# CHAPTER 1

# INTRODUCTION TO BIG DEPENDENT DATA



Big data are common nowadays everywhere. Statistical methods capable of extracting useful information embedded in those data, including machine learning and artificial intelligence, have attracted much interest among researchers and practitioners in recent years. Most of the available statistical methods for analyzing big data were developed under the assumption that the observations are from independent samples. See, for instance, most methods discussed in Bühlmann and van de Geer (2011). Observations of big data, however, are dependent in many applications. The dependence may occur in the order by which the data were taken (such as time series data) or in space by which the sampling units reside (such as spatial data). Monthly civilian unemployment rates (16 years and older) of the 50 states in the United States is an example. Unemployment rates tend to be sticky over time and geographically neighboring states may share similar industries and, hence, have similar unemployment patterns. For dependent data, the spatial and/or temporal dependence is often the focus of statistical analysis. Consequently, there is a need to study analysis of big dependent data.

The main focus of this book is to provide readers a comprehensive treatment of statistical methods that can be used to analyze big dependent data. We start with some examples and simple methods for their descriptive analysis. More sophisticated methods will be introduced in other chapters. Keep in mind, however, that the analysis of big dependent data is gaining more attention as time advances. The methods discussed in this book reflect, to a large degree, our personal experience and preferences. We hope that the book can be helpful to many researchers and practitioners

*Statistical Learning for Big Dependent Data*, First Edition. Daniel Peña and Ruey S. Tsay.

in different scientific fields in their data analysis. Also, it could attract more attention to and motivate further research in the challenging problem of analyzing big dependent data.

## 1.1 EXAMPLES OF DEPENDENT DATA

In this section, we present some examples of big dependent data considered in the book, introduce some statistical methods for describing such data, and provide a framework for making statistical inference. Details of the methods introduced will be given in later chapters. Our goal is to demonstrate the data we intend to analyze and to highlight the consequences of ignoring their dependency. Whenever possible, we also illustrate the difficulties statistical methods developed for independent observations are facing when they are applied to big dependent data.

By dependent data, we mean observations of the variables of interest were taken over time or across space or both. Consequently, in these data the order in time or space matters. By big data, we mean the number of variables, $k$, is large or the number of data points, $T$, is large or both, and it could be that $k > T$. Figure 1.1 shows a data set of three time series representing the relative average temperature of November in Europe, North America, and South America from 1910 to 2014. The data are in the file `Temperatures.csv` of the book web site. The measurement is relative to the average temperature of November in the twentieth century. These data are dependent because observations of consecutive years tend to be closer to each other than observations that are many years apart. Furthermore, the relative temperature in South America shows a growing trend since 1980, whereas the trend can only be seen in Europe starting from 2000, yet it does not appear in the North America data.

Statistical analysis of these monthly temperature series is relatively easy, as the number of series is only 3 and the sample size is small. It belongs to the multivariate

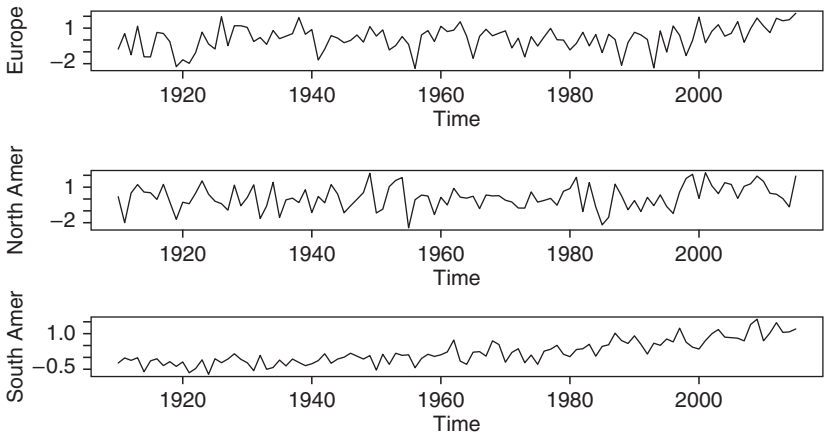


**Figure 1.1** Relative average temperatures of November in Europe, North America, and South America from 1910 to 2014. The measurement is relative to the average temperature of November in the twentieth century.

time series analysis in statistics. See, for instance, Tsay (2014). The main concern of this book is statistical analysis when the number of time series is large or the time intervals between observations are small, resulting in high-frequency data with large sample sizes. Modeling big temperature data is important in many applications. As a matter of fact, daily (even hourly) temperature is routinely taken at many locations around the world. These temperatures affect the demands of electricity and heating oil, are highly related to air pollution measurements, such as the particulate matter $PM_{2.5}$ and ozone concentration, and play an important role in commodity prices around the world.

Figure 1.2 shows the standardized daily stock indexes of the 99 most important financial markets around the world, from 3 January 2000 to 16 December 2015 for 4163 observations. The data are in the file `Stock_indexes_99world.csv`. The magnitudes of stock index vary markedly from one market to another so that we standardize the indexes in the plot. Specifically, each standardized index series has mean zero and variance one. In applications, we often analyze the returns of stock indexes. Figures 1.3 and 1.4 show the time plots of the first six indexes and their log returns, respectively. The log returns are obtained by taking the first difference of the logarithm of stock index.

These time plots demonstrate several features of multiple time series. First, they show the overall evolution and cross dependence of the financial markets. All markets exhibited a trough in 2003, had a steady increase reaching a peak around early 2008, then showed a dramatic drop caused by the 2008 financial crisis. Yet some markets experienced a gradual recovery after 2009. Second, the variabilities of the world financial markets appear to be higher when the markets were down; see Figure 1.4. This is not surprising as the fear factor, i.e. market volatility, is likely to be dominant during a bear market. On the other hand, the ranges of world market indexes at a given time seem to be smaller when the market were down; see Figure 1.2.

Figures 1.5 and 1.6 show the time plots of the daily financial indexes for 22 Asian and 47 European markets, respectively, for the same time period, but with $x$-axis now

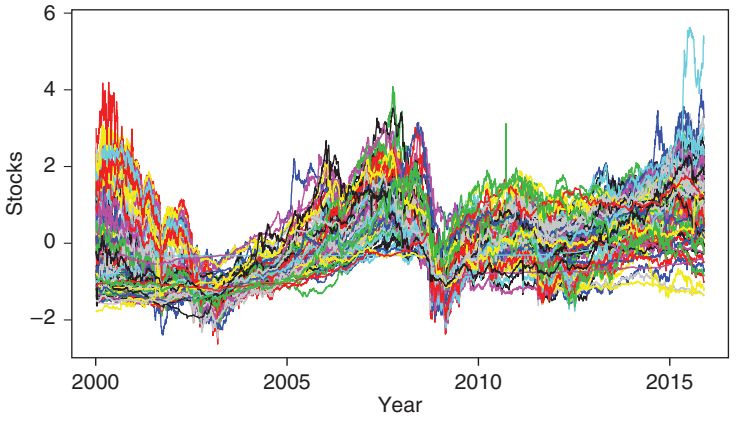


**Figure 1.2** Time plots of standardized daily stock market indexes of the 99 most important financial markets around the world from 3 January 2000 to 16 December 2015.

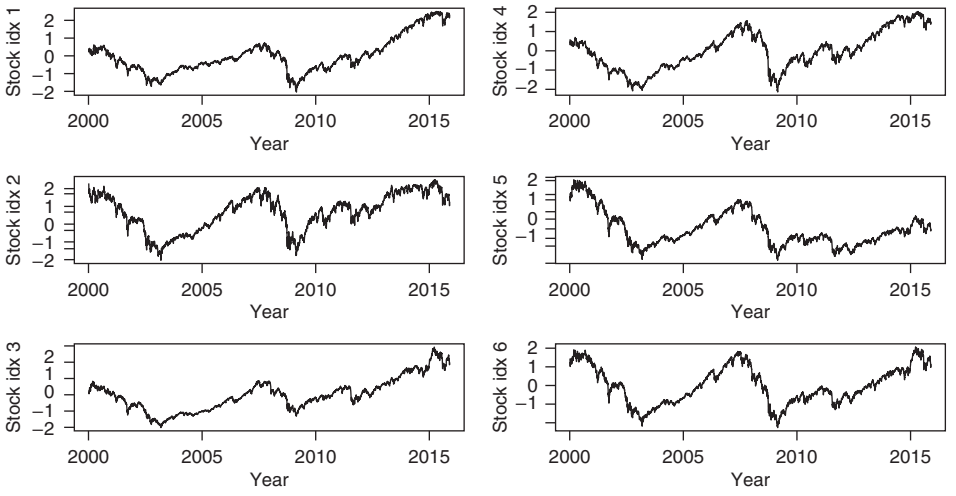


**Figure 1.3** Time plots of the first six standardized daily stock market indexes: 3 January 2000 to 16 December 2015.

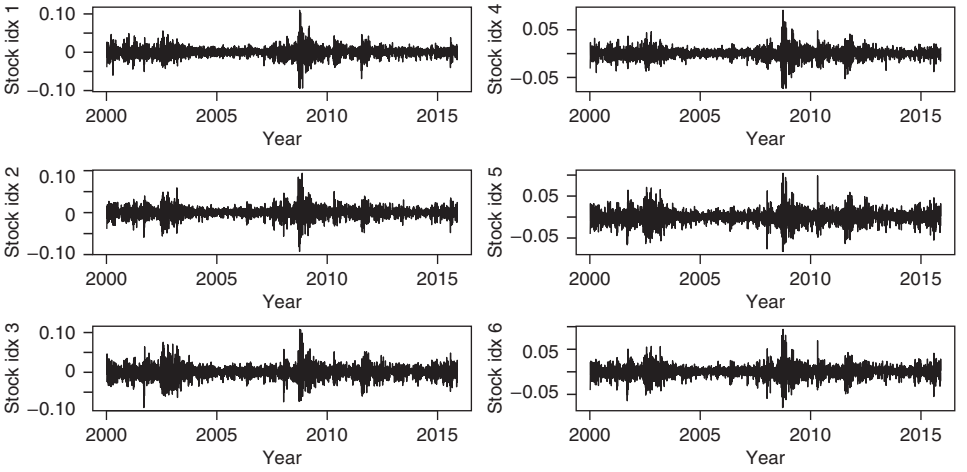


**Figure 1.4** Time plots of the log returns of the first six daily stock market indexes: 3 January 2000 to 16 December 2015.

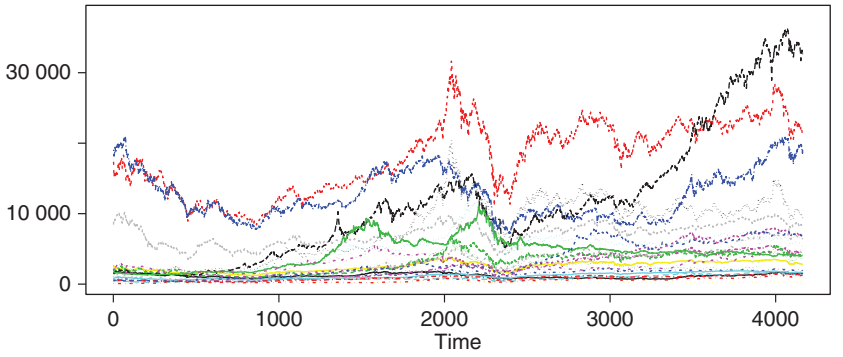


**Figure 1.5** Time plots of daily stock market indexes of 22 Asian financial markets from 3 January 2000 to 16 December 2015.

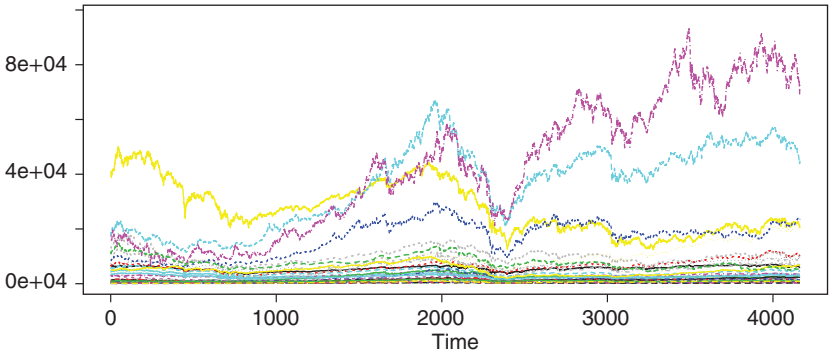


**Figure 1.6** Time plots of daily stock market indexes of 47 European financial markets from 3 January 2000 to 16 December 2015.

being in time index. The difference between Asian and European financial markets is easily seen, implying that care must be exercised when analyzing similar time series across different continents.

As a third example, Figure 1.7 shows 33 monthly consumer price indexes of European countries from January 2000 to October 2015. The data are in the file `CPIEurope2000-15.csv`. In the plot, the series have been standardized to have zero mean and unit variance. As expected, the plot shows an upward trend of the price indexes, but it also demonstrates sufficient differences between the series. For example, some series show a clear seasonal pattern, but others do not.

Large sets of data are available in marketing research. As an example, Figure 1.8 shows the time plots of daily sales, in natural logarithms, of a clothing brand in 25 provinces in China from 1 January 2008 to 9 December 2012 for 1805 observations. The data are from Chang et al. (2014) and are given in the file `clothing.csv`. The plots exhibit certain annual pattern, referring to as seasonal behavior in time

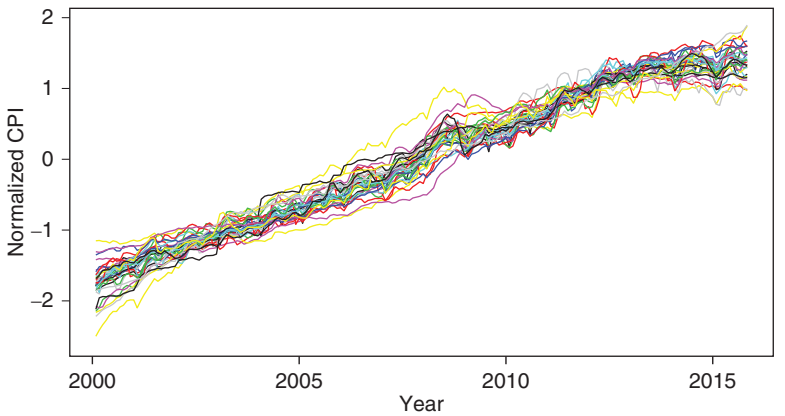


**Figure 1.7** Time plots of 33 monthly price indexes of European countries from January 2000 to October 2015.

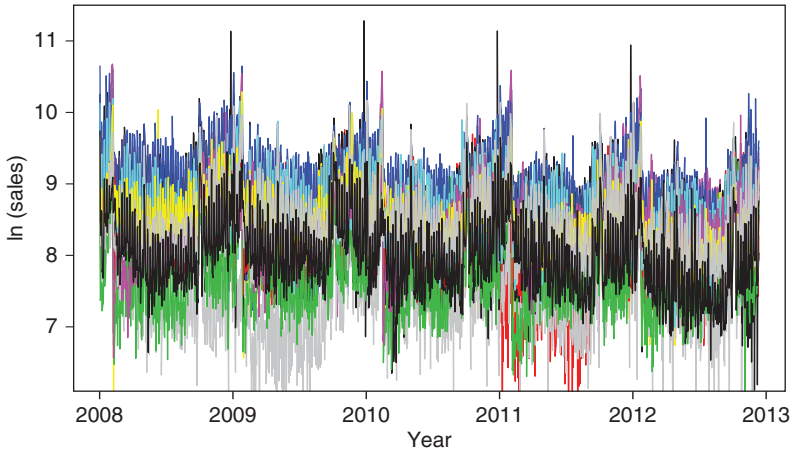


**Figure 1.8** Time plots of daily sales in natural logarithms of a clothing brand in 25 provinces in China from 1 January 2008 to 9 December 2012.

series analysis. To gain further insight, Figure 1.9 shows the same time plots for the first eight provinces. They are Beijing, Fujian, Guangdong, Guangxi, Hainan, Hebei, Henan, and Hubei. The levels of the plots are adjusted so that there is no overlapping in the figure. A special characteristic of the plots is that local peaks occur irregularly in the early part of each year, followed by certain drops in sales. This is caused by the Chinese New Year holidays that vary from year to year. In addition, the peaks do not occur for all provinces; see the second plot (from the top) of Figure 1.9. Analyzing these series jointly requires modeling the common features as well as the variations from one province to another.

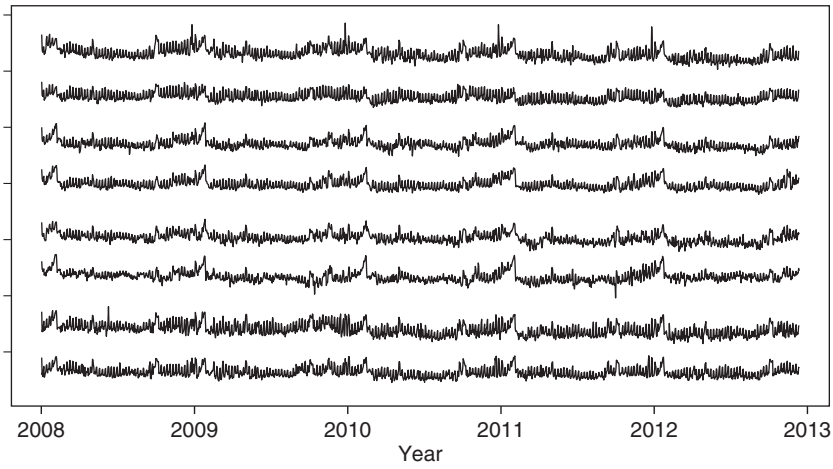


**Figure 1.9** Time plots of daily sales in natural logarithms of a clothing brand in eight provinces in China from 1 January 2008 to 9 December 2012.

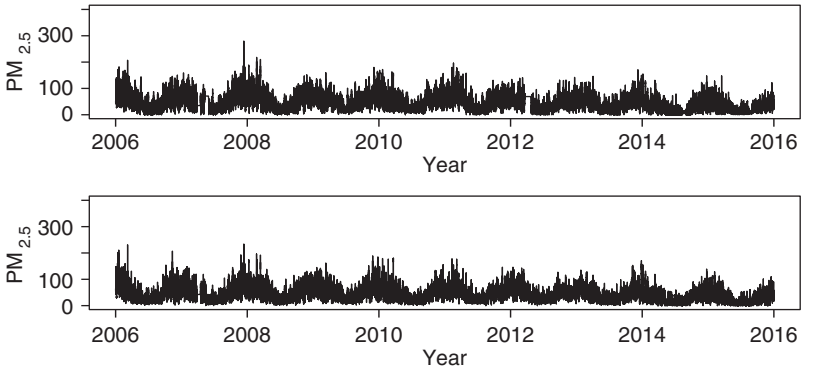


**Figure 1.10** Time plots of hourly observations of the particulate matter $PM_{2.5}$ from two monitoring stations in southern Taiwan from 1 January 2006 to 31 December 2015.

Figure 1.10 shows the time plots of hourly observations of the particulate matter $PM_{2.5}$ from two monitoring stations in southern Taiwan from 1 January 2006 to 31 December 2015. We drop the data of February 29 for simplicity, resulting in a sample size 87 600. The data are given in the file `TaiwanPM25.csv`. The two series are parts of many monitoring stations throughout the island. Also available at each station are measurements of temperature, dew points, and other air pollution indexes. Thus, this is a simple example of big dependent data. The time plots exhibit certain strong annual patterns and seem to indicate a minor decreasing trend. Figure 1.11 shows the first 30 000 sample autocorrelations of the $PM_{2.5}$ series of Station 1. The annual cycle with periodicity $s = 24 \times 365 = 8760$ is clearly seen from the autocorrelations. Figure 1.12, on the other hand, shows the first 120 sample autocorrelations of the same time series. The plot shows that there also exists a daily cyclic pattern in the

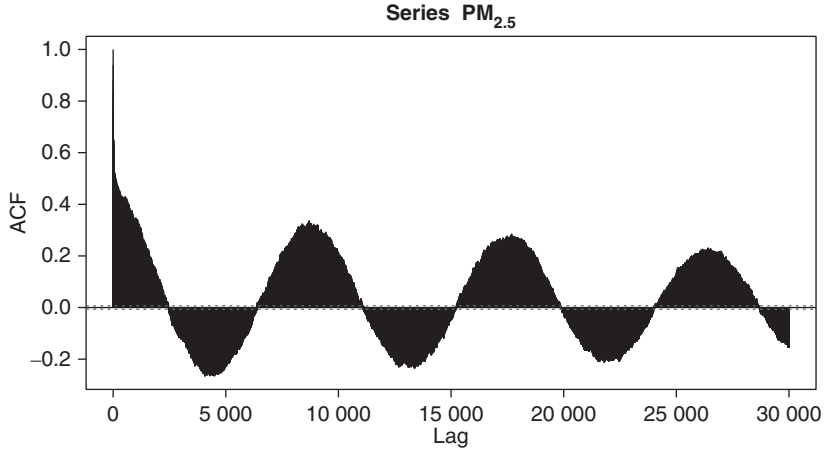


**Figure 1.11** Sample autocorrelations of $PM_{2.5}$ measurement of Station 1 in southern Taiwan. The first 30 000 autocorrelations are shown. The annual cycle with periodicity 8760 is clearly seen.

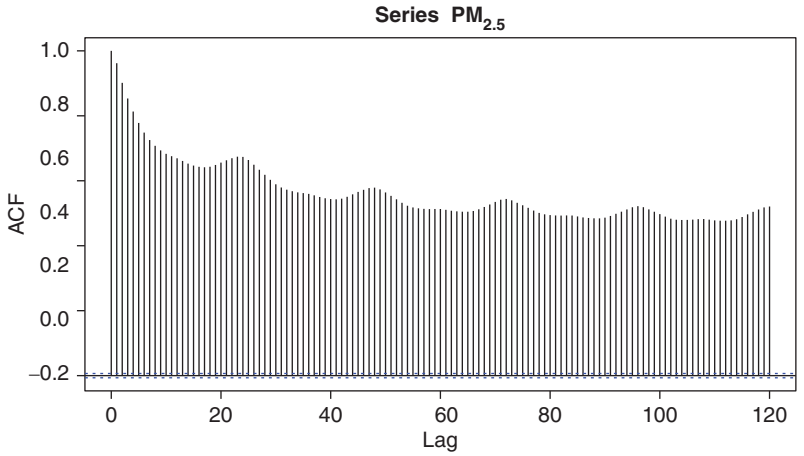


**Figure 1.12** Sample autocorrelations of $PM_{2.5}$ measurement of Station 1 in southern Taiwan. The first 120 autocorrelations are shown. The daily cyclical pattern with periodicity 24 is clearly seen.

$PM_{2.5}$ series with a periodicity $\lambda = 24$. Consequently, analysis of such a big dependent data should consider not only the overall trend, but also various cyclic patterns with different frequencies.

Finally, Figure 1.13 plots the monitoring stations (in circle) and the ozone levels (in color-coded squares) of US Midwestern states on 20 June 1987. The data set is available from the R package **fields**. The states are added to the plot so that readers can understand the geographical features of ozone levels in the Midwest. Clearly, on this particular day, high ozone readings appeared in the southwest of Illinois and south of Indiana. Also, ozone levels of neighboring monitoring stations are similar. This is an example of spatial process discussed in Chapter 9.

The six data sets presented are examples of multivariate spatio-temporal series, where we observe the values of several variables of interest over time and space.

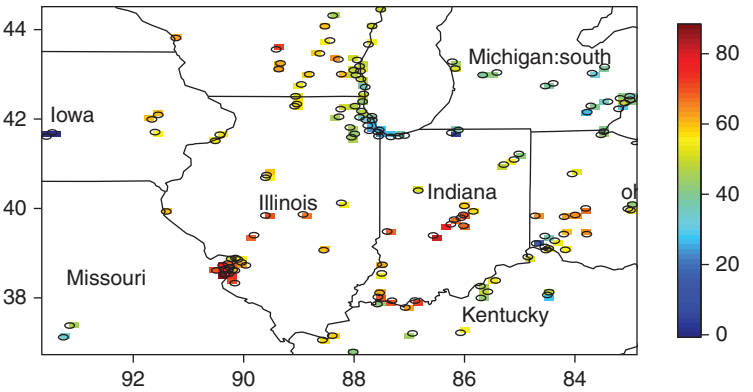


**Figure 1.13** Locations and ozone levels at various monitoring stations in US Midwestern states. The measurements are 8-hour averages (9 a.m. to 4 p.m.) on 20 June 1987.

The time sequence introduces dynamic dependency in a given series, which may be time varying. The data also show cross dependence between series and the strength of cross dependence may, in turn, depend on geographical locations or economic relations among countries. One of the main focuses of the book is to investigate both the dynamic and cross dependence between multiple time series.

The objectives of analyzing big dependent data include (a) to study the relationships between the series, (b) to explore the dynamic dependence within a series and across series, and (c) to predict the future values of the series. For spatial processes, one might be interested in predicting the variable of interest at a location where no observations are available. The first two objectives are referred to as data analysis and modeling, whereas the third objective involves statistical inference. Both modeling and inference require not only methods for handling data, but also the theory on which proper inference can be drawn. To this end, we need a framework for understanding the characteristics of the time series under study and for making sound predictions. We discuss this framework by starting with stochastic processes in the next section.

## 1.2 STOCHASTIC PROCESSES

Theory and properties of stochastic processes form the foundation for statistical analysis and inference of dependent data.

### 1.2.1 Scalar Processes

A scalar stochastic process is a sequence of random variables $\{z_t\}$, where the subscript $t$ takes values in a certain set $C$. In most cases, the elements of $C$ are ordered and often correspond to certain calendar time (days, months, years, etc.). The resulting stochastic process $z_t$ is called a time series. Unless specifically mentioned, we assume that the time index $t$ is equally spaced in the book. For a spatial process, $C$ consists of locations of certain geographical regions. The locations may denote the latitude and longitude of an observation station or a city. For each value $t$ in $C$ (e.g. each point in time), $z_t$ is a real-valued random variable defined on a common measurable space. The observed values, also denoted by $\{z_t\}$ for simplicity, form a realization of the stochastic process. We shall denote the realization by $\boldsymbol{z}_1^T = (z_1, \ldots, z_t, \ldots, z_T)'$, where $T$ is the sample size and $\boldsymbol{A}'$ denotes the transpose of the matrix or vector $\boldsymbol{A}$. The sample size $T$ denotes the number of observations for a time series or the number of locations for a pure spatial process.

Statistical properties of a stochastic process $z_t$ are characterized by its probability distribution. More specifically, for a given $T$ (finite and fixed), properties of $\boldsymbol{z}_1^T$ are determined by the joint probability distribution of its elements $\{z_t | t = 1, \ldots, T\}$. The marginal distribution of any non-empty subset of $\boldsymbol{z}_1^T$ can be obtained from the joint distribution by integrating out elements not in the subset. In particular, the probability distribution of the individual element $z_t$ is called the marginal distribution of $z_t$. Thus, similar to the general finite-dimensional random variable, we say that the probabilistic structure of $\boldsymbol{z}_1^T$ is known when we know its joint probability distribution.

In this book, we assume that a probability distribution has a well-defined density function. Let $f_t(z)$ be the marginal probability density function of $z_t$. Define the

*expected value* or *mean* of $z_t$ as

$$\mu_t = E(z_t) = \int_R z f_t(z) dz,$$

provided that the integral exists, where $R$ denotes the real line. The sequence $\boldsymbol{\mu}_1^T = (\mu_1, \mu_2, \ldots, \mu_T)$ is the *mean function* of $\boldsymbol{z}_1^T$. Define the *variance* of $z_t$ as

$$\sigma_t^2 = E(z_t - \mu_t)^2 = \int_R (z - \mu_t)^2 f_t(z) dz$$

provided that the integral exists. The sequence $\{\boldsymbol{\sigma}_t^2\}_1^T = (\sigma_1^2, \ldots, \sigma_T^2)$ is the *variance function* of $\boldsymbol{z}_1^T$. Higher order moments can be defined similarly, assuming that they exist.

A special case of interest is that all elements of $\boldsymbol{z}_1^T$ have the same mean. In this case, $\boldsymbol{\mu}_1^T$ is a constant function. If all elements of $\boldsymbol{z}_1^T$ have the same variance, then $\{\boldsymbol{\sigma}_t^2\}_1^T$ is a constant function. If both $\boldsymbol{\mu}_1^T$ and $\{\boldsymbol{\sigma}_t^2\}_1^T$ are constants for all finite $T$, then the stochastic process $z_t$ is stable and we say that it is a homogeneous process. In some applications, $\boldsymbol{\mu}_1^T$ or $\{\boldsymbol{\sigma}_t^2\}_1^T$ may not exist.

In what follows, we assume that $\sigma_t^2$ exists for all $t$. The linear dynamic dependence of a stochastic process $z_t$ is determined by its *autocovariance* or *autocorrelation* function (ACF). For $z_t$ and $z_{t-j}$, where $j$ is a given integer, we define their autocovariance as

$$\gamma(t, t-j) = \text{Cov}(z_t, z_{t-j}) = E[(z_t - \mu_t)(z_{t-j} - \mu_{t-j})]. \tag{1.1}$$

When $j = 0$, $\gamma(t, t) = \text{Var}(z_t) = \sigma_t^2$. Also, from the definition, it is clear that $\gamma(t, t-j) = \gamma(t-j, t)$. The *autocorrelation* between $z_t$ and $z_{t-j}$ is defined as

$$\rho(t, t-j) = \frac{\gamma(t, t-j)}{\sigma_t \sigma_{t-j}}. \tag{1.2}$$

Of particular importance in real applications is the case in which the autocovariance and, hence, the autocorrrelation between $z_t$ and $z_{t-j}$ is a function of the lag $j$, and not a function of $t$. In this case, the autocovariance and autocorrelation are time-invariant and we write $\gamma(t, t-j) = \gamma_j$ and $\rho(t, t-j) = \rho_j$.

***1.2.1.1 Stationarity*** A stochastic process $z_t$ is *strictly stationary* if the joint distributions of the $m$-dimensional random vectors $(z_{t_1}, z_{t_2}, \ldots, z_{t_m})$ and $(z_{t_1+h}, z_{t_2+h}, \ldots, z_{t_m+h})$ are the same, where $m$ is an arbitrary positive integer, $\{t_1, \ldots, t_m\}$ are arbitrary $m$ time indexes, and $h$ is an arbitrary integer. In other words, the joint distribution of arbitrary $m$ random variables is time-invariant. In particular, the marginal probability density $f_t(z)$ of a strictly stationary process $z_t$ does not depend on $t$, and we simply denote it by $f(z)$. A stochastic process $z_t$ is *weakly stationary* if its mean and autocovariance functions exist and are time-invariant. Specifically, a process $z_t$ is stationary in the weak sense if (1) $\mu_t = \mu$ for all $t$; (2) $\sigma_t^2 = \sigma^2$ for all $t$; and (3) $\gamma(t, t-j) = \gamma_j$ for all $t$, where $j$ is a given integer. The first two conditions say that the mean and variance are time-invariant. The third condition states that the autocovariance and, hence, the autocorrelation between two variables depend only on their time separation $j$. Unless specifically stated, we

use stationarity to denote weak stationarity of a stochastic process. For this process the autocovariance and autocorrelation functions depend only on the difference of lags and, in particular, we have $\gamma_0 = \sigma^2$, $\gamma_j = \gamma_{-j}$ and $\rho_j = \rho_{-j}$, where $j$ is an arbitrary integer.

Assume that $z_t$ is stationary. For a given positive integer $m$, the covariance matrix of $(z_t, z_{t-1}, \ldots, z_{t-m+1})'$ is defined as

$$\begin{aligned} \boldsymbol{M}_z(m) &= E\left\{ \begin{bmatrix} z_t - \mu \\ z_{t-1} - \mu \\ \vdots \\ z_{t-m+1} - \mu \end{bmatrix} [(z_t - \mu), (z_{t-1} - \mu), \ldots, (z_{t-m+1} - \mu)] \right\} \\ &= \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0 \end{bmatrix}. \end{aligned} \tag{1.3}$$

This covariance matrix has the special pattern that (a) all diagonal elements are the same, (b) all elements above and below the main diagonal are the same, and so on. Such a matrix is called a *Toeplitz* matrix. The corresponding *autocorrelation matrix* is

$$\boldsymbol{R}_z(m) = \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{m-1} \\ \rho_1 & 1 & \cdots & \rho_{m-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{m-1} & \rho_{m-2} & \cdots & 1 \end{bmatrix}, \tag{1.4}$$

which is also a Toeplitz matrix.

An important property of a stationary process is that any non-trivial finite linear combination of its elements remains stationary. In other words, a process obtained by any non-zero linear combination (with finite elements) of a stationary process is also stationary. For example, if $z_t$ is stationary, the process $w_t$ defined by $w_t = z_t - z_{t-1}$ is also stationary. A consequence of this property is that the autocovariances of $z_t$ must satisfy certain conditions that are useful for studying stochastic processes. In particular, we can consider a non-trivial linear combination of a stationary process $z_t$ and its lagged values, namely

$$y_t = c_1 z_t + c_2 z_{t-1} + \cdots + c_m z_{t-m+1} = \boldsymbol{c}' \boldsymbol{z}_{t,m}$$

where $\boldsymbol{c} = (c_1, \ldots, c_m)' \neq \boldsymbol{0}$ and $\boldsymbol{z}_{t,m} = (z_t, z_{t-1}, \ldots, z_{t-m+1})'$. Since $y_t$ is stationary, its variance exists and is given by $\boldsymbol{c}' \boldsymbol{M}_z(m) \boldsymbol{c} \geq 0$, where $\boldsymbol{M}_z(m)$ is defined in Eq. (1.3). Consequently, $\boldsymbol{M}_z(m)$ must be a nonnegative definite matrix. Similarly, the correlation matrix $\boldsymbol{R}_z(m)$ must also be nonnegative definite.

Strict stationarity implies weak stationarity provided that the first two moments of the process exist. But weak stationarity does not guarantee strict stationarity of a stochastic process. Consider the following simple example. Let $\{z_t\}$ be a sequence of independent and identically distributed standard normal random variables,

i.e. $z_t \sim_{iid} N(0,1)$. Define a stochastic process $x_t$ by

$$x_t = \begin{cases} 0.5(z_t^2 - 1), & \text{if } t \text{ is odd} \\ z_t, & \text{if } t \text{ is even.} \end{cases}$$

It is easy to show that (a) $E(x_t) = 0$, (b) $\text{Var}(x_t) = 1$, and (c) $E(x_t x_{t-j}) = 0$ for $j \neq 0$. Therefore, $\{x_t\}$ is a weakly stationary process. On the other hand, $x_t$ is Gaussian if $t$ is even and it is a shifted $\chi^2$ with one degree of freedom when $t$ is odd. The process $\{x_t\}$ is not identically distributed and, hence, is not strictly stationary. If we assume that $\boldsymbol{z}_1^T$ follows a multivariate normal distribution for all $T$, then weak stationarity is equivalent to strict stationarity, because a normal (or Gaussian) distribution is determined by its first two moments.

#### 1.2.1.2 White Noise Process

An important stationary process is the *white noise process*. A stochastic process $z_t$ is a white noise process if it satisfies the following conditions: (1) $E(z_t) = 0$; (2) $\text{Var}(z_t) = \sigma^2 < \infty$; and (3) $\text{Cov}(z_t, z_{t-j}) = 0$ for all $j \neq 0$. In other words, $z_t$ is a white noise series if and only if it has a zero mean and finite variance, and is not serially correlated. Thus, a white noise is not necessarily strictly stationary. If we impose the additional condition that $z_t$ and $z_{t-j}$ are independent and have the same distribution, where $j \neq 0$, then $z_t$ is a *strict white noise* process.

#### 1.2.1.3 Conditional Distribution

In addition to marginal distributions, conditional distributions also play an important role in studying stochastic processes. Let $\boldsymbol{z}_\ell^k = (z_\ell, z_{\ell+1}, \ldots, z_k)'$, where $\ell$ and $k$ are positive integers and $\ell \leq k$. In prediction, we are interested in the conditional distribution of $\boldsymbol{z}_{T+1}^{T+h}$ given $\boldsymbol{z}_1^T$, where $h \geq 1$ is referred to as the *forecast horizon*. In estimation, we express the joint probability density function of $\boldsymbol{z}_1^T$ as a product of the conditional density functions of $z_{t+1}$ given $\boldsymbol{z}_1^t$ for $t = T-1, T-2, \ldots, 1$. More details are given in later chapters.

If the stochastic process $z_t$ has the following property

$$f(z_{t+1}|\boldsymbol{z}_1^t) = f(z_{t+1}|z_t), \qquad t = 1, 2, \ldots,$$

then $z_t$ is a first-order *Markov* process. The aforementioned property is referred to as the memoryless of a Markov process as the conditional distribution only depends on the variable one period before.

### 1.2.2 Vector Processes

A stochastic vector process is a sequence of random vectors $\{\boldsymbol{z}_t\}$, where the index $t$ assumes values in a certain set $C$ and $\boldsymbol{z}_t = (z_{1t}, \ldots, z_{kt})'$ is a $k$-dimensional vector. Each component $z_{it}$ follows a scalar stochastic process, and the vector process is characterized by the joint probability distribution of the random variables $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$. Similar to the scalar case, we define the mean vector of $\boldsymbol{z}_t$ as

$$E(\boldsymbol{z}_t) = [E(z_{1t}), E(z_{2t}), \ldots, E(z_{kt})]' = (\mu_{1t}, \ldots, \mu_{kt})' \equiv \boldsymbol{\mu}_t,$$

provided that the expectations exist, and the lag-$\ell$ autocovariance function as

$$\boldsymbol{\Gamma}(t, t-\ell) = \text{Cov}(\boldsymbol{z}_t, \boldsymbol{z}_{t-\ell}) = E[(\boldsymbol{z}_t - \boldsymbol{\mu}_t)(\boldsymbol{z}_{t-\ell} - \boldsymbol{\mu}_{t-\ell})'] \tag{1.5}$$

provided that the covariances involved all exist, where $\ell$ is an integer. We say that the vector process $\{\boldsymbol{z}_t\}$ is weakly stationary if (a) $\boldsymbol{\mu}_t = \boldsymbol{\mu}$, a constant vector, and (b) $\boldsymbol{\Gamma}(t, t-\ell)$ depends only on $\ell$. In other words, a $k$-dimensional stochastic process $\boldsymbol{z}_t$ is weakly stationary if its first two moments are time-invariant. In this case, we write $\boldsymbol{\Gamma}(t, t-\ell) = E[(\boldsymbol{z}_t - \boldsymbol{\mu})(\boldsymbol{z}_{t-\ell} - \boldsymbol{\mu})'] = \boldsymbol{\Gamma}(\ell) = [\gamma_{ij}(\ell)]$, where $1 \le i, j \le k$. In particular, the lag-0 autocovariance matrix of a weakly stationary $k$-dimensional stochastic process $\boldsymbol{z}_t$ is given by

$$\boldsymbol{\Gamma}(0) = \begin{bmatrix} \gamma_{11}(0) & \gamma_{12}(0) & \cdots & \gamma_{1k}(0) \\ \gamma_{21}(0) & \gamma_{22}(0) & \cdots & \gamma_{2k}(0) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{k1}(0) & \gamma_{k2}(0) & \cdots & \gamma_{kk}(0) \end{bmatrix} \tag{1.6}$$

which is symmetric, because $\gamma_{ij}(0) = \gamma_{ji}(0)$. Note that the diagonal element $\gamma_{ii}(0) = \sigma_i^2 = \text{Var}(z_{it})$.

For a stationary vector process $\boldsymbol{z}_t$, its lag-$\ell$ autocovariance matrix $\boldsymbol{\Gamma}(\ell) = [\gamma_{ij}(\ell)]$ measures the linear dependence among the component series $z_{it}$ and their lagged variables $z_{j,t-\ell}$. It pays to study the exact meaning of individual element $\gamma_{ij}(\ell)$ of $\boldsymbol{\Gamma}(\ell)$. First, the diagonal element $\gamma_{ii}(\ell)$ is the lag-$\ell$ autocovariance of the scalar process $z_{it}$ ($i = 1, \ldots, k$). Second, the off-diagonal element $\gamma_{ij}(\ell)$ is

$$\gamma_{ij}(\ell) = E[(z_{it} - \mu_i)(z_{j,t-\ell} - \mu_j)],$$

which measures the linear dependence of $z_{it}$ on the lagged value $z_{j,t-\ell}$ for $\ell > 0$. Thus, $\gamma_{ij}(\ell)$ quantifies the linear dependence of $z_{it}$ on the $\ell$th past value of $z_{jt}$. Third, in general, $\boldsymbol{\Gamma}(\ell)$ is not symmetric if $\ell > 0$. As a matter of fact, $\gamma_{ij}(\ell) \neq \gamma_{ji}(\ell)$ for most stochastic process $\boldsymbol{z}_t$ because, as stated before, $\gamma_{ij}(\ell)$ denotes the linear dependence of $z_{it}$ on $z_{j,t-\ell}$, whereas $\gamma_{ji}(\ell)$ quantifies the linear dependence of $z_{jt}$ on $z_{i,t-\ell}$. Finally, we have

$$\begin{aligned} \gamma_{ij}(-\ell) &= E[(z_{it} - \mu_i)(z_{j,t+\ell} - \mu_j)] \\ &= E[(z_{j,v} - \mu_j)(z_{i,v-\ell} - \mu_i)] \quad (v = \mathrm{t} + \ell) \\ &= \gamma_{ji}(\ell). \quad \text{(by stationarity)} \end{aligned}$$

Therefore, we have $\gamma_{ij}(-\ell) = \gamma_{ji}(\ell)$ for all $\ell$. Consequently, the autocovariance matrices of a stationary vector process $\boldsymbol{z}_t$ satisfy

$$\boldsymbol{\Gamma}(-\ell) = [\boldsymbol{\Gamma}(\ell)]',$$

and, hence, it suffices to consider $\boldsymbol{\Gamma}(\ell)$ for $\ell \ge 0$ in practice. If the dimension of $\boldsymbol{\Gamma}(\ell)$ is needed to avoid any confusion, we write $\boldsymbol{\Gamma}(\ell) \equiv \boldsymbol{\Gamma}_k(\ell)$ with the subscript $k$ denoting the dimension of the underlying stochastic process $\boldsymbol{z}_t$.

A global measure of the overall variability of the process $\boldsymbol{z}_t$ is given by its generalized variance, which is the determinant of the lag-0 covariance matrix $\boldsymbol{\Gamma}(0)$, i.e. $|\boldsymbol{\Gamma}(0)|$. The standardized, by the dimension, measure is the effective variance, defined by

$$\text{EV}(\boldsymbol{z}_t) = |\boldsymbol{\Gamma}(0)|^{1/k}. \tag{1.7}$$

For instance, if $\boldsymbol{z}_t$ is two dimensional, then $\text{EV}(\boldsymbol{z}_t) = [\gamma_{11}(0)\gamma_{22}(0)\{1 - \rho_{12}^2(0)\}]^{1/2}$, where $\rho_{12}(0)$ is the contemporaneous correlation coefficient between $z_{1t}$ and $z_{2t}$.

Clearly, $\text{EV}(\boldsymbol{z}_t)$ assumes its largest value when the two series are uncorrelated, i.e. $\rho_{12}(0) = 0$. For a $k$-dimensional process $\boldsymbol{z}_t$, as the determinant of a square-matrix is the product of its eigenvalues, the effective variance is the geometrical mean of the eigenvalues of $\boldsymbol{\Gamma}(0)$.

We can summarize the properties of a stationary vector process $\boldsymbol{z}_t$ by using linear combinations of its components. Consider, for example, the linear process

$$y_t = \boldsymbol{c}'\boldsymbol{z}_t = \sum_{i=1}^{k} c_i z_{it},$$

where $\boldsymbol{c} = (c_1, \ldots, c_k)' \neq \boldsymbol{0}$. This process is stationary because it is a finite linear combination of the stationary components $z_{it}$ $(i = 1, \ldots, k)$. More specifically, the expectation is $E(y_t) = c_1 E(z_{1t}) + \cdots + c_k E(z_{kt}) = \boldsymbol{c}'E(\boldsymbol{z}_t) = \boldsymbol{c}'\boldsymbol{\mu}$, which is time-invariant. Similarly, the variance of $y_t$ is given by $\text{Var}(y_t) = \text{Var}(\boldsymbol{c}'\boldsymbol{z}_t) = \boldsymbol{c}'\text{Var}(\boldsymbol{z}_t)\boldsymbol{c} = \boldsymbol{c}'\boldsymbol{\Gamma}(0)\boldsymbol{c}$, which is also time-invariant. In general, we have $\text{Cov}(y_t, y_{t-\ell}) = \boldsymbol{c}'\boldsymbol{\Gamma}(\ell)\boldsymbol{c}$, which depends only on $\ell$. Thus, $y_t$ is stationary. In addition, since $\text{Var}(y_t)$ is non-negative and $\boldsymbol{c}$ is an arbitrary non-zero vector, we conclude that $\boldsymbol{\Gamma}(0)$ is nonnegative definite.

Similar to the scalar case, the autocovariance matrices in Eq. (1.5) of a stationary vector process $\boldsymbol{z}_t$ must satisfy certain conditions. To see this, we consider a non-zero linear combination of $\boldsymbol{z}_t$ and its $m$ lagged values, say,

$$y_t = \sum_{i=1}^{m} \boldsymbol{b}_i'\boldsymbol{z}_{t-i+1} = \boldsymbol{b}_1'\boldsymbol{z}_t + \boldsymbol{b}_2'\boldsymbol{z}_{t-1} + \cdots + \boldsymbol{b}_m'\boldsymbol{z}_{t-m+1}$$

where $\boldsymbol{b}_j = (b_{j1}, \ldots, b_{jk})'$ is a $k$-dimensional real-valued vector for $j = 1, \ldots, m$. Defining $\boldsymbol{b} = (\boldsymbol{b}_1', \boldsymbol{b}_2', \ldots, \boldsymbol{b}_m')'$, which is a non-zero $km$-dimensional constant vector, and $\boldsymbol{z}_{t,m} = (\boldsymbol{z}_t', \boldsymbol{z}_{t-1}', \ldots, \boldsymbol{z}_{t-m+1}')'$, we have $\text{Var}(y_t) = \boldsymbol{b}'\text{Var}(\boldsymbol{z}_{t,m})\boldsymbol{b} = \boldsymbol{b}'\boldsymbol{M}_{\mathbf{z}}(m)\boldsymbol{b}$, where

$$\boldsymbol{M}_{\mathbf{z}}(m) = \begin{bmatrix} \boldsymbol{\Gamma}(0) & \boldsymbol{\Gamma}(1) & \cdots & \boldsymbol{\Gamma}(m-1) \\ \boldsymbol{\Gamma}(-1) & \boldsymbol{\Gamma}(0) & \cdots & \boldsymbol{\Gamma}(m-2) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}(-m+1) & \boldsymbol{\Gamma}(-m+2) & \cdots & \boldsymbol{\Gamma}(0) \end{bmatrix} \tag{1.8}$$

is a nonnegative block Toeplitz matrix. This $\boldsymbol{M}_{\mathbf{z}}(m)$ matrix is the generalization of $\boldsymbol{M}_z(m)$ in Eq. (1.3) for the scalar series $z_t$.

The cross-correlation matrices (CCMs) of $\boldsymbol{z}_t$ are defined as the autocovariances of the standardized process $\boldsymbol{x}_t = \boldsymbol{D}^{-1/2}(\boldsymbol{z}_t - \boldsymbol{\mu})$, where the matrix $\boldsymbol{D}$ is defined as $\text{diag}(\gamma_{11}(0), \ldots, \gamma_{kk}(0))$, consisting of the variances of the components of $\boldsymbol{z}_t$. Specifically, the lag-$\ell$ CCM of $\boldsymbol{z}_t$ is

$$\begin{aligned} \mathbf{R}_{\mathbf{z}}(\ell) &= \text{Cov}(\boldsymbol{x}_t, \boldsymbol{x}_{t-\ell}) = \boldsymbol{D}^{-1/2}\text{Cov}(\boldsymbol{z}_t, \boldsymbol{z}_{t-\ell})\boldsymbol{D}^{-1/2} \\ &= \begin{bmatrix} \rho_{11}(\ell) & \rho_{12}(\ell) & \cdots & \rho_{1k}(\ell) \\ \rho_{21}(\ell) & \rho_{22}(\ell) & \cdots & \rho_{2k}(\ell) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k1}(\ell) & \rho_{k2}(\ell) & \cdots & \rho_{kk}(\ell) \end{bmatrix}. \end{aligned} \tag{1.9}$$

From the definition, we have $\rho_{ii}(0) = 1$ and

$$\rho_{ij}(\ell) = \frac{\gamma_{ij}(\ell)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}}.$$

Similarly to the autocovariance matrices, we have $\boldsymbol{M}_{\mathbf{z}}(\ell) = [\boldsymbol{M}_{\mathbf{z}}(-\ell)]'$.

The *effective correlation* of $\boldsymbol{z}_t$ is given by

$$\mathrm{ER}_{\mathbf{z}} = 1 - |\mathbf{R}_{\mathbf{z}}(0)|^{1/k} \tag{1.10}$$

and it assumes a value between zero and one. For instance, for $k = 2$, $\mathrm{ER}_{\mathbf{z}} = 1 - [1 - \rho_{12}(0)]^{1/2}$. See Peña and Rodriguez (2003) for more properties.

***1.2.2.1 Vector White Noises*** A $k$-dimensional vector process $\boldsymbol{z}_t$ is a white noise process if (a) $E(\boldsymbol{z}_t) = \mathbf{0}$, (b) $\mathrm{Var}(\boldsymbol{z}_t) = \boldsymbol{\Gamma}(0)$ is positive-definite, and (c) $\boldsymbol{\Gamma}(\ell) = \mathbf{0}$ for all $\ell \neq 0$. This is a direct generalization of the scalar white noise. Thus, a vector white noise process consists of random vectors that have zero means, positive-definite covariance, and no lagged serial or cross correlations.

***1.2.2.2 Invertibility*** In many applications, we use linear regression type of models to describe the dynamic dependence of a vector process. The model can be written as

$$\boldsymbol{z}_t = \boldsymbol{\pi}_0 + \sum_{i=1}^{\infty} \boldsymbol{\pi}_i \boldsymbol{z}_{t-i} + \boldsymbol{a}_t \tag{1.11}$$

where $\{\boldsymbol{a}_t\}$ is a stationary vector white noise process, $\boldsymbol{\pi}_0$ is a $k$-dimensional vector of constants, and $\boldsymbol{\pi}_i$ are $k \times k$ real-valued matrices. For Eq. (1.11) to be meaningful, the coefficient matrices must satisfy the condition

$$\sum_{i=1}^{\infty} \|\boldsymbol{\pi}_i\| < \infty \tag{1.12}$$

where $\|\boldsymbol{A}\|$ denotes a norm of the matrix $\boldsymbol{A}$, e.g. $\|\boldsymbol{A}\| = \left(\sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}^2\right)^{1/2}$, which is the *Frobenius* norm. A vector process $\boldsymbol{z}_t$ is *invertible* if it can assume the model in (1.11) that satisfies the condition of Eq. (1.12). If the summation in Eq. (1.11) is truncated at a finite integer $p$, then $\boldsymbol{z}_t$ is invertible and the process follows a vector autoregressive (VAR) model.

Invertibility is another important concept in time series analysis. A necessary condition for invertible model is that $\|\boldsymbol{\pi}_i\| \to 0$ as $i \to \infty$. This means that the added contribution of $\boldsymbol{z}_{t-\ell}$ to $\boldsymbol{z}_t$ conditioned on the available information $\{\boldsymbol{z}_{t-1}, \boldsymbol{z}_{t-2}, \ldots, \boldsymbol{z}_{t-\ell+1}\}$ is diminishing as $\ell$ increases. Consider the scalar process $z_t$, any AR($p$) process with finite $p$ is invertible. On the other hand, the process $z_t = a_t - a_{t-1}$, where $\{a_t\}$ is a white noise, is stationary, but non-invertible.

## 1.3 SAMPLE MOMENTS OF STATIONARY VECTOR PROCESS

In real applications, we often observe a sequence of realizations, say $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T\}$, of a stationary vector process $\boldsymbol{z}_t$. Our goal is to investigate properties of $\boldsymbol{z}_t$ based on the observed data. To this end, certain exploratory data analysis is useful. In particular, much can be learned about $\boldsymbol{z}_t$ by studying its sample moments.

### 1.3.1 Sample Mean

An unbiased estimator of the population mean of the scalar process $z_{it}$ is the sample mean

$$\bar{z}_i = \frac{1}{T}\sum_{t=1}^{T} z_{it}. \tag{1.13}$$

Putting them together, the sample mean of the vector process $\boldsymbol{z}_t$ is

$$\bar{\boldsymbol{z}} = (\bar{z}_1, \ldots, \bar{z}_k)' = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{z}_t. \tag{1.14}$$

For an independent and identically distributed random sample, it is easy to see that $E(\bar{z}_i) = E(z_{it}) = \mu_i$, and the variance of $\bar{z}_i$ is $\sigma_i^2/T$, where $\sigma_i^2 = \text{Var}(z_{it})$. Therefore, the sample mean in this case is a consistent estimator of the population mean $\mu_i$ as $T \to \infty$.

Note that for a stationary process, the consistency of the sample mean does not necessarily hold. For example, let $z_{11}$ be a random sample from a normal distribution with mean zero and variance $\sigma_1^2$. Let $z_{1t} = z_{11}$ for $t = 2, \ldots, T$. This special process $\{z_{1t}\}$ is stationary because (1) $E(z_{1t}) = 0$, which is time-invariant, (2) $\text{Var}(z_{1t}) = \sigma_1^2$, which is also time-invariant, and $\text{Cov}(z_{1t}, z_{1,t-j}) = \sigma_1^2$, which is time-invariant for all $j$. However, in this particular case we have $\bar{z}_1 = z_{11}$ and $\text{Var}(\bar{z}_1) = \sigma_1^2$, no matter what the sample size $T$ is. Therefore, the sample mean is not a consistent estimator for this special process as $T \to \infty$.

The condition under which the sample mean $\bar{z}_i$ is a consistent estimator of the population mean $\mu_i$ is referred to as the *ergodic* condition of a stochastic process. The ergodic theory is concerned with conditions under which the time average of a long realization (sample mean) converges to the space average (average of many random samples at a given time index) of a dynamic system. For a scalar weakly stationary $z_t$, the ergodic condition for $\bar{z}$ to converge to $\mu = E(z_t)$ is that $\sum_{i=0}^{\infty} |\gamma_i| < \infty$. We assume this condition in the book from now on.

For a weakly stationary vector process $\mathrm{z}_t$, the covariance matrix of the sample mean $\bar{\boldsymbol{z}}$ is given by

$$\begin{aligned}\text{Var}(\bar{\boldsymbol{z}}) &= E\left[(\bar{\boldsymbol{z}} - \boldsymbol{\mu})(\bar{\boldsymbol{z}} - \boldsymbol{\mu})'\right] = E\left[\left\{\frac{1}{T}\sum_{t=1}^{T}(\mathbf{z}_t - \boldsymbol{\mu})\right\}\left\{\frac{1}{T}\sum_{j=1}^{T}(\mathbf{z}_j - \boldsymbol{\mu})'\right\}\right] \\ &= \frac{1}{T^2}\sum_{t=1}^{T} E\left[(\mathbf{z}_t - \boldsymbol{\mu})\left\{\sum_{j=1}^{T}(\mathbf{z}_j - \boldsymbol{\mu})'\right\}\right] = \frac{1}{T^2}\left[\sum_{j=-(T-1)}^{T-1}(T - |j|)\boldsymbol{\Gamma}(j)\right]. \end{aligned} \tag{1.15}$$

This implies that the asymptotic variance of the scalar sample mean $\bar{z}_i$ is

$$T \times \text{Var}(\bar{z}_i) \to \gamma_{ii}(0) + 2\sum_{\ell=1}^{\infty}\gamma_{ii}(\ell), \quad \text{as} \quad T \to \infty. \tag{1.16}$$

where, again, $\gamma_{ii}(\ell)$ is the lag-$\ell$ autocovariance of $z_{it}$. From Equation (1.2), if $\{z_{it}\}$ is white noise the variance of the sample mean $\bar{z}_i$ is $\sigma_i^2/T$. In general, we have

$$\text{Var}(\bar{z}_i) \approx \frac{1}{T}\left[\gamma_{ii}(0) + 2\sum_{l=1}^{T}\gamma_{ii}(l)\right], \quad \text{as} \quad T \to \infty.$$

Therefore, $\text{Var}(\bar{z}_i)$ could be large if the sum of the autocovariances of $z_{it}$ is large. A necessary (although not sufficient) condition for the sum of autocovariances of $z_{it}$ to converge is

$$\lim_{l\to\infty}\gamma_{ii}(l) \to 0,$$

which implies that the dependence of $z_{it}$ on its past values $z_{i,t-l}$ vanishes as $l$ increases.

### 1.3.2 Sample Covariance and Correlation Matrices

For a vector process $\boldsymbol{z}_t$, the lag-$\ell$ sample autocovariance matrix is

$$\hat{\boldsymbol{\Gamma}}(\ell) = \frac{1}{T}\sum_{t=\ell+1}^{T}(\boldsymbol{z}_t - \bar{\mathrm{z}})(\boldsymbol{z}_{t-\ell} - \bar{\mathrm{z}})'. \tag{1.17}$$

In particular, the sample covariance matrix is

$$\hat{\boldsymbol{\Gamma}}(0) = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{z}_t - \bar{\mathrm{z}})(\boldsymbol{z}_t - \bar{\mathrm{z}})'.$$

The lag-$\ell$ sample CCM of $\boldsymbol{z}_t$ is then

$$\hat{\boldsymbol{R}}(\ell) = \hat{\boldsymbol{D}}^{-1/2}\hat{\boldsymbol{\Gamma}}(\ell)\hat{\boldsymbol{D}}^{-1/2}, \tag{1.18}$$

where $\hat{\boldsymbol{D}} = \text{diag}\{\hat{\gamma}_{11}(0), \ldots, \hat{\gamma}_{kk}(0)\}$. From Eq. (1.18), we have

$$\hat{\boldsymbol{R}}(\ell) = [\hat{\rho}_{ij}(\ell)] \quad \text{with} \quad \hat{\rho}_{ij}(\ell) = \frac{\hat{\gamma}_{ij}(\ell)}{\sqrt{\hat{\gamma}_{ii}(0)\hat{\gamma}_{jj}(0)}}.$$

Under some mild conditions, $\hat{\boldsymbol{R}}(\ell)$ is a consistent estimator of $\boldsymbol{R}(\ell)$ for a weakly stationary process $\boldsymbol{z}_t$. The conditions are met if $\boldsymbol{z}_t$ is a stationary Gaussian process. The asymptotic properties of $\hat{\boldsymbol{R}}(\ell)$, however, are rather complicated. They depend on the dynamic dependence of the process $\boldsymbol{z}_t$. However, the results for some special cases are easier to understand. Suppose the vector time series $\boldsymbol{z}_t$ is a white noise process so that $\boldsymbol{R}(\ell) = \boldsymbol{0}$ for $\ell \neq 0$. Then, we have

$$\text{Var}[\hat{\rho}_{ij}(\ell)] \approx \frac{1}{T}, \quad \ell \neq 0.$$

and, for $i \neq j$,

$$\text{Var}[\hat{\rho}_{ij}(0)] \approx \frac{[1-\rho_{ij}^2(0)]^2}{T}. \tag{1.19}$$

For further details, see, for instance, Reinsel (1993, section 4.1.2). These simple results are useful in exploratory data analysis of vector time series. This is particularly so when the dimension $k$ is large. An important feature in analyzing time series data is to explore the dynamic dependence of the observed time series. For a $k$-dimensional series, each sample CCM $\hat{\boldsymbol{R}}(\ell)$ is a $k \times k$ matrix. If $k = 10$, then each CCM contains 100 sample correlations. It would not be easy to decipher the pattern embedded in the sample CCM $\hat{\boldsymbol{R}}(\ell)$ simultaneously for $\ell = 1, \ldots, m$, where $m$ is a given positive

integer. To aid the reading of sample CCM, Tiao and Box (1981) device a simplified approach to inspect the features of sample CCM. Specifically, consider $\hat{\boldsymbol{R}}(\ell)$ for $\ell > 0$. Tiao and Box define a corresponding simplified matrix $\boldsymbol{S}(\ell) = [S_{ij}(\ell)]$, where

$$S_{ij}(\ell) = \begin{cases} + & \text{if } \hat{\rho}_{ij}(\ell) \geq 2/\sqrt{T}, \\ \cdot & \text{if } |\hat{\rho}_{ij}(\ell)| < 2/\sqrt{T}, \\ - & \text{if } \hat{\rho}_{ij}(\ell) \leq -2/\sqrt{T}, \end{cases}$$

where it is understood that $1/\sqrt{T}$ is the sample standard error of elements of $\hat{\boldsymbol{R}}(\ell)$ provided that $\boldsymbol{z}_t$ is a Gaussian white noise. It is much easier to comprehend the dynamic dependence of $\boldsymbol{z}_t$ by inspecting the simplified CCM $\boldsymbol{S}(\ell)$ for $\ell = 1, \ldots, m$. We demonstrate the effectiveness of the simplified CCM via an example.

**Example 1.1**

Consider the three temperature series of Figure 1.1. The sample means of the three series are 0.15, 0.14, and 0.12, respectively, for Europe, North America, and South America. The sample standard deviations are 1.02, 1.06, and 0.47, respectively. The average temperatures of November for South America appear to be lower and relatively less variable compared with those of Europe and North America. Turn to sample CCMs. In this particular instance, $k = 3$ so that each CCM contains 9 real numbers. Suppose we like to examine the dynamic dependence of the three series using the first 12 lags of sample CCM. This would require us to decipher 108 numbers simultaneously. On the other hand, the dynamic dependence pattern is relatively easy to comprehend by examining the simplified CCMs, which are given below.

```
Simplified matrix:
CCM at lag: 1
. + +
. . +
. . +
CCM at lag: 2
. . +
. . +
+ . +
CCM at lag: 3
. . +
. . .
. + +
CCM at lag: 4
. . .
. . .
. + +
CCM at lag: 5
. . .
. . +
. + +
```

```
CCM at lag: 6
. . +
. . .
. . +
CCM at lag: 7
. . .
. . +
. . +
CCM at lag: 8
. . .
. . .
. . +
CCM at lag: 9
. . +
. . .
. . +
CCM at lag: 10
. . .
. . .
. . +
CCM at lag: 11
. . .
. . +
. . +
CCM at lag: 12
. . .
. . +
. . +
```

From the simplified CCM, we make the following observations. First, the dynamic dependence of the temperature series of South America appears to be persistent as all of its sample ACFs are large (showing by the + sign). This is not surprising as the time plot of the series in Figure 1.1 shows an upward trend. Second, the temperature series of Europe and North America are dynamically correlated with that of South America because there exist several + signs at the (1,3)th and (2,3)th elements of the CCMs. Finally, there seems to have no strong dynamic dependence in November temperatures between Europe and North America, because most sample cross-correlations at the (1,2)th and (2,1)th positions are small. ■

**Remarks.** Some remarks are in order.

- The demonstration of Example 1.1 is carried out by the **MTS** package in R. The command used is `ccm` with default options.

```
da <- read.table("temperatures.txt",header=TRUE)
da <- da[,-1] # remove time index
require(MTS)
ccm(da)
```

- If the dimension $k$ is close to or larger than the sample size $T$, then the aforementioned properties of sample CCMs do not hold. We shall discuss the situation in later chapters, e.g. Chapter 6.
- When the dimension $k$ is large, even the simplified CCM $\boldsymbol{S}(\ell)$ becomes hard to comprehend and some summary statistics must be used. In this book, we provide three summary statistics to extract helpful information embedded in the sample CCMs. They are given below:
  1. $P$-value plot: A scatterplot of the $p$-values of the null hypothesis $H_0 : \boldsymbol{R}(\ell) = \mathbf{0}$ versus lag $\ell$. The test statistic used asymptotically follows a $\chi^2_{k^2}$ distribution under the null hypothesis that there are no serial or cross correlations in $\boldsymbol{z}_t$ for $\ell > 0$. Details of the test statistic is given in Chapter 3.
  2. Diagonal element plot: A scatterplot of the fraction of significant diagonal elements of $\widehat{\boldsymbol{R}}(\ell)$ versus $\ell$. Here an element of $\widehat{\boldsymbol{R}}(\ell)$ is said to be significant if it is greater than $2/\sqrt{T}$ in absolute value.
  3. Off-diagonal element plot: A scatterplot of the fraction of significant off-diagonal elements of $\widehat{\boldsymbol{R}}(\ell)$ versus $\ell$. Again, the significance of an element of $\widehat{\boldsymbol{R}}(\ell)$ is with respect to $2/\sqrt{T}$.

The R command for the three summary plots is `Summaryccm` of the **SLBDD** package, which is a package associated with this book.

**Example 1.2**

Consider the daily log returns of the 99 financial market indexes of the world in Figure 1.2. Here $k = 99$ and it would be hard to comprehend any 99-by-99 cross correlation matrix. Figure 1.14 shows the three summary statistics of the daily log returns. The upper plot is the scatterplot of $p$-values. From the plot, all $p$-values are close to zero indicating that there are serial and cross-correlations in the 99-dimensional log returns for lags 1 to 12. The middle plot shows the fraction of significant diagonal elements of $\widehat{\boldsymbol{R}}(\ell)$. This plot shows that there are significant serial correlations in the daily log returns of 99 world financial indexes. But, as expected, the serial correlations decay quickly so that the fraction approaches zero as $\ell$ increases. The bottom plot shows the fraction of significant off-diagonal elements of $\widehat{\boldsymbol{R}}(\ell)$ versus $\ell$. The plot suggests that there is significant cross dependence among the daily log returns of world financial market indexes when $\ell$ is small. As expected, the cross dependence also decays to zero quickly as $\ell$ increases.

**R command used for the summary CCM plots**:

```
require(SLBDD)
da <- read.csv2("Stock_indexes_99world.csv",header=FALSE)
da <- as.matrix(da[,-1]) # remove the time index
rt <- diff(log(da)) # log returns
Summaryccm(rt)
```

■

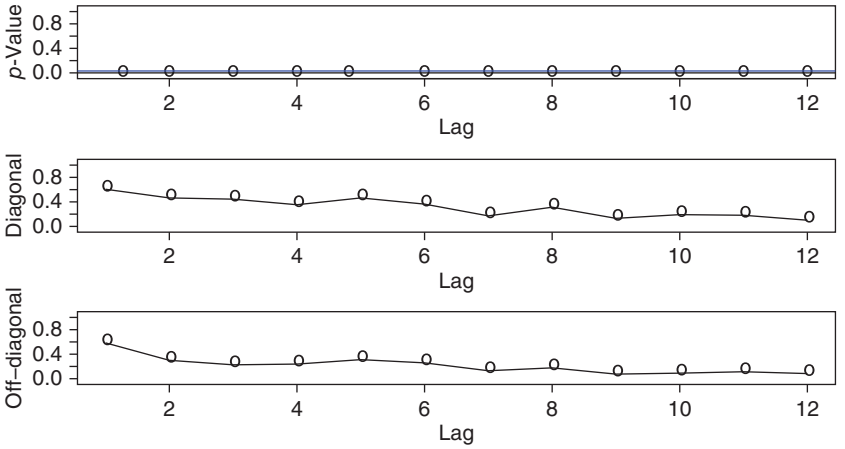


**Figure 1.14** Scatterplots of three summary statistics of lagged CCMs of the daily log returns of 99 world financial market indexes. The summary statistics are (*a*) *p*-values of testing $\boldsymbol{R}(\ell) = \boldsymbol{0}$, (*b*) fractions of significant diagonal elements of $\widehat{\boldsymbol{R}}(\ell)$, and (*c*) fractions of significant off-diagonal elements of $\widehat{\boldsymbol{R}}(\ell)$.

## 1.4 NONSTATIONARY PROCESSES

A vector process $\boldsymbol{z}_t$ is nonstationary if it fails to meet the stationarity conditions. Simply put, if some joint distributions of $\boldsymbol{z}_t$ change over time, then $\boldsymbol{z}_t$ is nonstationary. In this case, the moments of $\boldsymbol{z}_t$ may depend on time. Since we typically have a single observation at a given time index $t$, further assumptions are needed to render estimation possible for a nonstationary process. The most commonly employed nonstationary processes are the *integrated processes*. Such processes are commonly referred to as unit-root nonstationary series.

A scalar time series $z_t$ is called an integrated process of order 1 if its increment $w_t = z_t - z_{t-1}$ is a stationary and invertible process. In the econometric literature, a stationary and invertible process is called an $I(0)$ process, and an integrated process of order 1 is called an $I(1)$ process. A random-walk series is an example of an $I(1)$ process. Let $B$ denote the back-shift operator, such that $Bz_t = z_{t-1}$. Then, $z_t$ is an $I(d)$ process if $w_t = (1 - B)^d z_t$ is an $I(0)$ process. Following the convention, we may also use the notation $\nabla = (1 - B)$ in describing integrated processes. The operator $(1 - B)$ is called the first-difference operator, and the process $w_t = (1 - B)z_t$ is the first-differenced series of $z_t$.

In a similar way, a vector process $\boldsymbol{z}_t$ is called an $I(d)$ process if $\boldsymbol{w}_t = (1 - B)^d \boldsymbol{z}_t$ is a stationary and invertible series, where $(1 - B)^d$ applies to each and every component of $\boldsymbol{z}_t$. Of course, there is no particular reason to expect that all components of $\boldsymbol{z}_t$ are integrated of the same order $d$. Therefore, let $\boldsymbol{d} = (d_1, \ldots, d_k)'$, where $d_i$ are nonnegative integers and $\max\{d_i\} > 0$. We call the vector process $\boldsymbol{z}_t$ an $I(\boldsymbol{d})$ process if $\boldsymbol{w}_t = (w_{1t}, \ldots, w_{kt})'$, where $w_{it} = (1 - B)^{d_i} z_{it}$, is a stationary and invertible process. In this case, we may define $\boldsymbol{D} = \text{diag}\{\nabla^{d_1}, \ldots, \nabla^{d_k}\}$ and write $\boldsymbol{w}_t = \boldsymbol{D}\boldsymbol{z}_t$.

An important concept of the integrated vector process $\boldsymbol{z}_t$ is *co-integration*. Suppose that the components $z_{it}$ are $I(1)$ processes. If there exists a

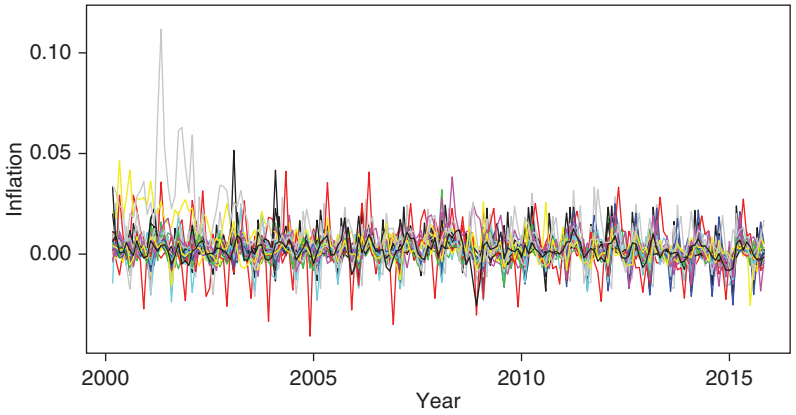


**Figure 1.15** Time plots of the first-differenced series of the 33 log monthly price indexes of European countries from January 2000 to October 2015.

$k$-dimensional vector $\boldsymbol{c} = (c_1, \ldots, c_k)' \neq \mathbf{0}$ such that $w_t = \boldsymbol{c}'\boldsymbol{z}_t$ is an $I(0)$ process, then $\boldsymbol{z}_t$ is called a co-integrated process and $\boldsymbol{c}$ is the co-integrating vector. The $I(0)$ process $w_t$ is the co-integrated series, which is stationary and invertible. In general, one can have more than one co-integrated process for a $k$-dimensional unit-root process $\boldsymbol{z}_t$. We shall discuss co-integration and its associated properties later.

An important class of nonstationary processes consists of seasonal time series. For a seasonal time series, the mean function is not constant but varies over time accordingly to certain periodic pattern. This pattern is repeated every season in a given year such as every quarter, month, or week. The number of seasons in a year is referred to as *periodicity*. For monthly series, the pattern is often repeated year after year and we have a periodicity $s = 12$. For quarterly data, we have $s = 4$. Consider the 33 monthly price indexes of Figure 1.7. All of the series exhibit an upward trend so that they are not stationary. Figure 1.15 shows the time plots of the first-differenced series of the log price indexes. From the plots, as expected, the trends disappear and the series (i.e. inflation rate) now fluctuate around zero. Therefore, there is evidence that the log price indexes are $I(1)$ processes. The cyclic patterns shown in Figure 1.15 indicate that some months have higher values than others so that the processes belong to the class of seasonal time series. We shall discuss seasonal time series later. For now, it suffices to introduce the idea of *seasonal difference*. The seasonal difference is given by $w_t = (1 - B^s)z_t = z_t - z_{t-s}$, which denotes the annual increment. For convenience, we refer to $1 - B$ as the regular difference.

Figure 1.16 shows the time plots of the regularly and seasonally differenced series of the 40 price indexes of Figure 1.7. Specifically, the time series shown are $w_t = (1 - B)(1 - B^{12})z_t$. From the plots, the seasonal patterns are largely removed. A seasonal time series that requires seasonal difference to achieve stationarity yet remains invertible is referred to as a seasonally integrated process. The idea of co-integration also generalizes to seasonal co-integration. Again, details are given in Chapter 3.

Another type of nonstationarity that attracts much attention is the locally stationary process. A stochastic process is locally stationary if its mean and variance functions evolve slowly over time. A formal definition of such processes can be

found in Dahlhaus (2012). Different from the conventional time series analysis, statistical inference on locally stationary processes is based on in-filled asymptotics. To illustrate, consider the time-varying parameter autoregressive process

$$z_t = \phi_t z_{t-1} + \sigma_t \epsilon_t, \tag{1.20}$$

where $\{\epsilon_t\}$ is a sequence of standard normal random variates, and $\phi_t$ and $\sigma_t$ are rescaled smooth functions such that $\phi_t \equiv \phi(t/T) : [0,1] \to (-1,1)$ and $\sigma_t \equiv \sigma(t/T) : [0,1] \to (0,\infty)$, where $T$ is the sample size. When $T$ increases, the observations on $\phi(t/T)$ and $\sigma(t/T)$ become denser in their domain [0,1] so that we can obtain better estimates of the time-varying parameters. Figure 1.17 shows the time plot of a realization with 300 observations from the time-varying parameter AR(1) model of Eq. (1.20), where $\phi(t/T) = 0.2 + 0.4(t/T) - 0.2(t/T)^2$ and $\sigma(t/T) = \exp(t/T)$ with $T = 300$. As expected, the series fluctuates around zero with an increasing variability. The process is locally stationary because it can be approximated by a stationary series within a small time interval. For instance, in this particular example, the first 120 observations of Figure 1.17 appears to be weakly stationary.

## 1.5 PRINCIPAL COMPONENT ANALYSIS

Curse of dimensionality is a well-known difficulty in analyzing big dependent data. This is so because it is hard to decipher useful information embedded in high-dimensional time series. Some exploratory data analyses are helpful in this situation. A powerful tool to explore the dynamic dependence in a vector time series is the *principal component analysis* (PCA). Even though PCA was developed primarily for independent data, it has been shown to be useful in time series analysis too. See, for instance, Peña and Box (1987), Tsay (2014, chapter 6) and the references therein. See also Taniguchi and Krishnaiah (1987) and Chang et al. (2014) for some theoretical developments. However, there are differences between applying PCA to independent data and to time series data. The main difference is that PCA is

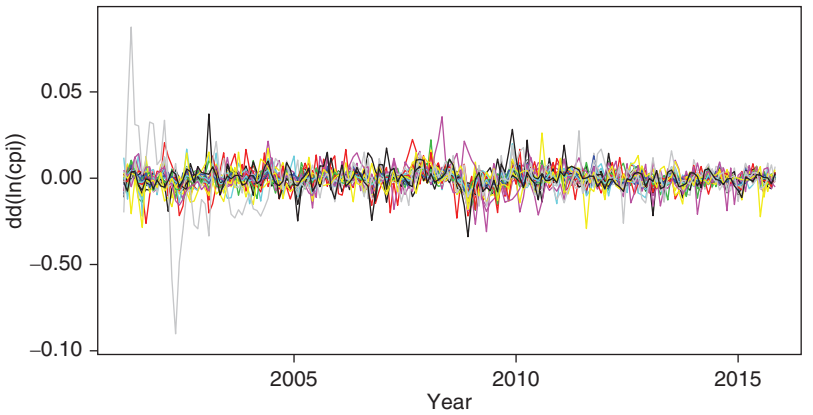


**Figure 1.16** Time plots of the first and seasonally differenced series of the 30 monthly price indexes of European countries from January 2000 to October 2015.

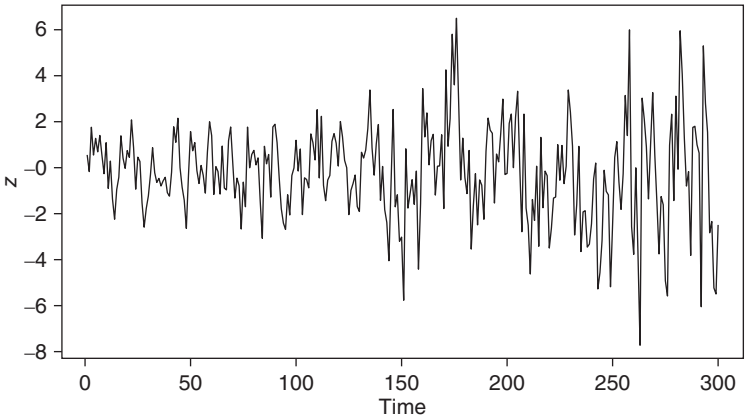


**Figure 1.17** Time plot of 300 realizations from a time-varying parameter autoregressive model of order 1 in Eq. (1.20), where $\phi(t/T) = 0.2 + 0.4(t/T) - 0.2(t/T)^2$ and $\sigma(t/T) = \exp(t/T)$.

mainly concerned with the covariance matrix, not the lag covariance matrices, which are important in time series analysis. Therefore, the limiting properties of the eigenvalues and eigenvectors of PCA are different for time series data. See a recent paper by Zhang and Tong (2020) and some discussions in Section 1.6. Also, we present in Chapter 6 a more general procedure, called the dynamic principal component analysis (DPCA) for time series data, that takes into account the lag dependence of a vector time series.

There are two main ways to introduce PCA. The first way is variance decomposition, and the second one optimal reconstruction or interpolation. We start with the variance decomposition of independent data at the population level. Let $\boldsymbol{z} = (z_1, \ldots, z_k)'$ be a $k$-dimensional random vector with mean zero and positive-definite covariance $\boldsymbol{\Gamma}(0)$. The first population principal component (PC) $y_1$ of $\boldsymbol{z}$ is a linear combination $y_1 = \boldsymbol{c}_1'\boldsymbol{z}$ satisfying $\boldsymbol{c}_1'\boldsymbol{c}_1 = 1$ such that $\mathrm{Var}(y_1)$ attends the maximum among all possible linear combinations of $\boldsymbol{z}$ built with vectors of unit length. The second PC of $\boldsymbol{z}$ is defined as a linear combination of $\boldsymbol{z}$, say $y_2 = \boldsymbol{c}_2'\boldsymbol{z}$ satisfying (i) $\boldsymbol{c}_2'\boldsymbol{c}_2 = 1$ and (ii) $\boldsymbol{c}_2'\boldsymbol{c}_1 = 0$ such that $\mathrm{Var}(\ y_2)$ assumes the maximum variance among all linear combinations $\boldsymbol{z}$ built with vectors of unit length and such that the second PC is uncorrelated with the first PC. In general, the $i$th PC of $\boldsymbol{z}$ is a non-zero linear combination of $\boldsymbol{z}$, say $y_i = \boldsymbol{c}_i'\boldsymbol{z}$ using a vector of unit length that has the maximum variance among all linear combinations of $\boldsymbol{z}$ and that is uncorrelated to all the previous PCs, that is, $\boldsymbol{c}_i$ satisfies (iii) $\boldsymbol{c}_i'\boldsymbol{c}_i = 1$ and (iv) $\boldsymbol{c}_i'\boldsymbol{c}_j = 0$ for $j = i - 1, \ldots, 1$. The unit length normalization, $\boldsymbol{c}_i'\boldsymbol{c}_i = 1$, is needed to control the scaling effect; otherwise the maximum has no meaning.

Using properties of positive-definite matrices given in the Appendix, the PCA of a stationary process $\boldsymbol{z}_t$ with positive-definite covariance matrix $\boldsymbol{\Gamma}(0)$ can be obtained from the spectral decomposition of $\boldsymbol{\Gamma}(0)$. Specifically, let $(\lambda_i, \boldsymbol{e}_i)$ be the $i$th eigenvalue-eigenvector pair of $\boldsymbol{\Gamma}(0)$, where the eigenvalues satisfy $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. The spectral decomposition of $\boldsymbol{\Gamma}(0)$ is

$$\boldsymbol{\Gamma}(0) = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1' + \cdots + \lambda_k \boldsymbol{e}_k \boldsymbol{e}_k'.$$

Then, the $i$th PCA is $y_{it} = \boldsymbol{e}_i'\boldsymbol{z}_t$. From the definition, $\text{Var}(y_{it}) = \text{Var}(\boldsymbol{e}_i'\boldsymbol{z}) = \boldsymbol{e}_i'\boldsymbol{\Gamma}(0)\boldsymbol{e}_i = \lambda_i\boldsymbol{e}_i'\boldsymbol{e}_i = \lambda_i$. Since $\text{tr}[\boldsymbol{\Gamma}(0)] = \sum_{i=1}^k \sigma_{ii} = \sum_{i=1}^k \lambda_i$, the proportion of variability of $\boldsymbol{z}$ explained by its $i$th PC is $\lambda_i/(\sum_{j=1}^k \lambda_j)$. Similarly, the proportion of variability of $\boldsymbol{z}_t$ explained by the first $m$ PCs is $(\sum_{i=1}^m \lambda_i)/(\sum_{j=1}^k \lambda_j)$. If a small $m$ can be found such that the first $m$ PCs explain a large proportion of the variability in $\boldsymbol{z}$, then one can focus analysis on the first $m$ PCs.

Suppose now that, instead of independent variables, we have a random realization $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T\}$ of a stationary vector time series $\boldsymbol{z}_t$. For simplicity, we assume that the process is mean-adjusted, i.e. $\sum_{t=1}^T \boldsymbol{z}_t/T = \boldsymbol{0}$. Define the $T \times k$ matrix $\boldsymbol{Z} = [\boldsymbol{z}_{(1)}, \ldots, \boldsymbol{z}_{(k)}]$, where $\boldsymbol{z}_{(j)} = (z_{j1}, \ldots, z_{jT})'$ is the vector of observations of the $j$th component $z_{jt}$. Then, the covariance matrix of $\boldsymbol{z}_t$ is estimated by the sample covariance matrix $\widehat{\boldsymbol{\Gamma}}(0) = \boldsymbol{Z}'\boldsymbol{Z}/T$ and the eigenvalue-eigenvector pairs $(\lambda_i, \boldsymbol{c}_i)$ of this matrix can be used to construct PCs. The first PC is given by $\boldsymbol{y}_1 = \boldsymbol{Z}\boldsymbol{c}_1$, where

$$\frac{1}{T}\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{c}_1 = \lambda_1\boldsymbol{c}_1, \quad \boldsymbol{c}_1'\boldsymbol{c}_1 = 1.$$

Pre-multiplying the above equation by $\boldsymbol{Z}$, we have

$$\frac{1}{T}\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{y}_1 = \lambda_1\boldsymbol{y}_1. \tag{1.21}$$

Therefore, the first PC $\boldsymbol{y}_1$ of $\boldsymbol{z}_t$ is proportional to the eigenvector corresponding to the largest eigenvalue of the matrix $\boldsymbol{Z}\boldsymbol{Z}'$. Other PCs can be obtained in a similar manner.

An alternative approach to introduce PCA is from an optimal reconstruction or interpolation of the data. For simplicity, we consider the sample version with the mean-adjusted stationary time series data as before, i.e. $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T\}$. We want to obtain a new time series, $f_{1t}$, defined as a linear combination of $\boldsymbol{z}_t$ that minimizes the error in the reconstruction of the observed data. More formally, let $f_{1t} = \boldsymbol{z}_t'\boldsymbol{\beta}_1$ and $\boldsymbol{f}_1 = \boldsymbol{Z}\boldsymbol{\beta}_1$, where $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{k1})'$ and $\boldsymbol{f}_1 = (f_{11}, \ldots, f_{1T})'$. The coefficients $\beta_{i1}$ are obtained by minimizing the objective function

$$L(\boldsymbol{\beta}_1, \boldsymbol{f}_1) = \sum_{t=1}^T \sum_{i=1}^k (z_{it} - \beta_{i1}f_{1t})^2. \tag{1.22}$$

Since both $f_{1t}$ and $\boldsymbol{\beta}_1$ are unknown, the solution to Eq. (1.22) is not well-defined, because $\beta_{i1}f_{1t} = (\beta_{i1}/h)(hf_{1t}) = \beta_{i1}^* f_{1t}^*$, where $h$ is an arbitrary non-zero real number. To overcome the difficulty of non-uniqueness, we assume that $f_{1t}$ satisfies $\sum_{t=1}^T f_{1t}^2 = 1$. The scale of $f_{1t}$ is then identified, and we can take the partial derivative of $L$ with respective to $\beta_{i1}$. By equating the partial derivatives to zero, we have

$$\hat{\beta}_{i1} = \frac{\sum_{t=1}^T z_{it}f_{1t}}{\sum_{t=1}^T f_{1t}^2} = \sum_{t=1}^T z_{it}f_{1t}, \quad i = 1, \ldots, k, \tag{1.23}$$

or, in vector form, $\widehat{\boldsymbol{\beta}}_1 = \boldsymbol{Z}'\boldsymbol{f}_1$. Using $\boldsymbol{f}_1 = \boldsymbol{Z}\boldsymbol{\beta}_1$, we obtain

$$\widehat{\boldsymbol{\beta}}_1 = \boldsymbol{Z}'\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_1,$$

and, hence, $\widehat{\boldsymbol{\beta}}_1$ is an un-normalized eigenvector of the matrix $\boldsymbol{Z}'\boldsymbol{Z}$.

Similarly, we can take the partial derivative of $L$ with respective to $f_{1t}$ and obtain

$$f_{1t} = \frac{\sum_{i=1}^{k} z_{it}\beta_{i1}}{\sum_{i=1}^{k} \beta_{i1}^2}, \quad t = 1, \ldots, T. \tag{1.24}$$

The numerator of Eq. (1.24) can be written as $\boldsymbol{z}_t'(\boldsymbol{Z}'\boldsymbol{f}_1)$. Letting $c = \sum_{i=1}^{k} \beta_{i1}^2$, plugging Eq. (1.23) into Eq. (1.24), and using matrix notation, we have

$$\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{f}_1 = c\boldsymbol{f}_1. \tag{1.25}$$

This implies that $\boldsymbol{f}_1$ is an eigenvector of the matrix $\boldsymbol{Z}\boldsymbol{Z}'$ associated with eigenvalue $c$.

The function in Eq. (1.22) is an objective function of least squares estimation, and the fitted value is given by $\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{f}_1$. Therefore, the value of objective function evaluated at the estimates becomes

$$\begin{aligned}
\widehat{L} &= \sum_{t=1}^{T}\sum_{i=1}^{k} z_{it}^2 - [\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{f}_1]'(\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{f}_1) \\
&= \sum_{t=1}^{T}\sum_{i=1}^{k} z_{it}^2 - \boldsymbol{f}_1'\boldsymbol{Z}\boldsymbol{Z}'(c\boldsymbol{f}_1) \\
&= \sum_{t=1}^{T}\sum_{i=1}^{k} z_{it}^2 - c^2,
\end{aligned}$$

where we have used $\boldsymbol{f}_1'\boldsymbol{f}_1 = \sum_{t=1}^{T} f_{1t}^2 = 1$. Consequently, to achieve the minimum value of the objective function $\widehat{L}$, we choose $c$ to be the largest eigenvalue of $\boldsymbol{Z}\boldsymbol{Z}'$. The vector $\boldsymbol{f}_1$ is not equal to the first PC $\boldsymbol{y}_1$, because it has variance 1, whereas the PC has variance $\lambda_1$, but apart from this scaling effects they are identical, as both are proportional to the eigenvalues of the matrix $\boldsymbol{Z}\boldsymbol{Z}'$, as shown in (1.21) and (1.25).

For the second sample PC, we can follow a similar procedure by replacing the observed data $\boldsymbol{z}_t$ by the residual $\boldsymbol{\epsilon}_t = \boldsymbol{z}_t - \widehat{\boldsymbol{\beta}}_1 f_{1t}$. This process can be repeated until all sample PCs are introduced.

In summary, both approaches led to the same result with a difference in scale. In the first approach, the linear combinations of maximum variance are defined by unit norm vectors that are the eigenvectors of the covariance matrix and the PCs have different variances. In the second approach, the optimal interpolations are defined by vectors that are proportional to the eigenvalues of the covariance matrix and the linear combinations are standardized to have unit variance.

### 1.5.1 Discussion

We have introduced PCA using both the variance decomposition and the reconstruction approaches. However, there is a subtle difference between the two versions. In the first case we assume stationarity. On the other hand, for a given data set, the reconstruction approach can be applied to any time series.

### 1.5.2 Properties of the PCs

From the introduction, PCs $\boldsymbol{y}_t = (y_{1t}, \ldots, y_{kt})'$ of a stationary vector process $\boldsymbol{z}_t$ are linear transforms of the observed data, and they are ordered according to their variabilities such that $y_{1t}$ has the largest variance. In particular, $\text{Var}(y_{it}) = \lambda_i$, which is the $i$th largest eigenvalue of the covariance matrix of $\boldsymbol{z}_t$. Since $\sum_{i=1}^{k} \text{Var}(z_{it}) = \sum_{i=1}^{k} \gamma_{ii}(0) = \text{tr}[\boldsymbol{\Gamma}(0)]$, where $\text{tr}[\boldsymbol{\Gamma}(0)]$ denotes the trace of the matrix $\boldsymbol{\Gamma}(0)$, and $\text{tr}[\boldsymbol{\Gamma}(0)] = \sum_{i=1}^{k} \lambda_i$, we see that PCs retain the total variability of the vector process $\boldsymbol{z}_t$.

Next, let $\boldsymbol{e}_i$ be the eigenvector associated with eigenvalue $\lambda_i$ of $\boldsymbol{\Gamma}(0)$ such that $\boldsymbol{e}_i'\boldsymbol{e}_i = 1$. We have $\boldsymbol{y}_t = \boldsymbol{P}'\boldsymbol{z}_t$, where $\boldsymbol{P} = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k]$ is the matrix of eigenvectors. This transformation matrix is an orthonormal matrix so that $\boldsymbol{P}\boldsymbol{P}' = \boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}$. Consequently, we have $1 = |\boldsymbol{P}\boldsymbol{P}'| = |\boldsymbol{P}||\boldsymbol{P}'|$. From the transformation, $\text{Cov}(\boldsymbol{y}_t) = \boldsymbol{P}'\text{Cov}(\boldsymbol{z}_t)\boldsymbol{P}$. Using $|\boldsymbol{P}||\boldsymbol{P}'| = 1$, we have $|\text{Cov}(\boldsymbol{y}_t)| = |\text{Cov}(\boldsymbol{z}_t)|$, implying that the generalized variance of $\boldsymbol{y}_t$ is the same as that of $\boldsymbol{z}_t$.

An important application of PCA is to reduce the dimension of a vector process. A scalar stochastic process would be close to a constant series if its variance is close to zero. Therefore, a PC associated with an eigenvalue $\lambda_i \approx 0$ would carry little information about the underlying process $\boldsymbol{z}_t$. As discussed before, the proportion of variability in $\boldsymbol{z}_t$ explained by the first $m$ PCs is $(\sum_{i=1}^{m} \lambda_i)/(\sum_{i=1}^{k} \lambda_k)$. In practice, one can select an $m$ such that the first $m$ PCs, $\{y_{it} | i = 1, \ldots, m\}$, explain a substantial portion of the variability of $\boldsymbol{z}_t$. In the literature, some methods have been proposed to aid the choice of $m$. A commonly used, yet informal, method is the *scree plot*, which is simply the scatterplot of $\lambda_i$ versus $i$. To choose $m$, one looks for an elbow (bend) in the scree plot.

**Example 1.3**

To demonstrate PCA, we consider, again, the 33 monthly price indexes shown in Figure 1.7. In this example, we employ the logarithms of the price indexes. Let $\boldsymbol{z}_t$ be the log series of the 33 monthly price indexes from January 2000 to October 2015 for 190 observations. A key feature of the price series shown in Figure 1.7 is the upward trends. These trends do not move in unison as different price series may increase at different paces. We apply PCA to $\boldsymbol{z}_t$ and obtain the results in Table 1.1.

From Table 1.1, we see that (a) the eigenvalue $\lambda_i$ decays quickly as $i$ increases, (b) the first PC alone explains about 95.8% of the variability in the data, and (c) the second PC explains about 3.3% of the total variability. As a matter of fact, the first

**TABLE 1.1 The Results of PCA Applied to the Log Series of Monthly Price Indexes of 33 European Countries. Only the Results of the First Six PCs Are Shown, Where $\lambda_i$ Is the $i$th Largest Eigenvalue**

| Component | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\sqrt{\lambda_i}$ | 0.968 | 0.176 | 0.075 | 0.042 | 0.031 | 0.020 |
| Proportion | 0.958 | 0.033 | 0.006 | 0.002 | 0.001 | 0.001 |
| Cumulative | 0.958 | 0.990 | 0.996 | 0.997 | 0.998 | 0.999 |

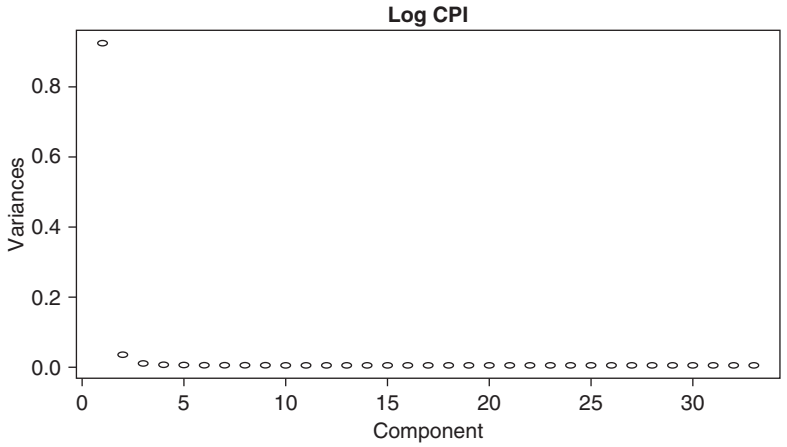


**Figure 1.18** Scree plot for the log series of 33 monthly price indexes of European countries from January 2000 to October 2015.

**TABLE 1.2 The Results of PCA Applied to the First-Differenced Log Series of Monthly Price Indexes. Only the Results of the First Six PCs Are Shown, Where $\lambda_i$ Is the *i*th Largest Eigenvalue**

| Component | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\sqrt{\lambda_i}$ | 0.0217 | 0.0165 | 0.0124 | 0.0120 | 0.0108 | 0.0075 |
| Proportion | 0.3011 | 0.1725 | 0.0975 | 0.0909 | 0.0736 | 0.0360 |
| Cumulative | 0.3011 | 0.4736 | 0.5711 | 0.6620 | 0.7356 | 0.7716 |

six PCs explain more than 99.9% of the sample variability in $\boldsymbol{z}_t$. Figure 1.18 shows the scree plot of the PCA of $\boldsymbol{z}_t$ and it confirms the findings that the first PC is the dominating factor and the first few PCs capture most of the variabilities in the log price indexes. Figure 1.19 shows the time plots of the first six PCs. From the plots, we see that the first PC captures the overall upward trend of the log price indexes. The second and third PCs also exhibit certain trending behavior, but they indicate that the trends do not necessarily move in a monotone manner. The fourth and sixth PCs show some periodic patterns of the price indexes, signifying the seasonal behavior of price indexes. Figure 1.20 shows the time plots of the 7th to 12th PCs. Even though those plots show certain variabilities, they are in a smaller scale compared with those of Figure 1.19.

To demonstrate that PCA are simply for variance decomposition, we also consider the first-differenced series $\boldsymbol{x}_t = (1 - B)\boldsymbol{z}_t$. The component $x_{it}$ of the series $\boldsymbol{x}_t$ represents the inflation rates (with respective to the previous month) of the $i$th price index. Table 1.2 gives the results of applying PCA to $\boldsymbol{x}_t$. The eigenvalues are much smaller compared with those of Table 1.1. This is understandable as trends are the dominating factors of the monthly price indexes. The inflation rates, on the other hand, are in a much smaller scale. For the inflation rates, the first six PCs explain

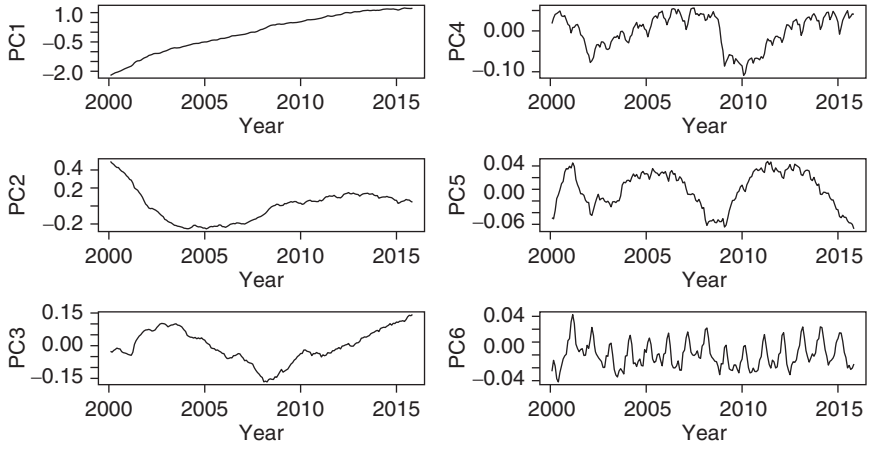


**Figure 1.19** Time plots of the first six PCs of the log series of 33 monthly price indexes of European countries.

about 77.16% of the data variabilities. Figure 1.21 shows the corresponding scree plot. From the plot, the first five PCs seem to distance themselves from the others and the decay of the eigenvalues $\lambda_i$ appears to be slow for $i \geq 6$. Finally, Figure 1.22 shows the time plots of the first six PCs of $\boldsymbol{x}_t$. It is interesting to see that the seasonal pattern of the monthly price indexes is now clearly captured by the first PC. In addition, as expected, there is no upward trend in the PCs.

This simple example shows that PCA can be applied to unit-root nonstationary time series. It also demonstrates that, as a tool for variance decomposition, results of PCA depend on the scale. The method is useful in capturing the dominating factors of variability for a given data set, but care must be exercised in interpreting the results of PCA. If one thought that the first two PCs of the log price indexes, which explain about 99.0% of the variabilities, are sufficient for the data, she would miss the important characteristics of seasonality.

### Main R commands used in Example 1.3

```
da <- read.csv2("CPIEurope2000-15.csv",header=TRUE)
x <- log(as.matrix(da))
m1 <- princomp(x)
m1
names(m1)
#[1] "sdev"    "loadings" "center"   "scale"   "n.obs" "scores"   "call"
# sdev: the square-root of eigenvalues
# loadings: the eigenvectors
# scores: the principal components
plot(1:33,m1$sdev^2,xlab="component",ylab="variance", pch="o",
main="log CPI") # Figure 1.18.
```

■

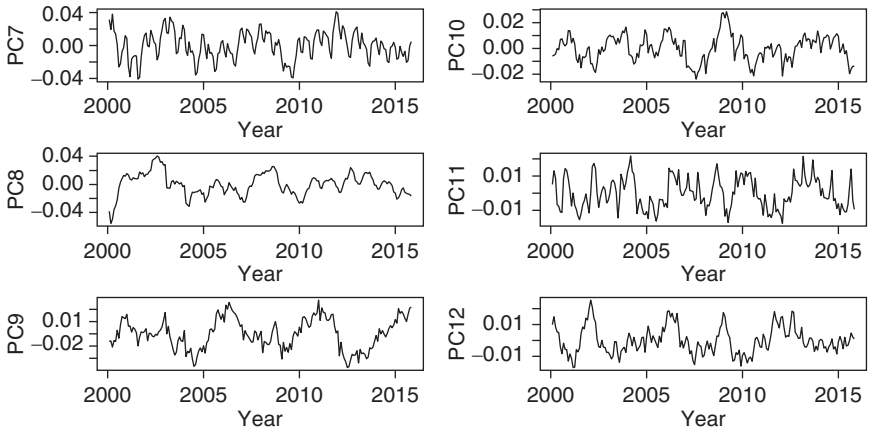


**Figure 1.20** Time plots of the 7th to 12th PCs of the log series of 33 monthly price indexes from European countries.

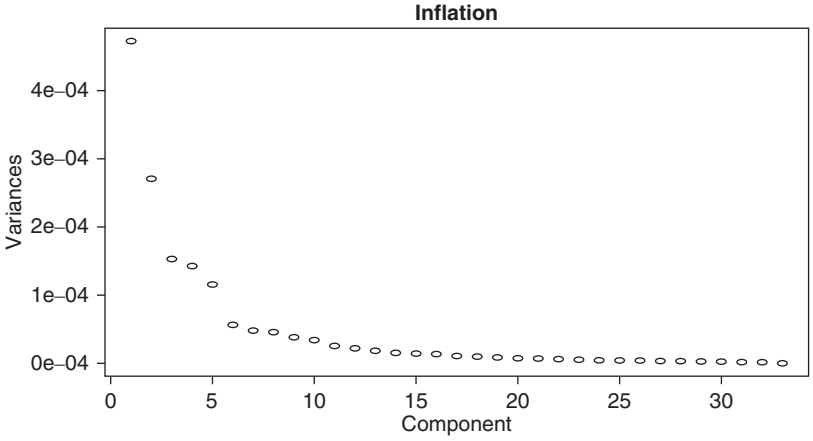


**Figure 1.21** Scree plot for the first-differenced log series of 33 monthly price indexes of European countries from January 2000 to October 2015.

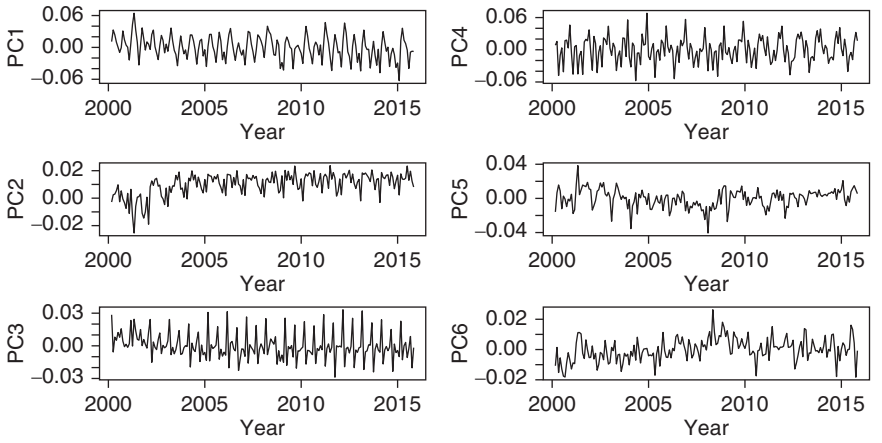


**Figure 1.22** Time plots of the first six PCs of the first-differenced log series of 33 monthly price indexes of European countries.

## 1.6 EFFECTS OF SERIAL DEPENDENCE

In the previous sections, we have introduced some examples of big dependent data and some basic concepts of stochastic processes. In this section, we use some simple examples to illustrate that statistical methods developed for analysis of independent data may encounter difficulties when they are applied to dependent data.

As stated before, PCA was proposed originally for independent data and the sample version of PCA is applicable to dependent data. Let $\boldsymbol{y}_t = \boldsymbol{P}'\boldsymbol{z}_t$ be the PCs of a vector process $\boldsymbol{z}_t$, where $\boldsymbol{P}$ is the orthonormal transformation matrix consisting of the eigenvectors of the sample covariance matrix of $\boldsymbol{z}_t$. From the orthogonality, we have $\text{Cov}(y_{it}, y_{jt}) = 0$ for $i \neq j$. That is, the PCs are contemporaneously uncorrelated. For independent data, $\text{Cov}(\boldsymbol{z}_t, \boldsymbol{z}_{t-j}) = \boldsymbol{0}$ provided $j \neq 0$. Consequently, for independent data, the PCs $\boldsymbol{y}_t$ are contemporaneously and serially uncorrelated. Therefore, we may analyze, in this case, each component $y_{it}$ separately if the analysis focuses on the first two moments of the data. In particular, under the Gaussian assumption, PCs are mutually independent.

Turn to dependent data. It is true that the PCs $\boldsymbol{y}_t$ are contemporaneously uncorrelated, but they can be dynamically (or serially) correlated. To demonstrate, Figure 1.23 shows the plots of autocorrelations and cross-correlations of $y_{2t}$ and $y_{3t}$ of the growth rates of the 40 price indexes of Figure 1.7. These two PCs are shown in Figure 1.22. From the correlation plots, $y_{2t}$ and $y_{3t}$ are serially correlated and, hence, they must be analyzed jointly. Consequently, PCA for dependent data may differ markedly from that for independent data. As a matter of fact, Zhang and Tong (2020) show that (a) the eigenvalues of the sample covariance matrix of a stationary and serially correlated vector time series are asymptotically correlated and (b) those eigenvalues are asymptotically dependent on all eigenvectors. On the other hand, for independent data, (a) eigenvalues and eigenvectors are asymptotically

independent and (b) eigenvalues are asymptotically uncorrelated under certain conditions.

Turn to statistical methods proposed for big independent data. One of the most commonly used test statistics is the one-sample $t$-test for testing that the mean of a scalar random variable is zero. For an observed data set $\{z_1, \ldots, z_T\}$, one computes the test statistic

$$t = \frac{\bar{z}}{\sqrt{s^2/T}}, \tag{1.26}$$

where $\bar{z} = \sum_{t=1}^{T} z_t/T$ is the sample mean and $s^2 = \sum_{t=1}^{T} (z_t - \bar{z})^2/(T-1)$ is the sample variance. For i.i.d. sample, the $t$ statistic of Eq. (1.26) follows asymptotically the $N(0,1)$ distribution so that statistical inference can be made. Suppose, on the other hand, $z_t$ follows an AR(1) model, say $z_t = \phi_0 + \phi_1 z_{t-1} + a_t$, where $\phi_0$ is a constant, $\phi_1 \neq 0$, and $\{a_t\}$ is a scalar white noise series. Then, the test statistic in Eq. (1.26) does not follow asymptotically the $N(0,1)$ distribution, because it is easy to verify that $\text{Var}(\sqrt{T}\bar{z})$ converges to $\text{Var}(z_t) \times \frac{1+\phi}{1-\phi}$ as $T \to \infty$. Therefore, the proper test statistic to use in this case is

$$t_{\text{dep}} = \frac{\bar{z}}{\sqrt{s^2(1+\phi)/[T(1-\phi)]}}. \tag{1.27}$$

From Eqs. (1.26) and (1.27), we have $t = t_{\text{dep}} \times \sqrt{(1+\phi)/(1-\phi)}$, so that $t$ and $t_{\text{dep}}$ can differ markedly, indicating that overlooking the serial dependence can lead to erroneous inference for the one-sample $t$-test. For example, if $\phi = 0.9$, then $t = t_{\text{dep}} \times \sqrt{19}$, and for $\phi = -0.9$, $t = t_{\text{dep}}/\sqrt{19}$. As another example, a widely used statistical method for analyzing big data in recent years is the Lasso regression of Tibshirani (1996). The method emphasizes on sparsity, and is applicable even when the sample size is smaller than the number of variables. Details of the method and its extensions will be given in Chapter 7. The Lasso was developed for independent data. Here we use a simple simulated example to demonstrate that Lasso regression may fail when it is applied to serially dependent data.

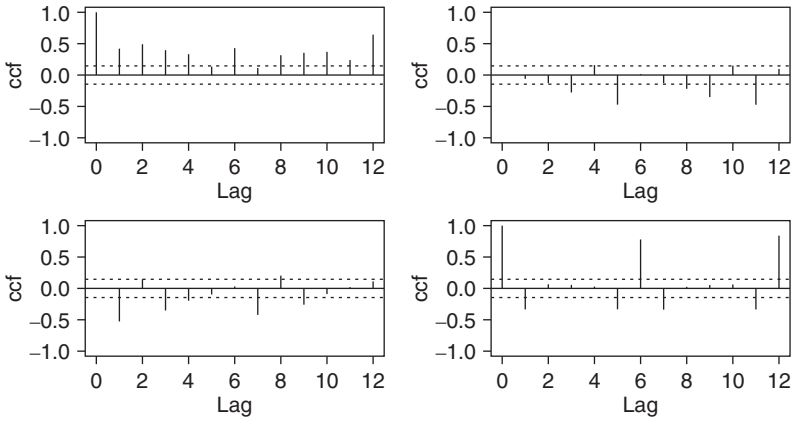


**Figure 1.23** The sample autocorrelations and cross-correlations of the second and third PCs of the inflation rates of 33 monthly price indexes. The diagonal plots are sample autocorrelations and the off-diagonal plots are sample cross-correlations.

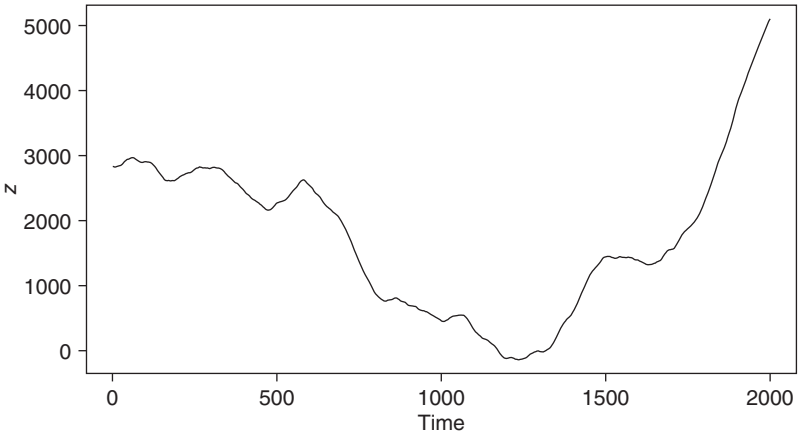


**Figure 1.24** Time plot of a simulated series with 2000 observations based on the model in Eq. (1.28).

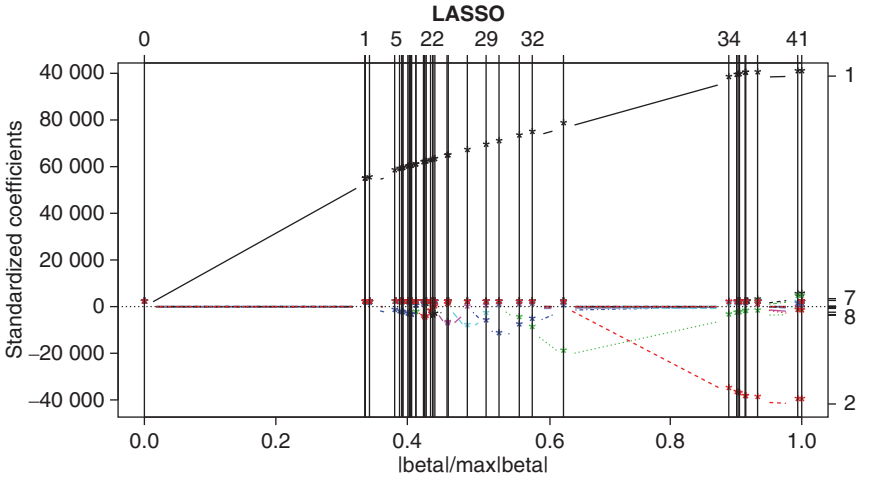


**Figure 1.25** Result of Lasso regression for the data in Example 1.4. The plot is obtained by the `lars` package of R.

### Example 1.4

We generated 2000 observations using the model

$$x_t = 1.9x_{t-1} - 0.8x_{t-2} - 0.1x_{t-3} + \epsilon_t, \quad t = 1, \ldots, 2000, \tag{1.28}$$

where $\epsilon_t$ are random draws from the standard normal distribution, i.e. $\epsilon_t \sim_{iid} N(0, 1)$. The model in Eq. (1.28) is a Gaussian ARIMA(1,2,0) model. The series is shown in Figure 1.24. We then consider a linear regression

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t, \quad t = 11, \ldots, 2000,$$

where the design vector is given by

$$\boldsymbol{x}_t = (x_{t-1}, x_{t-2}, \ldots, x_{t-10}, z_{1t}, \ldots, z_{10,t})',$$

where $z_{it} \sim_{iid} N(0,1)$ for all $i$ and $t$. Thus, we have a linear regression model with $p = 20$ and $T = 1990$. The data generating model is sparse with three non-zero coefficients. Our goal is to estimate the regression model and to identify the non-zero coefficients. In this particular case, a proper statistical method would identify $x_{t-1}, x_{t-2}$ and $x_{t-3}$ as the only predictors with non-zero coefficients. Figure 1.25 shows the result of the Lasso estimation. The plot is called a coefficient profile plot, which is a scatterplot of standardized coefficient estimates versus a properly scaled norm of the coefficient estimates. The numbers on top of the plot indicate the number of non-zero coefficients. The numbers on the right side of the plot identify the key predictors. Details and use of the coefficient profile plot are given in Chapter 7. It suffices now to note that large non-zero coefficient estimates are shown in the plot. From the plot, it is clear that the Lasso can easily identify $x_{t-1}$ and $x_{t-2}$ as important explanatory variables, but it fails to pin down $x_{t-3}$. This is not surprising because the AR(3) model in Eq. (1.28) is an $I(2)$ process which has strong serial dependence, leading to the difficulty of multicollinearity in linear regression analysis. In this particular case, the explanatory variables $x_{t-1}$ to $x_{t-10}$ are strongly correlated with sample correlations being close to 1. In addition, estimation of Lasso regression often requires normalization of each column of the design matrix. For an $I(2)$ series, the sample standard deviation grows at a higher order than the sample size. The normalization can easily lead to focusing purely on the unit-root dependence. This, in turn, may overlook the serial dependence of the stationary part of the process. Details can be justified by using the theory of unit-root processes. See, for instance, Tsay (2014) and the references therein. It suffices here to say that Lasso regression may fail for dependent data when the serial dependence is strong. We shall discuss ways to overcome such a difficulty in Chapter 7. Consequently, there is a need to develop statistical methods for analyzing big dependent data. ■

## APPENDIX 1.A: SOME MATRIX THEORY

In this appendix, we provide some properties of matrix useful in multivariate statistical analysis. Let $\boldsymbol{A}$ be a $k \times k$ real-valued matrix. The solutions of the determinant equation $|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0$ are the eigenvalues of $\boldsymbol{A}$. A $k$-dimension vector $\boldsymbol{x} = (x_1, \ldots, x_k)'$ is an eigenvector associated with an eigenvalue $\lambda$ of $\boldsymbol{A}$ if and only if $\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$. In general, eigenvalues and eigenvectors of $\boldsymbol{A}$ may assume complex values.

The matrix $\boldsymbol{A}$ is positive-definite if (a) $\boldsymbol{A} = \boldsymbol{A}'$, i.e. $\boldsymbol{A}$ is symmetric, and (b) for any non-zero $k$-dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_k)'$, $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} > 0$. Positive-definite matrices play an important role in many statistical applications and it pays to study their properties. It is easy to see that all eigenvalues of a positive-definite matrix $\boldsymbol{A}$ are positive.

**Spectral decomposition**: Assume that $\boldsymbol{A}$ is a $k \times k$ real-valued positive-definite matrix. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ be the eigenvalues of $\boldsymbol{A}$. Let $\boldsymbol{e}_i$ be an eigenvector of $\boldsymbol{A}$ associated with eigenvalue $\lambda_i$ such that $\boldsymbol{e}_i'\boldsymbol{e}_1 = 1$. That is the Euclidean norm of $\boldsymbol{e}_i$ is 1. Let $\boldsymbol{P} = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k]$ be the $k \times k$ matrix of eigenvectors. Then, we have (1)

$\boldsymbol{AP} = \boldsymbol{P\Lambda}$, where $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_k\}$; (2) $\boldsymbol{P'P} = \boldsymbol{PP'} = \boldsymbol{I}$, the $k$-dimensional identity matrix; (3)

$$\boldsymbol{A} = \boldsymbol{P\Lambda P'} = \sum_{i=1}^{k} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i'.$$

Property (3) is referred to as the spectral decomposition of the positive-definite matrix $\boldsymbol{A}$. Using the spectral decomposition and properties of eigenvalues and eigenvectors of $\boldsymbol{A}$, we have the following properties.

**Property I**: Assume that $\boldsymbol{A}$ is a positive-definite $k \times k$ real-valued matrix with spectral decomposition given by eigenvalue-eigenvector pairs $(\lambda_i, \boldsymbol{e}_i)$. Then,

$$\max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x'Ax}}{\boldsymbol{x'x}} = \lambda_1$$

$$\min_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x'Ax}}{\boldsymbol{x'x}} = \lambda_k,$$

where the maximum and minimum are attended with $\boldsymbol{x} = \boldsymbol{e}_1$ and $\boldsymbol{x} = \boldsymbol{e}_k$, respectively. Furthermore,

$$\max_{\boldsymbol{x} \neq \boldsymbol{0}, \boldsymbol{x'e}_j = 0; j = 1, \ldots, i} \frac{\boldsymbol{x'Ax}}{\boldsymbol{x'x}} = \lambda_{i+1}$$

where the maximum is attended when $\boldsymbol{x} = \boldsymbol{e}_{i+1}$, where $i = 2, \ldots, k$.

## EXERCISES

1. Consider the 99 world financial market indexes. Compute the log returns of the indexes. Obtain a time plot of all series and perform a PCA of the log returns. Summarize the results of PCA, including scree plot and time plots of the first six PCs. [An R command for PCA is `princomp`].
2. Consider the clothing data set of Figure 1.8. Perform PCA on the sales data and summarize the results. Obtain time plots of the first 12 PCs with 6 series on one page.
3. Consider, again, the clothing data set. Obtain the three summary plots of the sample cross-correlations for lags 1 to 21.
4. Consider the temperature data of Figure 1.1. (a) Obtain the sample mean and sample covariance matrix of the data. (b) Obtain the lag-1 to lag-10 sample CCMs of the data.
5. Consider the hourly $PM_{2.5}$ measurements at 15 monitoring stations in the southern Taiwan; columns 4 to 18 of the file `TaiwanPM25.csv`. (a) Compute the sample mean and sample covariance matrix of the 15 time series. (b) Obtain a time plot of the 15 time series.
6. Prove Eq. (1.16) and discuss the increase in the variance of sample mean with respect to the independent case when the series has a non-zero first-order autocorrelation coefficient and zero autocorrelations for all higher lags. That is, the series follows an MA(1) model.

## REFERENCES

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer, New York, NY.

Chang, J., Guo, B., and Yao, Q. (2014). Segmenting multiple time series by contemporaneous linear transformation: PCA for time series. Working paper, London School of Economics.

Dahlhaus, R. (2012). Locally stationary processes. In *Handbook of statistics*, **30**: 351–413.

Peña, D. and Box, G. (1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, **82**: 836–843.

Peña, D. and Rodriguez, J. (2003). Descriptive measures of multivariate scatter and linear dependence. *Journal of Multivariate Analysis*, **85**: 361–374.

Reinsel, G. (1993). *Elements of Multivariate Time Series Analysis*. Springer-Verlag, New York, NY.

Taniguchi, M. and Krishnaiah, P. (1987). Asymptotic distributions of functions of the eigenvalues of sample covariance matrix and canonical correlation matrix in multivariate time series. *Journal of Multivariate Analysis*, **22**: 156–176.

Tiao, G. C. and Box, G. E. P. (1981). Modeling multivariate time series with applications. *Journal of American Statistical Association*, **72**: 802–816.

Tibshirani, R. (1996). Regression shrinkage and selection via Lasso. *Journal of the Royal Statistical Society, Series B*, **58**: 267–288.

Tsay, R. S. (2014). *Multivariate Time Series Analysis with R and Financial Applications*. John Wiley & Sons, Hoboken, NJ.

Zhang, X. and Tong, H. (2020). Some cautionary comments on PCA for time series data. Working paper, London School of Economics and Political Science, UK.