is the cumulative distribution function (CDF) of $x = \ln(P_T)$ evaluated at $\ln(K)$, $\Phi(.)$ is the CDF of the standard normal random variable, and

$$
\begin{aligned}
-h_- &= \frac{\ln(K) - \ln(P_t) - (r - \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}} \\
&= \frac{-\ln(P_t/K) - (r - \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}.
\end{aligned}
$$

The integration of the first term of Eq. (6.36) can be written as

$$
\int_{\ln(K)}^\infty \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2(T-t)}} \exp\left[ x - \frac{[x - \ln(P_t) - (r - \sigma^2/2)(T-t)]^2}{2\sigma^2(T-t)} \right] dx,
$$

where the exponent can be simplified to

$$
\begin{aligned}
& x - \frac{[x - \{\ln(P_t) + (r - \sigma^2/2)(T-t)\}]^2}{2\sigma^2(T-t)} \\
&\quad = -\frac{[x - \{\ln(P_t) + (r + \sigma^2/2)(T-t)\}]^2}{2\sigma^2(T-t)} + \ln(P_t) + r(T-t).
\end{aligned}
$$

Consequently, the first integration becomes

$$
\begin{aligned}
& \int_{\ln(K)}^\infty e^x f(x)\, dx \\
&= P_t e^{r(T-t)} \int_{\ln(K)}^\infty \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2(T-t)}} \exp\left[ -\frac{[x - \{\ln(P_t) + (r + \sigma^2/2)(T-t)\}]^2}{2\sigma^2(T-t)} \right] dx,
\end{aligned}
$$

which involves the CDF of a normal distribution with mean $\ln(P_t) + (r + \sigma^2/2)(T - t)$ and variance $\sigma^2(T - t)$. By using the same techniques as those of the second intergation shown before, we have

$$\int_{\ln(K)}^{\infty} e^x f(x)\, dx = P_t e^{r(T-t)} \Phi(h_+),$$

where $h_+$ is given by

$$h_+ = \frac{\ln(P_t/K) + (r + \sigma^2/2)(T - t)}{\sigma\sqrt{T - t}}.$$

Putting the two integration results together, we have

$$c_t = e^{-r(T-t)}[P_t e^{r(T-t)} \Phi(h_+) - K\Phi(h_-)] = P_t \Phi(h_+) - K e^{-r(T-t)} \Phi(h_-).$$

## APPENDIX B. APPROXIMATION TO STANDARD NORMAL PROBABILITY

The CDF $\Phi(x)$ of a standard normal random variable can be approximated by

$$\Phi(x) = \begin{cases} 1 - f(x)[c_1 k + c_2 k^2 + c_3 k^3 + c_4 k^4 + c_5 k^5] & \text{if } x \geq 0 \\ 1 - \Phi(-x) & \text{if } x < 0, \end{cases}$$

where $f(x) = \exp(-x^2/2)/\sqrt{2\pi}$, $k = 1/(1 + 0.2316419x)$, $c_1 = 0.319381530$, $c_2 = -0.356563782$, $c_3 = 1.781477937$, $c_4 = -1.821255978$, and $c_5 = 1.330274429$.

For illustration, using the earlier approximation, we obtain $\Phi(1.96) = 0.975002$, $\Phi(0.82) = 0.793892$, and $\Phi(-0.61) = 0.270931$. These probabilities are very close to that obtained from a typical normal probability table.

## EXERCISES

**1.** Assume that the log price $p_t = \ln(P_t)$ follows a stochastic differential equation

$$dp_t = \gamma dt + \sigma\, dw_t,$$

where $w_t$ is a Wiener process. Derive the stochastic equation for the price $P_t$.

**2.** Considering the forward price $F$ of a nondividend-paying stock, we have

$$F_{t,T} = P_t e^{r(T-t)},$$

where $r$ is the risk-free interest rate, which is constant and $P_t$ is the current stock price. Suppose $P_t$ follows the geometric Brownian motion $dP_t = \mu P_t dt + \sigma P_t\, dw_t$. Derive a stochastic differential equation for $F_{t,T}$.

**3.** Assume that the price of IBM stock follows the Ito's process

$$dP_t = \mu P_t dt + \sigma P_t \, dw_t,$$

where $\mu$ and $\sigma$ are constant and $w_t$ is a standard Brownian motion. Consider the daily log returns of IBM stock in 1997. The average return and the sample standard deviation are 0.00131 and 0.02215, respectively. Use the data to estimate the parameters $\mu$ and $\sigma$ assuming that there were 252 trading days in 1997.

**4.** Suppose that the current price of a stock is \$120 per share with volatility $\sigma = 50\%$ per annum. Suppose further that the risk-free interest rate is 7% per annum and the stock pays no dividend. (a) What is the price of a European call option contingent on the stock with a strike price of \$125 that will expire in 3 months? (b) What is the price of a European put option on the same stock with a strike price of \$118 that will expire in 3 months? If the volatility $\sigma$ is increased to 80% per annum, then what are the prices of the two options?

**5.** Derive the limiting marginal effects of the five variables $K$, $P_t$, $T-t$, $\sigma$, and $r$ on a European put option contingent on a stock.

**6.** A stock price is currently \$60 per share and follows the geometric Brownian motion $dP_t = \mu P_t \, dt + \sigma P_t \, dt$. Assume that the expected return $\mu$ from the stock is 20% per annum and its volatility is 40% per annum. What is the probability distribution for the stock price in 2 years? Obtain the mean and standard deviation of the distribution and construct a 95% confidence interval for the stock price.

**7.** A stock price is currently \$60 per share and follows the geometric Brownian motion $dP_t = \mu P_t \, dt + \sigma P_t \, dt$. Assume that the expected return $\mu$ from the stock is 20% per annum and its volatility is 40% per annum. What is the probability distribution for the continuously compounded rate of return of the stock over 2 years? Obtain the mean and standard deviation of the distribution.

**8.** Suppose that the current price of Stock A is \$70 per share and the price follows the jump diffusion model in Eq. (6.26). Assume that the risk-free interest rate is 8% per annum and the stock volatility is 30% per annum. In addition, the price on average has about 15 jumps per year with average jump size $-2\%$ and jump volatility 3%. What is the price of a European call option with strike price \$75 that will expire in 3 months? What is the price of the corresponding European put option?

## REFERENCES

Abramowitz, M., and Stegun, I. A. (1972), *Handbook of Mathematical Functions*, 10th ed., U.S. National Bureau of Standards.

Ait-Sahalia, Y. (1996), "Testing continuous-time models for the spot interest rate," *Review of Financial Studies*, 9, 385–426.

Ait-Sahalia, Y. (1997), "Maximum likelihood estimation of discretely sampled diffusions: a closed-form approach," working paper, Economics Department, Princeton University.

Bakshi, G., Cao, C., and Chen, Z. (1997), "Empirical performance of alternative option pricing models," *Journal of Finance*, 52, 2003–2049.

Billingsley, P. (1986), *Probability and Measure*, 2nd ed., Wiley: New York.

Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley: New York.

Black, F. and Scholes, M. (1973), "The pricing of options and corporate liabilities," *Journal of Political Economy*, 81, 637–654.

Conley, T. G., Hansen, L. P., Luttmer, E. G. J., and Scheinkman, J. A. (1997), "Short-term interest rates as subordinated diffusions," *Review of Financial Studies*, 10, 525–577.

Cox, J. C., and Rubinstein, M. (1985), *Options Markets*, Prentice-Hall: Englewood Cliffs, New Jersey.

Donsker, M. (1951), "An invariance principle for certain probability limit theorems," *Mem. American Mathematical Society*, No. 6.

Duffie, D. (1995), *Dynamic Asset Pricing Theory*, 2nd ed., Princeton University Press: Princeton, New Jersey.

Elerian, O., Chib, S., and Shephard, N. (2001), "Likelihood inference for discretely observed non-linear diffusions," *Econometrica*, 69, 959–993.

Eraker, B. (2001), "MCMC analysis of diffusion models with application to finance," *Journal of Business & Economic Statistics*, 19, 177–191.

Gallant, A. R., and Long, J. R. (1997), "Estimating stochastic diffusion equations efficiently by minimum chi-squared, " *Biometrika*, 84, 125–141.

Gallant, A. R., and Tauchen, G. (1997), "The relative efficiency of method of moments estimators," Working paper, Economics Department, University of North Carolina.

Hull, J. C. (1997), *Options, Futures, and Other Derivatives*, 3rd ed. Prentice-Hall: Upper Saddle River, New Jersey.

Kessler, M. (1997), "Estimation of an ergodic diffusion from discrete observations," *Scandinavian Journal of Statistics*, 24, 1–19.

Kou, S. (2000), "A jump diffusion model for option pricing with three properties: Leptokurtic feature, volatility smile, and analytic tractability," working paper, Columbia University.

Lo, A. W. (1988), "Maximum likelihood estimation of generalized Ito's processes with discretely sampled data," *Econometric Theory*, 4, 231–247.

Merton, R. C. (1976), "Option pricing when the underlying stock returns are discontinuous," *Journal of Financial Economics*, 5, 125–144.

*Analysis of Financial Time Series.* Ruey S. Tsay

ISBN: 0-471-41544-8

CHAPTER 7

# Extreme Values, Quantile Estimation, and Value at Risk

Extreme price movements in the financial markets are rare, but important. The stock market crash on Wall Street in October 1987 and other big financial crises such as the Long Term Capital Management have attracted a great deal of attention among practitioners and researchers, and some people even called for government regulations on the derivative markets. In recent years, the seemingly large daily price movements in high-tech stocks have further generated discussions on market risk and margin setting for financial institutions. As a result, value at risk (VaR) has become a widely used measure of market risk in risk management.

In this chapter, we discuss various methods for calculating VaR and the statistical theories behind these methods. In particular, we consider the extreme value theory developed in the statistical literature for studying rare (or extraordinary) events and its application to VaR. Both unconditional and conditional concepts of extreme values are discussed. The unconditional approach to VaR calculation for a financial position uses the historical returns of the instruments involved to compute VaR. However, a conditional approach uses the historical data and explanatory variables to calculate VaR.

Other approaches to VaR calculation discussed in the chapter are RiskMetrics, econometric modeling using volatility models, and empirical quantile. We use daily log returns of IBM stock to illustrate the actual calculation of all the methods discussed. The results obtained can therefore be used to compare the performance of different methods. Figure 7.1 shows the time plot of daily log returns of IBM stock from July 3, 1962 to December 31, 1998 for 9190 observations.

## 7.1 VALUE AT RISK

There are several types of risk in financial markets. Credit risk, liquidity risk, and market risk are three examples. Value at risk (VaR) is mainly concerned with market risk. It is a single estimate of the amount by which an institution's position in a risk category could decline due to general market movements during a given holding

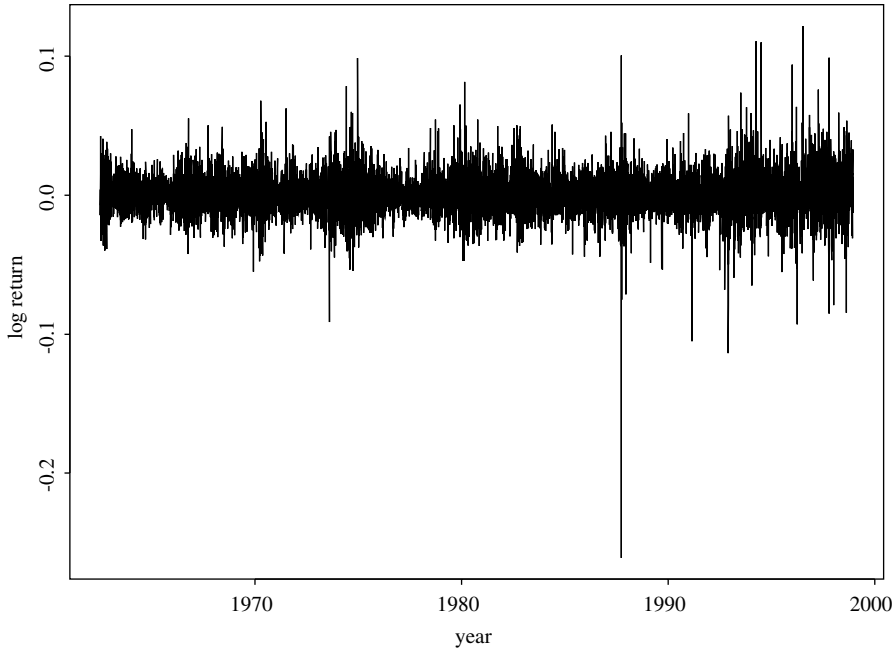


**Figure 7.1.** Time plot of daily log returns of IBM stock from July 3, 1962 to December 31, 1998.

period; see Duffie and Pan (1997) and Jorion (1997) for a general exposition of VaR. 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. From the viewpoint of a financial institution, VaR can be defined as the maximal loss of a financial position during a given time period for a given probability. In this view, one treats VaR as a measure of loss associated with a rare (or extraordinary) event under normal market conditions. Alternatively, from the viewpoint of a regulatory committee, VaR can be defined as the minimal loss under extraordinary market circumstances. Both definitions will lead to the same VaR measure, even though the concepts appear to be different.

In what follows, we define VaR under a probabilistic framework. Suppose that at the time index $t$ we are interested in the risk of a financial position for the next $\ell$ periods. Let $\Delta V(\ell)$ be the change in value of the assets in the financial position from time $t$ to $t + \ell$. This quantity is measured in dollars and is a random variable at the time index $t$. Denote the cumulative distribution function (CDF) of $\Delta V(\ell)$ by $F_\ell(x)$. We define the VaR of a long position over the time horizon $\ell$ with probability $p$ as

$$p = \Pr[\Delta V(\ell) \leq \text{VaR}] = F_\ell(\text{VaR}). \tag{7.1}$$

Since the holder of a long financial position suffers a loss when $\Delta V(\ell) < 0$, the VaR defined in Eq. (7.1) typically assumes a negative value when $p$ is small. The negative sign signifies a loss. From the definition, the probability that the holder would encounter a loss greater than or equal to VaR over the time horizon $\ell$ is $p$. Alternatively, VaR can be interpreted as follows. With probability $(1 - p)$, the potential loss encountered by the holder of the financial position over the time horizon $\ell$ is less than or equal to VaR.

The holder of a short position suffers a loss when the value of the asset increases [i.e., $\Delta V(\ell) > 0$]. The VaR is then defined as

$$p = \Pr[\Delta V(\ell) \geq \text{VaR}] = 1 - \Pr[\Delta V(\ell) \leq \text{VaR}] = 1 - F_\ell(\text{VaR}).$$

For a small $p$, the VaR of a short position typically assumes a positive value. The positive sign signifies a loss.

The previous definitions show that VaR is concerned with tail behavior of the CDF $F_\ell(x)$. For a long position, the left tail of $F_\ell(x)$ is important. Yet a short position focuses on the right tail of $F_\ell(x)$. Notice that the definition of VaR in Eq. (7.1) continues to apply to a short position if one uses the distribution of $-\Delta V(\ell)$. Therefore, it suffices to discuss methods of VaR calculation using a long position.

Fo any univariate CDF $F_\ell(x)$ and probability $p$, such that $0 < p < 1$, the quantity

$$x_p = \inf\{x \mid F_\ell(x) \geq p\}$$

is called the $p$th quantile of $F_\ell(x)$, where inf denotes the smallest real number satisfying $F_\ell(x) \geq p$. If the CDF $F_\ell(x)$ of Eq. (7.1) is known, then VaR is simply its $p$th quantile (i.e., VaR $= x_p$). The CDF is unknown in practice, however. Studies of VaR are essentially concerned with estimation of the CDF and/or its quantile, especially the tail behavior of the CDF.

In practical applications, calculation of VaR involves several factors:

1. The probability of interest $p$, such as $p = 0.01$ or $p = 0.05$.
2. The time horizon $\ell$. It might be set by a regulatory committee, such as 1 day or 10 days.
3. The frequency of the data, which might not be the same as the time horizon $\ell$. Daily observations are often used.
4. The CDF $F_\ell(x)$ or its quantiles.
5. The amount of the financial position or the mark-to-market value of the portfolio.

Among these factors, the CDF $F_\ell(x)$ is the focus of econometric modeling. Different methods for estimating the CDF give rise to different approaches to VaR calculation.

***Remark:*** The definition of VaR in Eq. (7.1) is in dollar amount. Since log returns correspond approximately to percentage changes in value of a financial position,

we use log returns $r_t$ in data analysis. The VaR calculated from the quantile of the distribution of $r_{t+1}$ given information available at time $t$ is therefore in percentage. The dollar amount of VaR is then the cash value of the financial position times the VaR of the log return series.

***Remark:*** VaR is a prediction concerning possible loss of a portfolio in a given time horizon. It should be computed using the *predictive distribution* of future returns of the financial position. For example, the VaR for a 1-day horizon of a portfolio using daily returns $r_t$ should be calculated using the predictive distribution of $r_{t+1}$ given information available at time $t$. From a statistical viewpoint, predictive distribution takes into account the parameter uncertainty in a properly specified model. However, predictive distribution is hard to obtain, and most of the available methods for VaR calculation ignore the effects of parameter uncertainty.

## 7.2 RISKMETRICS

J.P. Morgan developed the RiskMetrics$^{TM}$ methodology to VaR calculation; see Longerstaey and More (1995). In its simple form, RiskMetrics assumes that the continuously compounded daily return of a portfolio follows a conditional normal distribution. Denote the daily log return by $r_t$ and the information set available at time $t-1$ by $F_{t-1}$. RiskMetrics assumes that $r_t \mid F_{t-1} \sim N(\mu_t, \sigma_t^2)$, where $\mu_t$ is the conditional mean and $\sigma_t^2$ is the conditional variance of $r_t$. In addition, the method assumes that the two quantities evolve over time according to the simple model:

$$\mu_t = 0, \quad \sigma_t^2 = \alpha\sigma_{t-1}^2 + (1-\alpha)r_{t-1}^2, \quad 1 > \alpha > 0. \tag{7.2}$$

Therefore, the method assumes that the logarithm of the daily price, $p_t = \ln(P_t)$, of the portfolio satisfies the difference equation $p_t - p_{t-1} = a_t$, where $a_t = \sigma_t\epsilon_t$ is an IGARCH(1, 1) process without a drift. The value of $\alpha$ is often in the interval (0.9, 1).

A nice property of such a special random-walk IGARCH model is that the conditional distribution of a multiperiod return is easily available. Specifically, for a $k$-period horizon, the log return from time $t+1$ to time $t+k$ (inclusive) is $r_t[k] = r_{t+1} + \cdots + r_{t+k-1} + r_{t+k}$. We use the square bracket $[k]$ to denote a $k$-horizon return. Under the special IGARCH(1,1) model in Eq. (7.2), the conditional distribution $r_t[k] \mid F_t$ is normal with mean zero and variance $\sigma_t^2[k]$, where $\sigma_t^2[k]$ can be computed using the forecasting method discussed in Chapter 3. Using the independence assumption of $\epsilon_t$ and model (7.2), we have

$$\sigma_t^2[k] = \text{Var}(r_t[k] \mid F_t) = \sum_{i=1}^{k} \text{Var}(a_{t+i} \mid F_t),$$

where $\text{Var}(a_{t+i} \mid F_t) = E(\sigma_{t+i}^2 \mid F_t)$ can be obtained recursively. Using $r_{t-1} = a_{t-1} = \sigma_{t-1}\epsilon_{t-1}$, we can rewrite the volatility equation of the IGARCH(1, 1) model

in Eq. (7.2) as

$$\sigma_t^2 = \sigma_{t-1}^2 + (1-\alpha)\sigma_{t-1}^2(\epsilon_{t-1}^2 - 1) \quad \text{for all } t.$$

In particular, we have

$$\sigma_{t+i}^2 = \sigma_{t+i-1}^2 + (1-\alpha)\sigma_{t+i-1}^2(\epsilon_{t+i-1}^2 - 1) \quad \text{for} \quad i = 2, \ldots, k.$$

Since $E(\epsilon_{t+i-1}^2 - 1 \mid F_t) = 0$ for $i \geq 2$, the prior equation shows that

$$E(\sigma_{t+i}^2 \mid F_t) = E(\sigma_{t+i-1}^2 \mid F_t) \quad \text{for} \quad i = 2, \ldots, k. \tag{7.3}$$

For the 1-step ahead volatility forecast, Eq. (7.2) shows that $\sigma_{t+1}^2 = \alpha\sigma_t^2 + (1-\alpha)r_t^2$. Therefore, Eq. (7.3) shows that $\text{Var}(r_{t+i} \mid F_t) = \sigma_{t+1}^2$ for $i \geq 1$ and hence, $\sigma_t^2[k] = k\sigma_{t+1}^2$. The results show that $r_t[k] \mid F_t \sim N(0, k\sigma_{t+1}^2)$. Consequently, under the special IGARCH(1, 1) model in Eq. (7.2) the conditional variance of $r_t[k]$ is proportional to the time horizon $k$. The conditional standard deviation of a $k$-period horizon log return is then $\sqrt{k}\sigma_{t+1}$.

Suppose that the financial position is a long position so that loss occurs when there is a big price drop (i.e., a large negative return). If the probability is set to 5%, then RiskMetrics uses $1.65\sigma_{t+1}$ to measure the risk of the portfolio—that is, it uses the one-sided 5% quantile of a normal distribution with mean zero and standard deviation $\sigma_{t+1}$. The actual 5% quantile is $-1.65\sigma_{t+1}$, but the negative sign is ignored with the understanding that it signifies a loss. Consequently, if the standard deviation is measured in percentage, then the daily VaR of the portfolio under RiskMetrics is

$$\text{VaR} = \text{Amount of Position} \times 1.65\sigma_{t+1},$$

and that of a $k$-day horizon is

$$\text{VaR}(k) = \text{Amount of Position} \times 1.65\sqrt{k}\sigma_{t+1},$$

where the argument $(k)$ of VaR is used to denote the time horizon. Consequently, under RiskMetrics, we have

$$\text{VaR}(k) = \sqrt{k} \times \text{VaR}.$$

This is referred to as the *square root of time rule* in VaR calculation under RiskMetrics.

**Example 7.1.** The sample standard deviation of the continuously compounded daily return of the German Mark/U.S. Dollar exchange rate was about 0.53% in June 1997. Suppose that an investor was long in \$10 million worth of Mark/Dollar exchange rate contract. Then the 5% VaR for a 1-day horizon of the

investor is

$$\$10{,}000{,}000 \times (1.65 \times 0.0053) = \$87{,}450.$$

The corresponding VaR for 1-month horizon (30 days) is

$$\$10{,}000{,}000 \times (\sqrt{30} \times 1.65 \times 0.0053) \approx \$478{,}983.$$

**Example 7.2.** Consider the daily IBM log returns of Figure 7.1. As mentioned in Chapter 1, the sample mean of the returns is significantly different from zero. However, for demonstration of VaR calculation using RiskMetrics, we assume in this example that the conditional mean is zero and the volatility of the returns follows an IGARCH(1, 1) model without a drift. The fitted model is

$$r_t = a_t, \quad a_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = 0.9396\sigma_{t-1}^2 + (1 - 0.9396)a_{t-1}^2, \tag{7.4}$$

where $\{\epsilon_t\}$ is a standard Gaussian white noise series. As expected, this model is rejected by the Q-statistics. For instance, we have a highly significant statistic $Q(10) = 56.19$ for the squared standardized residuals.

From the data and the fitted model, we have $r_{9190} = -0.0128$ and $\hat{\sigma}_{9190}^2 = 0.0003472$. Therefore, the 1-step ahead volatility forecast is $\hat{\sigma}_{9190}^2(1) = 0.000336$. The 5% quantile of the conditional distribution $r_{9191} \mid F_{9190}$ is $-1.65 \times \sqrt{0.000336} = -0.03025$, where it is understood that the negative sign signifies a loss. Consequently, the 1-day horizon 5% VaR of a long position of \$10 million is

$$\text{VaR} = \$10{,}000{,}000 \times 0.03025 = \$302{,}500.$$

The 1% quantile is $-2.3262 \times \sqrt{0.000336} = -0.04265$, and the corresponding 1% VaR for the same long position is \$426,500.

### 7.2.1 Discussion

An advantage of RiskMetrics is simplicity. It is easy to understand and apply. Another advantage is that it makes risk more transparent in the financial markets. However, as security returns tend to have heavy tails (or fat tails), the normality assumption used often results in underestimation of VaR. Other approaches to VaR calculation avoid making such an assumption.

The square root of time rule is a consequence of the special model used by RiskMetrics. If either the zero mean assumption or the special IGARCH(1, 1) model assumption of the log returns fails, then the rule is invalid. Consider the simple model:

$$\begin{aligned} r_t &= \mu + a_t, \quad a_t = \sigma_t \epsilon_t, \quad \mu \neq 0 \\ \sigma_t^2 &= \alpha\sigma_{t-1}^2 + (1 - \alpha)a_{t-1}^2, \end{aligned}$$

where $\{\epsilon_t\}$ is a standard Gaussian white noise series. The assumption that $\mu \neq 0$ holds for returns of many heavily traded stocks on the NYSE; see Chapter 1. For this simple model, the distribution of $r_{t+1}$ given $F_t$ is $N(\mu, \sigma_{t+1}^2)$. The 5% quantile used to calculate the 1-period horizon VaR becomes $\mu - 1.65\sigma_{t+1}$. For a $k$-period horizon, the distribution of $r_t[k]$ given $F_t$ is $N(k\mu, k\sigma_{t+1}^2)$, where as before $r_t[k] = r_{t+1} + \cdots + r_{t+k}$. The 5% quantile used in $k$-period horizon VaR calculation is $k\mu - 1.65\sqrt{k}\sigma_{t+1} = \sqrt{k}(\sqrt{k}\mu - 1.65\sigma_{t+1})$. Consequently, $\text{VaR}(k) \neq \sqrt{k} \times \text{VaR}$ when the mean return is not zero. It is also easy to show that the rule fails when the volatility model of the return is not an IGARCH(1, 1) model without a drift.

### 7.2.2 Multiple Positions

In some applications, an investor may hold multiple positions and needs to compute the overall VaR of the positions. RiskMetrics adopts a simple approach for doing such a calculation under the assumption that daily log returns of each position follow a random-walk IGARCH(1, 1) model. The additional quantities needed are the cross-correlation coefficients between the returns. Consider the case of two positions. Let $\text{VaR}_1$ and $\text{VaR}_2$ be the VaR for the two positions and $\rho_{12}$ be the cross-correlation coefficient between the two returns—that is,

$$\rho_{12} = \text{Cov}(r_{1t}, r_{2t})/[\text{Var}(r_{1t})\text{Var}(r_{2t})]^{0.5}.$$

Then the overall VaR of the investor is

$$\text{VaR} = \sqrt{\text{VaR}_1^2 + \text{VaR}_2^2 + 2\rho_{12}\text{VaR}_1\text{VaR}_2}.$$

The generalization of VaR to a position consisting of $m$ instruments is straightforward as

$$\text{VaR} = \sqrt{\sum_{i=1}^{m} \text{VaR}_i^2 + 2\sum_{i<j}^{m} \rho_{ij}\text{VaR}_i\text{VaR}_j},$$

where $\rho_{ij}$ is the cross-correlation coefficient between returns of the $i$th and $j$th instruments and $\text{VaR}_i$ is the VaR of the $i$th instrument.

## 7.3 AN ECONOMETRIC APPROACH TO VAR CALCULATION

A general approach to VaR calculation is to use the time-series econometric models of Chapters 2 to 4. For a log return series, the time series models of Chapter 2 can be used to model the mean equation, and the conditional heteroscedastic models of Chapter 3 or 4 are used to handle the volatility. For simplicity, we use GARCH models in our discussion and refer to the approach as an *econometric approach* to VaR calculation. Other volatility models, including the nonlinear ones in Chapter 4, can also be used.

Consider the log return $r_t$ of an asset. A general time series model for $r_t$ can be written as

$$r_t = \phi_0 + \sum_{i=1}^{p} \phi_i r_{t-i} + a_t - \sum_{j=1}^{q} \theta_j a_{t-j} \tag{7.5}$$

$$a_t = \sigma_t \epsilon_t$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{u} \alpha_i a_{t-i}^2 + \sum_{j=1}^{v} \beta_j \sigma_{t-j}^2. \tag{7.6}$$

Equations (7.5) and (7.6) are the mean and volatility equations for $r_t$. These two equations can be used to obtain 1-step ahead forecasts of the conditional mean and conditional variance of $r_t$ assuming that the parameters are known. Specifically, we have

$$\hat{r}_t(1) = \phi_0 + \sum_{i=1}^{p} \phi_i r_{t+1-i} - \sum_{j=1}^{q} \theta_j a_{t+1-j}$$

$$\hat{\sigma}_t^2(1) = \alpha_0 + \sum_{i=1}^{u} \alpha_i a_{t+1-i}^2 + \sum_{j=1}^{v} \beta_j \sigma_{t+1-j}^2.$$

If one further assumes that $\epsilon_t$ is Gaussian, then the conditional distribution of $r_{t+1}$ given the information available at time $t$ is $N[\hat{r}_t(1), \hat{\sigma}_t^2(1)]$. Quantiles of this conditional distribution can easily be obtained for VaR calculation. For example, the 5% quantile is $\hat{r}_t(1) - 1.65\hat{\sigma}_t(1)$. If one assumes that $\epsilon_t$ is a standardized Student-$t$ distribution with $v$ degrees of freedom, then the quantile is $\hat{r}_t(1) - t_v^*(p)\hat{\sigma}_t(1)$, where $t_v^*(p)$ is the $p$th quantile of a standardized Student-$t$ distribution with $v$ degrees of freedom.

The relationship between quantiles of a Student-$t$ distribution with $v$ degrees of freedom, denoted by $t_v$, and those of its standardized distribution, denoted by $t_v^*$, is

$$p = \Pr(t_v \le q) = \Pr\left(\frac{t_v}{\sqrt{v/(v-2)}} \le \frac{q}{\sqrt{v/(v-2)}}\right) = \Pr\left(t_v^* \le \frac{q}{\sqrt{v/(v-2)}}\right),$$

where $v > 2$. That is, if $q$ is the $p$th quantile of a Student-$t$ distribution with $v$ degrees of freedom, then $q/\sqrt{v/(v-2)}$ is the $p$th quantile of a standardized Student-$t$ distribution with $v$ degrees of freedom. Therefore, if $\epsilon_t$ of the GARCH model in Eq. (7.6) is a standardized Student-$t$ distribution with $v$ degrees of freedom and the probability is $p$, then the quantile used to calculate the 1-period horizon VaR at time index $t$ is

$$\hat{r}_t(1) - \frac{t_v(p)\hat{\sigma}_t(1)}{\sqrt{v/(v-2)}},$$

where $t_v(p)$ is the $p$th quantile of a Student-$t$ distribution with $v$ degrees of freedom.

**Example 7.3.** Consider again the daily IBM log returns of Example 7.2. We use two volatility models to calculate VaR of 1-day horizon at $t = 9190$ for a long position of \$10 million. These econometric models are reasonable based on the modeling techniques of Chapters 2 and 3.

***Case 1***

Assume that $\epsilon_t$ is standard normal. The fitted model is

$$r_t = 0.00066 - 0.0247 r_{t-2} + a_t, \quad a_t = \sigma_t \epsilon_t$$
$$\sigma_t^2 = 0.00000389 + 0.0799 a_{t-1}^2 + 0.9073 \sigma_t^2.$$

From the data, we have $r_{9189} = -0.00201$, $r_{9190} = -0.0128$, and $\sigma_{9190}^2 = 0.00033455$. Consequently, the prior AR(2)-GARCH(1, 1) model produces 1-step ahead forecasts as

$$\hat{r}_{9190}(1) = 0.00071 \quad \text{and} \quad \hat{\sigma}_{9190}^2(1) = 0.0003211.$$

The 5% quantile is then

$$0.00071 - 1.6449 \times \sqrt{0.0003211} = -0.02877,$$

where it is understood that the negative sign denotes left tail of the conditional normal distribution. The VaR for a long position of \$10 million with probability 0.05 is VaR $= \$10{,}000{,}000 \times 0.02877 = \$287{,}700$. The result shows that, with probability 95%, the potential loss of holding that position next day is \$287,200 or less assuming that the AR(2)-GARCH(1, 1) model holds. If the probability is 0.01, then the 1% quantile is

$$0.00071 - 2.3262 \times \sqrt{0.0003211} = -0.0409738.$$

The VaR for the position becomes \$409,738.

***Case 2***

Assume that $\epsilon_t$ is a standardized Student-$t$ distribution with 5 degrees of freedom. The fitted model is

$$r_t = 0.0003 - 0.0335 r_{t-2} + a_t, \quad a_t = \sigma_t \epsilon_t$$
$$\sigma_t^2 = 0.000003 + 0.0559 a_{t-1}^2 + 0.9350 \sigma_{t-1}^2.$$

From the data, we have $r_{9189} = -0.00201$, $r_{9190} = -0.0128$, and $\sigma_{9190}^2 = 0.000349$. Consequently, the prior Student-$t$ AR(2)-GARCH(1, 1) model produces 1-step ahead forecasts

$$\hat{r}_{9190}(1) = 0.000367 \quad \text{and} \quad \hat{\sigma}_{9190}^2(1) = 0.0003386.$$

The 5% quantile of a Student-$t$ distribution with 5 degrees of freedom is $-2.015$ and that of its standardized distribution is $-2.015/\sqrt{5/3} = -1.5608$. Therefore, the 5% quantile of the conditional distribution of $r_{9191}$ given $F_{9190}$ is

$$0.000367 - 1.5608\sqrt{0.0003386} = -0.028354.$$

The VaR for a long position of \$10 million is

$$\text{VaR} = \$10{,}000{,}000 \times 0.028352 = \$283{,}520,$$

which is essentially the same as that obtained under the normality assumption. The 1% quantile of the conditional distribution is

$$0.000367 - (3.3649/\sqrt{5/3})\sqrt{0.0003386} = -0.0475943.$$

The corresponding VaR is \$475,943. Comparing with that of Case I, we see the heavy-tail effect of using a Student-$t$ distribution with 5 degrees of freedom; it increases the VaR when the tail probability becomes smaller.

### 7.3.1 Multiple Periods

Suppose that at time $h$ we like to compute the $k$-horizon VaR of an asset whose log return is $r_t$. The variable of interest is the $k$-period log return at the forecast origin $h$ (i.e., $r_h[k] = r_{h+1} + \cdots + r_{h+k}$). If the return $r_t$ follows the time series model in Eqs. (7.5) and (7.6), then the conditional mean and variance of $r_h[k]$ given the information set $F_h$ can be obtained by the forecasting methods discussed in Chapters 2 and 3.

***Expected Return and Forecast Error***

The conditional mean $E(r_h[k] \mid F_h)$ can be obtained by the forecasting method of ARMA models in Chapter 2. Specifically, we have

$$\hat{r}_h[k] = r_h(1) + \cdots + r_h(k),$$

where $r_h(\ell)$ is the $\ell$-step ahead forecast of the return at the forecast origin $h$. These forecasts can be computed recursively as discussed in subsection 2.6.4. Using the MA representation

$$r_t = \mu + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots$$

of the ARMA model in Eq. (7.5), we can write the $\ell$-step ahead forecast error at the forecast origin $h$ as

$$e_h(\ell) = r_{h+\ell} - r_h(\ell) = a_{h+\ell} + \psi_1 a_{h+\ell-1} + \cdots + \psi_{\ell-1} a_{h+1};$$

see Eq. (2.30) and the associated forecast error. The forecast error of the expected $k$-period return $\hat{r}_h[k]$ is the sum of 1-step to $k$-step forecast errors of $r_t$ at the forecast origin $h$ and can be written as

$$\begin{aligned} e_h[k] &= e_h(1) + e_h(2) + \cdots + e_h(k) \\ &= a_{h+1} + (a_{h+2} + \psi_1 a_{h+1}) + \cdots + \sum_{i=0}^{k-1} \psi_i a_{h+k-i} \\ &= a_{h+k} + (1 + \psi_1) a_{h+k-1} + \cdots + \left( \sum_{i=0}^{k-1} \psi_i \right) a_{h+1} \end{aligned} \tag{7.7}$$

where $\psi_0 = 1$.

***Expected Volatility***

The volatility forecast of the $k$-period return at the forecast origin $h$ is the conditional variance of $e_h[k]$ given $F_h$. Using the independent assumption of $\epsilon_{t+i}$ for $i = 1, \ldots, k$, where $a_{t+i} = \sigma_{t+i}\epsilon_{t+i}$, we have

$$\begin{aligned} \text{Var}(e_h[k] \mid F_h) &= \text{Var}(a_{h+k} \mid F_h) + (1 + \psi_1)^2 \,\text{Var}(a_{h+k-1} \mid F_h) + \cdots \\ &\quad + \left( \sum_{i=0}^{k-1} \psi_i \right)^2 \text{Var}(a_{h+1} \mid F_h) \\ &= \sigma_h^2(k) + (1 + \psi_1)^2 \sigma_h^2(k-1) + \cdots + \left( \sum_{i=0}^{k-1} \psi_i \right)^2 \sigma_h^2(1), \end{aligned}$$

where $\sigma_h^2(\ell)$ is the $\ell$-step ahead volatility forecast at the forecast origin $h$. If the volatility model is the GARCH model in Eq. (7.6), then these volatility forecasts can be obtained recursively by the methods discussed in Chapter 3.

As an illustration, consider the special time series model

$$\begin{aligned} r_t &= \mu + a_t, \quad a_t = \sigma_t \epsilon_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2. \end{aligned}$$

Then we have $\psi_i = 0$ for all $i > 0$. The point forecast of the $k$-period return at the forecast origin $h$ is $\hat{r}_h[k] = k\mu$ and the associated forecast error is

$$e_h[k] = a_{h+k} + a_{h+k-1} + \cdots + a_{h+1}.$$

Consequently, the volatility forecast for the $k$-period return at the forecast origin $h$ is

$$\text{Var}(e_h[k] \mid F_h) = \sum_{\ell=1}^{k} \sigma_h^2(\ell).$$

Using the forecasting method of GARCH(1, 1) models in Section 3.4, we have

$$\sigma_h^2(1) = \alpha_0 + \alpha_1 a_h^2 + \beta_1 \sigma_h^2$$
$$\sigma_h^2(\ell) = \alpha_0 + (\alpha_1 + \beta_1)\sigma_h^2(\ell - 1), \quad \ell = 2, \ldots, k. \tag{7.8}$$

Thus, $\text{Var}(e_h[k] \mid F_h)$ can be obtained by the prior recursion. If $\epsilon_t$ is Gaussian, then the conditional distribution of $r_h[k]$ given $F_h$ is normal with mean $k\mu$ and variance $\text{Var}(e_h[k] \mid F_h)$. The quantiles needed in VaR calculation are readily available.

**Example 7.3.** (continued) Consider the Gaussian AR(2)-GARCH(1, 1) model of Example 7.3 for the daily log returns of IBM stock. Suppose that we are interested in the VaR of a 15-day horizon starting at the forecast origin 9190 (i.e., December 31, 1998). We can use the fitted model to compute the conditional mean and variance for the 15-day log return via $r_{9190}[15] = \sum_{i=1}^{15} r_{9190+i}$ given $F_{9190}$. The conditional mean is 0.00998 and the conditional variance is 0.0047948, which is obtained by the recursion in Eq. (7.8). The 5% quantile of the conditional distribution is then $0.00998 - 1.6449\sqrt{0.0047948} = -0.1039191$. Consequently, the 15-day horizon VaR for a long position of \$10 million is VaR $=$ \$10,000,000 $\times$ 0.1039191 $=$ \$1,039,191. This amount is smaller than \$287,700 $\times \sqrt{15} =$ \$1,114,257. This example further demonstrates that the square root of time rule used by RiskMetrics holds only for the special white-noise IGARCH(1, 1) model used. When the conditional mean is not zero, proper steps must be taken to compute the $k$-horizon VaR.

## 7.4 QUANTILE ESTIMATION

Quantile estimation provides a nonparametric approach to VaR calculation. It makes no specific distributional assumption on the return of a portfolio except that the distribution continues to hold within the prediction period. There are two types of quantile methods. The first method is to use empirical quantile directly, and the second method uses quantile regression.

### 7.4.1 Quantile and Order Statistics

Assuming that the distribution of return in the prediction period is the same as that in the sample period, one can use the empirical quantile of the return $r_t$ to calculate VaR. Let $r_1, \ldots, r_n$ be the returns of a portfolio in the sample period. The *order statistics* of the sample are these values arranged in increasing order. We use the notation

$$r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$$

to denote the arrangement and refer to $r_{(i)}$ as the $i$th order statistic of the sample. In particular, $r_{(1)}$ is the sample minimum and $r_{(n)}$ the sample maximum.

Assume that the returns are independent and identically distributed random variables that have a continuous distribution with probability density function (pdf) $f(x)$ and CDF $F(x)$. Then we have the following asymptotic result from the statistical literature (e.g., Cox and Hinkley, 1974, Appendix 2), for the order statistic $r_{(\ell)}$, where $\ell = np$ with $0 < p < 1$.

***Result:*** Let $x_p$ be the $p$th quantile of $F(x)$ [i.e., $x_p = F^{-1}(p)$]. Assume that the pdf $f(x)$ is not zero at $x_p$ [i.e., $f(x_p) \neq 0$]. Then the order statistic $r_{(\ell)}$ is asymptotically normal with mean $x_p$ and variance $p(1-p)/[nf^2(x_p)]$. That is,

$$r_{(\ell)} \sim N\left[x_p, \frac{p(1-p)}{n[f(x_p)]^2}\right], \quad \ell = np. \tag{7.9}$$

Based on the prior result, one can use $r_{(\ell)}$ to estimate the quantile $x_p$, where $\ell = np$. In practice, the probability of interest $p$ may not satisfy that $np$ is a positive integer. In this case, one can use simple interpolation to obtain quantile estimates. More specifically, for noninteger $np$, let $\ell_1$ and $\ell_2$ be the two neighboring positive integers such that $\ell_1 < np < \ell_2$. Define $p_i = \ell_i/n$. The previous result shows that $r_{(\ell_i)}$ is a consistent estimate of the quantile $x_{p_i}$. From the definition, $p_1 < p < p_2$. Therefore, the quantile $x_p$ can be estimated by

$$\hat{x}_p = \frac{p_2 - p}{p_2 - p_1} r_{(\ell_1)} + \frac{p - p_1}{p_2 - p_1} r_{(\ell_2)}. \tag{7.10}$$

**Example 7.4.** Consider the daily log returns of Intel stock from December 15, 1972 to December 31, 1997. There are 6329 observations. The empirical 5% quantile of the data can be obtained as

$$\hat{x}_{0.05} = 0.55 r_{(316)} + 0.45 r_{(317)} = -4.229\%,$$

where $np = 6329 \times 0.05 = 316.45$ and $r_{(i)}$ is the $i$th order statistic of the sample. In this particular instance, $r_{(316)} = -4.237\%$ and $r_{(317)} = -4.220\%$. Here we use the lower tail of the empirical distribution because it is relevant to holding a long position in VaR calculation.

**Example 7.5.** Consider again the daily log returns of IBM stock from July 3, 1962 to December 31, 1998. Using all the 9190 observations, the empirical 5% quantile can be obtained as $(r_{(459)} + r_{(460)})/2 = -0.021603$, where $r_{(i)}$ is the $i$th order statistic and $np = 9190 \times 0.05 = 459.5$. The VaR of a long position of \$10 million is \$216,030, which is much smaller than those obtained by the econometric approach discussed before. Because the sample size is 9190, we have $91 < 9190 \times 0.01 < 92$. Let $p_1 = 91/9190 = 0.0099$ and $p_2 = 92/9190 = 0.01001$. The empirical 1% quantile can be obtained as

$$\begin{aligned}\hat{x}_{0.01} &= \frac{p_2 - 0.01}{p_2 - p_1} r_{(91)} + \frac{0.01 - p_1}{p_2 - p_1} r_{(92)} \\ &= \frac{0.00001}{0.00011}(-3.658) + \frac{0.0001}{0.00011}(-3.657) \\ &\approx -3.657.\end{aligned}$$

The 1% 1-day horizon VaR of the long position is \$365,709. Again, this amount is lower than those obtained before by other methods.

**Discussion:** Advantages of using the prior quantile method to VaR calculation include (a) simplicity, and (b) using no specific distributional assumption. However, the approach has several drawbacks. First, it assumes that the distribution of the return $r_t$ remains unchanged from the sample period to the prediction period. Given that VaR is concerned mainly with tail probability, this assumption implies that the predicted loss cannot be greater than that of the historical loss. It is definitely not so in practice. Second, for extreme quantiles (i.e., when $p$ is close to zero or unity), the empirical quantiles are not efficient estimates of the theoretical quantiles. Third, the direct quantile estimation fails to take into account the effect of explanatory variables that are relevant to the portfolio under study. In real application, VaR obtained by the empirical quantile can serve as a lower bound for the actual VaR.

### 7.4.2 Quantile Regression

In real application, one often has explanatory variables available that are important to the problem under study. For example, the action taken by Federal Reserve Banks on interest rates could have important impacts on the returns of U.S. stocks. It is then more appropriate to consider the distribution function $r_{t+1} \mid F_t$, where $F_t$ includes the explanatory variables. In other words, we are interested in the quantiles of the distribution function of $r_{t+1}$ given $F_t$. Such a quantile is referred to as a *regression quantile* in the literature; see Koenker and Bassett (1978).

To understand regression quantile, it is helpful to cast the empirical quantile of the previous subsection as an estimation problem. For a given probability $p$, the $p$th quantile of $\{r_t\}$ is obtained by

$$\hat{x}_p = \text{argmin}_\beta \sum_{i=1}^{n} w_p(r_i - \beta),$$

where $w_p(z)$ is defined by

$$w_p(z) = \begin{cases} pz & \text{if } z \geq 0 \\ (p-1)z & \text{if } z < 0. \end{cases}$$

Regression quantile is a generalization of such an estimate.

To see the generalization, suppose that we have the linear regression

$$r_t = \boldsymbol{\beta}' \boldsymbol{x}_t + a_t, \tag{7.11}$$

where $\boldsymbol{\beta}$ is a $k$-dimensional vector of parameters and $\boldsymbol{x}_t$ is a vector of predictors that are elements of $F_{t-1}$. The conditional distribution of $r_t$ given $F_{t-1}$ is a translation of the distribution of $a_t$ because $\boldsymbol{\beta}' \boldsymbol{x}_t$ is known. Viewing the problem this way, Koenker and Bassett (1978) suggest to estimate the conditional quantile $x_p \mid F_{t-1}$ of $r_t$ given $F_{t-1}$ as

$$\hat{x}_p \mid F_{t-1} \equiv \inf\{\boldsymbol{\beta}_o' \boldsymbol{x} \mid R_p(\boldsymbol{\beta}_o) = \min\}, \tag{7.12}$$

where "$R_p(\boldsymbol{\beta}_o) = \min$" means that $\boldsymbol{\beta}_o$ is obtained by

$$\boldsymbol{\beta}_o = \text{argmin}_{\boldsymbol{\beta}} \sum_{t=1}^{n} w_p(r_t - \boldsymbol{\beta}' \boldsymbol{x}_t),$$

where $w_p(.)$ is defined as before. A computer program to obtain such an estimated quantile can be found in Koenker and D'Orey (1987).

## 7.5 EXTREME VALUE THEORY

In this section, we review some extreme value theory in the statistical literature. Denote the return of an asset, measured in a fixed time interval such as daily, by $r_t$. Consider the collection of $n$ returns, $\{r_1, \ldots, r_n\}$. The minimum return of the collection is $r_{(1)}$, that is, the smallest order statistic, whereas the maximum return is $r_{(n)}$, the maximum order statistic. Specifically, $r_{(1)} = \min_{1 \le j \le n}\{r_j\}$ and $r_{(n)} = \max_{1 \le j \le n}\{r_j\}$. We focus on properties of the minimum return $r_{(1)}$ because this minimum is highly relevant to VaR calculation for a long position. However, the theory discussed also applies to the maximum return of an asset over a given time period because properties of the maximum return can be obtained from those of the minimum by a simple sign change. Specifically, we have $r_{(n)} = -\min_{1 \le j \le n}\{-r_j\} = -r^c_{(1)}$, where $r^c_t = -r_t$ with the superscript $c$ denoting sign change. The maximum return is relevant to holding a short financial position.

### 7.5.1 Review of Extreme Value Theory

Assume that the returns $r_t$ are serially independent with a common cumulative distribution function $F(x)$ and that the range of the return $r_t$ is $[l, u]$. For log returns, we have $l = -\infty$ and $u = \infty$. Then the CDF of $r_{(1)}$, denoted by $F_{n,1}(x)$, is given by

$$\begin{aligned} F_{n,1}(x) &= \Pr[r_{(1)} \le x] = 1 - \Pr[r_{(1)} > x] \\ &= 1 - \Pr(r_1 > x, r_2 > x, \ldots, r_n > x) \end{aligned}$$

$$
\begin{aligned}
&= 1 - \prod_{j=1}^{n} \Pr(r_j > x), \quad \text{(by independence)} \\
&= 1 - \prod_{j=1}^{n} [1 - \Pr(r_j \le x)] \\
&= 1 - \prod_{j=1}^{n} [1 - F(x)] \quad \text{(by common distribution)} \\
&= 1 - [1 - F(x)]^n. \qquad (7.13)
\end{aligned}
$$

In practice, the CDF $F(x)$ of $r_t$ is unknown and, hence, $F_{n,1}(x)$ of $r_{(1)}$ is unknown. However, as $n$ increases to infinity, $F_{n,1}(x)$ becomes degenerated—namely, $F_{n,1}(x) \to 0$ if $x \le l$ and $F_{n,1}(x) \to 1$ if $x > l$ as $n$ goes to infinity. This degenerated CDF has no practical value. Therefore, the extreme value theory is concerned with finding two sequences $\{\beta_n\}$ and $\{\alpha_n\}$, where $\alpha_n > 0$, such that the distribution of $r_{(1*)} \equiv (r_{(1)} - \beta_n)/\alpha_n$ converges to a nondegenerated distribution as $n$ goes to infinity. The sequence $\{\beta_n\}$ is a location series and $\{\alpha_n\}$ is a series of scaling factors. Under the independent assumption, the limiting distribution of the normalized minimum $r_{(1*)}$ is given by

$$
F_*(x) = \begin{cases} 1 - \exp[-(1 + kx)^{1/k}] & \text{if } k \neq 0 \\ 1 - \exp[-\exp(x)] & \text{if } k = 0 \end{cases} \qquad (7.14)
$$

for $x < -1/k$ if $k < 0$ and for $x > -1/k$ if $k > 0$, where the subscript $*$ signifies the minimum. The case of $k = 0$ is taken as the limit when $k \to 0$. The parameter $k$ is referred to as the *shape parameter* that governs the tail behavior of the limiting distribution. The parameter $\alpha = -1/k$ is called the *tail index* of the distribution.

The limiting distribution in Eq. (7.14) is the *generalized extreme value distribution* of Jenkinson (1955) for the minimum. It encompasses the three types of limiting distribution of Gnedenko (1943):

- Type I: $k = 0$, the Gumbel family. The CDF is

$$
F_*(x) = 1 - \exp[-\exp(x)], \quad -\infty < x < \infty. \qquad (7.15)
$$

- Type II: $k < 0$, the Fréchet family. The CDF is

$$
F_*(x) = \begin{cases} 1 - \exp[-(1 + kx)^{1/k}] & \text{if } x < -1/k \\ 1 & \text{otherwise.} \end{cases} \qquad (7.16)
$$

- Type III: $k > 0$, the Weibull family. The CDF here is

$$
F_*(x) = \begin{cases} 1 - \exp[-(1 + kx)^{1/k}] & \text{if } x > -1/k \\ 0 & \text{otherwise.} \end{cases}
$$

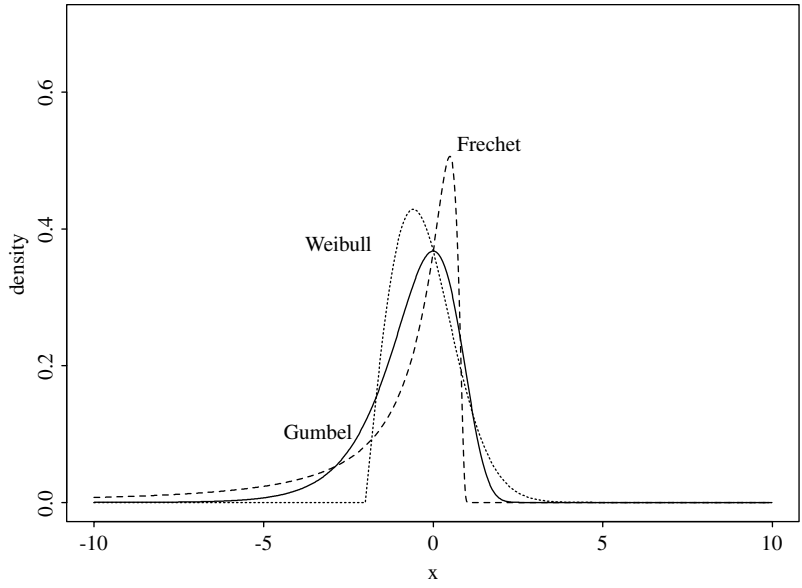


**Figure 7.2.** Probability density functions of extreme value distributions for minimum: The solid line is for a Gumbel distribution, the dotted line is for the Weibull distribution with $k = 0.5$, and the dashed line for the Fréchet distribution with $k = -0.9$.

Gnedenko (1943) gave necessary and sufficient conditions for the CDF $F(x)$ of $r_t$ to be associated with one of the three types of limiting distribution. Briefly speaking, the tail behavior of $F(x)$ determines the limiting distribution $F_*(x)$ of the minimum. The (left) tail of the distribution declines exponentially for the Gumbel family, by a power function for the Fréchet family, and is finite for the Weibull family. Readers are referred to Embrechts, Kuppelberg, and Mikosch (1997) for a comprehensive treatment of the extreme value theory. For risk management, we are mainly interested in the Fréchet family that includes stable and Student-$t$ distributions. The Gumbel family consists of thin-tailed distributions such as normal and log-normal distributions. The probability density function (pdf) of the generalized limiting distribution in Eq. (7.14) can be obtained easily by differentiation:

$$f_*(x) = \begin{cases} (1+kx)^{1/k-1} \exp[-(1+kx)^{1/k}] & \text{if } k \neq 0 \\ \exp[x - \exp(x)] & \text{if } k = 0, \end{cases} \tag{7.17}$$

where $-\infty < x < \infty$ for $k = 0$, $x < -1/k$ for $k < 0$, and $x > -1/k$ for $k > 0$.

The aforementioned extreme value theory has two important implications. First, the tail behavior of the CDF $F(x)$ of $r_t$, not the specific distribution, determines the limiting distribution $F_*(x)$ of the (normalized) minimum. Thus, the theory is generally applicable to a wide range of distributions for the return $r_t$. The sequences $\{\beta_n\}$ and $\{\alpha_n\}$, however, may depend on the CDF $F(x)$. Second, Feller (1971, p. 279)

shows that the tail index $k$ does not depend on the time interval of $r_t$. That is, the tail index (or equivalently the shape parameter) is invariant under time aggregation. This second feature of the limiting distribution becomes handy in the VaR calculation.

The extreme value theory has been extended to serially dependent observations $\{r_t\}_{t=1}^n$ provided that the dependence is weak. Berman (1964) shows that the same form of the limiting extreme value distribution holds for stationary normal sequences provided that the autocorrelation function of $r_t$ is squared summable (i.e., $\sum_{i=1}^{\infty} \rho_i^2 < \infty$), where $\rho_i$ is the lag-$i$ autocorrelation function of $r_t$. For further results concerning the effect of serial dependence on the extreme value theory, readers are referred to Leadbetter, Lindgren, and Rootzén (1983, Chapter 3).

### 7.5.2 Empirical Estimation

The extreme value distribution contains three parameters—$k$, $\beta_n$, and $\alpha_n$. These parameters are referred to as the *shape*, *location*, and *scale parameter*, respectively. They can be estimated by using either parametric or nonparametric methods. We review some of the estimation methods.

For a given sample, there is only a single minimum or maximum, and we cannot estimate the three parameters with only an extreme observation. Alternative ideas must be used. One of the ideas used in the literature is to divide the sample into subsamples and apply the extreme value theory to the subsamples. Assume that there are $T$ returns $\{r_j\}_{j=1}^T$ available. We divide the sample into $g$ non-overlapping subsamples each with $n$ observations, assuming for simplicity that $T = ng$. In other words, we divide the data as

$$\{r_1, \ldots, r_n \mid r_{n+1}, \ldots, r_{2n} \mid r_{2n+1}, \ldots, r_{3n} \mid \cdots \mid r_{(g-1)n+1}, \ldots, r_{ng}\}$$

and write the observed returns as $r_{in+j}$, where $1 \leq j \leq n$ and $i = 0, \ldots, g-1$. Notice that each subsample corresponds to a subperiod of the data span. When $n$ is sufficiently large, we hope that the extreme value theory applies to each subsample. In application, the choice of $n$ can be guided by practical considerations. For example, for daily returns, $n = 21$ corresponds approximately to the number of trading days in a month and $n = 63$ denotes the number of trading days in a quarter.

Let $r_{n,i}$ be the minimum of the $i$th subsample (i.e., $r_{n,i}$ is the smallest return of the $i$th subsample), where the subscript $n$ is used to denote the size of the subsample. When $n$ is sufficiently large, $x_{n,i} = (r_{n,i} - \beta_n)/\alpha_n$ should follow an extreme value distribution, and the collection of subsample minima $\{r_{n,i} \mid i = 1, \ldots, g\}$ can then be regarded as a sample of $g$ observations from that extreme value distribution. Specifically, we define

$$r_{n,i} = \min_{1 \leq j \leq n} \{r_{(i-1)n+j}\}, \quad i = 1, \ldots, g. \tag{7.18}$$

The collection of subsample minima $\{r_{n,i}\}$ are the data we use to estimate the unknown parameters of the extreme value distribution. Clearly, the estimates obtained may depend on the choice of subperiod length $n$.

#### 7.5.2.1 *The Parametric Approach*

Two parametric approaches are available. They are the maximum likelihood and regression methods.

*Maximum likelihood method*

Assuming that the subperiod minima $\{r_{n,i}\}$ follow a generalized extreme value distribution such that the pdf of $x_i = (r_{n,i} - \beta_n)/\alpha_n$ is given in Eq. (7.17), we can obtain the pdf of $r_{n,i}$ by a simple transformation as

$$f(r_{n,i}) = \begin{cases} \dfrac{1}{\alpha_n}\left[1 + \dfrac{k_n(r_{n,i} - \beta_n)}{\alpha_n}\right]^{1/k_n - 1} \exp\left\{-\left[1 + \dfrac{k_n(r_{n,i} - \beta_n)}{\alpha_n}\right]^{1/k_n}\right\} & \text{if } k_n \neq 0 \\ \dfrac{1}{\alpha_n} \exp\left\{\dfrac{r_{n,i} - \beta_n}{\alpha_n} - \exp\left[\dfrac{r_{n,i} - \beta_n}{\alpha_n}\right]\right\} & \text{if } k_n = 0, \end{cases}$$

where it is understood that $1 + k_n(r_{n,i} - \beta_n)/\alpha_n > 0$ if $k_n \neq 0$. The subscript $n$ is added to the shape parameter $k$ to signify that its estimate depends on the choice of $n$. Under the independence assumption, the likelihood function of the subperiod minima is

$$\ell(r_{n,1}, \ldots, r_{n,g} \mid k_n, \alpha_n, \beta_n) = \prod_{i=1}^{g} f(r_{n,i}).$$

Nonlinear estimation procedures can then be used to obtain maximum likelihood estimates of $k_n$, $\beta_n$, and $\alpha_n$. These estimates are unbiased, asymptotically normal, and of minimum variance under proper assumptions. We apply this approach to some stock return series later.

*Regression method*

This method assumes that $\{r_{n,i}\}_{i=1}^{g}$ is a random sample from the generalized extreme value distribution in Eq. (7.14) and make uses of properties of order statistics; see Gumbel (1958). Denote the order statistics of the subperiod minima $\{r_{n,i}\}_{i=1}^{g}$ as

$$r_{n(1)} \leq r_{n(2)} \leq \cdots \leq r_{n(g)}.$$

Using properties of order statistics (e.g., Cox and Hinkley, 1974, p. 467), we have

$$E\{F_*[r_{n(i)}]\} = \frac{i}{g+1}, \quad i = 1, \ldots, g. \tag{7.19}$$

For simplicity, we separate the discussions into two cases depending on the value of $k$. First, consider the case of $k \neq 0$. From Eq. (7.14), we have

$$F_*[r_{n(i)}] = 1 - \exp\left\{-\left[1 + k_n \frac{r_{n(i)} - \beta_n}{\alpha_n}\right]^{1/k_n}\right\} \tag{7.20}$$

Consequently, using Eqs. (7.19) and (7.20) and approximating expectation by an observed value, we have

$$\frac{i}{g+1} = 1 - \exp\left\{-\left[1 + k_n \frac{r_{n(i)} - \beta_n}{\alpha_n}\right]^{1/k_n}\right\}.$$

Therefore,

$$\exp\left\{-\left[1 + k_n \frac{r_{n(i)} - \beta_n}{\alpha_n}\right]^{1/k_n}\right\} = 1 - \frac{i}{g+1} = \frac{g+1-i}{g+1}, \quad i = 1, \ldots, g.$$

Taking natural logarithm twice, the prior equation gives

$$\ln\left[-\ln\left(\frac{g+1-i}{g+1}\right)\right] = \frac{1}{k_n} \ln\left[1 + k_n \frac{r_{n(i)} - \beta_n}{\alpha_n}\right], \quad i = 1, \ldots, g.$$

In practice, letting $e_i$ be the deviation between the previous two quantities and assuming that the series $\{e_t\}$ is not serially correlated, we have a regression setup

$$\ln\left[-\ln\left(\frac{g+1-i}{g+1}\right)\right] = \frac{1}{k_n} \ln\left[1 + k_n \frac{r_{n(i)} - \beta_n}{\alpha_n}\right] + e_i, \quad i = 1, \ldots, g. \tag{7.21}$$

The least squares estimates of $k_n$, $\beta_n$, and $\alpha_n$ can be obtained by minimizing the sum of squares of $e_i$.

When $k_n = 0$, the regression setup reduces to

$$\ln\left[-\ln\left(\frac{g+1-i}{g+1}\right)\right] = \frac{1}{\alpha_n} r_{n(i)} - \frac{\beta_n}{\alpha_n} + e_i, \quad i = 1, \ldots, g.$$

The least squares estimates are consistent, but less efficient than the likelihood estimates. We use the likelihood estimates in this chapter.

#### *7.5.2.2 The Nonparametric Approach*

The shape parameter $k$ can be estimated using some nonparametric methods. We mention two such methods here. These two methods are proposed by Hill (1975) and Pickands (1975) and are referred to as the Hill estimator and Pickands estimator, respectively. Both estimators apply directly to the returns $\{r_t\}_{t=1}^T$. Thus, there is no need to consider subsamples. Denote the order statistics of the sample as

$$r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(T)}.$$

Let $q$ be a positive integer. The two estimators of $k$ are defined as

$$k_p(q) = -\frac{1}{\ln(2)} \ln\left[\frac{-r_{(q)} + r_{(2q)}}{-r_{(2q)} + r_{(4q)}}\right] \tag{7.22}$$

$$k_h(q) = \frac{-1}{q} \sum_{i=1}^{q} \left\{\ln[-r_{(i)}] - \ln[-r_{(q+1)}]\right\}, \tag{7.23}$$

where the argument $(q)$ is used to emphasize that the estimators depend on $q$. The choice of $q$ differs between Hill and Pickands estimators. It has been investigated by several researchers, but there is no general consensus on the best choice available. Dekkers and De Haan (1989) show that $k_p(q)$ is consistent if $q$ increases at a properly chosen pace with the sample size $T$. In addition, $\sqrt{q}[k_p(q) - k]$ is asymptotically normal with mean zero and variance $k^2(2^{-2k+1} + 1)/[2(2^{-k} - 1)\ln(2)]^2$. The Hill estimator is applicable to the Fréchet distribution only, but it is more efficient than the Pickands estimator when applicable. Goldie and Smith (1987) show that $\sqrt{q}[k_h(q) - k]$ is asymptotically normal with mean zero and variance $k^2$. In practice, one may plot the Hill estimator $k_h(q)$ against $q$ and find a proper $q$ such that the estimate appears to be stable.

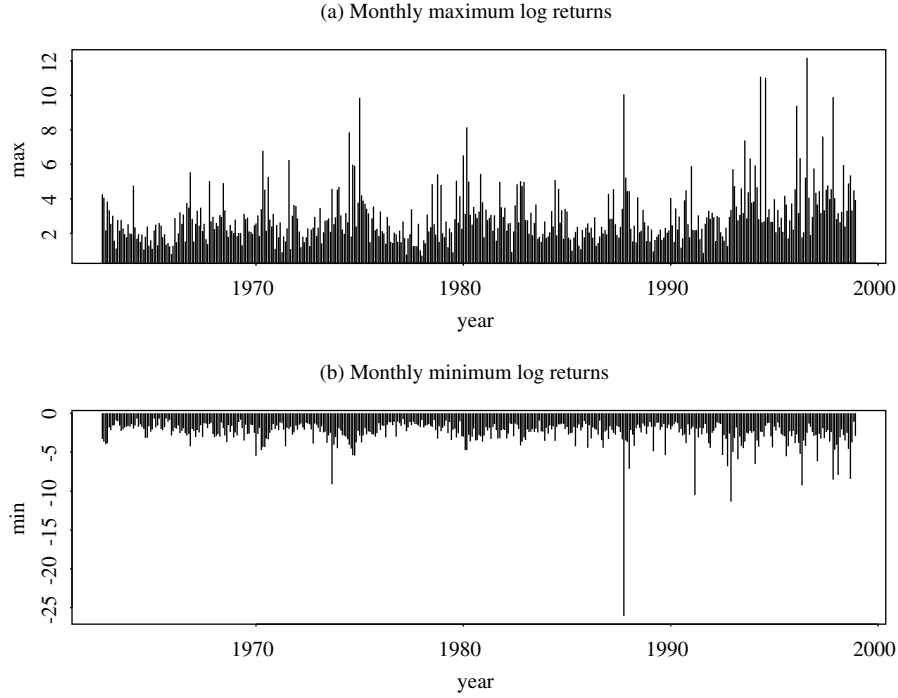


**Figure 7.3.** Maximum and minimum daily log returns of IBM stock when the subperiod is 21 trading days. The data span is from July 3, 1962 to December 31, 1998: (a) positive returns, and (b) negative returns.

### 7.5.3 Application to Stock Returns

We apply the extreme value theory to the daily log returns of IBM stock from July 3, 1962 to December 31, 1998. The returns are measured in percentages, and the sample size is 9190 (i.e., $T = 9190$). Figure 7.3 shows the time plots of extreme daily log returns when the length of the subperiod is 21, which corresponds approximately to a month. The October 1987 crash is clearly seen from the plot. Excluding the 1987 crash, the range of extreme daily log returns is between 0.5% and 13%.

Table 7.1 summarizes some estimation results of the shape parameter $k$ via the Hill estimator. Three choices of $q$ are reported in the table, and the results are stable. To provide an overall picture of the performance of Hill estimator, Figure 7.4 shows

**Table 7.1. Results of the Hill Estimator for Daily Log Returns of IBM Stock From July 3, 1962 to December 31, 1998. Standard Errors are in Parentheses.**

| $q$ | 190 | 200 | 210 |
|---|---|---|---|
| Maximum | −0.300(0.022) | −0.297(0.021) | −0.303(0.021) |
| Minimum | −0.290(0.021) | −0.292(0.021) | −0.289(0.020) |

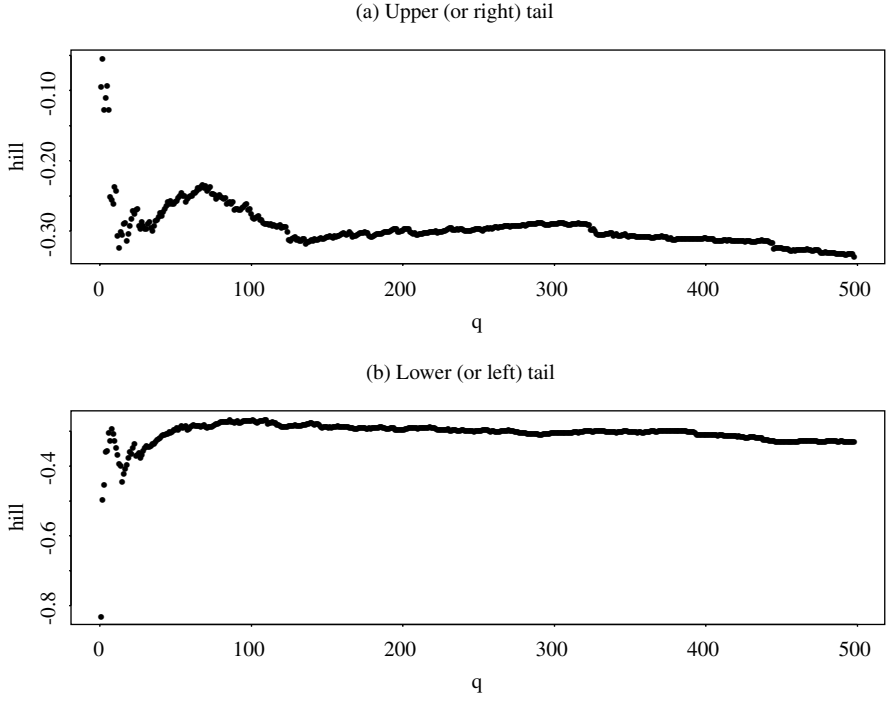


**Figure 7.4.** Scatterplots of the Hill estimator for the daily log returns of IBM stock. The sample period is from July 3, 1962 to December 31, 1998: (a) positive returns, and (b) negative returns.

the scatter plots of the Hill estimator $k_h(q)$ against $q$. For both positive and negative extreme daily log returns, the estimator is stable except for cases when $q$ is small. The estimated shape parameters are about $-0.30$ and are significantly different from zero at the asymptotic 5% level. The plots also indicate that the shape parameter $k$ appears to be smaller for the negative extremes, indicating that the daily log return may have a heavier left tail. Overall, the result indicates that the distribution of daily log returns of IBM stock belongs to the Fréchet family. The analysis thus rejects the normality assumption commonly used in practice. Such a conclusion is in agreement with that of Longin (1996), who used a U.S. stock market index series.

Next we apply the maximum likelihood method to estimate parameters of the generalized extreme value distribution for IBM daily log returns. Table 7.2 summarizes the estimation results for different choices of the length of subperiods ranging from 1 month ($n = 21$) to 1 year ($n = 252$). From the table, we make the following observations:

- Estimates of the location and scale parameters $\beta_n$ and $\alpha_n$ increase in modulus as $n$ increases. This is expected as magnitudes of the subperiod minimum and maximum are nondecreasing functions of $n$.
- Estimates of the shape parameter (or equivalently the tail index) are stable for the negative extremes when $n \geq 63$ and are approximately $-0.33$.
- Estimates of the shape parameter are less stable for the positive extremes. The estimates are smaller in magnitude, but remain significantly different from zero.
- The results for $n = 252$ have higher variabilities as the number of subperiods $g$ is relatively small.

Again the conclusion obtained is similar to that of Longin (1996), who provided a good illustration of applying the extreme value theory to stock market returns.

**Table 7.2. Maximum Likelihood Estimates of the Extreme Value Distribution for Daily Log Returns of IBM Stock From July 3, 1962 to December 31, 1998. Standard Errors are in Parentheses.**

| Length of subperiod | Scale $\alpha_n$ | Location $\beta_n$ | Shape Par. $k_n$ |
|---|---|---|---|
| (a) Minimal returns | | | |
| 1 mon.($n = 21$, $g = 437$) | 0.823(0.035) | −1.902(0.044) | −0.197(0.036) |
| 1 qur($n = 63$, $g = 145$) | 0.945(0.077) | −2.583(0.090) | −0.335(0.076) |
| 6 mon.($n = 126$, $g = 72$) | 1.147(0.131) | −3.141(0.153) | −0.330(0.101) |
| 1 year($n = 252$, $g = 36$) | 1.542(0.242) | −3.761(0.285) | −0.322(0.127) |
| (b) Maximal returns | | | |
| 1 mon.($n = 21$, $g = 437$) | 0.931(0.039) | 2.184(0.050) | −0.168(0.036) |
| 1 qur($n = 63$, $g = 145$) | 1.157(0.087) | 3.012(0.108) | −0.217(0.066) |
| 6 mon.($n = 126$, $g = 72$) | 1.292(0.158) | 3.471(0.181) | −0.349(0.130) |
| 1 year($n = 252$, $g = 36$) | 1.624(0.271) | 4.475(0.325) | −0.264(0.186) |

## 7.6 AN EXTREME VALUE APPROACH TO VAR

In this section, we discuss an approach to VaR calculation using the extreme value theory. The approach is similar to that of Longin (1999a, 1999b), who proposed an eight-step procedure for the same purpose. We divide the discussion into two parts. The first part is concerned with parameter estimation using the method discussed in the previous subsections. The second part focuses on VaR calculation by relating the probabilities of interest associated with different time intervals.

***Part I***

Assume that there are $T$ observations of an asset return available in the sample period. We partition the sample period into $g$ nonoverlapping subperiods of length $n$ such that $T = ng$. If $T = ng + m$ with $1 \leq m < n$, then we delete the first $m$ observations from the sample. The extreme value theory discussed in the previous section enables us to obtain estimates of the location, scale, and shape parameters $\beta_n$, $\alpha_n$, and $k_n$ for the subperiod minima $\{r_{n,i}\}$. Plugging the MLE estimates into the CDF in Eq. (7.14) with $x = (r - \beta_n)/\alpha_n$, we can obtain the quantile of a given probability of the generalized extreme value distribution. Because we focus on holding a long financial position, the lower probability (or left) quantiles are of interest. Let $p^*$ be a small probability that indicates the potential loss of a long position and $r_n^*$ be the $p^*$th quantile of the subperiod minimum under the limiting generalized extreme value distribution. Then we have

$$p^* = \begin{cases} 1 - \exp\left[-\left(1 + \dfrac{k_n(r_n^* - \beta_n)}{\alpha_n}\right)^{1/k_n}\right] & \text{if } k_n \neq 0 \\ 1 - \exp\left[-\exp\left(\dfrac{r_n^* - \beta_n}{\alpha_n}\right)\right] & \text{if } k_n = 0, \end{cases}$$

where it is understood that $1 + k_n(r_n^* - \beta_n)/\alpha_n > 0$ for $k_n \neq 0$. Rewriting this equation as

$$\ln(1 - p^*) = \begin{cases} -\left[1 + \dfrac{k_n(r_n^* - \beta_n)}{\alpha_n}\right]^{1/k_n} & \text{if } k_n \neq 0 \\ -\exp\left[\dfrac{r_n^* - \beta_n}{\alpha_n}\right] & \text{if } k_n = 0, \end{cases}$$

we obtain the quantile as

$$r_n^* = \begin{cases} \beta_n - \frac{\alpha_n}{k_n}\left\{1 - \left[-\ln(1 - p^*)\right]^{k_n}\right\} & \text{if } k_n \neq 0 \\ \beta_n + \alpha_n \ln[-\ln(1 - p^*)] & \text{if } k_n = 0. \end{cases} \tag{7.24}$$

In financial applications, the case of $k_n \neq 0$ is of major interest.

***Part II***

For a given lower (or left tail) probability $p^*$, the quantile $r_n^*$ of Eq. (7.24) is the VaR based on the extreme value theory for the subperiod minima. The next step is to make explicit the relationship between subperiod minima and the observed return $r_t$ series.

Because most asset returns are either serially uncorrelated or have weak serial correlations, we may use the relationship in Eq. (7.13) and obtain

$$p^* = P(r_{n,i} \le r_n^*) = 1 - [1 - P(r_t \le r_n^*)]^n$$

or, equivalently,

$$1 - p^* = [1 - P(r_t \le r_n^*)]^n. \tag{7.25}$$

This relationship between probabilities allows us to obtain VaR for the original asset return series $r_t$. More precisely, for a specified small lower probability $p$, the $p$th quantile of $r_t$ is $r_n^*$ if the probability $p^*$ is chosen based on Eq. (7.25), where $p = P(r_t \le r_n^*)$. Consequently, for a given small probability $p$, the VaR of holding a long position in the asset underlying the log return $r_t$ is

$$\text{VaR} = \begin{cases} \beta_n - \dfrac{\alpha_n}{k_n}\left\{1 - [-n \ln(1-p)]^{k_n}\right\} & \text{if } k_n \neq 0 \\ \beta_n + \alpha_n \ln[-n \ln(1-p)] & \text{if } k_n = 0. \end{cases} \tag{7.26}$$

***Summary***

We summarize the approach of applying the traditional extreme value theory to VaR calculation as follows:

1. Select the length of the subperiod $n$ and obtain subperiod minima $\{r_{n,i}\}$, $i = 1, \ldots, g$, where $g = T/n$.
2. Obtain the maximum likelihood estimates of $\beta_n$, $\alpha_n$, and $k_n$.
3. Check the adequacy of the fitted extreme value model; see the next section for some methods of model checking.
4. If the extreme value model is adequate, apply Eq. (7.26) to calculate VaR.

***Remark:*** Since we focus on holding a long financial position and, hence, on the quantile in the left tail of a return distribution, the quantile is negative. Yet it is customary in practice to use a positive number for VaR calculation. Thus, in using Eq. (7.26), one should be aware that the negative sign signifies a loss.

**Example 7.6.** Consider the daily log return, in percentage, of IBM stock from July 7, 1962 to December 31, 1998. From Table 7.2, we have $\hat{\alpha}_n = 0.945$, $\hat{\beta}_n = -2.583$, and $\hat{k}_n = -0.335$ for $n = 63$. Therefore, for the left-tail probability $p = 0.01$, the corresponding VaR is

$$\begin{aligned}\text{VaR} &= -2.583 - \frac{0.945}{-0.335}\left\{1 - [-63\ln(1-0.01)]^{-0.335}\right\} \\ &= -3.04969.\end{aligned}$$

Thus, for daily log returns of the stock, the 1% quantile is $-3.04969$. If one holds a long position on the stock worth \$10 million, then the estimated VaR with probability 1% is $\$10,000,000 \times 0.0304969 = \$304{,}969$. If the probability is 0.05, then the corresponding VaR is \$166,641.

If we chose $n = 21$ (i.e., approximately 1 month), then $\hat{\alpha}_n = 0.823$, $\hat{\beta}_n = -1.902$, and $\hat{k}_n = -0.197$. The 1% quantile of the extreme value distribution is

$$\text{VaR} = -1.902 - \frac{0.823}{-0.197}\{1 - [-21\ln(1-0.01)]^{-0.197}\} = -3.40013.$$

Therefore, for a long position of \$10,000,000, the corresponding 1-day horizon VaR is \$340,013 at the 1% risk level. If the probability is 0.05, then the corresponding VaR is \$184,127. In this particular case, the choice of $n = 21$ gives higher VaR values.

It is somewhat surprising to see that the VaR values obtained in Example 7.6 using the extreme value theory are smaller than those of Example 7.3 that uses a GARCH(1, 1) model. In fact, the VaR values of Example 7.6 are even smaller than those based on the empirical quantile in Example 7.5. This is due in part to the choice of probability 0.05. If one chooses probability $0.001 = 0.1\%$ and considers the same financial position, then we have VaR = \$546,641 for the Gaussian AR(2)-GARCH(1, 1) model and VaR = \$666,590 for the extreme value theory with $n = 21$. Furthermore, the VaR obtained here via the traditional extreme value theory may not be adequate because the independent assumption of daily log returns is often rejected by statistical testings. Finally, the use of subperiod minima overlooks the fact of volatility clustering in the daily log returns. The new approach of extreme value theory discussed in the next section overcomes these weaknesses.

***Remark:*** As shown by the results of Example 7.6, the VaR calculation based on the traditional extreme value theory depends on the choice of $n$, which is the length of subperiods. For the limiting extreme value distribution to hold, one would prefer a large $n$. But a larger $n$ means a smaller $g$ when the sample size $T$ is fixed, where $g$ is the effective sample size used in estimating the three parameters $\alpha_n$, $\beta_n$, and $k_n$. Therefore, some compromise between the choices of $n$ and $g$ is needed. A proper choice may depend on the returns of the asset under study. We recommend that one should check the stability of the resulting VaR in applying the traditional extreme value theory.

### 7.6.1 Discussion

We have applied various methods of VaR calculation to the daily log returns of IBM stock for a long position of \$10 million. Consider the VaR of the position for the

next trading day. If the probability is 5%, which means that with probability 0.95 the loss will be less than or equal to the VaR for the next trading day, then the results obtained are

1. \$302,500 for the RiskMetrics,
2. \$287,200 for a Gaussian AR(2)-GARCH(1, 1) model,
3. \$283,520 for an AR(2)-GARCH(1, 1) model with a standardized Student-$t$ distribution with 5 degrees of freedom,
4. \$216,030 for using the empirical quantile, and
5. \$184,127 for applying the traditional extreme value theory using monthly minima (i.e., subperiod length $n = 21$).

If the probability is 1%, then the VaR is

1. \$426,500 for the RiskMetrics,
2. \$409,738 for a Gaussian AR(2)-GARCH(1, 1) model,
3. \$475,943 for an AR(2)-GARCH(1, 1) model with a standardized Student-$t$ distribution with 5 degrees of freedom,
4. \$365,709 for using the empirical quantile, and
5. \$340,013 for applying the traditional extreme value theory using monthly minima (i.e., subperiod length $n = 21$).

If the probability is 0.1%, then the VaR becomes

1. \$566,443 for the RiskMetrics,
2. \$546,641 for a Gaussian AR(2)-GARCH(1, 1) model,
3. \$836,341 for an AR(2)-GARCH(1, 1) model with a standardized Student-$t$ distribution with 5 degrees of freedom,
4. \$780,712 for using the empirical quantile, and
5. \$666,590 for applying the traditional extreme value theory using monthly minima (i.e., subperiod length $n = 21$).

There are substantial differences among different approaches. This is not surprising because there exists substantial uncertainty in estimating tail behavior of a statistical distribution. Since there is no true VaR available to compare the accuracy of different approaches, we recommend that one applies several methods to gain insight into the range of VaR.

The choice of tail probability also plays an important role in VaR calculation. For the daily IBM stock returns, the sample size is 9190 so that the empirical quantiles of 5% and 1% are decent estimates of the quantiles of the return distribution. In this case, we can treat the results based on empirical quantiles as conservative estimates of the true VaR (i.e., lower bounds). In this view, the approach based on the traditional extreme value theory seems to underestimate the VaR for the daily log returns of IBM

stock. The conditional approach of extreme value theory discussed in the next section overcomes this weakness.

When the tail probability is small (e.g., 0.1%), the empirical quantile is a less reliable estimate of the true quantile. The VaR based on empirical quantiles can no longer serve as a lower bound of the true VaR. Finally, the earlier results show clearly the effects of using a heavy-tail distribution in VaR calculation when the tail probability is small. The VaR based on either a Student-$t$ distribution with 5 degrees of freedom or the extreme value distribution is greater than that based on the normal assumption when the probability is 0.1%.

### 7.6.2 Multiperiod VaR

The square root of time rule of the RiskMetrics methodology becomes a special case under the extreme value theory. The proper relationship between $\ell$-day and 1-day horizons is

$$\text{VaR}(\ell) = \ell^{1/\alpha}\text{VaR} = \ell^{-k}\text{VaR},$$

where $\alpha$ is the tail index and $k$ is the shape parameter of the extreme value distribution; see Danielsson and de Vries (1997a). This relationship is referred to as the $\alpha$-root of time rule. Here $\alpha = \frac{1}{k}$, not the scale parameter $\alpha_n$.

For illustration, consider the daily log returns of IBM stock in Example 7.6. If we use $p = 0.05$ and the results of $n = 21$, then for a 30-day horizon we have

$$\text{VaR}(30) = (30)^{0.335}\text{VaR} = 3.125 \times \$184{,}127 = \$575{,}397.$$

Because $\ell^{0.335} < \ell^{0.5}$, the $\alpha$-root of time rule produces lower $\ell$-day horizon VaR than does the square root of time rule.

### 7.6.3 VaR for a Short Position

In this subsection, we give the formulas of VaR calculation for holding short positions. Here the quantity of interest is the subperiod maximum and the limiting extreme value distribution becomes

$$F_*(r) = \begin{cases} \exp\left\{-\left[1 - \dfrac{k_n(r-\beta_n)}{\alpha_n}\right]^{1/k_n}\right\} & \text{if } k_n \neq 0 \\ \exp\left[-\exp\left(\dfrac{r-\beta_n}{\alpha_n}\right)\right] & \text{if } k_n = 0, \end{cases} \tag{7.27}$$

where $r$ denotes a value of the subperiod maximum and it is understood that $1 - k_n(r - \beta_n)/\alpha_n > 0$ for $k_n \neq 0$.

Following similar procedures as those of long positions, we obtain the $(1 - p)$th quantile of the return $r_t$ as

$$\text{VaR} = \begin{cases} \beta_n + \dfrac{\alpha_n}{k_n}\left\{1 - [-n \ln(1-p)]^{k_n}\right\} & \text{if } k_n \neq 0 \\ \beta_n + \alpha_n \ln[-n \ln(1-p)] & \text{if } k_n = 0, \end{cases} \tag{7.28}$$

where $p$ is a small probability denoting the chance of loss for holding a short position.

## 7.7 A NEW APPROACH BASED ON THE EXTREME VALUE THEORY

The aforementioned approach to VaR calculation using the extreme value theory encounters some difficulties. First, the choice of subperiod length $n$ is not clearly defined. Second, the approach is unconditional and, hence, does not take into consideration effects of other explanatory variables. To overcome these difficulties, a modern approach to extreme value theory has been proposed in the statistical literature; see Davison and Smith (1990) and Smith (1989). Instead of focusing on the extremes (maximum or minimum), the new approach focuses on exceedances of the measurement over some high threshold and the times at which the exceedances occur. For instance, consider the daily log returns $r_t$ of IBM stock used in this chapter and a long position on the stock. Let $\eta$ be a prespecified high threshold. We may choose $\eta = -2.5\%$. Suppose that the $i$th exceedance occurs at day $t_i$ (i.e., $r_{t_i} \leq \eta$). Then the new approach focuses on the data $(t_i, r_{t_i} - \eta)$. Here $r_{t_i} - \eta$ is the exceedance over the threshold $\eta$ and $t_i$ is the time at which the $i$th exceedance occurs. Similarly, for a short position, we may choose $\eta = 2\%$ and focus on the data $(t_i, r_{t_i} - \eta)$ for which $r_{t_i} \geq \eta$.

In practice, the occurrence times $\{t_i\}$ provide useful information about the intensity of the occurrence of important "rare events" (e.g., less than the threshold $\eta$ for a long position). A cluster of $t_i$ indicates a period of large market declines. The exceeding amount (or exceedance) $r_{t_i} - \eta$ is also of importance as it provides the actual quantity of interest.

Based on the prior introduction, the new approach does not require the choice of a subperiod length $n$, but it requires the specification of threshold $\eta$. Different choices of the threshold $\eta$ lead to different estimates of the shape parameter $k$ (and hence the tail index $-1/k$). In the literature, some researchers believe that the choice of $\eta$ is a statistical problem as well as a financial one, and it cannot be determined purely based on the statistical theory. For example, different financial institutions (or investors) have different risk tolerances. As such, they may select different thresholds even for an identical financial position. For the daily log returns of IBM stock considered in this chapter, the calculated VaR is not sensitive to the choice of $\eta$.

The choice of threshold $\eta$ also depends on the observed log returns. For a stable return series, $\eta = -2.5\%$ may fare well for a long position. For a volatile return

series (e.g., daily returns of a dot-com stock), $\eta$ may be as low as $-10\%$. Limited experience shows that $\eta$ can be chosen so that the number of exceedances is sufficiently large (e.g., about 5% of the sample). For a more formal study on the choice of $\eta$, see Danielsson and de Vries (1997b).

### 7.7.1 Statistical Theory

Again consider the log return $r_t$ of an asset. Suppose that the $i$th exceedance occurs at $t_i$. Focusing on the exceedance $r_t - \eta$ and exceeding time $t_i$ results in a fundamental change in statistical thinking. Instead of using the marginal distribution (e.g., the limiting distribution of the minimum or maximum), the new approach employs a conditional distribution to handle the magnitude of exceedance given that the measurement exceeds a threshold. The chance of exceeding the threshold is governed by a probability law. In other words, the new approach considers the conditional distribution of $x = r_t - \eta$ given $r_t \leq \eta$ for a long position. Occurrence of the event $\{r_t \leq \eta\}$ follows a point process (e.g., a Poisson process). See Section 6.9 for the definition of a Poisson process. In particular, if the intensity parameter $\lambda$ of the process is time-invariant, then the Poisson process is homogeneous. If $\lambda$ is time-variant, then the process is nonhomogeneous. The concept of Poisson process can be generalized to the multivariate case.

For ease in presentation, in what follows we use a positive threshold and the right-hand side of a return distribution to discuss the statistical theory behind the new approach of extreme value theory. This corresponds to holding a short financial position. However, the theory applies equally well to holding a long position if it is applied to the $r_t^c$ series, where $r_t^c = -r_t$. This is easily seen because $r_t^c \geq \eta$ for a positive threshold is equivalent to $r_t \leq -\eta$, where $-\eta$ becomes a negative threshold.

The basic theory of the new approach is to consider the conditional distribution of $r = x + \eta$ given $r > \eta$ for the limiting distribution of the maximum given in Eq. (7.27). Since there is no need to choose the subperiod length $n$, we do not use it as a subscript of the parameters. Then the conditional distribution of $r \leq x + \eta$ given $r > \eta$ is

$$\Pr(r \leq x + \eta \mid r > \eta) = \frac{\Pr(\eta \leq r \leq x + \eta)}{\Pr(r > \eta)} = \frac{\Pr(r \leq x + \eta) - \Pr(r \leq \eta)}{1 - \Pr(r \leq \eta)}. \tag{7.29}$$

Using the CDF $F_*(.)$ of Eq. (7.27) and the approximation $e^{-y} \approx 1 - y$ and after some algebra, we obtain that

$$\begin{aligned}\Pr(r \leq x + \eta \mid r > \eta) &= \frac{F_*(x + \eta) - F_*(\eta)}{1 - F_*(\eta)} \\ &= \frac{\exp\left\{-\left[1 - \frac{k(x+\eta-\beta)}{\alpha}\right]^{1/k}\right\} - \exp\left\{-\left[1 - \frac{k(\eta-\beta)}{\alpha}\right]^{1/k}\right\}}{1 - \exp\left\{-\left[1 - \frac{k(\eta-\beta)}{\alpha}\right]^{1/k}\right\}}\end{aligned}$$

$$\approx 1 - \left[1 - \frac{kx}{\alpha - k(\eta - \beta)}\right]^{1/k}, \tag{7.30}$$

where $x > 0$ and $1 - k(\eta - \beta)/\alpha > 0$. As is seen later, this approximation makes explicit the connection of the new approach to the traditional extreme value theory. The case of $k = 0$ is taken as the limit of $k \to 0$ so that

$$\Pr(r \leq x + \eta \mid r > \eta) \approx 1 - \exp(-x/\alpha).$$

### 7.7.2 A New Approach

Using the statistical result in Eq. (7.30) and considering jointly the exceedances and exceeding times, Smith (1989) proposes a two-dimensional Poisson process to model $(t_i, r_{t_i})$. This approach was used by Tsay (1999) to study VaR in risk management. We follow the same approach.

Assume that the baseline time interval is $T$, which is typically a year. In the United States, $T = 252$ is used as there are typically 252 trading days in a year. Let $t$ be the time interval of the data points (e.g., daily), and denote the data span by $t = 1, 2, \ldots, N$, where $N$ is the total number of data points. For a given threshold $\eta$, the exceeding times over the threshold are denoted by $\{t_i,\ i = 1, \ldots, N_\eta\}$ and the observed log return at $t_i$ is $r_{t_i}$. Consequently, we focus on modeling $\{(t_i, r_{t_i})\}$ for $i = 1, \ldots, N_\eta$, where $N_\eta$ depends on the threshold $\eta$.

The new approach to applying the extreme value theory is to postulate that the exceeding times and the associated returns [i.e., $(t_i, r_{t_i})$] jointly form a two-dimensional Poisson process with intensity measure given by

$$\Lambda[(T_2, T_1) \times (r, \infty)] = \frac{T_2 - T_1}{T} S(r; k, \alpha, \beta), \tag{7.31}$$

where

$$S(r; k, \alpha, \beta) = \left[1 - \frac{k(r - \beta)}{\alpha}\right]_+^{1/k},$$

$0 \leq T_1 \leq T_2 \leq N$, $r > \eta$, $\alpha > 0$, $\beta$, and $k$ are parameters, and the notation $[x]_+$ is defined as $[x]_+ = \max(x, 0)$. This intensity measure says that the occurrence of exceeding the threshold is proportional to the length of the time interval $[T_1, T_2]$ and the probability is governed by a survival function similar to the exponent of the CDF $F_*(r)$ in Eq. (7.27). A survival function of a random variable $X$ is defined as $S(x) = \Pr(X > x) = 1 - \Pr(X \leq x) = 1 - \text{CDF}(x)$. When $k = 0$, the intensity measure is taken as the limit of $k \to 0$—that is,

$$\Lambda[(T_2, T_1) \times (r, \infty)] = \frac{T_2 - T_1}{T} \exp\left[\frac{-(r - \beta)}{\alpha}\right].$$

In Eq. (7.31), the length of time interval is measured with respect to the baseline interval $T$.

The idea of using the intensity measure in Eq. (7.31) becomes clear when one considers its implied conditional probability of $r = x + \eta$ given $r > \eta$ over the time interval $[0, T]$, where $x > 0$,

$$\frac{\Lambda[(0,T)\times(x+\eta,\infty)]}{\Lambda[(0,T)\times(\eta,\infty)]} = \left[\frac{1-k(x+\eta-\beta)/\alpha}{1-k(\eta-\beta)/\alpha}\right]^{1/k} = \left[1-\frac{kx}{\alpha-k(\eta-\beta)}\right]^{1/k},$$

which is precisely the survival function of the conditional distribution given in Eq. (7.30). This survival function is obtained from the extreme limiting distribution for maximum in Eq. (7.27). We use survival function here because it denotes the probability of exceedance.

The relationship between the limiting extreme value distribution in Eq. (7.27) and the intensity measure in Eq. (7.31) directly connects the new approach of extreme value theory to the traditional one.

Mathematically, the intensity measure in Eq. (7.31) can be written as an integral of an intensity function:

$$\Lambda[(T_2,T_1)\times(r,\infty)] = \int_{T_1}^{T_2}\int_{r}^{\infty}\lambda(t,z;k,\alpha,\beta)\,dt\,dz,$$

where the intensity function $\lambda(t, z; k, \alpha, \beta)$ is defined as

$$\lambda(t,z;k,\alpha,\beta) = \frac{1}{T}g(z;k,\alpha,\beta) \tag{7.32}$$

where

$$g(z;k,\alpha,\beta) = \begin{cases} \dfrac{1}{\alpha}\left[1-\dfrac{k(z-\beta)}{\alpha}\right]^{1/k-1} & \text{if } k \neq 0 \\ \dfrac{1}{\alpha}\exp\left[\dfrac{-(z-\beta)}{\alpha}\right] & \text{if } k = 0. \end{cases}$$

Using the results of a Poisson process, we can write down the likelihood function for the observed exceeding times and their corresponding returns $\{(t_i, r_{t_i})\}$ over the two-dimensional space $[0, N] \times (\eta, \infty)$ as

$$L(k,\alpha,\beta) = \left(\prod_{i=1}^{N_\eta}\frac{1}{T}g(r_{t_i};k,\alpha,\beta)\right)\times\exp\left[-\frac{N}{T}S(\eta;k,\alpha,\beta)\right]. \tag{7.33}$$

The parameters $k, \alpha, \beta$ can then be estimated by maximizing the logarithm of this likelihood function. Since the scale parameter $\alpha$ is nonnegative, we use $\ln(\alpha)$ in the estimation.

**Table 7.3. Estimation Results of a Two-Dimensional Homogeneous Poisson Model for the Daily Negative Log Returns of IBM Stock From July 3, 1962 to December 31, 1998. The Baseline Time Interval is 252 (i.e., One Year). The Numbers in Parentheses Are Standard Errors, Where "Thr." and "Exc." Stand For Threshold and the Number of Exceedings.**

| Thr. | Exc. | Shape Par. $k$ | Log(Scale) $\ln(\alpha)$ | Location $\beta$ |
|---|---|---|---|---|
| | | (a) Original log returns | | |
| 3.0% | 175 | −0.30697(0.09015) | 0.30699(0.12380) | 4.69204(0.19058) |
| 2.5% | 310 | −0.26418(0.06501) | 0.31529(0.11277) | 4.74062(0.18041) |
| 2.0% | 554 | −0.18751(0.04394) | 0.27655(0.09867) | 4.81003(0.17209) |
| | | (b) Removing the sample mean | | |
| 3.0% | 184 | −0.30516(0.08824) | 0.30807(0.12395) | 4.73804(0.19151) |
| 2.5% | 334 | −0.28179(0.06737) | 0.31968(0.12065) | 4.76808(0.18533) |
| 2.0% | 590 | −0.19260(0.04357) | 0.27917(0.09913) | 4.84859(0.17255) |

**Example 7.7.** Consider again the daily log returns of IBM stock from July 3, 1962 to December 31, 1998. There are 9190 daily returns. Table 7.3 gives some estimation results of the parameters $k, \alpha, \beta$ for three choices of the threshold when the negative series $\{-r_t\}$ is used. We use the negative series $\{-r_t\}$, instead of $\{r_t\}$, because we focus on holding a long financial position. The table also shows the number of exceeding times for a given threshold. It is seen that the chance of dropping 2.5% or more in a day for IBM stock occurred with probability $310/9190 \approx 3.4\%$. Because the sample mean of IBM stock returns is not zero, we also consider the case when the sample mean is removed from the original daily log returns. From the table, removing the sample mean has little impact on the parameter estimates. These parameter estimates are used next to calculate VaR, keeping in mind that in a real application one needs to check carefully the adequacy of a fitted Poisson model. We discuss methods of model checking in the next subsection.

### 7.7.3 VaR Calculation Based on the New Approach

As shown in Eq. (7.30), the two-dimensional Poisson process model used, which employs the intensity measure in Eq. (7.31), has the same parameters as those of the extreme value distribution in Eq. (7.27). Therefore, one can use the same formula as that of the Eq. (7.28) to calculate VaR of the new approach. More specifically, for a given upper tail probability $p$, the $(1-p)$th quantile of the log return $r_t$ is

$$\text{VaR} = \begin{cases} \beta + \frac{\alpha}{k}\left\{1 - [-T\ln(1-p)]^k\right\} & \text{if } k \neq 0 \\ \beta + \alpha \ln[-T\ln(1-p)] & \text{if } k = 0, \end{cases} \tag{7.34}$$

where $T$ is the baseline time interval used in estimation. Typically, $T = 252$ in the United States for the approximate number of trading days in a year.

**Example 7.8.** Consider again the case of holding a long position of IBM stock valued at \$10 million. We use the estimation results of Table 7.3 to calculate 1-day horizon VaR for the tail probabilities of 0.05 and 0.01.

- Case I: Use the original daily log returns. The three choices of threshold $\eta$ result in the following VaR values:

  **1.** $\eta = 3.0\%$: VaR(5%) = \$228,239, VaR(1%) = \$359,303.
  **2.** $\eta = 2.5\%$: VaR(5%) = \$219,106, VaR(1%) = \$361,119.
  **3.** $\eta = 2.0\%$: VaR(5%) = \$212,981, VaR(1%) = \$368,552.

- Case II: The sample mean of the daily log returns is removed. The three choices of threshold $\eta$ result in the VaR values:

  **1.** $\eta = 3.0\%$: VaR(5%) = \$232,094, VaR(1%) = \$363,697.
  **2.** $\eta = 2.5\%$: VaR(5%) = \$225,782, VaR(1%) = \$364,254.
  **3.** $\eta = 2.0\%$: VaR(5%) = \$217,740, VaR(1%) = \$372,372.

As expected, removing the sample mean, which is positive, increases slightly the VaR. However, the VaR is rather stable among the three threshold values used. In practice, we recommend that one removes the sample mean first before applying this new approach to VaR calculation.

**Discussion:** Compared with the VaR of Example 7.6 that uses the traditional extreme value theory, the new approach provides a more stable VaR calculation. The traditional approach is rather sensitive to the choice of the subperiod length $n$.

### 7.7.4 Use of Explanatory Variables

The two-dimensional Poisson process model discussed earlier is *homogeneous* because the three parameters $k$, $\alpha$, and $\beta$ are constant over time. In practice, such a model may not be adequate. Furthermore, some explanatory variables are often available that may influence the behavior of the log returns $r_t$. A nice feature of the new extreme value theory approach to VaR calculation is that it can easily take explanatory variables into consideration. We discuss such a framework in this subsection. In addition, we also discuss methods that can be used to check the adequacy of a fitted two-dimensional Poisson process model.

Suppose that $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{vt})'$ is a vector of $v$ explanatory variables that are available *prior to* time $t$. For asset returns, the volatility $\sigma_t^2$ of $r_t$ discussed in Chapter 3 is an example of explanatory variables. Another example of explanatory variables in the U.S. equity markets is an indicator variable denoting the meetings of Federal Open Market Committee. A simple way to make use of explanatory variables is to postulate that the three parameters $k$, $\alpha$, and $\beta$ are time-varying and are linear functions of the explanatory variables. Specifically, when explanatory variables $\boldsymbol{x}_t$

are available, we assume that

$$\begin{aligned} k_t &= \gamma_0 + \gamma_1 x_{1t} + \cdots + \gamma_v x_{vt} \equiv \gamma_0 + \boldsymbol{\gamma}' \boldsymbol{x}_t \\ \ln(\alpha_t) &= \delta_0 + \delta_1 x_{1t} + \cdots + \delta_v x_{vt} \equiv \delta_0 + \boldsymbol{\delta}' \boldsymbol{x}_t \\ \beta_t &= \theta_0 + \theta_1 x_{1t} + \cdots + \theta_v x_{vt} \equiv \theta_0 + \boldsymbol{\theta}' \boldsymbol{x}_t. \end{aligned} \tag{7.35}$$

If $\boldsymbol{\gamma} = \mathbf{0}$, then the shape parameter $k_t = \gamma_0$, which is time-invariant. Thus, testing the significance of $\boldsymbol{\gamma}$ can provide information about the contribution of the explanatory variables to the shape parameter. Similar methods apply to the scale and location parameters. In Eq. (7.35), we use the same explanatory variables for all the three parameters $k_t$, $\ln(\alpha_t)$, and $\beta_t$. In an application, different explanatory variables may be used for different parameters.

When the three parameters of the extreme value distribution are time-varying, we have an *inhomogeneous* Poisson process. The intensity measure becomes

$$\Lambda[(T_1, T_2) \times (r, \infty)] = \frac{T_2 - T_1}{T} \left(1 - \frac{k_t(r - \beta_t)}{\alpha_t}\right)_+^{1/k_t}, \quad r > \eta. \tag{7.36}$$

The likelihood function of the exceeding times and returns $\{(t_i, r_{t_i})\}$ becomes

$$L = \left(\prod_{i=1}^{N_\eta} \frac{1}{T} g(r_{t_i}; k_{t_i}, \alpha_{t_i}, \beta_{t_i})\right) \times \exp\left[-\frac{1}{T}\int_0^N S(\eta; k_t, \alpha_t, \beta_t)dt\right],$$

which reduces to

$$L = \left(\prod_{i=1}^{N_\eta} \frac{1}{T} g(r_{t_i}; k_{t_i}, \alpha_{t_i}, \beta_{t_i})\right) \times \exp\left[-\frac{1}{T}\sum_{t=1}^{N} S(\eta; k_t, \alpha_t, \beta_t)\right] \tag{7.37}$$

if one assumes that the parameters $k_t$, $\alpha_t$, and $\beta_t$ are constant within each trading day, where $g(z; k_t, \alpha_t, \beta_t)$ and $S(\eta; k_t, \alpha_t, \beta_t)$ are given in Eqs. (7.32) and (7.31), respectively. For given observations $\{r_t, \boldsymbol{x}_t \mid t = 1, \ldots, N\}$, the baseline time interval $T$, and the threshold $\eta$, the parameters in Eq. (7.35) can be estimated by maximizing the logarithm of the likelihood function in Eq. (7.37). Again we use $\ln(\alpha_t)$ to satisfy the positive constraint of $\alpha_t$.

***Remark:*** The parameterization in Eq. (7.35) is similar to that of the volatility models of Chapter 3 in the sense that the three parameters are exact functions of the available information at time $t$. Other functions can be used if necessary.

### 7.7.5 Model Checking

Checking an entertained two-dimensional Poisson process model for exceedance times and excesses involves examining three key features of the model. The first

feature is to verify the adequacy of the exceedance rate, the second feature is to examine the distribution of exceedances, and the final feature is to check the independence assumption of the model. We discuss briefly some statistics that are useful for checking these three features. These statistics are based on some basic statistical theory concerning distributions and stochastic processes.

***Exceedance Rate***

A fundamental property of univariate Poisson processes is that the time durations between two consecutive events are independent and exponentially distributed. To exploit a similar property for checking a two-dimensional process model, Smith and Shively (1995) propose to examine the time durations between consecutive exceedances. If the two-dimensional Poisson process model is appropriate for the exceedance times and excesses, the time duration between the $i$th and $(i-1)$th exceedances should follow an exponential distribution. More specifically, letting $t_0 = 0$, we expect that

$$z_{t_i} = \int_{t_{i-1}}^{t_i} \frac{1}{T} g(\eta; k_s, \alpha_s, \beta_s) ds, \quad i = 1, 2, \ldots$$

are independent and identically distributed (iid) as a standard exponential distribution. Because daily returns are discrete-time observations, we employ the time durations

$$z_{t_i} = \frac{1}{T} \sum_{t=t_{i-1}+1}^{t_i} S(\eta; k_t, \alpha_t, \beta_t) \tag{7.38}$$

and use the quantile-to-quantile (QQ) plot to check the validity of the iid standard exponential distribution. If the model is adequate, the QQ-plot should show a straight line through the origin with unit slope.

***Distribution of Excesses***

Under the two-dimensional Poisson process model considered, the conditional distribution of the excess $x_t = r_t - \eta$ over the threshold $\eta$ is a generalized Pareto distribution (GPD) with shape parameter $k_t$ and scale parameter $\psi_t = \alpha_t - k_t(\eta - \beta_t)$. Therefore, we can make use of the relationship between a standard exponential distribution and GPD, and define

$$w_{t_i} = \begin{cases} \dfrac{-1}{k_{t_i}} \ln\left(1 - k_{t_i} \dfrac{r_{t_i} - \eta}{\psi_{t_i}}\right)_+ & \text{if } k_{t_i} \neq 0 \\ \dfrac{r_{t_i} - \eta}{\psi_{t_i}} & \text{if } k_{t_i} = 0. \end{cases} \tag{7.39}$$

If the model is adequate, $\{w_{t_i}\}$ are independent and exponentially distributed with mean 1; see also Smith (1999). We can then apply the QQ-plot to check the validity of the GPD assumption for excesses.

***Independence***

A simple way to check the independence assumption, after adjusting for the effects of explanatory variables, is to examine the sample autocorrelation functions of $z_{t_i}$ and $w_{t_i}$. Under the independence assumption, we expect zero serial correlations for both $z_{t_i}$ and $w_{t_i}$.

### 7.7.6 An Illustration

In this subsection, we apply a two-dimensional inhomogeneous Poisson process model to the daily log returns, in percentages, of IBM stock from July 3, 1962 to December 31, 1998. We focus on holding a long position of \$10 million. The analysis enables us to compare the results with those obtained before by using other approaches to calculating VaR.

We begin by pointing out that the two-dimensional homogeneous model of Example 7.7 needs further refinements because the fitted model fails to pass the model checking statistics of the previous subsection. Figures 7.5(a) and (b) show the auto-

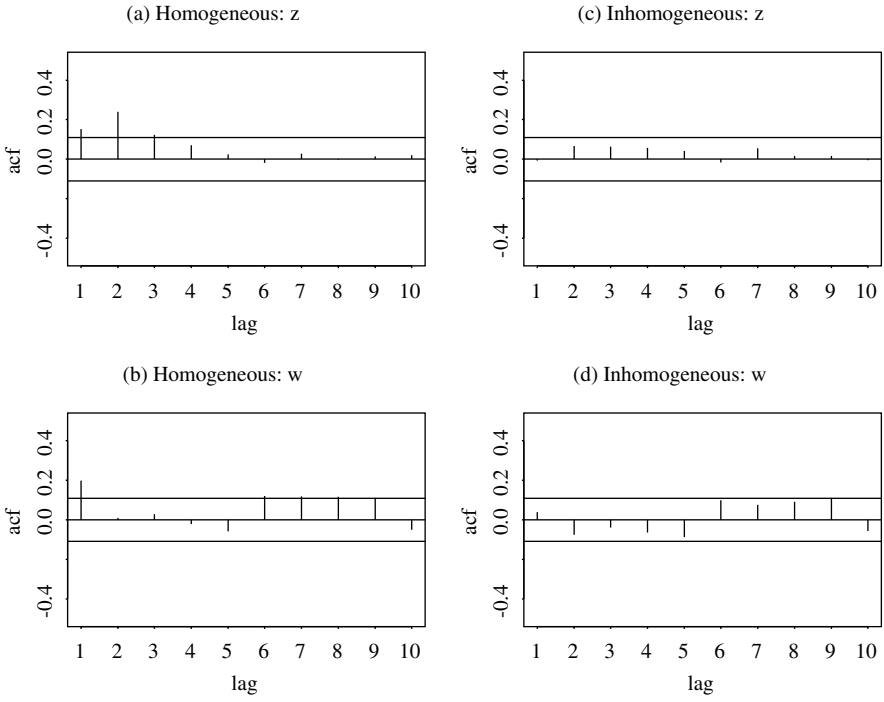


**Figure 7.5.** Sample autocorrelation functions of the $z$ and $w$ measures for two-dimensional Poisson models. (a) and (b) are for the homogeneous model, and (c) and (d) are for the inhomogeneous model. The data are daily mean-corrected log returns, in percentages, of IBM stock from July 3, 1962 to December 31, 1998, and the threshold is 2.5%. A long financial position is used.

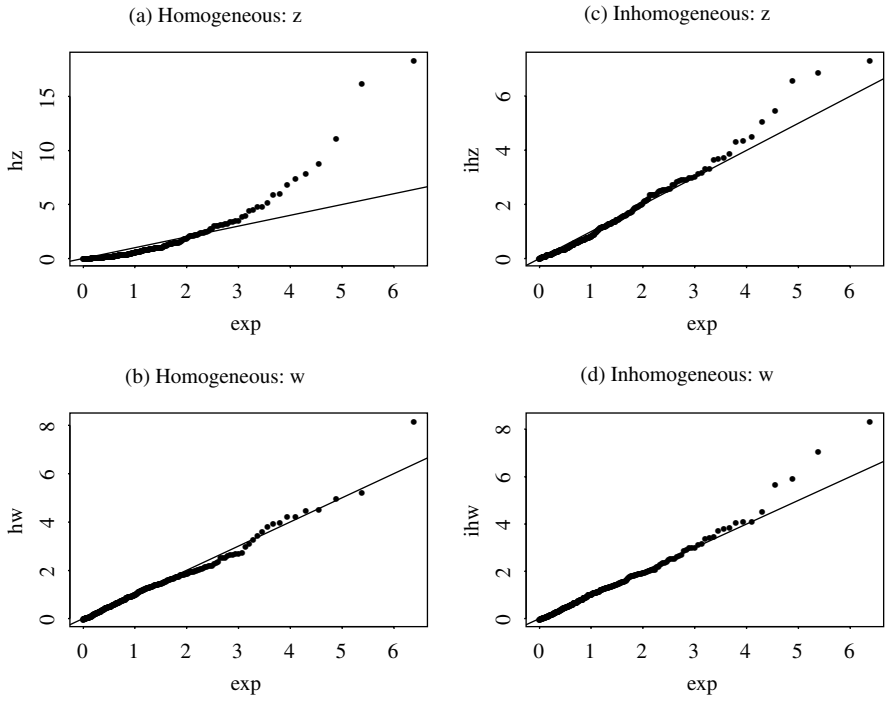


**Figure 7.6.** Quantile-to-quantile plot of the $z$ and $w$ measures for two-dimensional Poisson models. (a) and (b) are for the homogeneous model, and (c) and (d) are for the inhomogeneous model. The data are daily mean-corrected log returns, in percentages, of IBM stock from July 3, 1962 to December 31, 1998, and the threshold is 2.5%. A long financial position is used.

correlation functions of the statistics $z_{t_i}$ and $w_{t_i}$, defined in Eqs. (7.38) and (7.39), of the homogeneous model when the threshold is $\eta = 2.5\%$. The horizontal lines in the plots denote asymptotic limits of two standard errors. It is seen that both $z_{t_i}$ and $w_{t_i}$ series have some significant serial correlations. Figures 7.6(a) and (b) show the QQ-plots of $z_{t_i}$ and $w_{t_i}$ series. The straight line in each plot is the theoretical line, which passes through the origin and has a unit slope under the assumption of a standard exponential distribution. The QQ-plot of $z_{t_i}$ shows some discrepancy.

To refine the model, we use the mean-corrected log return series

$$r_t^o = r_t - \bar{r}, \quad \bar{r} = \frac{1}{9190} \sum_{t=1}^{9190} r_t,$$

where $r_t$ is the daily log return in percentages, and employ the following explanatory variables:

1. $x_{1t}$: an indicator variable for October, November, and December. That is, $x_{1t} = 1$ if $t$ is in October, November, or December. This variable is chosen to take care of the fourth-quarter effect (or year-end effect), if any, on the daily IBM stock returns.
2. $x_{2t}$: an indicator variable for the behavior of the previous trading day. Specifically, $x_{2t} = 1$ if and only if the log return $r^o_{t-1} \leq -2.5\%$. Since we focus on holding a long position with threshold 2.5%, an exceedance occurs when the daily price drops over 2.5%. Therefore, $x_{2t}$ is used to capture the possibility of panic selling when the price of IBM stock dropped 2.5% or more on the previous trading day.
3. $x_{3t}$: a qualitative measurement of volatility, which is the number of days between $t-1$ and $t-5$ (inclusive) that has a log return with magnitude exceeding the threshold. In our case, $x_{3t}$ is the number of $r^o_{t-i}$ satisfying $|r^o_{t-i}| \geq 2.5\%$ for $i = 1, \ldots, 5$.
4. $x_{4t}$: an annual trend defined as $x_{4t} =$ (year of time $t - 1961)/38$. This variable is used to detect any trend in the behavior of extreme returns of IBM stock.
5. $x_{5t}$: a volatility series based on a Gaussian GARCH(1, 1) model for the mean-corrected series $r^o_t$. Specifically, $x_{5t} = \sigma_t$, where $\sigma^2_t$ is the conditional variance of the GARCH(1, 1) model

$$r^o_t = a_t, \quad a_t = \sigma_t \epsilon_t, \quad \epsilon_t \sim N(0, 1)$$
$$\sigma^2_t = 0.04565 + 0.0807 a^2_{t-1} + 0.9031 \sigma^2_{t-1}.$$

These five explanatory variables are all available at time $t-1$. We use two volatility measures ($x_{3t}$ and $x_{5t}$) to study the effect of market volatility on VaR. As shown in Example 7.3 by the fitted AR(2)-GARCH(1, 1) model, the serial correlations in $r_t$ are weak so that we do not entertain any ARMA model for the mean equation.

Using the prior five explanatory variables and deleting insignificant parameters, we obtain the estimation results shown in Table 7.4. Figures 7.5(c) and (d) and Figures 7.6(c) and (d) show the model checking statistics for the fitted two-dimensional inhomogeneous Poisson process model when the threshold is $\eta = 2.5\%$. All autocorrelation functions of $z_{t_i}$ and $w_{t_i}$ are within the asymptotic two standard-error limits. The QQ-plots also show marked improvements as they indicate no model inadequacy. Based on these checking results, the inhomogeneous model seems adequate.

Consider the case of threshold 2.5%. The estimation results show the following:

1. All three parameters of the intensity function depend significantly on the annual time trend. In particular, the shape parameter has a negative annual trend, indicating that the log returns of IBM stock are moving farther away from normality as time passes. Both the location and scale parameters increase over time.
2. Indicators for the fourth quarter, $x_{1t}$, and for panic selling, $x_{2t}$, are not significant for all three parameters.

**Table 7.4. Estimation Results of a Two-Dimensional Inhomogeneous Poisson Process Model for Daily Log Returns, in Percentages, of IBM Stock From July 3, 1962 to December 31, 1998. Four Explanatory Variables Defined in the Text Are Used. The Model is for Holding a Long Position on IBM Stock. The Sample Mean of the Log Returns is Removed From the Data.**

| Parameter | Constant | Coef. of $x_{3t}$ | Coef. of $x_{4t}$ | Coef. of $x_{5t}$ |
|---|---|---|---|---|
| | (a) Threshold 2.5% with 334 exceedances | | | |
| $\beta_t$ | 0.3202 | | 1.4772 | 2.1991 |
| (Std.err) | (0.3387) | | (0.3222) | (0.2450) |
| $\ln(\alpha_t)$ | −0.8119 | 0.3305 | 1.0324 | |
| (Std.err) | (0.1798) | (0.0826) | (0.2619) | |
| $k_t$ | −0.1805 | −0.2118 | −0.3551 | 0.2602 |
| (Std.err) | (0.1290) | (0.0580) | (0.1503) | (0.0461) |
| | (b) Threshold 3.0% with 184 exceedances | | | |
| $\beta_t$ | 1.1569 | | | 2.1918 |
| (Std.err) | (0.4082) | | | (0.2909) |
| $\ln(\alpha_t)$ | −0.0316 | 0.3336 | | |
| (Std.err) | (0.1201) | (0.0861) | | |
| $k_t$ | −0.6008 | −0.2480 | | 0.3175 |
| (Std.err) | (0.1454) | (0.0731) | | (0.0685) |

3. The location and shape parameters are positively affected by the volatility of the GARCH(1, 1) model; see the coefficients of $x_{5t}$. This is understandable because the variability of log returns increases when the volatility is high. Consequently, the dependence of log returns on the tail index is reduced.
4. The scale and shape parameters depend significantly on the qualitative measure of volatility. The signs of the estimates are also plausible.

The explanatory variables for December 31, 1998 assumed the values $x_{3,9190} = 0$, $x_{4,9190} = 0.9737$, and $x_{5,9190} = 1.9766$. Using these values and the fitted model in Table 7.4, we obtain

$$k_{9190} = -0.01195, \quad \ln(\alpha_{9190}) = 0.19331, \quad \beta_{9190} = 6.105.$$

Assume that the tail probability is 0.05. The VaR quantile shown in Eq. (7.34) gives VaR = 3.03756%. Consequently, for a long position of \$10 million, we have

$$\text{VaR} = \$10{,}000{,}000 \times 0.0303756 = \$303{,}756.$$

If the tail probability is 0.01, the VaR is \$497,425. The 5% VaR is slightly larger than that of Example 7.3, which uses a Gaussian AR(2)-GARCH(1, 1) model. The 1% VaR is larger than that of Case I of Example 7.3. Again, as expected, the effect of extreme values (i.e., heavy tails) on VaR is more pronounced when the tail probability used is small.

An advantage of using explanatory variables is that the parameters are adaptive to the change in market conditions. For example, the explanatory variables for December 30, 1998 assumed the values $x_{3,9189} = 1$, $x_{4,9189} = 0.9737$, and $x_{5,9189} = 1.8757$. In this case, we have

$$k_{9189} = -0.2500, \quad \ln(\alpha_{9189}) = 0.52385, \quad \beta_{9189} = 5.8834.$$

The 95% quantile (i.e., the tail probability is 5%) then becomes 2.69139%. Consequently, the VaR is

$$\text{VaR} = \$10{,}000{,}000 \times 0.0269139 = \$269{,}139.$$

If the tail probability is 0.01, then VaR becomes \$448,323. Based on this example, the homogeneous Poisson model shown in Example 7.8 seems to underestimate the VaR.

## EXERCISES

**1.** Consider the daily log returns of GE stock from July 3, 1962 to December 31, 1999. The data can be obtained from CRSP or the file "d-geln.dat." Suppose that you hold a long position on the stock valued at \$1 million. Use the tail probability 0.05. Compute the value at risk of your position for 1-day horizon and 15-day horizon using the following methods:

**(a)** The RiskMetrics method.

**(b)** A Gaussian ARMA-GARCH model.

**(c)** An ARMA-GARCH model with a Student-$t$ distribution. You should also estimate the degrees of freedom.

**(d)** The traditional extreme value theory with subperiod length $n = 21$.

**2.** The file "d-csco9199.dat" contains the daily log returns of Cisco Systems stock from 1991 to 1999 with 2275 observations. Suppose that you hold a long position of Cisco stock valued at \$1 million. Compute the Value at Risk of your position for the next trading day using probability $p = 0.01$.

**(a)** Use the RiskMetrics method.

**(b)** Use a GARCH model with a conditional Gaussian distribution.

**(c)** Use a GARCH model with a Student-$t$ distribution. You may also estimate the degrees of freedom.

**(d)** Use the unconditional sample quantile.

**(e)** Use a two-dimensional homogeneous Poisson process with threshold 2%. That is, focusing on the exceeding times and exceedances that the daily stock price drops 2% or more. Check the fitted model.

**(f)** Use a two-dimensional nonhomogeneous Poisson process with threshold 2%. The explanatory variables are (1) an annual time trend, (2) a dummy variable for October, November, and December, and (3) a fitted volatility based on

a Gaussian GARCH(1, 1) model. Perform a diagnostic check on the fitted model.

**(g)** Repeat the prior two-dimensional nonhomogeneous Poisson process with threshold 2.5% or 3%. Comment on the selection of threshold.

**3.** Use Hill's estimator and the data "d-csco9199.dat" to estimate the tail index for daily stock returns of Cisco Systems.

**4.** The file "d-hwp3dx8099.dat" contains the daily log returns of Hewlett-Packard, CRSP value-weighted index, equal-weighted index, and S&P 500 index from 1980 to 1999. All returns are in percentages and include dividend distributions. Assume that the tail probability of interest is 0.01. Calculate Value at Risk for the following financial positions for the first trading day of year 2000.

**(a)** Long on Hewlett-Packard stock of \$1 million and S&P 500 index of \$1 million, using RiskMetrics. The $\alpha$ coefficient of the IGARCH(1, 1) model for each series should be estimated.

**(b)** The same position as part (a), but using a univariate ARMA-GARCH model for each return series.

**(c)** A long position on Hewlett-Packard stock of \$1 million using a two-dimensional nonhomogeneous Poisson model with the following explanatory variables: (1) an annual time trend, (2) a fitted volatility based on a Gaussian GARCH model for Hewlett-Packard stock, (3) a fitted volatility based on a Gaussian GARCH model for S&P 500 index returns, and (4) a fitted volatility based on a Gaussian GARCH model for the value-weighted index return. Perform a diagnostic check for the fitted models. Are the market volatility as measured by S&P 500 index and value-weighted index returns helpful in determining the tail behavior of stock returns of Hewlett-Packard? You may choose several thresholds.

## REFERENCES

Berman, S. M. (1964), "Limiting theorems for the maximum term in stationary sequences," *Annals of Mathematical Statistics*, 35, 502–516.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman and Hall.

Danielsson, J., and De Vries, C. G. (1997a), "Value at risk and extreme returns," working paper, London School of Economics, London, U.K.

Danielsson, J., and De Vries, C. G. (1997b), "Tail index and quantile estimation with very high frequency data," *Journal of Empirical Finance*, 4, 241–257.

Davison, A. C., and Smith, R. L. (1990), "Models for exceedances over high thresholds," (with discussion), *Journal of the Royal Statistical Society, Series B*, 52, 393–442.

De Haan L., Resnick, I. S., Rootzén, and De Vries, C. G. (1989), "Extremal behavior of solutions to a stochastic difference equation with applications to ARCH process," *Stochastic Processes and Their Applications*, 32, 213–224.

Dekkers, A. L. M., and De Haan, L. (1989), "On the estimation of extreme value index and large quantile estimation," *Annals of Statistics*, 17, 1795–1832.

Duffie, D., and Pan, J. (1997), "An overview of value at risk," *Journal of Derivatives*, Spring, 7–48.

Embrechts, P., Kuppelberg, C., and Mikosch, T. (1997), *Modelling Extremal Events*, Berlin: Springer Verlag.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Vol. 2, New York: Wiley.

Goldie, C. M., and Smith, R. L. (1987), "Slow variation with remainder: Theory and applications," *Quarterly Journal of Mathematics*, Oxford 2nd series, 38, 45–71.

Gnedenko, B. V. (1943), "Sur la distribution limite du terme maximum of d'une série Aléatorie," *Annals of Mathematics*, **44**, 423–453.

Gumbel, E. J. (1958), *Statistics of Extremes*, New York: Columbia University Press.

Hill, B. M. (1975), "A simple general approach to inference about the tail of a distribution," *Annals of Statistics*, 3, 1163–1173.

Jenkinson, A. F. (1955), "The frequency distribution of the annual maximum (or minimum) of meteorological elements," *Quarterly Journal of the Royal Meteorological Society*, **81**, 158–171.

Jorion, P. (1997), *Value at Risk: The New Benchmark for Controlling Market Risk*. The McGraw-Hill Company: Chicago.

Koenker, R. W., and Bassett, G. W. (1978), "Regression quantiles," *Econometrica*, 46, 33–50.

Koenker, R. W., and D'Orey, V. (1987), "Computing regression quantiles," *Applied Statistics*, 36, 383–393.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983), *Extremes and Related Properties of Random Sequences and Processes*, New York: Springer Verlag.

Longerstaey, J., and More, L. (1995), "Introduction to RiskMetrics$^{TM}$," 4th edition, Morgan Guaranty Trust Company: New York.

Longin, F. M. (1996), "The asymptotic distribution of extreme stock market returns," *Journal of Business*, 69, 383–408.

Longin, F. M. (1999a), "Optimal margin level in futures markets: Extreme price movements," *The Journal of Futures Markets*, 19, 127–152.

Longin, F. M. (1999b), "From value at risk to stress testing: The extreme value approach," working paper, Centre for Economic Policy Research, London, UK.

Pickands, J. (1975), "Statistical inference using extreme order statistics," *Annals of Statistics*, 3, 119–131.

Smith, R. L. (1989), "Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone" (with discussion), *Statistical Science*, 4, 367–393.

Smith, R. L. (1999), "Measuring risk with extreme value theory," working paper, Department of Statistics, University of North Carolina at Chapel Hill.

Smith, R. L., and Shively, T. S. (1995), "A point process approach to modeling trends in troposheric ozone," *Atmospheric Environment*, **29**, 3489–3499.

Tsay, R. S. (1999), "Extreme value analysis of financial data," working paper, Graduate School of Business, University of Chicago.

*Analysis of Financial Time Series.* Ruey S. Tsay

ISBN: 0-471-41544-8

CHAPTER 8

# Multivariate Time Series Analysis and Its Applications

Economic globalization and internet communication have accelerated the integration of world financial markets in recent years. Price movements in one market can spread easily and instantly to another market. For this reason, financial markets are more dependent on each other than ever before, and one must consider them jointly to better understand the dynamic structure of the global finance. One market may lead the other market under some circumstances, yet the relationship may be reversed under other circumstances. Consequently, knowing how the markets are interrelated is of great importance in finance. Similarly, for an investor or a financial institution holding multiple assets, the dynamic relationships between returns of the assets play an important role in decision making. In this and the next chapters, we introduce econometric models and methods useful for studying jointly multiple return series. In the statistical literature, these models and methods belong to vector or multivariate time series analysis.

A multivariate time series consists of multiple single series referred to as *components*. As such, concepts of vector and matrix are important in multivariate time series analysis. We use boldfaced notation to indicate vectors and matrixes. If necessary, readers may consult Appendix A of the chapter for some basic operations and properties of vector and matrix. Appendix B provides some results of multivariate normal distribution, which is widely used in multivariate statistical analysis (e.g., Johnson and Wichern, 1998).

Let $\boldsymbol{r}_t = (r_{1t}, r_{2t}, \ldots, r_{kt})'$ be the log returns of $k$ assets at time $t$, where $\boldsymbol{a}'$ denotes the transpose of $\boldsymbol{a}$. For example, an investor holding stocks of IBM, Microsoft, Exxon Mobil, General Motors, and Wal-Mart Stores may consider the five-dimensional daily log returns of these companies. Here $r_{1t}$ denotes the daily log return of IBM stock, $r_{2t}$ is that of Microsoft, and so on. As a second example, an investor who is interested in global investment may consider the return series of the S&P 500 index of United States, the FTSE 100 index of United Kingdom, and the Nikkei 225 index of Japan. Here the series is three-dimensional, with $r_{1t}$ denoting the return of S&P 500 index, $r_{2t}$ the return of FTSE 100 index, and $r_{3t}$ the return of Nikkei 225. The goal of this chapter is to study econometric models for analyzing

the multivariate process $\boldsymbol{r}_t$. The chapter also discusses methods that can simplify the dynamic structure or reduce the dimension of $\boldsymbol{r}_t$.

Many of the models and methods discussed in the previous chapters can be generalized directly to the multivariate case. But there are situations in which the generalization requires some attention. In some situations, one needs new models and methods to handle the complicated relationships between multiple returns. We also discuss methods that search for common factors affecting the returns of different assets. Our discussion emphasizes intuition and applications. For statistical theory of multivariate time series analysis, readers are referred to Lütkepohl (1991) and Reinsel (1993).

## 8.1 WEAK STATIONARITY AND CROSS-CORRELATION MATRIXES

Consider a $k$-dimensional time series $\boldsymbol{r}_t = (r_{1t}, \ldots, r_{kt})'$. The series $\boldsymbol{r}_t$ is *weakly stationary* if its first and second moments are time-invariant. In particular, the mean vector and covariance matrix of a weakly stationary series are constant over time. Unless stated explicitly to the contrary, we assume that the return series of financial assets are weakly stationary.

For a weakly stationary time series $\boldsymbol{r}_t$, we define its mean vector and covariance matrix as

$$\boldsymbol{\mu} = E(\boldsymbol{r}_t), \quad \boldsymbol{\Gamma}_0 = E[(\boldsymbol{r}_t - \boldsymbol{\mu}_t)(\boldsymbol{r}_t - \boldsymbol{\mu}_t)'], \tag{8.1}$$

where the expectation is taken element by element over the joint distribution of $\boldsymbol{r}_t$. The mean $\boldsymbol{\mu}$ is a $k$-dimensional vector consisting of the unconditional expectations of the components of $\boldsymbol{r}_t$. The covariance matrix $\boldsymbol{\Gamma}_0$ is a $k \times k$ matrix. The $i$th diagonal element of $\boldsymbol{\Gamma}_0$ is the variance of $r_{it}$, whereas the $(i, j)$th element of $\boldsymbol{\Gamma}_0$ is the covariance between $r_{it}$ and $r_{jt}$. We write $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ and $\boldsymbol{\Gamma}_0 = [\Gamma_{ij}(0)]$ when the elements are needed.

### 8.1.1 Cross-Correlation Matrixes

Let $\boldsymbol{D}$ be a $k \times k$ diagonal matrix consisting of the standard deviations of $r_{it}$ for $i = 1, \ldots, k$. In other words, $\boldsymbol{D} = \text{diag}\{\sqrt{\Gamma_{11}(0)}, \ldots, \sqrt{\Gamma_{kk}(0)}\}$. The concurrent, or lag-zero, cross-correlation matrix of $\boldsymbol{r}_t$ is defined as

$$\boldsymbol{\rho}_0 \equiv [\rho_{ij}(0)] = \boldsymbol{D}^{-1}\boldsymbol{\Gamma}_0\boldsymbol{D}^{-1}.$$

More specifically, the $(i, j)$th element of $\boldsymbol{\rho}_0$ is

$$\rho_{ij}(0) = \frac{\Gamma_{ij}(0)}{\sqrt{\Gamma_{ii}(0)\Gamma_{jj}(0)}} = \frac{\text{Cov}(r_{it}, r_{jt})}{\text{std}(r_{it})\text{std}(r_{jt})},$$

which is the correlation coefficient between $r_{it}$ and $r_{jt}$. In a time series analysis, such a correlation coefficient is referred to as a concurrent, or contemporaneous,

correlation coefficient because it is the correlation of the two series at time $t$. It is easy to see that $\rho_{ij}(0) = \rho_{ji}(0)$, $-1 \leq \rho_{ij}(0) \leq 1$, and $\rho_{ii}(0) = 1$ for $1 \leq i, j \leq k$. Thus, $\boldsymbol{\rho}(0)$ is a symmetric matrix with unit diagonal elements.

An important topic in multivariate time series analysis is the lead-lag relationships between component series. To this end, the cross-correlation matrixes are used to measure the strength of linear dependence between time series. The lag-$\ell$ cross-covariance matrix of $\boldsymbol{r}_t$ is defined as

$$\boldsymbol{\Gamma}_\ell \equiv [\Gamma_{ij}(\ell)] = E[(\boldsymbol{r}_t - \boldsymbol{\mu})(\boldsymbol{r}_{t-\ell} - \boldsymbol{\mu})'], \tag{8.2}$$

where $\boldsymbol{\mu}$ is the mean vector of $\boldsymbol{r}_t$. Therefore, the $(i, j)$th element of $\boldsymbol{\Gamma}_\ell$ is the covariance between $r_{it}$ and $r_{j,t-\ell}$. For a weakly stationary series, the cross-covariance matrix $\boldsymbol{\Gamma}_\ell$ is a function of $\ell$, not the time index $t$.

The lag-$\ell$ cross-correlation matrix (CCM) of $\boldsymbol{r}_t$ is defined as

$$\boldsymbol{\rho}_\ell \equiv [\rho_{ij}(\ell)] = \boldsymbol{D}^{-1}\boldsymbol{\Gamma}_\ell\boldsymbol{D}^{-1}, \tag{8.3}$$

where, as before, $\boldsymbol{D}$ is the diagonal matrix of standard deviations of the individual series $r_{it}$. From the definition,

$$\rho_{ij}(\ell) = \frac{\Gamma_{ij}(\ell)}{\sqrt{\Gamma_{ii}(0)\Gamma_{jj}(0)}} = \frac{\text{Cov}(r_{it}, r_{j,t-\ell})}{\text{std}(r_{it})\text{std}(r_{jt})}, \tag{8.4}$$

which is the correlation coefficient between $r_{it}$ and $r_{j,t-\ell}$. When $\ell > 0$, this correlation coefficient measures the linear dependence of $r_{it}$ on $r_{j,t-\ell}$, which occurred prior to time $t$. Consequently, if $\rho_{ij}(\ell) \neq 0$ and $\ell > 0$, we say that the series $r_{jt}$ *leads* the series $r_{it}$ at lag $\ell$. Similarly, $\rho_{ji}(\ell)$ measures the linear dependence of $r_{jt}$ and $r_{i,t-\ell}$, and we say that the series $r_{it}$ *leads* the series $r_{jt}$ at lag $\ell$ if $\rho_{ji}(\ell) \neq 0$ and $\ell > 0$. Equation (8.4) also shows that the diagonal element $\rho_{ii}(\ell)$ is simply the lag-$\ell$ autocorrelation coefficient of $r_{it}$.

Based on this discussion, we obtain some important properties of the cross-correlations when $\ell > 0$. First, in general, $\rho_{ij}(\ell) \neq \rho_{ji}(\ell)$ for $i \neq j$ because the two correlation coefficients measure different linear relationships between $\{r_{it}\}$ and $\{r_{jt}\}$. Therefore, $\boldsymbol{\Gamma}_\ell$ and $\boldsymbol{\rho}_\ell$ are in general not symmetric. Second, from $\text{Cov}(r_{it}, r_{j,t-\ell}) = \text{Cov}(r_{j,t-\ell}, r_{it})$ and by the weak stationarity assumption

$$\text{Cov}(r_{j,t-\ell}, r_{it}) = \text{Cov}(r_{j,t}, r_{i,t+\ell}) = \text{Cov}[r_{jt}, r_{i,t-(-\ell)}],$$

we have $\Gamma_{ij}(\ell) = \Gamma_{ji}(-\ell)$. Because $\Gamma_{ji}(-\ell)$ is the $(j, i)$th element of the matrix $\boldsymbol{\Gamma}_{-\ell}$ and the equality holds for $1 \leq i, j \leq k$, we have $\boldsymbol{\Gamma}_\ell = \boldsymbol{\Gamma}'_{-\ell}$ and $\boldsymbol{\rho}_\ell = \boldsymbol{\rho}'_{-\ell}$. Consequently, unlike the univariate case, $\boldsymbol{\rho}_\ell \neq \boldsymbol{\rho}_{-\ell}$ for a general vector time series when $\ell > 0$. Because $\boldsymbol{\rho}_\ell = \boldsymbol{\rho}'_{-\ell}$, it suffices in practice to consider the cross-correlation matrixes $\boldsymbol{\rho}_\ell$ for $\ell \geq 0$.

### 8.1.2 Linear Dependence

Considered jointly, the cross-correlation matrixes $\{\boldsymbol{\rho}_\ell \mid \ell = 0, 1, \ldots\}$ of a weakly stationary vector time series contain the following information:

1. The diagonal elements $\{\rho_{ii}(\ell) \mid \ell = 0, 1, \ldots\}$ are the autocorrelation function of $r_{it}$.
2. The off-diagonal element $\rho_{ij}(0)$ measures the concurrent linear relationship between $r_{it}$ and $r_{jt}$.
3. For $\ell > 0$, the off-diagonal element $\rho_{ij}(\ell)$ measures the linear dependence of $r_{it}$ on the past value $r_{j,t-\ell}$.

Therefore, if $\rho_{ij}(\ell) = 0$ for all $\ell > 0$, then $r_{it}$ does not depend linearly on any past value $r_{j,t-\ell}$ of the $r_{jt}$ series.

In general, the linear relationship between two time series $\{r_{it}\}$ and $\{r_{jt}\}$ can be summarized as follows:

1. $r_{it}$ and $r_{jt}$ have no linear relationship if $\rho_{ij}(\ell) = \rho_{ji}(\ell) = 0$ for all $\ell \geq 0$.
2. $r_{it}$ and $r_{jt}$ are concurrently correlated if $\rho_{ij}(0) \neq 0$.
3. $r_{it}$ and $r_{jt}$ have no lead-lag relationship if $\rho_{ij}(\ell) = 0$ and $\rho_{ji}(\ell) = 0$ for all $\ell > 0$. In this case, we say the two series are uncoupled.
4. There is a *unidirectional relationship* from $r_{it}$ to $r_{jt}$ if $\rho_{ij}(\ell) = 0$ for all $\ell > 0$, but $\rho_{ji}(v) \neq 0$ for some $v > 0$. In this case, $r_{it}$ does not depend on any past value of $r_{jt}$, but $r_{jt}$ depends on some past values of $r_{it}$.
5. There is a *feedback relationship* between $r_{it}$ and $r_{jt}$ if $\rho_{ij}(\ell) \neq 0$ for some $\ell > 0$ and $\rho_{ji}(v) \neq 0$ for some $v > 0$.

The conditions stated earlier are sufficient conditions. A more informative approach to study the relationship between time series is to build a multivariate model for the series because a properly specified model considers simultaneously the serial and cross correlations among the series.

### 8.1.3 Sample Cross-Correlation Matrixes

Given the data $\{\boldsymbol{r}_t \mid t = 1, \ldots, T\}$, the cross-covariance matrix $\boldsymbol{\Gamma}_\ell$ can be estimated by

$$\widehat{\boldsymbol{\Gamma}}_\ell = \frac{1}{T} \sum_{t=\ell+1}^{T} (\boldsymbol{r}_t - \bar{\boldsymbol{r}})(\boldsymbol{r}_{t-\ell} - \bar{\boldsymbol{r}})', \quad \ell \geq 0, \tag{8.5}$$

where $\bar{\boldsymbol{r}} = \sum_{t=1}^{T} \boldsymbol{r}_t / T$ is the vector of sample means. The cross-correlation matrix $\boldsymbol{\rho}_\ell$ is estimated by

$$\hat{\boldsymbol{\rho}}_\ell = \widehat{\boldsymbol{D}}^{-1} \widehat{\boldsymbol{\Gamma}}_\ell \widehat{\boldsymbol{D}}^{-1}, \quad \ell \geq 0, \tag{8.6}$$

where $\widehat{\boldsymbol{D}}$ is the $k \times k$ diagonal matrix of the sample standard deviations of the component series.

Similar to the univariate case, asymptotic properties of the sample cross-correlation matrix $\hat{\boldsymbol{\rho}}_\ell$ have been investigated under various assumptions; see, for instance, Fuller (1976, Chapter 6). The estimate is consistent, but is biased in a finite sample. For asset return series, the finite sample distribution of $\hat{\boldsymbol{\rho}}_\ell$ is rather complicated partly because of the presence of conditional heteroscedasticity and high kurtosis. If the finite-sample distribution of cross-correlations is needed, we recommend that proper bootstrap resampling methods be used to obtain an approximate estimate of the distribution. For many applications, a crude approximation of the variance of $\hat{\rho}_{ij}(\ell)$ is sufficient.

**Example 8.1.** Consider the monthly log returns of IBM stock and the S&P 500 index from January 1926 to December 1999 with 888 observations. The returns include dividend payments and are in percentages. Denote the returns of IBM stock and the S&P 500 index by $r_{1t}$ and $r_{2t}$, respectively. These two returns form a bivariate time series $\boldsymbol{r}_t = (r_{1t}, r_{2t})'$. Figure 8.1 shows the time plots of $\boldsymbol{r}_t$ using the same scale. Figure 8.2 shows some scatterplots of the two series. The plots show that the two return series are concurrently correlated. Indeed, the sample concurrent correlation

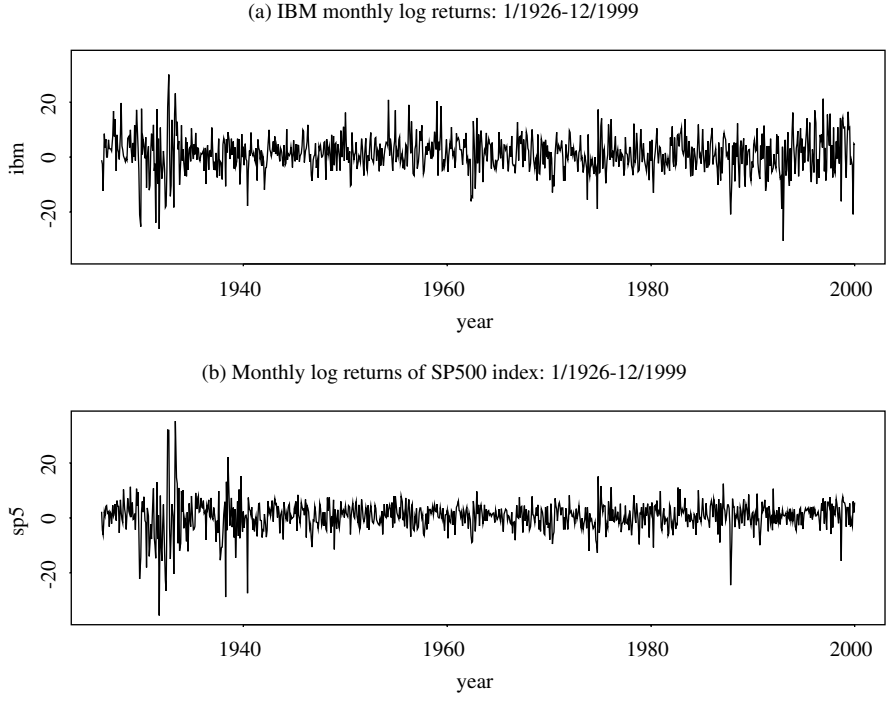


**Figure 8.1.** Time plot of monthly log returns in percentages for IBM stock and the S&P 500 index from January 1926 to December 1999.

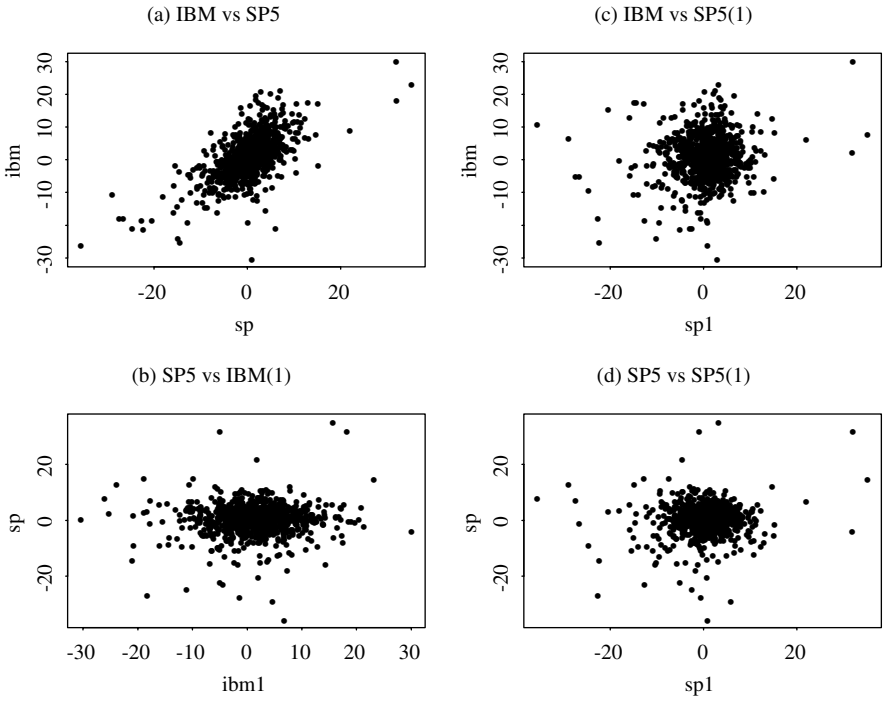


**Figure 8.2.** Some scatterplots for monthly log returns of IBM stock and the S&P 500 index: (a) Concurrent plot, (b) S&P 500 versus lag-1 IBM, (c) IBM versus lag-1 S&P 500, and (d) S&P 500 versus lag-1 S&P 500.

coefficient between the two returns is 0.64, which is statistically significant at the 5% level. However, the cross-correlations at lag 1 are weak if any.

Table 8.1 provides some summary statistics and cross-correlation matrixes of the two series. For a bivariate series, each CCM is a $2 \times 2$ matrix with four correlations. Empirical experience indicates that it is rather hard to absorb simultaneously many cross-correlation matrixes, especially when the dimension $k$ is greater than 3. To overcome this difficulty, we use the simplifying notation of Tiao and Box (1981) and define a simplified cross-correlation matrix consisting of three symbols "+," "−," and "." where

1. "+" means that the corresponding correlation coefficient is greater than or equal to $2/\sqrt{T}$,
2. "−" means that the corresponding correlation coefficient is less than or equal to $-2/\sqrt{T}$, and
3. "." means that the corresponding correlation coefficient is between $-2/\sqrt{T}$ and $2/\sqrt{T}$,

where $1/\sqrt{T}$ is the asymptotic 5% critical value of the sample correlation under the assumption that $\boldsymbol{r}_t$ is a white noise series.

**Table 8.1. Summary Statistics and Cross-Correlation Matrixes of Monthly Log Returns of IBM Stock and the S&P 500 Index. The Data Span is From January 1926 to December 1999.**

(a) Summary statistics

| Ticker | Mean | St. Error | Skewness | Exc.Kurt. | Minimum | Maximum |
|---|---|---|---|---|---|---|
| IBM | 1.240 | 6.729 | −0.237 | 1.917 | −30.37 | 30.10 |
| SP5 | 0.537 | 5.645 | −0.521 | 8.117 | −35.58 | 35.22 |

(b) Cross-correlation matrixes

| Lag 1 | | Lag 2 | | Lag 3 | | Lag 4 | | Lag 5 | |
|---|---|---|---|---|---|---|---|---|---|
| .08 | .10 | .02 | −.06 | −.02 | −.07 | −.02 | −.03 | .00 | .07 |
| .04 | .08 | .02 | −.02 | −.07 | −.11 | .04 | .02 | .00 | .08 |

(c) Simplified notation

$$\begin{bmatrix} + & + \\ . & + \end{bmatrix} \quad \begin{bmatrix} . & . \\ . & . \end{bmatrix} \quad \begin{bmatrix} . & - \\ - & - \end{bmatrix} \quad \begin{bmatrix} . & . \\ . & . \end{bmatrix} \quad \begin{bmatrix} . & . \\ . & + \end{bmatrix}$$

Table 8.1(c) shows the simplified CCM for the monthly log returns of IBM stock and the S&P 500 index. It is easily seen that significant cross-correlations at the approximate 5% level appear mainly at lags 1 and 3. An examination of the sample CCMs at these two lags indicates that (a) S&P 500 index returns have some marginal autocorrelations at lags 1 and 3, and (b) IBM stock returns depend weakly on the previous returns of the S&P 500 index. The latter observation is based on the significance of cross-correlations at the (1, 2)th element of lag-1 and lag-3 CCMs.

Figure 8.3 shows the sample autocorrelations and cross-correlations of the two series. Since ACF is symmetric with respect to lag 0, only those of positive lags are shown. Because lagged values of the S&P 500 index return are used to compute the cross-correlations, the plot associated with positive lags in Figure 8.3(c) shows the dependence of IBM stock return on the past S&P 500 index returns, and the plot associated with negative lags shows the linear dependence of the index return on the past IBM stock returns. The horizontal lines in the plots are the asymptotic two standard-error limits of the sample auto- and cross-correlation coefficients. From the plots, the dynamic relationship is weak between the two return series, but their contemporaneous correlation is statistically significant.

**Example 8.2.** Consider the simple returns of monthly indexes of U.S. government bonds with maturities in 30 years, 20 years, 10 years, 5 years, and 1 year. The data obtained from CRSP database have 696 observations starting from January 1942 to December 1999. Let $\boldsymbol{r}_t = (r_{1t}, \ldots, r_{5t})'$ be the return series with decreasing time to maturity. Figure 8.4 shows the time plots of $\boldsymbol{r}_t$ on the same scale. The variability of the 1-year bond returns is much smaller than that of returns with

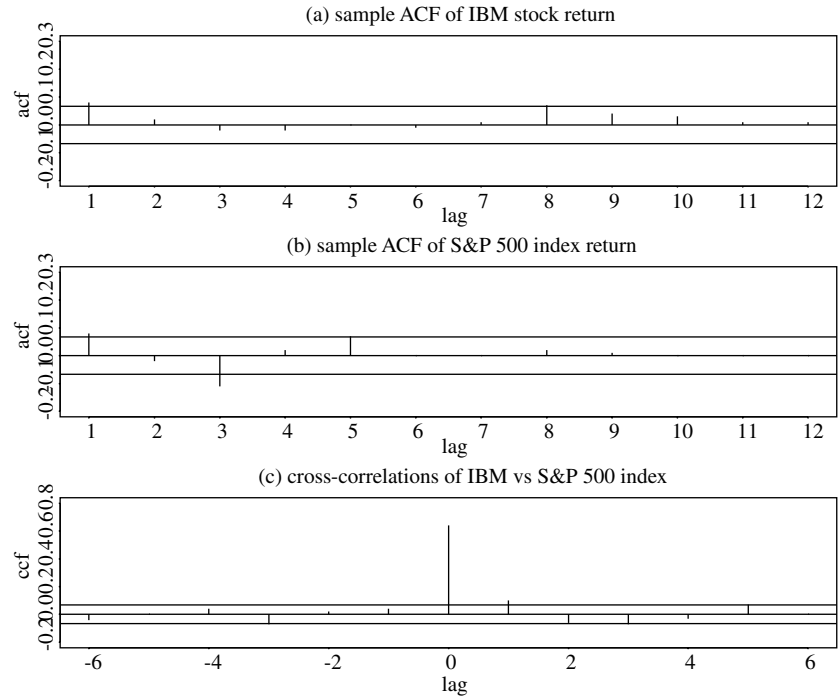


**Figure 8.3.** Sample auto- and cross-correlation functions of two monthly log returns: (a) sample ACF of IBM stock returns, (b) sample ACF of S&P 500 index returns, and (c) cross-correlations between IBM stock return and lagged S&P 500 index returns.

longer maturities. The sample means and standard deviations of the data are $\widehat{\boldsymbol{\mu}} = 10^{-2}(0.43, 0.45, 0.45, .0.46, 0.44)'$ and $\widehat{\boldsymbol{\sigma}} = 10^{-2}(2.53, 2.43, 1.97, 1.39, 0.53)'$. The concurrent correlation matrix of the series is

$$\hat{\boldsymbol{\rho}}_0 = \begin{bmatrix} 1.00 & 0.98 & 0.92 & 0.85 & 0.63 \\ 0.98 & 1.00 & 0.91 & 0.86 & 0.64 \\ 0.92 & 0.91 & 1.00 & 0.90 & 0.68 \\ 0.85 & 0.86 & 0.90 & 1.00 & 0.82 \\ 0.63 & 0.64 & 0.68 & 0.82 & 1.00 \end{bmatrix}.$$

It is not surprising that (a) the series have high concurrent correlations, and (b) the correlations between long-term bonds are higher than those between short-term bonds.

Table 8.2 gives the lag-1 and lag-2 cross-correlation matrixes of $\boldsymbol{r}_t$ and the corresponding simplified matrixes. Most of the significant cross-correlations are at lag 1, and the five return series appear to be intercorrelated. In addition, lag-1 and lag-2 sample ACFs of the 1-year bond returns are substantially higher than those of other series with longer maturities.

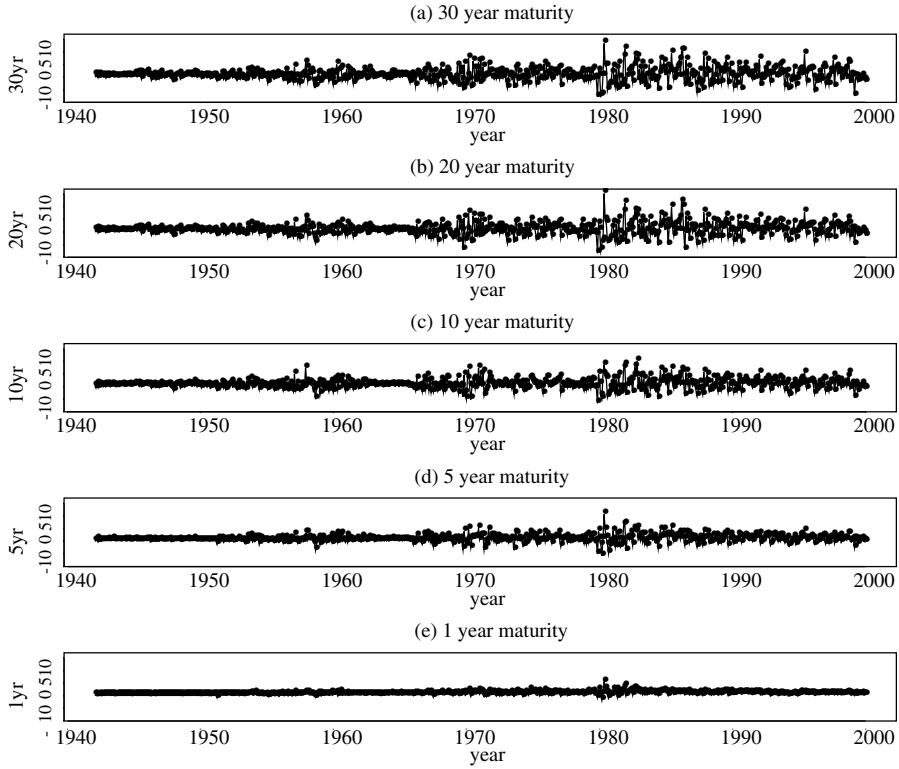


**Figure 8.4.** Time plots of monthly simple returns of five indexes of U.S. government bonds with maturities in 30 years, 20 years, 10 years, 5 years, and 1 year. The sample period is from January 1942 to December 1999.

**Table 8.2. Sample Cross-Correlation Matrixes of Monthly Simple Returns of Five Indexes of U.S. Government Bonds. The Data Span is From January 1942 to December 1999.**

| Lag 1 | | | | | Lag 2 | | | | |
|---|---|---|---|---|---|---|---|---|---|
| (a) Cross-correlations | | | | | | | | | |
| .10 | .08 | .11 | .12 | .16 | −.01 | .00 | .00 | −.03 | .03 |
| .10 | .08 | .12 | .14 | .17 | −.01 | .00 | .00 | −.04 | .02 |
| .09 | .08 | .09 | .13 | .18 | .01 | .01 | .01 | −.02 | .07 |
| .14 | .12 | .15 | .14 | .22 | −.02 | −.01 | .00 | −.04 | .07 |
| .17 | .15 | .21 | .22 | .40 | −.02 | .00 | .02 | .02 | .22 |

(b) Simplified cross-correlation matrixes

$$\begin{bmatrix} + & + & + & + & + \\ + & + & + & + & + \\ + & + & + & + & + \\ + & + & + & + & + \\ + & + & + & + & + \end{bmatrix} \qquad \begin{bmatrix} . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & + \end{bmatrix}$$

### 8.1.4 Multivariate Portmanteau Tests

The univariate Ljung–Box statistic $Q(m)$ has been generalized to the multivariate case by Hosking (1980, 1981) and Li and McLeod (1981). For a multivariate series, the null hypothesis of the test statistic is $H_0 : \boldsymbol{\rho}_1 = \cdots = \boldsymbol{\rho}_m = \mathbf{0}$, and the alternative hypothesis $H_a : \boldsymbol{\rho}_i \neq \mathbf{0}$ for some $i \in \{1, \ldots, m\}$. Thus, the statistic is used to test that there are no auto- and cross-correlations in the vector series $\boldsymbol{r}_t$. The test statistic assumes the form

$$Q_k(m) = T^2 \sum_{\ell=1}^{m} \frac{1}{T-\ell} tr(\widehat{\boldsymbol{\Gamma}}_\ell' \widehat{\boldsymbol{\Gamma}}_0^{-1} \widehat{\boldsymbol{\Gamma}}_\ell \widehat{\boldsymbol{\Gamma}}_0^{-1}), \tag{8.7}$$

where $T$ is the sample size, $k$ is the dimension of $\boldsymbol{r}_t$, and $tr(\boldsymbol{A})$ is the trace of the matrix $\boldsymbol{A}$, which is the sum of the diagonal elements of $\boldsymbol{A}$. Under the null hypothesis and some regularity conditions, $Q_k(m)$ follows asymptotically a chi-squared distribution with $k^2 m$ degrees of freedom.

***Remark:*** The $Q_k(m)$ statistics can be rewritten in terms of the sample cross-correlation matrixes $\hat{\boldsymbol{\rho}}_\ell$, but the expression involves Kronecker product $\otimes$ and vectorization of matrixes discussed in Appendix A of the chapter. Using these operators, we have

$$Q_k(m) = T^2 \sum_{\ell=1}^{m} \frac{1}{T-\ell} \boldsymbol{b}_\ell' (\hat{\boldsymbol{\rho}}_0^{-1} \otimes \hat{\boldsymbol{\rho}}_0^{-1}) \boldsymbol{b}_\ell,$$

where $\boldsymbol{b}_\ell = \text{vec}(\hat{\boldsymbol{\rho}}_\ell')$. More specifically, the test statistic proposed by Li and McLeod (1981) is

$$Q_k^*(m) = T \sum_{\ell=1}^{m} \boldsymbol{b}_\ell' (\hat{\boldsymbol{\rho}}_0^{-1} \otimes \hat{\boldsymbol{\rho}}_0^{-1}) \boldsymbol{b}_\ell + \frac{k^2 m(m+1)}{2T},$$

which is asymptotically equivalent to $Q_k(m)$.

Applying the $Q_k(m)$ statistics to the bivariate monthly log returns of IBM stock and the S&P 500 index of Example 8.1, we have $Q_2(1) = 9.81$, $Q_2(5) = 47.06$, and $Q_2(10) = 71.65$. Based on asymptotic chi-squared distributions with degrees of freedom 4, 20, and 40, the $p$ values of these $Q_2(m)$ statistics are all close to zero. Consequently, the Portmanteau tests confirm the existence of serial dependence in the bivariate return series. For the five-dimensional monthly simple returns of bond indexes in Example 8.2, we have $Q_5(5) = 1065.63$, which is highly significant compared with a chi-squared distribution with 125 degrees of freedom.

The $Q_k(m)$ statistic is a joint test for checking the first $m$ cross-correlation matrixes of $\boldsymbol{r}_t$. If it rejects the null hypothesis, then we must build a multivariate model for the series to study the lead-lag relationships between the component series. In what follows, we discuss some simple vector models useful for modeling the linear dynamic structure of a multivariate financial time series.

## 8.2 VECTOR AUTOREGRESSIVE MODELS

A simple vector model useful in modeling asset returns is the vector autoregressive (VAR) model. A multivariate time series $\boldsymbol{r}_t$ is a VAR process of order 1, or VAR(1) for short, if it follows the model

$$\boldsymbol{r}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}\boldsymbol{r}_{t-1} + \boldsymbol{a}_t, \tag{8.8}$$

where $\boldsymbol{\phi}_0$ is a $k$-dimensional vector, $\boldsymbol{\Phi}$ is a $k \times k$ matrix, and $\{\boldsymbol{a}_t\}$ is a sequence of serially uncorrelated random vectors with mean zero and covariance matrix $\boldsymbol{\Sigma}$. In application, the covariance matrix $\boldsymbol{\Sigma}$ is required to be positive definite; otherwise, the dimension of $\boldsymbol{r}_t$ can be reduced. In the literature, it is often assumed that $\boldsymbol{a}_t$ is multivariate normal.

Consider the bivariate case [i.e., $k = 2$, $\boldsymbol{r}_t = (r_{1t}, r_{2t})'$ and $\boldsymbol{a}_t = (a_{1t}, a_{2t})'$]. The VAR(1) model consists of the following two equations:

$$\begin{aligned} r_{1t} &= \phi_{10} + \Phi_{11} r_{1,t-1} + \Phi_{12} r_{2,t-1} + a_{1t} \\ r_{2t} &= \phi_{20} + \Phi_{21} r_{1,t-1} + \Phi_{22} r_{2,t-1} + a_{2t}, \end{aligned}$$

where $\Phi_{ij}$ is the $(i, j)$th element of $\boldsymbol{\Phi}$ and $\phi_{i0}$ is the $i$th element of $\boldsymbol{\phi}_0$. Based on the first equation, $\Phi_{12}$ denotes the linear dependence of $r_{1t}$ on $r_{2,t-1}$ in the presence of $r_{1,t-1}$. Therefore, $\Phi_{12}$ is the conditional effect of $r_{2,t-1}$ on $r_{1t}$ given $r_{1,t-1}$. If $\Phi_{12} = 0$, then $r_{1t}$ does not depend on $r_{2,t-1}$, and the model shows that $r_{1t}$ only depends on its own past. Similarly, if $\Phi_{21} = 0$, then the second equation shows that $r_{2t}$ does not depend on $r_{1,t-1}$ when $r_{2,t-1}$ is given.

Consider the two equations jointly. If $\Phi_{12} = 0$ and $\Phi_{21} \neq 0$, then there is a unidirectional relationship from $r_{1t}$ to $r_{2t}$. If $\Phi_{12} = \Phi_{21} = 0$, then $r_{1t}$ and $r_{2t}$ are uncoupled. If $\Phi_{12} \neq 0$ and $\Phi_{21} \neq 0$, then there is a feedback relationship between the two series.

In general, the coefficient matrix $\boldsymbol{\Phi}$ measures the dynamic dependence of $\boldsymbol{r}_t$. The concurrent relationship between $r_{1t}$ and $r_{2t}$ is shown by the off-diagonal element $\sigma_{12}$ of the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{a}_t$. If $\sigma_{12} = 0$, then there is no concurrent linear relationship between the two component series. In the econometric literature, the VAR(1) model in Eq. (8.8) is called a *reduced-form* model because it does not show explicitly the concurrent dependence between the component series. If necessary, an explicit expression involving the concurrent relationship can be deduced from the reduced-form model by a simple linear transformation. Because $\boldsymbol{\Sigma}$ is positive definite, there exists a lower triangular matrix $\boldsymbol{L}$ with unit diagonal elements and a diagonal matrix $\boldsymbol{G}$ such that $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{G}\boldsymbol{L}'$; see Appendix A on Cholesky Decomposition. Therefore, $\boldsymbol{L}^{-1}\boldsymbol{\Sigma}(\boldsymbol{L}')^{-1} = \boldsymbol{G}$.

Define $\boldsymbol{b}_t = (b_{1t}, \ldots, b_{kt})' = \boldsymbol{L}^{-1}\boldsymbol{a}_t$. Then

$$E(\boldsymbol{b}_t) = \boldsymbol{L}^{-1}E(\boldsymbol{a}_t) = \boldsymbol{0}, \quad \text{Cov}(\boldsymbol{b}_t) = \boldsymbol{L}^{-1}\boldsymbol{\Sigma}(\boldsymbol{L}^{-1})' = \boldsymbol{L}^{-1}\boldsymbol{\Sigma}(\boldsymbol{L}')^{-1} = \boldsymbol{G}.$$

Since $\boldsymbol{G}$ is a diagonal matrix, the components of $\boldsymbol{b}_t$ are uncorrelated. Multiplying $\boldsymbol{L}^{-1}$ from left to model (8.8), we obtain

$$\boldsymbol{L}^{-1}\boldsymbol{r}_t = \boldsymbol{L}^{-1}\phi_0 + \boldsymbol{L}^{-1}\boldsymbol{\Phi}\boldsymbol{r}_{t-1} + \boldsymbol{L}^{-1}\boldsymbol{a}_t = \phi_0^* + \boldsymbol{\Phi}^*\boldsymbol{r}_{t-1} + \boldsymbol{b}_t, \tag{8.9}$$

where $\phi_0^* = \boldsymbol{L}^{-1}\phi_0$ is a $k$-dimensional vector and $\boldsymbol{\Phi}^* = \boldsymbol{L}^{-1}\boldsymbol{\Phi}$ is a $k \times k$ matrix. Because of the special matrix structure, the $k$th row of $\boldsymbol{L}^{-1}$ is in the form $(w_{k1}, w_{k2}, \ldots, w_{k,k-1}, 1)$. Consequently, the $k$th equation of model (8.9) is

$$r_{kt} + \sum_{i=1}^{k-1} w_{ki} r_{it} = \phi_{k,0}^* + \sum_{i=1}^{k} \Phi_{ki}^* r_{i,t-1} + b_{kt}, \tag{8.10}$$

where $\phi_{k,0}^*$ is the $k$th element of $\phi_0^*$ and $\Phi_{ki}^*$ is the $(k, i)$th element of $\boldsymbol{\Phi}^*$. Because $b_{kt}$ is uncorrelated with $b_{it}$ for $1 \le i < k$, Eq. (8.10) shows explicitly the concurrent linear dependence of $r_{kt}$ on $r_{it}$, where $1 \le i \le k-1$. This equation is referred to as a structural equation for $r_{kt}$ in the econometric literature.

For any other component $r_{it}$ of $\boldsymbol{r}_t$, we can rearrange the VAR(1) model so that $r_{it}$ becomes the last component of $\boldsymbol{r}_t$. The prior transformation method can then be applied to obtain a structural equation for $r_{it}$. Therefore, the reduced-form model (8.8) is equivalent to the structural form used in the econometric literature. In time series analysis, the reduced-form model is commonly used for two reasons. The first reason is ease in estimation. The second and main reason is that the concurrent correlations cannot be used in forecasting.

**Example 8.3.** To illustrate the transformation from a reduced-form model to structural equations, consider the bivariate AR(1) model

$$\begin{bmatrix} r_{1t} \\ r_{2t} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 \\ -0.6 & 1.1 \end{bmatrix} \begin{bmatrix} r_{1,t-1} \\ r_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}.$$

For this particular covariance matrix $\boldsymbol{\Sigma}$, the lower triangular matrix

$$\boldsymbol{L}^{-1} = \begin{bmatrix} 1.0 & 0.0 \\ -0.5 & 1.0 \end{bmatrix}$$

provides a Cholesky decomposition (i.e., $\boldsymbol{L}^{-1}\boldsymbol{\Sigma}(\boldsymbol{L}')^{-1}$ is a diagonal matrix). Premultiplying $\boldsymbol{L}^{-1}$ to the previous bivariate AR(1) model, we obtain

$$\begin{bmatrix} 1.0 & 0.0 \\ -0.5 & 1.0 \end{bmatrix} \begin{bmatrix} r_{1t} \\ r_{2t} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.3 \end{bmatrix} + \begin{bmatrix} 0.2 & 0.3 \\ -0.7 & 0.95 \end{bmatrix} \begin{bmatrix} r_{1,t-1} \\ r_{2,t-1} \end{bmatrix} + \begin{bmatrix} b_{1t} \\ b_{2t} \end{bmatrix},$$

$$\boldsymbol{G} = \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix},$$

where $\boldsymbol{G} = \text{Cov}(\boldsymbol{b}_t)$. The second equation of this transformed model gives

$$r_{2t} = 0.3 + 0.5r_{1t} - 0.7r_{1,t-1} + 0.95r_{2,t-1} + b_{2t},$$

which shows explicitly the linear dependence of $r_{2t}$ on $r_{1t}$.

Rearranging the order of elements in $\boldsymbol{r}_t$, the bivariate AR(1) model becomes

$$\begin{bmatrix} r_{2t} \\ r_{1t} \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.2 \end{bmatrix} + \begin{bmatrix} -0.6 & 1.1 \\ 0.2 & 0.3 \end{bmatrix} \begin{bmatrix} r_{2,t-1} \\ r_{1,t-1} \end{bmatrix} + \begin{bmatrix} a_{2t} \\ a_{1t} \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}.$$

The lower triangular matrix needed in the Cholesky decomposition of $\boldsymbol{\Sigma}$ becomes

$$\boldsymbol{L}^{-1} = \begin{bmatrix} 1.0 & 0.0 \\ -1.0 & 1.0 \end{bmatrix}.$$

Premultiplying $\boldsymbol{L}^{-1}$ to the earlier rearranged VAR(1) model, we obtain

$$\begin{bmatrix} 1.0 & 0.0 \\ -1.0 & 1.0 \end{bmatrix} \begin{bmatrix} r_{2t} \\ r_{1t} \end{bmatrix} = \begin{bmatrix} 0.4 \\ -0.2 \end{bmatrix} + \begin{bmatrix} -0.6 & 1.1 \\ 0.8 & -0.8 \end{bmatrix} \begin{bmatrix} r_{2,t-1} \\ r_{1,t-1} \end{bmatrix} + \begin{bmatrix} c_{1t} \\ c_{2t} \end{bmatrix},$$

$$\boldsymbol{G} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

where $\boldsymbol{G} = \text{Cov}(\boldsymbol{c}_t)$. The second equation now gives

$$r_{1t} = -0.2 + 1.0r_{2t} + 0.8r_{2,t-1} - 0.8r_{1,t-1} + c_{2t}.$$

Again this equation shows explicitly the concurrent linear dependence of $r_{1t}$ on $r_{2t}$.

### 8.2.1 Stationarity Condition and Moments of a VAR(1) Model

Assume that the VAR(1) model in Eq. (8.8) is weakly stationary. Taking expectation of the model and using $E(\boldsymbol{a}_t) = \boldsymbol{0}$, we obtain

$$E(\boldsymbol{r}_t) = \boldsymbol{\phi}_0 + \boldsymbol{\Phi} E(\boldsymbol{r}_{t-1}).$$

Since $E(\boldsymbol{r}_t)$ is time-invariant, we have

$$\boldsymbol{\mu} \equiv E(\boldsymbol{r}_t) = (\boldsymbol{I} - \boldsymbol{\Phi})^{-1} \boldsymbol{\phi}_0$$

provided that the matrix $\boldsymbol{I} - \boldsymbol{\Phi}$ is nonsingular, where $\boldsymbol{I}$ is the $k \times k$ identity matrix.

Using $\boldsymbol{\phi}_0 = (\boldsymbol{I} - \boldsymbol{\Phi})\boldsymbol{\mu}$, the VAR(1) model in Eq. (8.8) can be written as

$$(\boldsymbol{r}_t - \boldsymbol{\mu}) = \boldsymbol{\Phi}(\boldsymbol{r}_{t-1} - \boldsymbol{\mu}) + \boldsymbol{a}_t.$$

Let $\tilde{\boldsymbol{r}}_t = \boldsymbol{r}_t - \boldsymbol{\mu}$ be the mean-corrected time series. Then the VAR(1) model becomes

$$\tilde{\boldsymbol{r}}_t = \boldsymbol{\Phi} \tilde{\boldsymbol{r}}_{t-1} + \boldsymbol{a}_t. \tag{8.11}$$

This model can be used to derive properties of a VAR(1) model. By repeated substitutions, we can rewrite Eq. (8.11) as

$$\tilde{\boldsymbol{r}}_t = \boldsymbol{a}_t + \boldsymbol{\Phi}\boldsymbol{a}_{t-1} + \boldsymbol{\Phi}^2\boldsymbol{a}_{t-2} + \boldsymbol{\Phi}^3\boldsymbol{a}_{t-3} + \cdots.$$

This expression shows several characteristics of a VAR(1) process. First, since $\boldsymbol{a}_t$ is serially uncorrelated, it follows that $\text{Cov}(\boldsymbol{a}_t, \boldsymbol{r}_{t-1}) = \boldsymbol{0}$. In fact, $\boldsymbol{a}_t$ is not correlated with $\boldsymbol{r}_{t-\ell}$ for all $\ell > 0$. For this reason, $\boldsymbol{a}_t$ is referred to as the *shock* or *innovation* of the series at time $t$. It turns out that, similar to the univariate case, $\boldsymbol{a}_t$ is uncorrelated with the past value $\boldsymbol{r}_{t-j}$ $(j > 0)$ for all time series models. Second, postmultiplying the expression by $\boldsymbol{a}_t'$, taking expectation, and using the fact of no serial correlations in the $\boldsymbol{a}_t$ process, we obtain $\text{Cov}(\boldsymbol{r}_t, \boldsymbol{a}_t) = \boldsymbol{\Sigma}$. Third, for a VAR(1) model, $\boldsymbol{r}_t$ depends on the past innovation $\boldsymbol{a}_{t-j}$ with coefficient matrix $\boldsymbol{\Phi}^j$. For such dependence to be meaningful, $\boldsymbol{\Phi}^j$ must converge to zero as $j \to \infty$. This means that the $k$ eigenvalues of $\boldsymbol{\Phi}$ must be less than 1 in modulus; otherwise, $\boldsymbol{\Phi}^j$ will either explode or converge to a nonzero matrix as $j \to \infty$. As a matter of fact, the requirement that all eigenvalues of $\boldsymbol{\Phi}$ are less than 1 in modulus is the necessary and sufficient condition for weak stationarity of $\boldsymbol{r}_t$ provided that the covariance matrix of $\boldsymbol{a}_t$ exists. Notice that this stationarity condition reduces to that of the univariate AR(1) case in which the condition is $|\phi| < 1$. Fourth, using the expression, we have

$$\text{Cov}(\boldsymbol{r}_t) = \boldsymbol{\Gamma}_0 = \boldsymbol{\Sigma} + \boldsymbol{\Phi}\boldsymbol{\Sigma}\boldsymbol{\Phi}' + \boldsymbol{\Phi}^2\boldsymbol{\Sigma}(\boldsymbol{\Phi}^2)' + \cdots = \sum_{i=0}^{\infty} \boldsymbol{\Phi}^i \boldsymbol{\Sigma}(\boldsymbol{\Phi}^i)',$$

where it is understood that $\boldsymbol{\Phi}^0 = \boldsymbol{I}$, the $k \times k$ identity matrix.

Postmultiplying $\tilde{\boldsymbol{r}}_{t-\ell}'$ to Eq. (8.11), taking expectation, and using the result $\text{Cov}(\boldsymbol{a}_t, \boldsymbol{r}_{t-j}) = E(\boldsymbol{a}_t\tilde{\boldsymbol{r}}_{t-j}') = \boldsymbol{0}$ for $j > 0$, we obtain

$$E(\tilde{\boldsymbol{r}}_t\tilde{\boldsymbol{r}}_{t-\ell}') = \boldsymbol{\Phi}E(\tilde{\boldsymbol{r}}_{t-1}, \tilde{\boldsymbol{r}}_{t-\ell})', \quad \ell > 0.$$

Therefore,

$$\boldsymbol{\Gamma}_\ell = \boldsymbol{\Phi}\boldsymbol{\Gamma}_{\ell-1}, \quad \ell > 0, \tag{8.12}$$

where $\boldsymbol{\Gamma}_j$ is the lag-$j$ cross-covariance matrix of $\boldsymbol{r}_t$. Again this result is a generalization of that of a univariate AR(1) process. By repeated substitutions, Eq. (8.12) shows that

$$\boldsymbol{\Gamma}_\ell = \boldsymbol{\Phi}^\ell\boldsymbol{\Gamma}_0, \quad \text{for} \quad \ell > 0.$$

### 8.2.2 Vector AR($p$) Models

The generalization of VAR(1) to VAR($p$) models is straightforward. The time series $\boldsymbol{r}_t$ follows a VAR($p$) model if it satisfies

$$\boldsymbol{r}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Phi}_1\boldsymbol{r}_{t-1} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{r}_{t-p} + \boldsymbol{a}_t, \quad p > 0, \tag{8.13}$$

where $\phi_0$ and $\boldsymbol{a}_t$ are defined as before, and $\boldsymbol{\Phi}_j$ are $k \times k$ matrixes. Using the back-shift operator $B$, the VAR($p$) model can be written as

$$(\boldsymbol{I} - \boldsymbol{\Phi}_1 B - \cdots - \boldsymbol{\Phi}_p B^p)\boldsymbol{r}_t = \phi_0 + \boldsymbol{a}_t,$$

where $\boldsymbol{I}$ is the $k \times k$ identity matrix. This representation can be written in a compact form as

$$\boldsymbol{\Phi}(B)\boldsymbol{r}_t = \phi_0 + \boldsymbol{a}_t,$$

where $\boldsymbol{\Phi}(B) = \boldsymbol{I} - \boldsymbol{\Phi}_1 B - \cdots - \boldsymbol{\Phi}_p B^p$ is a matrix polynomial. If $\boldsymbol{r}_t$ is weakly stationary, then we have

$$\boldsymbol{\mu} = E(\boldsymbol{r}_t) = (\boldsymbol{I} - \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_p)^{-1}\phi_0 = [\boldsymbol{\Phi}(1)]^{-1}\phi_0$$

provided that the inverse exists. Let $\tilde{\boldsymbol{r}}_t = \boldsymbol{r}_t - \boldsymbol{\mu}$. The VAR($p$) model becomes

$$\tilde{\boldsymbol{r}}_t = \boldsymbol{\Phi}_1\tilde{\boldsymbol{r}}_{t-1} + \cdots + \boldsymbol{\Phi}_p\tilde{\boldsymbol{r}}_{t-p} + \boldsymbol{a}_t. \tag{8.14}$$

Using this equation and the same techniques as those for VAR(1) models, we obtain that

- $\text{Cov}(\boldsymbol{r}_t, \boldsymbol{a}_t) = \boldsymbol{\Sigma}$, the covariance matrix of $\boldsymbol{a}_t$;
- $\text{Cov}(\boldsymbol{r}_{t-\ell}, \boldsymbol{a}_t) = \boldsymbol{0}$ for $\ell > 0$;
- $\boldsymbol{\Gamma}_\ell = \boldsymbol{\Phi}_1\boldsymbol{\Gamma}_{\ell-1} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{\Gamma}_{\ell-p}$ for $\ell > 0$.

The last property is called the moment equations of a VAR($p$) model. It is a multivariate version of the Yule–Walker equation of a univariate AR($p$) model.

A simple approach to understanding properties of the VAR($p$) model in Eq. (8.13) is to make use of the results of the VAR(1) model in Eq. (8.8). This can be achieved by transforming the VAR($p$) model of $\boldsymbol{r}_t$ into a $kp$-dimensional VAR(1) model. Specifically, let $\boldsymbol{x}_t = (\tilde{\boldsymbol{r}}'_{t-p+1}, \tilde{\boldsymbol{r}}'_{t-p+2}, \ldots, \tilde{\boldsymbol{r}}'_t)'$ and $\boldsymbol{b}_t = (0, \ldots, 0, \boldsymbol{a}'_t)'$ be two $kp$-dimensional processes. The mean of $\boldsymbol{b}_t$ is zero and the covariance matrix of $\boldsymbol{b}_t$ is a $kp \times kp$ matrix with zero everywhere except for the lower right corner, which is $\boldsymbol{\Sigma}$. The VAR($p$) model for $\boldsymbol{r}_t$ can then be written in the form

$$\boldsymbol{x}_t = \boldsymbol{\Phi}^*\boldsymbol{x}_{t-1} + \boldsymbol{b}_t, \tag{8.15}$$

where $\boldsymbol{\Phi}^*$ is a $kp \times kp$ matrix given by

$$\boldsymbol{\Phi}^* = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{I} \\ \boldsymbol{\Phi}_p & \boldsymbol{\Phi}_{p-1} & \boldsymbol{\Phi}_{p-2} & \boldsymbol{\Phi}_{p-3} & \cdots & \boldsymbol{\Phi}_1 \end{bmatrix},$$

where $\mathbf{0}$ and $\boldsymbol{I}$ are the $k \times k$ zero and identity matrix, respectively. In the literature, $\boldsymbol{\Phi}^*$ is called the *companion* matrix of the matrix polynomial $\boldsymbol{\Phi}(B)$.

Equation (8.15) is a VAR(1) model for $\boldsymbol{x}_t$, which contains $\boldsymbol{r}_t$ as its last $k$ components. The results of a VAR(1) model shown in the previous subsection can now be used to derive properties of the VAR($p$) model via Eq. (8.15). For example, from the definition, $\boldsymbol{x}_t$ is weakly stationary if and only if $\boldsymbol{r}_t$ is weakly stationary. Therefore, the necessary and sufficient condition of weak stationarity for the VAR($p$) model in Eq. (8.13) is that all eigenvalues of $\boldsymbol{\Phi}^*$ in Eq. (8.15) are less than 1 in modulus.

Of particular relevance to financial time series analysis is the structure of the coefficient matrixes $\boldsymbol{\Phi}_\ell$ of a VAR($p$) model. For instance, if the $(i, j)$th element $\Phi_{ij}(\ell)$ of $\boldsymbol{\Phi}_\ell$ is zero for all $\ell$, then $r_{it}$ does not depend on the past values of $r_{jt}$. The structure of the coefficient matrixes $\boldsymbol{\Phi}_\ell$ thus provides information on the lead-lag relationship between the components of $\boldsymbol{r}_t$.

### 8.2.3 Building a VAR($p$) Model

We continue to use the iterative procedure of order specification, estimation, and model checking to build a vector AR model for a given time series. The concept of partial autocorrelation function of a univariate series can be generalized to specify the order $p$ of a vector series. Consider the following consecutive VAR models:

$$\begin{aligned}
\boldsymbol{r}_t &= \boldsymbol{\phi}_0 + \boldsymbol{\Phi}_1 \boldsymbol{r}_{t-1} + \boldsymbol{a}_t \\
\boldsymbol{r}_t &= \boldsymbol{\phi}_0 + \boldsymbol{\Phi}_1 \boldsymbol{r}_{t-1} + \boldsymbol{\Phi}_2 \boldsymbol{r}_{t-2} + \boldsymbol{a}_t \\
&\vdots = \vdots \\
\boldsymbol{r}_t &= \boldsymbol{\phi}_0 + \boldsymbol{\Phi}_1 \boldsymbol{r}_{t-1} + \cdots + \boldsymbol{\Phi}_i \boldsymbol{r}_{t-i} + \boldsymbol{a}_t \\
&\vdots = \vdots
\end{aligned} \tag{8.16}$$

Parameters of these models can be estimated by the ordinary least squares (OLS) method. This is called the multivariate linear regression estimation in multivariate statistical analysis; see Johnson and Wichern (1998).

For the $i$th equation in Eq. (8.16), let $\widehat{\boldsymbol{\Phi}}_j^{(i)}$ be the OLS estimate of $\boldsymbol{\Phi}_j$ and $\widehat{\boldsymbol{\phi}}_0^{(i)}$ be the estimate of $\boldsymbol{\phi}_0$, where the superscript $(i)$ is used to denote that the estimates are for a VAR($i$) model. Then the residual is

$$\widehat{\boldsymbol{a}}_t^{(i)} = \boldsymbol{r}_t - \widehat{\boldsymbol{\phi}}_0^{(i)} - \widehat{\boldsymbol{\Phi}}_1^{(i)} \boldsymbol{r}_{t-1} - \cdots - \widehat{\boldsymbol{\Phi}}_i^{(i)} \boldsymbol{r}_{t-i}.$$

For $i = 0$, the residual is defined as $\widehat{\boldsymbol{r}}_t^{(0)} = \boldsymbol{r}_t - \bar{\boldsymbol{r}}$, where $\bar{\boldsymbol{r}}$ is the sample mean of $\boldsymbol{r}_t$. The residual covariance matrix is defined as

$$\widehat{\boldsymbol{\Sigma}}_i = \frac{1}{T - 2i - 1} \sum_{t=i+1}^{T} \widehat{\boldsymbol{a}}_t^{(i)} (\widehat{\boldsymbol{a}}_t^{(i)})', \quad i \geq 0. \tag{8.17}$$

To specify the order $p$, one can test the hypothesis $H_0 : \mathbf{\Phi}_\ell = \mathbf{0}$ versus the alternative hypothesis $H_a : \mathbf{\Phi}_\ell \neq \mathbf{0}$ sequentially for $\ell = 1, 2, \ldots$. For example, using the first equation in Eq. (8.16), we can test the hypothesis $H_0 : \mathbf{\Phi}_1 = \mathbf{0}$ versus the alternative hypothesis $H_a : \mathbf{\Phi}_1 \neq \mathbf{0}$. The test statistic is

$$M(1) = -\left(T - k - 2\frac{1}{2}\right) \ln\left(\frac{|\,\widehat{\mathbf{\Sigma}}_1\,|}{|\,\widehat{\mathbf{\Sigma}}_0\,|}\right),$$

where $\widehat{\mathbf{\Sigma}}_i$ is defined in Eq. (8.17) and $|\boldsymbol{A}|$ denotes the determinant of the matrix $\boldsymbol{A}$. Under some regularity conditions, the test statistic $M(1)$ is asymptotically a chi-squared distribution with $k^2$ degrees of freedom; see Tiao and Box (1981).

In general, we use the $i$th and $(i-1)$th equations in Eq. (8.16) to test $H_0 : \mathbf{\Phi}_i = \mathbf{0}$ versus $H_a : \mathbf{\Phi}_i \neq \mathbf{0}$—that is, testing a VAR($i$) model versus a VAR($i-1$) model. The test statistic is

$$M(i) = -\left(T - k - i - \frac{3}{2}\right) \ln\left(\frac{|\,\widehat{\mathbf{\Sigma}}_i\,|}{|\,\widehat{\mathbf{\Sigma}}_{i-1}\,|}\right). \tag{8.18}$$

Asymptotically, $M(i)$ is distributed as a chi-squared distribution with $k^2$ degrees of freedom.

Alternatively, one can use the Akaike information criterion (AIC) or its variants to select the order $p$. Assume that $\boldsymbol{a}_t$ is multivariate normal and consider the $i$th equation in Eq. (8.16). One can estimate the model by the maximum likelihood (ML) method. For AR models, the OLS estimates $\widehat{\boldsymbol{\phi}}_0$ and $\widehat{\mathbf{\Phi}}_j$ are equivalent to the (conditional) ML estimates. However, there are differences between the estimates of $\mathbf{\Sigma}$. The ML estimate of $\mathbf{\Sigma}$ is

$$\tilde{\mathbf{\Sigma}}_i = \frac{1}{T} \sum_{t=i+1}^{T} \widehat{\boldsymbol{a}}_t^{(i)} [\widehat{\boldsymbol{a}}_t^{(i)}]'. \tag{8.19}$$

The AIC of a VAR($i$) model under the normality assumption is defined as

$$AIC(i) = \ln(|\,\tilde{\mathbf{\Sigma}}_i\,|) + \frac{2k^2 i}{T}.$$

For a given vector time series, one selects the AR order $p$ such that $AIC(p) = \min_{1 \le i \le p_0} AIC(i)$, where $p_0$ is a prespecified positive integer.

**Example 8.4.** Assuming that the bivariate series of monthly log returns of IBM stock and the S&P 500 index discussed in Example 8.1 follows a VAR model, we apply the $M(i)$ statistics and AIC to the data. Table 8.3 shows the results of these statistics. Both statistics indicate that a VAR(3) model might be adequate for the data. The $M(i)$ statistics are marginally significant at lags 1, 3, and 5 at the 5% level. The minimum of AIC occurs at order 3. For this particular instance, the $M(i)$

**Table 8.3. Order-Specification Statistics for the Monthly Log Returns of IBM Stock and the S&P 500 Index from January 1926 to December 1999. The 5% and 1% Critical Values of a chi-Squared Distribution with 4 Degrees of Freedom are 9.5 and 13.3.**

| Order | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $M(i)$ | 9.81 | 8.93 | 12.57 | 6.08 | 9.56 | 2.80 |
| AIC | 6.757 | 6.756 | 6.750 | 6.753 | 6.751 | 6.756 |

statistics are nonsignificant at the 1% level, confirming the previous observation that the dynamic linear dependence between the two return series is weak.

***Estimation and Model Checking***

For a specified VAR model, one can estimate the parameters using either the ordinary least squares method or the maximum likelihood method. The two methods are asymptotically equivalent. Under some regularity conditions, the estimates are asymptotically normal; see Reinsel (1993). A fitted model should then be checked carefully for any possible inadequacy. The $Q_k(m)$ statistic can be applied to the residual series to check the assumption that there are no serial or cross-correlations in the residuals. For a fitted VAR($p$) model, the $Q_k(m)$ statistic of the residuals is asymptotically a chi-squared distribution with $k^2m - g$ degrees of freedom, where $g$ is the number of estimated parameters in the AR coefficient matrixes.

**Example 8.4.** (continued) Table 8.4(a) shows the estimation results of a VAR(3) model for the bivariate series of monthly log returns of IBM stock and

**Table 8.4. Estimation Results of a VAR(3) Model for the Monthly Log Returns, in Percentages, of IBM Stock and the S&P 500 Index from January 1926 to December 1999.**

| Param. | $\phi_0$ | $\Phi_1$ | | $\Phi_3$ | | $\Sigma$ | |
|---|---|---|---|---|---|---|---|
| | | | (a) Full model | | | | |
| Estimate | 1.20 | 0.011 | 0.108 | 0.039 | −0.112 | 44.44 | 23.51 |
| | 0.58 | −0.013 | 0.084 | −0.007 | −0.105 | 23.51 | 31.29 |
| St. Error | 0.23 | 0.043 | 0.051 | 0.044 | 0.052 | | |
| | 0.19 | 0.036 | 0.043 | 0.037 | 0.044 | | |
| | | | (b) Simplified model | | | | |
| Estimate | 1.24 | 0 | 0.117 | 0 | −0.083 | 44.48 | 23.51 |
| | 0.57 | 0 | 0.073 | 0 | −0.109 | 23.51 | 31.29 |
| St. Error | 0.23 | – | 0.040 | – | 0.040 | | |
| | 0.19 | – | 0.033 | – | 0.033 | | |

the S&P 500 index. The specified model is in the form

$$\boldsymbol{r}_t = \phi_0 + \boldsymbol{\Phi}_1 \boldsymbol{r}_{t-1} + \boldsymbol{\Phi}_3 \boldsymbol{r}_{t-3} + \boldsymbol{a}_t, \tag{8.20}$$

where the first component of $\boldsymbol{r}_t$ denotes IBM stock returns. For this particular instance, we only use AR coefficient matrixes at lags 1 and 3 because of the weak serial dependence of the data. In general, when the $M(i)$ statistics and the AIC criterion specify a VAR(3) model, all three AR lags should be used. Table 8.4(b) shows the estimation results after some statistically insignificant parameters are set to zero. The $Q_k(m)$ statistics of the residual series for the fitted model in Table 8.4(b) give $Q_2(4) = 18.17$ and $Q_2(8) = 41.26$. Since the fitted VAR(3) model has four parameters in the AR coefficient matrixes, these two $Q_k(m)$ statistics are distributed asymptotically as a chi-squared distribution with degrees of freedom 12 and 28, respectively. The $p$ values of the test statistics are 0.111 and 0.051, and hence the fitted model is adequate at the 5% significance level. As shown by the univariate analysis, the return series are likely to have conditional heteroscedasticity. We discuss multivariate volatility in Chapter 9.

From the fitted model in Table 8.4(b), we make the following observations. (a) The concurrent correlation coefficient between the two innovational series is $23.51/\sqrt{44.48 \times 31.29} = 0.63$, which, as expected, is close to the sample correlation coefficient between $r_{1t}$ and $r_{2t}$. (b) The two log return series have positive and significant means, implying that the log prices of the two series had an upward trend over the data span. (c) The model shows that

$$\begin{aligned} \text{IBM}_t &= 1.24 + 0.117\text{SP5}_{t-1} - 0.083\text{SP5}_{t-3} + a_{1t} \\ \text{SP5}_t &= 0.57 + 0.073\text{SP5}_{t-1} - 0.109\text{SP5}_{t-3} + a_{2t}. \end{aligned}$$

Consequently, at the 5% significant level, there is a unidirectional dynamic relationship from the monthly S&P 500 index return to the IBM return. If the S&P 500 index represents the U.S. stock market, then IBM return is affected by the past movements of the market. However, past movements of IBM stock returns do not significantly affect the U.S. market, even though the two returns have substantial concurrent correlation. Finally, the fitted model can be written as

$$\begin{bmatrix} \text{IBM}_t \\ \text{SP5}_t \end{bmatrix} = \begin{bmatrix} 1.24 \\ 0.57 \end{bmatrix} + \begin{bmatrix} 0.117 \\ 0.073 \end{bmatrix} \text{SP5}_{t-1} - \begin{bmatrix} 0.083 \\ 0.109 \end{bmatrix} \text{SP5}_{t-3} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix},$$

indicating that $\text{SP5}_t$ is the driving factor of the bivariate series.

***Forecasting***

Treating a properly built model as the true model, one can apply the same techniques as those in the univariate analysis to produce forecasts and standard deviations of the associated forecast errors. For a VAR($p$) model, the 1-step ahead forecast at the time origin $h$ is $\boldsymbol{r}_h(1) = \phi_0 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{r}_{h+1-i}$, and the associated forecast error

is $\boldsymbol{e}_h(1) = \boldsymbol{a}_{h+1}$. The covariance matrix of the forecast error is $\boldsymbol{\Sigma}$. If $\boldsymbol{r}_t$ is weakly stationary, then the $\ell$-step ahead forecast $\boldsymbol{r}_h(\ell)$ converges to its mean vector $\boldsymbol{\mu}$ as the forecast horizon $\ell$ increases.

In summary, building a VAR model involves three steps: (a) use the test statistic $M(i)$ or the Akaike information criterion to identify the order, (b) estimate the specified model by using the least squares method (in some cases, one can reestimate the model by removing statistically insignificant parameters), and (c) use the $Q_k(m)$ statistic of the residuals to check the adequacy of a fitted model. Other characteristics of the residual series, such as conditional heteroscedasticity and outliers, can also be checked. If the fitted model is adequate, then it can be used to obtain forecasts.

## 8.3 VECTOR MOVING-AVERAGE MODELS

A vector moving-average model of order $q$, or VMA($q$), is in the form

$$\boldsymbol{r}_t = \boldsymbol{\theta}_0 + \boldsymbol{a}_t - \boldsymbol{\Theta}_1 \boldsymbol{a}_{t-1} - \cdots - \boldsymbol{\Theta}_q \boldsymbol{a}_{t-q} \quad \text{or} \quad \boldsymbol{r}_t = \boldsymbol{\theta}_0 + \boldsymbol{\Theta}(B)\boldsymbol{a}_t, \tag{8.21}$$

where $\boldsymbol{\theta}_0$ is a $k$-dimensional vector, $\boldsymbol{\Theta}_i$ are $k \times k$ matrixes, and $\boldsymbol{\Theta}(B) = \boldsymbol{I} - \boldsymbol{\Theta}_1 B - \cdots - \boldsymbol{\Theta}_q B^q$ is the MA matrix polynomial in the back-shift operator $B$. Similar to the univariate case, VMA($q$) processes are weakly stationary provided that the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{a}_t$ exists. Taking expectation of Eq. (8.21), we obtain that $\boldsymbol{\mu} = E(\boldsymbol{r}_t) = \boldsymbol{\theta}_0$. Thus, the constant vector $\boldsymbol{\theta}_0$ is the mean vector of $\boldsymbol{r}_t$ for a VMA model.

Let $\tilde{\boldsymbol{r}}_t = \boldsymbol{r}_t - \boldsymbol{\theta}_0$ be the mean-corrected VAR($q$) process. Then using Eq. (8.21) and the fact that $\{\boldsymbol{a}_t\}$ has no serial correlations, we have

1. $\text{Cov}(\boldsymbol{r}_t, \boldsymbol{a}_t) = \boldsymbol{\Sigma}$,
2. $\boldsymbol{\Gamma}_0 = \boldsymbol{\Sigma} + \boldsymbol{\Theta}_1 \boldsymbol{\Sigma} \boldsymbol{\Theta}_1' + \cdots + \boldsymbol{\Theta}_q \boldsymbol{\Sigma} \boldsymbol{\Theta}_q'$,
3. $\boldsymbol{\Gamma}_\ell = \boldsymbol{0}$ if $\ell > q$, and
4. $\boldsymbol{\Gamma}_\ell = \sum_{j=\ell}^{q} \boldsymbol{\Theta}_j \boldsymbol{\Sigma} \boldsymbol{\Theta}_{j-\ell}'$ if $1 \le \ell \le q$, where $\boldsymbol{\Theta}_0 = -\boldsymbol{I}$.

Since $\boldsymbol{\Gamma}_\ell = \boldsymbol{0}$ for $\ell > q$, the cross-correlation matrixes (CCM) of a VMA($q$) process $\boldsymbol{r}_t$ satisfy

$$\boldsymbol{\rho}_\ell = \boldsymbol{0}, \quad \ell > q. \tag{8.22}$$

Therefore, similar to the univariate case, the sample CCMs can be used to identify the order of a VMA process.

To better understand the VMA processes, let us consider the bivariate MA(1) model

$$\boldsymbol{r}_t = \boldsymbol{\theta}_0 + \boldsymbol{a}_t - \boldsymbol{\Theta} \boldsymbol{a}_{t-1} = \boldsymbol{\mu} + \boldsymbol{a}_t - \boldsymbol{\Theta} \boldsymbol{a}_{t-1}, \tag{8.23}$$

where, for simplicity, the subscript of $\boldsymbol{\Theta}_1$ is removed. This model can be written explicitly as

$$\begin{bmatrix} r_{1t} \\ r_{2t} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} - \begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix} \begin{bmatrix} a_{1,t-1} \\ a_{2,t-1} \end{bmatrix}. \tag{8.24}$$

It says that the current return series $\boldsymbol{r}_t$ only depends on the current and past shocks. Therefore, the model is a finite-memory model.

Consider the equation for $r_{1t}$ in Eq. (8.24). The parameter $\Theta_{12}$ denotes the linear dependence of $r_{1t}$ on $a_{2,t-1}$ in the presence of $a_{1,t-1}$. If $\Theta_{12} = 0$, then $r_{1t}$ does not depend on the lagged values of $a_{2t}$ and, hence, the lagged values of $r_{2t}$. Similarly, if $\Theta_{21} = 0$, then $r_{2t}$ does not depend on the past values of $r_{1t}$. The off-diagonal elements of $\boldsymbol{\Theta}$ thus show the dynamic dependence between the component series. For this simple VMA(1) model, we can classify the relationships between $r_{1t}$ and $r_{2t}$ as follows:

1. They are uncoupled series if $\Theta_{12} = \Theta_{21} = 0$.
2. There is a unidirectional dynamic relationship from $r_{1t}$ to $r_{2t}$ if $\Theta_{12} = 0$, but $\Theta_{21} \neq 0$. The opposite unidirectional relationship holds if $\Theta_{21} = 0$, but $\Theta_{12} \neq 0$.
3. There is a feedback relationship between $r_{1t}$ and $r_{2t}$ if $\Theta_{12} \neq 0$ and $\Theta_{21} \neq 0$.

Finally, the concurrent correlation between $r_{it}$ is the same as that between $a_{it}$. The previous classification can be generalized to a VMA($q$) model.

***Estimation***

Unlike the VAR models, estimation of VMA models is much more involved; see Hillmer and Tiao (1979), Lütkepohl (1991), and the references therein. For the likelihood approach, there are two methods available. The first method is the conditional likelihood method that assumes that $\boldsymbol{a}_t = \boldsymbol{0}$ for $t \leq 0$. The second method is the exact likelihood method that treats $\boldsymbol{a}_t$ with $t \leq 0$ as additional parameters of the model. To gain some insight into the problem of estimation, we consider the VMA(1) model in Eq. (8.23). Suppose that the data are $\{\boldsymbol{r}_t \mid t = 1, \ldots, T\}$ and $\boldsymbol{a}_t$ is multivariate normal. For a VMA(1) model, the data depend on $\boldsymbol{a}_0$.

***Conditional MLE***

The conditional likelihood method assumes that $\boldsymbol{a}_0 = \boldsymbol{0}$. Under such an assumption and rewriting the model as $\boldsymbol{a}_t = \boldsymbol{r}_t - \boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{a}_{t-1}$, we can compute the shock $\boldsymbol{a}_t$ recursively as

$$\boldsymbol{a}_1 = \boldsymbol{r}_1 - \boldsymbol{\theta}_0, \quad \boldsymbol{a}_2 = \boldsymbol{r}_2 - \boldsymbol{\theta}_0 + \boldsymbol{\Theta}_1\boldsymbol{a}_1, \quad \ldots.$$

Consequently, the likelihood function of the data becomes

$$f(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_T \mid \boldsymbol{\theta}_0, \boldsymbol{\Theta}_1, \boldsymbol{\Sigma}) = \prod_{t=1}^{T} \frac{1}{(2\pi)^{k/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}\boldsymbol{a}_t'\boldsymbol{\Sigma}^{-1}\boldsymbol{a}_t\right],$$

which can be evaluated to obtain the parameter estimates.

***Exact MLE***

For the exact likelihood method, $\boldsymbol{a}_0$ is an unknown vector that must be estimated from the data to evaluate the likelihood function. For simplicity, let $\tilde{r}_t = \boldsymbol{r}_t - \boldsymbol{\theta}_0$ be the mean-corrected series. Using $\tilde{\boldsymbol{r}}_t$ and Eq. (8.23), we have

$$\boldsymbol{a}_t = \tilde{\boldsymbol{r}}_t + \boldsymbol{\Theta}\boldsymbol{a}_{t-1}. \tag{8.25}$$

By repeated substitutions, $\boldsymbol{a}_0$ is related to all $\tilde{\boldsymbol{r}}_t$ as

$$\begin{aligned}
\boldsymbol{a}_1 &= \tilde{\boldsymbol{r}}_1 + \boldsymbol{\Theta}\boldsymbol{a}_0 \\
\boldsymbol{a}_2 &= \tilde{\boldsymbol{r}}_2 + \boldsymbol{\Theta}\boldsymbol{a}_1 = \tilde{\boldsymbol{r}}_2 + \boldsymbol{\Theta}\tilde{\boldsymbol{r}}_1 + \boldsymbol{\Theta}^2\boldsymbol{a}_0 \\
\vdots &= \vdots \\
\boldsymbol{a}_T &= \tilde{\boldsymbol{r}}_T + \boldsymbol{\Theta}\tilde{\boldsymbol{r}}_{T-1} + \cdots + \boldsymbol{\Theta}^{T-1}\tilde{\boldsymbol{r}}_1 + \boldsymbol{\Theta}^T\boldsymbol{a}_0.
\end{aligned} \tag{8.26}$$

Thus, $\boldsymbol{a}_0$ is a linear function of the data if $\boldsymbol{\theta}_0$ and $\boldsymbol{\Theta}$ are given. This result enables us to estimate $\boldsymbol{a}_0$ using the data and initial estimates of $\boldsymbol{\theta}_0$ and $\boldsymbol{\Theta}$. More specifically, given $\boldsymbol{\theta}_0$, $\boldsymbol{\Theta}$, and the data, we can define

$$\boldsymbol{r}_t^* = \tilde{\boldsymbol{r}}_t + \boldsymbol{\Theta}\tilde{\boldsymbol{r}}_{t-1} + \cdots + \boldsymbol{\Theta}^{t-1}\tilde{\boldsymbol{r}}_1, \quad \text{for} \quad t = 1, 2, \ldots, T.$$

Equation (8.26) can then be rewritten as

$$\begin{aligned}
\boldsymbol{r}_1^* &= -\boldsymbol{\Theta}\boldsymbol{a}_0 + \boldsymbol{a}_1 \\
\boldsymbol{r}_2^* &= -\boldsymbol{\Theta}^2\boldsymbol{a}_0 + \boldsymbol{a}_2 \\
\vdots &= \vdots \\
\boldsymbol{r}_T^* &= -\boldsymbol{\Theta}^T\boldsymbol{a}_0 + \boldsymbol{a}_T.
\end{aligned}$$

This is in the form of a multiple linear regression with parameter vector $\boldsymbol{a}_0$, even though the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{a}_t$ may not be a diagonal matrix. If initial estimate of $\boldsymbol{\Sigma}$ is also available, one can premultiply each equation of the prior system by $\boldsymbol{\Sigma}^{-1/2}$, which is the square-root matrix of $\boldsymbol{\Sigma}$. The resulting system is indeed a multiple linear regression, and the ordinary least squares method can be used to obtain an estimate of $\boldsymbol{a}_0$. Denote the estimate by $\widehat{\boldsymbol{a}}_0$.

Using the estimate $\widehat{\boldsymbol{a}}_0$, we can compute the shocks $\boldsymbol{a}_t$ recursively as

$$\boldsymbol{a}_1 = \boldsymbol{r}_1 - \boldsymbol{\theta}_0 + \boldsymbol{\Theta}\widehat{a}_0, \quad \boldsymbol{a}_2 = \boldsymbol{r}_2 - \boldsymbol{\theta}_0 + \boldsymbol{\Theta}\boldsymbol{a}_1, \quad \ldots$$

In addition, we can also derive the exact likelihood function of the data from the joint distribution of $\{\boldsymbol{a}_t \mid t = 0, \ldots, T\}$. The resulting likelihood function can then be evaluated to obtain the exact ML estimates.

In summary, the exact likelihood method works as follows. Given initial estimates of $\boldsymbol{\theta}_0$, $\boldsymbol{\Theta}$, and $\boldsymbol{\Sigma}$, one uses Eq. (8.26) to derive an estimate of $\boldsymbol{a}_0$. This estimate is in

turn used to compute $\boldsymbol{a}_t$ recursively using Eq. (8.25) and starting with $\boldsymbol{a}_1 = \tilde{\boldsymbol{r}}_1 + \boldsymbol{\Theta}\widehat{\boldsymbol{a}}_0$. The resulting $\{\boldsymbol{a}_t\}_{t=1}^{T}$ are then used to evaluate the exact likelihood function of the data to update the estimates of $\boldsymbol{\theta}_0$, $\boldsymbol{\Theta}$, and $\boldsymbol{\Sigma}$. The whole process is then repeated until the estimates converge. This iterative method to evaluate the exact likelihood function applies to the general VMA($q$) models.

From the previous discussion, the exact likelihood method requires more intensive computation than the conditional likelihood approach does. But it provides more accurate parameter estimates, especially when some eigenvalues of $\boldsymbol{\Theta}$ is close to 1 in modulus. Hillmer and Tiao (1979) provide some comparison between the conditional and exact likelihood estimations of VMA models. In multivariate time series analysis, the exact maximum likelihood method becomes important if one suspects that the data might have been overdifferenced. Overdifferencing may occur in many situations (e.g., differencing individual components of a co-integrated system; see discussion later on co-integration).

In summary, building a VMA model involves three steps: (a) use the sample cross-correlation matrixes to specify the order $q$—for a VMA($q$) model, $\boldsymbol{\rho}_\ell = \mathbf{0}$ for $\ell > q$; (b) estimate the specified model by using either the conditional or exact likelihood method—the exact method is preferred when the sample size is not large; and (c) the fitted model should be checked for adequacy (e.g., applying the $Q_k(m)$ statistics to the residual series). Finally, forecasts of a VMA model can be obtained by using the same procedure as a univariate MA model.

**Example 8.5.** Consider again the bivariate series of monthly log returns in percentages of IBM stock and the S&P 500 index from January 1926 to December 1999. Since significant cross-correlations occur mainly at lags 1 and 3, we employ the VMA(3) model

$$\boldsymbol{r}_t = \boldsymbol{\theta}_0 + \boldsymbol{a}_t - \boldsymbol{\Theta}_1\boldsymbol{a}_{t-1} - \boldsymbol{\Theta}_3\boldsymbol{a}_{t-3} \tag{8.27}$$

for the data. Table 8.5 shows the estimation results of the model. The $Q_k(m)$ statistics for the residuals of the simplified model give $Q_2(4) = 17.25$ and $Q_2(8) = 39.30$. Compared with chi-squared distributions with 12 and 28 degrees of freedom, the $p$ values of these statistics are 0.1404 and 0.0762, respectively. Thus, the model is adequate at the 5% significance level.

From Table 8.5, we make the following observations:

1. The difference between conditional and exact likelihood estimates is small for this particular example. This is not surprising because the sample size is not small and, more important, the dynamic structure of the data is weak.
2. The VMA(3) model provides essentially the same dynamic relationship for the series as that of the VAR(3) model in Example 8.4. The monthly log return of IBM stock depends on the previous returns of the S&P 500 index. The market return, in contrast, does not depend on lagged returns of IBM stock. In other words, the dynamic structure of the data is driven by the market return, not by IBM return. The concurrent correlation between the two returns remains strong, however.

**Table 8.5. Estimation Results for Monthly Log Returns of IBM Stock and the S&P 500 Index Using the Vector Moving-Average Model in Eq. (8.27). The Data Span is from January 1926 to December 1999.**

| Parameter | $\theta_0$ | $\Theta_1$ | | $\Theta_3$ | | $\Sigma$ | |
|---|---|---|---|---|---|---|---|
| | (a) Full model with conditional likelihood method | | | | | | |
| Estimate | 1.24 | −.013 | −.121 | −.038 | .108 | 44.48 | 23.52 |
| | 0.54 | .020 | −.101 | .014 | .105 | 23.52 | 31.20 |
| St. Error | 0.24 | .043 | .051 | .044 | .052 | | |
| | 0.18 | .036 | .043 | .036 | .043 | | |
| | (b) Full model with exact likelihood method | | | | | | |
| Estimate | 1.24 | −.013 | −.121 | −.038 | .108 | 44.48 | 23.52 |
| | 0.54 | .020 | −.101 | .013 | .105 | 23.52 | 31.20 |
| St. Error | 0.24 | .043 | .051 | .044 | .052 | | |
| | 0.18 | .036 | .043 | .036 | .043 | | |
| | (c) Simplified model with exact likelihood method | | | | | | |
| Estimate | 1.24 | .000 | −.126 | .000 | .082 | 44.54 | 23.51 |
| | 0.54 | .000 | −.084 | .000 | .114 | 23.51 | 31.21 |
| St. Error | 0.23 | – | .040 | – | .040 | | |
| | 0.18 | – | .033 | – | .033 | | |

## 8.4 VECTOR ARMA MODELS

Univariate ARMA models can also be generalized to handle vector time series. The resulting models are called VARMA models. The generalization, however, encounters some new issues that do not occur in developing VAR and VMA models. One of the issues is the *identifiability* problem. Unlike the univariate ARMA models, VARMA models may not be uniquely defined. For example, the VMA(1) model

$$\begin{bmatrix} r_{1t} \\ r_{2t} \end{bmatrix} = \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} - \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a_{1,t-1} \\ a_{2,t-1} \end{bmatrix}$$

is *identical* to the VAR(1) model

$$\begin{bmatrix} r_{1t} \\ r_{2t} \end{bmatrix} - \begin{bmatrix} 0 & -2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} r_{1,t-1} \\ r_{2,t-1} \end{bmatrix} = \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}.$$

The equivalence of the two models can easily be seen by examining their component models. For the VMA(1) model, we have

$$r_{1t} = a_{1t} - 2a_{2,t-1}, \quad r_{2t} = a_{2t}.$$

For the VAR(1) model, the equations are

$$r_{1t} + 2r_{2,t-1} = a_{1t}, \quad r_{2t} = a_{2t}.$$

From the model for $r_{2t}$, we have $r_{2,t-1} = a_{2,t-1}$. Therefore, the models for $r_{1t}$ are identical. This type of identifiability problem is harmless because either model can be used in a real application.

Another type of identifiability problem is more troublesome. Consider the VARMA(1, 1) model

$$\begin{bmatrix} r_{1t} \\ r_{2t} \end{bmatrix} - \begin{bmatrix} 0.8 & -2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} r_{1,t-1} \\ r_{2,t-1} \end{bmatrix} = \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} - \begin{bmatrix} -0.5 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a_{1,t-1} \\ a_{2,t-1} \end{bmatrix}.$$

This model is identical to the VARMA(1,1) model

$$\begin{bmatrix} r_{1t} \\ r_{2t} \end{bmatrix} - \begin{bmatrix} 0.8 & -2+\eta \\ 0 & \omega \end{bmatrix} \begin{bmatrix} r_{1,t-1} \\ r_{2,t-1} \end{bmatrix} = \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} - \begin{bmatrix} -0.5 & \eta \\ 0 & \omega \end{bmatrix} \begin{bmatrix} a_{1,t-1} \\ a_{2,t-1} \end{bmatrix},$$

for any nonzero $\omega$ and $\eta$. In this particular instance, the equivalence occurs because we have $r_{2t} = a_{2t}$ in both models. The effects of the parameters $\omega$ and $\eta$ on the system cancel out between AR and MA parts of the second model. Such an identifiability problem is serious because, without proper constraints, the likelihood function of a vector ARMA(1,1) model for the data is not uniquely defined, resulting in a situation similar to the exact multicollinearity in a regression analysis. This type of identifiability problem can occur in a vector model even if none of the components is a white noise series.

These two simple examples highlight the new issues involved in the generalization to VARMA models. Building a VARMA model for a given data set thus requires some attention. In the time series literature, methods of *structural specification* have been proposed to overcome the identifiability problem; see Tiao and Tsay (1989), Tsay (1991), and the references therein. We do not discuss the detail of structural specification here because VAR and VMA models are sufficient in most financial applications. When VARMA models are used, only lower order models are entertained (e.g., a VARMA(1, 1) or VARMA(2, 1) model) especially when the time series involved are not seasonal.

A VARMA$(p, q)$ model can be written as

$$\boldsymbol{\Phi}(B)\boldsymbol{r}_t = \boldsymbol{\phi}_0 + \boldsymbol{\Theta}(B)\boldsymbol{a}_t,$$

where $\boldsymbol{\Phi}(B) = \boldsymbol{I} - \boldsymbol{\Phi}_1 B - \cdots - \boldsymbol{\Phi}_p B^p$ and $\boldsymbol{\Theta}(B) = \boldsymbol{I} - \boldsymbol{\Theta}_1 B - \cdots - \boldsymbol{\Theta}_q B^q$ are two $k \times k$ matrix polynomials. We assume that the two matrix polynomials have no left common factors; otherwise, the model can be simplified. The necessary and sufficient condition of weak stationarity for $\boldsymbol{r}_t$ is the same as that for the VAR$(p)$ model with matrix polynomial $\boldsymbol{\Phi}(B)$. For $v > 0$, the $(i, j)$th elements of the coefficient matrixes $\boldsymbol{\Phi}_v$ and $\boldsymbol{\Theta}_v$ measure the linear dependence of $r_{1t}$ on $r_{j,t-v}$ and $a_{j,t-v}$, respectively. If the $(i, j)$th element is zero for all AR and MA coefficient matrixes,

then $r_{it}$ does not depend on the lagged values of $r_{jt}$. However, the converse proposition does not hold in a VARMA model. In other words, nonzero coefficients at the $(i, j)$th position of AR and MA matrixes may exist even when $r_{it}$ does not depend on any lagged value of $r_{jt}$.

To illustrate, consider the following bivariate model

$$\begin{bmatrix} \Phi_{11}(B) & \Phi_{12}(B) \\ \Phi_{21}(B) & \Phi_{22}(B) \end{bmatrix} \begin{bmatrix} r_{1t} \\ r_{2t} \end{bmatrix} = \begin{bmatrix} \Theta_{11}(B) & \Theta_{12}(B) \\ \Theta_{21}(B) & \Theta_{22}(B) \end{bmatrix} \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}.$$

Here the necessary and sufficient conditions for the existence of a unidirectional dynamic relationship from $r_{1t}$ to $r_{2t}$ are

$$\begin{aligned} &\Phi_{22}(B)\Theta_{12}(B) - \Phi_{12}(B)\Theta_{22}(B) = 0, \quad \text{but} \\ &\Phi_{11}(B)\Theta_{21}(B) - \Phi_{21}(B)\Theta_{11}(B) \neq 0. \end{aligned} \tag{8.28}$$

These conditions can be obtained as follows. Letting

$$\Omega(B) = |\,\mathbf{\Phi}(B)\,| = \Phi_{11}(B)\Phi_{22}(B) - \Phi_{12}(B)\Phi_{21}(B)$$

be the determinant of the AR matrix polynomial and premultiplying the model by the matrix

$$\begin{bmatrix} \Phi_{22}(B) & -\Phi_{12}(B) \\ -\Phi_{21}(B) & \Phi_{11}(B) \end{bmatrix},$$

we can rewrite the bivariate model as

$$\mathbf{\Omega}(B) \begin{bmatrix} r_{1t} \\ r_{2t} \end{bmatrix} = \begin{bmatrix} \Phi_{22}(B)\Theta_{11}(B) - \Phi_{12}(B)\Theta_{21}(B) & \Phi_{22}(B)\Theta_{12}(B) - \Phi_{12}(B)\Theta_{22}(B) \\ \Phi_{11}(B)\Theta_{21}(B) - \Phi_{21}(B)\Theta_{11}(B) & \Phi_{11}(B)\Theta_{22}(B) - \Phi_{21}(B)\Theta_{12}(B) \end{bmatrix} \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}.$$

Consider the equation for $r_{1t}$. The first condition in Eq. (8.28) shows that $r_{1t}$ does not depend on any past value of $a_{2t}$ or $r_{2t}$. From the equation for $r_{2t}$, the second condition in Eq. (8.28) implies that $r_{2t}$ indeed depends on some past values of $a_{1t}$. Based on Eq. (8.28), $\Theta_{12}(B) = \Phi_{12}(B) = 0$ is a sufficient, but not necessary, condition for the unidirectional relationship from $r_{1t}$ to $r_{2t}$.

Estimation of a VARMA model can be carried out by either the conditional or exact maximum likelihood method. The $Q_k(m)$ statistic continues to apply to the residual series of a fitted model, but the degrees of freedom of its asymptotic chi-squared distribution are $k^2m - g$, where $g$ is the number of estimated parameters in both the AR and MA coefficient matrixes.

**Example 8.6.** To demonstrate VARMA modeling, we consider two U.S. monthly interest-rate series. The first series is the 1-year Treasury constant maturity

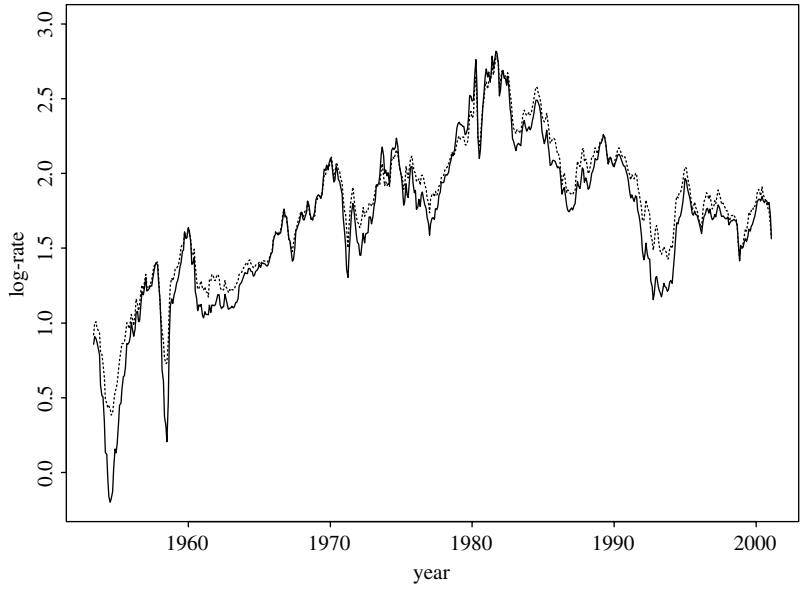


**Figure 8.5.** Time plots of log U.S. monthly interest rates from April 1953 to January 2001. The solid line denotes the 1-year Treasury constant maturity rate, and the dashed line denotes the 3-year rate.

rate, and the second series is the 3-year Treasury constant maturity rate. The data are obtained from the Federal Reserve Bank of St Louis, and the sampling period is from April 1953 to January 2001. There are 574 observations. To ensure the positiveness of U.S. interest rates, we analyze the log series. Figure 8.5 shows the time plots of the two log interest-rate series. The solid line denotes the 1-year maturity rate. The two series moved closely in the sampling period.

The $M(i)$ statistics and AIC criterion specify a VAR(4) model for the data. However, we employ a VARMA(2, 1) model because the two models provide similar fits. Table 8.6 shows the parameter estimates of the VARMA(2, 1) model obtained by the exact likelihood method. We removed the insignificant parameters and reestimated the simplified model. The residual series of the fitted model has some minor serial

**Table 8.6. Parameter Estimates of a VARMA(2, 1) Model for Two Monthly U.S. Interest-Rate Series.**

| Par. | $\boldsymbol{\Phi}_1$ | | $\boldsymbol{\Phi}_2$ | | $\phi_0$ | $\boldsymbol{\Theta}_1$ | | $\boldsymbol{\Sigma} \times 10^3$ | |
|---|---|---|---|---|---|---|---|---|---|
| Est. | 1.57 | −0.54 | −0.60 | 0.56 | .020 | 0.60 | −1.17 | 3.58 | 2.50 |
| | | 0.99 | | | .025 | | −0.47 | 2.50 | 2.19 |
| Std. | 0.10 | 0.16 | 0.09 | 0.15 | .013 | 0.11 | 0.18 | | |
| | | 0.01 | | | .011 | | 0.04 | | |

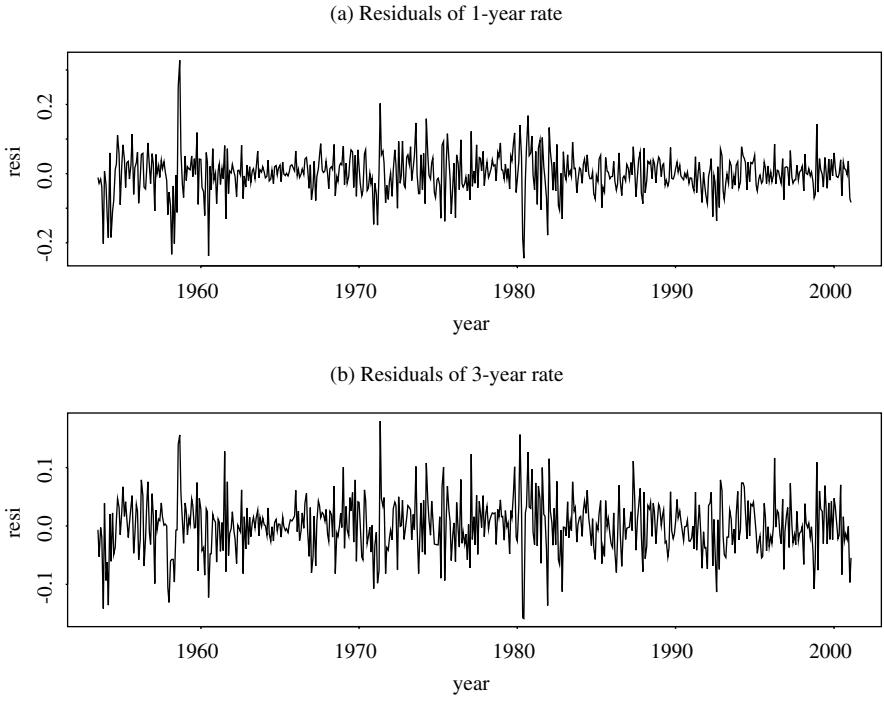


**Figure 8.6.** Residual plots for log U.S. monthly interest-rate series. The fitted model is a VARMA(2, 1).

and cross-correlations at lags 7 and 11. Figure 8.6 shows the residual plots and indicates the existence of some outlying data points. The model can be further improved, but it seems to capture the dynamic structure of the data reasonably well.

The final VARMA(2, 1) model shows some interesting characteristics of the data. First, the interest-rate series are highly contemporaneously correlated. The concurrent correlation coefficient is $2.5/\sqrt{3.58 \times 2.19} = 0.893$. Second, there is a unidirectional linear relationship from the 3-year rate to the 1-year rate because the (2, 1)th elements of all AR and MA matrixes are zero, but some (2, 1)th element is not zero. As a matter of fact, the model in Table 8.6 shows that

$$\begin{aligned} r_{3t} &= 0.025 + 0.99 r_{3,t-1} + a_{3t} + 0.47 a_{3,t-1} \\ r_{1t} &= 0.020 + 1.57 r_{1,t-1} - 0.60 r_{1,t-2} - 0.54 r_{3,t-1} + 0.56 r_{3,t-2} \\ &\quad + a_{1t} - 0.60 a_{1,t-1} + 1.17 a_{3,t-1}, \end{aligned}$$

where $r_{it}$ is the log series of $i$-year interest rate and $a_{it}$ is the corresponding shock series. Therefore, the 3-year interest rate does not depend on the past values of 1-year rate, but the 1-year rate depends on the past values of 3-year rate. Third, the two

interest-rate series appear to be unit-root nonstationary. Using the back-shift operator $B$, the model can be rewritten approximately as

$$(1-B)r_{3t} = 0.025 + (1+0.47B)a_{3t}$$
$$(1-B)(1-0.6B)r_{1t} = 0.02 - 0.55B(1-B)r_{3,t} + (1-0.6B)a_{1t} + 1.17Ba_{3,t}.$$

### 8.4.1 Marginal Models of Components

Given a vector model for $\boldsymbol{r}_t$, the implied univariate models for the components $r_{it}$ are the *marginal* models. For a $k$-dimensional ARMA$(p, q)$ model, the marginal models are ARMA$[kp, (k-1)p+q]$. This result can be obtained in two steps. First, the marginal models of a VMA$(q)$ model is univariate MA$(q)$. Assume that $\boldsymbol{r}_t$ is a VMA$(q)$ process. Because the cross-correlation matrix of $\boldsymbol{r}_t$ vanishes after lag $q$ (i.e., $\boldsymbol{\rho}_\ell = \mathbf{0}$ for $\ell > q$), the ACF of $r_{it}$ is zero beyond lag $q$. Therefore, $r_{it}$ is an MA process and its univariate model is in the form $r_{it} = \theta_{i,0} + \sum_{j=1}^{q} \theta_{i,j} b_{i,t-j}$, where $\{b_{it}\}$ is a sequence of uncorrelated random variables with mean zero and variance $\sigma_{ib}^2$. The parameters $\theta_{i,j}$ and $\sigma_{ib}$ are functions of the parameters of the VMA model for $\boldsymbol{r}_t$.

The second step to obtain the result is to diagonalize the AR matrix polynomial of a VARMA$(p, q)$ model. For illustration, consider the bivariate AR(1) model

$$\begin{bmatrix} 1-\Phi_{11}B & -\Phi_{12}B \\ -\Phi_{21}B & 1-\Phi_{22}B \end{bmatrix} \begin{bmatrix} r_{1t} \\ r_{2t} \end{bmatrix} = \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}.$$

Premultiplying the model by the matrix polynomial

$$\begin{bmatrix} 1-\Phi_{22}B & \Phi_{12}B \\ \Phi_{21}B & 1-\Phi_{11}B \end{bmatrix},$$

we obtain

$$[(1-\Phi_{11}B)(1-\Phi_{22}B) - \Phi_{12}\Phi_{22}B^2] \begin{bmatrix} r_{1t} \\ r_{2t} \end{bmatrix} = \begin{bmatrix} 1-\Phi_{22}B & -\Phi_{12}B \\ -\Phi_{21}B & 1-\Phi_{11}B \end{bmatrix} \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}.$$

The left-hand side of the prior equation shows that the univariate AR polynomials for $r_{it}$ are of order 2. In contrast, the right-hand side of the equation is in a VMA(1) form. Using the result of VMA models in step 1, we show that the univariate model for $r_{it}$ is ARMA(2, 1). The technique generalizes easily to the $k$-dimensional VAR(1) model, and the marginal models are ARMA$(k, k-1)$. More generally, for a $k$-dimensional VAR$(p)$ model, the marginal models are ARMA$[kp, (k-1)p]$. The result for VARMA models follows directly from those of VMA and VAR models.

The order $[kp, (k-1)p+q]$ is the maximum order (i.e., the upper bound) for the marginal models. The actual marginal order of $r_{it}$ can be much lower.

## 8.5 UNIT-ROOT NONSTATIONARITY AND CO-INTEGRATION

When modeling several unit-root nonstationary time series jointly, one may encounter the case of *co-integration*. Consider the bivariate ARMA(1, 1) model

$$\begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} - \begin{bmatrix} 0.5 & -1.0 \\ -0.25 & 0.5 \end{bmatrix} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \end{bmatrix} = \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} - \begin{bmatrix} 0.2 & -0.4 \\ -0.1 & 0.2 \end{bmatrix} \begin{bmatrix} a_{1,t-1} \\ a_{2,t-1} \end{bmatrix}, \quad (8.29)$$

where the covariance matrix $\boldsymbol{\Sigma}$ of the shock $\boldsymbol{a}_t$ is positive definite. This is not a weakly stationary model because the two eigenvalues of the AR coefficient matrix are 0 and 1. Figure 8.7 shows the time plots of a simulated series of the model with 200 data points and $\boldsymbol{\Sigma} = \boldsymbol{I}$, whereas Figure 8.8 shows that sample autocorrelations of the two component series $x_{it}$. It is easy to see that the two series have high autocorrelations and exhibit features of unit-root nonstationarity. The two marginal models of $\boldsymbol{x}_t$ are indeed unit-root nonstationary. Rewrite the model as

$$\begin{bmatrix} 1 - 0.5B & B \\ 0.25B & 1 - 0.5B \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} = \begin{bmatrix} 1 - 0.2B & 0.4B \\ 0.1B & 1 - 0.2B \end{bmatrix} \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}.$$

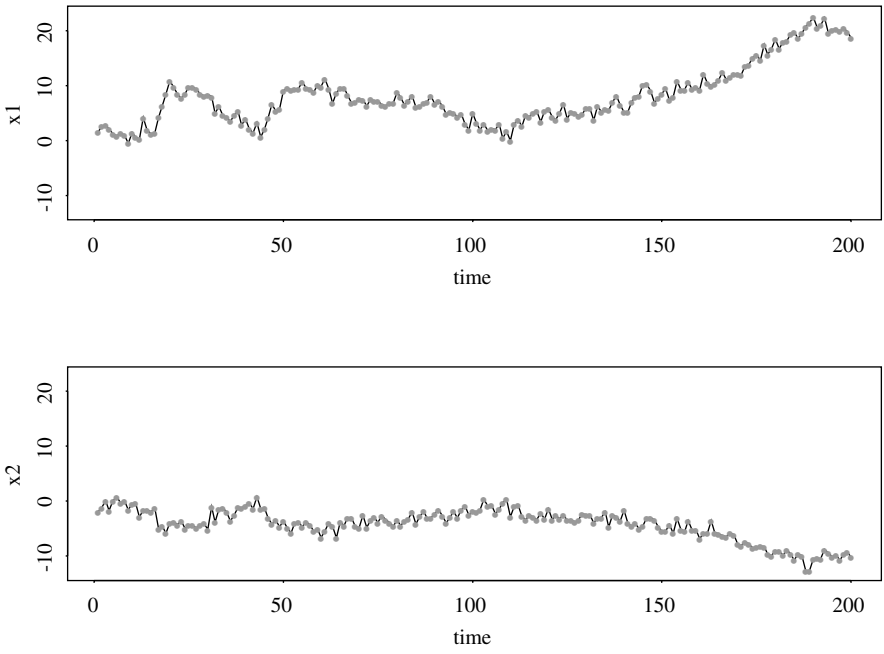


**Figure 8.7.** Time plots of a simulated series based on model (8.29) with identity covariance matrix for the shocks.

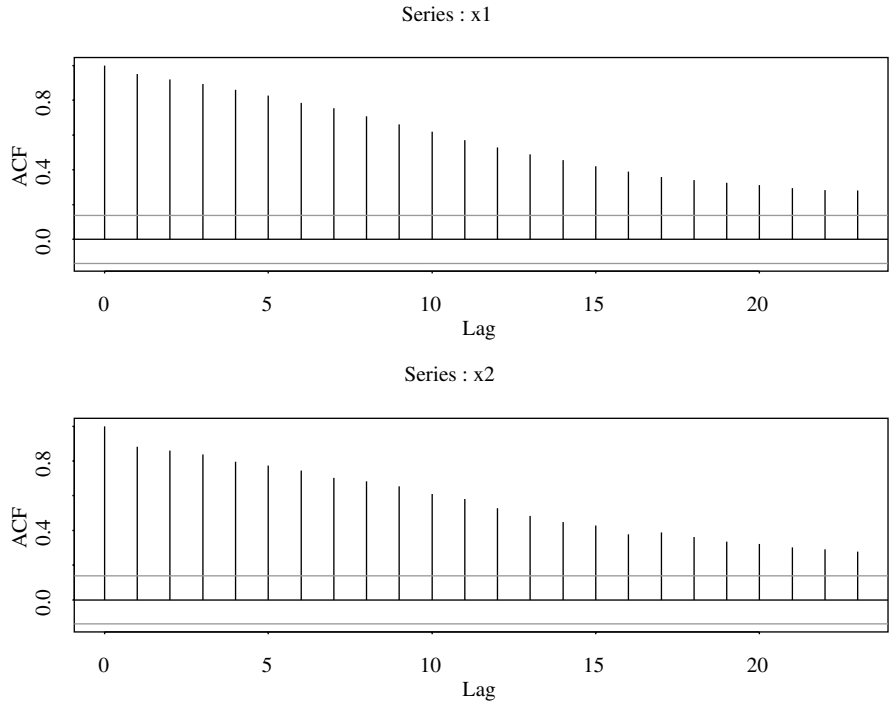


**Figure 8.8.** Sample autocorrelation functions of two simulated component series. There are 200 observations, and the model is given by Eq. (8.29) with identity covariance matrix for the shocks.

Premultiplying the prior equation by

$$\begin{bmatrix} 1-0.5B & -B \\ -0.25B & 1-0.5B \end{bmatrix},$$

we obtain the result

$$\begin{bmatrix} 1-B & 0 \\ 0 & 1-B \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} = \begin{bmatrix} 1-0.7B & -0.6B \\ -0.15B & 1-0.7B \end{bmatrix} \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix}.$$

Therefore, each component $x_{it}$ of the model is unit-root nonstationary and follows an ARIMA(0, 1, 1) model.

However, we can consider a linear transformation by defining

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} 1.0 & -2.0 \\ 0.5 & 1.0 \end{bmatrix} \begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} \equiv \boldsymbol{L}\boldsymbol{x}_t,$$

$$\begin{bmatrix} b_{1t} \\ b_{2t} \end{bmatrix} = \begin{bmatrix} 1.0 & -2.0 \\ 0.5 & 1.0 \end{bmatrix} \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} \equiv \boldsymbol{L}\boldsymbol{a}_t.$$

The VARMA model of the transformed series $\boldsymbol{y}_t$ can be obtained as follows:

$$\begin{aligned}\boldsymbol{L}\boldsymbol{x}_t &= \boldsymbol{L}\boldsymbol{\Phi}\boldsymbol{x}_{t-1} + \boldsymbol{L}\boldsymbol{a}_t - \boldsymbol{L}\boldsymbol{\Theta}\boldsymbol{a}_{t-1} \\ &= \boldsymbol{L}\boldsymbol{\Phi}\boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{x}_{t-1} + \boldsymbol{L}\boldsymbol{a}_t - \boldsymbol{L}\boldsymbol{\Theta}\boldsymbol{L}^{-1}\boldsymbol{L}\boldsymbol{a}_{t-1} \\ &= \boldsymbol{L}\boldsymbol{\Phi}\boldsymbol{L}^{-1}(\boldsymbol{L}\boldsymbol{x}_{t-1}) + \boldsymbol{b}_t - \boldsymbol{L}\boldsymbol{\Theta}\boldsymbol{L}^{-1}\boldsymbol{b}_{t-1}.\end{aligned}$$

Thus, the model for $\boldsymbol{y}_t$ is

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} - \begin{bmatrix} 1.0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} = \begin{bmatrix} b_{1t} \\ b_{2t} \end{bmatrix} - \begin{bmatrix} 0.4 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} b_{1,t-1} \\ b_{2,t-1} \end{bmatrix}. \tag{8.30}$$

From the prior model, we see that (a) $y_{1t}$ and $y_{2t}$ are uncoupled series with concurrent correlation equal to that between the shocks $b_{1t}$ and $b_{2t}$, (b) $y_{1t}$ follows a univariate ARIMA(0,1,1) model, and (c) $y_{2t}$ is a white noise series (i.e., $y_{2t} = b_{2t}$). In particular, the model in Eq. (8.30) shows that there is *only* a single unit root in the system. Consequently, the unit roots of $x_{1t}$ and $x_{2t}$ are introduced by the unit root of $y_{1t}$.

The phenomenon that both $x_{1t}$ and $x_{2t}$ are unit-root nonstationary, but there is only a single unit root in the vector series, is referred to as *co-integration* in the econometric and time series literature. Another way to define co-integration is to focus on linear transformations of unit-root nonstationary series. For the simulated example of model (8.29), the transformation shows that the linear combination $y_{2t} = 0.5x_{1t} + x_{2t}$ does not have a unit root. Consequently, $x_{1t}$ and $x_{2t}$ are co-integrated if (a) both of them are unit-root nonstationary, and (b) they have a linear combination that is unit-root stationary.

Generally speaking, for a $k$-dimensional unit-root nonstationary time series, co-integration exists if there are less than $k$ unit roots in the system. Let $h$ be the number of unit roots in the $k$-dimensional series $\boldsymbol{x}_t$. Co-integration exists if $0 < h < k$, and the quantity $k - h$ is called the number of co-integrating factors. Alternatively, the number of co-integrating factors is the number of different linear combinations that are unit-root stationary. The linear combinations are called the co-integrating vectors. For the prior simulated example, $y_{2t} = (0.5, 1)\boldsymbol{x}_t$ so that $(0.5, 1)'$ is a co-integrating vector for the system. For more discussions on co-integration and co-integration tests, see Box and Tiao (1977), Engle and Granger (1987), Stock and Watson (1988), and Johansen (1989).

The concept of co-integration is interesting and has attracted a lot of attention in the literature. However, there are difficulties in testing for co-integration in a real application. The main source of difficulties is that co-integration tests overlook the scaling effects of the component series. Interested readers are referred to Cochrane (1988) and Tiao, Tsay, and Wang (1993) for further discussion.

While I have some misgivings on the practical value of co-integration tests, the idea of co-integration is highly relevant in financial study. For example, consider the stock of Finnish Nokia Corporation. Its price on the Helsinki Stock Market must move in unison with the price of its American Depositary Receipts on the New York

Stock Exchange; otherwise there exists some arbitrage opportunity for investors. If the stock price has a unit root, then the two price series must be co-integrated. In practice, such a co-integration can exist after adjusting for transaction costs and exchange-rate risk. We discuss issues like this later in Section 8.6.

### 8.5.1 An Error-Correction Form

Because there are more unit-root nonstationary components than the number of unit roots in a co-integrated system, differencing individual components to achieve stationarity results in overdifferencing. Overdifferencing leads to the problem of unit roots in the MA matrix polynomial, which in turn may encounter difficulties in parameter estimation. If the MA matrix polynomial contains unit roots, the vector time series is said to be noninvertible.

Engle and Granger (1987) discuss an error-correction representation for a co-integrated system that overcomes the difficulty of estimating noninvertible VARMA models. Consider the co-integrated system in Eq. (8.29). Let $\nabla \boldsymbol{x}_t = \boldsymbol{x}_t - \boldsymbol{x}_{t-1}$ be the differenced series. Subtracting $\boldsymbol{x}_{t-1}$ from both sides of the equation, we obtain a model for $\nabla \boldsymbol{x}_t$ as

$$\begin{bmatrix} \nabla x_{1t} \\ \nabla x_{2t} \end{bmatrix} = \begin{bmatrix} -0.5 & -1.0 \\ -0.25 & -0.5 \end{bmatrix} \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} - \begin{bmatrix} 0.2 & -0.4 \\ -0.1 & 0.2 \end{bmatrix} \begin{bmatrix} a_{1,t-1} \\ a_{2,t-1} \end{bmatrix}$$
$$= \begin{bmatrix} -1 \\ -0.5 \end{bmatrix} [0.5, 1.0] \begin{bmatrix} x_{1,t-1} \\ x_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix} - \begin{bmatrix} 0.2 & -0.4 \\ -0.1 & 0.2 \end{bmatrix} \begin{bmatrix} a_{1,t-1} \\ a_{2,t-1} \end{bmatrix}.$$

This is a stationary model because both $\nabla \boldsymbol{x}_t$ and $[0.5, 1.0]\boldsymbol{x}_t = y_{2t}$ are unit-root stationary. Because $\boldsymbol{x}_{t-1}$ is used in the right-hand side of the previous equation, the MA matrix polynomial is the same as before and, hence, the model does not encounter the problem of noninvertibility. Such a formulation is referred to as an error-correction model for $\nabla \boldsymbol{x}_t$, and it can be extended to the general co-integrated VARMA model. For a co-integrated VARMA$(p, q)$ model with $m$ co-integrating factors, an error-correction representation is

$$\nabla \boldsymbol{x}_t = \boldsymbol{\alpha}\boldsymbol{\beta}\boldsymbol{x}_{t-1} + \sum_{i=1}^{p-1} \boldsymbol{\Phi}_i^* \nabla \boldsymbol{x}_{t-i} + \boldsymbol{a}_t - \sum_{j=1}^{q} \boldsymbol{\Theta}_j \boldsymbol{a}_{t-j}, \tag{8.31}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $k \times m$ and $m \times k$ full-rank matrixes. The AR coefficient matrixes $\boldsymbol{\Phi}_i^*$ are functions of the original coefficient matrixes $\boldsymbol{\Phi}_j$; see the Remark later. The time series $\boldsymbol{\beta}\boldsymbol{x}_t$ is unit-root stationary, and the rows of $\boldsymbol{\beta}$ are the co-integrating vectors of $\boldsymbol{x}_t$.

Existence of the stationary series $\boldsymbol{\beta}\boldsymbol{x}_{t-1}$ in the error-correction representation (8.31) is natural. It can be regarded as a "compensation" term for the over differenced system $\nabla \boldsymbol{x}_t$. The stationarity of $\boldsymbol{\beta}\boldsymbol{x}_{t-1}$ can be justified as follows. The theory of unit-root time series shows that the sample correlation coefficient between a unit-root nonstationary series and a stationary series converges to zero as the sample size goes

to infinity; see Tsay and Tiao (1990) and the references therein. In an error-correction representation, $\boldsymbol{x}_{t-1}$ is unit-root nonstationary, but $\nabla \boldsymbol{x}_t$ is stationary. Therefore, the only way that $\nabla \boldsymbol{x}_t$ can relate meaningfully to $\boldsymbol{x}_{t-1}$ is through a stationary series $\boldsymbol{\beta} x_{t-1}$.

***Remark:*** The AR coefficient matrixes of the error-correction model in Eq. (8.31) are given by

$$\begin{aligned}
\boldsymbol{\Phi}_{p-1}^* &= -\boldsymbol{\Phi}_p \\
\boldsymbol{\Phi}_{p-2}^* &= -\boldsymbol{\Phi}_{p-1} - \boldsymbol{\Phi}_p \\
&\vdots = \vdots \\
\boldsymbol{\Phi}_1^* &= -\boldsymbol{\Phi}_2 - \cdots - \boldsymbol{\Phi}_p \\
\boldsymbol{\alpha}\boldsymbol{\beta} &= \boldsymbol{\Phi}_p + \boldsymbol{\Phi}_{p-1} + \cdots + \boldsymbol{\Phi}_1 - \boldsymbol{I} = -\boldsymbol{\Phi}(1).
\end{aligned}$$

***Remark:*** Our discussion of co-integration assumes that all unit roots are of multiplicity 1, but the concept can be extended to cases in which the unit roots have different multiplicities. Also, if the number of co-integrating factors $m$ is given, then the error-correction model in Eq. (8.31) can be estimated by likelihood methods. Finally, there are many ways to construct an error-correction representation. In fact, one can use any $\boldsymbol{\alpha}\boldsymbol{\beta}\boldsymbol{x}_{t-v}$ for $1 \leq v \leq p$ in Eq. (8.31) with some modifications to the AR coefficient matrixes $\boldsymbol{\Phi}_i^*$.

## 8.6 THRESHOLD CO-INTEGRATION AND ARBITRAGE

In this section, we focus on detecting arbitrage opportunities in index trading by using multivariate time series methods. We also demonstrate that simple univariate nonlinear models of Chapter 4 can be extended naturally to the multivariate case in conjunction with the idea of co-integration.

Our study considers the relationship between the price of S&P 500 index futures and the price of the shares underlying the index on the cash market. Let $f_{t,\ell}$ be the log price of the index futures at time $t$ with maturity $\ell$, and let $s_t$ be the log price of the shares underlying the index on the cash market at time $t$. A version of the *cost-of-carry model* in the finance literature states

$$f_{t,\ell} - s_t = (r_{t,\ell} - q_{t,\ell})(\ell - t) + z_t^*, \tag{8.32}$$

where $r_{t,\ell}$ is the risk-free interest rate, $q_{t,\ell}$ is the dividend yield with respect to the cash price at time $t$, and $(\ell - t)$ is the time to maturity of the futures contract; see Brenner and Kroner (1995), Dwyer, Locke, and Yu (1996), and the references therein.

The $z_t^*$ process of model (8.32) must be unit-root stationary; otherwise there exist *persistent* arbitrage opportunities. Here an arbitrage trading consists of simultane-

ously buying (short-selling) the security index and selling (buying) the index futures whenever the log prices diverge by more than the cost of carrying the index over time until maturity of the futures contract. Under the weak stationarity of $z_t^*$, for arbitrage to be profitable, $z_t^*$ must exceed a certain value in modulus determined by transaction costs and other economic and risk factors.

It is commonly believed that the $f_{t,\ell}$ and $s_t$ series of the S&P 500 index contain a unit root, but Eq. (8.32) indicates that they are co-integrated after adjusting for the effect of interest rate and dividend yield. The co-integrating vector is $(1, -1)$ after the adjustment, and the co-integrated series is $z_t^*$. Therefore, one should use an error-correction form to model the return series $\boldsymbol{r}_t = (\nabla f_t, \nabla s_t)'$, where $\nabla f_t = f_{t,\ell} - f_{t-1,\ell}$ and $\nabla s_t = s_t - s_{t-1}$, where for ease in notation we drop the maturity time $\ell$ from the subscript of $\nabla f_t$.

### 8.6.1 Multivariate Threshold Model

In practice, arbitrage tradings affect the dynamic of the market, and hence the model for $\boldsymbol{r}_t$ may vary over time depending on the presence or absence of arbitrage tradings. Consequently, the prior discussions lead naturally to the model

$$\boldsymbol{r}_t = \begin{cases} \boldsymbol{c}_1 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i^{(1)} \boldsymbol{r}_{t-i} + \boldsymbol{\beta}_1 z_{t-1} + \boldsymbol{a}_t^{(1)} & \text{if } z_{t-1} \le \gamma_1 \\ \boldsymbol{c}_2 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i^{(2)} \boldsymbol{r}_{t-i} + \boldsymbol{\beta}_2 z_{t-1} + \boldsymbol{a}_t^{(2)} & \text{if } \gamma_1 < z_{t-1} \le \gamma_2 \\ \boldsymbol{c}_3 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i^{(3)} \boldsymbol{r}_{t-i} + \boldsymbol{\beta}_3 z_{t-1} + \boldsymbol{a}_t^{(3)} & \text{if } \gamma_2 < z_{t-1}, \end{cases} \tag{8.33}$$

where $z_t = 100 z_t^*$, $\gamma_1 < 0 < \gamma_2$ are two real numbers, and $\{\boldsymbol{a}_t^{(i)}\}$ are sequences of two-dimensional white noises and are independent of each other. Here we use $z_t = 100 z_t^*$ because the actual value of $z_t^*$ is relatively small.

The model in Eq. (8.33) is referred to as a multivariate threshold model with three regimes. The two real numbers $\gamma_1$ and $\gamma_2$ are the thresholds and $z_{t-1}$ is the threshold variable. The threshold variable $z_{t-1}$ is supported by the data; see Tsay (1998). In general, one can select $z_{t-d}$ as a threshold variable by considering $d \in \{1, \ldots, d_0\}$, where $d_0$ is a prespecified positive integer.

Model (8.33) is a generalization of the threshold autoregressive model of Chapter 4. It is also a generalization of the error-correlation model of Eq. (8.31). As mentioned earlier, an arbitrage trading is profitable only when $z_t^*$ or, equivalently, $z_t$ is large in modulus. Therefore, arbitrage tradings only occurred in regimes 1 and 3 of model (8.33). As such, the dynamic relationship between $f_{t,\ell}$ and $s_t$ in regime 2 is determined mainly by the normal market force, and hence the two series behave more or less like a random walk. In other words, the two log prices in the middle regime should be free from arbitrage effects and, hence, free from the co-integration constraint. From an econometric viewpoint, this means that the estimate of $\boldsymbol{\beta}_2$ in the middle regime should be insignificant.

In summary, we expect that the co-integration effects between the log price of the futures and the log price of security index on the cash market are significant in

regimes 1 and 3, but insignificant in regime 2. This phenomenon is referred to as a *threshold co-integration*; see Balke and Fomby (1997).

### 8.6.2 The Data

The data used in this case study are the intraday transaction data of the S&P 500 index in May 1993 and its June futures contract traded at the Chicago Mercantile Exchange; see Forbes, Kalb, and Kofman (1999), who used the data to construct a minute-by-minute bivariate price series with 7060 observations. To avoid the unduly influence of unusual returns, I replaced 10 extreme values (5 on each side) by the simple average of their two nearest neighbors. This step does not affect the qualitative conclusion of the analysis, but may affect the conditional heteroscedasticity in the data. For simplicity, we do not consider conditional heteroscedasticity in the study. Figure 8.9 shows the time plots of the log returns of the index futures and cash prices and the associated threshold variable $z_t = 100z_t^*$ of model (8.32).

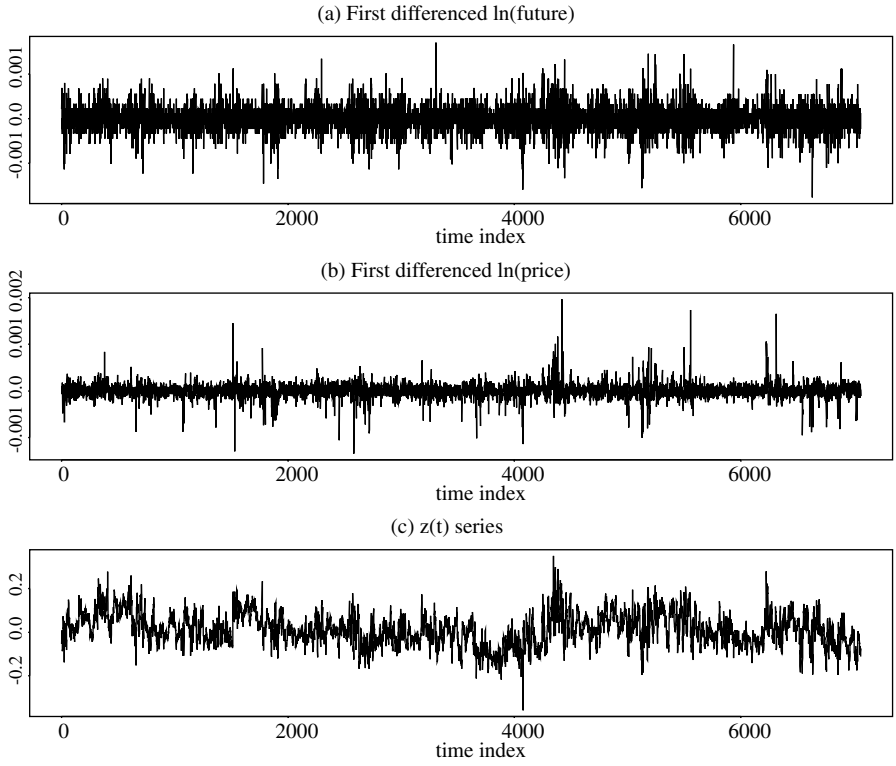


**Figure 8.9.** Time plots of 1-minute log returns of the S&P 500 index futures and cash prices and the associated threshold variable in May 1993: (a) log returns of the index futures, (b) log returns of the index cash prices, and (c) the $z_t$ series.

### 8.6.3 Estimation

A formal specification of the multivariate threshold model in Eq. (8.33) includes selecting the threshold variable, determining the number of regimes, and choosing the order $p$ for each regime. Interested readers are referred to Tsay (1998) and Forbes, Kalb, and Kofman (1999). The thresholds $\gamma_1$ and $\gamma_2$ can be estimated by using some information criteria (e.g., the Akaike information criterion [AIC] or the sum of squares of residuals). Assuming $p = 8$, $d \in \{1, 2, 3, 4\}$, $\gamma_1 \in [-0.15, -0.02]$, and $\gamma_2 \in [0.025, 0.145]$, and using a grid search method with 300 points on each of the two intervals, the AIC selects $z_{t-1}$ as the threshold variable with thresholds $\hat{\gamma}_1 = -0.0226$ and $\hat{\gamma}_2 = 0.0377$. Details of the parameter estimates are given in Table 8.7.

From Table 8.7, we make the following observations. First, the $t$ ratios of $\widehat{\beta}_2$ in the middle regime show that, as expected, the estimates are insignificant at the 5% level, confirming that there is no co-integration between the two log prices in the absence of arbitrage opportunities. Second, $\nabla f_t$ depends negatively on $\nabla f_{t-1}$ in all three regimes. This is in agreement with the bid-ask bounce discussed in Chapter 5. Third, past log returns of the index futures seem to be more informative than the past log returns of the cash prices because there are more significant $t$ ratios in $\nabla f_{t-i}$ than in $\nabla s_{t-i}$. This is reasonable because futures series are in general more liquid. For more information on index arbitrage, see Dwyer, Locke, and Yu (1996).

## 8.7 PRINCIPAL COMPONENT ANALYSIS

We have focused on modeling the dynamic structure of a vector time series in the previous sections. Of equal importance in multivariate time series analysis is the covariance (or correlation) structure of the series. For example, the covariance structure of a vector return series plays an important role in portfolio selection. In what follows, we discuss some statistical methods useful in studying the covariance structure of a vector time series.

Given a $k$-dimensional random variable $\boldsymbol{r} = (r_1, \ldots, r_k)'$ with covariance matrix $\boldsymbol{\Sigma}_r$, a *principal component analysis* (PCA) is concerned with using a few linear combinations of $r_i$ to explain the structure of $\boldsymbol{\Sigma}_r$. If $\boldsymbol{r}$ denotes the monthly log returns of $k$ assets, then PCA can be used to study the source of variations of these $k$ asset returns. Here the key word is *few* so that simplification can be achieved in multivariate analysis.

### 8.7.1 Theory of PCA

PCA applies to either the covariance matrix $\boldsymbol{\Sigma}_r$ or the correlation matrix $\boldsymbol{\rho}_r$ of $\boldsymbol{r}$. Since the correlation matrix is the covariance matrix of the standardized random vector $\boldsymbol{r}^* = \boldsymbol{D}^{-1}\boldsymbol{r}$, where $\boldsymbol{D}$ is the diagonal matrix of standard deviations of the components of $\boldsymbol{r}$, we use covariance matrix in our theoretical discussion. Let $\boldsymbol{c}_i =$

**Table 8.7. Least Squares Estimates and Their *t* Ratios of the Multivariate Threshold Model in Eq. (8.33) for the S&P 500 Index Data in May 1993. The Numbers of Data Points for the Three Regimes are 2234, 2410, and 2408, Respectively.**

| | Regime 1 | | Regime 2 | | Regime 3 | |
|---|---|---|---|---|---|---|
| | $\nabla f_t$ | $\nabla s_t$ | $\nabla f_t$ | $\nabla s_t$ | $\nabla f_t$ | $\nabla s_t$ |
| $\phi_0$ | 0.00002 | 0.00005 | 0.00000 | 0.00000 | −0.00001 | −0.00005 |
| $t$ | ( 1.47) | ( 7.64) | (−0.07) | ( 0.53) | (−0.74) | (−6.37) |
| $\nabla f_{t-1}$ | −0.08468 | 0.07098 | −0.03861 | 0.04037 | −0.04102 | 0.02305 |
| $t$ | (−3.83) | ( 6.15) | (−1.53) | ( 3.98) | (−1.72) | ( 1.96) |
| $\nabla f_{t-2}$ | −0.00450 | 0.15899 | 0.04478 | 0.08621 | −0.02069 | 0.09898 |
| $t$ | (−0.20) | ( 13.36) | ( 1.85) | ( 8.88) | (−0.87) | ( 8.45) |
| $\nabla f_{t-3}$ | 0.02274 | 0.11911 | 0.07251 | 0.09752 | 0.00365 | 0.08455 |
| $t$ | (0.95) | ( 9.53) | ( 3.08) | ( 10.32) | ( 0.15) | ( 7.02) |
| $\nabla f_{t-4}$ | 0.02429 | 0.08141 | 0.01418 | 0.06827 | −0.02759 | 0.07699 |
| $t$ | (0.99) | ( 6.35) | (0.60) | ( 7.24) | (−1.13) | ( 6.37) |
| $\nabla f_{t-5}$ | 0.00340 | 0.08936 | 0.01185 | 0.04831 | −0.00638 | 0.05004 |
| $t$ | (0.14) | ( 7.10) | (0.51) | ( 5.13) | (−0.26) | ( 4.07) |
| $\nabla f_{t-6}$ | 0.00098 | 0.07291 | 0.01251 | 0.03580 | −0.03941 | 0.02615 |
| $t$ | (0.04) | ( 5.64) | (0.54) | ( 3.84) | (−1.62) | ( 2.18) |
| $\nabla f_{t-7}$ | −0.00372 | 0.05201 | 0.02989 | 0.04837 | −0.02031 | 0.02293 |
| $t$ | (−0.15) | ( 4.01) | ( 1.34) | ( 5.42) | (−0.85) | ( 1.95) |
| $\nabla f_{t-8}$ | 0.00043 | 0.00954 | 0.01812 | 0.02196 | −0.04422 | 0.00462 |
| $t$ | (0.02) | (0.76) | (0.85) | ( 2.57) | (−1.90) | (0.40) |
| $\nabla s_{t-1}$ | −0.08419 | 0.00264 | −0.07618 | −0.05633 | 0.06664 | 0.11143 |
| $t$ | (−2.01) | (0.12) | (−1.70) | (−3.14) | ( 1.49) | ( 5.05) |
| $\nabla s_{t-2}$ | −0.05103 | 0.00256 | −0.10920 | −0.01521 | 0.04099 | −0.01179 |
| $t$ | (−1.18) | ( 0.11) | (−2.59) | (−0.90) | (0.92) | (−0.53) |
| $\nabla s_{t-3}$ | 0.07275 | −0.03631 | −0.00504 | 0.01174 | −0.01948 | −0.01829 |
| $t$ | ( 1.65) | (−1.58) | (−0.12) | (0.71) | (−0.44) | (−0.84) |
| $\nabla s_{t-4}$ | 0.04706 | 0.01438 | 0.02751 | 0.01490 | 0.01646 | 0.00367 |
| $t$ | ( 1.03) | (0.60) | (0.71) | (0.96) | (0.37) | (0.17) |
| $\nabla s_{t-5}$ | 0.08118 | 0.02111 | 0.03943 | 0.02330 | −0.03430 | −0.00462 |
| $t$ | ( 1.77) | ( 0.88) | (0.97) | ( 1.43) | (−0.83) | (−0.23) |
| $\nabla s_{t-6}$ | 0.04390 | 0.04569 | 0.01690 | 0.01919 | 0.06084 | −0.00392 |
| $t$ | (0.96) | ( 1.92) | (0.44) | ( 1.25) | ( 1.45) | (−0.19) |
| $\nabla s_{t-7}$ | −0.03033 | 0.02051 | −0.08647 | 0.00270 | −0.00491 | 0.03597 |
| $t$ | (−0.70) | (0.91) | (−2.09) | (0.16) | (−0.13) | ( 1.90) |
| $\nabla s_{t-8}$ | −0.02920 | 0.03018 | 0.01887 | −0.00213 | 0.00030 | 0.02171 |
| $t$ | (−0.68) | ( 1.34) | (0.49) | (−0.14) | (-0.01) | ( 1.14) |
| $z_{t-1}$ | 0.00024 | 0.00097 | −0.00010 | 0.00012 | 0.00025 | 0.00086 |
| $t$ | ( 1.34) | ( 10.47) | (−0.30) | (0.86) | ( 1.41) | ( 9.75) |

$(c_{i1}, \ldots, c_{ik})'$ be a $k$-dimensional vector, where $i = 1, \ldots, k$. Then

$$y_i = \boldsymbol{c}_i' \boldsymbol{r} = \sum_{j=1}^{k} c_{ij} r_j$$

is a linear combination of the random vector $\boldsymbol{r}$. If $\boldsymbol{r}$ consists of the simple returns of $k$ stocks, then $y_i$ is the return of a portfolio that assigns weight $c_{ij}$ to the $j$th stock. Since multiplying a constant to $\boldsymbol{c}_i$ does not affect the proportion of allocation assigned to the $j$th stock, we standardize the vector $\boldsymbol{c}_i$ so that $\boldsymbol{c}_i' \boldsymbol{c}_i = \sum_{j=1}^{k} c_{ij}^2 = 1$.

Using properties of a linear combination of random variables, we have

$$\text{Var}(y_i) = \boldsymbol{c}_i' \boldsymbol{\Sigma}_r \boldsymbol{c}_i, \qquad i = 1, \ldots, k \tag{8.34}$$

$$\text{Cov}(y_i, y_j) = \boldsymbol{c}_i' \boldsymbol{\Sigma}_r \boldsymbol{c}_j, \quad i, j = 1, \ldots, k. \tag{8.35}$$

The idea of PCA is to find linear combinations $\boldsymbol{c}_i$ such that $y_i$ and $y_j$ are uncorrelated for $i \neq j$ and the variances of $y_i$ are as large as possible. More specifically:

1. the first principal component of $\boldsymbol{r}$ is the linear combination $y_1 = \boldsymbol{c}_1' \boldsymbol{r}$ that maximizes $\text{Var}(y_1)$ subject to the constraint $\boldsymbol{c}_1' \boldsymbol{c}_1 = 1$,
2. the second principal component of $\boldsymbol{r}$ is the linear combination $y_2 = \boldsymbol{c}_2' \boldsymbol{r}$ that maximizes $\text{Var}(y_2)$ subject to the constraints $\boldsymbol{c}_2' \boldsymbol{c}_2 = 1$ and $\text{Cov}(y_2, y_1) = 0$, and
3. the $i$th principal component of $\boldsymbol{r}$ is the linear combination $y_i = \boldsymbol{c}_i' \boldsymbol{r}$ that maximizes $\text{Var}(y_i)$ subject to the constraints $\boldsymbol{c}_i' \boldsymbol{c}_i = 1$ and $\text{Cov}(y_i, y_j) = 0$ for $j = 1, \ldots, i-1$.

Since the covariance matrix $\boldsymbol{\Sigma}_r$ is non-negative definite, it has a spectral decomposition; see Appendix A. Let $(\lambda_1, \boldsymbol{e}_1), \ldots, (\lambda_k, \boldsymbol{e}_k)$ be the eigenvalue-eigenvector pairs of $\boldsymbol{\Sigma}_r$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$. We have the following statistical result.

***Result:*** The $i$th principal component of $\boldsymbol{r}$ is $y_i = \boldsymbol{e}_i' \boldsymbol{r} = \sum_{j=1}^{k} e_{ij} r_j$ for $i = 1, \ldots, k$. Moreover,

$$\text{Var}(y_i) = \boldsymbol{e}_i' \boldsymbol{\Sigma}_r \boldsymbol{e}_i = \lambda_i, \quad i = 1, \ldots, k$$

$$\text{Cov}(y_i, y_j) = \boldsymbol{e}_i' \boldsymbol{\Sigma}_r \boldsymbol{e}_j = 0, \quad i \neq j.$$

If some eigenvalues $\lambda_i$ are equal, the choices of the corresponding eigenvectors $\boldsymbol{e}_i$ and hence $y_i$ are not unique. In addition, we have

$$\sum_{i=1}^{k} \text{Var}(r_i) = tr(\boldsymbol{\Sigma}_r) = \sum_{i=1}^{k} \lambda_i = \sum_{i=1}^{k} \text{Var}(y_i). \tag{8.36}$$

The result of Eq. (8.36) says that

$$\frac{\text{Var}(y_i)}{\sum_{i=1}^{k} \text{Var}(r_i)} = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_k}.$$

Consequently, the proportion of total variance in $\boldsymbol{r}$ explained by the $i$th principal component is simply the ratio between the $i$th eigenvalue and the sum of all eigenvalues of $\boldsymbol{\Sigma}_r$. One can also compute the cumulative proportion of total variance explained by the first $i$ principal components [i.e., $(\sum_{j=1}^{i} \lambda_j)/(\sum_{j=1}^{k} \lambda_j)$]. In practice, one selects a small $i$ such that the prior cumulative proportion is large.

Since $tr(\boldsymbol{\rho}_r) = k$, the proportion of variance explained by the $i$th principal component becomes $\lambda_i/k$ when the correlation matrix is used to perform the PCA.

A byproduct of the PCA is that a zero eigenvalue of $\boldsymbol{\Sigma}_r$, or $\boldsymbol{\rho}_r$, indicates the existence of an *exact* linear relationship between the components of $\boldsymbol{r}$. For instance, if the smallest eigenvalue $\lambda_k = 0$, then by the prior Result $\text{Var}(y_k) = 0$. Therefore, $y_k = \sum_{j=1}^{k} e_{kj} r_j$ is a constant and there are only $k - 1$ random quantities in $\boldsymbol{r}$. In this case, the dimension of $\boldsymbol{r}$ can be reduced. For this reason, PCA has been used in the literature as a tool for dimension reduction.

### 8.7.2 Empirical PCA

In application, the covariance matrix $\boldsymbol{\Sigma}_r$ and the correlation matrix $\boldsymbol{\rho}_r$ of the return vector $\boldsymbol{r}$ are unknown, but they can be estimated consistently by the sample covariance and correlation matrixes under some regularity conditions. Assuming that the returns are weakly stationary and the data consist of $\{\boldsymbol{r}_t \mid t = 1, \ldots, T\}$, we have the following estimates

$$\widehat{\boldsymbol{\Sigma}}_r \equiv [\hat{\sigma}_{ij,r}] = \frac{1}{T-1} \sum_{t=1}^{T} (\boldsymbol{r}_t - \bar{\boldsymbol{r}})(\boldsymbol{r}_t - \bar{\boldsymbol{r}})', \quad \bar{\boldsymbol{r}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{r}_t \tag{8.37}$$

$$\hat{\boldsymbol{\rho}}_r = \widehat{\boldsymbol{D}}^{-1} \widehat{\boldsymbol{\Sigma}}_r \widehat{\boldsymbol{D}}^{-1} \tag{8.38}$$

where $\widehat{\boldsymbol{D}} = \text{diag}\{\sqrt{\hat{\sigma}_{11,r}}, \ldots, \sqrt{\hat{\sigma}_{kk,r}}\}$ is the diagonal matrix of sample standard errors of $\boldsymbol{r}_t$. Methods to compute eigenvalues and eigenvectors of a symmetric matrix can then be used to perform the PCA. Most statistical packages now have the capability to perform principal component analysis.

**Example 8.7.** Consider the monthly log returns of International Business Machines, Hewlett-Packard, Intel Corporation, Merrill Lynch, and Morgan Stanley Dean Witter from January 1990 to December 1999. The returns are in percentages and include dividends. The data set has 120 observations. Figure 8.10 shows the time plots of these five monthly return series. As expected, returns of companies in the same industrial sector tend to exhibit similar patterns.

Denote the returns by $\boldsymbol{r}' = (\text{IBM}, \text{HWP}, \text{INTC}, \text{MER}, \text{MWD})$. The sample mean vector of the returns is $(1.47, 1.97, 3.05, 2.30, 2.36)'$ and the sample covariance and

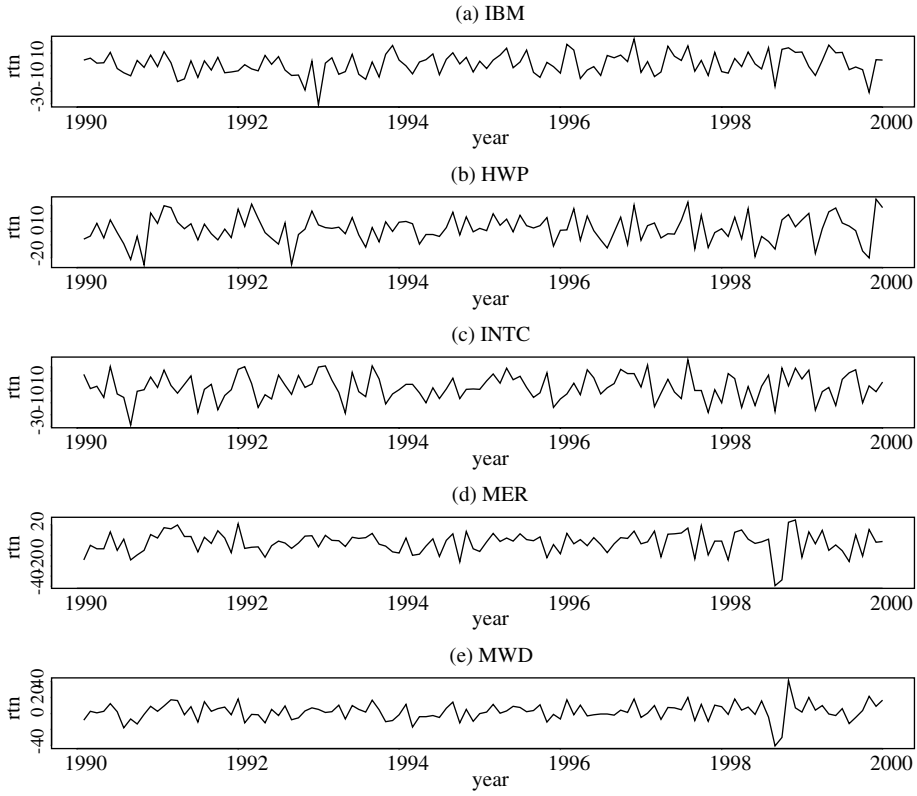


**Figure 8.10.** Time plots of monthly log returns in percentages and including dividends for International Business Machines, Hewlett-Packard, Intel, Merrill Lynch, and Morgan Stanley Dean Witter from January 1990 to December 1999.

correlation matrixes are

$$\widehat{\Sigma}_r = \begin{bmatrix} 73.10 & & & & \\ 36.48 & 103.60 & & & \\ 27.08 & 48.86 & 113.96 & & \\ 16.06 & 37.59 & 27.96 & 105.56 & \\ 16.33 & 40.72 & 26.86 & 85.47 & 109.91 \end{bmatrix},$$

$$\hat{\rho}_r = \begin{bmatrix} 1.00 & & & & \\ 0.42 & 1.00 & & & \\ 0.30 & 0.45 & 1.00 & & \\ 0.18 & 0.36 & 0.26 & 1.00 & \\ 0.18 & 0.38 & 0.24 & 0.79 & 1.00 \end{bmatrix}.$$

Table 8.8 gives the results of PCA using both the covariance and correlation matrixes. Also given are eigenvalues, eigenvectors, and proportions of variabilities

**Table 8.8. Results of Principal Component Analysis for the Monthly Log Returns, Including Dividends, of Stocks of IBM, Hewlett-Packard, Intel, Merrill Lynch, and Morgan Stanley Dean Witter from January 1990 to December 1999. The Eigenvectors Are in Columns.**

| | | | | | |
|---|---|---|---|---|---|
| (a) Using sample covariance matrix | | | | | |
| Eigenvalue | 256.16 | 116.14 | 64.91 | 46.82 | 22.11 |
| Proportion | 0.506 | 0.229 | 0.128 | 0.093 | 0.044 |
| Cumulative | 0.506 | 0.736 | 0.864 | 0.956 | 1.000 |
| Eigenvector | 0.246 | 0.327 | 0.586 | −0.700 | 0.018 |
| | 0.461 | 0.360 | 0.428 | 0.687 | −0.050 |
| | 0.409 | 0.585 | −0.683 | −0.153 | 0.033 |
| | 0.522 | −0.452 | −0.082 | −0.115 | −0.710 |
| | 0.536 | −0.467 | −0.036 | −0.042 | 0.701 |
| (b) Using sample correlation matrix | | | | | |
| Eigenvalue | 2.4563 | 1.1448 | 0.6986 | 0.4950 | 0.2053 |
| Proportion | 0.491 | 0.229 | 0.140 | 0.099 | 0.041 |
| Cumulative | 0.491 | 0.720 | 0.860 | 0.959 | 1.000 |
| Eigenvector | 0.342 | 0.525 | 0.691 | −0.362 | −0.012 |
| | 0.474 | 0.314 | −0.043 | 0.820 | 0.050 |
| | 0.387 | 0.405 | −0.717 | −0.414 | −0.034 |
| | 0.503 | −0.481 | 0.052 | −0.147 | 0.701 |
| | 0.505 | −0.481 | 0.071 | −0.062 | −0.711 |

explained by the principal components. Consider the correlation matrix and denote the sample eigenvalues and eigenvectors by $\hat{\lambda}_i$ and $\widehat{\boldsymbol{e}}_i$. We have

$$\hat{\lambda}_1 = 2.456, \quad \widehat{\boldsymbol{e}}_1 = (0.342, 0.474, 0.387, 0.503, 0.505)'$$
$$\hat{\lambda}_2 = 1.145, \quad \widehat{\boldsymbol{e}}_2 = (0.525, 0.314, 0.405, -0.481, -0.481)'$$

for the first two principal components. These two components explain about 72% of the total variability of the data, and they have interesting interpretations. The first component is a roughly equally weighted linear combination of the stock returns. This component might represent the general movement of the stock market and hence is a *market component*. The second component represents the difference between the two industrial sectors—namely, technologies versus financial services. It might be an *industrial component*. Similar interpretations of principal components can also be found by using the covariance matrix of $\boldsymbol{r}$.

An informal but useful procedure to determine the number of principal components needed in an application is to examine the *scree plot*, which is the time plot of the eigenvalues $\hat{\lambda}_i$ ordered from the largest to the smallest (i.e., a plot of $\hat{\lambda}_i$ vs $i$). Figure 8.11(a) shows the scree plot for the five stock returns of Example 8.7. By looking for an elbow in the scree plot, indicating that the remaining eigenvalues

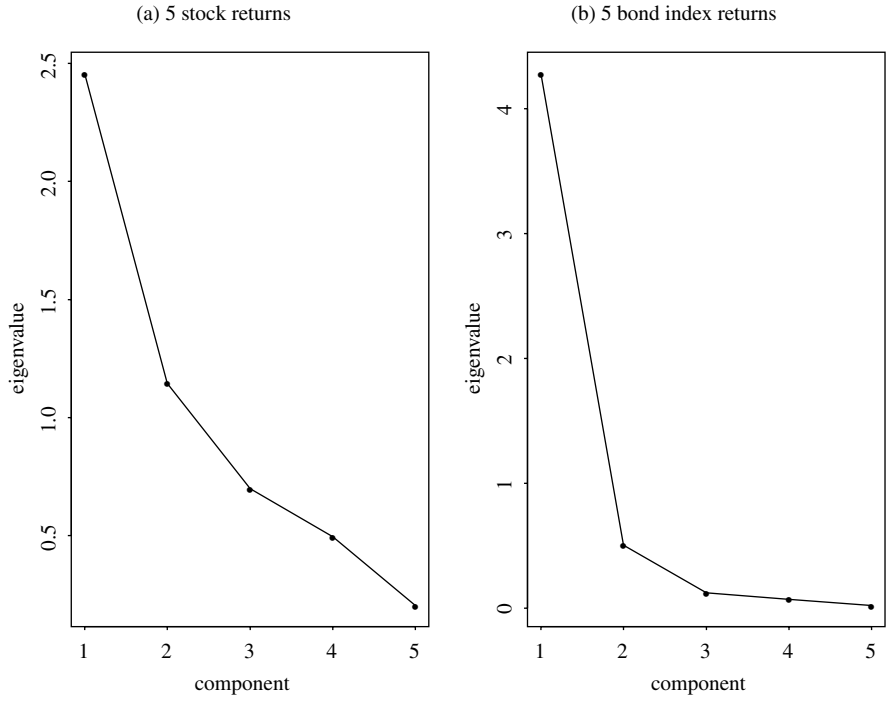


**Figure 8.11.** Scree plots for two 5-dimensional asset returns: (a) series of Example 8.7, and (b) bond index returns of Example 8.9.

are relatively small and all about the same size, one can determine the appropriate number of components. For both plots in Figure 8.11, two components appear to be appropriate. Finally, except for the case in which $\lambda_j = 0$ for $j > i$, selecting the first $i$ principal components only provides an approximation to the total variance of the data. If a small $i$ can provide a good approximation, then the simplification becomes valuable.

## 8.8 FACTOR ANALYSIS

One of the main difficulties in multivariate analysis is the "curse of dimensionality." In particular, the number of parameters of a parametric model often increases dramatically when the order of the model or the dimension of the time series is increased. Simplifying methods are often sought to overcome the difficulty of curse of dimensionality. From an empirical viewpoint, multivariate data often exhibit similar patterns indicating the existence of common structure hidden in the data. Factor analysis is one of those simplifying methods available in the literature. The aim of factor analysis is to identify a few factors that can account for most of the variations in the covariance or correlation matrix of the data.

Traditional factor analysis assumes that the data have no serial correlations. This assumption is often violated by financial data taken with frequency less than or equal to a week. However, the assumption appears to be reasonable for asset returns with lower frequencies (e.g., monthly returns of stocks or market indexes). If the assumption is violated, then one can use the parametric models discussed in this book to remove the linear dynamic dependence of the data and apply factor analysis to the residual series.

In what follows, we discuss factor analysis based on the *orthogonal factor model*. Consider the $k$-dimensional log returns $\boldsymbol{r} = (r_1, \ldots, r_k)'$ and assume that the mean and covariance matrix of $\boldsymbol{r}$ are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. For a return series, this assumption is equivalent to requiring that $\boldsymbol{r}$ is weakly stationary. The factor model postulates that $\boldsymbol{r}$ is linearly dependent on a few *unobservable* random variables $\boldsymbol{F} = (f_1, f_2, \ldots, f_m)'$ and $k$ additional noises $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_k)'$. Here $m < k$, $f_i$ are the common factors, and the $\epsilon_i$ are the errors. Mathematically, the factor model is

$$\begin{aligned} r_1 - \mu_1 &= \ell_{11} f_1 + \cdots + \ell_{1m} f_m + \epsilon_1 \\ r_2 - \mu_2 &= \ell_{21} f_1 + \cdots + \ell_{2m} f_m + \epsilon_2 \\ &\vdots = \vdots \\ r_k - \mu_k &= \ell_{k1} f_1 + \cdots + \ell_{km} f_m + \epsilon_k, \end{aligned}$$

or equivalently in matrix notation

$$\boldsymbol{r} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\epsilon}, \tag{8.39}$$

where $\boldsymbol{L} = [\ell_{ij}]_{k \times m}$ is the *matrix of factor loadings*, $\ell_{ij}$ is the loading of the $i$th variable on the $j$th factor, and $\epsilon_i$ is the *specific error* of $r_i$. A key feature of the prior factor model is that the $m$ factors $f_i$ and the $k$ errors $\epsilon_i$ are *unobservable*. As such, Eq. (8.39) is not a multivariate linear regression model, even though it has a similar appearance.

The factor model in Eq. (8.39) is an orthogonal factor model if it satisfies the following assumptions:

1. $E(\boldsymbol{F}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{F}) = \boldsymbol{I}_m$, the $m \times m$ identity matrix;
2. $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{\Psi} = \text{diag}\{\Psi_1, \ldots, \Psi_k\}$ (i.e., $\boldsymbol{\Psi}$ is a $k \times k$ diagonal matrix); and
3. $\boldsymbol{F}$ and $\boldsymbol{\epsilon}$ are independent so that $\text{Cov}(\boldsymbol{F}, \boldsymbol{\epsilon}) = E(\boldsymbol{F}\boldsymbol{\epsilon}') = \boldsymbol{0}_{m \times k}$.

Under the previous assumptions, it is easy to see that

$$\begin{aligned} \boldsymbol{\Sigma} = \text{Cov}(\boldsymbol{r}) &= E[(\boldsymbol{r} - \boldsymbol{\mu})(\boldsymbol{r} - \boldsymbol{\mu})'] \\ &= E[(\boldsymbol{L}\boldsymbol{F} + \boldsymbol{\epsilon})(\boldsymbol{L}\boldsymbol{F} + \boldsymbol{\epsilon})'] \\ &= \boldsymbol{L}E(\boldsymbol{F}\boldsymbol{F}')\boldsymbol{L}' + E(\boldsymbol{\epsilon}\boldsymbol{F}')\boldsymbol{L}' + \boldsymbol{L}E(\boldsymbol{F}\boldsymbol{\epsilon}') + E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') \\ &= \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi} \end{aligned} \tag{8.40}$$

and

$$\text{Cov}(\boldsymbol{r}, \boldsymbol{F}) = E[(\boldsymbol{r} - \boldsymbol{\mu})\boldsymbol{F}'] = \boldsymbol{L}E(\boldsymbol{F}\boldsymbol{F}') + E(\boldsymbol{\epsilon}\boldsymbol{F}') = \boldsymbol{L}. \tag{8.41}$$

Using Eqs. (8.40) and (8.41), we see that for the orthogonal factor model in Eq. (8.39)

$$\begin{aligned} \text{Var}(r_i) &= \ell_{i1}^2 + \cdots + \ell_{im}^2 + \Psi_i \\ \text{Cov}(r_i, r_j) &= \ell_{i1}\ell_{j1} + \cdots + \ell_{im}\ell_{jm} \\ \text{Cov}(r_i, f_j) &= \ell_{ij}. \end{aligned}$$

The quantity $\ell_{i1}^2 + \cdots + \ell_{im}^2$, which is the portion of the variance of $r_i$ contributed by the $m$ common factors, is called the *communality*. The remaining portion $\Psi_i$ of the variance of $r_i$ is called the *uniqueness* or *specific variance*. Let $c_i^2 = \ell_{i1}^2 + \cdots + \ell_{im}^2$ be the communality, which is the sum of squares of the loadings of the $i$th variable on the $m$ common factors. The variance of component $r_i$ becomes $\text{Var}(r_i) = c_i^2 + \Psi_i$.

In practice, not every covariance matrix has an orthogonal factor representation. In other words, there exists a random variable $\boldsymbol{r}$ that does not have any orthogonal factor representation. Furthermore, the orthogonal factor representation of a random variable is not unique. In fact, for any $m \times m$ orthogonal matrix $\boldsymbol{P}$ satisfying $\boldsymbol{P}\boldsymbol{P}' = \boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}$, let $\boldsymbol{L}^* = \boldsymbol{L}\boldsymbol{P}$ and $\boldsymbol{F}^* = \boldsymbol{P}'\boldsymbol{F}$. Then

$$\boldsymbol{r} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\epsilon} = \boldsymbol{L}\boldsymbol{P}\boldsymbol{P}'\boldsymbol{F} + \boldsymbol{\epsilon} = \boldsymbol{L}^*\boldsymbol{F}^* + \boldsymbol{\epsilon}.$$

In addition, $E(\boldsymbol{F}^*) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{F}^*) = \boldsymbol{P}'\,\text{Cov}(\boldsymbol{F})\boldsymbol{P} = \boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}$. Thus, $\boldsymbol{L}^*$ and $\boldsymbol{F}^*$ form another orthogonal factor model for $\boldsymbol{r}$. This nonuniqueness of orthogonal factor representation is a weakness as well as an advantage for factor analysis. It is a weakness because it makes the meaning of factor loading arbitrary. It is an advantage because it allows us to perform rotations to find common factors that have nice interpretations. Because $\boldsymbol{P}$ is an orthogonal matrix, the transformation $\boldsymbol{F}^* = \boldsymbol{P}'\boldsymbol{F}$ is a rotation in the $m$-dimensional space.

### 8.8.1 Estimation

The orthogonal factor model in Eq. (8.39) can be estimated by two methods. The first estimation method uses the principal component analysis of the previous section. This method does not require the normality assumption of the data nor the prespecification of the number of common factors. It applies to both the covariance and correlation matrixes. But as mentioned in PCA, the solution is often an approximation. The second estimation method is the maximum likelihood method that uses normal density and requires a prespecification for the number of common factors.

***Principal Component Method***

Again let $(\hat{\lambda}_1, \widehat{\boldsymbol{e}}_1), \ldots, (\hat{\lambda}_k, \widehat{\boldsymbol{e}}_k)$ be pairs of the eigenvalues and eigenvectors of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}_r$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_k$. Let $m < k$ be the number

of common factors. Then the matrix of factor loadings is given by

$$\widehat{\boldsymbol{L}} \equiv [\hat{\ell}_{ij}] = \left[\sqrt{\hat{\lambda}_1}\widehat{\boldsymbol{e}}_1 \mid \sqrt{\hat{\lambda}_2}\widehat{\boldsymbol{e}}_2 \mid \cdots \mid \sqrt{\hat{\lambda}_m}\widehat{\boldsymbol{e}}_m\right]. \tag{8.42}$$

The estimated specific variances are the diagonal elements of the matrix $\widehat{\boldsymbol{\Sigma}}_r - \widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}'$. That is, $\widehat{\boldsymbol{\Psi}} = \text{diag}\{\hat{\Psi}_1, \ldots, \hat{\Psi}_k\}$, where $\hat{\Psi}_i = \hat{\sigma}_{ii,r} - \sum_{j=1}^{m} \hat{\ell}_{ij}^2$, where $\hat{\sigma}_{ii,r}$ is the $(i, i)$th element of $\widehat{\boldsymbol{\Sigma}}_r$. The communalities are estimated by

$$\hat{c}_i^2 = \hat{\ell}_{i1}^2 + \cdots + \hat{\ell}_{im}^2.$$

The error matrix due to approximation is

$$\widehat{\boldsymbol{\Sigma}}_r - (\widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}' + \widehat{\boldsymbol{\Psi}}).$$

Ideally, we would like this matrix to be close to zero. It can be shown that the sum of squared elements of $\widehat{\boldsymbol{\Sigma}}_r - (\widehat{\boldsymbol{L}}\widehat{\boldsymbol{L}}' + \widehat{\boldsymbol{\Psi}})$ is less than or equal to $\hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_k^2$. Therefore, the approximation error is bounded by the sum of squares of the neglected eigenvalues.

From the solution in Eq. (8.42), the estimated factor loadings based on the principal component method do not change as the number of common factors $m$ is increased.

***Maximum Likelihood Method***

If the common factors $\boldsymbol{F}$ and the specific factors $\boldsymbol{\epsilon}$ are jointly normal, then $\boldsymbol{r}$ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}_r = \boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi}$. The maximum likelihood method can then be used to obtain estimates of $\boldsymbol{L}$ and $\boldsymbol{\Psi}$ under the constraint $\boldsymbol{L}'\boldsymbol{\Psi}^{-1}\boldsymbol{L} = \boldsymbol{\Delta}$, which is a diagonal matrix. Here $\boldsymbol{\mu}$ is estimated by the sample mean. For more details on this method, readers are referred to Johnson and Wichern (1998). In using the maximum likelihood method, the number of common factors must be given *a priori*.

### 8.8.2 Factor Rotation

As mentioned before, for any $m \times m$ orthogonal matrix $\boldsymbol{P}$,

$$\boldsymbol{r} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\epsilon} = \boldsymbol{L}^*\boldsymbol{F}^* + \boldsymbol{\epsilon},$$

where $\boldsymbol{L}^* = \boldsymbol{L}\boldsymbol{P}$ and $\boldsymbol{F}^* = \boldsymbol{P}'\boldsymbol{F}$. In addition,

$$\boldsymbol{L}\boldsymbol{L}' + \boldsymbol{\Psi} = \boldsymbol{L}\boldsymbol{P}\boldsymbol{P}'\boldsymbol{L}' + \boldsymbol{\Psi} = \boldsymbol{L}^*(\boldsymbol{L}^*)' + \boldsymbol{\Psi}.$$

This result states that the communalities and the specific variances remain unchanged under an orthogonal transformation. It is then reasonable to find an orthogonal

matrix $\boldsymbol{P}$ to transform the factor model so that the common factors have nice interpretations. Such a transformation is equivalent to rotating the common factors in the $m$-dimensional space. In fact, there are infinite possible factor rotations available. Kaiser (1958) proposes a *varimax* criterion to select the rotation that works well in many applications. Denote the rotated matrix of factor loadings by $\boldsymbol{L}^* = [\ell_{ij}^*]$ and the $i$th communality by $c_i^2$. Define $\tilde{\ell}_{ij}^* = \ell_{ij}^*/c_i$ to be the rotated coefficients scaled by the (positive) square root of communalities. The varimax procedure selects the orthogonal matrix $\boldsymbol{P}$ that maximizes the quantity

$$V = \frac{1}{k}\sum_{j=1}^{m}\left[\sum_{i=1}^{k}(\tilde{\ell}_{ij}^*)^4 - \frac{1}{k}\left(\sum_{i=1}^{k}\tilde{\ell}_{ij}^{*2}\right)^2\right].$$

This complicated expression has a simple interpretation. Maximizing $V$ corresponds to spreading out the squares of the loadings on each factor as much as possible. Consequently, the procedure is to find groups of large and negligible coefficients in any column of the rotated matrix of factor loadings. In a real application, factor rotation is used to aid the interpretations of common factors. It may be helpful in some applications, but not so in others.

### 8.8.3 Applications

Given the data $\{\boldsymbol{r}_t\}$ of asset returns, the factor analysis enables us to search for common factors that explain the variabilities of the returns. Since factor analysis assumes no serial correlations in the data, one should check the validity of this assumption before using factor analysis. The multivariate Portmanteau statistics can be used for this purpose. If serial correlations are found, one can build a VARMA model to remove the dynamic dependence in the data and apply the factor analysis to the residual series. For many returns series, the correlation matrix of the residuals of a linear model is often very close to the correlation matrix of the original data. In this case, the effect of dynamic dependence on factor analysis is negligible.

**Example 8.8.** Consider again the monthly log stock returns of IBM, Hewlett-Parkard, Intel, Merrill Lynch, and Morgan Stanley Dean Witter used in Example 8.7. To check the assumption of no serial correlations, we compute the Portmanteau statistics and obtain $Q_5(1) = 34.28$, $Q_5(4) = 114.30$, and $Q_5(8) = 216.78$. Compared with chi-squared distributions with 25, 100, and 200 degrees of freedom, the $p$ values of these test statistics are 0.102, 0.156, and 0.198, respectively. Therefore, the assumption of no serial correlations cannot be rejected even at the 10% level.

Table 8.9 shows the results of factor analysis based on the correlation matrix using both the principal component and maximum likelihood methods. We assume that the number of common factors is 2, which is reasonable according to the principal component analysis of Example 8.7. From the table, the factor analysis reveals several interesting findings:

**Table 8.9. Factor Analysis of the Monthly Log Stock Returns of IBM, Hewlett-Packard, Intel, Merrill Lynch, and Morgan Stanley Dean Witter. The Returns Include Dividends and Are from January 1990 to December 1999. The Analysis Is Based on the Sample Cross-Correlation Matrix and Assumes Two Common Factors.**

(a) The principal component method

| Variable | Estimates of factor loadings $f_1$ | $f_2$ | Rotated factor loadings $f_1^*$ | $f_2^*$ | Communalities $1-\Psi_i$ |
|---|---|---|---|---|---|
| IBM | 0.536 | 0.561 | 0.011 | 0.776 | 0.602 |
| HWP | 0.744 | 0.335 | 0.317 | 0.752 | 0.665 |
| INTC | 0.607 | 0.433 | 0.151 | 0.730 | 0.556 |
| MER | 0.788 | −0.515 | 0.928 | 0.158 | 0.887 |
| MWD | 0.791 | −0.514 | 0.930 | 0.161 | 0.891 |
| Variance | 2.4563 | 1.1448 | 1.8502 | 1.7509 | 3.6011 |
| Proportion | 0.491 | 0.229 | 0.370 | 0.350 | 0.720 |

(b) The maximum likelihood method

| Variable | Estimates of factor loadings $f_1$ | $f_2$ | Rotated factor loadings $f_1^*$ | $f_2^*$ | Communalities $1-\Psi_i$ |
|---|---|---|---|---|---|
| IBM | 0.191 | 0.496 | 0.087 | 0.524 | 0.282 |
| HWP | 0.394 | 0.689 | 0.247 | 0.755 | 0.630 |
| INTC | 0.250 | 0.511 | 0.141 | 0.551 | 0.323 |
| MER | 0.800 | 0.072 | 0.769 | 0.232 | 0.645 |
| MWD | 0.994 | −0.015 | 0.976 | 0.186 | 0.988 |
| Variance | 1.8813 | 0.9866 | 1.6324 | 1.2355 | 2.8679 |
| Proportion | 0.376 | 0.197 | 0.326 | 0.247 | 0.574 |

- The two factors identified by the principal component method explain more variability than those identified by the maximum likelihood method.
- Based on the rotated factor loadings, the two estimation methods identify essentially the same two common factors for the data. The financial stocks (Merrill Lynch and Morgan Stanley Dean Witter) load heavily on the first factor, whereas the technology stocks (IBM, Hewlett-Packard, and Intel) load highly on the second factor. These two rotated factors jointly differentiate the industrial sectors.
- In this particular instance, the varimax rotation does not change much the two factors identified by the maximum likelihood method. Yet the first unrotated factor identified by the principal component method was destroyed by the rotation. This is not surprising in view of the idea behind the varimax criterion.
- The specific variances of IBM and Intel stock returns are relatively large based on the maximum likelihood method, indicating that these two stocks have their own features that are worth further investigation.

**Example 8.9.** In this example, we consider the monthly log returns of U.S. bond indexes with maturities in 30 years, 20 years, 10 years, 5 years, and 1 year. The data are described in Example 8.2, but have been transformed into log returns. There are 696 observations. As shown in Example 8.2, there is serial dependence in the data. However, removing serial dependence by fitting a VARMA(2, 1) model has hardly any effects on the concurrent correlation matrix. As a matter of fact, the correlation matrixes before and after fitting a VARMA(2, 1) model are

$$\hat{\rho}_o = \begin{bmatrix} 1.0 & & & & \\ .98 & 1.0 & & & \\ .92 & .91 & 1.0 & & \\ .85 & .86 & .90 & 1.0 & \\ .63 & .64 & .67 & .81 & 1.0 \end{bmatrix}, \quad \hat{\rho} = \begin{bmatrix} 1.0 & & & & \\ .98 & 1.0 & & & \\ .92 & .92 & 1.0 & & \\ .85 & .86 & .90 & 1.0 & \\ .66 & .67 & .71 & .84 & 1.0 \end{bmatrix},$$

**Table 8.10. Factor Analysis of the Monthly Log Returns of U.S. Bond Indexes With Maturities in 30 Years, 20 Years, 10 Years, 5 Years, and 1 Year. The Data Are from January 1942 to December 1999. The Analysis Is Based on the Sample Cross-Correlation Matrix and Assumes Two Common Factors.**

(a) The principal component method

| Variable | Estimates of factor loadings $f_1$ | $f_2$ | Rotated factor loadings $f_1^*$ | $f_2^*$ | Communalities $1 - \Psi_i$ |
|---|---|---|---|---|---|
| 30 years | 0.952 | 0.253 | 0.927 | 0.333 | 0.970 |
| 20 years | 0.954 | 0.240 | 0.922 | 0.345 | 0.968 |
| 10 years | 0.956 | 0.140 | 0.866 | 0.429 | 0.934 |
| 5 years | 0.955 | −0.142 | 0.704 | 0.660 | 0.931 |
| 1 year | 0.800 | −0.585 | 0.325 | 0.936 | 0.982 |
| Variance | 4.2812 | 0.5039 | 3.0594 | 1.7256 | 4.7851 |
| Proportion | 0.856 | 0.101 | 0.612 | 0.345 | 0.957 |

(b) The maximum likelihood method

| Variable | Estimates of factor loadings $f_1$ | $f_2$ | Rotated factor loadings $f_1^*$ | $f_2^*$ | Communalities $1 - \Psi_i$ |
|---|---|---|---|---|---|
| 30 years | 0.849 | −0.513 | 0.895 | 0.430 | 0.985 |
| 20 years | 0.857 | −0.486 | 0.876 | 0.451 | 0.970 |
| 10 years | 0.896 | −0.303 | 0.744 | 0.584 | 0.895 |
| 5 years | 1.000 | 0.000 | 0.547 | 0.837 | 1.000 |
| 1 year | 0.813 | 0.123 | 0.342 | 0.747 | 0.675 |
| Variance | 3.91783 | 0.6074 | 2.5378 | 1.9874 | 4.5252 |
| Proportion | 0.784 | 0.121 | 0.508 | 0.397 | 0.905 |

where $\hat{\boldsymbol{\rho}}_o$ is the correlation matrix of the original log returns. Therefore, we apply factor analysis directly to the return series.

Table 8.10 shows the results of factor analysis of the data. For both estimation methods, the first two common factors explain more than 90% of the total variability of the data. Indeed, the high communalities indicate that the specific variances are very small for the five bond index returns. Because the results of the two methods are close, we only discuss that of the principal component method. The unrotated factor loadings indicate that (a) all five return series load roughly equally on the first factor, and (b) the loadings on the second factor are positively correlated with the time to maturity. Therefore, the first common factor represents the general U.S. bond returns, and the second factor shows the "time-to-maturity" effect. Furthermore, the loadings of the second factor sum approximately to zero. Therefore, this common factor can also be interpreted as the contrast between long-term and short-term bonds. Here a long-term bond means one with maturity 10 years or longer. For the rotated factors, the loadings are also interesting. The loadings for the first rotated factor are proportional to the time to maturity, whereas the loadings of the second factor are inversely proportional to the time to maturity.

***Remark:*** The factor analyses of this section are carried out using the Minitab computer program.

## APPENDIX A. REVIEW OF VECTORS AND MATRIXES

In this appendix, we briefly review some algebra and properties of vectors and matrixes. No proofs are given as they can be found in standard textbooks on matrixes (e.g., Graybill, 1969).

A $m \times n$ real-valued matrix is an $m$ by $n$ array of real numbers. For example,

$$\boldsymbol{A} = \begin{bmatrix} 2 & 5 & 8 \\ -1 & 3 & 4 \end{bmatrix}$$

is a $2 \times 3$ matrix. This matrix has two rows and three columns. In general, an $m \times n$ matrix is written as

$$\boldsymbol{A} \equiv [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1,n-1} & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2,n-1} & a_{2n} \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{m,n-1} & a_{mn} \end{bmatrix}. \tag{8.43}$$

The positive integers $m$ and $n$ are the *row dimension* and *column dimension* of $\boldsymbol{A}$. The real number $a_{ij}$ is referred to as the $(i, j)$th element of $\boldsymbol{A}$. In particular, the elements $a_{ii}$ are the *diagonal elements* of the matrix.

A $m \times 1$ matrix forms a $m$-dimensional column vector, and a $1 \times n$ matrix is an $n$-dimensional row vector. In the literature, a vector is often meant to be a column

vector. If $m = n$, then the matrix is a square matrix. If $a_{ij} = 0$ for $i \neq j$ and $m = n$, then the matrix $\boldsymbol{A}$ is a *diagonal matrix*. If $a_{ij} = 0$ for $i \neq j$ and $a_{ii} = 1$ for all $i$, then $\boldsymbol{A}$ is the $m \times m$ *identity matrix*, which is commonly denoted by $\boldsymbol{I}_m$ or simply $\boldsymbol{I}$ if the dimension is clear.

The $n \times m$ matrix

$$\boldsymbol{A}' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m-1,1} & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m-1,2} & a_{m2} \\ \vdots & \vdots & & \vdots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{m-1,n} & a_{mn} \end{bmatrix}$$

is the *transpose* of the matrix $\boldsymbol{A}$. For example,

$$\begin{bmatrix} 2 & -1 \\ 5 & 3 \\ 8 & 4 \end{bmatrix} \quad \text{is the transpose of} \quad \begin{bmatrix} 2 & 5 & 8 \\ -1 & 3 & 4 \end{bmatrix}.$$

We use the notation $\boldsymbol{A}' = [a'_{ij}]$ to denote the transpose of $\boldsymbol{A}$. From the definition, $a'_{ij} = a_{ji}$ and $(\boldsymbol{A}')' = \boldsymbol{A}$. If $\boldsymbol{A}' = \boldsymbol{A}$, then $\boldsymbol{A}$ is a *symmetric matrix*.

### Basic Operations

Suppose that $\boldsymbol{A} = [a_{ij}]_{m \times n}$ and $\boldsymbol{C} = [c_{ij}]_{p \times q}$ are two matrixes with dimensions given in the subscript. Let $b$ be a real number. Some basic matrix operations are defined next:

- Addition: $\boldsymbol{A} + \boldsymbol{C} = [a_{ij} + c_{ij}]_{m \times n}$ if $m = p$ and $n = q$;
- Subtraction: $\boldsymbol{A} - \boldsymbol{C} = [a_{ij} - c_{ij}]_{m \times n}$ if $m = p$ and $n = q$;
- Scalar multiplication: $b\boldsymbol{A} = [ba_{ij}]_{m \times n}$; and
- Multiplication: $\boldsymbol{AC} = [\sum_{v=1}^{n} a_{iv} c_{vj}]_{m \times q}$ provided that $n = p$.

When the dimensions of matrixes satisfy the condition for multiplication to take place, the two matrixes are said to be *conformable*. An example of matrix multiplication is

$$\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ -1 & 2 & -4 \end{bmatrix} = \begin{bmatrix} 2 \cdot 1 - 1 \cdot 1 & 2 \cdot 2 + 1 \cdot 2 & 2 \cdot 3 - 1 \cdot 4 \\ 1 \cdot 1 - 1 \cdot 1 & 1 \cdot 2 + 1 \cdot 2 & 1 \cdot 3 - 1 \cdot 4 \end{bmatrix}$$
$$= \begin{bmatrix} 1 & 6 & 2 \\ 0 & 4 & -1 \end{bmatrix}.$$

Important rules of matrix operations include (a) $(\boldsymbol{AC})' = \boldsymbol{C}'\boldsymbol{A}'$ and (b) $\boldsymbol{AC} \neq \boldsymbol{CA}$ in general.

### Inverse, Trace, Eigenvalue, and Eigenvector

A square matrix $\boldsymbol{A}_{m\times m}$ is *nonsingular* or *invertible* if there exists a unique matrix $\boldsymbol{C}_{m\times m}$ such that $\boldsymbol{AC} = \boldsymbol{CA} = \boldsymbol{I}_m$, the $m \times m$ identity matrix. In this case, $\boldsymbol{C}$ is called the *inverse* matrix of $\boldsymbol{A}$ and is denoted by $\boldsymbol{C} = \boldsymbol{A}^{-1}$.

The trace of $\boldsymbol{A}_{m\times m}$ is the sum of its diagonal elements [i.e., $tr(\boldsymbol{A}) = \sum_{i=1}^{m} a_{ii}$]. It is easy to see that (a) $tr(\boldsymbol{A}+\boldsymbol{C}) = tr(\boldsymbol{A})+tr(\boldsymbol{C})$, (b) $tr(\boldsymbol{A}) = tr(\boldsymbol{A}')$, and (c) $tr(\boldsymbol{AC}) = tr(\boldsymbol{CA})$ provided that the two matrixes are conformable.

A number $\lambda$ and a $m \times 1$ vector $\boldsymbol{b}$, possibly complex-valued, are a right *eigenvalue* and *eigenvector* pair of the matrix $\boldsymbol{A}$ if $\boldsymbol{Ab} = \lambda\boldsymbol{b}$. There are $m$ possible eigenvalues for the matrix $\boldsymbol{A}$. For a real-valued matrix $\boldsymbol{A}$, complex eigenvalues occur in conjugated pairs. The matrix $\boldsymbol{A}$ is nonsingular if and only if all of its eigenvalues are nonzero. Denote the eigenvalues by $\{\lambda_i \mid i = 1, \ldots, m\}$, we have $tr(\boldsymbol{A}) = \sum_{i=1}^{m} \lambda_i$. In addition, the *determinant* of the matrix $\boldsymbol{A}$ can be defined as $|\boldsymbol{A}| = \prod_{i=1}^{m} \lambda_i$. For a general definition of determinant of a matrix, see a standard textbook on matrix (e.g., Graybill, 1969).

Finally, the rank of the matrix $\boldsymbol{A}_{m\times n}$ is the number of nonzero eigenvalues of the symmetric matrix $\boldsymbol{AA}'$. Also, for a nonsingular matrix $\boldsymbol{A}$, $(\boldsymbol{A}^{-1})' = (\boldsymbol{A}')^{-1}$.

### Positive Definite Matrix

A square matrix $\boldsymbol{A}$ ($m \times m$) is a *positive definite* matrix if (a) $\boldsymbol{A}$ is symmetric, and (b) all eigenvalues of $\boldsymbol{A}$ are positive. Alternatively, $\boldsymbol{A}$ is a positive definite matrix if for any nonzero $m$-dimensional vector $\boldsymbol{b}$, we have $\boldsymbol{b}'\boldsymbol{Ab} > 0$.

Useful properties of a positive definite matrix $\boldsymbol{A}$ include (a) all eigenvalues of $\boldsymbol{A}$ are real and positive, and (b) the matrix can be decomposed as

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}',$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix consisting of all eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{P}$ is a $m \times m$ matrix consisting of the $m$ right eigenvectors of $\boldsymbol{A}$. It is common to write the eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ and the eigenvectors as $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m$ such that $\boldsymbol{Ae}_i = \lambda_i\boldsymbol{e}_i$ and $\boldsymbol{e}_i'\boldsymbol{e}_i = 1$. In addition, these eigenvectors are orthogonal to each other—namely, $\boldsymbol{e}_i'\boldsymbol{e}_j = 0$ if $i \neq j$—if the eigenvalues are distinct. The matrix $\boldsymbol{P}$ is an *orthogonal* matrix and the decomposition is referred to as the *spectral decomposition* of the matrix $\boldsymbol{A}$. Consider, for example, the simple $2 \times 2$ matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix},$$

which is positive definite. Simple calculations show that

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = 3\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Therefore, 3 and 1 are eigenvalues of $\boldsymbol{\Sigma}$ with normalized eigenvectors $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})'$ and $(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})'$, respectively. It is easy to verify that the spectral decomposition holds—that is,

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}.$$

For a symmetric matrix $\boldsymbol{A}$, there exists a lower triangular matrix $\boldsymbol{L}$ with diagonal elements being 1 and a diagonal matrix $\boldsymbol{G}$ such that $\boldsymbol{A} = \boldsymbol{LGL}'$; see Chapter 1 of Strang (1980). If $\boldsymbol{A}$ is positive definite, then the diagonal elements of $\boldsymbol{G}$ are positive. In this case,

$$\boldsymbol{A} = \boldsymbol{L}\sqrt{\boldsymbol{G}}\sqrt{\boldsymbol{G}}\boldsymbol{L}' = (\boldsymbol{L}\sqrt{\boldsymbol{G}})(\boldsymbol{L}\sqrt{\boldsymbol{G}})',$$

where $\boldsymbol{L}\sqrt{\boldsymbol{G}}$ is again a lower triangular matrix and the square root is taking element by element. Such a decomposition is called the *Cholesky Decomposition* of $\boldsymbol{A}$. This decomposition shows that a positive definite matrix $\boldsymbol{A}$ can be diagonalized as

$$\boldsymbol{L}^{-1}\boldsymbol{A}(\boldsymbol{L}')^{-1} = \boldsymbol{L}^{-1}\boldsymbol{A}(\boldsymbol{L}^{-1})' = \boldsymbol{G}.$$

Since $\boldsymbol{L}$ is a lower triangular matrix with unit diagonal elements, $\boldsymbol{L}^{-1}$ is also lower triangular matrix with unit diagonal elements. Consider again the prior $2 \times 2$ matrix $\boldsymbol{\Sigma}$. It is easy to verify that

$$\boldsymbol{L} = \begin{bmatrix} 1.0 & 0.0 \\ 0.5 & 1.0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{G} = \begin{bmatrix} 2.0 & 0.0 \\ 0.0 & 1.5 \end{bmatrix}$$

satisfy that $\boldsymbol{\Sigma} = \boldsymbol{LGL}'$. In addition,

$$\boldsymbol{L}^{-1} = \begin{bmatrix} 1.0 & 0.0 \\ -0.5 & 1.0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{L}^{-1}\boldsymbol{\Sigma}(\boldsymbol{L}^{-1})' = \boldsymbol{G}.$$

**Vectorization and Kronecker Product**

Writing a $m \times n$ matrix $\boldsymbol{A}$ in its columns as $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$, we define the stacking operation as $\text{vec}(\boldsymbol{A}) = (\boldsymbol{a}_1', \boldsymbol{a}_2', \ldots, \boldsymbol{a}_m')'$, which is a $mn \times 1$ vector. For two matrixes $\boldsymbol{A}_{m\times n}$ and $\boldsymbol{C}_{p\times q}$, the Kronecker product between $\boldsymbol{A}$ and $\boldsymbol{C}$ is

$$\boldsymbol{A} \otimes \boldsymbol{C} = \begin{bmatrix} a_{11}\boldsymbol{C} & a_{12}\boldsymbol{C} & \cdots & a_{1n}\boldsymbol{C} \\ a_{21}\boldsymbol{C} & a_{22}\boldsymbol{C} & \cdots & a_{2n}\boldsymbol{C} \\ \vdots & \vdots & & \vdots \\ a_{m1}\boldsymbol{C} & a_{m2}\boldsymbol{C} & \cdots & a_{mn}\boldsymbol{C} \end{bmatrix}_{mp\times nq}.$$

For example, assume that

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ -1 & 3 \end{bmatrix}, \quad \boldsymbol{C} = \begin{bmatrix} 4 & -1 & 3 \\ -2 & 5 & 2 \end{bmatrix}.$$

Then $\text{vec}(\boldsymbol{A}) = (2, -1, 1, 3)'$, $\text{vec}(\boldsymbol{C}) = (4, -2, -1, 5, 3, 2)'$, and

$$\boldsymbol{A} \otimes \boldsymbol{C} = \begin{bmatrix} 8 & -2 & 6 & 4 & -1 & 3 \\ -4 & 10 & 4 & -2 & 5 & 2 \\ -4 & 1 & -3 & 12 & -3 & 9 \\ 2 & -5 & -2 & -6 & 15 & 6 \end{bmatrix}.$$

Assuming that the dimensions are appropriate, then we have the following useful properties for the two operators:

1. $\boldsymbol{A} \otimes \boldsymbol{C} \neq \boldsymbol{C} \otimes \boldsymbol{A}$ in general;
2. $(\boldsymbol{A} \otimes \boldsymbol{C})' = \boldsymbol{A}' \otimes \boldsymbol{C}'$;
3. $\boldsymbol{A} \otimes (\boldsymbol{C} + \boldsymbol{D}) = \boldsymbol{A} \otimes \boldsymbol{C} + \boldsymbol{A} \otimes \boldsymbol{D}$;
4. $(\boldsymbol{A} \otimes \boldsymbol{C})(\boldsymbol{F} \otimes \boldsymbol{G}) = (\boldsymbol{AF}) \otimes (\boldsymbol{CG})$;
5. If $\boldsymbol{A}$ and $\boldsymbol{C}$ are invertible, then $(\boldsymbol{A} \otimes \boldsymbol{C})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{C}^{-1}$;
6. For square matrixes $\boldsymbol{A}$ and $\boldsymbol{C}$, $tr(\boldsymbol{A} \otimes \boldsymbol{C}) = tr(\boldsymbol{A})tr(\boldsymbol{C})$;
7. $\text{vec}(\boldsymbol{A} + \boldsymbol{C}) = \text{vec}(\boldsymbol{A}) + \text{vec}(\boldsymbol{C})$;
8. $\text{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}' \otimes \boldsymbol{A})\text{vec}(\boldsymbol{B})$;
9. $tr(\boldsymbol{AC}) = \text{vec}(\boldsymbol{C'})'\text{vec}(\boldsymbol{A}) = \text{vec}(\boldsymbol{A'})'\text{vec}(\boldsymbol{C})$;
10. and

$$\begin{aligned} tr(\boldsymbol{ABC}) &= \text{vec}(\boldsymbol{A}')'(\boldsymbol{C}' \otimes \boldsymbol{I})\text{vec}(\boldsymbol{B}) = \text{vec}(\boldsymbol{A}')'(\boldsymbol{I} \otimes \boldsymbol{B})\text{vec}(\boldsymbol{C}) \\ &= \text{vec}(\boldsymbol{B}')'(\boldsymbol{A}' \otimes \boldsymbol{I})\text{vec}(\boldsymbol{C}) = \text{vec}(\boldsymbol{B}')'(\boldsymbol{I} \otimes \boldsymbol{C})\text{vec}(\boldsymbol{A}) \\ &= \text{vec}(\boldsymbol{C}')'(\boldsymbol{B}' \otimes \boldsymbol{I})\text{vec}(\boldsymbol{A}) = \text{vec}(\boldsymbol{C}')'(\boldsymbol{I} \otimes \boldsymbol{A})\text{vec}(\boldsymbol{B}). \end{aligned}$$

In multivariate statistical analysis, we often deal with symmetric matrixes. It is therefore convenient to generalize the stacking operation to the *half-stacking* operation, which consists of elements on or below the main diagonal. Specifically, for a symmetric square matrix $\boldsymbol{A} = [a_{ij}]_{k \times k}$, define

$$\text{vech}(\boldsymbol{A}) = (\boldsymbol{a}'_{1.}, \boldsymbol{a}'_{2*}, \ldots, \boldsymbol{a}'_{k*})',$$

where $\boldsymbol{a}_{1.}$ is the first column of $\boldsymbol{A}$, and $\boldsymbol{a}_{i*} = (a_{ii}, a_{i+1,i}, \ldots, a_{ki})'$ is a $(k - i + 1)$-dimensional vector. The dimension of $\text{vech}(\boldsymbol{A})$ is $k(k + 1)/2$. For example, suppose that $k = 3$. Then we have $\text{vech}(\boldsymbol{A}) = (a_{11}, a_{21}, a_{31}, a_{22}, a_{32}, a_{33})'$, which is a six-dimensional vector.

## APPENDIX B. MULTIVARIATE NORMAL DISTRIBUTIONS

A $k$-dimensional random vector $\boldsymbol{x} = (x_1, \ldots, x_k)'$ follows a multivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ and positive definite covariance matrix $\boldsymbol{\Sigma} = [\sigma_{ij}]$ if its probability density function (pdf) is

$$f(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right]. \tag{8.44}$$

We use the notation $\boldsymbol{x} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote that $\boldsymbol{x}$ follows such a distribution. This normal distribution plays an important role in multivariate statistical analysis and it has several nice properties. Here we only consider those properties that are relevant to our study. Interested readers are referred to Johnson and Wichern (1998) for details.

To gain insight into multivariate normal distributions, consider the bivariate case (i.e., $k = 2$). In this case, we have

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}, \quad \boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}.$$

Using the correlation coefficient $\rho = \sigma_{12}/(\sigma_1\sigma_2)$, where $\sigma_i = \sqrt{\sigma_{ii}}$ is the standard deviation of $x_i$, we have $\sigma_{12} = \rho\sqrt{\sigma_{11}\sigma_{22}}$ and $\mid \boldsymbol{\Sigma} \mid = \sigma_{11}\sigma_{22}(1 - \rho^2)$. The pdf of $\boldsymbol{x}$ then becomes

$$f(x_1, x_2 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}} \exp\left[-\frac{1}{2(1 - \rho^2)}[Q(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})]\right],$$

where

$$Q(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1 - \mu_1}{\sigma_1}\right)\left(\frac{x_2 - \mu_2}{\sigma_2}\right).$$

Chapter 4 of Johnson and Wichern (1998) contains some plots of this pdf function.

Let $\boldsymbol{c} = (c_1, \ldots, c_k)'$ be a nonzero $k$-dimensional vector. Partition the random vector as $\boldsymbol{x} = (\boldsymbol{x}_1', \boldsymbol{x}_2')'$, where $\boldsymbol{x}_1 = (x_1, \ldots, x_p)'$ and $\boldsymbol{x}_2 = (x_{p+1}, \ldots, x_k)'$ with $1 \leq p < k$. Also partition $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ accordingly as

$$\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right).$$

Some properties of $\boldsymbol{x}$ are as follows:

1. $\boldsymbol{c}'\boldsymbol{x} \sim N(\boldsymbol{c}'\boldsymbol{\mu}, \boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c})$. That is, any nonzero linear combination of $\boldsymbol{x}$ is univariate normal. The inverse of this property also holds. Specifically, if $\boldsymbol{c}'\boldsymbol{x}$ is univariate normal for any nonzero vector $\boldsymbol{c}$, then $\boldsymbol{x}$ is multivariate normal.
2. The marginal distribution of $\boldsymbol{x}_i$ is normal. In fact, $\boldsymbol{x}_i \sim N_{k_i}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii})$ for $i = 1$ and 2, where $k_1 = p$ and $k_2 = k - p$.

3. $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ if and only if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are independent.
4. The random variable $y = (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$ follows a chi-squared distribution with $m$ degrees of freedom.
5. The conditional distribution of $\boldsymbol{x}_1$ given $\boldsymbol{x}_2 = \boldsymbol{b}$ is also normally distributed as

$$(\boldsymbol{x}_1 \mid \boldsymbol{x}_2 = \boldsymbol{b}) \sim N_p[\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{b} - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}].$$

The last property is useful in many scientific areas. For instance, it forms the basis for time series forecasting under the normality assumption and for recursive least squares estimation.

## EXERCISES

1. Consider the monthly log stock returns, in percentages and including dividends, of Merck & Company, Johnson & Johnson, General Electric, General Motors, Ford Motor Company, and value-weighted index from January 1960 to December 1999; see the file "m-mrk2vw.dat," which has six columns in the order listed before.
   - **(a)** Compute the sample mean, covariance matrix, and correlation matrix of the data.
   - **(b)** Test the hypothesis $H_0 : \boldsymbol{\rho}_1 = \cdots = \boldsymbol{\rho}_6 = \mathbf{0}$, where $\boldsymbol{\rho}_i$ is the lag-$i$ cross-correlation matrix of the data. Draw conclusion based on the 5% significance level.
   - **(c)** Is there any lead-lag relationship among the six return series?
   - **(d)** Perform a principal component analysis of the data using the sample covariance matrix.
   - **(e)** Perform a principal component analysis of the data using the sample correlation matrix.
   - **(f)** Perform a factor analysis on the data. Identify the number of common factors. Obtain estimates of factor loadings using both the principal component and maximum likelihood methods.
2. The Federal Reserve Bank of St Louis publishes selected interest rates and U.S. financial data on its Web site:

   http://www.stls.frb.org/fred/index.html

   Consider the monthly 1-year and 10-year Treasury constant maturity rates from April 1953 to October 2000 for 571 observations; see the file "m-gs1n10.dat." The rates are in percentages.
   - **(a)** Let $c_t = r_t - r_{t-1}$ be the change series of the monthly interest rate $r_t$. Build a bivariate autoregressive model for the two change series. Discuss the implications of the model. Transform the model into a structural form.

**(b)** Build a bivariate moving-average model for the two change series. Discuss the implications of the model and compare it with the bivariate AR model built earlier.

**(c)** Are the two monthly interest rate series co-integrated?

**3.** Again consider the monthly 1-year and 10-year Treasury constant maturity rates from April 1953 to October 2000. Consider the log series of the data and build a VARMA model for the series. Discuss the implications of the model obtained.

**4.** Again consider the monthly 1-year and 10-year Treasury constant maturity rates from April 1953 to October 2000. Are the two interest rate series threshold-cointegrated? Use the interest spread $s_t = r_{10,t} - r_{1,t}$ as the threshold variable, where $r_{it}$ is the $i$-year Treasury constant maturity rate. If they are threshold-cointegrated, build a multivariate threshold model for the two series.

**5.** The bivariate AR(4) model $\boldsymbol{x}_t - \boldsymbol{\Phi}_4 \boldsymbol{x}_{t-4} = \boldsymbol{\phi}_0 + \boldsymbol{a}_t$ is a special seasonal model with periodicity 4, where $\{\boldsymbol{a}_t\}$ is a sequence of independent and identically distributed normal random vectors with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Such a seasonal model may be useful in studying quarterly earning of a company. (a) Assume that $\boldsymbol{x}_t$ is weakly stationary. Derive the mean vector and covariance matrix of $\boldsymbol{x}_t$. (b) Derive the necessary and sufficient condition of weak stationarity for $\boldsymbol{x}_t$. (c) Show that $\boldsymbol{\Gamma}_\ell = \boldsymbol{\Phi}_4 \boldsymbol{\Gamma}_{\ell-4}$ for $\ell > 0$, where $\boldsymbol{\Gamma}_\ell$ is the lag-$\ell$ autocovariance matrix of $\boldsymbol{x}_t$.

**6.** The bivariate MA(4) model $\boldsymbol{x}_t = \boldsymbol{a}_t - \boldsymbol{\Theta}_4 \boldsymbol{a}_{t-4}$ is another seasonal model with periodicity 4, where $\{\boldsymbol{a}_t\}$ is a sequence of independent and identically distributed normal random vectors with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Derive the covariance matrixes $\boldsymbol{\Gamma}_\ell$ of $\boldsymbol{x}_t$ for $\ell = 0, \ldots, 5$.

## REFERENCES


Balke, N. S., and Fomby, T. B. (1997), "Threshold cointegration," *International Economic Review*, 38, 627–645.

Box, G. E. P., and Tiao, G. C. (1977), "A canonical analysis of multiple time series," *Biometrika*, 64, 355–366.

Brenner, R. J., and Kroner, K. F. (1995), "Arbitrage, cointegration, and testing the unbiasedness hypothesis in financial markets," *Journal of Financial and Quantitative Analysis*, 30, 23–42.

Cochrane, J. H. (1988), "How big is the random walk in the GNP?" *Journal of Political Economy*, 96, 893–920.

Dwyer, Jr., G. P., Locke, P., and Yu, W. (1996), "Index arbitrage and nonlinear dynamics between the S&P 500 futures and cash," *Review of Financial Studies*, 9, 301–332.

Engle, R. F., and Granger, C. W. J. (1987), "Co-integration and error correction representation, estimation and testing," *Econometrica*, 55, 251–276.

Forbes, C. S., Kalb, G. R. J., and Kofman, P. (1999), "Bayesian arbitrage threshold analysis," *Journal of Business & Economic Statistics*, 17, 364–372.

Fuller, W. A. (1976), *Introduction to Statistical Time Series*, Wiley: New York.

Graybill, F. A. (1969), *Introduction to Matrixes with Applications in Statistics*, Wadsworth: Belmont, California.

Hillmer, S. C., and Tiao, G. C. (1979), "Likelihood function of stationary multiple autoregressive moving average models," *Journal of the American Statistical Association*, 74, 652–660.

Hosking, J. R. M. (1980), "The multivariate portmanteau statistic," *Journal of the American Statistical Association*, 75, 602–608.

Hosking, J. R. M. (1981), "Lagrange-multiplier tests of multivariate time series models," *Journal of the Royal Statistical Society, Series B*, 43, 219–230.

Johansen, S. (1989), "Statistical analysis of co-integration vectors," *Journal of Economic Dynamics and Control*, 12,231–254.

Johnson, R. A. and Wichern, D. W. (1998), *Applied Multivariate Statistical Analysis*, 4th ed., Prentice Hall: Upper Saddle River, New Jersey.

Kaiser, H. F. (1958), "The varimax criterion for analytic rotation in factor analysis," *Psychometrika*, 23, 187–200.

Li, W. K., and McLeod, A. I. (1981), "Distribution of the residual autocorrelations in multivariate ARMA time series models," *Journal of the Royal Statistical Society, Series B*, 43, 231–239.

Lütkepohl, H. (1991), *Introduction to Multiple Time Series Analysis*, Springer-Verlag: New York.

Reinsel, G. C. (1993), *Elements of Multivariate Time Series Analysis*, Springer-Verlag: New York.

Stock, J. H., and Watson, M. W. (1988), "Testing for common trends," *Journal of the American Statistical Association*, 83, 1097–1107.

Strang, G. (1980), *Linear Algebra and its Applications*, 2nd ed., Harcourt Brace Jovanovich: Chicago.

Tiao, G. C., and Box, G. E. P. (1981), "Modeling multiple time series with applications," *Journal of the American Statistical Association*, 76, 802–816.

Tiao, G. C., and Tsay, R. S. (1989), "Model specification in multivariate time series" (with discussions), *Journal of the Royal Statistical Society, Series B*, 51, 157–213.

Tiao, G. C., Tsay, R. S., and Wang, T. (1993), "Usefulness of linear transformations in multivariate time series analysis," *Empirical Economics*, 18, 567–593.

Tsay, R. S. (1991), "Two canonical forms for vector ARMA processes," *Statistica Sinica*, 1, 247–269.

Tsay, R. S. (1998), "Testing and modeling multivariate threshold models," *Journal of the American Statistical Association*, 93, 1188–1202.

Tsay, R. S., and Tiao, G. C. (1990), "Asymptotic properties of multivariate nonstationary processes with applications to autoregressions," *Annals of Statistics*, 18, 220–250.

*Analysis of Financial Time Series.* Ruey S. Tsay

ISBN: 0-471-41544-8

CHAPTER 9

# Multivariate Volatility Models and Their Applications

In this chapter, we generalize the univariate volatility models of Chapter 3 to the multivariate case and discuss some methods for simplifying the dynamic relationships between volatility processes of multiple asset returns. Multivariate volatilities have many important financial applications. They play an important role in portfolio selection and asset allocation, and they can be used to compute the Value at Risk of a financial position consisting of multiple assets.

Consider a multivariate return series $\{\boldsymbol{r}_t\}$. We adopt the same approach as the univariate case by rewriting the series as

$$\boldsymbol{r}_t = \boldsymbol{\mu}_t + \boldsymbol{a}_t,$$

where $\boldsymbol{\mu}_t = E(\boldsymbol{r}_t \mid \boldsymbol{F}_{t-1})$ is the conditional expectation of $\boldsymbol{r}_t$ given the past information $\boldsymbol{F}_{t-1}$, and $\boldsymbol{a}_t = (a_{1t}, \ldots, a_{kt})'$ is the shock, or innovation, of the series at time $t$. The $\boldsymbol{\mu}_t$ process is assumed to follow the conditional expectation of a multivariate time series model of Chapter 8. For most return series, it suffices to employ a simple vector ARMA structure for $\boldsymbol{\mu}_t$—that is,

$$\boldsymbol{\mu}_t = \boldsymbol{\phi}_0 + \sum_{i=1}^{p} \boldsymbol{\Phi}_i \boldsymbol{r}_{t-i} - \sum_{i=1}^{q} \boldsymbol{\Theta}_i \boldsymbol{a}_{t-i}, \tag{9.1}$$

where $p$ and $q$ are non-negative integers. Explanatory variables can be added to the prior equation if necessary. We refer to Eq. (9.1) as the mean equation of $\boldsymbol{r}_t$.

The conditional covariance matrix of $\boldsymbol{a}_t$ given $\boldsymbol{F}_{t-1}$ is a $k \times k$ positive-definite matrix $\boldsymbol{\Sigma}_t$ defined by $\boldsymbol{\Sigma}_t = \text{Cov}(\boldsymbol{a}_t \mid \boldsymbol{F}_{t-1})$. Multivariate volatility modeling is concerned with the time evolution of $\boldsymbol{\Sigma}_t$. We refer to a model for $\boldsymbol{\Sigma}_t$ as a volatility model for the return series $\boldsymbol{r}_t$.

There are many ways to generalize univariate volatility models to the multivariate case, but the curse of dimensionality quickly becomes a major obstacle in applications because there are $k(k+1)/2$ quantities in $\boldsymbol{\Sigma}_t$ for a $k$-dimensional return series. To illustrate, there are 15 conditional variances and covariances in $\boldsymbol{\Sigma}_t$ for a

five-dimensional return series. The goal of this chapter is to introduce some relatively simple multivariate volatility models that are useful, yet remain manageable in real application. In particular, we discuss some models that allow for time-varying correlation coefficients between asset returns. Time-varying correlations are useful in finance. For example, they can be used to estimate the time-varying beta of the market model of a return series.

We begin by introducing two methods to reparameterize $\mathbf{\Sigma}_t$ for volatility modeling in Section 9.1. The reparameterization based on the Cholesky decomposition is found to be very useful. We then study volatility models for bivariate returns in Section 9.2 using the GARCH model as an example. In this particular case, the volatility model can be bivariate or three-dimensional. Section 9.3 is concerned with volatility models for higher dimensional returns, and Section 9.4 addresses the issue of dimension reduction. We demonstrate some applications of multivariate volatility models in Section 9.5. Finally, Section 9.6 gives a multivariate student-$t$ distribution useful for volatility modeling.

## 9.1 REPARAMETERIZATION

An important step in multivariate volatility modeling is to reparameterize $\mathbf{\Sigma}_t$ by making use of its symmetric property. We consider two reparameterizations of $\mathbf{\Sigma}_t$.

### 9.1.1 Use of Correlations

The first reparameterization of $\mathbf{\Sigma}_t$ is to use the conditional correlation coefficients and variances of $\boldsymbol{a}_t$. Specifically, we write $\mathbf{\Sigma}_t$ as

$$\mathbf{\Sigma}_t \equiv [\sigma_{ij,t}] = \boldsymbol{D}_t \boldsymbol{\rho}_t \boldsymbol{D}_t, \tag{9.2}$$

where $\boldsymbol{\rho}_t$ is the conditional correlation matrix of $\boldsymbol{a}_t$, and $\boldsymbol{D}_t$ is a $k \times k$ diagonal matrix consisting of the conditional standard deviations of elements of $\boldsymbol{a}_t$ (i.e., $\boldsymbol{D}_t = \text{diag}\{\sqrt{\sigma_{11,t}}, \ldots, \sqrt{\sigma_{kk,t}}\}$).

Because $\boldsymbol{\rho}_t$ is symmetric with unit diagonal elements, the time evolution of $\mathbf{\Sigma}_t$ is governed by that of the conditional variances $\sigma_{ii,t}$ and the elements $\rho_{ij,t}$ of $\boldsymbol{\rho}_t$, where $j < i$ and $1 \leq i \leq k$. Therefore, to model the volatility of $\boldsymbol{a}_t$, it suffices to consider the conditional variances and correlation coefficients of $a_{it}$. Define the $k(k+1)/2$-dimensional vector

$$\mathbf{\Xi}_t = (\sigma_{11,t}, \ldots, \sigma_{kk,t}, \boldsymbol{\varrho}_t')', \tag{9.3}$$

where $\boldsymbol{\varrho}_t$ is a $k(k-1)/2$-dimensional vector obtained by stacking columns of the correlation matrix $\boldsymbol{\rho}_t$, but using only elements below the main diagonal. Specifically, for a $k$-dimensional return series,

$$\boldsymbol{\varrho}_t = (\rho_{21,t}, \ldots, \rho_{k1,t} | \, \rho_{32,t}, \ldots, \rho_{k2,t} \, | \cdots | \, \rho_{k,k-1,t})'.$$

To illustrate, for $k = 2$, we have $\boldsymbol{\varrho}_t = \rho_{21,t}$ and

$$\boldsymbol{\Xi}_t = (\sigma_{11,t}, \sigma_{22,t}, \rho_{21,t})', \tag{9.4}$$

which is a 3-dimensional vector, and for $k = 3$, we have $\boldsymbol{\varrho}_t = (\rho_{21,t}, \rho_{31,t}, \rho_{32,t})'$ and

$$\boldsymbol{\Xi}_t = (\sigma_{11,t}, \sigma_{22,t}, \sigma_{33,t}, \rho_{21,t}, \rho_{31,t}, \rho_{32,t})', \tag{9.5}$$

which is a six-dimensional vector.

If $\boldsymbol{a}_t$ is a bivariate normal random variable, then $\boldsymbol{\Xi}_t$ is given in Eq. (9.4) and the conditional density function of $\boldsymbol{a}_t$ given $\boldsymbol{F}_{t-1}$ is

$$f(a_{1t}, a_{2t} \mid \boldsymbol{\Xi}_t) = \frac{1}{2\pi\sqrt{\sigma_{11,t}\sigma_{22,t}(1-\rho_{21,t}^2)}} \exp\left[-\frac{Q(a_{1t}, a_{2t}, \boldsymbol{\Xi}_t)}{2(1-\rho_{21,t}^2)}\right],$$

where

$$Q(a_{1t}, a_{2t}, \boldsymbol{\Xi}_t) = \frac{a_{1t}^2}{\sigma_{11,t}} + \frac{a_{2t}^2}{\sigma_{22,t}} - \frac{2\rho_{21,t}a_{1t}a_{2t}}{\sqrt{\sigma_{11,t}\sigma_{22,t}}}.$$

The log probability density function of $\boldsymbol{a}_t$ relevant to the maximum likelihood estimation is

$$\ell(a_{1t}, a_{2t}, \boldsymbol{\Xi}_t) = -\frac{1}{2}\left\{ \ln[\sigma_{11,t}\sigma_{22,t}(1-\rho_{21,t}^2)] + \frac{1}{1-\rho_{21,t}^2}\left(\frac{a_{1t}^2}{\sigma_{11,t}} + \frac{a_{2t}^2}{\sigma_{22,t}} - \frac{2\rho_{21,t}a_{1t}a_{2t}}{\sqrt{\sigma_{11,t}\sigma_{22,t}}}\right)\right\}. \tag{9.6}$$

This reparameterization is useful because it models covariances and correlations directly. Yet the approach has several weaknesses. First, the likelihood function becomes complicated when $k \geq 3$. Second, the approach requires a constrained maximization in estimation to ensure the positive definiteness of $\boldsymbol{\Sigma}_t$. The constraint becomes complicated when $k$ is large.

### 9.1.2 Cholesky Decomposition

The second reparameterization of $\boldsymbol{\Sigma}_t$ is to use the Cholesky decomposition; see Appendix A of Chapter 8. This approach has some advantages in estimation as it requires no parameter constraints for the positive definiteness of $\boldsymbol{\Sigma}_t$; see Pourahmadi (1999). In addition, the reparameterization is an orthogonal transformation so that the resulting likelihood function is extremely simple. Details of the transformation are given next.

Because $\boldsymbol{\Sigma}_t$ is positive definite, there exist a lower triangular matrix $\boldsymbol{L}_t$ with unit diagonal elements and a diagonal matrix $\boldsymbol{G}_t$ with positive diagonal elements such that

$$\boldsymbol{\Sigma}_t = \boldsymbol{L}_t \boldsymbol{G}_t \boldsymbol{L}_t'. \tag{9.7}$$

This is the well-known Cholesky decomposition of $\boldsymbol{\Sigma}_t$. A nice feature of the decomposition is that the lower off-diagonal elements of $\boldsymbol{L}_t$ and the diagonal elements of $\boldsymbol{G}_t$ have nice interpretations. We demonstrate the decomposition by studying carefully the bivariate and three-dimensional cases. For the bivariate case, we have

$$\boldsymbol{\Sigma}_t = \begin{bmatrix} \sigma_{11,t} & \sigma_{21,t} \\ \sigma_{21,t} & \sigma_{22,t} \end{bmatrix}, \quad \boldsymbol{L}_t = \begin{bmatrix} 1 & 0 \\ q_{21,t} & 1 \end{bmatrix}, \quad \boldsymbol{G}_t = \begin{bmatrix} g_{11,t} & 0 \\ 0 & g_{22,t} \end{bmatrix},$$

where $g_{ii,t} > 0$ for $i = 1$ and 2. Using Eq. (9.7), we have

$$\boldsymbol{\Sigma}_t = \begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} \\ \sigma_{12,t} & \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} g_{11,t} & q_{21,t} g_{11,t} \\ q_{21,t} g_{11,t} & g_{22,t} + q_{21,t}^2 g_{11,t} \end{bmatrix}.$$

Equating elements of the previous matrix equation, we obtain

$$\sigma_{11,t} = g_{11,t}, \quad \sigma_{21,t} = q_{21,t} g_{11,t}, \quad \sigma_{22,t} = g_{22,t} + q_{21,t}^2 g_{11,t}. \tag{9.8}$$

Solving the prior equations, we have

$$g_{11,t} = \sigma_{11,t}, \quad q_{21,t} = \frac{\sigma_{21,t}}{\sigma_{11,t}}, \quad g_{22,t} = \sigma_{22,t} - \frac{\sigma_{21,t}^2}{\sigma_{11,t}}. \tag{9.9}$$

However, consider the simple conditional linear regression

$$a_{2t} = \beta a_{1t} + b_{2t}, \tag{9.10}$$

where $b_{2t}$ denotes the error term. From the well-known least squares theory, we have

$$\beta = \frac{\text{Cov}(a_{1t}, a_{2t})}{\text{Var}(a_{1t})} = \frac{\sigma_{21,t}}{\sigma_{11,t}},$$

$$\text{Var}(b_{2t}) = \text{Var}(a_{2t}) - \beta^2 \text{Var}(a_{1t}) = \sigma_{22,t} - \frac{\sigma_{21,t}^2}{\sigma_{11,t}}.$$

Furthermore, the error term $b_{2t}$ is uncorrelated with the regressor $a_{1t}$. Consequently, using Eq. (9.9), we obtain

$$g_{11,t} = \sigma_{11,t}, \quad q_{21,t} = \beta, \quad g_{22,t} = \text{Var}(b_{2t}), \quad b_{2t} \perp a_{1t},$$

where $\perp$ denotes no correlation. In summary, the Cholesky decomposition of the $2 \times 2$ matrix $\boldsymbol{\Sigma}_t$ amounts to performing an orthogonal transformation from $\boldsymbol{a}_t$ to $\boldsymbol{b}_t = (b_{1t}, b_{2t})'$ such that

$$b_{1t} = a_{1t} \quad \text{and} \quad b_{2t} = a_{2t} - q_{21,t} a_{1t},$$

where $q_{21,t} = \beta$ is obtained by the linear regression (9.10) and $\text{Cov}(\boldsymbol{b}_t)$ is a diagonal matrix with diagonal elements $g_{ii,t}$. The transformed quantities $q_{21,t}$ and $g_{ii,t}$ can be interpreted as follows:

1. The first diagonal element of $\boldsymbol{G}_t$ is simply the variance of $a_{1t}$.
2. The second diagonal element of $\boldsymbol{G}_t$ is the residual variance of the simple linear regression in Eq. (9.10).
3. The element $q_{21,t}$ of the lower triangular matrix $\boldsymbol{L}_t$ is the coefficient $\beta$ of the regression in Eq. (9.10).

The prior properties continue to hold for the higher dimensional case. For example, consider the three-dimensional case in which

$$\boldsymbol{L}_t = \begin{bmatrix} 1 & 0 & 0 \\ q_{21,t} & 1 & 0 \\ q_{31,t} & q_{32,t} & 1 \end{bmatrix}, \quad \boldsymbol{G}_t = \begin{bmatrix} g_{11,t} & 0 & 0 \\ 0 & g_{22,t} & 0 \\ 0 & 0 & g_{3,t} \end{bmatrix}.$$

From the decomposition in Eq. (9.7), we have

$$\begin{bmatrix} \sigma_{11,t} & \sigma_{21,t} & \sigma_{31,t} \\ \sigma_{21,t} & \sigma_{22,t} & \sigma_{32,t} \\ \sigma_{31,t} & \sigma_{32,t} & \sigma_{33,t} \end{bmatrix} =$$

$$\begin{bmatrix} g_{11,t} & q_{21,t} g_{11,t} & q_{31,t} g_{11,t} \\ q_{21,t} g_{22,t} & q_{21,t}^2 g_{11,t} + g_{22,t} & q_{31,t} q_{21,t} g_{11,t} + q_{32,t} g_{22,t} \\ q_{31,t} g_{11,t} & q_{31,t} q_{21,t} g_{11,t} + q_{32,t} g_{22,t} & q_{31,t}^2 g_{11,t} + q_{32,t}^2 g_{22,t} + g_{33,t} \end{bmatrix}.$$

Equating elements of the prior matrix equation, we obtain

$$\sigma_{11,t} = g_{11,t}, \sigma_{21,t} = q_{21,t} g_{11,t}, \quad \sigma_{22,t} = q_{21,t}^2 g_{11,t} + g_{22,t}, \sigma_{31,t} = q_{31,t} g_{11,t},$$

$$\sigma_{32,t} = q_{31,t} q_{21,t} g_{11,t} + q_{32,t} g_{22,t}, \quad \sigma_{33,t} = q_{31,t}^2 g_{11,t} + q_{32,t}^2 g_{22,t} + g_{33,t}$$

or, equivalently,

$$g_{11,t} = \sigma_{11,t}, \quad q_{21,t} = \frac{\sigma_{21,t}}{\sigma_{11,t}}, \quad g_{22,t} = \sigma_{22,t} - q_{21,t}^2 g_{11,t},$$

$$q_{31,t} = \frac{\sigma_{31,t}}{\sigma_{11,t}}, \quad q_{32,t} = \frac{1}{g_{22,t}} \left( \sigma_{32,t} - \frac{\sigma_{31,t} \sigma_{21,t}}{\sigma_{11,t}} \right),$$

$$g_{33,t} = \sigma_{33,t} - q_{31,t}^2 g_{11,t} - q_{32,t}^2 g_{22,t}.$$

These quantities look complicated, but they are simply the coefficients and residual variances of the orthogonal transformation

$$\begin{aligned} b_{1t} &= a_{1t} \\ b_{2t} &= a_{2t} - \beta_{21} b_{1t} \\ b_{3t} &= a_{3t} - \beta_{31} b_{1t} - \beta_{32} b_{2t}, \end{aligned}$$

where $\beta_{ij}$ are the coefficients of least squares regressions

$$\begin{aligned} a_{2t} &= \beta_{21} b_{1t} + b_{2t} \\ a_{3t} &= \beta_{31} b_{1t} + \beta_{32} b_{2t} + b_{3t}. \end{aligned}$$

In other words, we have $q_{ij,t} = \beta_{ij}$, $g_{ii,t} = \text{Var}(b_{it})$ and $b_{it} \perp b_{jt}$ for $i \neq j$.

Based on the prior discussion, using Cholesky decomposition amounts to doing an orthogonal transformation from $\boldsymbol{a}_t$ to $\boldsymbol{b}_t$, where $b_{1t} = a_{1t}$, and $b_{it}$, for $1 < i \leq k$, is defined recursively by the least squares regression

$$a_{it} = q_{i1,t} b_{1t} + q_{i2,t} b_{2t} + \cdots + q_{i(i-1),t} b_{(i-1)t} + b_{it}, \tag{9.11}$$

where $q_{ij,t}$ is the $(i, j)$th element of the lower triangular matrix $\boldsymbol{L}_t$ for $1 \leq j < i$. We can write this transformation as

$$\boldsymbol{b}_t = \boldsymbol{L}_t^{-1} \boldsymbol{a}_t, \quad \text{or} \quad \boldsymbol{a}_t = \boldsymbol{L}_t \boldsymbol{b}_t, \tag{9.12}$$

where, as mentioned before, $\boldsymbol{L}_t^{-1}$ is also a lower triangular matrix with unit diagonal elements. The covariance matrix of $\boldsymbol{b}_t$ is the diagonal matrix $\boldsymbol{G}_t$ of the Cholesky decomposition because

$$\text{Cov}(\boldsymbol{b}_t) = \boldsymbol{L}_t^{-1} \boldsymbol{\Sigma}_t (\boldsymbol{L}_t^{-1})' = \boldsymbol{G}_t.$$

The parameter vector relevant to volatility modeling under such a transformation becomes

$$\boldsymbol{\Xi}_t = (g_{11,t}, \ldots, g_{kk,t}, q_{21,t}, q_{31,t}, q_{32,t}, \ldots, q_{k1,t}, \ldots, q_{k(k-1),t})', \tag{9.13}$$

which is also a $k(k+1)/2$-dimensional vector.

The previous orthogonal transformation also dramatically simplifies the likelihood function of the data. Using the fact that $|\boldsymbol{L}_t| = 1$, we have

$$|\boldsymbol{\Sigma}_t| = |\boldsymbol{L}_t \boldsymbol{G}_t \boldsymbol{L}_t'| = |\boldsymbol{G}_t| = \prod_{i=1}^{k} g_{ii,t}. \tag{9.14}$$

If the conditional distribution of $\boldsymbol{a}_t$ given the past information is multivariate normal $N(\boldsymbol{0}, \boldsymbol{\Sigma}_t)$, then the conditional distribution of the transformed series $\boldsymbol{b}_t$ is multivariate

normal $N(\mathbf{0}, \boldsymbol{G}_t)$, and the log likelihood function of the data becomes extremely simple. Indeed, we have the log probability density of $\boldsymbol{a}_t$ as

$$\ell(\boldsymbol{a}_t, \boldsymbol{\Sigma}_t) = \ell(\boldsymbol{b}_t, \boldsymbol{\Xi}_t) = -\frac{1}{2}\sum_{i=1}^{k}\left[\ln(g_{ii,t}) + \frac{b_{it}^2}{g_{ii,t}}\right], \tag{9.15}$$

where for simplicity the constant term is omitted and $g_{ii,t}$ is the variance of $b_{it}$.

Using the Cholesky decomposition to reparameterize $\boldsymbol{\Sigma}_t$ has several advantages. First, from Eq. (9.14), $\boldsymbol{\Sigma}_t$ is positive definite if $g_{ii,t} > 0$ for all $i$. Consequently, the positive definite constraint of $\boldsymbol{\Sigma}_t$ can easily be achieved by modeling $\ln(g_{ii,t})$ instead of $g_{ii,t}$. Second, elements of the parameter vector $\boldsymbol{\Xi}_t$ in Eq. (9.13) have nice interpretations. They are the coefficients and residual variances of multiple linear regressions that orthogonalize the shocks to the returns. Third, the correlation coefficient between $a_{1t}$ and $a_{2t}$ is

$$\rho_{21,t} = \frac{\sigma_{21,t}}{\sqrt{\sigma_{11,t}\sigma_{22,t}}} = q_{21,t} \times \frac{\sqrt{\sigma_{11,t}}}{\sqrt{\sigma_{22,t}}},$$

which is time-varying if $q_{21,t} \neq \rho\sqrt{\sigma_{22,t}}/\sqrt{\sigma_{11,t}}$ where $\rho$ is a constant. For example, if $q_{21,t} = c \neq 0$, then $\rho_{21,t} = c\sqrt{\sigma_{11,t}}/\sqrt{\sigma_{22,t}}$, which continues to be time-varying provided that the variance ratio $\sigma_{11,t}/\sigma_{22,t}$ is not a constant. This time-varying property applies to other correlation coefficients when the dimension of $\boldsymbol{r}_t$ is greater than 2 and is a major difference between the two approaches for reparameterizing $\boldsymbol{\Sigma}_t$.

Using Eq. (9.11) and the orthogonality among the transformed shocks $b_{it}$, we obtain

$$\sigma_{ii,t} = \text{Var}(a_{it} \mid F_{t-1}) = \sum_{v=1}^{i} q_{iv,t}^2 g_{vv,t}, \quad i = 1, \ldots, k,$$

$$\sigma_{ij,t} = \text{Cov}(a_{it}, a_{jt} \mid F_{t-1}) = \sum_{v=1}^{j} q_{iv,t} q_{jv,t} g_{vv,t}, \quad j < i, \quad i = 2, \ldots, k,$$

where $q_{vv,t} = 1$ for $v = 1, \ldots, k$. These equations show the parameterization of $\boldsymbol{\Sigma}_t$ under the Cholesky decomposition.

## 9.2 GARCH MODELS FOR BIVARIATE RETURNS

Since the same techniques can be used to generalize many univariate volatility models to the multivariate case, we focus our discussion on the multivariate GARCH model. Other multivariate volatility models can also be used.

For a $k$-dimensional return series $\boldsymbol{r}_t$, a multivariate GARCH model uses "exact equations" to describe the evolution of the $k(k+1)/2$-dimensional vector $\boldsymbol{\Xi}_t$ over time. By exact equation, we mean that the equation does not contain any stochastic

shock. However, the exact equation may become complicated even in the simplest case of $k = 2$ for which $\Xi_t$ is three-dimensional. To keep the model simple, some restrictions are often imposed on the equations.

### 9.2.1 Constant-Correlation Models

To keep the number of volatility equations low, Bollerslev (1990) considers the special case in which the correlation coefficient $\rho_{21,t} = \rho_{21}$ is time invariant, where $|\rho_{21}| < 1$. Under such an assumption, $\rho_{21}$ is a constant parameter and the volatility model consists of two equations for $\Xi_t^*$, which is defined as $\Xi_t^* = (\sigma_{11,t}, \sigma_{22,t})'$. A GARCH(1, 1) model for $\Xi_t^*$ becomes

$$\Xi_t^* = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 \boldsymbol{a}_{t-1}^2 + \boldsymbol{\beta}_1 \Xi_{t-1}^*, \tag{9.16}$$

where $\boldsymbol{a}_{t-1}^2 = (a_{1,t-1}^2, a_{2,t-1}^2)'$, $\boldsymbol{\alpha}_0$ is a two-dimensional positive vector, and $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ are $2 \times 2$ non-negative definite matrixes. More specifically, the model can be expressed in detail as

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} \alpha_{10} \\ \alpha_{20} \end{bmatrix} + \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{bmatrix} \begin{bmatrix} a_{1,t-1}^2 \\ a_{2,t-1}^2 \end{bmatrix} + \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{22,t-1} \end{bmatrix}, \tag{9.17}$$

where $\alpha_{i0} > 0$ for $i = 1$ and 2. Defining $\boldsymbol{\eta}_t = \boldsymbol{a}_t^2 - \Xi_t^*$, we can rewrite the prior model as

$$\boldsymbol{a}_t^2 = \boldsymbol{\alpha}_0 + (\boldsymbol{\alpha}_1 + \boldsymbol{\beta}_1)\boldsymbol{a}_{t-1}^2 + \boldsymbol{\eta}_t - \boldsymbol{\beta}_1 \boldsymbol{\eta}_{t-1},$$

which is a bivariate ARMA(1, 1) model for the $\boldsymbol{a}_t^2$ process. This result is a direct generalization of the univariate GARCH(1, 1) model of Chapter 3. Consequently, some properties of model (9.17) are readily available from those of the bivariate ARMA(1, 1) model of Chapter 8. In particular, we have the following results:

1. If all of the eigenvalues of $\boldsymbol{\alpha}_1 + \boldsymbol{\beta}_1$ are positive, but less than 1, then the bivariate ARMA(1, 1) model for $\boldsymbol{a}_t^2$ is weakly stationary and, hence, $E(\boldsymbol{a}_t^2)$ exists. This implies that the shock process $\boldsymbol{a}_t$ of the returns has a positive-definite unconditional covariance matrix. The unconditional variances of the elements of $\boldsymbol{a}_t$ are $(\sigma_1^2, \sigma_2^2)' = (\boldsymbol{I} - \boldsymbol{\alpha}_1 - \boldsymbol{\beta}_1)^{-1}\phi_0$, and the unconditional covariance between $a_{1t}$ and $a_{2t}$ is $\rho_{21}\sigma_1\sigma_2$.
2. If $\alpha_{12} = \beta_{12} = 0$, then the volatility of $a_{1t}$ does not depend on the past volatility of $a_{2t}$. Similarly, if $\alpha_{21} = \beta_{21} = 0$, then the volatility of $a_{2t}$ does not depend on the past volatility of $a_{1t}$.
3. If both $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}_1$ are diagonal, then the model reduces to two univariate GARCH(1, 1) models. In this case, the two volatility processes are not dynamically related.

**4.** Volatility forecasts of the model can be obtained by using forecasting methods similar to those of a vector ARMA(1, 1) model; see the univariate case in Chapter 3. The 1-step ahead volatility forecast at the forecast origin $h$ is

$$\Xi_h^*(1) = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 \boldsymbol{a}_h^2 + \boldsymbol{\beta}_1 \Xi_h^*.$$

For the $\ell$-step ahead forecast, we have

$$\Xi_h^*(\ell) = \boldsymbol{\alpha}_0 + (\boldsymbol{\alpha}_1 + \boldsymbol{\beta}_1)\Xi_h^*(\ell - 1), \quad \ell > 1.$$

These forecasts are for the marginal volatilities of $a_{it}$. The $\ell$-step ahead forecast of the covariance between $a_{1t}$ and $a_{2t}$ is $\hat{\rho}_{21}[\sigma_{11,h}(\ell)\sigma_{22,h}(\ell)]^{0.5}$, where $\hat{\rho}_{21}$ is the estimate of $\rho_{21}$ and $\sigma_{ii,h}(\ell)$ are the elements of $\Xi_h^*(\ell)$.

**Example 9.1.** As an illustration, consider the daily log returns of the stock market indexes for Hong Kong and Japan from January 1, 1996 to October 16, 1997 for 469 observations. The indexes are dollar denominated and the returns are in percentages. We select the sample period to avoid the effect of Asian financial crisis,

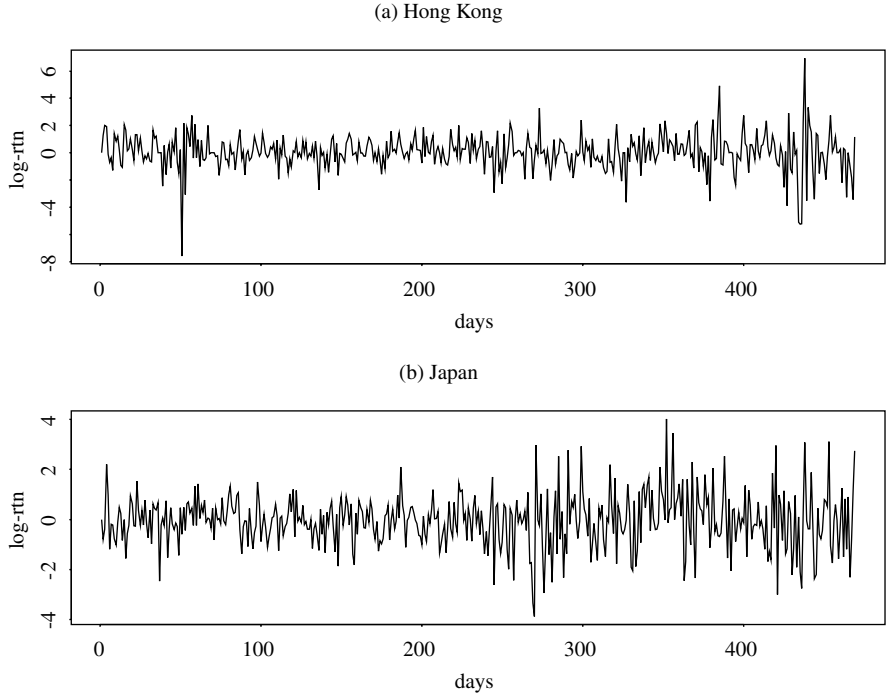


**Figure 9.1.** Time plots of daily log returns in percentages of stock market indexes for Hong Kong and Japan from January 1, 1996 to October 16, 1997: (a) the Hong Kong market, and (b) the Japanese market.

which hit the Hong Kong Market on October 17, 1997. The data are obtained from Datastream. Figure 9.1 shows the time plots of the two index returns. Let $r_{1t}$ be the index return for the Hong Kong stock market and $r_{2t}$ for the Japanese stock market. If univariate GARCH models are entertained, we obtain the models

$$\begin{aligned} r_{1t} &= 0.137 r_{1,t-1} + a_{1t}, \quad a_{1t} = \sigma_{1t}\epsilon_{1t}, \\ \sigma_{1t}^2 &= 0.164 + 0.142 a_{1,t-1}^2 + 0.765\sigma_{1,t-1}^2 \end{aligned} \tag{9.18}$$

$$\begin{aligned} r_{2t} &= a_{2t}, \quad a_{2t} = \sigma_{2t}\epsilon_{2t}, \\ \sigma_{2t}^2 &= 0.085 + 0.128 a_{2,t-1}^2 + 0.807\sigma_{2,t-1}^2, \end{aligned} \tag{9.19}$$

where all of the parameter estimates are highly significant except for the AR(1) coefficient of the $r_{1t}$ series, which has a $p$ value of 0.029. The Ljung–Box statistics of the standardized residuals and their squared series of the prior two models fail to indicate any model inadequacy. Figure 9.2 shows the estimated volatilities of the previous two univariate GARCH(1, 1) models. The Hong Kong stock market appears to be more volatile than the Japanese stock market, but the Japanese market exhibits an increas-

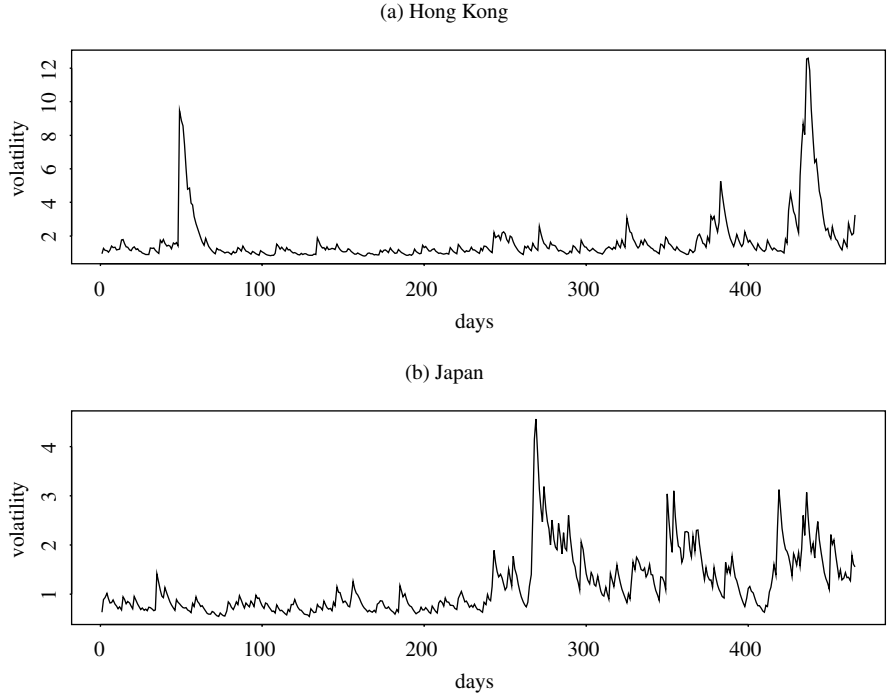


**Figure 9.2.** Estimated volatilities for daily log returns in percentages of stock market indexes for Hong Kong and Japan from January 1, 1996 to October 16, 1997: (a) the Hong Kong market, and (b) the Japanese market. Univariate models are used.

ing trend in volatility. The unconditional innovational variance of the Hong Kong market is about 1.76 and that of the Japanese market is 1.31.

Turning to bivariate GARCH models, we obtain two models that fit the data well. The mean equations of the first bivariate model are

$$r_{1t} = -0.118 r_{1,t-6} + a_{1t}$$
$$r_{2t} = a_{2t},$$

where the standard error of the AR(6) coefficient is 0.044. The volatility equations of the first model are

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} 0.275 \\ (0.079) \\ 0.051 \\ (0.014) \end{bmatrix} + \begin{bmatrix} 0.112 & \cdot \\ (0.032) & \\ \cdot & 0.091 \\ & (0.026) \end{bmatrix} \begin{bmatrix} a_{1,t-1}^2 \\ a_{2,t-1}^2 \end{bmatrix}$$
$$+ \begin{bmatrix} 0.711 & \cdot \\ (0.068) & \\ \cdot & 0.869 \\ & (0.028) \end{bmatrix} \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{22,t-1} \end{bmatrix}, \tag{9.20}$$

where the numbers in parentheses are standard errors. The estimated correlation coefficients between $a_{1t}$ and $a_{2t}$ is 0.226 with standard error 0.047.

Let $\tilde{\boldsymbol{a}}_t = (\tilde{a}_{1t}, \tilde{a}_{2t})'$ be the standardized residuals, where $\tilde{a}_{it} = a_{it}/\sqrt{\sigma_{ii,t}}$. The Ljung–Box statistics of $\tilde{\boldsymbol{a}}_t$ give $Q(4) = 22.29(0.10)$ and $Q(8) = 34.83(0.29)$, where the number in parentheses denotes $p$ value. Here the $p$ values are based on chi-squared distributions with 15 and 31 degrees of freedom, respectively, because an AR(6) coefficient is used in the mean equation. The Ljung–Box statistics for the $\tilde{\boldsymbol{a}}_t^2$ process give $Q(4) = 9.54(0.85)$ and $Q(8) = 18.58(0.96)$. Consequently, there are no serial correlations or conditional heteroscedasticities in the bivariate standardized residuals of model (9.20). The unconditional innovational variances of the two residuals are 1.55 and 1.28, respectively, for the Hong Kong and Japanese markets.

The model in Eq. (9.20) shows two uncoupled volatility equations, indicating that the volatilities of the two markets are not dynamically related, but they are contemporaneously correlated. We refer to the model as a bivariate *diagonal constant-correlation* model.

The mean equations of the second bivariate GARCH model are

$$r_{1t} = -0.143 r_{1,t-6} + a_{1t}$$
$$r_{2t} = a_{2t},$$

where the standard error of the AR(6) coefficient is 0.042, and the volatility equations of the second model are

$$\begin{bmatrix}\sigma_{11,t}\\ \sigma_{22,t}\end{bmatrix} = \begin{bmatrix}0.378\\ (0.103)\\ \cdot\end{bmatrix} + \begin{bmatrix}0.108 & \cdot\\ (0.030) & \\ \cdot & 0.172\\ & (0.035)\end{bmatrix}\begin{bmatrix}a_{1,t-1}^2\\ a_{2,t-1}^2\end{bmatrix}$$

$$+ \begin{bmatrix}\cdot & 0.865\\ & (0.109)\\ 0.321 & 0.869\\ (0.135) & (0.028)\end{bmatrix}\begin{bmatrix}\sigma_{11,t-1}\\ \sigma_{22,t-1}\end{bmatrix}, \tag{9.21}$$

where the numbers in parentheses are standard errors. The estimated correlation coefficient between $a_{1t}$ and $a_{2t}$ is 0.236 with standard error 0.045. Defining the standardized residuals as before, we obtain $Q(4) = 24.22(0.06)$ and $Q(8) = 35.52(0.26)$ for the standardized residuals of the prior model and $Q(4) = 17.45(0.29)$ and $Q(8) = 24.55(0.79)$ for the squared standardized residuals. These Ljung–Box statistics are insignificant at the 5% level, and hence the model in Eq. (9.21) is also adequate. The unconditional innovational variances of the prior model are 1.71 and 1.32, respectively, for the Hong Kong and Japanese markets.

In contrast with model (9.20), this second bivariate GARCH(1, 1) model shows a feedback relationship between the two markets. It is then interesting to compare the two volatility models. First, the unconditional innovational variances of model (9.21) are closer to those of the univariate models in Eqs. (9.18) and (9.19). Second, Figure 9.3 shows the fitted volatility processes of model (9.20), whereas Figure 9.4 shows those of model (9.21). Because model (9.20) implies no dynamic volatility dependence between the two markets, Figure 9.3 is similar to that of Figure 9.2. In contrast, Figure 9.4 shows evidence of mutual impacts between the two markets. Third, the maximized log likelihood function for model (9.20) is $-535.13$ for $t = 8, \ldots, 469$, whereas that of model (9.21) is $-540.32$; see the log probability density function in Eq. (9.6). Therefore, model (9.20) is preferred if one uses the likelihood principal. Finally, because practical implications of the two bivariate volatility models differ dramatically, further investigation is needed to separate them. Such an investigation may use a longer sample period or include more variables (e.g., using some U.S. market returns).

**Example 9.2.** As a second illustration, consider the monthly log returns, in percentages, of IBM stock and the S&P 500 index from January 1926 to December 1999 used in Chapter 8. Let $r_{1t}$ and $r_{2t}$ be the monthly log returns for IBM stock and the S&P 500 index, respectively. If a constant-correlation GARCH(1, 1) model is entertained, we obtain the mean equations

$$r_{1t} = 1.351 + 0.072r_{1,t-1} + 0.055r_{1,t-2} - 0.119r_{2,t-2} + a_{1t}$$
$$r_{2t} = 0.703 + a_{2t},$$

where standard errors of the parameters in the first equation are 0.225, 0.029, 0.034, and 0.044, respectively, and that of the parameter in the second equation is 0.155.

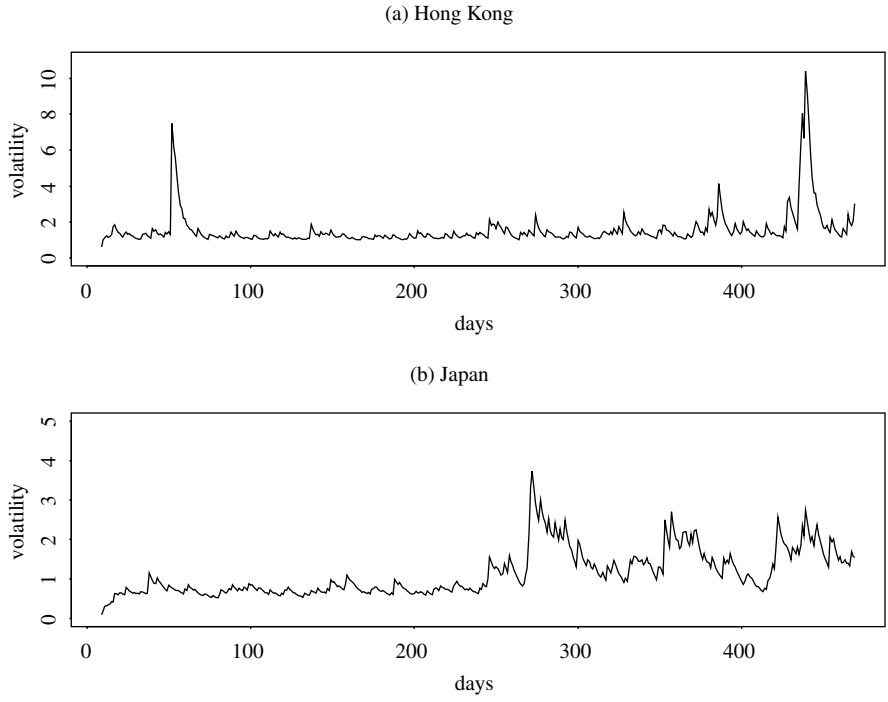


**Figure 9.3.** Estimated volatilities for daily log returns in percentages of stock market indexes for Hong Kong and Japan from January 1, 1996 to October 16, 1997: (a) the Hong Kong market, and (b) the Japanese market. The model used is in Eq. (9.20).

The volatility equations are

$$\begin{bmatrix}\sigma_{11,t}\\ \sigma_{22,t}\end{bmatrix} = \begin{bmatrix}2.98\\ (0.59)\\ 2.09\\ (0.47)\end{bmatrix} + \begin{bmatrix}0.079 & \cdot\\ (0.013) & \\ 0.042 & 0.045\\ (0.009) & (0.010)\end{bmatrix}\begin{bmatrix}a^2_{1,t-1}\\ a^2_{2,t-1}\end{bmatrix} + \begin{bmatrix}0.873 & -0.031\\ (0.020) & (0.009)\\ -0.066 & 0.913\\ (0.015) & (0.014)\end{bmatrix}\begin{bmatrix}\sigma_{11,t-1}\\ \sigma_{22,t-1}\end{bmatrix}, \tag{9.22}$$

where the numbers in parentheses are standard errors. The constant correlation coefficient is 0.614 with standard error 0.020. Using the standardized residuals, we obtain the Ljung–Box statistics $Q(4) = 16.77(0.21)$ and $Q(8) = 32.40(0.30)$, where the $p$ values shown in parentheses are obtained from chi-squared distributions with 13 and 29 degrees of freedom, respectively. Here the degrees of freedom have been adjusted because the mean equations contain three lagged predictors. For the squared stan-

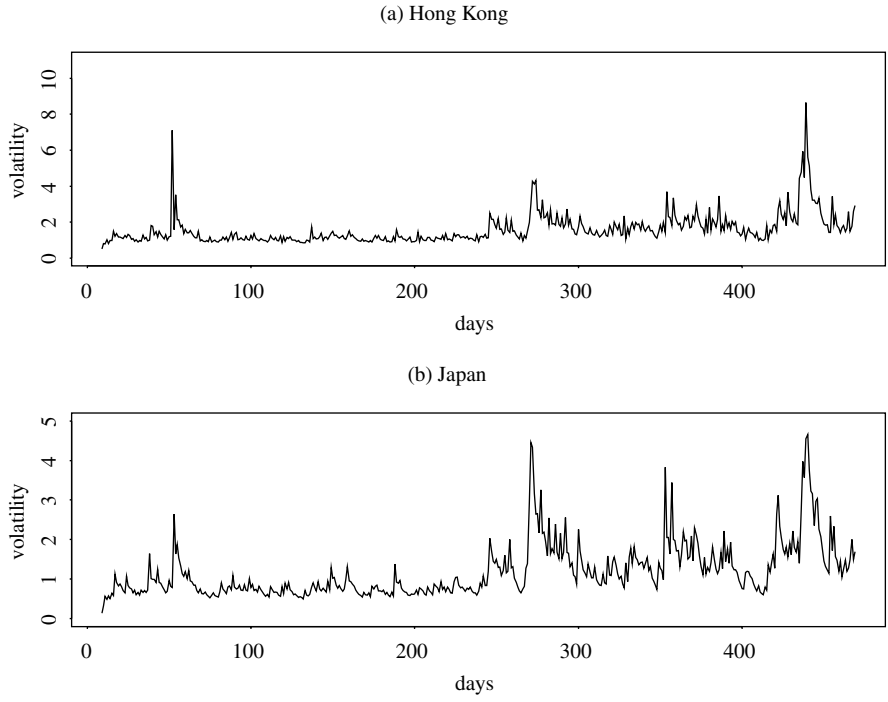


**Figure 9.4.** Estimated volatilities for daily log returns in percentages of stock market indexes for Hong Kong and Japan from January 1, 1996 to October 16, 1997: (a) the Hong Kong market, and (b) the Japanese market. The model used is in Eq. (9.21).

dardized residuals, we have $Q(4) = 18.00(0.16)$ and $Q(8) = 39.09(0.10)$. Therefore, at the 5% significance level, the standardized residuals $\tilde{\boldsymbol{a}}_t$ have no serial correlations or conditional heteroscedasticities. This bivariate GARCH(1, 1) model shows a feedback relationship between the volatilities of the two monthly log returns.

### 9.2.2 Time-Varying Correlation Models

A major drawback of the constant-correlation volatility models is that the correlation coefficient tends to change over time in a real application. Consider the monthly log returns of IBM stock and the S&P 500 index used in Example 9.2. It is hard to justify that the S&P 500 index return, which is a weighted average, can maintain a constant correlation coefficient with IBM return over the past 70 years. Figure 9.5 shows the sample correlation coefficient between the two monthly log return series using a moving window of 120 observations (i.e., 10 years). The correlation changes over time and appears to be decreasing in recent years. The decreasing trend in correlation is not surprising because the ranking of IBM market capitalization among large U.S. industrial companies has changed in recent years. A Lagrange multiplier statistic

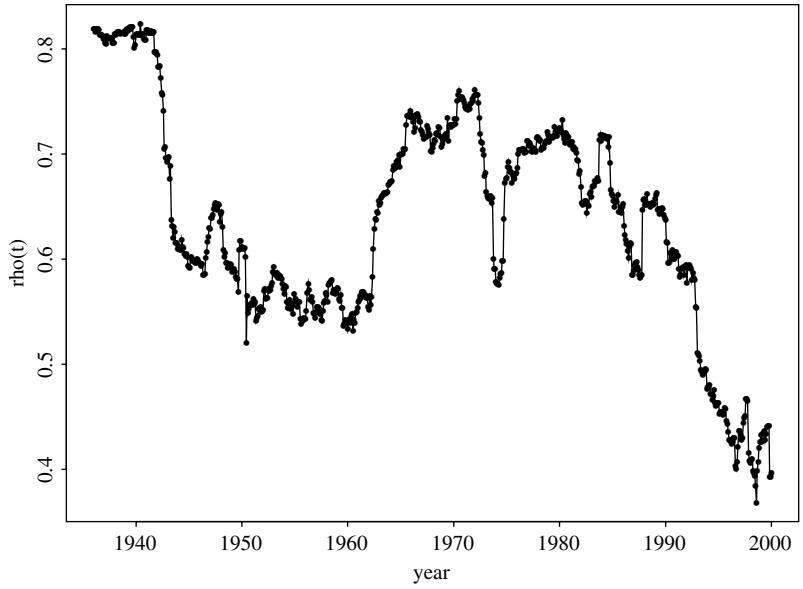


**Figure 9.5.** The sample correlation coefficient between monthly log returns of IBM stock and the S&P 500 index. The correlation is computed by a moving window of 120 observations. The sample period is from January 1926 to December 1999.

was proposed recently by Tse (2000) to test constant correlation coefficients in a multivariate GARCH model.

A simple way to relax the constant-correlation constraint within the GARCH framework is to specify an exact equation for the conditional correlation coefficient. This can be done by two methods using the two reparameterizations of $\mathbf{\Sigma}_t$ discussed in Section 9.1. First, we use the correlation coefficient directly. Because the correlation coefficient between the returns of IBM stock and S&P 500 index is positive and must be in the interval [0, 1], we employ the equation

$$\rho_{21,t} = \frac{\exp(q_t)}{1 + \exp(q_t)}, \tag{9.23}$$

where

$$q_t = \varpi_0 + \varpi_1 \rho_{21,t-1} + \varpi_2 \frac{a_{1,t-1}a_{2,t-1}}{\sqrt{\sigma_{11,t-1}\sigma_{22,t-1}}},$$

where $\sigma_{ii,t-1}$ is the conditional variance of the shock $a_{i,t-1}$. We refer to this equation as a GARCH(1, 1) model for the correlation coefficient because it uses the lag-1 cross-correlation and the lag-1 cross-product of the two shocks. If $\varpi_1 = \varpi_2 = 0$, then model (9.23) reduces to the case of constant correlation.

In summary, a time-varying correlation bivariate GARCH(1, 1) model consists of two sets of equations. The first set of equations consists of a bivariate GARCH(1, 1) model for the conditional variances and the second set of equation is a GARCH(1, 1) model for the correlation in Eq. (9.23). In practice, a negative sign can be added to Eq. (9.23) if the correlation coefficient is negative. In general, when the sign of correlation is unknown, we can use the Fisher transformation for correlation

$$q_t = \ln\left(\frac{1+\rho_{21,t}}{1-\rho_{21,t}}\right) \quad \text{or} \quad \rho_{21,t} = \frac{\exp(q_t)-1}{\exp(q_t)+1}$$

and employ a GARCH model for $q_t$ to model the time-varying correlation between two returns.

**Example 9.2.** (continued). Augmenting Eq. (9.23) to the GARCH(1, 1) model in Eq. (9.22) for the monthly log returns of IBM stock and the S&P 500 index and performing a joint estimation, we obtain the following model for the two series:

$$\begin{aligned} r_{1t} &= 1.318 + 0.076 r_{1,t-1} - 0.068 r_{2,t-2} + a_{1t} \\ r_{2t} &= 0.673 + a_{2t}, \end{aligned}$$

where standard errors of the three parameters in the first equation are 0.215, 0.026, and 0.034, respectively, and that of the parameter in the second equation is 0.151. The volatility equations are

$$\begin{bmatrix} \sigma_{11,t} \\ \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} 2.80 \\ (0.58) \\ 1.71 \\ (0.40) \end{bmatrix} + \begin{bmatrix} 0.084 & \cdot \\ (0.013) & \\ 0.037 & 0.054 \\ (0.009) & (0.010) \end{bmatrix} \begin{bmatrix} a^2_{1,t-1} \\ a^2_{2,t-1} \end{bmatrix} + \begin{bmatrix} 0.864 & -0.020 \\ (0.021) & (0.009) \\ -0.058 & 0.914 \\ (0.014) & (0.013) \end{bmatrix} \begin{bmatrix} \sigma_{11,t-1} \\ \sigma_{22,t-1} \end{bmatrix}, \tag{9.24}$$

where, as before, standard errors are in parentheses. The conditional correlation equation is

$$\rho_t = \frac{\exp(q_t)}{1+\exp(q_t)}, \quad q_t = -2.024 + 3.983\rho_{t-1} + 0.088\frac{a_{1,t-1}a_{2,t-1}}{\sqrt{\sigma_{11,t-1}\sigma_{22,t-1}}}, \tag{9.25}$$

where standard errors of the estimates are 0.050, 0.090, and 0.019, respectively. The parameters of the prior correlation equation are highly significant. Applying the Ljung–Box statistics to the standardized residuals $\tilde{\boldsymbol{a}}_t$, we have $Q(4) = 20.57(0.11)$ and $Q(8) = 36.08(0.21)$. For the squared standardized residuals, we have $Q(4) =$

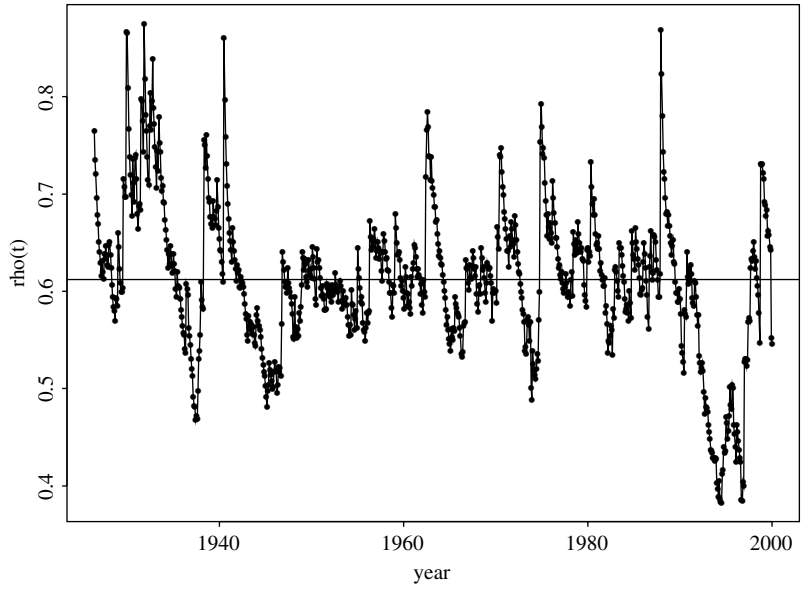


**Figure 9.6.** The fitted conditional correlation coefficient between monthly log returns of IBM stock and the S&P 500 index using the time-varying correlation GARCH(1, 1) model of Example 9.2. The horizontal line denotes the average 0.612 of the correlation coefficients.

16.69(0.27) and $Q(8) = 36.71(0.19)$. Therefore, the standardized residuals of the model have no significant serial correlations or conditional heteroscedasticities.

It is interesting to compare this time-varying correlation GARCH(1, 1) model with the constant-correlation GARCH(1, 1) model in Eq. (9.22). First, the mean and volatility equations of the two models are close. Second, Figure 9.6 shows the fitted conditional-correlation coefficient between the monthly log returns of IBM stock and the S&P 500 index based on model (9.25). The plot shows that the correlation coefficient fluctuated over time and became smaller in recent years. This latter characteristic is in agreement with that of Figure 9.5. Third, the average of the fitted correlation coefficients is 0.612, which is essentially the estimate 0.614 of the constant-correlation model in Eq. (9.22). Fourth, using the sample variances of $r_{it}$ as the starting values for the conditional variances and the observations from $t = 4$ to $t = 888$, the maximized log likelihood function is $-3691.21$ for the constant-correlation GARCH(1, 1) model and $-3679.64$ for the time-varying correlation GARCH(1, 1) model. Thus, the time-varying correlation model shows some significant improvement over the constant-correlation model. Finally, consider the 1-step ahead volatility forecasts of the two models at the forecast origin $h = 888$. For the constant-correlation model in Eq. (9.22), we have $a_{1,888} = 3.075$, $a_{2,888} = 4.931$, $\sigma_{11,888} = 77.91$, and $\sigma_{22,888} = 21.19$. Therefore, the 1-step ahead forecast for the conditional covariance matrix is

$$\widehat{\boldsymbol{\Sigma}}_{888}(1) = \begin{bmatrix} 71.09 & 21.83 \\ 21.83 & 17.79 \end{bmatrix},$$

where the covariance is obtained by using the constant correlation coefficient 0.614. For the time-varying correlation model in Eqs. (9.24) and (9.25), we have $a_{1,888} = 3.287$, $a_{2,888} = 4.950$, $\sigma_{11,888} = 83.35$, $\sigma_{22,888} = 28.56$, and $\rho_{888} = 0.546$. The 1-step ahead forecast for the covariance matrix is

$$\widehat{\boldsymbol{\Sigma}}_{888}(1) = \begin{bmatrix} 75.15 & 23.48 \\ 23.48 & 24.70 \end{bmatrix},$$

where the forecast of the correlation coefficient is 0.545.

In the second method, we use the Cholesky decomposition of $\boldsymbol{\Sigma}_t$ to model time-varying correlations. For the bivariate case, the parameter vector is $\boldsymbol{\Xi}_t = (g_{11,t}, g_{22,t}, q_{21,t})'$; see Eq. (9.13). A simple GARCH(1, 1)-type model for $\boldsymbol{a}_t$ is

$$\begin{aligned} g_{11,t} &= \alpha_{10} + \alpha_{11} b_{1,t-1}^2 + \beta_{11} g_{11,t-1} \\ q_{21,t} &= \gamma_0 + \gamma_1 q_{21,t-1} + \gamma_2 a_{2,t-1} \\ g_{22,t} &= \alpha_{20} + \alpha_{21} b_{1,t-1}^2 + \alpha_{22} b_{2,t-1}^2 + \beta_{21} g_{11,t-1} + \beta_{22} g_{22,t-1}, \end{aligned} \tag{9.26}$$

where $b_{1t} = a_{1t}$ and $b_{2t} = a_{2t} - q_{21,t} a_{1t}$. Thus, $b_{1t}$ assumes a univariate GARCH(1, 1) model, $b_{2t}$ uses a bivariate GARCH(1, 1) model, and $q_{21,t}$ is auto-correlated and uses $a_{2,t-1}$ as an additional explanatory variable. The probability density function relevant to maximum likelihood estimation is given in Eq. (9.15) with $k = 2$.

**Example 9.2.** (continued). Again we use the monthly log returns of IBM stock and the S&P 500 index to demonstrate the volatility model in Eq. (9.26). Using the same specification as before, we obtain the fitted mean equations as

$$\begin{aligned} r_{1t} &= 1.364 + 0.075 r_{1,t-1} - 0.058 r_{2,t-2} + a_{1t} \\ r_{2t} &= 0.643 + a_{2t}, \end{aligned}$$

where standard errors of the parameters in the first equation are 0.219, 0.027, and 0.032, respectively, and that of the parameter in the second equation is 0.154. These two mean equations are close to what we obtained before. The fitted volatility model is

$$\begin{aligned} g_{11,t} &= 3.714 + 0.113 b_{1,t-1}^2 + 0.804 g_{11,t-1} \\ q_{21,t} &= 0.0029 + 0.9915 q_{21,t-1} - 0.0041 a_{2,t-1} \\ g_{22,t} &= 1.023 + 0.021 b_{1,t-1}^2 + 0.052 b_{2,t-1}^2 - 0.040 g_{11,t-1} + 0.937 g_{22,t-1}, \end{aligned} \tag{9.27}$$

where $b_{1t} = a_{1t}$, $b_{2t} = a_{2t} - q_{21,t}b_{1t}$. Standard errors of the parameters in the equation of $g_{11,t}$ are 1.033, 0.022, and 0.037, respectively, those of the parameters in the equation of $q_{21,t}$ are 0.001, 0.002, and 0.0004, respectively, and those of the parameters in the equation of $g_{22,t}$ are 0.344, 0.007, 0.013, and 0.015, respectively. All estimates are statistically significant at the 1% level.

The conditional covariance matrix $\mathbf{\Sigma}_t$ can be obtained from model (9.27) by using the Cholesky decomposition in Eq. (9.7). For the bivariate case, the relationship is specifically given in Eq. (9.8). Consequently, we obtain the time-varying correlation coefficient as

$$\rho_t = \frac{\sigma_{21,t}}{\sqrt{\sigma_{11,t}\sigma_{22,t}}} = \frac{q_{21,t}\sqrt{g_{11,t}}}{\sqrt{g_{22,t} + q_{21,t}^2 g_{11,t}}}. \tag{9.28}$$

Using the fitted values of $\sigma_{11,t}$ and $\sigma_{22,t}$, we can compute the standardized residuals to perform model checking. The Ljung–Box statistics for the standardized residuals of model (9.27) give $Q(4) = 19.77(0.14)$ and $Q(8) = 34.22(0.27)$. For the squared standardized residuals, we have $Q(4) = 15.34(0.36)$ and $Q(8) = 31.87(0.37)$. Thus, the fitted model is adequate in describing the conditional mean and volatility. The model shows a strong dynamic dependence in the correlation; see the coefficient 0.9915 in Eq. (9.27).

Figure 9.7 shows the fitted time-varying correlation coefficient in Eq. (9.28). It shows a smoother correlation pattern than that of Figure 9.6 and confirms the decreasing trend of the correlation coefficient. In particular, the fitted correlation coefficients in recent years are smaller than those of the other models. The two time-varying correlation models for the monthly log returns of IBM stock and the S&P 500 index have comparable maximized likelihood functions of about $-3672$, indicating the fits are similar. However, the approach based on the Cholesky decomposition may have some advantages. First, it does not require any parameter constraint in estimation to ensure the positive definiteness of $\mathbf{\Sigma}_t$. If one also uses log transformation for $g_{ii,t}$, then no constraints are needed for the entire volatility model. Second, the log likelihood function becomes simple under the transformation. Third, the time-varying parameters $q_{ij,t}$ and $g_{ii,t}$ have nice interpretations. However, the transformation makes inference a bit more complicated because the fitted model may depend on the ordering of elements in $\boldsymbol{a}_t$; recall that $a_{1t}$ is not transformed. In theory, the ordering of elements in $\boldsymbol{a}_t$ should have no impact on volatility.

Finally, the 1-step ahead forecast of the conditional covariance matrix at the forecast origin $t = 888$ for the new time-varying correlation model is

$$\widehat{\mathbf{\Sigma}}_{888}(1) = \begin{bmatrix} 73.45 & 7.34 \\ 7.34 & 17.87 \end{bmatrix}.$$

The correlation coefficient of the prior forecast is 0.203, which is substantially smaller than those of the previous two models. However, forecasts of the conditional variances are similar as before.

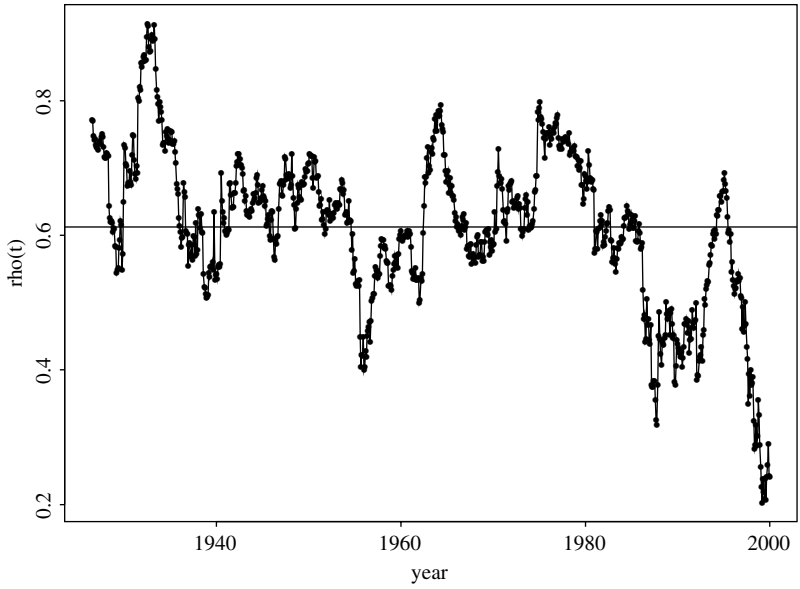


**Figure 9.7.** The fitted conditional correlation coefficient between monthly log returns of IBM stock and the S&P 500 index using the time-varying correlation GARCH(1, 1) model of Example 9.2 with Cholesky decomposition. The horizontal line denotes the average 0.612 of the estimated coefficients.

***Remark:*** In a recent manuscript, Tse and Tsui (1998) consider a multivariate GARCH model with time-varying correlations. For a $k$-dimensional returns, these authors assume that the conditional correlation matrix $\boldsymbol{\rho}_t$ follows the model

$$\boldsymbol{\rho}_t = (1 - \theta_1 - \theta_2)\boldsymbol{\rho} + \theta_1 \boldsymbol{\rho}_{t-1} + \theta_2 \boldsymbol{\psi}_{t-1},$$

where $\theta_1$ and $\theta_2$ are scalar parameters, $\boldsymbol{\rho}$ is a $k \times k$ positive definite matrix with unit diagonal elements, and $\boldsymbol{\psi}_{t-1}$ is the $k \times k$ sample correlation matrix using shocks from $t - m, \ldots, t - 1$ for a prespecified $m$. Estimation of the two scalar parameters $\theta_1$ and $\theta_2$ requires special constraints to ensure positive definiteness of the correlation matrix. This approach seems much more complicated than the two methods considered in this chapter.

## 9.3 HIGHER DIMENSIONAL VOLATILITY MODELS

In this section, we make use of the sequential nature of Cholesky decomposition to suggest a strategy for building a high-dimensional volatility model. Again, write the vector return series as $\boldsymbol{r}_t = \boldsymbol{\mu}_t + \boldsymbol{a}_t$. The mean equations for $\boldsymbol{r}_t$ can be specified by

using the methods of Chapter 8. A simple vector AR model is often sufficient. Here we focus on building a volatility model using the shock process $\boldsymbol{a}_t$.

Based on the discussion of Cholesky decomposition in Section 9.1, the orthogonal transformation from $a_{it}$ to $b_{it}$ only involves $b_{jt}$ for $j < i$. In addition, the time-varying volatility models built in Section 9.2 appear to be nested in the sense that the model for $g_{ii,t}$ only depends on quantities related to $b_{jt}$ for $j < i$. Consequently, we consider the following sequential procedure to build a multivariate volatility model:

1. Select a market index or a stock return that is of major interest. Build a univariate volatility model for the selected return series.
2. Augment a second return series to the system, perform the orthogonal transformation on the shock process of this new return series, and build a bivariate volatility model for the system. The parameter estimates of the univariate model in Step 1 can be used as the starting values in bivariate estimation.
3. Augment a third return series to the system, perform the orthogonal transformation on this newly added shock process, and build a three-dimensional volatility model. Again parameter estimates of the bivariate model can be used as the starting values in the three-dimensional estimation.
4. Continue the augmentation until a joint volatility model is built for all the return series of interest.

Finally, model checking should be performed in each step to ensure the adequacy of the fitted model. Experience shows that this sequential procedure can simplify substantially the complexity involved in building a high-dimensional volatility model. In particular, it can markedly reduce the computing time in estimation.

**Example 9.3.** We demonstrate the proposed sequential procedure by building a volatility model for the daily log returns of S&P 500 index and the stocks of Cisco Systems and Intel Corporation. The data span is from January 2, 1991 to December 31, 1999 with 2275 observations. The log returns are in percentages and shown in Figure 9.8. Components of the return series are ordered as $\boldsymbol{r}_t = (\text{SP5}_t, \text{CSCO}_t, \text{INTC}_t)'$. The sample means, standard errors, and correlation matrix of the data are

$$\widehat{\boldsymbol{\mu}} = \begin{bmatrix} 0.066 \\ 0.257 \\ 0.156 \end{bmatrix}, \quad \begin{bmatrix} \hat{\sigma}_1 \\ \hat{\sigma}_2 \\ \hat{\sigma}_3 \end{bmatrix} = \begin{bmatrix} 0.875 \\ 2.853 \\ 2.464 \end{bmatrix}, \quad \widehat{\boldsymbol{\rho}} = \begin{bmatrix} 1.00 & 0.52 & 0.50 \\ 0.52 & 1.00 & 0.47 \\ 0.50 & 0.47 & 1.00 \end{bmatrix}.$$

Using the Ljung–Box statistics to detect any serial dependence in the return series, we obtain $Q(1) = 26.20$, $Q(4) = 79.73$, and $Q(8) = 123.68$. These test statistics are highly significant with $p$ values close to zero as compared with chi-squared distributions with degrees of freedom 9, 36, and 72, respectively. There is indeed some serial dependence in the data. Table 9.1 gives the first five lags of sample cross-correlation matrixes shown in the simplified notation of Chapter 8. An examination

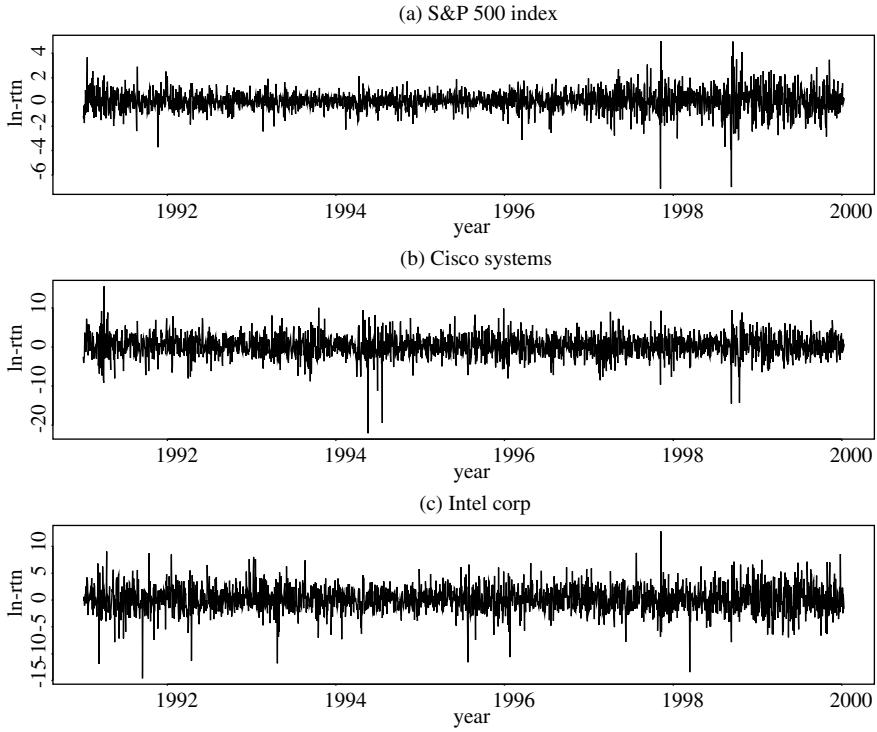


**Figure 9.8.** Time plots of daily log returns in percentages of the S&P 500 index and stocks of Cisco Systems and Intel Corporation from January 2, 1991 to December 31, 1999.

of the table shows that (a) the daily log return of S&P 500 index does not depend on the past returns of Cisco or Intel, (b) the log return of Cisco stock has some serial correlations and depends on the past return of S&P 500 index (see lags 2 and 5), and (c) the log return of Intel stock depends on the past returns of S&P 500 index (see lags 1 and 5). These observations are similar to that between the returns of IBM stock and S&P 500 index analyzed in Chapter 8. They suggest that returns of individual

**Table 9.1. Sample Cross-Correlation Matrixes of Daily Log Returns of the S&P 500 Index and the Stocks of Cisco Systems and Intel Corporation from January 2, 1991 to December 31, 1999.**

| Lag | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | 2 | | | 3 | | | 4 | | | 5 | | | 6 | | |
| . | . | . | . | . | . | − | . | . | . | . | . | − | . | . | . | . | . |
| . | . | . | . | − | . | . | . | . | . | . | . | − | . | . | . | − | . |
| − | . | . | . | . | . | . | . | . | . | . | . | − | . | . | . | . | . |

large-cap companies tend to be affected by the past behavior of the market. However, the market return is not significantly affected by the past returns of individual companies.

Turning to volatility modeling and following the suggested procedure, we start with the log return of S&P 500 index and obtain the model

$$\begin{aligned} r_{1t} &= 0.078 + 0.042r_{1,t-1} - 0.062r_{1,t-3} - 0.048r_{1,t-4} - 0.052r_{1,t-5} + a_{1t} \\ \sigma_{11,t} &= 0.013 + 0.092a_{1,t-1}^2 + 0.894\sigma_{11,t-1}, \end{aligned} \tag{9.29}$$

where standard errors of the parameters in the mean equation are 0.016, 0.023, 0.020, 0.022, and 0.020, respectively, and those of the parameters in the volatility equation are 0.002, 0.006, and 0.007, respectively. Univariate Ljung–Box statistics of the standardized residuals and their squared series fail to detect any remaining serial correlation or conditional heteroscedasticity in the data. Indeed, we have $Q(10) = 7.38(0.69)$ for the standardized residuals and $Q(10) = 3.14(0.98)$ for the squared series.

Augmenting the daily log returns of Cisco stock to the system, we build a bivariate model with mean equations given by

$$\begin{aligned} r_{1t} &= 0.065 - 0.046r_{1,t-3} + a_{1t} \\ r_{2t} &= 0.325 + 0.195r_{1,t-2} - 0.091r_{2,t-2} + a_{2t}, \end{aligned} \tag{9.30}$$

where all of the estimates are statistically significant at the 1% level. Using the notation of Cholesky decomposition, we obtain the volatility equations as

$$\begin{aligned} g_{11,t} &= 0.006 + 0.051b_{1,t-1}^2 + 0.943g_{11,t-1} \\ q_{21,t} &= 0.331 + 0.790q_{21,t-1} - 0.041a_{2,t-1} \\ g_{22,t} &= 0.177 + 0.082b_{2,t-1}^2 + 0.890g_{22,t-1}, \end{aligned} \tag{9.31}$$

where $b_{1t} = a_{1t}$, $b_{2t} = a_{2t} - q_{21,t}b_{1t}$, standard errors of the parameters in the equation of $g_{11,t}$ are 0.001, 0.005, and 0.006, those of the parameters in the equation of $q_{21,t}$ are 0.156, 0.099, and 0.011, and those of the parameters in the equation of $g_{22,t}$ are 0.029, 0.008, and 0.011, respectively. The bivariate Ljung–Box statistics of the standardized residuals fail to detect any remaining serial dependence or conditional heteroscedasticity. The bivariate model is adequate. Comparing with Eq. (9.29), we see that the difference between the marginal and univariate models of $r_{1t}$ is small.

The next and final step is to augment the daily log returns of Intel stock to the system. The mean equations become

$$\begin{aligned} r_{1t} &= 0.065 - 0.043r_{1,t-3} + a_{1t} \\ r_{2t} &= 0.326 + 0.201r_{1,t-2} - 0.089r_{2,t-2} + a_{2t} \\ r_{3t} &= 0.192 - 0.264r_{1,t-1} + 0.059r_{3,t-1} + a_{3t}, \end{aligned} \tag{9.32}$$

where standard errors of the parameters in the first equation are 0.016 and 0.017, those of the parameters in the second equation are 0.052, 0.059, and 0.021, and those of the parameters in the third equation are 0.050, 0.057, and 0.022, respectively. All estimates are statistically significant at about the 1% level. As expected, the mean equations for $r_{1t}$ and $r_{2t}$ are essentially the same as those in the bivariate case.

The three-dimensional time-varying volatility model becomes a bit more complicated, but it remains manageable as

$$\begin{aligned}
g_{11,t} &= 0.006 + 0.050b_{1,t-1}^2 + 0.943g_{11,t-1} \\
q_{21,t} &= 0.277 + 0.824q_{21,t-1} - 0.035a_{2,t-1} \\
g_{22,t} &= 0.178 + 0.082b_{2,t-1}^2 + 0.889g_{22,t-1} \\
q_{31,t} &= 0.039 + 0.973q_{31,t-1} + 0.010a_{3,t-1} \\
q_{32,t} &= 0.006 + 0.981q_{32,t-1} + 0.004a_{2,t-1} \\
g_{33,t} &= 1.188 + 0.053b_{3,t-1}^2 + 0.687g_{33,t-1} - 0.019g_{22,t-1},
\end{aligned} \tag{9.33}$$

where $b_{1t} = a_{1t}$, $b_{2t} = a_{2t} - q_{21,t}b_{1t}$, $b_{3t} = a_{3t} - q_{31,t}b_{1t} - q_{32,t}b_{2t}$, and standard errors of the parameters are given in Table 9.2. Except for the constant term of the $q_{32,t}$ equation, all estimates are significant at the 5% level. Let $\tilde{\boldsymbol{a}}_t = (a_{1t}/\hat{\sigma}_{1t}, a_{2t}/\hat{\sigma}_{2t}, a_{3t}/\hat{\sigma}_{3t})'$ be the standardized residual series, where $\hat{\sigma}_{it} = \sqrt{\hat{\sigma}_{ii,t}}$ is the fitted conditional standard error of the $i$th return. The Ljung–Box statistics of $\tilde{\boldsymbol{a}}_t$ give $Q(4) = 34.48(0.31)$ and $Q(8) = 60.42(0.70)$, where the degrees of freedom of the chi-squared distributions are 31 and 67, respectively, after adjusting for the number of parameters used in the mean equations. For the squared standardized residual series $\tilde{\boldsymbol{a}}_t^2$, we have $Q(4) = 28.71(0.58)$ and $Q(8) = 52.00(0.91)$. Therefore, the fitted model appears to be adequate in modeling the conditional means and volatilities.

The three-dimensional volatility model in Eq. (9.33) shows some interesting features. First, it is essentially a time-varying correlation GARCH(1, 1) model because only lag-1 variables are used in the equations. Second, the volatility of the daily log return of S&P 500 index does not depend on the past volatilities of Cisco or Intel stock return. Third, by taking the inverse transformation of the Cholesky decomposition, the volatilities of daily log returns of Cisco and Intel stocks depend on the past

**Table 9.2. Standard Errors of Parameter Estimates of a Three-Dimensional Volatility Model for the Daily Log Returns in Percentages of S&P 500 Index and Stocks of Cisco Systems and Intel Corporation from January 2, 1991 to December 31, 1999. The Ordering of the Parameter Is the Same As That Appears in Eq. (9.33).**

| Equation | Standard error | | | | Equation | Standard error | | |
|---|---|---|---|---|---|---|---|---|
| $g_{11,t}$ | 0.001 | 0.005 | 0.006 | | $q_{21,t}$ | 0.135 | 0.086 | 0.010 |
| $g_{22,t}$ | 0.029 | 0.009 | 0.011 | | $q_{31,t}$ | 0.017 | 0.012 | 0.004 |
| $g_{33,t}$ | 0.407 | 0.015 | 0.100 | 0.008 | $q_{32,t}$ | 0.004 | 0.013 | 0.001 |

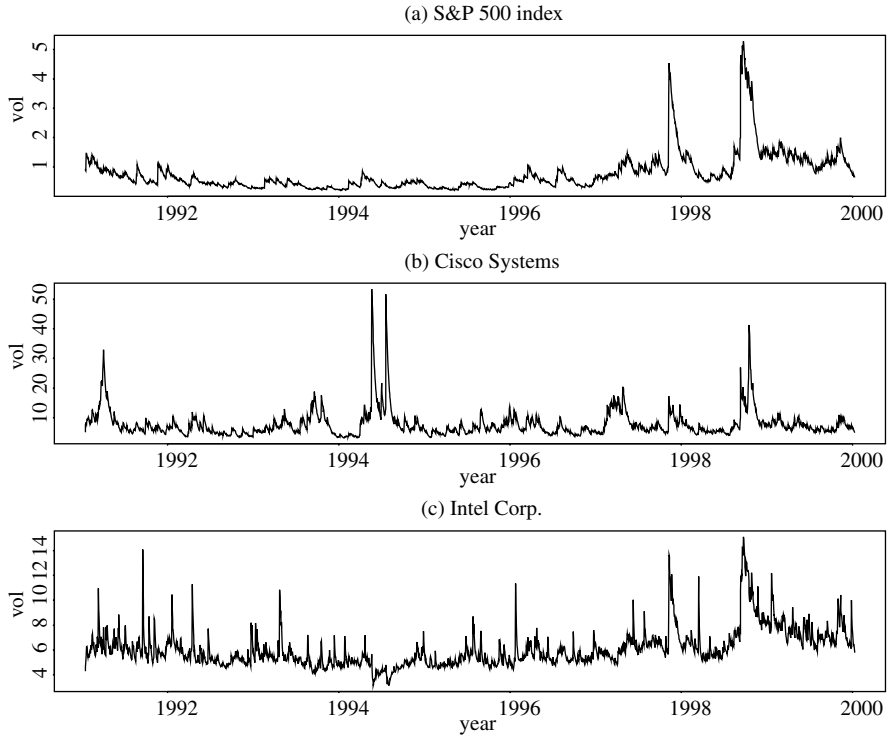


**Figure 9.9.** Time plots of fitted volatilities for daily log returns, in percentages, of the S&P 500 index and stocks of Cisco Systems and Intel Corporation from January 2, 1991 to December 31, 1999.

volatility of the market return; see the relationships between elements of $\boldsymbol{\Sigma}_t$, $\boldsymbol{L}_t$, and $\boldsymbol{G}_t$ given in Section 9.1. Fourth, the correlation quantities $q_{ij,t}$ have high persistence with large AR(1) coefficients.

Figure 9.9 shows the fitted volatility processes of the model (i.e., $\hat{\sigma}_{ii,t}$) for the data. The volatility of the index return is much smaller than those of the two individual stock returns. The plots also show that the volatility of the index return has increased in recent years, but this is not the case for the return of Cisco Systems. Figure 9.10 shows the time-varying correlation coefficients between the three return series. Of particular interest is to compare Figures 9.9 and 9.10. They show that the correlation coefficient between two return series increases when the returns are volatile. This is in agreement with empirical study of relationships between international stock market indexes for which the correlation between two markets tends to increase during a financial crisis.

The volatility model in Eq. (9.33) consists of two sets of equations. The first set of equations describes the time evolution of conditional variances (i.e. $g_{ii,t}$), and the second set of equations deals with correlation coefficients (i.e. $q_{ij,t}$ with $i > j$). For this particular data set, an AR(1) model might be sufficient for the cor-

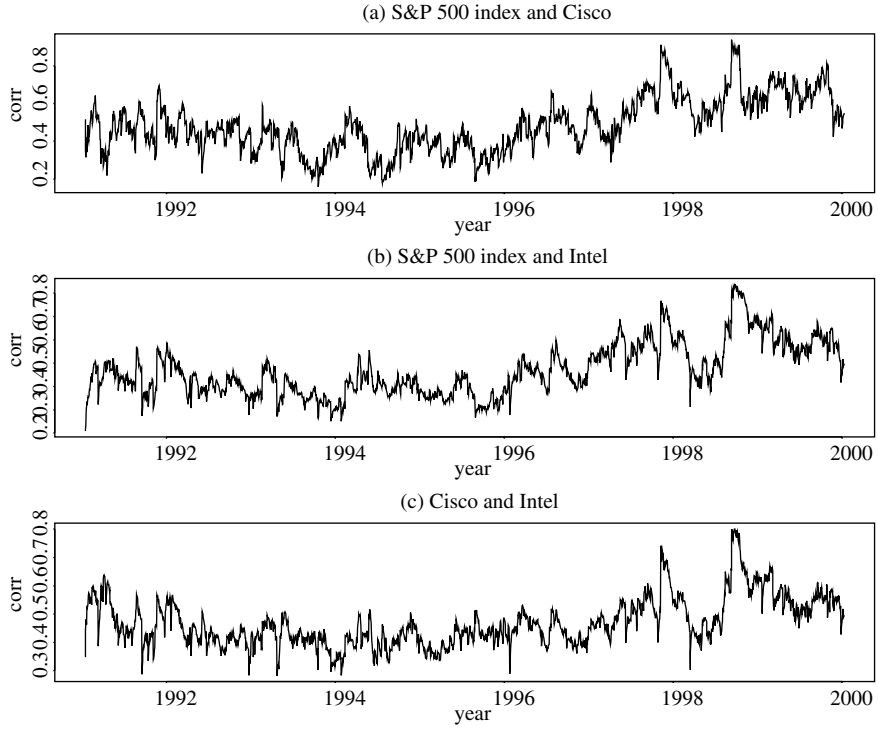


**Figure 9.10.** Time plots of fitted time-varying correlation coefficients between daily log returns of S&P 500 index and stocks of Cisco Systems and Intel Corporation from January 2, 1991 to December 31, 1999.

relation equations. Similarly, a simple AR model might also be sufficient for the conditional variances. Define $\boldsymbol{v}_t = (v_{11,t}, v_{22,t}, v_{33,t})'$, where $v_{ii,t} = \ln(g_{ii,t})$, and $\boldsymbol{q}_t = (q_{21,t}, q_{31,t}, q_{32,t})'$. The previous discussion suggests that we can use the simple lag-1 models

$$\boldsymbol{v}_t = \boldsymbol{c}_1 + \boldsymbol{\beta}_1 \boldsymbol{v}_{t-1}, \quad \boldsymbol{q}_t = \boldsymbol{c}_2 + \boldsymbol{\beta}_2 \boldsymbol{q}_{t-1}$$

as exact functions to model the volatility of asset returns, where $\boldsymbol{c}_i$ are constant vectors and $\boldsymbol{\beta}_i$ are $3 \times 3$ real-valued matrixes. If a noise term is also included in the prior equations, then the models become

$$\boldsymbol{v}_t = \boldsymbol{c}_1 + \boldsymbol{\beta}_1 \boldsymbol{v}_{t-1} + \boldsymbol{e}_{1t}, \quad \boldsymbol{q}_t = \boldsymbol{c}_2 + \boldsymbol{\beta}_2 \boldsymbol{q}_{t-1} + \boldsymbol{e}_{2t},$$

where $\boldsymbol{e}_{it}$ are random shocks with mean zero and a positive definite covariance matrix, and we have a simple multivariate stochastic volatility model. In a recent manuscript, Chib, Nardari, and Shephard (1999) use Markov Chain Monte Carlo (MCMC) methods to study high-dimensional stochastic volatility models. The model

considered there allows for time-varying correlations, but in a relatively restrictive manner. Additional references of multivariate volatility model include Harvey, Ruiz, and Shephard (1995). We discuss MCMC methods to volatility modeling in Chapter 10.

## 9.4 FACTOR-VOLATILITY MODELS

Another approach to simplifying the dynamic structure of a multivariate volatility process is to use factor models. In practice, the "common factors" can be determined *a priori* by substantive matter or empirical methods. As an illustration, we use the factor analysis of Chapter 8 to discuss factor-volatility models. Because volatility models are concerned with the evolution over time of the conditional covariance matrix of $\boldsymbol{a}_t$, where $\boldsymbol{a}_t = \boldsymbol{r}_t - \boldsymbol{\mu}_t$, a simple way to identify the "common factors" in volatility is to perform a principal component analysis (PCA) on $\boldsymbol{a}_t$; see the PCA of Chapter 8. Building a factor volatility model thus involves a three-step procedure:

- select the first few principal components that explain a high percentage of variability in $\boldsymbol{a}_t$,
- build a volatility model for the selected principal components, and
- relate the volatility of each $a_{it}$ series to the volatilities of the selected principal components.

The objective of such a procedure is to reduce the dimension, but maintain an accurate approximation of the multivariate volatility.

**Example 9.4.** Consider again the monthly log returns, in percentages, of IBM stock and the S&P 500 index of Example 9.2. Using the bivariate AR(3) model of Example 8.4, we obtain an innovational series $\boldsymbol{a}_t$. Performing a PCA on $\boldsymbol{a}_t$ based on its covariance matrix, we obtained eigenvalues 63.373 and 13.489. The first eigenvalue explains 82.2% of the generalized variance of $\boldsymbol{a}_t$. Therefore, we may choose the first principal component $x_t = 0.797a_{1t} + 0.604a_{2t}$ as the common factor. Alternatively, as shown by the model in Example 8.4, the serial dependence in $\boldsymbol{r}_t$ is weak and hence, one can perform the PCA on $\boldsymbol{r}_t$ directly. For this particular instance, the two eigenvalues of the sample covariance matrix of $\boldsymbol{r}_t$ are 63.625 and 13.513, which are essentially the same as those based on $\boldsymbol{a}_t$. The first principal component explains approximately 82.5% of the generalized variance of $\boldsymbol{r}_t$, and the corresponding common factor is $x_t = 0.796r_{1t} + 0.605r_{2t}$. Consequently, for the two monthly log return series considered, the effect of the conditional mean equations on PCA is negligible.

Based on the prior discussion and for simplicity, we use $x_t = 0.796r_{1t} + 0.605r_{2t}$ as a common factor for the two monthly return series. Figure 9.11(a) shows the time plot of this common factor. If univariate Gaussian GARCH models are entertained, we obtain the following model for $x_t$:

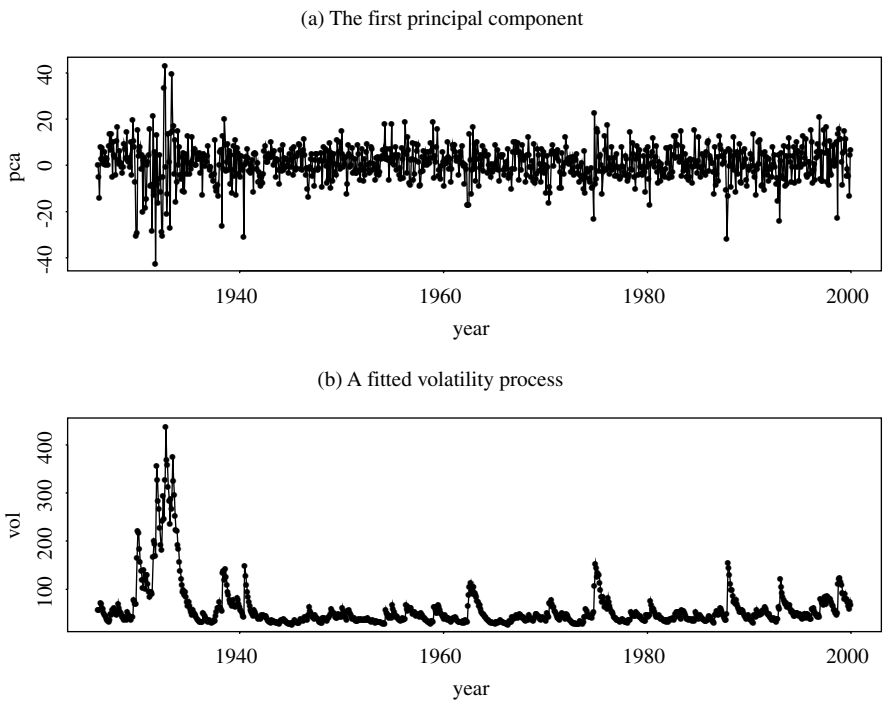


**Figure 9.11.** (a) Time plot of the first principal component of the monthly log returns of IBM stock and the S&P 500 index. (b) The fitted volatility process based on a GARCH(1, 1) model.

$$x_t = 1.317 + 0.096x_{t-1} + a_t, \quad a_t = \sigma_t \epsilon_t$$
$$\sigma_t^2 = 3.834 + 0.110a_{t-1}^2 + 0.825\sigma_{t-1}^2. \tag{9.34}$$

All parameter estimates of the previous model are highly significant at the 1% level, and the Ljung–Box statistics of the standardized residuals and their squared series fail to detect any model inadequacy. Figure 9.11(b) shows the fitted volatility of $x_t$ [i.e., the sample $\sigma_t^2$ series in Eq. (9.34)].

Using $\sigma_t^2$ of model (9.34) as a common volatility factor, we obtain the following model for the original monthly log returns. The mean equations are

$$r_{1t} = 1.140 + 0.079r_{1,t-1} + 0.067r_{1,t-2} - 0.122r_{2,t-2} + a_{1t}$$
$$r_{2t} = 0.537 + a_{2t},$$

where standard errors of the parameters in the first equation are 0.211, 0.030, 0.031, and 0.043, respectively, and that of the parameter in the second equation is 0.165. The conditional variance equation is

$$\begin{bmatrix}\sigma_{11,t}\\ \sigma_{22,t}\end{bmatrix} = \begin{bmatrix}19.08\\ (3.70)\\ -5.62\\ (2.36)\end{bmatrix} + \begin{bmatrix}0.098 & \cdot\\ (0.044) & \\ \cdot & \cdot\end{bmatrix}\begin{bmatrix}a_{1,t-1}^2\\ a_{2,t-1}^2\end{bmatrix} + \begin{bmatrix}0.333\\ (0.076)\\ 0.596\\ (0.050)\end{bmatrix}\sigma_t^2, \tag{9.35}$$

where, as before, standard errors are in parentheses, and $\sigma_t^2$ is obtained from model (9.34). The conditional correlation equation is

$$\rho_t = \frac{\exp(q_t)}{1+\exp(q_t)}, \quad q_t = -2.098 + 4.120\rho_{t-1} + 0.078\frac{a_{1,t-1}a_{2,t-1}}{\sqrt{\sigma_{11,t-1}\sigma_{22,t-1}}}, \tag{9.36}$$

where standard errors of the three parameters are 0.025, 0.038, and 0.015, respectively. Defining the standardized residuals as before, we obtain $Q(4) = 15.37(0.29)$ and $Q(8) = 34.24(0.23)$, where the number in parentheses denotes $p$ value. Therefore, the standardized residuals have no serial correlations. Yet we have $Q(4) = 20.25(0.09)$ and $Q(8) = 61.95(0.0004)$ for the squared standardized residuals. The volatility model in Eq. (9.35) does not adequately handle the conditional heteroscedasticity of the data especially at higher lags. This is not surprising as the single common factor only explains about 82.5% of the generalized variance of the data.

Comparing the factor model in Eqs. (9.35) and (9.36) with the time-varying correlation model in Eqs. (9.24) and (9.25), we see that (a) the correlation equations of the two models are essential the same, (b) as expected the factor model uses fewer parameters in the volatility equation, and (c) the common-factor model provides a reasonable approximation to the volatility process of the data.

***Remark:*** In Example 9.4, we used a two-step estimation procedure. In the first step, a volatility model is built for the common factor. The estimated volatility is treated as given in the second step to estimate the multivariate volatility model. Such an estimation procedure is simple, but may not be efficient. A more efficient estimation procedure is to perform a joint estimation. This can be done relatively easily provided that the common factors are known. For example, for the monthly log returns of Example 9.4, a joint estimation of Eqs. (9.34)–(9.36) can be performed if the common factor $x_t = 0.769r_{1t} + 0.605r_{2t}$ is treated as given.

## 9.5 APPLICATION

We illustrate the application of multivariate volatility models by considering the Value at Risk (VaR) of a financial position with multiple assets. Suppose that an investor holds a long position in the stocks of Cisco Systems and Intel Corporation each worth \$1 million. We use the daily log returns for the two stocks from January 2, 1991 to December 31, 1999 to build volatility models. The VaR is computed using the 1-step ahead forecasts at the end of data span and 5% critical values.

Let $\text{VaR}_1$ be the value at risk for holding the position on Cisco Systems stock and $\text{VaR}_2$ for holding Intel stock. Results of Chapter 7 show that the overall daily VaR for the investor is

$$\text{VaR} = \sqrt{\text{VaR}_1^2 + \text{VaR}_2^2 + 2\rho \text{VaR}_1 \text{VaR}_2}.$$

In this illustration, we consider three approaches to volatility modeling for calculating VaR. For simplicity, we do not report standard errors for the parameters involved or model checking statistics. Yet all of the estimates are statistically significant at the 5% level and the models are adequate based on the Ljung–Box statistics of the standardized residual series and their squared series. The log returns are in percentages so that the quantiles are divided by 100 in VaR calculation. Let $r_{1t}$ be the return of Cisco stock and $r_{2t}$ the return of Intel stock.

***(A) Univariate Models***

This approach uses a univariate volatility model for each stock return and uses the sample correlation coefficient of the stock returns to estimate $\rho$. The univariate volatility models for the two stock returns are

$$\begin{aligned} r_{1t} &= 0.380 + 0.034 r_{1,t-1} - 0.061 r_{1,t-2} - 0.055 r_{1,t-3} + a_{1t} \\ \sigma_{1t}^2 &= 0.599 + 0.117 a_{1,t-1}^2 + 0.814 \sigma_{1,t-1}^2 \\ r_{2t} &= 0.187 + a_{2t} \\ \sigma_{2t}^2 &= 0.310 + 0.032 a_{2,t-1}^2 + 0.918 \sigma_{2,t-1}^2. \end{aligned}$$

The sample correlation coefficient is 0.473. The 1-step ahead forecasts needed in VaR calculation at the forecast origin $t = 2275$ are

$$\hat{r}_1 = 0.626, \quad \hat{\sigma}_1^2 = 4.152, \quad \hat{r}_2 = 0.187, \quad \hat{\sigma}_2^2 = 6.087, \quad \hat{\rho} = 0.473.$$

The 5% quantiles for both daily returns are

$$q_1 = 0.626 - 1.65\sqrt{4.152} = -2.736, \quad q_2 = 0.187 - 1.65\sqrt{6.087} = -3.884,$$

where the negative sign denotes loss. The VaR for the individual stocks are $\text{VaR}_1 = \$1000000 q_1/100 = \$27360$ and $\text{VaR}_2 = \$1000000 q_2/100 = \$38840$. Consequently, the overall VaR for the investor is VaR = \$57117.

***(B) Constant Correlation Bivariate Model***

This approach employs a bivariate GARCH(1, 1) model for the stock returns. The correlation coefficient is assumed to be constant over time, but it is estimated jointly with other parameters. The model is

$$\begin{aligned} r_{1t} &= 0.385 + 0.038 r_{1,t-1} - 0.060 r_{1,t-2} - 0.047 r_{1,t-3} + a_{1t} \\ r_{2t} &= 0.222 + a_{2t} \end{aligned}$$

$$\sigma_{11,t} = 0.624 + 0.110a_{1,t-1}^2 + 0.816\sigma_{11,t-1}$$
$$\sigma_{22,t} = 0.664 + 0.038a_{2,t-1}^2 + 0.853\sigma_{22,t-1}$$

and $\hat{\rho} = 0.475$. This is a diagonal bivariate GARCH(1, 1) model. The 1-step ahead forecasts for VaR calculation at the forecast origin $t = 2275$ are

$$\hat{r}_1 = 0.373, \quad \hat{\sigma}_1^2 = 4.287, \quad \hat{r}_2 = 0.222, \quad \hat{\sigma}_2^2 = 5.706, \quad \hat{\rho} = 0.475.$$

Consequently, we have $\text{VaR}_1 = \$30432$ and $\text{VaR}_2 = \$37195$. The overall 5% VaR for the investor is VaR = \$58180.

***(C) Time-Varying Correlation Model***

Finally, we allow the correlation coefficient to evolve over time by using the Cholesky decomposition. The fitted model is

$$r_{1t} = 0.355 + 0.039r_{1,t-1} - 0.057r_{1,t-2} - 0.038r_{1,t-3} + a_{1t}$$
$$r_{2t} = 0.206 + a_{2t}$$
$$g_{11,t} = 0.420 + 0.091b_{1,t-1}^2 + 0.858g_{11,t-1}$$
$$q_{21,t} = 0.123 + 0.689q_{21,t-1} - 0.014a_{2,t-1}$$
$$g_{22,t} = 0.080 + 0.013b_{2,t-1}^2 + 0.971g_{22,t-1},$$

where $b_{1t} = a_{1t}$ and $b_{2t} = a_{2t} - q_{21,t}a_{1t}$. The 1-step ahead forecasts for VaR calculation at the forecast origin $t = 2275$ are

$$\hat{r}_1 = 0.352, \quad \hat{r}_2 = 0.206, \quad \hat{g}_{11} = 4.252, \quad \hat{q}_{21} = 0.421, \quad \hat{g}_{22} = 5.594.$$

Therefore, we have $\hat{\sigma}_1^2 = 4.252$, $\hat{\sigma}_{21} = 1.791$ and $\hat{\sigma}_2^2 = 6.348$. The correlation coefficient is $\hat{\rho} = 0.345$. Using these forecasts, we have $\text{VaR}_1 = \$30504$, $\text{VaR}_2 = \$39512$, and the overall value at risk VaR = \$57648.

The estimated VaR of the three approaches are similar. The univariate models give the lowest VaR, whereas the constant-correlation model produces the highest VaR. The range of the difference is about \$1100. The time-varying volatility model seems to produce a compromise between the two extreme models.

## 9.6 MULTIVARIATE $t$ DISTRIBUTION

Empirical analysis indicates that the multivariate Gaussian innovations used in the previous sections may fail to capture the kurtosis of asset returns. In this situation, multivariate Student-$t$ distribution might be useful. There are many versions of multivariate Student-$t$ distribution. We give a simple version here for volatility modeling.

A $k$-dimensional random vector $\boldsymbol{x} = (x_1, \ldots, x_k)'$ has a multivariate Student-$t$ distribution with $v$ degrees of freedom and parameters $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$ (the

indentify matrix) if its probability density function (pdf) is

$$f(\boldsymbol{x} \mid v) = \frac{\Gamma((v+k)/2)}{(\pi v)^{k/2}\Gamma(v/2)}(1 + v^{-1}\boldsymbol{x}'\boldsymbol{x})^{-(v+k)/2}, \tag{9.37}$$

where $\Gamma(y)$ is the Gamma function; see Mardia, Kent, and Bibby (1979, p. 57). The variance of each component $x_i$ in Eq. (9.37) is $v/(v-2)$ and hence we define $\boldsymbol{\epsilon}_t = \sqrt{(v-2)/v}\,\boldsymbol{x}$ as the standardized multivariate Student-$t$ distribution with $v$ degrees of freedom. By transformation, the pdf of $\boldsymbol{\epsilon}_t$ is

$$f(\boldsymbol{\epsilon}_t \mid v) = \frac{\Gamma((v+k)/2)}{[\pi(v-2)]^{k/2}\Gamma(v/2)}[1 + (v-2)^{-1}\boldsymbol{\epsilon}_t'\boldsymbol{\epsilon}_t]^{-(v+k)/2}. \tag{9.38}$$

For volatility modeling, we write $\boldsymbol{a}_t = \boldsymbol{\Sigma}_t^{1/2}\boldsymbol{\epsilon}_t$ and assume that $\boldsymbol{\epsilon}_t$ follows the multivariate Student-$t$ distribution in Eq. (9.38). By transformation, the pdf of $\boldsymbol{a}_t$ is

$$f(\boldsymbol{a}_t \mid v, \boldsymbol{\Sigma}_t) = \frac{\Gamma((v+k)/2)}{[\pi(v-2)]^{k/2}\Gamma(v/2)|\,\boldsymbol{\Sigma}_t\,|^{1/2}}[1 + (v-2)^{-1}\boldsymbol{a}_t'\boldsymbol{\Sigma}_t^{-1}\boldsymbol{a}_t]^{-(v+k)/2}.$$

Furthermore, if we use the Cholesky decomposition of $\boldsymbol{\Sigma}_t$, then the pdf of the transformed shock $\boldsymbol{b}_t$ becomes

$$f(\boldsymbol{b}_t \mid v, \boldsymbol{L}_t, \boldsymbol{G}_t) = \frac{\Gamma((v+k)/2)}{[\pi(v-2)]^{k/2}\Gamma(v/2)\prod_{j=1}^{k} g_{jj,t}^{1/2}}\left[1 + (v-2)^{-1}\sum_{j=1}^{k}\frac{b_{jt}^2}{g_{jj,t}}\right]^{-(v+k)/2},$$

where $\boldsymbol{a}_t = \boldsymbol{L}_t\boldsymbol{b}_t$ and $g_{jj,t}$ is the conditional variance of $b_{jt}$. Because this pdf does not involve any matrix inversion, the conditional likelihood function of the data is easy to evaluate.

## APPENDIX A. SOME REMARKS ON ESTIMATION

The estimation of multivariate ARMA models in this chapter is done by using the time series program SCA of Scientific Computing Associates. The estimation of multivariate volatility models is done by using the Regression Analysis for Time Series (RATS) program. Below are some runstreams for estimating multivariate volatility models using the RATS program. A line starting with "*" means "comment" only.

**(A):** Estimation of the diagonal constant-correlation AR(2)-GARCH(1, 1) model for Example 9.2. The program includes some Ljung–Box statistics for each component and some fitted values for the last few observations. The data file is "m-ibmspln.dat," which has two columns, and there are 888 observations.

```
all 0 888:1
open data m-ibmspln.dat
data(org=obs) / r1 r2
set h1 = 0.0
set h2 = 0.0
nonlin a0 a1 b1 a00 a11 b11 rho c1 c2 p1
frml a1t = r1(t)-c1-p1*r2(t-1)
frml a2t = r2(t)-c2
frml gvar1 = a0+a1*a1t(t-1)**2+b1*h1(t-1)
frml gvar2 = a00+a11*a2t(t-1)**2+b11*h2(t-1)
frml gdet = -0.5*(log(h1(t)=gvar1(t))+log(h2(t)=gvar2(t)) $
 +log(1.0-rho**2))
frml gln = gdet(t)-0.5/(1.0-rho**2)*((a1t(t)**2/h1(t)) $
 +(a2t(t)**2/h2(t))-2*rho*a1t(t)*a2t(t)/sqrt(h1(t)*h2(t)))
smpl 3 888
compute c1 = 1.22, c2 = 0.57, p1 = 0.1, rho = 0.1
compute a0 = 3.27, a1 = 0.1, b1 = 0.6
compute a00 = 1.17, a11 = 0.13, b11 = 0.8
maximize(method=bhhh,recursive,iterations=150) gln
set fv1 = gvar1(t)
set resi1 = a1t(t)/sqrt(fv1(t))
set residsq = resi1(t)*resi1(t)
* Checking standardized residuals *
cor(qstats,number=12,span=4) resi1
* Checking squared standardized residuals *
cor(qstats,number=12,span=4) residsq
set fv2 = gvar2(t)
set resi2 = a2t(t)/sqrt(fv2(t))
set residsq = resi2(t)*resi2(t)
* Checking standardized residuals *
cor(qstats,number=12,span=4) resi2
* Checking squared standardized residuals *
cor(qstats,number=12,span=4) residsq
* Last few observations needed for computing forecasts *
set shock1 = a1t(t)
set shock2 = a2t(t)
print 885 888 shock1 shock2 fv1 fv2
```

**(B):** Estimation of the time-varying correlation model in Example 9.2.

```
all 0 888:1
open data m-ibmspln.dat
data(org=obs) / r1 r2
set h1 = 45.0
set h2 = 31.0
set rho = 0.8
nonlin a0 a1 b1 f1 a00 a11 b11 d11 f11 c1 c2 p1 p3 q0 q1 q2
frml a1t = r1(t)-c1-p1*r1(t-1)-p3*r2(t-2)
frml a2t = r2(t)-c2
frml gvar1 = a0+a1*a1t(t-1)**2+b1*h1(t-1)+f1*h2(t-1)
frml gvar2 = a00+a11*a2t(t-1)**2+b11*h2(t-1)+f11*h1(t-1) $
 +d11*a1t(t-1)**2
frml rh1 = q0 + q1*rho(t-1) $
 + q2*a1t(t-1)*a2t(t-1)/sqrt(h1(t-1)*h2(t-1))
```

```
frml rh = exp(rh1(t))/(1+exp(rh1(t)))
frml gdet = -0.5*(log(h1(t)=gvar1(t))+log(h2(t)=gvar2(t)) $
 +log(1.0-(rho(t)=rh(t))**2))
frml gln = gdet(t)-0.5/(1.0-rho(t)**2)*((a1t(t)**2/h1(t)) $
 +(a2t(t)**2/h2(t))-2*rho(t)*a1t(t)*a2t(t)/sqrt(h1(t)*h2(t)))
smpl 4 888
compute c1 = 1.4, c2 = 0.7, p1 = 0.1, p3 = -0.1
compute a0 = 2.95, a1 = 0.08, b1 = 0.87, f1 = -.03
compute a00 = 2.05, a11 = 0.05
compute b11 = 0.92, f11=-.06, d11=.04, q0 = -2.0
compute q1 = 3.0, q2 = 0.1
nlpar(criterion=value,cvcrit=0.00001)
maximize(method=bhhh,recursive,iterations=150) gln
set fv1 = gvar1(t)
set resi1 = a1t(t)/sqrt(fv1(t))
set residsq = resi1(t)*resi1(t)
* Checking standardized residuals *
cor(qstats,number=16,span=4) resi1
* Checking squared standardized residuals *
cor(qstats,number=16,span=4) residsq
set fv2 = gvar2(t)
set resi2 = a2t(t)/sqrt(fv2(t))
set residsq = resi2(t)*resi2(t)
* Checking standardized residuals *
cor(qstats,number=16,span=4) resi2
* Checking squared standardized residuals *
cor(qstats,number=16,span=4) residsq
* Last few observations needed for computing forecasts *
set rhohat = rho(t)
set shock1 = a1t(t)
set shock2 = a2t(t)
print 885 888 shock1 shock2 fv1 fv2 rhohat
```

**(C):** Estimation of the time-varying correlation model in Example 9.2 using Cholesky decomposition.

```
all 0 888:1
open data m-ibmspln.dat
data(org=obs) / r1 r2
set h1 = 45.0
set h2 = 20.0
set q = 0.8
nonlin a0 a1 b1 a00 a11 b11 d11 f11 c1 c2 p1 p3 t0 t1 t2
frml a1t = r1(t)-c1-p1*r1(t-1)-p3*r2(t-2)
frml a2t = r2(t)-c2
frml v1 = a0+a1*a1t(t-1)**2+b1*h1(t-1)
frml qt = t0 + t1*q(t-1) + t2*a2t(t-1)
frml bt = a2t(t) - (q(t)=qt(t))*a1t(t)
frml v2 = a00+a11*bt(t-1)**2+b11*h2(t-1)+f11*h1(t-1) $
 +d11*a1t(t-1)**2
frml gdet = -0.5*(log(h1(t) = v1(t))+ log(h2(t)=v2(t)))
frml garchln = gdet-0.5*(a1t(t)**2/h1(t)+bt(t)**2/h2(t))
smpl 5 888
compute c1 = 1.4, c2 = 0.7, p1 = 0.1, p3 = -0.1
```

```
compute a0 = 1.0, a1 = 0.08, b1 = 0.87
compute a00 = 2.0, a11 = 0.05, b11 = 0.8
compute d11=.04, f11=-.06, t0 =0.2, t1 = 0.1, t2 = 0.1
nlpar(criterion=value,cvcrit=0.00001)
maximize(method=bhhh,recursive,iterations=150) garchln
set fv1 = v1(t)
set resi1 = a1t(t)/sqrt(fv1(t))
set residsq = resi1(t)*resi1(t)
* Checking standardized residuals *
cor(qstats,number=16,span=4) resi1
* Checking squared standardized residuals *
cor(qstats,number=16,span=4) residsq
set fv2 = v2(t)+qt(t)**2*v1(t)
set resi2 = a2t(t)/sqrt(fv2(t))
set residsq = resi2(t)*resi2(t)
* Checking standardized residuals *
cor(qstats,number=16,span=4) resi2
* Checking squared standardized residuals *
cor(qstats,number=16,span=4) residsq
* Last few observations needed for forecasts *
set rhohat = qt(t)*sqrt(v1(t)/fv2(t))
set shock1 = a1t(t)
set shock2 = a2t(t)
set g22 = v2(t)
set q21 = qt(t)
set b2t = bt(t)
print 885 888 shock1 shock2 fv1 fv2 rhohat g22 q21 b2t
```

**(D):** Estimation of the three-dimensional time-varying correlation volatility model in Example 9.3 using Cholesky decomposition. Initial estimates are obtained by a sequential modeling procedure.

```
all 0 2275:1
open data d-cscointc.dat
data(org=obs) / r1 r2 r3
set h1 = 1.0
set h2 = 4.0
set h3 = 3.0
set q21 = 0.8
set q31 = 0.3
set q32 = 0.3
nonlin c1 c2 c3 p3 p21 p22 p31 p33 a0 a1 a2 t0 t1 t2 b0 b1 $
       b2 u0 u1 u2 w0 w1 w2 d0 d1 d2 d5
frml a1t = r1(t)-c1-p3*r1(t-3)
frml a2t = r2(t)-c2-p21*r1(t-2)-p22*r2(t-2)
frml a3t = r3(t)-c3-p31*r1(t-1)-p33*r3(t-1)
frml v1 = a0+a1*a1t(t-1)**2+a2*h1(t-1)
frml q1t = t0 + t1*q21(t-1) + t2*a2t(t-1)
frml bt = a2t(t) - (q21(t)=q1t(t))*a1t(t)
frml v2 = b0+b1*bt(t-1)**2+b2*h2(t-1)
frml q2t = u0 + u1*q31(t-1) + u2*a3t(t-1)
frml q3t = w0 + w1*q32(t-1) + w2*a2t(t-1)
frml b1t = a3t(t)-(q31(t)=q2t(t))*a1t(t)-(q32(t)=q3t(t))*bt(t)
frml v3 = d0+d1*b1t(t-1)**2+d2*h3(t-1)+d5*h2(t-1)
```

```
frml gdet = -0.5*(log(h1(t) = v1(t))+ log(h2(t)=v2(t)) $
            +log(h3(t)=v3(t)))
frml garchln = gdet-0.5*(a1t(t)**2/h1(t)+bt(t)**2/h2(t) $
            +b1t(t)**2/h3(t))
smpl 8 2275
compute c1 = 0.07, c2 = 0.33, c3 = 0.19, p1 = 0.1, p3 = -0.04
compute p21 =0.2, p22 = -0.1, p31 = -0.26, p33 = 0.06
compute a0 = .01, a1 = 0.05, a2 = 0.94
compute t0 = 0.28, t1 =0.82, t2 = -0.035
compute b0 = .17, b1 = 0.08, b2 = 0.89
compute u0= 0.04, u1 = 0.97, u2 = 0.01
compute w0 =0.006, w1=0.98, w2=0.004
compute d0 =1.38, d1 = 0.06, d2 = 0.64, d5 = -0.027
nlpar(criterion=value,cvcrit=0.00001)
maximize(method=bhhh,recursive,iterations=250) garchln
set fv1 = v1(t)
set resi1 = a1t(t)/sqrt(fv1(t))
set residsq = resi1(t)*resi1(t)
* Checking standardized residuals *
cor(qstats,number=12,span=4) resi1
* Checking squared standardized residuals *
cor(qstats,number=12,span=4) residsq
set fv2 = v2(t)+q1t(t)**2*v1(t)
set resi2 = a2t(t)/sqrt(fv2(t))
set residsq = resi2(t)*resi2(t)
* Checking standardized residuals *
cor(qstats,number=12,span=4) resi2
* Checking squared standardized residuals *
cor(qstats,number=12,span=4) residsq
set fv3 = v3(t)+q2t(t)**2*v1(t)+q3t(t)**2*v2(t)
set resi3 = a3t(t)/sqrt(fv3(t))
set residsq = resi3(t)*resi3(t)
* Checking standardized residuals *
cor(qstats,number=12,span=4) resi3
* Checking squared standardized residuals *
cor(qstats,number=12,span=4) residsq
* print standardized residuals and correlation-coefficients
set rho21 = q1t(t)*sqrt(v1(t)/fv2(t))
set rho31 = q2t(t)*sqrt(v1(t)/fv3(t))
set rho32 = (q2t(t)*q1t(t)*v1(t) $
 +q3t(t)*v2(t))/sqrt(fv2(t)*fv3(t))
print 10 2275 resi1 resi2 resi3
print 10 2275 rho21 rho31 rho32
print 10 2275 fv1 fv2 fv3
```

## EXERCISES

**1.** The file "m-spibmge.dat" contains the monthly log returns in percentages of S&P 500 index, IBM stock, and General Electric stock from January 1926 to December 1999. The returns include dividends.

**(a)** Compute the sample mean, sample covariance matrix, and sample correlation matrix of the three return series.

**(b)** Compute the lag-1 and lag-2 cross-correlation matrixes of the three series. Draw inference concerning the linear relationships between the three series. Is there any lead-lag relation?

**(c)** Use the multivariate Ljung–Box statistics to test the null hypothesis that $H_0$ : $\boldsymbol{\rho}_1 = \cdots = \boldsymbol{\rho}_4 = \mathbf{0}$ at the 5% significance level. Draw your conclusion.

**2.** Focus on the monthly log returns in percentages of GE stock and the S&P 500 index. Build a constant correlation GARCH model for the bivariate series. Check the adequacy of the fitted model, and obtain 1-step ahead forecast of the covariance matrix at the forecast origin December 1999.

**3.** Focus on the monthly log returns in percentages of GE stock and the S&P 500 index. Build a time-varying correlation GARCH model for the bivariate series using a logistic function for the correlation coefficient. Check the adequacy of the fitted model, and obtain 1-step ahead forecast of the covariance matrix at the forecast origin December 1999.

**4.** Focus on the monthly log returns in percentages of GE stock and the S&P 500 index. Build a time-varying correlation GARCH model for the bivariate series using the Cholesky decomposition. Check the adequacy of the fitted model, and obtain 1-step ahead forecast of the covariance matrix at the forecast origin December 1999. Compare the model with the other two models built in the previous questions.

**5.** Consider the three-dimensional return series jointly. Build a multivariate time-varying correlation volatility model for the data, using the Cholesky decomposition. Discuss the implications of the model and compute 1-step ahead volatility forecast at the forecast origin $t = 888$.

**6.** An investor is interested in daily Value at Risk of his position on holding long \$0.5 million of Dell stock and \$1 million of Cisco Systems stock. Use 5% critical values and the daily log returns from January 2, 1990 to December 31, 1999 to do the calculation. The data are in the file "d-dellcsco9099.dat." Apply the three approaches to volatility modeling in Section 9.5 and compare the results.

## REFERENCES

Bollerslev, T. (1990), "Modeling the coherence in short-term nominal exchange rates: A multivariate generalized ARCH approach," *Review of Economics and Statistics*, 72, 498–505.

Chib, S., Nardari, F., and Shephard, N. (1999), "Analysis of high dimensional multivariate stochastic volatility models," Working paper, Washington University, St Louis.

Harvey, A., Ruiz, E., and Shephard, N. (1995), "Multivariate stochastic variance models," in *ARCH Selected Readings*, ed. R. F. Engle, pp. 253–276, Oxford University Press: Oxford, UK.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, Academic Press: New York.

Pourahmadi, M. (1999), "Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterization," *Biometrika*, 86, 677–690.

Tse, Y. K. (2000), "A test for constant correlations in a multivariate GARCH model," *Journal of Econometrics*, 98, 107–127.

Tse, Y. K., and Tsui, A. K. C. (1998), "A multivariate GARCH model with time-varying correlations," working paper, Department of Economics, National University of Singapore.

*Analysis of Financial Time Series.* Ruey S. Tsay

ISBN: 0-471-41544-8

CHAPTER 10

# Markov Chain Monte Carlo Methods with Applications

Advances in computing facilities and computational methods have dramatically increased our ability to solve complicated problems. The advances also extend the applicability of many existing econometric and statistical methods. Examples of such achievements in statistics include the Markov Chain Monte Carlo (MCMC) method and data augmentation. These techniques enable us to make some statistical inference that was not feasible just a few years ago. In this chapter, we introduce the ideas of MCMC methods and data augmentation that are widely applicable in finance. In particular, we discuss Bayesian inference via Gibbs sampling and demonstrate various applications of MCMC methods. Rapid developments in the MCMC methodology make it impossible to cover all the new methods available in the literature. Interested readers are referred to some recent books on Bayesian and empirical Bayesian statistics (e.g., Carlin and Louis, 2000; Gelman, Carlin, Stern, and Rubin, 1995).

For applications, we focus on issues related to financial econometrics. The demonstrations shown in this chapter only represent a small fraction of all possible applications of the techniques in finance. As a matter of fact, it is fair to say that Bayesian inference and the MCMC methods discussed here are applicable to most, if not all, of the studies in financial econometrics.

We begin the chapter by reviewing the concept of a *Markov process*. Consider a stochastic process $\{X_t\}$, where each $X_t$ assumes a value in the space $\mathbf{\Theta}$. The process $\{X_t\}$ is a Markov process if it has the property that, given the value of $X_t$, the values of $X_h$, $h > t$, do not depend on the values $X_s$, $s < t$. In other words, $\{X_t\}$ is a Markov process if its conditional distribution function satisfies

$$P(X_h \mid X_s, s \leq t) = P(X_h \mid X_t), \quad h > t.$$

If $\{X_t\}$ is a discrete-time stochastic process, then the prior property becomes

$$P(X_h \mid X_t, X_{t-1}, \ldots) = P(X_h \mid X_t), \quad h > t.$$

Let $\boldsymbol{A}$ be a subset of $\boldsymbol{\Theta}$. The function

$$P_t(\boldsymbol{\theta}, h, \boldsymbol{A}) = P(X_h \in \boldsymbol{A} \mid X_t = \boldsymbol{\theta}), \quad h > t$$

is called the transition probability function of the Markov process. If the transition probability depends on $h - t$, but not on $t$, then the process has a stationary transition distribution.

## 10.1 MARKOV CHAIN SIMULATION

Consider an inference problem with parameter vector $\boldsymbol{\theta}$ and data $\boldsymbol{X}$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. To make inference, we need to know the distribution $P(\boldsymbol{\theta} \mid \boldsymbol{X})$. The idea of Markov chain simulation is to simulate a Markov process on $\boldsymbol{\Theta}$, which converges to a stationary transition distribution that is $P(\boldsymbol{\theta} \mid \boldsymbol{X})$.

The key to Markov chain simulation is to create a Markov process whose stationary transition distribution is a specified $P(\boldsymbol{\theta} \mid \boldsymbol{X})$ and run the simulation sufficiently long so that the distribution of the current values of the process is close enough to the stationary transition distribution. It turns out that, for a given $P(\boldsymbol{\theta} \mid \boldsymbol{X})$, many Markov chains with the desired property can be constructed. We refer to methods that use Markov chain simulation to obtain the distribution $P(\boldsymbol{\theta} \mid \boldsymbol{X})$ as Markov Chain Monte Carlo (MCMC) methods.

The development of MCMC methods took place in various forms in the statistical literature. Consider the problem of "missing value" in data analysis. Most statistical methods discussed in this book were developed under the assumption of "complete data" (i.e., there is no missing value). For example, in modeling daily volatility of an asset return, we assume that the return data are available for all trading days in the sample period except for weekends and holidays. What should we do if there is a missing value?

Dempster, Laird, and Rubin (1977) suggest an iterative method called the EM algorithm to solve the problem. The method consists of two steps. First, if the missing value were available, then we could use methods of complete-data analysis to build a volatility model. Second, given the available data and the fitted model, we can derive the statistical distribution of the missing value. A simple way to fill in the missing value is to use the conditional expectation of the derived distribution of the missing value. In practice, one can start the method with an arbitrary value for the missing value and iterate the procedure for many many times until convergence. The first step of the prior procedure involves performing the maximum likelihood estimation of a specified model and is called the M-step. The second step is to compute the conditional expectation of the missing value and is called the E-step.

Tanner and Wong (1987) generalize the EM-algorithm in two ways. First, they introduce the idea of iterative simulation. For instance, instead of using the conditional expectation, one can simply replace the missing value by a random draw from its derived conditional distribution. Second, they extend the applicability of EM-algorithm by using the concept of data augmentation. By data augmentation, we

mean adding auxiliary variables to the problem under study. It turns out that many of the simulation methods can often be simplified or speeded up by data augmentation; see the application sections of this chapter.

## 10.2 GIBBS SAMPLING

Gibbs sampling (or Gibbs sampler) of Geman and Geman (1984) and Gelfand and Smith (1990) is perhaps the most popular MCMC method. We introduce the idea of Gibbs sampling by using a simple problem with three parameters. Here the word *parameter* is used in a very general sense. A missing data point can be regarded as a parameter under the MCMC framework. Similarly, an unobservable variable such as the "true" price of an asset can be regarded as $N$ parameters when there are $N$ transaction prices available. This concept of parameter is related to data augmentation and becomes apparent when we discuss applications of the MCMC methods.

Denote the three parameters by $\theta_1, \theta_2$, and $\theta_3$. Let $\boldsymbol{X}$ be the collection of available data and $M$ the entertained model. The goal here is to estimate the parameters so that the fitted model can be used to make inference. Suppose that the likelihood function of the model is hard to obtain, but the three conditional distributions of a single parameter given the others are available. In other words, we assume that the following three conditional distributions are known:

$$f_1(\theta_1 \mid \theta_2, \theta_3, \boldsymbol{X}, M); \quad f_2(\theta_2 \mid \theta_3, \theta_1, \boldsymbol{X}, M); \quad f_3(\theta_3 \mid \theta_1, \theta_2, \boldsymbol{X}, M), \tag{10.1}$$

where $f_i(\theta_i \mid \theta_{j\neq i}, \boldsymbol{X}, M)$ denotes the conditional distribution of the parameter $\theta_i$ given the data, the model, and the other two parameters. In application, we do not need to know the exact forms of the conditional distributions. What is needed is the ability to draw a random number from each of the three conditional distributions.

Let $\theta_{2,0}$ and $\theta_{3,0}$ be two arbitrary starting values of $\theta_2$ and $\theta_3$. The Gibbs sampler proceeds as follows:

1. Draw a random sample from $f_1(\theta_1 \mid \theta_{2,0}, \theta_{3,0}, \boldsymbol{X}, M)$. Denote the random draw by $\theta_{1,1}$.
2. Draw a random sample from $f_2(\theta_2 \mid \theta_{3,0}, \theta_{1,1}, \boldsymbol{X}, M)$. Denote the random draw by $\theta_{2,1}$.
3. Draw a random sample from $f_3(\theta_3 \mid \theta_{1,1}, \theta_{2,1}, \boldsymbol{X}, M)$. Denote the random draw by $\theta_{3,1}$.

This completes a Gibbs iteration and the parameters become $\theta_{1,1}$, $\theta_{2,1}$, and $\theta_{3,1}$.

Next, using the new parameters as starting values and repeating the prior iteration of random draws, we complete another Gibbs iteration to obtain the updated parameters $\theta_{1,2}$, $\theta_{2,2}$, and $\theta_{3,2}$. We can repeat the previous iterations for $m$ times to obtain a sequence of random draws:

$$(\theta_{1,1}, \theta_{2,1}, \theta_{3,1}), \ldots, (\theta_{1,m}, \theta_{2,m}, \theta_{3,m}).$$

Under some regularity conditions, it can be shown that, for a sufficiently large $m$, $(\theta_{1,m}, \theta_{2,m}, \theta_{3,m})$ is approximately equivalent to a random draw from the joint distribution $f(\theta_1, \theta_2, \theta_3 \mid \boldsymbol{X}, M)$ of the three parameters. The regularity conditions are weak; they essentially require that for an arbitrary starting value $(\theta_{1,0}, \theta_{2,0}, \theta_{3,0})$, the prior Gibbs iterations have a chance to visit the full parameter space. The actual convergence theorem involves using the Markov Chain theory; see Tierney (1994).

In practice, we use a sufficiently large $n$ and discard the first $m$ random draws of the Gibbs iterations to form a Gibbs sample, say

$$(\theta_{1,m+1}, \theta_{2,m+1}, \theta_{3,m+1}), \ldots, (\theta_{1,n}, \theta_{2,n}, \theta_{3,n}). \tag{10.2}$$

Since the previous realizations form a random sample from the joint distribution $f(\theta_1, \theta_2, \theta_3 \mid \boldsymbol{X}, M)$, they can be used to make inference. For example, a point estimate of $\theta_i$ and its variance are

$$\widehat{\theta_i} = \frac{1}{n-m} \sum_{j=m+1}^{n} \theta_{i,j}, \quad \widehat{\sigma}_i^2 = \frac{1}{n-m-1} \sum_{j=m+1}^{n} (\theta_{i,j} - \widehat{\theta_i})^2. \tag{10.3}$$

The Gibbs sample in Eq. (10.2) can be used in many ways. For example, if one is interested in testing the null hypothesis $H_o : \theta_1 = \theta_2$ versus the alternative hypothesis $H_a : \theta_1 \neq \theta_2$, then she can simply obtain point estimate of $\theta = \theta_1 - \theta_2$ and its variance as

$$\widehat{\theta} = \frac{1}{n-m} \sum_{j=m+1}^{n} (\theta_{1,j} - \theta_{2,j}), \quad \widehat{\sigma}^2 = \frac{1}{n-m-1} \sum_{j=m+1}^{n} (\theta_{1,j} - \theta_{2,j} - \widehat{\theta})^2.$$

The null hypothesis can then be tested by using the conventional $t$ ratio statistic $t = \widehat{\theta}/\widehat{\sigma}$.

***Remark:*** The first $m$ random draws of a Gibbs sampling, which are discarded, are commonly referred to as the *burn-ins* sample. The burn-ins are used to ensure that the Gibbs sample in Eq. (10.2) is indeed close enough to a random sample from the joint distribution $f(\theta_1, \theta_2, \theta_3 \mid \boldsymbol{X}, M)$.

***Remark:*** The method discussed before consists of running a single long chain and keeping all random draws after the burn-ins to obtain a Gibbs sample. Alternatively, one can run many relatively short chains using different starting values and a relatively small $n$. The random draw of the last Gibbs iteration in each chain is then used to form a Gibbs sample.

From the prior introduction, Gibbs sampling has the advantage to decompose a high-dimensional estimation problem into several lower dimensional ones via full conditional distributions of the parameters. At the extreme, a high-dimensional problem with $N$ parameters can be solved iteratively by using $N$ univariate conditional

distributions. This property makes the Gibbs sampling simple and widely applicable. However, it is often not efficient to reduce all the Gibbs draws into a univariate problem. When parameters are highly correlated, it pays to draw them jointly. Consider the three-parameter illustrative example. If $\theta_1$ and $\theta_2$ are highly correlated, then one should employ the conditional distributions $f(\theta_1, \theta_2 \mid \theta_3, \boldsymbol{X}, M)$ and $f_3(\theta_3 \mid \theta_1, \theta_2, \boldsymbol{X}, M)$ whenever possible. A Gibbs iteration then consists of (a) drawing jointly $(\theta_1, \theta_2)$ given $\theta_3$, and (b) drawing $\theta_3$ given $(\theta_1, \theta_2)$. For more information on the impact of parameter correlations on the convergence rate of a Gibbs sampler, see Liu, Wong, and Kong (1994).

In practice, convergence of a Gibbs sample is an important issue. The theory only states that the convergence occurs when the number of iterations $m$ is sufficiently large. It provides no specific guidance for choosing $m$. Many methods have been devised in the literature for checking the convergence of a Gibbs sample. But there is no consensus on which method performs best. In fact, none of the available methods can guarantee 100% that the Gibbs sample under study has converged for all applications. Performance of a checking method often depends on the problem at hand. Care must be exercised in a real application to ensure that there is no obvious violation of the convergence requirement; see Carlin and Louis (2000) and Gelman et al. (1995) for convergence checking methods. In application, it is important to repeat the Gibbs sampling several times with different starting values to ensure that the algorithm has converged.

## 10.3 BAYESIAN INFERENCE

Conditional distributions play a key role in Gibbs sampling. In the statistical literature, these conditional distributions are referred to as *conditional posterior distributions* because they are distributions of parameters given the data, other parameter values, and the entertained model. In this section, we review some well-known posterior distributions that are useful in using MCMC methods.

### 10.3.1 Posterior Distributions

There are two approaches to statistical inference. The first approach is the classical approach that bases on the maximum likelihood principle. Here a model is estimated by maximizing the likelihood function of the data, and the fitted model is used to make inference. The other approach is Bayesian inference that combines prior belief with data to obtain posterior distributions on which statistical inference is based. Historically, there were heated debates between the two schools of statistical inference. Yet both approaches have proved to be useful and are now widely accepted. The methods discussed so far in this book belong to the classical approach. However, Bayesian solutions exist for all of the problems considered. This is particularly so in recent years with the advances in MCMC methods, which greatly improve the feasibility of Bayesian analysis. Readers can revisit the previous chapters and derive MCMC solutions for the problems considered. In most cases, the Bayesian solutions

are similar to what we had before. In some cases, the Bayesian solutions might be advantageous. For example, consider the calculation of Value at Risk in Chapter 7. A Bayesian solution can easily take into consideration the parameter uncertainty in VaR calculation. However, the approach requires intensive computation.

Let $\boldsymbol{\theta}$ be the vector of unknown parameters of an entertained model and $\boldsymbol{X}$ be the data. Bayesian analysis seeks to combine knowledge about the parameters with the data to make inference. Knowledge of the parameters is expressed by specifying a *prior* distribution for the parameters, which is denoted by $P(\boldsymbol{\theta})$. For a given model, denote the likelihood function of the data by $f(\boldsymbol{X} \mid \boldsymbol{\theta})$. Then by the definition of conditional probability,

$$f(\boldsymbol{\theta} \mid \boldsymbol{X}) = \frac{f(\boldsymbol{\theta}, \boldsymbol{X})}{f(\boldsymbol{X})} = \frac{f(\boldsymbol{X} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta})}{f(\boldsymbol{X})}, \tag{10.4}$$

where the marginal distribution $f(\boldsymbol{X})$ can be obtained by

$$f(\boldsymbol{X}) = \int f(\boldsymbol{X}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int f(\boldsymbol{X} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

The distribution $f(\boldsymbol{\theta} \mid \boldsymbol{X})$ in Eq. (10.4) is called the *posterior distribution* of $\boldsymbol{\theta}$. In general, we can use Bayes' rule to obtain

$$f(\boldsymbol{\theta} \mid \boldsymbol{X}) \propto f(\boldsymbol{X} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}), \tag{10.5}$$

where $P(\boldsymbol{\theta})$ is the prior distribution and $f(\boldsymbol{X} \mid \boldsymbol{\theta})$ is the likelihood function. From Eq. (10.5), making statistical inference based on the likelihood function $f(\boldsymbol{X} \mid \boldsymbol{\theta})$ amounts to using Bayesian approach with a constant prior distribution.

### 10.3.2 Conjugate Prior Distributions

Obtaining the posterior distribution in Eq. (10.4) is not simple in general, but there are cases in which the prior and posterior distributions belong to the same family of distributions. Such a prior distribution is called a *conjugate* prior distribution. For MCMC methods, use of conjugate priors means that a closed-form solution for the conditional posterior distributions is available. Random draws of the Gibbs sampler can then be obtained by using the commonly available computer routines of probability distributions. In what follows, we review some well-known conjugate priors. For more information, readers are referred to textbooks on Bayesian statistics (e.g., DeGroot, 1970, Chapter 9).

***Result 1:*** Suppose that $x_1, \ldots, x_n$ form a random sample from a normal distribution with mean $\mu$, which is unknown, and variance $\sigma^2$, which is known and positive. Suppose that the prior distribution of $\mu$ is a normal distribution with mean $\mu_o$ and variance $\sigma_o^2$. Then the posterior distribution of $\mu$ given the data and prior is

normal with mean $\mu_*$ and variance $\sigma_*^2$ given by

$$\mu_* = \frac{\sigma^2\mu_o + n\sigma_o^2\bar{x}}{\sigma^2 + n\sigma_o^2} \quad \text{and} \quad \sigma_*^2 = \frac{\sigma^2\sigma_o^2}{\sigma^2 + n\sigma_o^2},$$

where $\bar{x} = \sum_{i=1}^n x_i/n$ is the sample mean.

In Bayesian analysis, it is often convenient to use the *precision* parameter $\eta = 1/\sigma^2$ (i.e., the inverse of the variance $\sigma^2$). Denote the precision parameter of the prior distribution by $\eta_o = 1/\sigma_o^2$ and that of the posterior distribution by $\eta_* = 1/\sigma_*^2$. Then Result 1 can be rewritten as

$$\eta_* = \eta_o + n\eta \quad \text{and} \quad \mu_* = \frac{\eta_o}{\eta_*} \times \mu_o + \frac{n\eta}{\eta_*} \times \bar{x}.$$

For the normal random sample considered, data information about $\mu$ is contained in the sample mean $\bar{x}$, which is the sufficient statistic of $\mu$. The precision of $\bar{x}$ is $n/\sigma^2 = n\eta$. Consequently, Result 1 says that (a) precision of the posterior distribution is the sum of the precisions of the prior and the data, and (b) the posterior mean is a weighted average of the prior mean and sample mean with weight proportional to the precision. The two formulas also show that the contribution of prior distribution is diminishing as the sample size $n$ increases.

A multivariate version of Result 1 is particularly useful in MCMC methods when linear regression models are involved; see Box and Tiao (1973).

***Result 1a:*** Suppose that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ form a random sample from a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and a known covariance matrix $\boldsymbol{\Sigma}$. Suppose also that the prior distribution of $\boldsymbol{\mu}$ is multivariate normal with mean vector $\boldsymbol{\mu}_o$ and covariance matrix $\boldsymbol{\Sigma}_o$. Then the posterior distribution of $\boldsymbol{\mu}$ is also multivariate normal with mean vector $\boldsymbol{\mu}_*$ and covariance matrix $\boldsymbol{\Sigma}_*$, where

$$\boldsymbol{\Sigma}_*^{-1} = \boldsymbol{\Sigma}_o^{-1} + n\boldsymbol{\Sigma}^{-1} \quad \text{and} \quad \boldsymbol{\mu}_* = \boldsymbol{\Sigma}_*(\boldsymbol{\Sigma}_o^{-1}\boldsymbol{\mu}_o + n\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}}),$$

where $\bar{\boldsymbol{x}} = \sum_{i=1}^n \boldsymbol{x}_i/n$ is the sample mean, which is distributed as a multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}/n$. Note that $n\boldsymbol{\Sigma}^{-1}$ is the precision matrix of $\bar{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_o^{-1}$ is the precision matrix of the prior distribution.

A random variable $\eta$ has a gamma distribution with positive parameters $\alpha$ and $\beta$ if its probability density function is

$$f(\eta \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\eta^{\alpha-1}e^{-\beta\eta}, \quad \eta > 0,$$

where $\Gamma(\alpha)$ is a Gamma function. For this distribution, $E(\eta) = \alpha/\beta$ and $\text{Var}(\eta) = \alpha/\beta^2$.

***Result 2:*** Suppose that $x_1, \ldots, x_n$ form a random sample from a normal distribution with a given mean $\mu$ and an unknown precision $\eta$. If the prior distribution of $\eta$ is a gamma distribution with positive parameters $\alpha$ and $\beta$, then the posterior distribution of $\eta$ is a gamma distribution with parameters $\alpha+(n/2)$ and $\beta+\sum_{i=1}^{n}(x_i-\mu)^2/2$.

A random variable $\theta$ has a beta distribution with positive parameters $\alpha$ and $\beta$ if its probability density function is

$$f(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0 < \theta < 1.$$

The mean and variance of $\theta$ are $E(\theta) = \alpha/(\alpha+\beta)$ and $\mathrm{Var}(\theta) = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$.

***Result 3:*** Suppose that $x_1, \ldots, x_n$ form a random sample from a Bernoulli distribution with parameter $\theta$. If the prior distribution of $\theta$ is a beta distribution with given positive parameters $\alpha$ and $\beta$, then the posterior of $\theta$ is a beta distribution with parameters $\alpha + \sum_{i=1}^{n} x_i$ and $\beta + n - \sum_{i=1}^{n} x_i$.

***Result 4:*** Suppose that $x_1, \ldots, x_n$ form a random sample from a Poisson distribution with parameter $\lambda$. Suppose also that the prior distribution of $\lambda$ is a gamma distribution with given positive parameters $\alpha$ and $\beta$. Then the posterior distribution of $\lambda$ is a gamma distribution with parameters $\alpha + \sum_{i=1}^{n} x_i$ and $\beta + n$.

***Result 5:*** Suppose that $x_1, \ldots, x_n$ form a random sample from an exponential distribution with parameter $\lambda$. If the prior distribution of $\lambda$ is a gamma distribution with given positive parameters $\alpha$ and $\beta$, then the posterior distribution of $\lambda$ is a gamma distribution with parameters $\alpha + n$ and $\beta + \sum_{i=1}^{n} x_i$.

A random variable $X$ has a negative binomial distribution with parameters $m$ and $\lambda$, where $m > 0$ and $0 < \lambda < 1$, if $X$ has a probability mass function

$$p(n \mid m, \lambda) = \begin{cases} \binom{m+n-1}{n}\lambda^m(1-\lambda)^n & \text{if } n = 0, 1, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

A simple example of negative binomial distribution in finance is how many MBA graduates a firm must interview before finding exactly $m$ "right candidates" for its $m$ openings, assuming that the applicants are independent and each applicant has a probability $\lambda$ of being a perfect fit. Denote the total number of interviews by $Y$. Then $X = Y - m$ is distributed as a negative binomial with parameters $m$ and $\lambda$.

***Result 6:*** Suppose that $x_1, \ldots, x_n$ form a random sample from a negative binomial distribution with parameters $m$ and $\lambda$, where $m$ is positive and fixed. If the prior

distribution of $\lambda$ is a beta distribution with positive parameters $\alpha$ and $\beta$, then the posterior distribution of $\lambda$ is a beta distribution with parameters $\alpha + mn$ and $\beta + \sum_{i=1}^{n} x_i$.

Next we consider the case of a normal distribution with an unknown mean $\mu$ and an unknown precision $\eta$. The two-dimensional prior distribution is partitioned as $P(\mu, \eta) = P(\mu \mid \eta)P(\eta)$.

***Result 7:*** Suppose that $x_1, \ldots, x_n$ form a random sample from a normal distribution with an unknown mean $\mu$ and an unknown precision $\eta$. Suppose also that the conditional distribution of $\mu$ given $\eta = \eta_o$ is a normal distribution with mean $\mu_o$ and precision $\tau_o \eta_o$ and the marginal distribution of $\eta$ is a gamma distribution with positive parameters $\alpha$ and $\beta$. Then the conditional posterior distribution of $\mu$ given $\eta = \eta_o$ is a normal distribution with mean $\mu_*$ and precision $\eta_*$,

$$\mu_* = \frac{\tau_o \mu_o + n\bar{x}}{\tau_o + n} \quad \text{and} \quad \eta_* = (\tau_o + n)\eta_o,$$

where $\bar{x} = \sum_{i=1}^{n} x_i / n$ is the sample mean, and the marginal posterior distribution of $\eta$ is a gamma distribution with parameters $\alpha + (n/2)$ and $\beta_*$, where

$$\beta_* = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\tau_o n(\bar{x} - \mu_o)^2}{2(\tau_o + n)}.$$

When the conditional variance of a random variable is of interest, inverted chi-squared distribution (or inverse chi-squared) is often used. A random variable $Y$ has an inverted chi-squared distribution with $v$ degrees of freedom if $1/Y$ follows a chi-squared distribution with the same degrees of freedom. The probability density function of $Y$ is

$$f(y \mid v) = \frac{2^{-v/2}}{\Gamma(v/2)} y^{-(v/2+1)} e^{-1/(2y)}, \quad y > 0.$$

For this distribution, we have $E(Y) = 1/(v-2)$ if $v > 2$ and $\text{Var}(Y) = 2/[(v-2)^2(v-4)]$ if $v > 4$.

***Result 8:*** Suppose that $a_1, \ldots, a_n$ form a random sample from a normal distribution with mean zero and variance $\sigma^2$. Suppose also that the prior distribution of $\sigma^2$ is an inverted chi-squared distribution with $v$ degrees of freedom [i.e., $(v\lambda)/\sigma^2 \sim \chi_v^2$, where $\lambda > 0$]. Then the posterior distribution of $\sigma^2$ is also an inverted chi-squared distribution with $v + n$ degrees of freedom—that is, $(v\lambda + \sum_{i=1}^{n} a_i^2)/\sigma^2 \sim \chi_{v+n}^2$.

## 10.4 ALTERNATIVE ALGORITHMS

In many applications, there are no closed-form solutions for the conditional posterior distributions. But many clever alternative algorithms have been devised in the statis-

tical literature to overcome this difficulty. In this section, we discuss some of these algorithms.

### 10.4.1 Metropolis Algorithm

This algorithm is applicable when the conditional posterior distribution is known except for a normalization constant; see Metropolis and Ulam (1949) and Metropolis et al. (1953). Suppose that we want to draw a random sample from the distribution $f(\boldsymbol{\theta} \mid \boldsymbol{X})$, which contains a complicated normalization constant so that a direct draw is either too time-consuming or infeasible. But there exists an approximate distribution for which random draws are easily available. The Metropolis algorithm generates a sequence of random draws from the approximate distribution whose distributions converge to $f(\boldsymbol{\theta} \mid \boldsymbol{X})$. The algorithm proceeds as follows:

1. Draw a random starting value $\boldsymbol{\theta}_0$ such that $f(\boldsymbol{\theta}_0 \mid \boldsymbol{X}) > 0$.
2. For $t = 1, 2, \ldots$
   - **(a)** Draw a candidate sample $\boldsymbol{\theta}_*$ from a *known* distribution at iteration $t$ given the previous draw $\boldsymbol{\theta}_{t-1}$. Denote the known distribution by $J_t(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1})$, which is called a *jumping distribution* in Gelman et al. (1995). The jumping distribution must be symmetric—that is, $J_t(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_j) = J_t(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_i)$ for all $\boldsymbol{\theta}_i$, $\boldsymbol{\theta}_j$, and $t$.
   - **(b)** Calculate the ratio
   $$r = \frac{f(\boldsymbol{\theta}_* \mid \boldsymbol{X})}{f(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{X})}.$$
   - **(c)** Set
   $$\boldsymbol{\theta}_t = \begin{cases} \boldsymbol{\theta}_* & \text{with probability } \min(r, 1) \\ \boldsymbol{\theta}_{t-1} & \text{otherwise.} \end{cases}$$

Under some regularity conditions, the sequence $\{\boldsymbol{\theta}_t\}$ converges in distribution to $f(\boldsymbol{\theta} \mid \boldsymbol{X})$; see Gelman et al. (1995).

Implementation of the algorithm requires the ability to calculate the ratio $r$ for all $\boldsymbol{\theta}_*$ and $\boldsymbol{\theta}_{t-1}$, to draw $\boldsymbol{\theta}_*$ from the jumping distribution, and to draw a random realization from a uniform distribution to determine the acceptance or rejection of $\boldsymbol{\theta}_*$. The normalization constant of $f(\boldsymbol{\theta} \mid \boldsymbol{X})$ is not needed because only ratio is used.

The acceptance and rejection rule of the algorithm can be stated as follows: (i) if the jump from $\boldsymbol{\theta}_{t-1}$ to $\boldsymbol{\theta}_*$ increases the conditional posterior density, then accept $\boldsymbol{\theta}_*$ as $\boldsymbol{\theta}_t$; (ii) if the jump decreases the posterior density, then set $\boldsymbol{\theta}_t = \boldsymbol{\theta}_*$ with probability equal to the density ratio $r$, and set $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1}$ otherwise. Such a procedure seems reasonable.

Examples of symmetric jumping distributions include the normal and Student-$t$ distributions for the mean parameter. For a given covariance matrix, we have $f(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_j) = f(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_i)$, where $f(\boldsymbol{\theta} \mid \boldsymbol{\theta}_o)$ denotes a multivariate normal density function with mean vector $\boldsymbol{\theta}_o$.

### 10.4.2 Metropolis–Hasting Algorithm

Hasting (1970) generalizes the Metropolis algorithm in two ways. First, the jumping distribution does not have to be symmetric. Second, the jumping rule is modified to

$$r = \frac{f(\boldsymbol{\theta}_* \mid \boldsymbol{X})/J_t(\boldsymbol{\theta}_* \mid \boldsymbol{\theta}_{t-1})}{f(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{X})/J_t(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{\theta}_*)} = \frac{f(\boldsymbol{\theta}_* \mid \boldsymbol{X})J_t(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{\theta}_*)}{f(\boldsymbol{\theta}_{t-1} \mid \boldsymbol{X})J_t(\boldsymbol{\theta}_* \mid \boldsymbol{\theta}_{t-1})}.$$

This modified algorithm is referred to as the Metropolis–Hasting algorithm.

### 10.4.3 Griddy Gibbs

In financial applications, an entertained model may contain some nonlinear parameters (e.g., the moving average parameters in an ARMA model or the GARCH parameters in a volatility model). Since conditional posterior distributions of nonlinear parameters do not have a closed-form expression, implementing a Gibbs sampler in this situation may become complicated even with the Metropolis–Hasting algorithm. Tanner (1996) describes a simple procedure to obtain random draws in a Gibbs sampling when the conditional posterior distribution is univariate. The method is called the *Griddy Gibbs sampler* and is widely applicable. However, the method could be inefficient in a real application.

Let $\theta_i$ be a scalar parameter with conditional posterior distribution $f(\theta_i \mid \boldsymbol{X}, \boldsymbol{\theta}_{-i})$, where $\boldsymbol{\theta}_{-i}$ is the parameter vector after removing $\theta_i$. For instance, if $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$, then $\boldsymbol{\theta}_{-1} = (\theta_2, \theta_3)'$. The Griddy Gibbs proceeds as follows:

1. Select a grid of points from a properly selected interval of $\theta_i$, say $\theta_{i1} \leq \theta_{i2} \leq \cdots \leq \theta_{im}$. Evaluate the conditional posterior density function to obtain $w_j = f(\theta_{ij} \mid \boldsymbol{X}, \boldsymbol{\theta}_{-i})$ for $j = 1, \ldots, m$.
2. Use $w_1, \ldots, w_m$ to obtain an approximation to the inverse cumulative distribution function (CDF) of $f(\theta_i \mid \boldsymbol{X}, \boldsymbol{\theta}_{-i})$.
3. Draw a uniform (0, 1) random variate and transform the observation via the approximate inverse CDF to obtain a random draw for $\theta_i$.

Some remarks on the Griddy Gibbs are in order. First, the normalization constant of the conditional posterior distribution $f(\theta_i \mid \boldsymbol{X}, \boldsymbol{\theta}_{-i})$ is not needed because the inverse CDF can be obtained from $\{w_j\}_{j=1}^m$ directly. Second, a simple approximation to the inverse CDF is a discrete distribution for $\{\theta_{ij}\}_{j=1}^m$ with probability $p(\theta_{ij}) = w_j / \sum_{v=1}^m w_v$. Third, in a real application, selection of the interval $[\theta_{i1}, \theta_{im}]$ for the parameter $\theta_i$ must be checked carefully. A simple checking procedure is to consider the histogram of the Gibbs draws of $\theta_i$. If the histogram indicates substantial probability around $\theta_{i1}$ or $\theta_{im}$, then the interval must be expanded. However, if the histogram shows a concentration of probability inside the interval $[\theta_{i1}, \theta_{im}]$, then the interval is too wide and can be shortened. If the interval is too wide, then the Griddy Gibbs becomes inefficient because most of $w_j$ would be zero. Finally, the Griddy Gibbs or Metropolis–Hasting algorithm can be used in a Gibbs sampling to obtain random draws of some parameters.

## 10.5 LINEAR REGRESSION WITH TIME-SERIES ERRORS

We are ready to consider some specific applications of MCMC methods. Examples discussed in the next few sections are for illustrative purposes only. The goal here is to highlight the applicability and usefulness of the methods. Understanding these examples can help readers gain insights into applications of MCMC methods in finance.

The first example is to estimate a regression model with serially correlated errors. This is a topic discussed in Chapter 2, where we use SCA to perform the estimation. A simple version of the model is

$$y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_k x_{kt} + z_t$$
$$z_t = \phi z_{t-1} + a_t,$$

where $y_t$ is the dependent variable, $x_{it}$ are explanatory variables that may contain lagged values of $y_t$, and $z_t$ follows a simple AR(1) model with $\{a_t\}$ being a sequence of independent and identically distributed normal random variables with mean zero and variance $\sigma^2$. Denote the parameters of the model by $\boldsymbol{\theta} = (\boldsymbol{\beta}', \phi, \sigma^2)'$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$, and let $\boldsymbol{x}_t = (1, x_{1t}, \ldots, x_{kt})'$ be the vector of all regressors at time $t$, including a constant of unity. The model becomes

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + z_t, \quad z_t = \phi z_{t-1} + a_t, \quad t = 1, \ldots, n, \tag{10.6}$$

where $n$ is the sample size.

A natural way to implement Gibbs sampling in this case is to iterate between regression estimation and time-series estimation. If the time-series model is known, then we can estimate the regression model easily by using the least squares method. However, if the regression model is known, then we can obtain the time series $z_t$ by using $z_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}$ and use the series to estimate the AR(1) model. Therefore, we need the following conditional posterior distributions:

$$f(\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{X}, \phi, \sigma^2); \quad f(\phi \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2); \quad f(\sigma^2 \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}, \phi),$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{X}$ denotes the collection of all observations of explanatory variables.

We use conjugate prior distributions to obtain closed-form expressions for the conditional posterior distributions. The prior distributions are

$$\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_o, \boldsymbol{\Sigma}_o), \quad \phi \sim N(\phi_o, \sigma_o^2), \quad \frac{v\lambda}{\sigma^2} \sim \chi_v^2, \tag{10.7}$$

where again $\sim$ denotes distribution, $\boldsymbol{\beta}_o$, $\boldsymbol{\Sigma}_o$, $\lambda$, $v$, $\phi_o$, and $\sigma_o^2$ are known quantities. These quantities are referred to as hyperparameters in Bayesian inference. Their exact values depend on the problem at hand. Typically, we assume that $\boldsymbol{\beta}_o = \boldsymbol{0}$,

$\phi_o = 0$, and $\boldsymbol{\Sigma}_o$ is a diagonal matrix with large diagonal elements. The prior distributions in Eq. (10.7) are assumed to be independent of each other. Thus, we use independent priors based on the partition of the parameter vector $\boldsymbol{\theta}$.

The conditional posterior distribution $f(\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{X}, \phi, \sigma^2)$ can be obtained by using Result 1a of Section 10.3. Specifically, given $\phi$, we define

$$y_{o,t} = y_t - \phi y_{t-1}, \quad \boldsymbol{x}_{o,t} = \boldsymbol{x}_t - \phi \boldsymbol{x}_{t-1}.$$

Using Eq. (10.6), we have

$$y_{o,t} = \boldsymbol{\beta}' \boldsymbol{x}_{o,t} + a_t, \quad t = 2, \ldots, n. \tag{10.8}$$

Under the assumption of $\{a_t\}$, Eq. (10.8) is a multiple linear regression. Therefore, information of the data about the parameter vector $\boldsymbol{\beta}$ is contained in its least squares estimate

$$\widehat{\boldsymbol{\beta}} = \left( \sum_{t=2}^{n} \boldsymbol{x}_{o,t} \boldsymbol{x}_{o,t}' \right)^{-1} \left( \sum_{t=2}^{n} \boldsymbol{x}_{o,t} y_{o,t} \right),$$

which has a multivariate normal distribution

$$\widehat{\boldsymbol{\beta}} \sim N \left[ \boldsymbol{\beta}, \quad \sigma^2 \left( \sum_{t=2}^{n} \boldsymbol{x}_{o,t} \boldsymbol{x}_{o,t}' \right)^{-1} \right].$$

Using Results 1a, the posterior distribution of $\boldsymbol{\beta}$, given the data, $\phi$, and $\sigma^2$, is multivariate normal. We write the result as

$$(\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{X}, \phi, \sigma) \sim N(\boldsymbol{\beta}_*, \boldsymbol{\Sigma}_*), \tag{10.9}$$

where the parameters are given by

$$\boldsymbol{\Sigma}_*^{-1} = \frac{\sum_{t=2}^{n} \boldsymbol{x}_{o,t} \boldsymbol{x}_{o,t}'}{\sigma^2} + \boldsymbol{\Sigma}_o^{-1}, \quad \boldsymbol{\beta}_* = \boldsymbol{\Sigma}_* \left( \frac{\sum_{t=2}^{n} \boldsymbol{x}_{o,t} \boldsymbol{x}_{o,t}'}{\sigma^2} \widehat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}_o^{-1} \boldsymbol{\beta}_o \right).$$

Next consider the conditional posterior distribution of $\phi$ given $\boldsymbol{\beta}, \sigma^2$, and the data. Because $\boldsymbol{\beta}$ is given, we can calculate $z_t = y_t - \boldsymbol{\beta}' \boldsymbol{x}_t$ for all $t$ and consider the AR(1) model

$$z_t = \phi z_{t-1} + a_t, \quad t = 2, \ldots, n.$$

The information of the likelihood function about $\phi$ is contained in the least squares estimate

$$\widehat{\phi} = \left( \sum_{t=2}^{n} z_{t-1}^2 \right)^{-1} \left( \sum_{t=2}^{n} z_{t-1} z_t \right),$$

which is normally distributed with mean $\phi$ and variance $\sigma^2(\sum_{t=2}^{n} z_{t-1}^2)^{-1}$. Based on Result 1, the posterior distribution of $\phi$ is also normal with mean $\phi_*$ and variance $\sigma_*^2$, where

$$\sigma_*^{-2} = \frac{\sum_{t=2}^{n} z_{t-1}^2}{\sigma^2} + \sigma_o^{-2}, \quad \phi_* = \sigma_*^2 \left( \frac{\sum_{t=2}^{n} z_{t-1}^2}{\sigma^2} \widehat{\phi} + \sigma_o^{-2} \phi_o \right). \tag{10.10}$$

Finally, turn to the posterior distribution of $\sigma^2$ given $\boldsymbol{\beta}$, $\phi$, and the data. Because $\boldsymbol{\beta}$ and $\phi$ are known, we can calculate

$$a_t = z_t - \phi z_{t-1}, \quad z_t = y_t - \boldsymbol{\beta}' \boldsymbol{x}_t, \quad t = 2, \ldots, n.$$

By Result 8 of Section 10.3, the posterior distribution of $\sigma^2$ is an inverted chi-squared distribution—that is,

$$\frac{v\lambda + \sum_{t=2}^{n} a_t^2}{\sigma^2} \sim \chi_{v+(n-1)}^2, \tag{10.11}$$

where $\chi_k^2$ denotes a chi-squared distribution with $k$ degrees of freedom.

Using the three conditional posterior distributions in Eqs. (10.9)–(10.11), we can estimate Eq. (10.6) via Gibbs sampling as follows:

1. Specify the hyperparameter values of the priors in Eq. (10.7).
2. Specify arbitrary starting values for $\boldsymbol{\beta}$, $\phi$, and $\sigma^2$ (e.g., the ordinary least squares estimate of $\boldsymbol{\beta}$ without time-series errors).
3. Use the multivariate normal distribution in Eq. (10.9) to draw a random realization for $\boldsymbol{\beta}$.
4. Use the univariate normal distribution in Eq. (10.10) to draw a random realization for $\phi$.
5. Use the chi-squared distribution in Eq. (10.11) to draw a random realization for $\sigma^2$.

Repeat Steps 3–5 for many iterations to obtain a Gibbs sample. The sample means are then used as point estimates of the parameters of model (10.6).

**Example 10.1.** As an illustration, we revisit the example of U.S. weekly interest rates of Chapter 2. The data are the 1-year and 3-year Treasury constant maturity rates from January 5, 1962 to September 10, 1999 and are obtained from the Federal Reserve Bank of St Louis. Because of unit-root nonstationarity, the dependent and independent variables are

1. $c_{3t} = r_{3t} - r_{3,t-1}$, which is the weekly change in 3-year maturity rate,
2. $c_{1t} = r_{1t} - r_{1,t-1}$, which is the weekly change in 1-year maturity rate,

where the original interest rates $r_{it}$ are measured in percentages. In Chapter 2, we employed a linear regression model with an MA(1) error for the data. Here we consider an AR(2) model for the error process. Using the traditional approach, we obtain the model

$$c_{3t} = 0.0002 + 0.782c_{1t} + z_t, \quad z_t = 0.205z_{t-1} - 0.068z_{t-2} + a_t, \qquad (10.12)$$

where $\widehat{\sigma}_a = 0.067$. Standard errors of the coefficient estimates of Eq. (10.12) are 0.0017, 0.008, 0.023, and 0.023, respectively. Except for a marginally significant residual ACF at lag 6, the prior model seems adequate.

Writing the model as

$$c_{3t} = \beta_0 + \beta_1 c_{1t} + z_t, \quad z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + a_t, \qquad (10.13)$$

where $\{a_t\}$ is an independent sequence of $N(0, \sigma^2)$ random variables, we estimate the parameters by Gibbs sampling. The prior distributions used are

$$\boldsymbol{\beta} \sim N(\mathbf{0}, 4\boldsymbol{I}_2), \quad \boldsymbol{\phi} \sim N[\mathbf{0}, \text{diag}(0.25, 0.16)], \quad (v\lambda)/\sigma^2 = (10 \times 0.1)/\sigma^2 \sim \chi^2_{10},$$

where $\boldsymbol{I}_2$ is the $2 \times 2$ identity matrix. The initial parameter estimates are obtained by the ordinary least squares method (i.e., by using a two-step procedure of fitting the linear regression model first, then fitting an AR(2) model to the regression residuals). Since the sample size 1966 is large, the initial estimates are close to those given in Eq. (10.12). We iterated the Gibbs sampling for 2100 iterations, but discard results of the first 100 iterations. Table 10.1 gives the posterior means and standard errors of the parameters. Figure 10.1 shows the histogram of the marginal posterior distribution of each parameter.

We repeated the Gibbs sampling with different initial values, but obtained similar results. The Gibbs sampling appears to have converged. From Table 10.1, the posterior means are close to the estimates of Eq. (10.12) except for the coefficient of $z_{t-2}$. However, the posterior standard errors of $\phi_1$ and $\phi_2$ are relatively large, indicating uncertainty in these two estimates. The histograms of Figure 10.1 are informative. In particular, they show that the distributions of $\widehat{\phi}_1$ and $\widehat{\phi}_2$ have not converged to the asymptotic normality; the distributions are skewed to the right. However, the asymptotic normality of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ seems reasonable.

**Table 10.1. Posterior Means and Standard Errors of Model (10.13) Estimated by a Gibbs Sampling with 2100 Iterations. The Results Are Based on the Last 2000 Iterations and the Prior Distributions Are Given in the Text.**

| Parameter | $\beta_0$ | $\beta_1$ | $\phi_1$ | $\phi_2$ | $\sigma$ |
|---|---|---|---|---|---|
| Mean | 0.025 | 0.784 | 0.305 | 0.032 | 0.074 |
| St. Error | 0.024 | 0.009 | 0.089 | 0.087 | 0.003 |

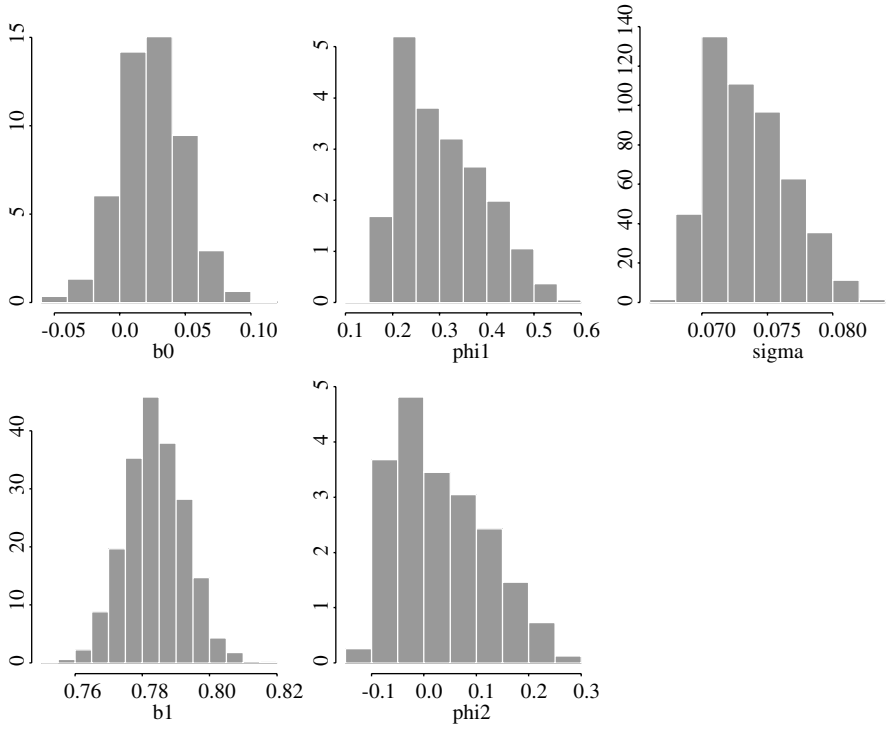


**Figure 10.1.** Histograms of Gibbs draws for the model in Eq. (10.13) with 2100 iterations. The results are based on the last 2000 draws. Prior distributions and starting parameter values are given in the text.

## 10.6 MISSING VALUES AND OUTLIERS

In this section, we discuss MCMC methods for handling missing values and detecting additive outliers. Let $\{y_t\}_{t=1}^n$ be an observed time series. A data point $y_h$ is an additive outlier if

$$y_t = \begin{cases} x_h + \omega & \text{if } t = h \\ x_t & \text{otherwise,} \end{cases} \tag{10.14}$$

where $\omega$ is the magnitude of the outlier and $x_t$ is an outlier-free time series. Examples of additive outliers include recording errors (e.g., typos and measurement errors). Outliers can seriously affect time-series analysis because they may induce substantial biases in parameter estimation and lead to model misspecification.

Consider a time series $x_t$ and a fixed time index $h$. We can learn a lot about $x_h$ by treating it as a missing value. If the model of $x_t$ were known, then we could derive the conditional distribution of $x_h$ given the other values of the series. By comparing the observed value $y_h$ with the derived distribution of $x_h$, we can determine whether

$y_h$ can be classified as an additive outlier. Specifically, if $y_h$ is a value that is likely to occur under the derived distribution, then $y_h$ is not an additive outlier. However, if the chance to observe $y_h$ is very small under the derived distribution, then $y_h$ can be classified as an additive outlier. Therefore, detection of additive outliers and treatment of missing values in time-series analysis are based on the same idea.

In the literature, missing values in a time series can be handled by using either the Kalman filter or MCMC methods; see Jones (1980) and McCulloch and Tsay (1994). Outlier detection has also been carefully investigated; see Chang, Tiao, and Chen (1988), Tsay (1988), Tsay, Peña, and Pankratz (2000), and the references therein. The outliers are classified into four categories depending on the nature of their impacts on the time series. Here we focus on additive outliers.

### 10.6.1 Missing Values

For ease in presentation, consider an AR($p$) time series

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t, \tag{10.15}$$

where $\{a_t\}$ is a Gaussian white noise series with mean zero and variance $\sigma^2$. Suppose that the sampling period is from $t = 1$ to $t = n$, but the observation $x_h$ is missing, where $1 < h < n$. Our goal is to estimate the model in the presence of a missing value.

In this particular instance, the parameters are $\boldsymbol{\theta} = (\boldsymbol{\phi}', x_h, \sigma^2)'$, where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$. Thus, we treat the missing value $x_h$ as an unknown parameter. If we assume that the prior distributions are

$$\boldsymbol{\phi} \sim N(\boldsymbol{\phi}_o, \boldsymbol{\Sigma}_o), \quad x_h \sim N(\mu_o, \sigma_o^2), \quad \frac{v\lambda}{\sigma^2} \sim \chi_v^2,$$

where the hyperparameters are known, then the conditional posterior distributions $f(\boldsymbol{\phi} \mid \boldsymbol{X}, x_h, \sigma^2)$ and $f(\sigma^2 \mid \boldsymbol{X}, x_h, \boldsymbol{\phi})$ are exactly as those given in the previous section, where $\boldsymbol{X}$ denotes the observed data. The conditional posterior distribution $f(x_h \mid \boldsymbol{X}, \boldsymbol{\phi}, \sigma^2)$ is univariate normal with mean $\mu_*$ and variance $\sigma_h^2$. These two parameters can be obtained by using a linear regression model. Specifically, given the model and the data, $x_h$ is only related to $\{x_{h-p}, \ldots, x_{h-1}, x_{h+1}, \ldots, x_{h+p}\}$. Keeping in mind that $x_h$ is an unknown parameter, we can write the relationship as follows:

**1.** For $t = h$, the model says

$$x_h = \phi_1 x_{h-1} + \cdots + \phi_p x_{h-p} + a_h.$$

Let $y_h = \phi_1 x_{h-1} + \cdots + \phi_p x_{h-p}$ and $b_h = -a_h$, the prior equation can be written as

$$y_h = x_h + b_h = \phi_0 x_h + b_h,$$

where $\phi_0 = 1$.

**2.** For $t = h + 1$, we have

$$x_{h+1} = \phi_1 x_h + \phi_2 x_{h-1} + \cdots + \phi_p x_{h+1-p} + a_{h+1}.$$

Let $y_{h+1} = x_{h+1} - \phi_2 x_{h-1} - \cdots - \phi_p x_{h+1-p}$ and $b_{h+1} = a_{h+1}$, the prior equation can be written as

$$y_{h+1} = \phi_1 x_h + b_{h+1}.$$

**3.** In general, for $t = h + j$ with $j = 1, \ldots, p$, we have

$$x_{h+j} = \phi_1 x_{h+j-1} + \cdots + \phi_j x_h + \phi_{j+1} x_{h-1} + \cdots + \phi_p x_{h+j-p} + a_{h+j}.$$

Let $y_{h+j} = x_{h+j} - \phi_1 x_{h+j-1} - \cdots - \phi_{j-1} x_{h+1} - \phi_{j+1} x_{h-1} - \cdots - \phi_p x_{h+j-p}$ and $b_{h+j} = a_{h+j}$. The prior equation reduces to

$$y_{h+j} = \phi_j x_h + b_{h+j}.$$

Consequently, for an AR($p$) model, the missing value $x_h$ is related to the model, and the data in $p + 1$ equations

$$y_{h+j} = \phi_j x_h + b_{h+j}, \quad j = 0, \ldots, p, \tag{10.16}$$

where $\phi_0 = 1$. Since a normal distribution is symmetric with respective to its mean, $a_h$ and $-a_h$ have the same distribution. Consequently, Eq. (10.16) is a special simple linear regression model with $p + 1$ data points. The least squares estimate of $x_h$ and its variance are

$$\widehat{x_h} = \frac{\sum_{j=0}^{p} \phi_j y_{h+j}}{\sum_{j=0}^{p} \phi_j^2}, \quad \text{Var}(\widehat{x_h}) = \frac{\sigma^2}{\sum_{j=0}^{p} \phi_j^2}.$$

For instance, when $p = 1$, we have $\widehat{x_h} = \frac{\phi_1}{1+\phi_1^2}(x_{h-1} + x_{h+1})$, which is referred to as the filtered value of $x_h$. Because a Gaussian AR(1) model is time reversible, equal weights are applied to the two neighboring observations of $x_h$ to obtain the filtered value.

Finally, using Result 1 of Section 10.3, we obtain that the posterior distribution of $x_h$ is normal with mean $\mu_*$ and variance $\sigma_*^2$, where

$$\mu_* = \frac{\sigma^2 \mu_o + \sigma_o^2 (\sum_{j=0}^{p} \phi_j^2) \widehat{x_h}}{\sigma^2 + \sigma_o^2 (\sum_{j=0}^{p} \phi_j^2)}, \quad \sigma_*^2 = \frac{\sigma^2 \sigma_o^2}{\sigma^2 + \sigma_o^2 \sum_{j=0}^{p} \phi_j^2}. \tag{10.17}$$

Missing values may occur in patches, resulting in the situation of multiple consecutive missing values. These missing values can be handled in two ways. First, we can generalize the prior method directly to obtain a solution for multiple filtered values.

Consider, for instance, the case that $x_h$ and $x_{h+1}$ are missing. These missing values are related to $\{x_{h-p}, \ldots, x_{h-1}; x_{h+2}, \ldots, x_{h+p+1}\}$. We can define a dependent variable $y_{h+j}$ in a similar manner as before to set up a multiple linear regression with parameters $x_h$ and $x_{h+1}$. The least squares method is then used to obtain estimates of $x_h$ and $x_{h+1}$. Combining with the specified prior distributions, we have a bivariate normal posterior distribution for $(x_h, x_{h+1})'$. In Gibbs sampling, this approach draws the consecutive missing values jointly. Second, we can apply the result of a single missing value in Eq. (10.17) multiple times within a Gibbs iteration. Again consider the case of missing $x_h$ and $x_{h+1}$. We can employ the conditional posterior distributions $f(x_h \mid \boldsymbol{X}, x_{h+1}, \boldsymbol{\phi}, \sigma^2)$ and $f(x_{h+1} \mid \boldsymbol{X}, x_h, \boldsymbol{\phi}, \sigma^2)$ separately. In Gibbs sampling, this means that we draw the missing value one at a time.

Because $x_h$ and $x_{h+1}$ are correlated in a time series drawing them jointly is preferred in a Gibbs sampling. This is particularly so if the number of consecutive missing values is large. Drawing one missing value at a time works well if the number of missing values is small.

***Remark:*** In the previous discussion, we assume $h - p \geq 1$ and $h + p \leq n$. If $h$ is close to the end points of the sample period, the number of data points available in the linear regression model must be adjusted.

### 10.6.2 Outlier Detection

Detection of additive outliers in Eq. (10.14) becomes straightforward under the MCMC framework. Except for the case of a patch of additive outliers with similar magnitudes, the simple Gibbs sampler of McCulloch and Tsay (1994) seems to work well; see Justel, Peña, and Tsay (2001). Again we use an AR model to illustrate the problem. The method applies equally well to other time series models when the Metropolis–Hasting algorithm, or the Griddy Gibbs is used to draw values of nonlinear parameters.

Assume that the observed time series is $y_t$, which may contain some additive outliers whose locations and magnitudes are unknown. We write the model for $y_t$ as

$$y_t = \delta_t \beta_t + x_t, \quad t = 1, \ldots, n, \tag{10.18}$$

where $\{\delta_t\}$ is a sequence of independent Bernoulli random variables such that $P(\delta_t = 1) = \epsilon$ and $P(\delta_t = 0) = 1 - \epsilon$, $\epsilon$ is a constant between 0 and 1, $\{\beta_t\}$ is a sequence of independent random variables from a given distribution, and $x_t$ is an outlier-free AR($p$) time series,

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + a_t,$$

where $\{a_t\}$ is a Gaussian white noise with mean zero and variance $\sigma^2$. This model seems complicated, but it allows additive outliers to occur at every time point. The chance of being an outlier for each observation is $\epsilon$.

Under the model in Eq. (10.18), we have $n$ data points, but there are $2n + p + 3$ parameters—namely, $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_p)'$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$, $\sigma^2$, and $\epsilon$. The binary parameters $\delta_t$ are governed by $\epsilon$ and $\beta_t$s are determined by the specified distribution. The parameters $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ are introduced by using the idea of data augmentation with $\delta_t$ denoting the presence or absence of an additive outlier at time $t$, and $\beta_t$ is the magnitude of the outlier at time $t$ when it is present.

Assume that the prior distributions are

$$\boldsymbol{\phi} \sim N(\boldsymbol{\phi}_o, \boldsymbol{\Sigma}_o), \quad \frac{v\lambda}{\sigma^2} \sim \chi^2_v, \quad \epsilon \sim \text{beta}(\gamma_1, \gamma_2), \quad \beta_t \sim N(0, \xi^2),$$

where the hyperparameters are known. These are conjugate prior distributions. To implement Gibbs sampling for model estimation with outlier detection, we need to consider the conditional posterior distributions of

$$f(\boldsymbol{\phi} \mid \boldsymbol{Y}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma^2), \quad f(\delta_h \mid \boldsymbol{Y}, \boldsymbol{\delta}_{-h}, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2), \quad f(\beta_h \mid \boldsymbol{Y}, \boldsymbol{\delta}, \boldsymbol{\beta}_{-h}, \boldsymbol{\phi}, \sigma^2),$$
$$f(\epsilon \mid \boldsymbol{Y}, \boldsymbol{\delta}), \quad f(\sigma^2 \mid \boldsymbol{Y}, \boldsymbol{\phi}, \boldsymbol{\delta}, \boldsymbol{\beta}),$$

where $1 \le h \le n$, $\boldsymbol{Y}$ denotes the data, and $\boldsymbol{\theta}_{-i}$ denotes that the $i$th element of $\boldsymbol{\theta}$ is removed.

Conditioned on $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$, the outlier-free time series $x_t$ can be obtained by $x_t = y_t - \delta_t \beta_t$. Information of the data about $\boldsymbol{\phi}$ is then contained in the least squares estimate

$$\widehat{\boldsymbol{\phi}} = \left( \sum_{t=p+1}^{n} \boldsymbol{x}_{t-1} \boldsymbol{x}'_{t-1} \right)^{-1} \left( \sum_{t=p+1}^{n} \boldsymbol{x}_{t-1} x_t \right),$$

where $\boldsymbol{x}_{t-1} = (1, x_{t-1}, \ldots, x_{t-p})'$, which is normally distributed with mean $\boldsymbol{\phi}$ and covariance matrix

$$\widehat{\boldsymbol{\Sigma}} = \sigma^2 \left( \sum_{t=p+1}^{n} \boldsymbol{x}_{t-1} \boldsymbol{x}'_{t-1} \right)^{-1}.$$

The conditional posterior distribution of $\boldsymbol{\phi}$ is therefore multivariate normal with mean $\boldsymbol{\phi}_*$ and covariance matrix $\boldsymbol{\Sigma}_*$, which are given in Eq. (10.9) with $\boldsymbol{\beta}$ being replaced by $\boldsymbol{\phi}$ and $\boldsymbol{x}_{o,t}$ by $\boldsymbol{x}_{t-1}$. Similarly, the conditional posterior distribution of $\sigma^2$ is an inverted chi-squared distribution—that is,

$$\frac{v\lambda + \sum_{t=p+1}^{n} a_t^2}{\sigma^2} \sim \chi^2_{v+(n-p)},$$

where $a_t = x_t - \boldsymbol{\phi}' \boldsymbol{x}_{t-1}$ and $x_t = y_t - \delta_t \beta_t$.

The conditional posterior distribution of $\delta_h$ can be obtained as follows. First, $\delta_h$ is only related to $\{y_j, \beta_j\}_{j=h-p}^{h+p}$, $\{\delta_j\}_{j=h-p}^{h+p}$ with $j \neq h$, $\boldsymbol{\phi}$, and $\sigma^2$. More specifically,

we have

$$x_j = y_j - \delta_j \beta_j, \quad j \neq h.$$

Second, $x_h$ can assume two possible values: $x_h = y_h - \beta_h$ if $\delta_h = 1$ and $x_h = y_h$ otherwise. Define

$$w_j = x_j^* - \phi_0 - \phi_1 x_{j-1}^* - \cdots - \phi_p x_{j-p}^*, \quad j = h, \ldots, h+p,$$

where $x_j^* = x_j$ if $j \neq h$ and $x_h^* = y_h$. The two possible values of $x_h$ give rise to two situations:

- Case I: $\delta_h = 0$. Here the $h$th observation is not an outlier and $x_h^* = y_h = x_h$. Hence, $w_j = a_j$ for $j = h, \ldots, h+p$. In other words, we have

$$w_j \sim N(0, \sigma^2), \quad j = h, \ldots, h+p.$$

- Case II: $\delta_h = 1$. Now the $h$th observation is an outlier and $x_h^* = y_h = x_h + \beta_h$. The $w_j$ defined before is contaminated by $\beta_h$. In fact, we have

$$w_h \sim N(\beta_h, \sigma^2) \quad \text{and} \quad w_j \sim N(-\phi_{j-h}\beta_h, \sigma^2), \quad j = h+1, \ldots, h+p.$$

  If we define $\psi_0 = -1$ and $\psi_i = \phi_i$ for $i = 1, \ldots, p$, then we have $w_j \sim N(-\psi_{j-h}\beta_h, \sigma^2)$ for $j = h, \ldots, h+p$.

Based on the prior discussion, we can summarize the situation as follows:

1. Case I: $\delta_h = 0$ with probability $1 - \epsilon$. In this case, $w_j \sim N(0, \sigma^2)$ for $j = h, \ldots, h+p$.
2. Case II: $\delta_h = 1$ with probability $\epsilon$. Here $w_j \sim N(-\psi_{j-h}\beta_h, \sigma^2)$ for $j = h, \ldots, h+p$.

Since there are $n$ data points, $j$ cannot be greater than $n$. Let $m = \min(n, h+p)$. The posterior distribution of $\delta_h$ is therefore

$$P(\delta_h = 1 \mid \boldsymbol{Y}, \boldsymbol{\delta}_{-h}, \boldsymbol{\beta}, \boldsymbol{\phi}, \sigma^2) = \frac{\epsilon \exp[-\sum_{j=h}^{m}(w_j + \psi_{j-h}\beta_h)^2/(2\sigma^2)]}{\epsilon \exp[-\sum_{j=h}^{m}(w_j + \psi_{j-h}\beta_h)^2/(2\sigma^2)] + (1-\epsilon)\exp[-\sum_{j=h}^{m} w_j^2/(2\sigma^2)]}. \tag{10.19}$$

This posterior distribution is simply to compare the weighted values of likelihood function under the two situations with weight being the probability of each situation.

Finally, the posterior distribution of $\beta_h$ is as follows.

- If $\delta_h = 0$, then $y_h$ is not an outlier and $\beta_h \sim N(0, \xi^2)$.
- If $\delta_h = 1$, then $y_h$ is contaminated by an outlier with magnitude $\beta_h$. The variable $w_j$ defined before contains information of $\beta_h$ for $j = h, h+1, \ldots,$ min $(h+p, n)$. Specifically, we have $w_j \sim N(-\psi_{j-h}\beta_h, \sigma^2)$ for $j = h, h+1, \ldots, \min(h+p, n)$. The information can be put in a linear regression framework as

$$w_j = -\psi_{j-h}\beta_h + a_j, \quad j = h, h+1, \ldots, \min(h+p, n).$$

Consequently, the information is embedded in the least squares estimate

$$\widehat{\beta}_h = \frac{\sum_{j=h}^{m} -\psi_{j-h} w_j}{\sum_{j=h}^{m} \psi_{j-h}^2}, \quad m = \min(h+p, n),$$

which is normally distributed with mean $\beta_h$ and variance $\sigma^2/(\sum_{j=h}^{m} \psi_{j-h}^2)$. By Result 1, the posterior distribution of $\beta_h$ is normal with mean $\beta_h^*$ and variance $\sigma_{h*}^2$, where

$$\beta_h^* = \frac{-(\sum_{j=h}^{m} \psi_{j-h} w_j)\xi^2}{\sigma^2 + (\sum_{j=h}^{m} \psi_{j-h}^2)\xi^2}, \quad \sigma_{h*}^2 = \frac{\sigma^2 \xi^2}{\sigma^2 + (\sum_{j=h}^{m} \psi_{j-h}^2)\xi^2}.$$

**Example 10.2.** Consider the weekly change series of U.S. 3-year Treasury constant maturity interest rate from March 18, 1988 to September 10, 1999 for 600 observations. The interest rate is in percentage and is a subseries of the dependent variable $c_{3t}$ of Example 10.1. The time series is shown in Figure 10.2(a). If AR models are entertained for the series, the partial autocorrelation function suggests an AR(3) model and we obtain

$$c_{3t} = 0.227 c_{3,t-1} + 0.006 c_{3,t-2} + 0.114 c_{3,t-2} + a_t, \quad \widehat{\sigma}_a^2 = 0.0128,$$

where standard errors of the coefficients are 0.041, 0.042, and 0.041, respectively. The Ljung–Box statistics of the residuals show $Q(12) = 11.4$, which is insignificant at the 5% level.

Next we apply the Gibbs sampling to estimate the AR(3) model and to detect simultaneously possible additive outliers. The prior distributions used are

$$\boldsymbol{\phi} \sim N(\mathbf{0}, 0.25\boldsymbol{I}_3), \quad \frac{\upsilon\lambda}{\sigma^2} = \frac{5 \times 0.00256}{\sigma^2} \sim \chi_5^2, \quad \gamma_1 = 5, \quad \gamma_2 = 95, \quad \xi^2 = 0.1,$$

where $0.00256 \approx \widehat{\sigma}^2/5$ and $\xi^2 \approx 9\widehat{\sigma}^2$. The expected number of additive outliers is 5%. Using initial values $\epsilon = 0.05$, $\sigma^2 = 0.012$, $\phi_1 = 0.2$, $\phi_2 = 0.02$, and

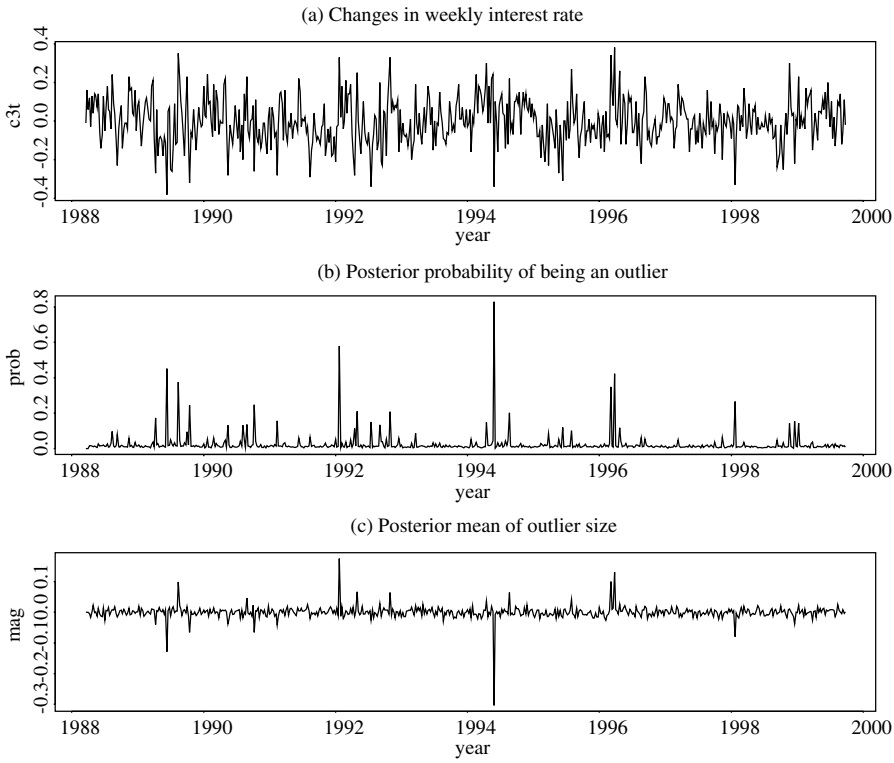


**Figure 10.2.** Time plots of weekly change series of U.S. 3-year Treasury constant maturity interest rate from March 18, 1988 to September 10, 1999: (a) the data, (b) the posterior probability of being an outlier, and (c) the posterior mean of outlier size. The estimation is based on a Gibbs sampling with 1050 iterations, but results of the first 50 iterations are discarded.

$\phi_3 = 0.1$, we run the Gibbs sampling for 1050 iterations, but discard results of the first 50 iterations. Using posterior means of the coefficients as parameter estimates, we obtain the fitted model

$$c_{3t} = 0.252c_{3,t-1} + 0.003c_{3,t-2} + 0.110c_{3,t-2} + a_t, \quad \widehat{\sigma}^2 = 0.0118,$$

where posterior standard deviations of the parameters are 0.046, 0.045, 0.046, and 0.0008, respectively. Thus, the Gibbs sampling produces results similar to that of the maximum likelihood method. Figure 10.2(b) shows the time plot of posterior probability of each observation being an additive outlier, and Figure 10.2(c) plots the posterior mean of outlier magnitude. From the probability plot, some observations have high probabilities of being an outlier. In particular, $t = 323$ has a probability of 0.83 and the associated posterior mean of outlier magnitude is $-0.304$. This point corresponds to May 20, 1994 when the $c_{3t}$ changed from 0.24 to $-0.34$ (i.e., about a 0.6% drop in the weekly interest rate within two weeks). The point with second

highest posterior probability of being an outlier is $t = 201$, which is January 17, 1992. The outlying posterior probability is 0.58 and the estimated outlier size is 0.176. At this second point, $c_{3t}$ changed from $-0.02$ to 0.33, corresponding to a jump of about 0.35% in the weekly interest rate.

***Remark:*** Outlier detection via Gibbs sampling requires intensive computation, but the approach performs a joint estimation of model parameters and outliers. Yet the traditional approach to outlier detection separates estimation from detection. It is much faster in computation, but may produce spurious detections when multiple outliers are present. For the data in Example 10.2, the SCA program also identifies $t = 323$ and $t = 201$ as the two most significant additive outliers. The estimated outlier sizes are $-0.39$ and 0.36, respectively.

## 10.7 STOCHASTIC VOLATILITY MODELS

An important financial application of MCMC methods is the estimation of stochastic volatility models; see Jacquier, Polson, and Rossi (1994) and the references therein. We start with a univariate stochastic volatility model. The mean and volatility equations of an asset return $r_t$ are

$$r_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_p x_{pt} + a_t, \quad a_t = \sqrt{h_t}\epsilon_t \tag{10.20}$$

$$\ln h_t = \alpha_0 + \alpha_1 \ln h_{t-1} + v_t \tag{10.21}$$

where $\{x_{it} \mid i = 1, \ldots, p\}$ are explanatory variables available at time $t-1$, $\beta_j$s are parameters, $\{\epsilon_t\}$ is a Gaussian white noise sequence with mean 0 and variance 1, $\{v_t\}$ is also a Gaussian white noise sequence with mean 0 and variance $\sigma_v^2$, and $\{\epsilon_t\}$ and $\{v_t\}$ are independent. The log transformation is used to ensure that $h_t$ is positive for all $t$. The explanatory variables $x_{it}$ may include lagged values of the return (e.g., $x_{it} = r_{t-i}$). In Eq. (10.21), we assume that $|\alpha_1| < 1$ so that the log volatility process $\ln h_t$ is stationary. If necessary, a higher order AR($p$) model can be used for $\ln h_t$.

Denote the coefficient vector of the mean equation by $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ and the parameter vector of the volatility equation by $\boldsymbol{\omega} = (\alpha_0, \alpha_1, \sigma_v^2)'$. Suppose that $\boldsymbol{R} = (r_1, \ldots, r_n)'$ is the collection of observed returns and $\boldsymbol{X}$ is the collection of explanatory variables. Let $\boldsymbol{H} = (h_1, \ldots, h_n)'$ be the vector of unobservable volatilities. Here $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ are the "traditional" parameters of the model and $\boldsymbol{H}$ is an auxiliary variable. Estimation of the model would be complicated via the maximum likelihood method because the likelihood function is a mixture over the $n$-dimensional $\boldsymbol{H}$ distribution as

$$f(\boldsymbol{R} \mid \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\omega}) = \int f(\boldsymbol{R} \mid \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{H}) f(\boldsymbol{H} \mid \boldsymbol{\omega}) d\boldsymbol{H}.$$

However, under the Bayesian framework, the volatility vector $\boldsymbol{H}$ consists of augmented parameters. Conditioning on $\boldsymbol{H}$, we can focus on the probability distribution functions $f(\boldsymbol{R} \mid \boldsymbol{H}, \boldsymbol{\beta})$ and $f(\boldsymbol{H} \mid \boldsymbol{\omega})$ and the prior distribution $p(\boldsymbol{\beta}, \boldsymbol{\omega})$. We assume

that the prior distribution can be partitioned as $p(\boldsymbol{\beta}, \boldsymbol{\omega}) = p(\boldsymbol{\beta})p(\boldsymbol{\omega})$, that is, prior distributions for the mean and volatility equations are independent. A Gibbs sampling approach to estimating the stochastic volatility in Eqs. (10.20) and (10.21) then involves drawing random samples from the following conditional posterior distributions:

$$f(\boldsymbol{\beta} \mid \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\omega}), \quad f(\boldsymbol{H} \mid \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\omega}), \quad f(\boldsymbol{\omega} \mid \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{H}).$$

In what follows, we give details of practical implementation of the Gibbs sampling used.

### 10.7.1 Estimation of Univariate Models

Given $\boldsymbol{H}$, the mean equation in Eq. (10.20) is a nonhomogeneous linear regression. Dividing the equation by $\sqrt{h_t}$, we can write the model as

$$r_{o,t} = \boldsymbol{x}'_{o,t}\boldsymbol{\beta} + \epsilon_t, \quad t = 1, \ldots, n, \tag{10.22}$$

where $r_{o,t} = r_t/\sqrt{h_t}$ and $\boldsymbol{x}_{o,t} = \boldsymbol{x}_t/\sqrt{h_t}$, with $\boldsymbol{x}_t = (1, x_{1t}, \ldots, x_{pt})'$ being the vector of explanatory variables. Suppose that the prior distribution of $\boldsymbol{\beta}$ is multivariate normal with mean $\boldsymbol{\beta}_o$ and covariance matrix $\boldsymbol{A}_o$. Then the posterior distribution of $\boldsymbol{\beta}$ is also multivariate normal with mean $\boldsymbol{\beta}_*$ and covariance matrix $\boldsymbol{A}_*$. These two quantities can be obtained as before via Result 1a and they are

$$A_*^{-1} = \sum_{t=1}^{n} \boldsymbol{x}_{o,t}\boldsymbol{x}'_{o,t} + \boldsymbol{A}_o^{-1}, \quad \boldsymbol{\beta}_* = \boldsymbol{A}_* \left( \sum_{t=1}^{n} \boldsymbol{x}_{o,t} r_{o,t} + \boldsymbol{A}_o^{-1}\boldsymbol{\beta}_o \right),$$

where it is understood that the summation starts with $p + 1$ if $r_{t-p}$ is the highest lagged return used in the explanatory variables.

The volatility vector $\boldsymbol{H}$ is drawn element by element. The necessary conditional posterior distribution is $f(h_t \mid \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{H}_{-t}, \boldsymbol{\beta}, \boldsymbol{\omega})$, which is produced by the normal distribution of $a_t$ and the lognormal distribution of the volatility,

$$\begin{aligned} & f(h_t \mid \boldsymbol{R}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{H}_{-t}, \boldsymbol{\omega}) \\ & \quad \propto f(a_t \mid h_t, r_t, \boldsymbol{x}_t, \boldsymbol{\beta}) f(h_t \mid h_{t-1}, \boldsymbol{\omega}) f(h_{t+1} \mid h_t, \boldsymbol{\omega}) \\ & \quad \propto h_t^{-0.5} \exp[-(r_t - \boldsymbol{x}'_t\boldsymbol{\beta})^2/(2h_t)] h_t^{-1} \exp[-(\ln h_t - \mu_t)^2/(2\sigma^2)] \\ & \quad \propto h_t^{-1.5} \exp[-(r_t - \boldsymbol{x}'_t\boldsymbol{\beta})^2/(2h_t) - (\ln h_t - \mu_t)^2/(2\sigma^2)], \end{aligned} \tag{10.23}$$

where $\mu_t = [\alpha_0(1-\alpha_1) + \alpha_1(\ln h_{t+1} + \ln h_{t-1})]/(1+\alpha_1^2)$ and $\sigma^2 = \sigma_v^2/(1+\alpha_1^2)$. Here we have used the following properties: (a) $a_t \mid h_t \sim N(0, h_t)$; (b) $\ln h_t \mid \ln h_{t-1} \sim N(\alpha_0 + \alpha_1 \ln h_{t-1}, \sigma_v^2)$; (c) $\ln h_{t+1} \mid \ln h_t \sim N(\alpha_0 + \alpha_1 \ln h_t, \sigma_v^2)$; (d) $d \ln h_t = h_t^{-1} dh_t$, where $d$ denotes differentiation; and (e) the equality

$$(x-a)^2 A + (x-b)^2 C = (x-c)^2(A+C) + (a-b)^2 AC/(A+C)$$

where $c = (Aa + Cb)/(A + C)$ provided that $A + C \neq 0$. This equality is a scalar version of Lemma 1 of Box and Tiao (1973, p. 418). In our application, $A = 1$, $a = \alpha_0 + \ln h_{t-1}$, $C = \alpha_1^2$, and $b = (\ln h_{t+1} - \alpha_0)/\alpha_1$. The term $(a-b)^2 AC/(A+C)$ does not contain the random variable $h_t$ and, hence, is integrated out in the derivation of the conditional posterior distribution. Jacquier, Polson, and Rossi (1994) use Metropolis algorithm to draw $h_t$. We use Griddy Gibbs in this section, and the range of $h_t$ is chosen to be a multiple of the unconditional sample variance of $r_t$.

To draw random samples of $\boldsymbol{\omega}$, we partition the parameters as $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)'$ and $\sigma_v^2$. The prior distribution of $\boldsymbol{\omega}$ is also partitioned accordingly [i.e., $p(\boldsymbol{\omega}) = p(\boldsymbol{\alpha})p(\sigma_v^2)$]. The conditional posterior distributions needed are

- $f(\boldsymbol{\alpha} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\beta}, \sigma_v^2) = f(\boldsymbol{\alpha} \mid \boldsymbol{H}, \sigma_v^2)$: Given $\boldsymbol{H}$, $\ln h_t$ follows an AR(1) model. Therefore, the result of AR models discussed in the previous two sections applies. Specifically, if the prior distribution of $\boldsymbol{\alpha}$ is multivariate normal with mean $\boldsymbol{\alpha}_o$ and covariance matrix $C_o$, then $f(\boldsymbol{\alpha} \mid \boldsymbol{H}, \sigma_v^2)$ is multivariate normal with mean $\boldsymbol{\alpha}_*$ and covariance matrix $\boldsymbol{C}_*$, where

$$\boldsymbol{C}_*^{-1} = \frac{\sum_{t=2}^{n} \boldsymbol{z}_t \boldsymbol{z}_t'}{\sigma_v^2} + \boldsymbol{C}_o^{-1}, \quad \boldsymbol{\alpha}_* = \boldsymbol{C}_* \left( \frac{\sum_{t=2}^{n} \boldsymbol{z}_t \ln h_t}{\sigma_v^2} + \boldsymbol{C}_o^{-1} \boldsymbol{\alpha}_o \right),$$

  where $\boldsymbol{z}_t = (1, \ln h_{t-1})'$.

- $f(\sigma_v^2 \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{H}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = f(\sigma_v^2 \mid \boldsymbol{H}, \boldsymbol{\alpha})$: Given $\boldsymbol{H}$ and $\boldsymbol{\alpha}$, we can calculate $v_t = \ln h_t - \alpha_0 - \alpha_1 \ln h_{t-1}$ for $t = 2, \ldots, n$. Therefore, if the prior distribution of $\sigma_v^2$ is $(m\lambda)/\sigma_v^2 \sim \chi_m^2$, then the conditional posterior distribution of $\sigma_v^2$ is an inverted chi-squared distribution with $m + n - 1$ degrees of freedom, i.e.

$$\frac{m\lambda + \sum_{t=2}^{n} v_t^2}{\sigma_v^2} \sim \chi_{m+n-1}^2.$$

***Remark:*** The formula (10.23) is for $1 < t < n$, where $n$ is the sample size. For the two end data points $h_1$ and $h_n$, some modifications are needed. A simple approach is to assume that $h_1$ is fixed so that the drawing of $h_t$ starts with $t = 2$. For $t = n$, one uses the result $\ln h_n \sim (\alpha_0 + \alpha_1 \ln h_{n-1}, \sigma_v^2)$. Alternatively, one can employ a forecast of $h_{n+1}$ and a backward prediction of $h_0$ and continue to apply the formula. Since $h_n$ is the variable of interest, we forecast $h_{n+1}$ by using a 2-step ahead forecast at the forecast origin $n - 1$. For the model in Eq. (10.21), the forecast of $h_{n+1}$ is

$$\widehat{h}_{n-1}(2) = \alpha_0 + \alpha_1(\alpha_0 + \alpha_1 \ln h_{n-1}).$$

The backward prediction of $h_0$ is based on the time reversibility of the model

$$(\ln h_t - \eta) = \alpha_1(\ln h_{t-1} - \eta) + v_t,$$

where $\eta = \alpha_0/(1-\alpha_1)$ and $|\alpha_1| < 1$. The model of the reversed series is

$$(\ln h_t - \eta) = \alpha_1(\ln h_{t+1} - \eta) + v_t^*,$$

where $\{v_t^*\}$ is also a Gaussian white noise series with mean zero and variance $\sigma_v^2$. Consequently, the 2-step backward prediction of $h_0$ at time $t = 2$ is

$$\widehat{h}_2(-2) = \alpha_1^2(\ln h_2 - \eta).$$

***Remark:*** The formula (10.23) can also be obtained by using results of a missing value in an AR(1) model; see subsection 10.6.1. Specifically, assume that $\ln h_t$ is missing. For the AR(1) model in Eq. (10.21), this missing value is related to $\ln h_{t-1}$ and $\ln h_{t+1}$ for $1 < t < n$. From the model, we have

$$\ln h_t = \alpha_0 + \alpha_1 \ln h_{t-1} + a_t.$$

Define $y_t = \alpha_0 + \alpha_1 \ln h_{t-1}$, $x_t = 1$, and $b_t = -a_t$. Then we obtain

$$y_t = x_t \ln h_t + b_t. \tag{10.24}$$

Next, from

$$\ln h_{t+1} = \alpha_0 + \alpha_1 \ln h_t + a_{t+1},$$

we define $y_{t+1} = \ln h_{t+1} - \alpha_0$, $x_{t+1} = \alpha_1$ and $b_{t+1} = a_{t+1}$, and obtain

$$y_{t+1} = x_{t+1} \ln h_t + b_{t+1}. \tag{10.25}$$

Now Eqs. (10.24) and (10.25) form a special simple linear regression with two observations and an unknown parameter $\ln h_t$. Note that $b_t$ and $b_{t+1}$ have the same distribution because $-a_t$ is also $N(0, \sigma_v^2)$. The least squares estimate of $\ln h_t$ is then

$$\widehat{\ln h_t} = \frac{x_t y_t + x_{t+1} y_{t+1}}{x_t^2 + x_{t+1}^2} = \frac{\alpha_0(1-\alpha_1) + \alpha_1(\ln h_{t+1} + \ln h_{t-1})}{1+\alpha_1^2},$$

which is precisely the conditional mean of $\ln h_t$ given in Eq. (10.23). In addition, this estimate is normally distributed with mean $\ln h_t$ and variance $\sigma_v^2/(1+\alpha_1^2)$. Formula (10.23) is simply the product of $a_t \sim N(0, h_t)$ and $\widehat{\ln h_t} \sim N[\ln h_t, \sigma_v^2/(1+\alpha_1^2)]$ with the transformation $d \ln h_t = h_t^{-1} dh_t$. This regression approach generalizes easily to other AR($p$) models for $\ln h_t$. We use this approach and assume that $\{h_t\}_{t=1}^p$ are fixed for a stochastic volatility AR($p$) model.

***Remark:*** Starting values of $h_t$ can be obtained by fitting a volatility model of Chapter 3 to the return series.

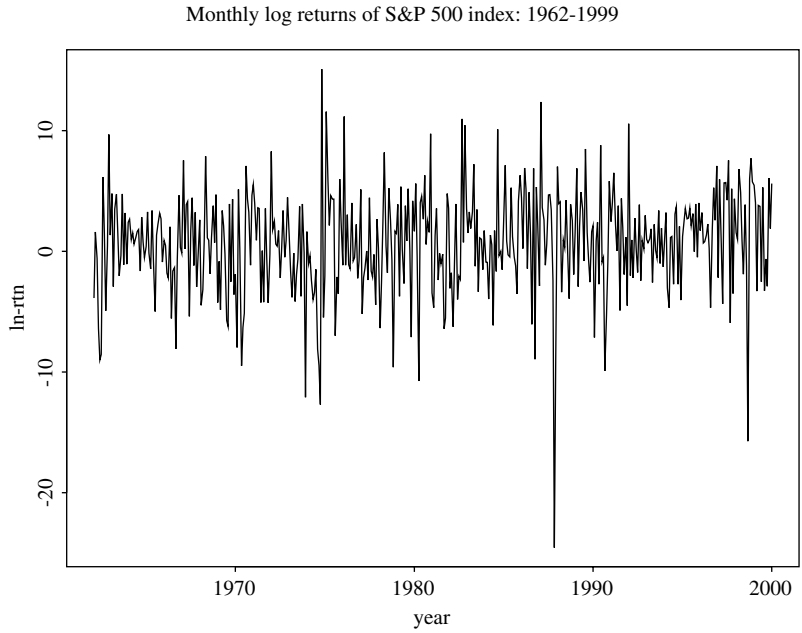


**Figure 10.3.** Time plot of monthly log returns of S&P 500 index from 1962 to 1999.

**Example 10.3.** Consider the monthly log returns of the S&P 500 index from January 1962 to December 1999 for 456 observations. Figure 10.3 shows the time plot of the return measured in percentage. If GARCH models are entertained for the series, we obtain a Gaussian GARCH(1, 1) model

$$\begin{aligned} r_t &= 0.658 + a_t, \quad a_t = \sqrt{h_t}\epsilon_t \\ h_t &= 3.349 + 0.086a_{t-1}^2 + 0.735h_{t-1}, \end{aligned} \tag{10.26}$$

where $t$ ratios of the coefficients are all greater than 2.52. The Ljung–Box statistics of the standardized residuals and their squared series fail to indicate any model inadequacy.

Next, consider the stochastic volatility model

$$\begin{aligned} r_t &= \mu + a_t, \quad a_t = \sqrt{h_t}\epsilon_t \\ \ln h_t &= \alpha_0 + \alpha_1 \ln h_{t-1} + v_t, \end{aligned} \tag{10.27}$$

where $v_t$s are iid $N(0, \sigma_v^2)$. To implement the Gibbs sampling, we use the prior distributions

$$\mu \sim N(0, 9), \quad \boldsymbol{\alpha} \sim N[\boldsymbol{\alpha}_o, \text{diag}(0.09, 0.04)], \quad \frac{5 \times 0.2}{\sigma_v^2} \sim \chi_5^2,$$

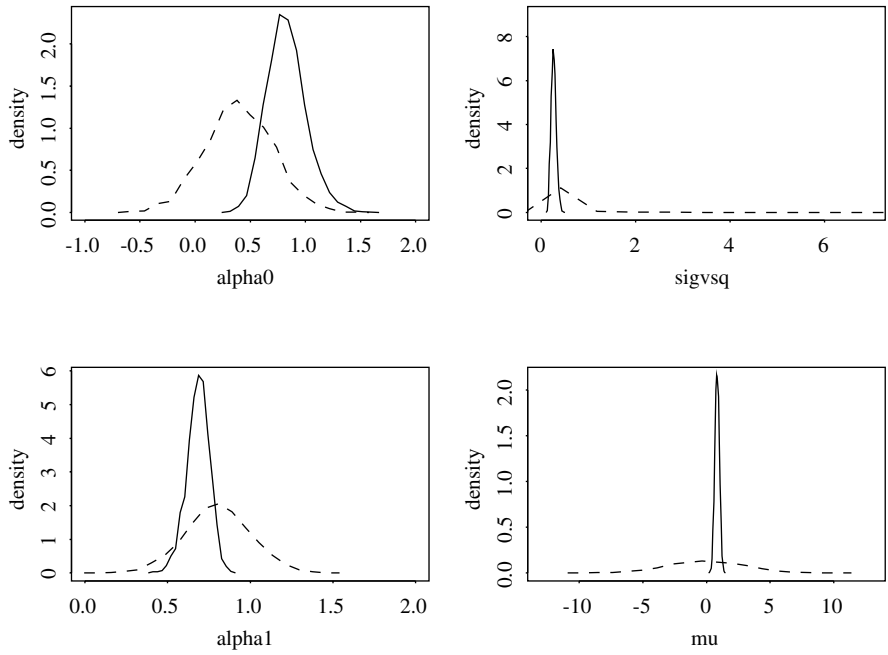


**Figure 10.4.** Density functions of prior and posterior distributions of parameters in a stochastic volatility model for the monthly log returns of S&P 500 index. The dashed line denotes prior density and solid line denotes the posterior density, which is based on results of Gibbs sampling with 5000 iterations. See the text for more details.

where $\boldsymbol{\alpha}_o = (0.4, 0.8)'$. For initial parameter values, we use the fitted values of the GARCH(1, 1) model in Eq. (10.26) for $\{h_t\}$ and set $\sigma_v^2 = 0.5$ and $\mu = 0.66$, which is the sample mean. In addition, $h_t$ is drawn by using the Griddy Gibbs with 500 grid points and the range of $h_t$ is $(0, 1.5s^2)$, where $s^2$ is the sample variance of the log return $r_t$.

We ran the Gibbs sampling for 5100 iterations, but discarded results of the first 100 iterations. Figure 10.4 shows the density functions of the prior and posterior distributions of the four coefficient parameters. The prior distributions used are relatively noninformative. The posterior distributions are concentrated especially for $\mu$ and $\sigma_v^2$. Figure 10.5 shows the time plots of fitted volatilities. The upper panel shows the posterior mean of $h_t$ over the 5000 iterations for each time point, whereas the lower panel shows the fitted values of the GARCH(1, 1) model in Eq. (10.26). The two plots exhibit a similar pattern.

The posterior mean and standard error of the four coefficients are as follows:

| Parameter | $\mu$ | $\alpha_0$ | $\alpha_1$ | $\sigma_v^2$ |
|---|---|---|---|---|
| Mean | 0.836 | 0.831 | 0.685 | 0.265 |
| St. Error | 0.177 | 0.183 | 0.069 | 0.056 |

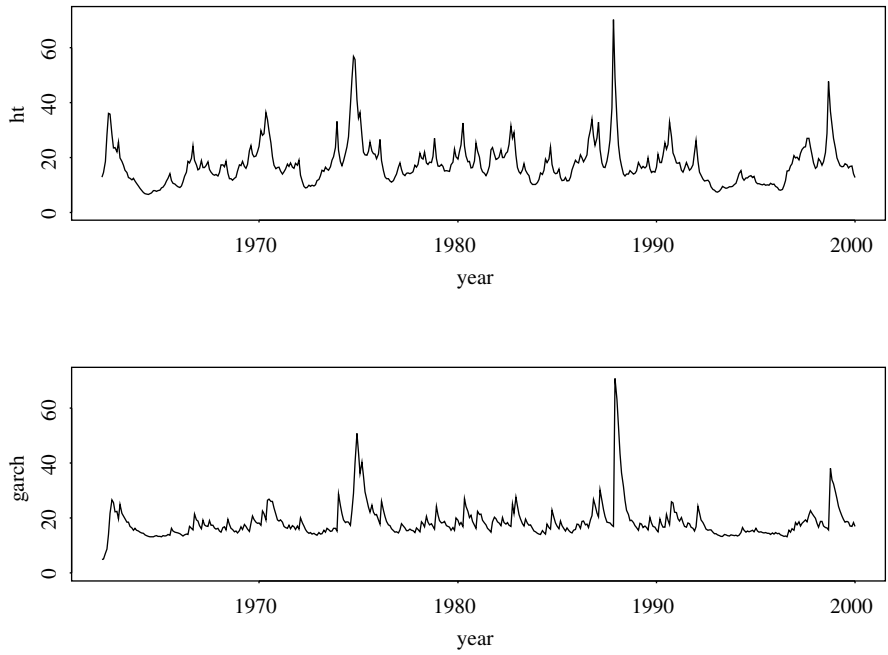


**Figure 10.5.** Time plots of fitted volatilities for monthly log returns of S&P 500 index from 1962 to 1999. The upper panel shows the posterior means of a Gibbs sampler with 5000 iterations. The lower panel shows the results of a GARCH(1, 1) model.

The posterior mean of $\alpha_1$ is 0.685, which is smaller than that obtained by Jacquier, Polson, and Rossi (1994) who used daily returns of the S&P 500 index. But it confirms the strong serial dependence in the volatility series. Finally, we have used different initial values and 3100 iterations for other Gibbs sampler, the posterior means of the parameters change slightly, but the series of posterior means of $h_t$ are stable.

### 10.7.2 Multivariate Stochastic Volatility Models

In this subsection, we study multivariate stochastic volatility models using the Cholesky decomposition of Chapter 9. We focus on the bivariate case, but the methods discussed also apply to the higher dimensional case. Based on the Cholesky decomposition, the innovation $\boldsymbol{a}_t$ of a return series $\boldsymbol{r}_t$ is transformed into $\boldsymbol{b}_t$ such that

$$b_{1t} = a_{1t}, \quad b_{2t} = a_{2t} - q_{21,t} b_{1t},$$

where $b_{2t}$ and $q_{21,t}$ can be interpreted as the residual and least squares estimate of the linear regression

$$a_{2t} = q_{21,t} a_{1t} + b_{2t}.$$

The conditional covariance matrix of $\boldsymbol{a}_t$ is parameterized by $\{g_{11,t}, g_{22,t}\}$ and $\{q_{21,t}\}$ as

$$\begin{bmatrix} \sigma_{11,t} & \sigma_{12,t} \\ \sigma_{21,t} & \sigma_{22,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ q_{21,t} & 1 \end{bmatrix} \begin{bmatrix} g_{11,t} & 0 \\ 0 & g_{22,t} \end{bmatrix} \begin{bmatrix} 1 & q_{21,t} \\ 0 & 1 \end{bmatrix}, \tag{10.28}$$

where $g_{ii,t} = \text{Var}(b_{it} \mid F_{t-1})$ and $b_{1t} \perp b_{2t}$. Thus, the quantities of interest are $g_{11,t}$, $g_{22,t}$, and $q_{21,t}$.

A simple bivariate stochastic volatility model for the return $\boldsymbol{r}_t = (r_{1t}, r_{2t})'$ is as follows:

$$\boldsymbol{r}_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{x}_t + \boldsymbol{a}_t \tag{10.29}$$

$$\ln g_{ii,t} = \alpha_{i0} + \alpha_{i1} \ln g_{ii,t-1} + v_{it}, \quad i = 1, 2 \tag{10.30}$$

$$q_{21,t} = \gamma_0 + \gamma_1 q_{21,t-1} + u_t, \tag{10.31}$$

where $\{\boldsymbol{a}_t\}$ is a sequence of serially uncorrelated Gaussian random vectors with mean zero and conditional covariance matrix $\boldsymbol{\Sigma}_t$ given by Eq. (10.28), $\boldsymbol{\beta}_0$ is a two-dimensional constant vector, $\boldsymbol{x}_t$ denotes the explanatory variables, and $\{v_{1t}\}$, $\{v_{2t}\}$, and $\{u_t\}$ are three independent Gaussian white noise series such that $\text{Var}(v_{it}) = \sigma_{iv}^2$ and $\text{Var}(u_t) = \sigma_u^2$. Again log transformation is used in Eq. (10.30) to ensure the positiveness of $g_{ii,t}$.

Let $\boldsymbol{G}_i = (g_{ii,1}, \ldots, g_{ii,n})'$, $\boldsymbol{G} = [\boldsymbol{G}_1, \boldsymbol{G}_2]$, and $\boldsymbol{Q} = (q_{21,1}, \ldots, q_{21,n})'$. The "traditional" parameters of the model in Eqs. (10.29)–(10.31) are $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, $\boldsymbol{\omega}_i = (\alpha_{i0}, \alpha_{i1}, \sigma_{iv}^2)$ for $i = 1, 2$, and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \sigma_u^2)$. The augmented parameters are $\boldsymbol{Q}$, $\boldsymbol{G}_1$, and $\boldsymbol{G}_2$. To estimate such a bivariate stochastic volatility model via Gibbs sampling, we use results of the univariate model in the previous subsection and two additional conditional posterior distributions. Specifically, we can draw random samples of

1. $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ row by row using the result (10.22);
2. $g_{11,t}$ using Eq. (10.23) with $a_t$ being replaced by $a_{1t}$;
3. $\boldsymbol{\omega}_1$ using exactly the same methods as those of the univariate case with $a_t$ replaced by $a_{1t}$.

To draw random samples of $\boldsymbol{\omega}_2$ and $g_{22,t}$, we need to compute $b_{2t}$. But this is easy because $b_{2t} = a_{2t} - q_{21,t} a_{1t}$ given the augmented parameter vector $\boldsymbol{Q}$. Furthermore, $b_{2t}$ is normally distributed with mean 0 and conditional variance $g_{22,t}$.

It remains to consider the conditional posterior distributions

$$f(\boldsymbol{\varpi} \mid \boldsymbol{Q}, \sigma_u^2), \quad f(\sigma_u^2 \mid \boldsymbol{Q}, \boldsymbol{\varpi}), \quad f(q_{21,t} \mid \boldsymbol{A}, \boldsymbol{G}, \boldsymbol{Q}_{-t}, \boldsymbol{\gamma}),$$

where $\boldsymbol{\varpi} = (\gamma_0, \gamma_1)'$ is the coefficient vector of Eq. (10.31) and $\boldsymbol{A}$ denotes the collection of $\boldsymbol{a}_t$, which is known if $\boldsymbol{R}$, $\boldsymbol{X}$, $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_1$ are given. Given $\boldsymbol{Q}$ and $\sigma_u^2$, model (10.31) is a simple Gaussian AR(1) model. Therefore, if the prior distribution of $\boldsymbol{\varpi}$

is bivariate normal with mean $\boldsymbol{\varpi}_o$ and covariance matrix $\boldsymbol{D}_o$, then the conditional posterior distribution of $\boldsymbol{\varpi}$ is also bivariate normal with mean $\boldsymbol{\varpi}_*$ and covariance matrix $\boldsymbol{D}_*$, where

$$\boldsymbol{D}_*^{-1} = \frac{\sum_{t=2}^{n} \boldsymbol{z}_t \boldsymbol{z}_t'}{\sigma_u^2} + \boldsymbol{D}_o^{-1}, \quad \boldsymbol{\varpi}_* = \boldsymbol{D}_* \left( \frac{\sum_{t=2}^{n} \boldsymbol{z}_t q_{21,t}}{\sigma_u^2} + \boldsymbol{D}_o^{-1} \boldsymbol{\varpi}_o \right),$$

where $\boldsymbol{z}_t = (1, q_{21,t-1})'$. Similarly, if the prior distribution of $\sigma_u^2$ is $(m\lambda)/\sigma_u^2 \sim \chi_m^2$, then the conditional posterior distribution of $\sigma_u^2$ is

$$\frac{m\lambda + \sum_{t=2}^{n} u_t^2}{\sigma_u^2} \sim \chi_{m+n-1}^2,$$

where $u_t = q_{21,t} - \gamma_0 - \gamma_1 q_{21,t-1}$. Finally,

$$\begin{aligned} & f(q_{21,t} \mid \boldsymbol{A}, \boldsymbol{G}, \boldsymbol{Q}_{-t}, \sigma_u^2, \boldsymbol{\varpi}) \qquad (10.32) \\ & \quad \propto f(b_{2t} \mid g_{22,t}) f(q_{21,t} \mid q_{21,t-1}, \boldsymbol{\varpi}, \sigma_u^2) f(q_{21,t+1} \mid q_{21,t}, \boldsymbol{\varpi}, \sigma_u^2) \\ & \quad \propto g_{22,t}^{-0.5} \exp[-(a_{2t} - q_{21,t} a_{1t})^2/(2g_{22,t})] \exp[-(q_{21,t} - \mu_t)^2/(2\sigma^2)], \end{aligned}$$

where $\mu_t = [\gamma_0(1-\gamma_1) + \gamma_1(q_{21,t-1} + q_{21,t+1})]/(1+\gamma_1^2)$ and $\sigma^2 = \sigma_u^2/(1+\gamma_1^2)$. In general, $\mu_t$ and $\sigma^2$ can be obtained by using the results of a missing value in an AR($p$) process. It turns out that Eq. (10.32) has a closed form distribution for $q_{21,t}$. Specifically, the first term of Eq. (10.32), which is the conditional distribution of $q_{21,t}$ given $g_{22,t}$ and $\boldsymbol{a}_t$, is normal with mean $a_{2t}/a_{1t}$ and variance $g_{22,t}/(a_{1t})^2$. The second term of the equation is also normal with mean $\mu_t$ and variance $\sigma^2$. Consequently, by Result 1 of Section 10.3, the conditional posterior distribution of $q_{21,t}$ is normal with mean $\mu_*$ and variance $\sigma_*^2$, where

$$\frac{1}{\sigma_*^2} = \frac{a_{1t}^2}{g_{22,t}} + \frac{1+\gamma_1^2}{\sigma_u^2}, \quad \mu_* = \sigma_*^2 \left( \frac{1+\gamma_1^2}{\sigma_u^2} \times \mu_t + \frac{a_{1t}^2}{g_{22,t}} \times \frac{a_{2t}}{a_{1t}} \right)$$

where $\mu_t$ is defined in Eq. (10.32).

**Example 10.4.** In this example, we study bivariate volatility models for the monthly log returns of IBM stock and the S&P 500 index from January 1962 to December 1999. This is an expanded version of Example 10.3 by adding the IBM returns. Figure 10.6 shows the time plots of the two return series. Let $\boldsymbol{r}_t = (\text{IBM}_t, \text{SP}_t)'$. If time-varying correlation GARCH models with Cholesky decomposition of Chapter 9 are entertained, we obtain the model

$$\boldsymbol{r}_t = \boldsymbol{\beta}_0 + \boldsymbol{a}_t \qquad (10.33)$$

$$g_{11,t} = \alpha_{10} + \alpha_{11} g_{11,t-1} + \alpha_{12} a_{1,t-1}^2 \qquad (10.34)$$

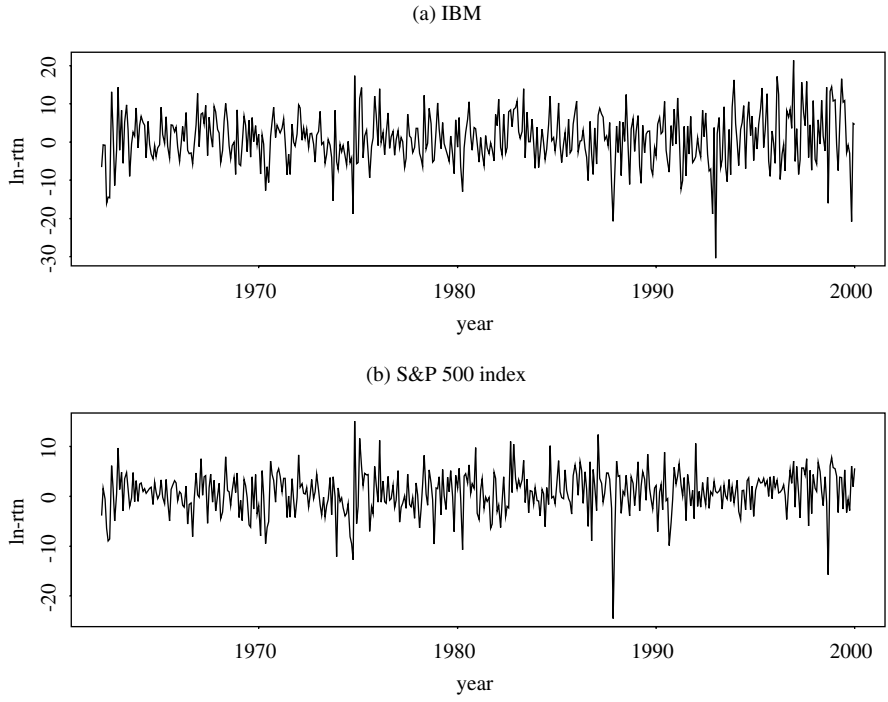


**Figure 10.6.** Time plots of monthly log returns of IBM stock and the S&P 500 index from 1962 to 1999.

$$g_{22,t} = \alpha_{20} + \alpha_{21} a_{1,t-1}^2 \tag{10.35}$$

$$q_{21,t} = \gamma_0, \tag{10.36}$$

where the estimates and their standard errors are given in Table 10.2(a). For comparison purpose, we employ the same mean equation in Eq. (10.33) and a stochastic volatility model similar to that in Eqs. (10.34)–(10.36). The volatility equations are

$$\ln g_{11,t} = \alpha_{10} + \alpha_{11} \ln g_{11,t-1} + v_{1t}, \quad \text{Var}(v_{1t}) = \sigma_{1v}^2 \tag{10.37}$$

$$\ln g_{22,t} = \alpha_{20} + v_{2t}, \quad \text{Var}(v_{2t}) = \sigma_{2v}^2 \tag{10.38}$$

$$q_{21,t} = \gamma_0 + u_t, \quad \text{Var}(u_t) = \sigma_u^2. \tag{10.39}$$

The prior distributions used are

$$\beta_{i0} \sim N(0.8, 4), \quad \boldsymbol{\alpha}_1 \sim N[(0.4, 0.8)', \text{diag}(0.16, 0.04)], \quad \alpha_{20} \sim N(5, 25),$$
$$\gamma_0 \sim N(0.4, .04), \quad \frac{10 \times 0.1}{\sigma_{1v}^2} \sim \chi_{10}^2, \quad \frac{5 \times 0.2}{\sigma_{2v}^2} \sim \chi_5^2, \quad \frac{5 \times 0.2}{\sigma_u^2} \sim \chi_5^2.$$

**Table 10.2. Estimation of Bivariate Volatility Models for Monthly Log Returns of IBM Stock and the S&P 500 Index from January 1962 to December 1999. The Stochastic Volatility Models Are Based on the Last 1000 Iterations of a Gibbs Sampling with 1300 Total Iterations.**

(a) Bivariate GARCH(1, 1) model with time-varying correlations

| Parameter | $\beta_{01}$ | $\beta_{02}$ | $\alpha_{10}$ | $\alpha_{11}$ | $\alpha_{12}$ | $\alpha_{20}$ | $\alpha_{21}$ | $\gamma_0$ |
|---|---|---|---|---|---|---|---|---|
| Estimate | 1.04 | 0.79 | 3.16 | 0.83 | 0.10 | 10.59 | 0.04 | 0.35 |
| Std. Error | 0.31 | 0.20 | 1.67 | 0.08 | 0.03 | 0.93 | 0.02 | 0.02 |

(b) Stochastic volatility model

| Parameter | $\beta_{01}$ | $\beta_{02}$ | $\alpha_{10}$ | $\alpha_{11}$ | $\sigma^2_{1v}$ | $\alpha_{20}$ | $\sigma^2_{2v}$ | $\gamma_0$ | $\sigma^2_u$ |
|---|---|---|---|---|---|---|---|---|---|
| Post.Mean | 0.86 | 0.84 | 0.52 | 0.86 | 0.08 | 1.81 | 0.39 | 0.39 | 0.08 |
| Std. Error | 0.30 | 0.18 | 0.18 | 0.05 | 0.03 | 0.11 | 0.06 | 0.03 | 0.02 |

These prior distributions are relatively noninformative. We ran the Gibbs sampling for 1300 iterations, but discarded results of the first 300 iterations. The random samples of $g_{ii,t}$ were drawn by Griddy Gibbs with 400 grid points in the intervals $[0, 1.5s_i^2]$, where $s_i^2$ is the sample variance of the log return $r_{it}$. Posterior means and standard errors of the "traditional" parameters of the bivariate stochastic volatility model are given in Table 10.2(b).

To check for convergence of the Gibbs sampling, we ran the procedure several times with different starting values and numbers of iterations. The results are stable. For illustration, Figure 10.7 shows the scatterplots of various quantities for two different Gibbs samples. The first Gibbs sample is based on $300 + 1000$ iterations, and the second Gibbs sample is based on $500 + 3000$ iterations, where $M + N$ denotes that the total number of Gibbs iterations is $M + N$, but results of the first $M$ iterations are discarded. The scatterplots shown are posterior means of $g_{11,t}$, $g_{22,t}$, $q_{21,t}$, $\sigma_{22,t}$, $\sigma_{21,t}$, and the correlation $\rho_{21,t}$. The line $y = x$ is added to each plot to show the closeness of the posterior means. The stability of the Gibbs sampling results is clearly seen.

It is informative to compare the GARCH model with time-varying correlations in Eqs. (10.33)–(10.36) with the stochastic volatility model. First, as expected, the mean equations of the two models are essentially identical. Second, Figure 10.8 shows the time plots of fitted volatilities for IBM stock return. The upper panel is for the GARCH model, and the lower panel shows the posterior mean of the stochastic volatility model. The two models show similar volatility characteristics; they exhibit volatility clusterings and indicate an increasing trend in volatility. However, the GARCH model produces higher peak volatility values. Third, Figure 10.9 shows the time plots of fitted volatilities for the S&P 500 index return. The GARCH model produces an extra volatility peak around 1993. This additional peak does not appear in the univariate analysis shown in Figure 10.5. It seems that for this particular instance the bivariate GARCH model produces a spurious volatility peak. This spurious peak is induced by its dependence on IBM returns and does not appear in the stochastic volatility model. Indeed, the fitted volatilities of S&P 500 index return by

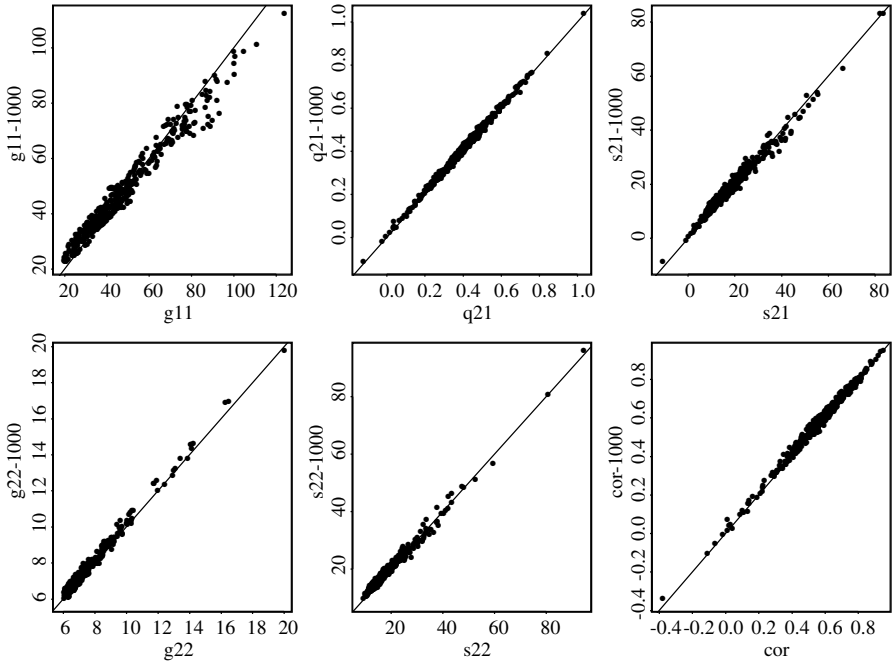


**Figure 10.7.** Scatterplots of posterior means of various statistics of two different Gibbs samples for the bivariate stochastic volatility model for monthly log returns of IBM stock and the S&P 500 index. The x-axis denotes results based on $500 + 3000$ iterations and the y-axis denotes results based on $300 + 1000$ iterations. The notation is defined in the text.

the bivariate stochastic volatility model are similar to that of the univariate analysis. Fourth, Figure 10.10 shows the time plots of fitted conditional correlations. Here the two models differ substantially. The correlations of the GARCH model are relatively smooth and positive with a mean value 0.55 and standard deviation 0.11. However, the correlations produced by the stochastic volatility model vary markedly from one month to another with a mean value 0.57 and standard deviation 0.17. Furthermore, there are isolated occasions in which the correlation is negative. The difference is understandable because $q_{21,t}$ contains the random shock $u_t$ in the stochastic volatility model.

***Remark:*** The Gibbs sampling estimation applies to other bivariate stochastic volatility models. The conditional posterior distributions needed require some extensions of those discussed in this section, but they are based on the same ideas.

## 10.8 MARKOV SWITCHING MODELS

The Markov switching model is another econometric model for which MCMC methods enjoy many advantages over the traditional likelihood method. McCulloch and

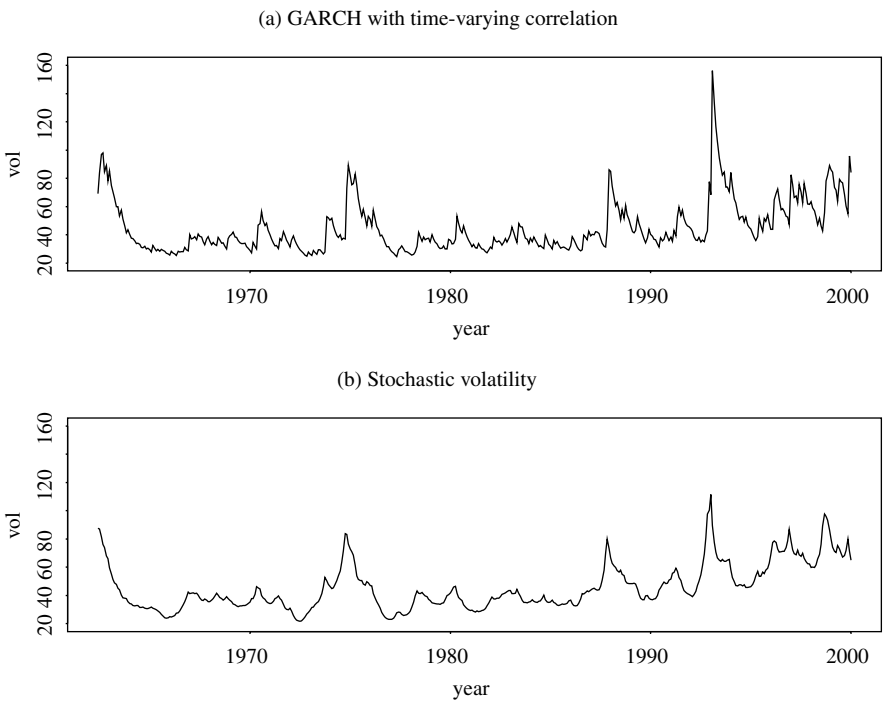


**Figure 10.8.** Time plots of fitted volatilities for monthly log returns of IBM stock from 1962 to 1999: (a) a GARCH model with time-varying correlations, and (b) a bivariate stochastic volatility model estimated by Gibbs sampling with 300 + 1000 iterations.

Tsay (1994) discuss a Gibbs sampling procedure to estimate such a model when the volatility in each state is constant over time. These authors applied the procedure to estimate a Markov switching model with different dynamics and mean levels for different states to the quarterly growth rate of U.S. real gross national product, seasonally adjusted, and obtained some interesting results. For instance, the dynamics of the growth rate are significantly different between periods of economic "contraction" and "expansion." Since this chapter is concerned with asset returns, we focus on models with volatility switching.

Suppose that an asset return $r_t$ follows a simple two-state Markov switching model with different risk premiums and different GARCH dynamics:

$$r_t = \begin{cases} \beta_1\sqrt{h_t} + \sqrt{h_t}\epsilon_t, & h_t = \alpha_{10} + \alpha_{11}h_{t-1} + \alpha_{12}a_{t-1}^2 \quad \text{if } s_t = 1 \\ \beta_2\sqrt{h_t} + \sqrt{h_t}\epsilon_t, & h_t = \alpha_{20} + \alpha_{21}h_{t-1} + \alpha_{22}a_{t-1}^2 \quad \text{if } s_t = 2, \end{cases} \tag{10.40}$$

where $a_t = \sqrt{h_t}\epsilon_t$, $\{\epsilon_t\}$ is a sequence of Gaussian white noises with mean zero and variance 1, and the parameters $\alpha_{ij}$ satisfy some regularity conditions so that the unconditional variance of $a_t$ exists. The probability transition from one state to

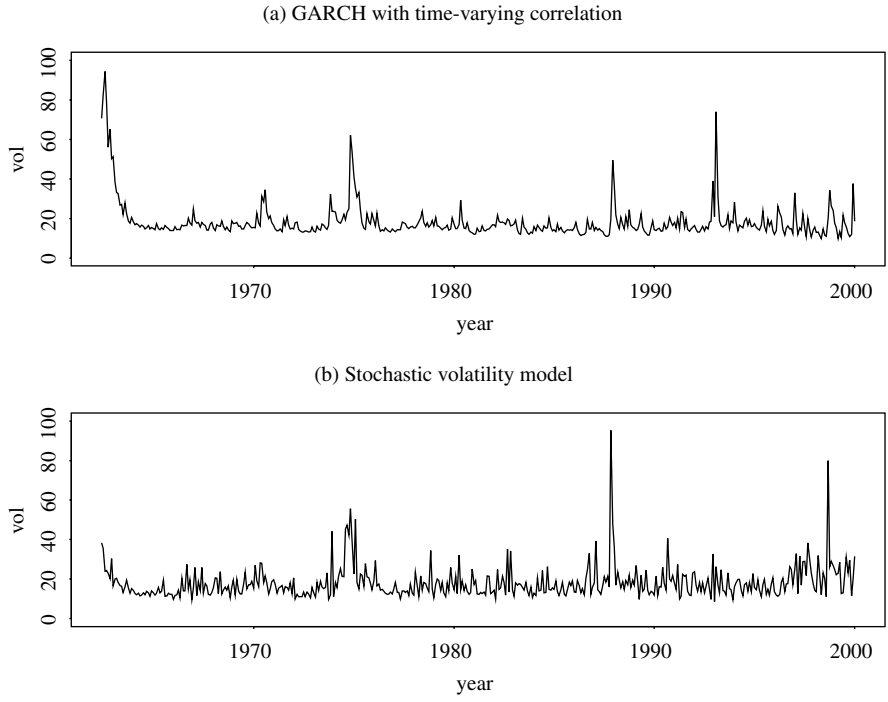


**Figure 10.9.** Time plots of fitted volatilities for monthly log returns of the S&P 500 index from 1962 to 1999: (a) a GARCH model with time-varying correlations, and (b) a bivariate stochastic volatility model estimated by Gibbs sampling with $300 + 1000$ iterations.

another is governed by

$$P(s_t = 2 \mid s_{t-1} = 1) = e_1, \quad P(s_t = 1 \mid s_{t-1} = 2) = e_2, \tag{10.41}$$

where $0 < e_i < 1$. A small $e_i$ means that the return series has a tendency to stay in the $i$th state with expected duration $1/e_i$. For the model in Eq. (10.40) to be identifiable, we assume that $\beta_2 > \beta_1$ so that State 2 is associated with higher risk premium. This is not a critical restriction because it is used to achieve uniqueness in labeling the states. A special case of the model results if $\alpha_{1j} = \alpha_{2j}$ for all $j$ so that the model assumes a GARCH model for all states. However, if $\beta_i \sqrt{h_t}$ is replaced by $\beta_i$, then model (10.40) reduces to a simple Markov switching GARCH model.

Model (10.40) is a Markov switching GARCH-M model. For simplicity, we assume that the initial volatility $h_1$ is given with value equal to the sample variance of $r_t$. A more sophisticated analysis is to treat $h_1$ as a parameter and estimate it jointly with other parameters. We expect the effect of fixing $h_1$ will be negligible in most applications, especially when the sample size is large. The "traditional" parameters of the Markov switching GARCH-M model are $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, $\boldsymbol{\alpha}_i = (\alpha_{i0}, \alpha_{i1}, \alpha_{i2})'$ for $i = 1$ and 2, and the transition probabilities $\boldsymbol{e} = (e_1, e_2)'$. The state vector

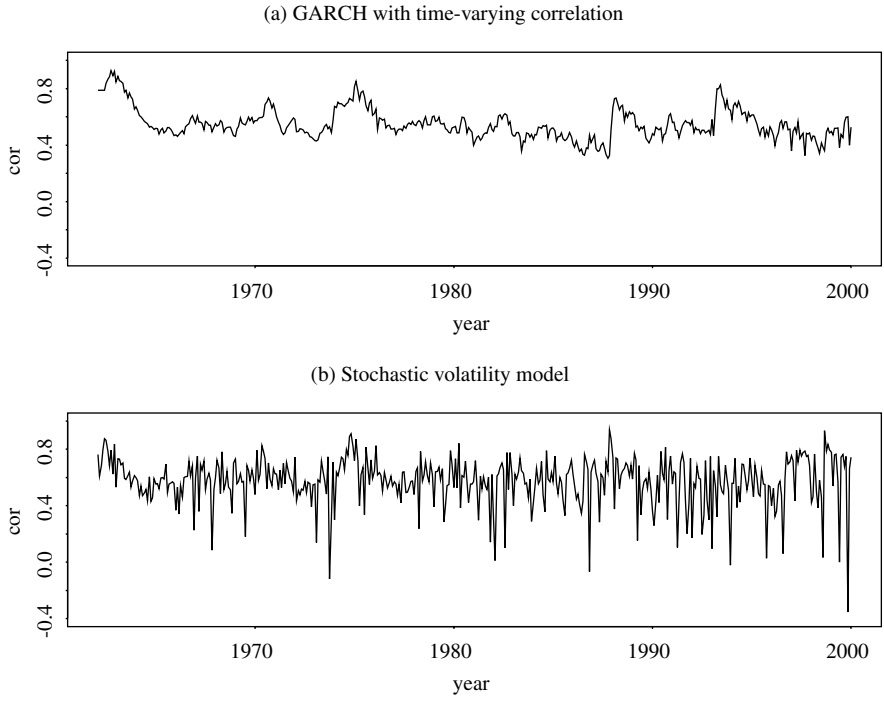


**Figure 10.10.** Time plots of fitted correlation coefficients between monthly log returns of IBM stock and the S&P 500 index from 1962 to 1999: (a) a GARCH model with time-varying correlations, and (b) a bivariate stochastic volatility model estimated by Gibbs sampling with $300 + 1000$ iterations.

$\boldsymbol{S} = (s_1, s_2, \ldots, s_n)'$ contains the augmented parameters. The volatility vector $\boldsymbol{H} = (h_2, \ldots, h_n)'$ can be computed recursively if $h_1$, $\boldsymbol{\alpha}_i$, and the state vector $\boldsymbol{S}$ are given.

Dependence of the return on volatility in model (10.40) implies that the return is also serially correlated. The model thus has some predictivity in the return. However, states of the future returns are unknown and a prediction produced by the model is necessarily a mixture of those over possible state configurations. This often results in high uncertainty in point prediction of future returns.

Turn to estimation. The likelihood function of model (10.40) is complicated as it is a mixture over all possible state configurations. Yet the Gibbs sampling approach only requires the following conditional posterior distributions:

$$f(\boldsymbol{\beta} \mid \boldsymbol{R}, \boldsymbol{S}, \boldsymbol{H}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2), \quad f(\boldsymbol{\alpha}_i \mid \boldsymbol{R}, \boldsymbol{S}, \boldsymbol{H}, \boldsymbol{\alpha}_{j \neq i}),$$
$$P(\boldsymbol{S} \mid \boldsymbol{R}, h_1, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2), \quad f(e_i \mid \boldsymbol{S}), \quad i = 1, 2,$$

where $\boldsymbol{R}$ is the collection of observed returns. For simplicity, we use conjugate prior distributions discussed in Section 10.3—that is,

$$\beta_i \sim N(\beta_{io}, \sigma_{io}^2), \quad e_i \sim \text{Beta}(\gamma_{i1}, \gamma_{i2}).$$

The prior distribution of parameter $\alpha_{ij}$ is uniform over a properly specified interval. Since $\alpha_{ij}$ is a nonlinear parameter of the likelihood function, we use the Griddy Gibbs to draw its random realizations. A uniform prior distribution simplifies the computation involved. Details of the conditional posterior distributions are given below:

**1.** The posterior distribution of $\beta_i$ only depends on the data in State $i$. Define

$$r_{it} = \begin{cases} r_t/\sqrt{h_t} & \text{if } s_t = i \\ 0 & \text{otherwise.} \end{cases}$$

Then we have

$$r_{it} = \beta_i + \epsilon_t, \quad \text{for} \quad s_t = i.$$

Therefore, information of the data on $\beta_i$ is contained in the sample mean of $r_{it}$. Let $\bar{r}_i = \sum_{s_t=i} r_{it}/n_i$, where the summation is over all data points in State $i$ and $n_i$ is the number of data points in State $i$. Then the conditional posterior distribution of $\beta_i$ is normal with mean $\beta_i^*$ and variance $\sigma_{i*}^2$, where

$$\frac{1}{\sigma_{i*}^2} = n_i + \frac{1}{\sigma_{io}^2}, \quad \beta_i^* = \sigma_{i*}^2 \left( n_i \bar{r}_i + \beta_{io}/\sigma_{io}^2 \right), \quad i = 1, 2.$$

**2.** Next, the parameter $\alpha_{ij}$ can be drawn one by one using the Griddy Gibbs method. Given $h_1$, $\boldsymbol{S}$, $\boldsymbol{\alpha}_{v \neq i}$ and $\alpha_{iv}$ with $v \neq j$, the conditional posterior distribution function of $\alpha_{ij}$ does not correspond to a well-known distribution, but it can be evaluated easily as

$$f(\alpha_{ij} \mid .) \propto -\frac{1}{2} \sum_{s_t=i} \left[ \ln h_t + \frac{(r_t - \beta_i \sqrt{h_t})^2}{h_t} \right],$$

where $h_t$ contains $\alpha_{ij}$. We evaluate this function over a grid of points for $\alpha_{ij}$ over a properly specified interval. For example, $0 \leq \alpha_{11} < 1 - \alpha_{12}$.

**3.** The conditional posterior distribution of $e_i$ only involves $\boldsymbol{S}$. Let $\ell_1$ be the number of switches from State 1 to State 2 and $\ell_2$ be the number of switches from State 2 to State 1 in $\boldsymbol{S}$. Also, let $n_i$ be the number of data points in State $i$. Then by Result 3 of conjugate prior distributions, the posterior distribution of $e_i$ is $\text{Beta}(\gamma_{i1} + \ell_i, \gamma_{i2} + n_i - \ell_i)$.

**4.** Finally, elements of $\boldsymbol{S}$ can be drawn one by one. Let $\boldsymbol{S}_{-j}$ be the vector obtained by removing $s_j$ from $\boldsymbol{S}$. Given $\boldsymbol{S}_{-j}$ and other information, $s_j$ can assume two possibilities (i.e., $s_j = 1$ or $s_j = 2$), and its conditional posterior distribution is

$$P(s_j \mid .) \propto \prod_{t=j}^{n} f(a_t \mid \boldsymbol{H}) P(s_j \mid \boldsymbol{S}_{-j}).$$

The probability

$$P(s_j = i \mid \boldsymbol{S}_{-j}) = P(s_j = i \mid s_{j-1}, s_{j+1}), \quad i = 1, 2$$

can be computed by the Markov transition probabilities in Eq. (10.41). In addition, assuming $s_j = i$, one can compute $h_t$ for $t \geq j$ recursively. The relevant likelihood function, denoted by $L(s_j)$, is given by

$$L(s_j = i) \equiv \prod_{t=j}^{n} f(a_t \mid \boldsymbol{H}) \propto \exp(f_{ji}), \quad f_{ji} = \sum_{t=j}^{n} -\frac{1}{2}\left[\ln(h_t) + \frac{a_t^2}{h_t}\right],$$

for $i = 1$ and 2, where $a_t = r_t - \beta_1\sqrt{h_t}$ if $s_t = 1$ and $a_t = r_t - \beta_2\sqrt{h_t}$ otherwise. Consequently, the conditional posterior probability of $s_j = 1$ is

$$P(s_j = 1 \mid .) = \frac{P(s_j = 1 \mid s_{j-1}, s_{j+1}) L(s_j = 1)}{P(s_j = 1 \mid s_{j-1}, s_{j+1}) L(s_j = 1) + P(s_j = 2 \mid s_{j-1}, s_{j+1}) L(s_j = 2)}.$$

The state $s_j$ can then be drawn easily using a uniform distribution on the unit interval [0, 1].

***Remark:*** Since $s_j$ and $s_{j+1}$ are highly correlated when $e_1$ and $e_2$ are small, it is more efficient to draw several $s_j$ jointly. However, the computation involved in enumerating the possible state configurations increases quickly with the number of states drawn jointly.

**Example 10.5.** In this example, we consider the monthly log stock returns of General Electric Company from January 1926 to December 1999 for 888 observations. The returns are in percentages and shown in Figure 10.11(a). For comparison purpose, we start with a GARCH-M model for the series and obtain

$$\begin{aligned} r_t &= 0.182\sqrt{h_t} + a_t, \quad a_t = \sqrt{h_t}\epsilon_t, \\ h_t &= 0.546 + 1.740h_{t-1} - 0.775h_{t-2} + 0.025a_{t-1}^2, \end{aligned} \tag{10.42}$$

where $r_t$ is the monthly log return and $\{\epsilon_t\}$ is a sequence of independent Gaussian white noises with mean zero and variance 1. All parameter estimates are highly significant with $p$ values less than 0.0006. The Ljung–Box statistics of the standardized residuals and their squared series fail to suggest any model inadequacy. It is reassuring to see that the risk premium is positive and significant. The GARCH model in

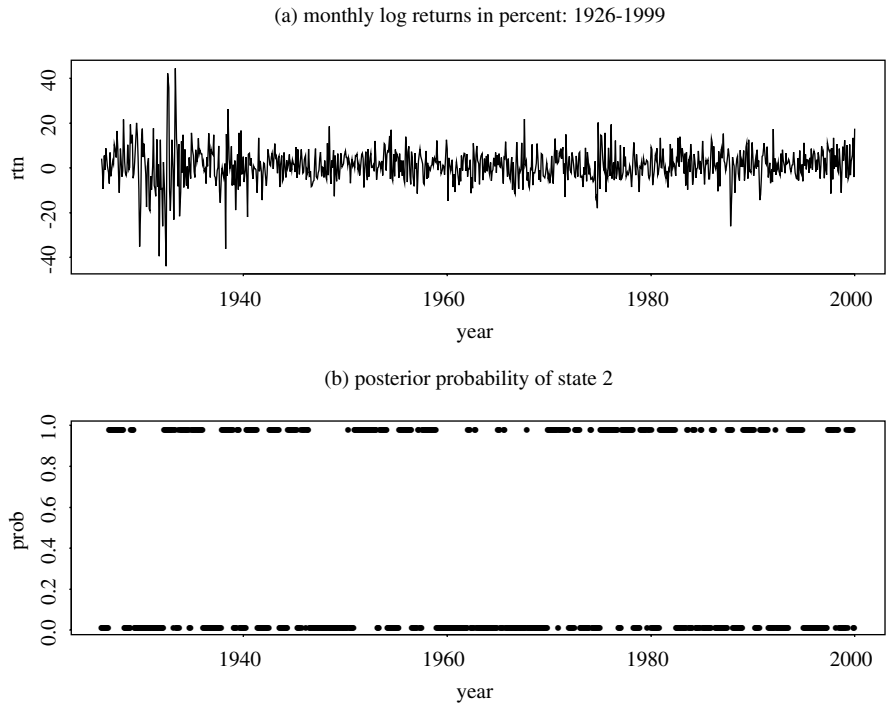


**Figure 10.11.** (a) Time plot of the monthly log returns, in percentages, of GE stock from 1926 to 1999. (b) Time plot of the posterior probability of being in State 2 based on results of the last 2000 iterations of a Gibbs sampling with 5000 + 2000 total iterations. The model used is a two-state Markov switching GARCH-M model.

Eq. (10.42) can be written as

$$(1 - 1.765B + 0.775B^2)a_t^2 = 0.546 + (1 - 0.025B)\eta_t,$$

where $\eta_t = a_t^2 - h_t$ and $B$ is the back-shift operator such that $Ba_t^2 = a_{t-1}^2$. As discussed in Chapter 3, the prior equation can be regarded as an ARMA(2, 1) model with nonhomogeneous innovations for the squared series $a_t^2$. The AR polynomial can be factorized as $(1 - 0.945B)(1 - 0.820B)$, indicating two real characteristic roots with magnitudes less than 1. Consequently, the unconditional variance of $r_t$ is finite and equal to $0.546/(1 - 1.765 + 0.775) \approx 49.64$.

Turn to Markov switching models. We use the following prior distributions:

$$\beta_1 \sim N(0.3, 0.09), \quad \beta_2 \sim N(1.3, 0.09), \quad \epsilon_i \sim \text{Beta}(5, 95).$$

The initial parameter values used are (a) $e_i = 0.1$, (b) $s_1$ is a Bernoulli trial with equal probabilities and $s_t$ is generated sequentially using the initial transition probabilities, and (c) $\boldsymbol{\alpha}_1 = (1.0, 0.6, 0.2)'$ and $\boldsymbol{\alpha}_2 = (2, 0.7, 0.1)'$. Gibbs samples of

**Table 10.3. A Fitted Markov Switching GARCH-M Model for the Monthly Log Returns of GE Stock from January 1926 to December 1999. The Numbers Shown Are the Posterior Means and Standard Deviations Based on a Gibbs Sampling With 5000 + 2000 Iterations. Results of the First 5000 Iterations Are Discarded. The Prior Distributions and Initial Parameter Estimates Are Given in the Text.**

| (a) | State 1 | | | | |
|---|---|---|---|---|---|
| Parameter | $\beta_1$ | $e_1$ | $\alpha_{10}$ | $\alpha_{11}$ | $\alpha_{12}$ |
| Post. Mean | 0.111 | 0.089 | 2.070 | 0.844 | 0.033 |
| Post. Std. | 0.043 | 0.012 | 1.001 | 0.038 | 0.033 |
| (b) | State 2 | | | | |
| Parameter | $\beta_2$ | $e_2$ | $\alpha_{20}$ | $\alpha_{21}$ | $\alpha_{22}$ |
| Post. Mean | 0.247 | 0.112 | 2.740 | 0.869 | 0.068 |
| Post. Std. | 0.050 | 0.014 | 1.073 | 0.031 | 0.024 |
| | Difference between States | | | | |
| Parameter | $\beta_2 - \beta_1$ | $e_2 - e_1$ | $\alpha_{20} - \alpha_{10}$ | $\alpha_{21} - \alpha_{11}$ | $\alpha_{22} - \alpha_{12}$ |
| Post. Mean | 0.135 | 0.023 | 0.670 | 0.026 | −0.064 |
| Post. Std. | 0.063 | 0.019 | 1.608 | 0.050 | 0.043 |

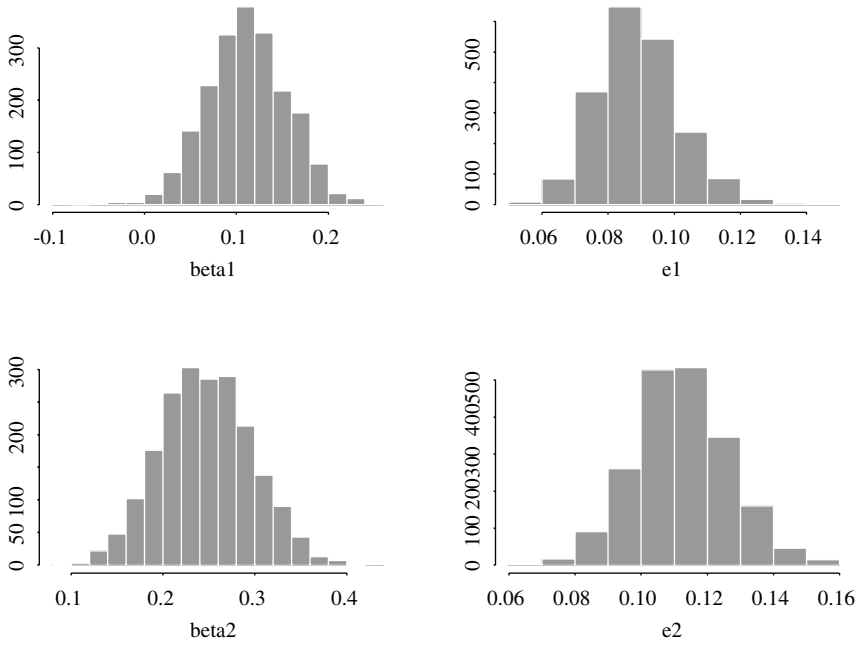


**Figure 10.12.** Histograms of the risk premium and transition probabilities of a two-state Markov switching GARCH-M model for the monthly log returns of GE stock from 1926 to 1999. The results are based on the last 2000 iterations of a Gibbs sampling with 5000 + 2000 total iterations.

$\alpha_{ij}$ are drawn using the Griddy Gibbs with 400 grid points, equally spaced over the following ranges: $\alpha_{i0} \in [0, 6.0]$, $\alpha_{i1} \in [0, 1]$, and $\alpha_{i2} \in [0, 0.5]$. In addition, we implement the constraints $\alpha_{i1} + \alpha_{i2} < 1$ for $i = 1, 2$. The Gibbs sampler is run for $5000 + 2000$ iterations, but only results of the last 2000 iterations are used to make inference.

Table 10.3 shows the posterior means and standard deviations of parameters of the Markov switching GARCH-M model in Eq. (10.40). In particular, it also contains some statistics showing the difference between the two states such as $\theta = \beta_2 - \beta_1$. The difference between the risk premiums is statistically significant at the 5% level. The differences in posterior means of the volatility parameters between the two states appear to be insignificant. Yet the posterior distributions of volatility parameters show some different characteristics. Figures 10.12 and 10.13 show the histograms of all parameters in the Markov switching GARCH-M model. They exhibit some differences between the two states. Figure 10.14 shows the time plot of the persistent parameter $\alpha_{i1} + \alpha_{i2}$ for the two states. It shows that the persistent parameter of State 1 reaches the boundary 1.0 frequently, but that of State 2 does not. The expected

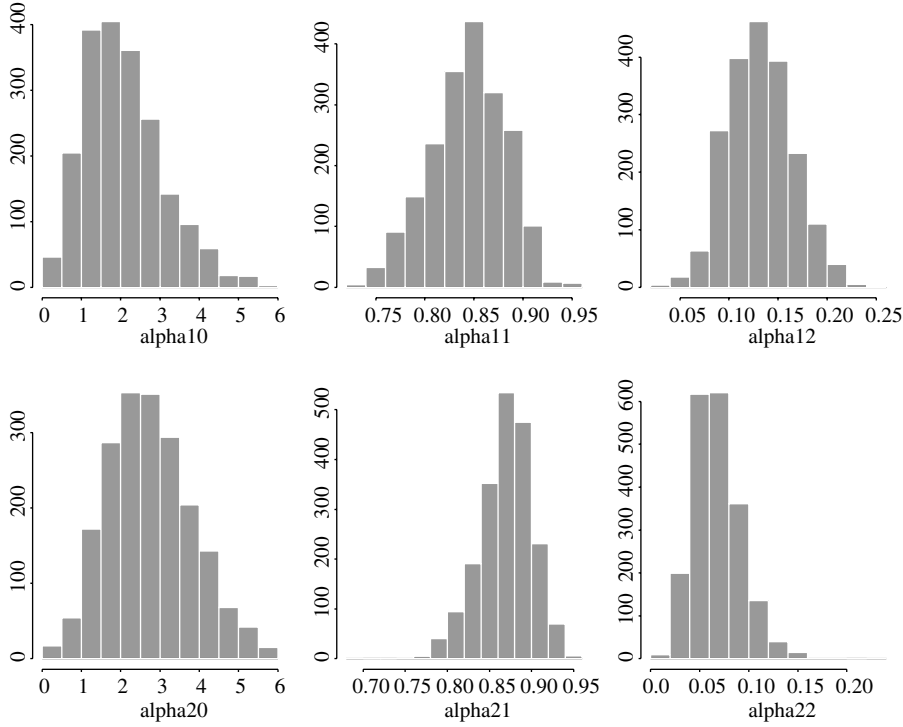


**Figure 10.13.** Histograms of volatility parameters of a two-state Markov switching GARCH-M model for the monthly log returns of GE stock from 1926 to 1999. The results are based on the last 2000 iterations of a Gibbs sampling with 5000 + 2000 total iterations.

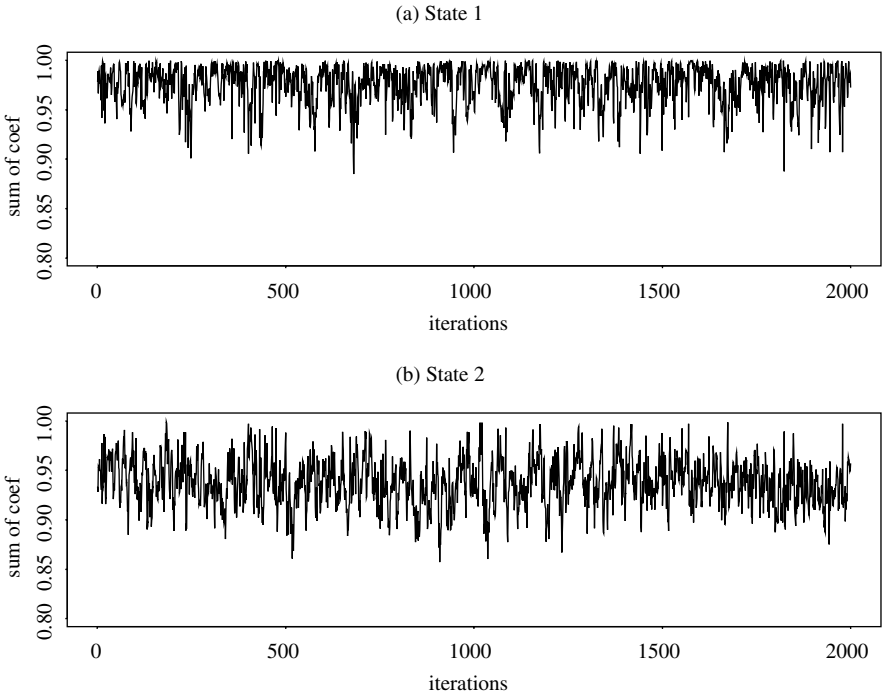


**Figure 10.14.** Time plots of the persistent parameter $\alpha_{i1}+\alpha_{i2}$ of a two-state Markov switching GARCH-M model for the monthly log returns of GE stock from 1926 to 1999. The results are based on the last 2000 iterations of a Gibbs sampling with 5000 + 2000 total iterations.

durations of the two states are about 11 and 9 months, respectively. Figure 10.11(b) shows the posterior probability of being in State 2 for each observation.

Finally, Figure 10.15 shows the fitted volatility series of the simple GARCH-M model in Eq. (10.42) and the Markov switching GARCH-M model in Eq. (10.40). The two fitted volatility series show similar pattern and are consistent with the behavior of the squared log returns. The simple GARCH-M model produces a smoother volatility series with lower estimated volatilities.

## 10.9 FORECASTING

Forecasting under the MCMC framework can be done easily. The procedure is simply to use the fitted model in each Gibbs iteration to generate samples for the forecasting period. In a sense, forecasting here is done by using the fitted model to simulate realizations for the forecasting period. We use the univariate stochastic volatility model to illustrate the procedure; forecasts of other models can be obtained by the same method.

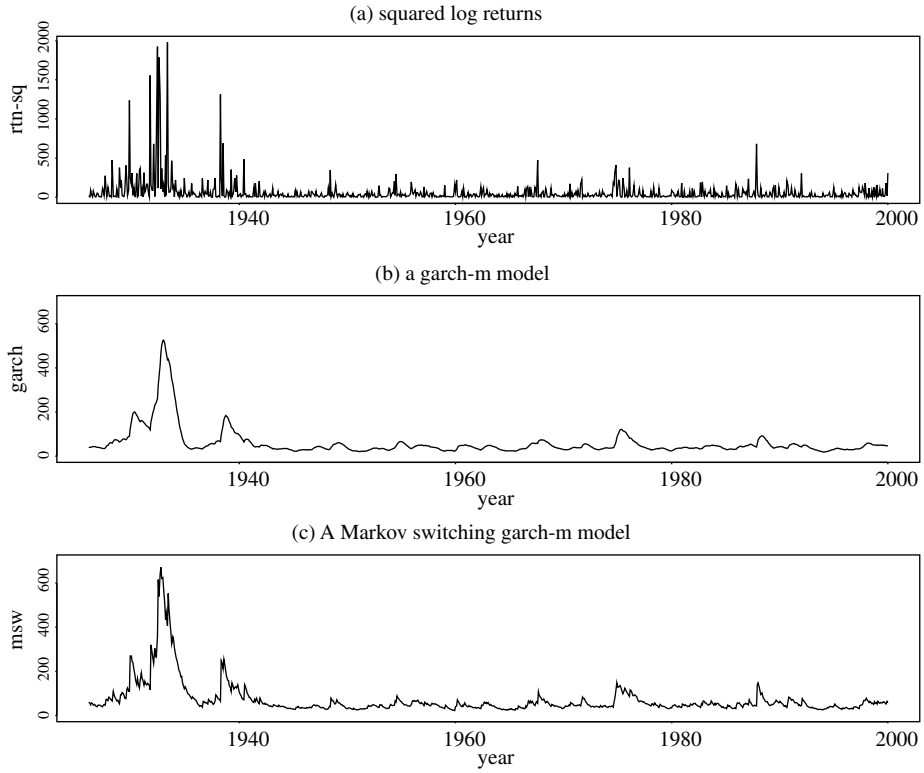


**Figure 10.15.** Fitted volatility series for the monthly log returns of GE stock from 1926 to 1999: (a) The squared log returns, (b) the GARCH-M model in Eq. (10.42), and (c) the two-state Markov switching GARCH-M model in Eq. (10.40).

Consider the stochastic volatility model in Eqs. (10.20) and (10.21). Suppose that there are $n$ returns available and we are interested in predicting the return $r_{n+i}$ and volatility $h_{n+i}$ for $i = 1, \ldots, \ell$, where $\ell > 0$. Assume that the explanatory variables $x_{jt}$ in Eq. (10.20) are either available or can be predicted sequentially during the forecasting period. Recall that estimation of the model under the MCMC framework is done by Gibbs sampling, which draws parameter values from their conditional posterior distributions iteratively. Denote the parameters by $\boldsymbol{\beta}_j = (\beta_{0,j}, \ldots, \beta_{p,j})'$, $\boldsymbol{\alpha}_j = (\alpha_{0,j}, \alpha_{1,j})'$, and $\sigma^2_{v,j}$ for the $j$th Gibbs iteration. In other words, at the $j$th Gibbs iteration, the model is

$$r_t = \beta_{0,j} + \beta_{1,j} x_{1t} + \cdots + \beta_{p,j} x_{pt} + a_t \tag{10.43}$$

$$\ln h_t = \alpha_{0,j} + \alpha_{1,j} \ln h_{t-1} + v_t, \quad \text{Var}(v_t) = \sigma^2_{v,j}. \tag{10.44}$$

We can use this model to generate a realization of $r_{n+i}$ and $h_{n+i}$ for $i = 1, \ldots, \ell$. Denote the simulated realizations by $r_{n+i,j}$ and $h_{n+i,j}$, respectively. These realizations are generated as follows:

- Draw a random sample $v_{n+1}$ from $N(0, \sigma^2_{v,j})$ and use Eq. (10.44) to compute $h_{n+1,j}$.
- Draw a random sample $\epsilon_{n+1}$ from $N(0, 1)$ to obtain $a_{n+1,j} = \sqrt{h_{n+1,j}}\epsilon_{n+1}$ and use Eq. (10.43) to compute $r_{n+1,j}$.
- Repeat the prior two steps sequentially for $n + i$ with $i = 2, \ldots, \ell$.

If we run a Gibbs sampling for $M + N$ iterations in model estimation, we only need to compute the forecasts for the last $N$ iterations. This results in a random sample for $r_{n+i}$ and $h_{n+i}$. More specifically, we obtain

$$\{r_{n+1,j}, \ldots, r_{n+\ell,j}\}_{j=1}^{N}, \quad \{h_{n+1,j}, \ldots, h_{n+\ell,j}\}_{j=1}^{N}.$$

These two random samples can be used to make inference. For example, point forecasts of the return $r_{n+i}$ and volatility $h_{n+i}$ are simply the sample means of the two random samples. Similarly, the sample standard deviations can be used as the variances of forecast errors. To improve the computational efficiency in volatility forecast, importance sampling can be used; see Gelman, Carlin, Stern, and Rubin (1995, p.307).

**Example 10.6.** (Example 10.3 continued.) As a demonstration, we consider the monthly log return series of S&P 500 index from 1962 to 1999. Table 10.4 gives the point forecasts of the return and its volatility for five forecast horizons starting with December 1999. Both the GARCH model in Eq. (10.26) and the stochastic volatility model in Eq. (10.27) are used in the forecasting. The volatility forecasts of the GARCH(1, 1) model increase gradually with the forecast horizon to the unconditional variance $3.349/(1 - 0.086 - 0.735) = 18.78$. The volatility forecasts of the stochastic volatility model are higher than those of the GARCH model. This is understandable because the stochastic volatility model takes into consideration the param-

**Table 10.4. Volatility Forecasts for the Monthly Log Return of S&P 500 Index. The Data Span Is From January 1962 to December 1999 and the Forecast Origin Is December 1999. Forecasts of the Stochastic Volatility Model Are Obtained by a Gibbs Sampling with 2000 + 2000 Iterations.**

| (a) | Log return | | | | |
|---|---|---|---|---|---|
| Horizon | 1 | 2 | 3 | 4 | 5 |
| GARCH | 0.66 | 0.66 | 0.66 | 0.66 | 0.66 |
| SVM | 0.53 | 0.78 | 0.92 | 0.88 | 0.84 |
| (b) | Volatility | | | | |
| Horizon | 1 | 2 | 3 | 4 | 5 |
| GARCH | 17.98 | 18.12 | 18.24 | 18.34 | 18.42 |
| SVM | 19.31 | 19.36 | 19.35 | 19.65 | 20.13 |

eter uncertainty in producing forecasts. In contrast, the GARCH model assumes that the parameters are fixed and given in Eq. (10.26). This is an important difference and is one of the reasons that GARCH models tend to underestimate the volatility in comparison with the implied volatility obtained from derivative pricing.

***Remark:*** Besides the advantage of taking into consideration parameter uncertainty in forecast, the MCMC method produces in effect a predictive distribution of the volatility of interest. The predictive distribution is more informative than a simple point forecast. It can be used, for instance, to obtain the quantiles needed in Value at Risk calculation.

## 10.10 OTHER APPLICATIONS

The MCMC method is applicable to many other financial problems. For example, Zhang, Russell, and Tsay (2000) use it to analyze information determinants of bid and ask quotes, McCulloch and Tsay (2000) use the method to estimate a hierarchical model for IBM transaction data, and Eraker (2001) and Elerian, Chib and Shephard (2001) use it to estimate diffusion equations. The method is also useful in Value at Risk calculation because it provides a natural way to evaluate predictive distributions. The main question is not whether the methods can be used in most financial applications, but how efficient the methods can become. Only time and experience can provide an adequate answer to the question.

## EXERCISES

**1.** Suppose that $x$ is normally distributed with mean $\mu$ and variance 4. Assume that the prior distribution of $\mu$ is also normal with mean 0 and variance 25. What is the posterior distribution of $\mu$ given the data point $x$?

**2.** Consider the linear regression model with time-series errors in Section 10.5. Assume that $z_t$ is an AR($p$) process (i.e., $z_t = \phi_1 z_{t-1} + \cdots + \phi_p z_{t-p} + a_t$). Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ be the vector of AR parameters. Derive the conditional posterior distributions of $f(\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\phi}, \sigma^2)$, $f(\boldsymbol{\phi} \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)$, and $f(\sigma^2 \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\phi})$ assuming that conjugate prior distributions are used—that is,

$$\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_o, \boldsymbol{\Sigma}_o), \quad \boldsymbol{\phi} \sim N(\boldsymbol{\phi}_o, \boldsymbol{A}_o), \quad (v\lambda)/\sigma^2 \sim \chi^2_v.$$

**3.** Consider the linear AR($p$) model in Subsection 10.6.1. Suppose that $x_h$ and $x_{h+1}$ are two missing values with a joint prior distribution being multivariate normal with mean $\boldsymbol{\mu}_o$ and covariance matrix $\boldsymbol{\Sigma}_o$. Other prior distributions are the same as that in the text. What is the conditional posterior distribution of the two missing values?

4. Consider the monthly log returns of General Motors stock from 1950 to 1999 with 600 observations: (a) build a GARCH model for the series, (b) build a stochastic volatility model for the series, and (c) compare and discuss the two volatility models.

5. Build a stochastic volatility model for the daily log return of Cisco Systems stock from January 1991 to December 1999. You may download the data from CRSP database or the file "d-csco9199.dat." Use the model to obtain a predictive distribution for 1-step ahead volatility forecast at the forecast origin December 1999. Finally, use the predictive distribution to compute the Value at Risk of a long position worth $1 million with probability 0.01 for the next trading day.

6. Build a bivariate stochastic volatility model for the monthly log returns of General Motors stock and the S&P 500 index for the sample period from January 1950 to December 1999. Discuss the relationship between the two volatility processes and compute the time-varying beta for GM stock.

## REFERENCES

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*. Addison-Wesley: Reading, MA.

Chang, I., Tiao, G. C., and Chen, C. (1988), "Estimation of time series parameters in the presence of outliers," *Technometrics*, **30**, 193–204.

Carlin, B. P., and Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., Chapman and Hall: London.

DeGroot, M. H. (1970), *Optimal Statistical Decisions*, McGraw-Hill: New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm" (with discussion), *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Elerian, O., Chib, S., and Shephard, N. (2001), "Likelihood inference for discretely observed nonlinear diffusions," *Econometrica*, **69**, 959–993.

Eraker, B. (2001), "Markov Chain Monte Carlo analysis of diffusion models with application to finance," *Journal of Business & Economic Statistics* **19**, 177–191.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, CRC Press: London.

Geman, S., and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Hasting, W. K. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, **57**, 97–109.

Jacquier, E., Polson, N. G., and Rossi, P. E. (1994), "Bayesian analysis of stochastic volatility models" (with discussion), *Journal of Business & Economic Statistics*, **12**, 371–417.

Jones, R. H.(1980), "Maximum likelihood fitting of ARMA models to time series with missing observations," *Technometrics*, **22**, 389–395.

Justel, A., Peña, D., and Tsay, R. S. (2001), "Detection of outlier patches in autoregressive time series," *Statistica Sinica*, **11** (to appear).

Liu, J., Wong, W. H., and Kong, A. (1994), "Correlation structure and convergence rate of the Gibbs samplers I: Applications to the comparison of estimators and augmentation schemes," *Biometrika*, **81**, 27–40.

McCulloch, R. E., and Tsay, R. S. (1994), "Bayesian analysis of autoregressive time series via the Gibbs sampler," *Journal of Time Series Analysis*, **15**, 235–250.

McCulloch, R. E., and Tsay, R. S. (1994), "Statistical analysis of economic time series via Markov switching models," *Journal of Time Series Analysis*, **15**, 523–539.

Metropolis, N., and Ulam, S. (1949), "The Monte Carlo method," *Journal of the American Statistical Association*, **44**, 335–341.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, **21**, 1087–1092.

Tanner, M. A. (1996), *Tools for Statistical Inference: Methods for the exploration of posterior distributions and likelihood functions*, 3rd ed., Springer-Verlag: New York.

Tanner, M. A., and Wong, W. H. (1987), "The calculation of posterior distributions by data augmentation" (with discussion), *Journal of the American Statistical Association*, **82**, 528–550.

Tierney, L. (1994), "Markov chains for exploring posterior distributions" (with discussion), *Annals of Statistics*, **22**, 1701–1762.

Tsay, R. S. (1988), "Outliers, level shifts, and variance changes in time series," *Journal of Forecasting*, **7**, 1–20.

Tsay, R. S., Peña, D., and Pankratz, A. (2000), "Outliers in multivariate time series," *Biometrika*, **87**, 789–804.

Zhang, M. Y., Russell, J. R. and Tsay, R. S. (2000), "Determinants of bid and ask quotes and implications for the cost of trading," Working paper, Statistics Research Center, Graduate School of Business, University of Chicago.

*Analysis of Financial Time Series.* Ruey S. Tsay

ISBN: 0-471-41544-8

# Index