# 8

# *Statistical Methods*

**Raghu Nandan Sengupta and Debasis Kundu**

**CONTENTS**

**ABSTRACT** The chapter of Statistical Methods starts with the basic concepts of data analysis and then leads into the concepts of probability, important properties of probability, limit theorems, and inequalities. The chapter also covers the basic tenets of estimation, desirable properties of estimates, before going on to the topic of maximum likelihood estimation, general methods of moments, Baye's estimation principle. Under linear and nonlinear regression different concepts of regressions are discussed. After which we discuss few important multivariate distributions and devote some time on copula theory also. In the later part of the chapter, emphasis is laid on both the theoretical content as well as the practical applications of a variety of multivariate techniques like Principle Component Analysis (PCA), Factor Analysis, Analysis of Variance (ANOVA), Multivariate Analysis of Variance (MANOVA), Conjoint Analysis, Canonical Correlation, Cluster Analysis, Multiple Discriminant Analysis, Multidimensional Scaling, Structural Equation Modeling, etc. Finally, the chapter ends with a good repertoire of information related to softwares, data sets, journals, etc., related to the topics covered in this chapter.

## 8.1 Introduction

Many people are familiar with the term *statistics*. It denotes recording of numerical facts and figures, for example, the daily prices of selected stocks on a stock exchange, the annual employment and unemployment of a country, the daily rainfall in the monsoon season, etc. However, statistics deals with situations in which the occurrence of some events cannot be predicted with certainty. It also provides methods for organizing and summarizing facts and for using information to draw various conclusions.

Historically, the word *statistics* is derived from the Latin word *status* meaning *state*. For several decades, statistics was associated solely with the display of facts and figures pertaining to economic, demographic, and political situations prevailing in a country. As a subject, statistics now encompasses concepts and methods that are of far-reaching importance in all enquires/questions that involve planning or designing of the experiment, gathering of data by a process of experimentation or observation, and finally making inference or conclusions by analyzing such data, which eventually helps in making the future decision.

Fact finding through the collection of data is not confined to professional researchers. It is a part of the everyday life of all people who strive, consciously or unconsciously, to know matters of interest concerning society, living conditions, the environment, and the world at large. Sources of factual information range from individual experience to reports in the news media, government records, and articles published in professional journals. Weather forecasts, market reports, costs of living indexes, and the results of public opinion are some other examples. Statistical methods are employed extensively in the production of such reports. Reports that are based on sound statistical reasoning and careful interpretation of conclusions are truly informative. However, the deliberate or inadvertent misuse of statistics leads to erroneous conclusions and distortions of truths.

## 8.2 Basic Concepts of Data Analysis

In order to clarify the preceding generalities, a few examples are provided:

*Socioeconomic surveys:* In the interdisciplinary areas of sociology, economics, and political science, such aspects are taken as the economic well-being of different ethnic groups, consumer expenditure patterns of different income levels, and attitudes toward pending legislation. Such studies are typically based on data oriented by interviewing or contacting a representative sample of person selected by statistical process from a large population that forms the domain of study. The data are then analyzed and interpretations of the issue in questions are made. See, for example, a recent monograph by Bandyopadhyay et al. (2011) on this topic.

*Clinical diagnosis:* Early detection is of paramount importance for the successful surgical treatment of many types of fatal diseases, say, for example, cancer or AIDS. Because frequent in-hospital checkups are expensive or inconvenient, doctors are searching for effective diagnosis process that patients can administer themselves. To determine the merits of a new process in terms of its rates of success in detecting true cases avoiding false detection, the process must be field tested on a large number of persons, who must then undergo in-hospital diagnostic test for comparison. Therefore, proper planning (designing the experiments) and data collection are required, which then need to be analyzed for final conclusions. An extensive survey of the different statistical methods used in clinical trial design can be found in Chen et al. (2015).

*Plant breeding:* Experiments involving the cross fertilization of different genetic types of plant species to produce high-yielding hybrids are of considerable interest to agricultural scientists. As a simple example, suppose that the yield of two hybrid varieties are to be compared under specific climatic conditions. The only way to learn about the relative performance of these two varieties is to grow them at a number of sites, collect data on their yield, and then analyze the data. Interested readers may refer to the edited volume by Kempton and Fox (2012) for further reading on this particular topic.

In recent years, attempts have been made to treat all these problems within the framework of a unified theory called decision theory. Whether or not statistical inference is viewed within the broader framework of decision theory depends heavily on the theory of probability. This is a mathematical theory, but the question of subjectivity versus objectivity arises in its applications and in its interpretations. We shall approach the subject of statistics as a science, developing each statistical idea as far as possible from its probabilistic foundation and applying each idea to different real-life problems as soon as it has been developed.

Statistical data obtained from surveys, experiments, or any series of measurements are often so numerous that they are virtually useless, unless they are condensed or reduced into a more suitable form. Sometimes, it may be satisfactory to present data just as they are, and let them speak for themselves; on other occasions, it may be necessary only to group the data and present results in the form of tables or in a graphical form. The summarization and exposition of the different important aspects of the data is commonly called descriptive statistics. This idea includes the condensation of the data in the form of tables, their graphical presentation, and computation of numerical indicators of the central tendency and variability.

There are mainly two main aspects of describing a data set:

1. Summarization and description of the overall pattern of the data by
   a. Presentation of tables and graphs
   b. Examination of the overall shape of the graphical data for important features, including symmetry or departure from it
   c. Scanning graphical data for any unusual observations, which seems to stick out from the major mass of the data
2. Computation of the numerical measures for
   a. A typical or representative value that indicates the center of the data
   b. The amount of spread or variation present in the data

Summarization and description of the data can be done in different ways. For a univariate data, the most popular methods are histogram, bar chart, frequency tables, box plot, or the stem and leaf plots. For bivariate or multivariate data, the useful methods are scatter plots or Chernoff faces. A wonderful exposition of the different exploratory data analysis techniques can be found in Tukey (1977), and for some recent development, see Theus and Urbanek (2008).

A typical or representative value that indicates the center of the data is the average value or the mean of the data. But since the mean is not a very robust estimate and is very much susceptible to the outliers, often, median can be used to represent the center of the data. In case of a symmetric distribution, both mean and median are the same, but in general, they are different. Other than mean or median, trimmed mean or the Windsorized mean can also be used to represent the central value of a data set. The amount of spread or the variation present in a data set can be measured using the standard deviation or the interquartile range.

## 8.3 Probability

The main aim of this section is to introduce the basic concepts of probability theory that are used quite extensively in developing different statistical inference procedures. We will try to provide the basic assumptions needed for the axiomatic development of the probability theory and will present some of the important results that are essential tools for statistical inference. For further study, the readers may refer to some of the classical books in probability theory such as Doob (1953) or Billingsley (1995), and for some recent development and treatment, readers are referred to Athreya and Lahiri (2006).

### 8.3.1 Sample Space and Events

The concept of probability is relevant to experiments that have somewhat uncertain outcomes. These are the situations in which, despite every effort to maintain fixed conditions, some variation of the result in repeated trials of the experiment is unavoidable. In probability, the term "experiment" is not restricted to laboratory experiments but includes any activity that results in the collection of data pertaining to the phenomena that exhibit variation. The domain of probability encompasses all phenomena for which outcomes cannot be exactly predicted in advance. Therefore, an experiment is the process of collecting data relevant to phenomena that exhibits variation in its outcomes. Let us consider the following examples:

*Experiment (a).* Let each of 10 persons taste a cup of instant coffee and a cup of percolated coffee. Report how many people prefer the instant coffee.

*Experiment (b).* Give 10 children a specific dose of multivitamin in addition to their normal diet. Observe the children's height and weight after 12 weeks.

*Experiment (c).* Note the sex of the first 2 new born babies in a particular hospital on a given day.

In all these examples, the experiment is described in terms of what is to be done and what aspect of the result is to be recorded. Although each experimental outcome is unpredictable, we can describe the collection of all possible outcomes.

**Definition**

The collection of all possible distinct outcomes of an experiment is called the sample space of the experiment, and each distinct outcome is called a simple event or an element of the sample space. The sample space is denoted by $\Omega$.

In a given situation, the sample space is presented either by listing all possible results of the experiments, using convenient symbols to identify the results or by making a descriptive statement characterizing the set of possible results. The sample space of the above three experiments can be described as follows:

*Experiment (a).* $\Omega = \{0, 1, \ldots, 10\}$.

*Experiment (b).* Here, the experimental result consists of the measurements of two characteristics, height and weight. Both of these are measured on a continuous scale. Denoting the measurements of gain in height and weight by $x$ and $y$, respectively, the sample space can be described as $\Omega = \{(x, y); x \text{ nonnegative}, y \text{ positive, negative or zero.}\}$

*Experiment (c).* $\Omega = \{\text{BB, BG, GB, GG}\}$, where, for example, BG denotes the birth of a boy first and then followed by a girl. Similarly, the other symbols are also defined.

In our study of probability, we are interested not only in the individual outcomes of $\Omega$ but also in any collection of outcomes of $\Omega$.

**Definition**

An event is any collection of outcomes contained in the sample space $\Omega$. An event is said to be simple, if it consists of exactly one outcome, and compound, if it consists of more than one outcome.

**Definition**

A sample space consisting of either a finite or a countably infinite number of elements is called a discrete sample space. When the sample space includes all the numbers in some interval (finite or infinite) of the real line, it is called continuous sample space.

### 8.3.2 Axioms, Interpretations, and Properties of Probability

Given an experiment and a sample space $\Omega$, the objective of probability is to assign to each event $A$, a number $P(A)$, called probability of the event $A$, which will give a precise measure of the chance that $A$ will occur. To ensure that the probability assignment will be consistent with our intuitive notion of probability, all assignments should satisfy the following axioms (basic properties) of probability:

- Axiom 1: For any event $A$, $0 \leq P(A) \leq 1$.
- Axiom 2: $P(\Omega) = 1$.
- Axiom 3: If $\{A_1, A_2, \ldots\}$ is an infinite collection of mutually exclusive events, then

$$P(A_1 \cup A_2 \cup A_3 \ldots) = \sum_{i=1}^{\infty} P(A_i).$$

Axiom 1 reflects the intuitive notion that the chance of $A$ occurring should be at least zero, so that negative probabilities are not allowed. The sample space is by definition an event that must occur when the experiment performed ($\Omega$) contains all possible outcomes. So, Axiom 2 says that the maximum probability of occurrence is assigned to $\Omega$. The third axiom formalizes the idea that if we wish the probability that at least one of a number of events will occur, and no two of the events can occur simultaneously, then the chance of at least one occurring is the sum of the chances of individual events.

Consider an experiment in which a single coin is tossed once. The sample space is $\Omega = \{H, T\}$. The axioms specify $P(\Omega) = 1$, so to complete the probability assignment, it remains only to determine $P(H)$ and $P(T)$. Since $H$ and $T$ are disjoint events, and $H \cup T = \Omega$, Axiom 3 implies that $1 = P(\Omega) = P(H) + P(T)$. So, $P(T) = 1 - P(H)$. Thus, the only freedom allowed by the axioms in this experiment is the probability assigned to $H$. One possible assignment of probabilities is $P(H) = 0.5$, $P(T) = 0.5$, while another possible assignment is $P(H) = 0.75$, $P(T) = 0.25$. In fact, letting $p$ represent any fixed number between 0 and 1, $P(H) = p$, $P(T) = 1 - p$ is an assignment consistent with the axioms.

### 8.3.3 Borel σ-Field, Random Variables, and Convergence

The basic idea of probability is to define a set function whose domain is a class of subsets of the sample space $\Omega$, whose range is $[0, 1]$, and it satisfies the three axioms mentioned in the previous subsection. If $\Omega$ is the collection of finite number or countable number of points, then it is quite easy to define the probability function always, for the class of all subsets of $\Omega$, so that it satisfies Axioms 1–3. If $\Omega$ is not countable, it is not always possible to define for the class of all subsets of $\Omega$. For example, if $\Omega = \mathbb{R}$, the whole real line, then the probability function (from now onward, we call it as a probability measure) is not possible to define for the class of all subsets of $\Omega$. Therefore, we define a particular class of subsets of $\mathbb{R}$, called Borel σ-field (it will be denoted by $\mathcal{B}$); see Billingsley (1995) for details, on which probability measure can be defined. The triplet $(\Omega, \mathcal{B}, P)$ is called the probability space, while $\Omega$ or $(\Omega, \mathcal{B})$ is called the sample space.

*Random variable:* A real-valued point function $X(\cdot)$ defined on the space $(\Omega, \mathcal{B}, P)$ is called a random variable of the set $\{\omega : X(\omega) \leq x\} \in \mathcal{B}$, for all $x \in \mathbb{R}$.

*Distribution function:* The point function

$$F(x) = P\{\omega : X(\omega) \leq x\} = P(X^{-1}(-\infty, x]),$$

defined on $\mathbb{R}$, is called the distribution function of $X$.

Now, we will define three important concepts of convergence of a sequence of random variables. Suppose $\{X_n\}$ is a sequence of random variables, and $X$ is also a random variable, and all are defined of the same probability space $(\Omega, \mathcal{B}, P)$.

*Convergence in probability or weakly:* The sequence of random variables $\{X_n\}$ is said to converge to $X$ in probability (denoted by $X_n \xrightarrow{p} X$) if for all $\epsilon > 0$,

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

*Almost sure convergence or strongly:* The sequence of random variables $\{X_n\}$ is said to converge to $X$ strongly (denoted by $X_n \xrightarrow{a.e.} X$), if

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

*Convergence in distribution:* The sequence of random variables $\{X_n\}$ is said to converge to $X$ in distribution (denoted by $X_n \xrightarrow{d} X$), if

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

for all $x$, such that $F$ is continuous at $x$. Here, $F_n$ and $F$ denote the distribution functions of $X_n$ and $X$, respectively.

### 8.3.4 Some Important Results

In this subsection, we present some of the most important results of probability theory that have direct relevance in statistical sciences. The books by Chung (1974) or Serfling (1980) are referred for details.

The characteristic function of a random variable $X$ with the distribution function $F(x)$ is defined as follows:

$$\phi_X(t) = E\left(e^{itX}\right) = \int_{-\infty}^{\infty} e^{itx} dF(x), \quad \text{for} \quad t \in \mathbb{R},$$

where $i = \sqrt{-1}$. The characteristic function uniquely defines a distribution function. For example, if $\phi_1(t)$ and $\phi_2(t)$ are the characteristic functions associated with the distribution functions $F_1(x)$ and $F_2(x)$, respectively, and $\phi_1(t) = \phi_2(t)$, for all $t \in \mathbb{R}$, then $F_1(x) = F_2(x)$, for all $x \in \mathbb{R}$.

*Chebyshev's theorem:* If $\{X_n\}$ is a sequence of random variables, such that $E(X_i) = \mu_i$, $V(X_i) = \sigma_i^2$, and they are uncorrelated, then

$$\lim_{n \to \infty} \frac{1}{n^2} \sigma_i^2 = 0 \Rightarrow \left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mu_i\right] \xrightarrow{p} 0.$$

*Khinchine's theorem:* If $\{X_n\}$ is a sequence of independent and identically distributed random variables, such that $E(X_1) = \mu < \infty$, then

$$\lim_{n \to \infty} \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu.$$

*Kolmogorov theorem 1:* If $\{X_n\}$ is a sequence of independent random variables, such that $E(X_i) = \mu_i$, $V(X_i) = \sigma_i^2$, then

$$\sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty \Rightarrow \left[ \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{1}{n} \sum_{i=1}^{n} \mu_i \right] \overset{a.s.}{\to} 0.$$

*Kolmogorov theorem 2:* If $\{X_n\}$ is a sequence of independent and identically distributed random variables, then a necessary and sufficient condition that

$$\frac{1}{n} \sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mu$$

is that $E(X_1) < \infty$, and it is equal to $\mu$.

*Central limit theorem:* If $\{X_n\}$ is a sequence of independent and identically distributed random variables, such that $E(X_1) = \mu$, and $V(X_1) = \sigma^2 < \infty$, then

$$\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \overset{d}{\to} Z.$$

Here, $Z$ is a standard normal random variable with mean zero and variance 1.

**Example 8.1**

Suppose $X_1, X_2, \ldots$ is a sequence of i.i.d. exponential random variable with the following probability density function for $x > 0$:

$$f(x) = \begin{cases} e^{-x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

In this case, $E(X_1) = V(X_1) = 1$. Therefore, by the weak law of large numbers (WLLN) of Khinchine, it immediately follows that

$$\frac{1}{n} \sum_{i=1}^{n} X_i \overset{p}{\to} 1,$$

and by Kolmogorov's strong law of large numbers (SLLN),

$$\frac{1}{n} \sum_{i=1}^{n} X_i \overset{a.e.}{\to} 1.$$

Further, by the central limit theorem (CLT), we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - 1) \overset{d}{\to} Z \sim N(0, 1).$$

## 8.4 Estimation

### 8.4.1 Introduction

The topic of parameter estimation deals with the estimation of some parameters from the data that characterizes the underlying process or phenomenon. For example, one is posed with the data taken repeatedly of the same temperature. These data are not equal, although the underlying *true* temperature was the same. In such a situation, one would like to obtain an estimate of the true temperature from the given data.

We may also be interested in finding the coefficient of resolution of a steel ball from the data on successive heights to which the ball rose. One may be interested in obtaining the face flow speed of vehicles from data on speed and density. All these estimation problems come under the purview of parameter estimation. The question of estimation arises because one always tries to obtain knowledge on the parameters of the population from the information available through the sample. The estimate obtained depends on the sample collected. Further, one could generally obtain more than one sample from a given population, and therefore the estimates of the same parameter could be different from one another. Most of the desirable properties of an estimate are defined keeping in mind the variability of the estimates.

In this discussion on the said topic, we will look into desirable properties of an estimator, and some methods for obtaining estimates. We will also see some examples that will help to clarify some of the salient features of parameter estimation. Finally, we will introduce the ideas of interval estimation, and illustrate its relevance to real-world problems.

### 8.4.2 Desirable Properties

The desirable properties of an estimator are defined keeping in mind that the estimates obtained are random. In the following discussion, $T$ will represent an estimate while $\theta$ will represent the true parameter value of a parameter. The properties that will be discussed are the following:

*Unbiasedness:* The unbiasedness property states that $E(T) = \theta$. The desirability of this property is self-evident. It basically implies that on an average the estimator should be equal to the parameter value.

*Minimum variance:* It is also desirable that any realization of $T$ (i.e., any estimate) may not be far off from the true value. Alternatively stated, it means that the probability of $\theta$ being near to $\theta$ should be high, or as high as possible. This is equivalent to saying that the variance of $T$ should be minimal. An estimator that has the minimum variance in the class of all unbiased estimators is called an efficient estimator.

*Sufficiency:* An estimator is sufficient if it uses all the information about the population parameter, $\theta$, that is available from the sample. For example, the sample median is not a sufficient estimator of the population mean, because median only utilizes the ranking of the sample values and not their relative distance. Sufficiency is important because it is a necessary condition for the minimum variance property (i.e., efficiency).

*Consistency:* The property of consistency demands that an estimate be very close to the true value of the parameter when the estimate is obtained from a large sample. More specifically, if $\lim_{n\to\infty} P(|T - \theta| < \epsilon) = 1$, for any $\epsilon > 0$, however small it might be, the estimator $T$ is said to be a consistent estimator of the parameter $\theta$. It may be noted that if $T$ has a zero bias, and the variance of $T$ tends to zero, then $T$ is a consistent estimator of $\theta$.

*Asymptotic properties:* The asymptotic properties of estimators relate to the behavior of the estimators based on a large sample. Consistency is thus an asymptotic property of an estimator. Other asymptotic properties include asymptotic unbiasedness and asymptotic efficiency.

As the nomenclature suggests, asymptotic unbiasedness refers to the unbiasedness of an estimator based on a large sample. Alternatively, it can be stated as follows:

$$\lim_{n\to\infty} E(T) = \theta.$$

For example, an estimator whose $E(T) = \theta - (1/n)$ is an asymptotically unbiased estimator of $\theta$. For small samples, however, this estimator has a finite negative bias.

Similarly, asymptotic efficiency suggests that an asymptotically efficient estimator is the minimum variance unbiased estimator of $\theta$ for large samples. Asymptotic efficiency may be thought of as the large sample equivalent of best unbiasedness, while asymptotic unbiasedness may be thought of as the large sample equivalent of unbiasedness property.

*Minimum mean square error:* The minimum mean square error (MSE) property states that the estimator $T$ should be such that the quantity MSE defined below is minimum:

$$MSE = E(T - \theta)^2.$$

Alternatively written,

$$MSE = Var(T) + (E(T) - \theta)^2.$$

Intuitively, it is appealing because it looks for an estimator that has small bias (may be zero) and small variance. This property is appealing because it does not constrain an estimator to be unbiased before looking at the variance of the estimator. Thus, the minimum MSE property does not give higher importance to unbiasedness than to variance. Both the factors are considered simultaneously.

*Robustness:* Another desirable property of an estimator is that the estimator should not be very sensitive to the presence of outliers or obviously erroneous points in the data set. Such an estimator is called a robust estimator. The robust property is important because, loosely speaking, it captures the reliability of an estimator. There are different ways in which robustness is quantified. Influence function and breakdown point are two such methods. Influence functions describe the effect of one outlier on the estimator. Breakdown point of an estimator is the proportion of incorrect observations (for example, arbitrarily large observations) an estimator can handle before giving an incorrect (that is arbitrarily large) result.

### 8.4.3 Methods of Estimation

One of the important questions in parameter estimation is, how does one estimate (the method of estimation) the unknown parameters so that the properties of the resulting estimators are in reasonable agreement with the desirable properties? There are many methods that are available in the literature, and needless to say that none of these methods provide estimators that satisfy all the desirable properties. As we will see later, some methods provide good estimators under certain assumptions and others provide good estimators with minor modifications. Although salient, one important aspect of developing a method for estimation that one should bear in mind the amount and complexity of the computation requirement associated with the methodology.

We will elaborate on four different methods, namely, (a) the method of maximum likelihood, (b) the method of least squares, (c) the method of moments, and (d) the method of minimum absolute deviation.

*The method of maximum likelihood:* Suppose $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is a random sample from a population that is characterized by $m$ parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. It is assumed that the population has the probability density function (PDF) or probability mass function (PMF) as $f(x; \boldsymbol{\theta})$. The principle of maximum likelihood estimation consists of choosing as an estimate of $\boldsymbol{\theta}$ a $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ that maximizes the likelihood function, which is defined as follows:

$$L(\boldsymbol{\theta}; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}).$$

Therefore,

$$L(\widehat{\boldsymbol{\theta}}; x_1, \ldots, x_n) = \sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; x_1, \ldots, x_n),$$

or in other words

$$\widehat{\boldsymbol{\theta}} = \text{argmax } L(\boldsymbol{\theta}, x_1, \ldots, x_n).$$

The notation "argmax" means that $L(\boldsymbol{\theta}, x_1, \ldots, x_n)$ achieves the maximum value at $\widehat{\boldsymbol{\theta}}$, and $\widehat{\boldsymbol{\theta}}$ is called the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$.

To motivate the use of the likelihood function, we begin with a simple example, and then provide with a theoretical justification.

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter $\theta$, which has the following probability mass function:

$$p(x) = \begin{cases} \theta^x (1-\theta)^x & x = 0, 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < 1$. Then

$$P(X_1 = x_1, \ldots, X_n = x_n) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n - \sum_{i=1}^{n} x_i},$$

where $x_i$ can be either 0 or 1. This probability, which is the joint probability mass function of $X_1, \ldots, X_n$, as a function of $\theta$, is the likelihood function $L(\theta)$ defined above, that is,

$$L(\theta) = \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n - \sum_{i=1}^{n} x_i}, \quad 0 < \theta < 1.$$

Now we may ask what should be the value of $\theta$ that would maximize the above probability $L(\theta)$ to obtain the specific observed sample $x_1, \ldots, x_n$. The value of $\theta$ that maximizes $L(\theta)$ seems to be a good estimate of $\theta$ as it provides the largest probability for this particular sample. Since

$$\widehat{\theta} = \text{argmax } L(\theta) = \text{argmax } \ln L(\theta),$$

it is often easier to maximize $l(\theta) = \ln L(\theta)$, rather than $L(\theta)$. In this case

$$l(\theta) = \ln L(\theta) = \left( \sum_{i=1}^{n} x_i \right) \ln \theta + \left( n - \sum_{i=1}^{n} x_i \right) \ln(1-\theta),$$

provided $\theta$ is not equal to 0 or 1. So we have

$$\frac{dl(\theta)}{d\theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n - \sum_{i=1}^{n} x_i}{1-\theta} = 0.$$

Therefore,

$$\widehat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Now we provide the theoretical justification to use the maximum likelihood estimator as a reasonable estimator of $\theta$. Suppose $\theta_0$ denotes the true value of $\theta$, then Theorem 8.1 provides a theoretical reason for maximizing the likelihood function. It says that the maximum of $L(\theta)$ asymptotically separates the true model at $\theta_0$ from models at $\theta \neq \theta_0$. We will state the main result without proof. For details, interested readers are referred to Lehmann and Casella (1998).

*Regularity conditions* A1. The PDFs are distinct, that is, for $\theta \neq \theta' \Rightarrow f(x;\theta) \neq f(x;\theta')$.

A2. The PDFs have common support for all $\theta$.

A3. The point $\theta_0$ is an interior point of the parameter space $\Omega$.

Note that the first assumption states that the parameters identify the PDFs. The second assumption implies that the support of the random variables does not depend on the parameter. Now, based on the above assumptions, we have the following important result regarding the existence and uniqueness of the maximum likelihood estimator of $\theta$.

**Theorem 8.1**

Let $\theta_0$ be the true parameter value, then under Assumptions A1–A3

$$\lim_{n\to\infty} P_{\theta_0}\left[L(\theta_0, X_1, \ldots, X_n) > L(\theta, X_1, \ldots, X_n)\right] = 1, \quad \text{for all } \theta \neq \theta_0.$$

Theorem 8.1 states that the likelihood function is asymptotically maximized at the true value $\theta_0$. The following theorem provides the consistency property of the maximum likelihood estimator under some suitable regularity conditions.

**Theorem 8.2**

Let $X_1, \ldots, X_n$ be a random sample from the probability density function $f(x;\theta)$, which satisfies Assumptions A1–A3. Further, it is assumed that $f(x;\theta)$ is differentiable with respect to $\theta \in \Omega$. If $\theta_0$ is the true parameter value, then the likelihood equation

$$\frac{\partial}{\partial\theta} L(\theta) = 0 \Leftrightarrow \frac{\partial}{\partial\theta} l(\theta) = 0$$

has a solution $\widehat{\theta}_n$, such that $\widehat{\theta}_n$ converges to $\theta_0$ in probability.

Finally, we state the asymptotic normality results based on certain regularity conditions. The details of the regularity conditions and the proof can be obtained in Lehmann and Casella (1998).

**Theorem 8.3**

Let $X_1, \ldots, X_n$ be a random sample from the probability density function $f(x;\theta)$, which satisfies the regularity conditions as stated in Lehmann and Casella (1998). Then

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{I}^{-1}).$$

Here, $\xrightarrow{D}$ means converges in distribution, and $\boldsymbol{I}$ is the Fisher information matrix.

Theorem 8.3 can be used for the construction of confidence intervals and also for testing purposes. One point should be mentioned that although the MLE is the most popular estimator, it may not be in the explicit form always. Let us consider the following example:

**Example 8.2**

Suppose $X_1, \ldots, X_n$ are i.i.d. random variables with the following PDF:

$$f(x;\theta) = \frac{e^{-(x-\theta)}}{(1+e^{-(x-\theta)})^2}; \quad -\infty < x < \infty, \quad -\infty < \theta < \infty. \tag{8.1}$$

It may be mentioned that Equation 8.1 is the PDF of a logistic distribution. The logarithm of the likelihood function can be written as

$$l(\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta) = n\theta - \sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} \ln\left(1 + e^{-(x_i-\theta)}\right). \tag{8.2}$$

The MLE of the unknown parameter $\theta$ can be obtained by maximizing Equation 8.2 with respect to the unknown parameter $\theta$. Setting the first partial derivative of Equation 8.2 equals to zero, we obtain

$$\frac{d}{d\theta} l(\theta) = n - 2\sum_{i=1}^{n} \frac{e^{-(x_i-\theta)}}{1+e^{-(x_i-\theta)}} = 0. \tag{8.3}$$

Rearranging Equation 8.3, it becomes

$$\sum_{i=1}^{n} \frac{e^{-(x_i-\theta)}}{1+e^{-(x_i-\theta)}} = \frac{n}{2}. \tag{8.4}$$

The MLE of $\theta$, $\widehat{\theta}$, can be obtained by solving Equation 8.4. Unfortunately, it cannot be obtained in explicit form. It can be shown that the solution exists and it is unique. It can be obtained as follows. The first derivative of the left-hand side of Equation 8.4 is

$$\frac{d}{d\theta}\sum_{i=1}^{n} \frac{e^{-(x_i-\theta)}}{1+e^{-(x_i-\theta)}} = \sum_{i=1}^{n} \frac{e^{-(x_i-\theta)}}{\left(1+e^{-(x_i-\theta)}\right)^2} > 0.$$

Therefore, the left-hand side of Equation 8.4 is a strictly increasing function of $\theta$, and it goes to zero, as $\theta$ goes to $\infty$ or $-\infty$. Hence, Equation 8.4 has a unique solution, and it has to be obtained numerically. The standard numerical analysis method like bisection or Newton's method may be used to compute $\widehat{\theta}$. It can be easily shown that the PDF of the logistic distribution satisfies all the regulatory conditions. Hence, we can conclude that as $n \to \infty$, $\widehat{\theta} \to \theta_0$; here, $\theta_0$ is the true value of the parameter. Moreover,

$$\sqrt{n}(\widehat{\theta} - \theta_0) \overset{D}{\to} N(0, 1.0/I(\theta_0)), \tag{8.5}$$

where

$$I(\theta_0) = -E\left[\frac{d^2}{d\theta^2} \ln f(x; \theta_0)\right] = \int\limits_{-\infty}^{\infty} \frac{e^{-2(x-\theta_0)}}{(1+e^{-(x-\theta_0)})^4} dx = \frac{1}{6}.$$

Therefore, using Equation 8.5, for large sample size, we can obtain an approximate 95% confidence interval of $\theta_0$ as $\left(\widehat{\theta} - 1.96 \times \sqrt{6/n}, \widehat{\theta} + 1.96 \times \sqrt{6/n}\right)$.

Now we will provide some of the numerical methods that can be used to compute the MLEs. In most of the cases, the MLEs have to be obtained by solving a set of nonlinear equations, or solving a multidimensional optimization problem. Some of the standard general-purpose algorithms can be used to compute the MLEs. For example, genetic algorithm of Goldberg (1989), simulated annealing of Kirkpatrick et al. (1983), downhill simplex method of Nelder and Mead (1965), etc. can be used to compute the MLEs of the unknown parameters by maximizing the likelihood function.

Another very important method that can be used very successfully to compute the MLEs of the unknown parameters, particularly if some of the data are missing or censored, is known as the expectation maximization (EM) algorithm introduced by Dempster et al. (1977); see the excellent book by McLachlan and Krishnan (1997) in this respect. The EM algorithm has two steps: (i) E-Step and (ii) M-Step. In E-Step, pseudo-likelihood function has been obtained by replacing the missing values with their corresponding expected values, and M-Step involves maximizing the pseudo-likelihood function. Although, EM algorithm has been used mainly when the complete data are not available, but it has been used in many cases in case of complete sample also by treating it as a missing value problem.

Before describing another very popular estimator, we will mention one very useful property of the MLE, and that is known as invariance property, which may not be true for most of the other estimators. It can be stated as follows. If $\widehat{\theta}$ is the MLE of $\theta$, and $h(\theta)$ is a "nice" function, then $h(\widehat{\theta})$ is the MLE of $h(\theta)$. Moreover, similar to Equation 8.5, in this case, we have

$$\sqrt{n}(g(\widehat{\theta}) - g(\theta_0)) \overset{D}{\to} N(0, (g'(\theta_0))^2/I(\theta_0)). \tag{8.6}$$

Hence, similar to $\theta_0$, using Equation 8.6, an asymptotic confidence interval of $g(\theta_0)$ can also be obtained along the same line.

### 8.4.4 Method of Moment Estimators

The method of moment estimators is the oldest method of finding point estimators. Dating goes back to Karl Pearson in the late 1800. It is very simple to use and most of the time it provides some sort of estimate. In many cases, it may happen that it can be improved upon; however, it is a good place to start with when other methods may be very difficult to implement.

Let $X_1, \ldots, X_n$ be a random sample from a PDF or PMF $f(x|\theta_1, \ldots, \theta_k)$. The method of moment estimators are found by equating the first $k$ sample moments to the corresponding $k$ population moments. The method of moment estimators are obtained by solving the resulting systems of equations simultaneously. To be more precise, for $j = 1, \ldots, k$, define

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i^1, \quad \mu_1 = E(X^1); \\ m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \mu_2 = E(X^2); \\ &\vdots \\ m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad \mu_1 = E(X^k). \end{aligned} \tag{8.7}$$

Usually, the population moment $\mu_j$, will be a function of $\theta_1, \ldots, \theta_k$, say $\mu_j(\theta_1, \ldots, \theta_k)$. Hence, the method of moment estimators $\widetilde{\theta}_1, \ldots, \widetilde{\theta}_k$ of $\theta_1, \ldots, \theta_k$, respectively, can be obtained by solving the following $k$-equations simultaneously:

$$\begin{aligned} m_1 &= \mu_1(\theta_1, \ldots, \theta_k); \\ m_2 &= \mu_2(\theta_1, \ldots, \theta_k); \\ &\vdots \\ m_k &= \mu_k(\theta_1, \ldots, \theta_k). \end{aligned} \tag{8.8}$$

The justification of the method of moment estimators mainly comes from the SLLN, and also from the CLT. Owing to SLLN, under some very mild conditions, it can be shown that the method of moment estimators are always consistent estimators of the corresponding parameters. Further, because of the CLT, asymptotically, the method of moment estimators follow multivariate normal distribution, whose covariance matrix can be easily obtained. For illustrative purposes, we provide a simple example where the method of moment estimators can be obtained explicitly. But it may not be the case always. Most of the times, we need to solve a system of nonlinear equations to compute the method of moment estimators.

**Example 8.3**

Suppose $X_1, \ldots, X_n$ are i.i.d. from a two-parameter exponential distribution with the following PDF for $\theta > 0$, and $\mu \in \mathbb{R}$:

$$f(x; \mu, \theta) = \begin{cases} \frac{1}{\theta} e^{-(1/\theta)(x-\mu)} & \text{if } x > \mu \\ 0 & \text{if } x \leq \mu. \end{cases}$$

In this case, using the same notation as above, we obtain

$$m_1 = \frac{1}{n} X_i, \quad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \mu_1 = \theta + \mu, \quad \mu_2 = (\mu + \theta)^2 + \theta^2.$$

Hence, in this case, the method of moment estimators of θ and μ can be easily obtained as

$$\widetilde{\theta} = \sqrt{m_2 - m_1^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}X_i\right)^2} \quad \text{and}$$

$$\widetilde{\mu} = m_1 - \widetilde{\theta} = \frac{1}{n}X_i - \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}X_i\right)^2}.$$

Note that the method of moment estimators are different from the MLEs. Finally, we will wind up this section with another estimator, which is becoming extremely popular in recent days.

### 8.4.5 Bayes Estimators

The Bayesian approach to statistics is philosophically different from the classical approach that we have just mentioned. First, let us describe the Bayesian approach to statistics. The main difference between the classical approach and the Bayesian approach is the following. In the classical approach, the parameter θ is assumed to be unknown but it is assumed to be a fixed quantity. A random sample is drawn from a population that is characterized by the parameter θ, and based on the random sample, a knowledge about the parameter θ is obtained. On the other hand, in the Bayesian approach, the parameter θ is not assumed to be a fixed quantity, and it is considered to be a quantity whose variation can be described by a probability distribution, known as the prior distribution. This prior distribution is purely a subjective distribution, and it depends on the choice of the experimenter. The most important aspect of the prior distribution is that it is formulated before any sample is observed. A sample is then obtained from the population indexed by the parameter θ, and then the prior distribution is updated based on the information of the present sample. The updated information about the parameter θ is known as the posterior distribution.

If we denote the prior distribution by $\pi(\theta)$ and the sampling distribution by $f(\boldsymbol{x}|\theta)$, then the posterior distribution of θ given the sample $\boldsymbol{x}$ becomes

$$\pi(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta)\pi(\theta)/m(\boldsymbol{x}).$$

Here, $m(\boldsymbol{x})$ is the marginal distribution of $\boldsymbol{x}$, and it can be obtained as

$$m(\boldsymbol{x}) = \int f(\boldsymbol{x}|\theta)\pi(\theta)d\theta.$$

Note that the posterior distribution of θ provides the complete information regarding the unknown parameter θ. The posterior distribution of θ can be used to obtain a point estimate of θ or to construct confidence interval (known as credible interval in the Bayesian terminology) of θ. The most popular point estimate of θ is the posterior mean. It has some other nice interpretation also. Other point estimates like median or mode can also be used in this case. Let us consider the following example to illustrate the Bayesian methodology.

**Example 8.4**

Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with parameter θ, and it has the following PDF:

$$f(x|\theta) = \begin{cases} \theta e^{-\theta x} & \text{if } x > 0 \\ 0 & \text{if } x \le 0. \end{cases}$$

Suppose the prior $\pi(\theta)$ on $\theta$ has a gamma distribution with the known shape and scale parameters as $a > 0$ and $b > 0$, respectively. Therefore, $\pi(\theta)$ for $\theta > 0$ has the following form:

$$\pi(\theta|a,b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}\,e^{-b\theta},$$

and zero otherwise. Therefore, for $\boldsymbol{x} = (x_1,\ldots,x_n)$, the posterior distribution of $\theta$ can be obtained as

$$\pi(\theta|\boldsymbol{x}) = \frac{b^a}{m(\boldsymbol{x})\Gamma(a)}\theta^{n+a-1}\,e^{-\theta(b+\sum_{i=1}^n x_i)} \tag{8.9}$$

for $\theta > 0$, and zero otherwise. Here

$$m(\boldsymbol{x}) = \int_0^\infty \frac{b^a}{\Gamma(a)}\theta^{n+a-1}\,e^{-\theta(b+\sum_{i=1}^n x_i)}d\theta = \frac{b^a}{\Gamma(a)} \times \frac{\Gamma(n+a)}{(b+\sum_{i=1}^n x_i)^{n+a}}.$$

Therefore,

$$\pi(\theta|\boldsymbol{x}) = \frac{(b+\sum_{i=1}^n x_i)^{n+a}}{\Gamma(n+a)}\theta^{n+a-1}e^{-\theta(b+\sum_{i=1}^n x_i)}.$$

Hence, the posterior distribution of $\theta$ becomes a gamma distribution with the shape and scale parameters as $n+a$ and $(b+\sum_{i=1}^n x_i)$, respectively. As we have mentioned before, the posterior distribution of $\theta$ provides complete information about the unknown parameter $\theta$. If we want a point estimate of $\theta$, the posterior mean can be considered as one such estimate and in this case it will be

$$\widehat{\theta}_{Bayes} = \frac{\Gamma(n+a)}{b+\sum_{i=1}^n x_i}.$$

Similarly, an associated confidence $100(1-\alpha)\%$ credible interval of $\theta$, say $(L,U)$, can also be constructed using posterior distribution as follows. Choose $L$ and $U$ such that

$$\int_L^U \pi(\theta|\boldsymbol{x})d\theta = 1-\alpha.$$

It is clear from the above discussions that integration techniques play a significant role in Bayesian inference. Here, we provide some illustration of some of the Monte Carlo techniques used for integration in Bayesian inference. We provide it with an example. Suppose $(X_1,\ldots,X_n)$, a random sample, is drawn from an $N(\theta,\sigma^2)$, where $\sigma^2$ is known. Then, $Y = \bar{X}$ is a sufficient statistic. Consider the Bayes model:

$$\begin{aligned} &Y|\theta \sim N(\theta,\sigma^2/n)\\ &\Theta \sim h(\theta) \propto \exp\{-(\theta-a)/b\}/(1+\exp\{-[(\theta-a)/b]^2\}); \quad -\infty < \theta < \infty. \end{aligned} \tag{8.10}$$

Here, $a$ and $b > 0$ are known hyperparameters. The distribution of $\Theta$ is known as the logistic distribution with parameter $a$ and $b$. The posterior PDF is

$$h(\theta|y) = \frac{\frac{1}{\sqrt{2\pi}\sigma/n}\exp\left\{-\frac{(y-\theta)^2}{2\sigma^2/n}\right\}e^{-(\theta-a)/b}/(1+e^{[-(\theta-a)/b]^2})}{\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma/n}\exp\left\{-\frac{(y-\theta)^2}{2\sigma^2/n}\right\}e^{-(\theta-a)/b}/(1+e^{[-(\theta-a)/b]^2})d\theta}.$$

Based on the squared error loss function, the Bayes estimate of $\theta$ becomes the mean of the posterior distribution. It involves computing two integrations, which cannot be obtained explicitly. In this case, Monte Carlo simulation technique can be used quite effectively to compute the Bayes estimate and the associated credible interval of $\theta$. Consider the following likelihood function as a function of $\theta$:

$$w(\theta) = \frac{1}{\sqrt{2\pi\sigma/n}} \exp\left\{-\frac{(y-\theta)^2}{2\sigma^2/n}\right\}.$$

Therefore, the Bayes estimate can be written as

$$\delta(y) = \frac{\int\limits_{-\infty}^{\infty} \theta w(\theta) b^{-1} e^{-(\theta-a)/b}/(1+e^{[-(\theta-a)/b]^2}) d\theta}{\int\limits_{-\infty}^{\infty} w(\theta) b^{-1} e^{-(\theta-a)/b}/(1+e^{[-(\theta-a)/b]^2}) d\theta} = \frac{E(\Theta w(\Theta))}{E(w(\Theta))},$$

where the expectation is taken with $\Theta$ having a logistic prior distribution. The computation of $\delta(y)$ can be carried out by the simple Monte Carlo technique as follows: Generate independently $\Theta_1, \Theta_2, \ldots, \Theta_M$ from the logistic distribution with PDF (8.10). This generation is straightforward, as the inverse function of the logistic distribution can be expressed in explicit form. Then compute

$$T_M = \frac{M^{-1}\sum_{i=1}^{M} \Theta_i w(\Theta_i)}{M^{-1}\sum_{i=1}^{M} w(\Theta_i)}.$$

By WLLN (Khinchine's theorem), it immediately follows that $T_M \xrightarrow{p} \delta(y)$, as $M \to \infty$. By bootstrapping this sample, the confidence interval of $\delta(y)$ can also be obtained, see, for example, Davison and Hinkley (1997). There are several very good books available on Bayesian theory and methodology. The readers are referred to Gilks et al. (1996) or Ghosh et al. (2006) for an easy reading on different Bayesian methodologies.

## 8.5 Linear and Nonlinear Regression Analysis

One of the most important problems in statistical analysis is to find the relationships, if any, that exist in a set of variables when at least one is random. In a regression problem, typically, one of the variables, usually called the dependent variable, is of particular interest, and it is denoted by $y$. The other variables $x_1, x_2, \ldots, x_k$, usually called explanatory variables or independent variables, are mainly used to predict or explain the behavior of $y$. If the prior experience or the plots of the data suggest some relationship between $y$ and $x_i$, then we would like to express this relationship via some function, $f$, namely

$$y \approx f(x_1, x_2, \ldots, x_k). \tag{8.11}$$

Now, using the functional Equation 8.11, from the given $x_1, x_2, \ldots, x_k$, we might be able to predict $y$. For example, $y$ could be the price of a used car of a certain make, $x_1$, the number of previous owners, $x_2$, the age of the car, and $x_3$ the mileage. As expected, the relationship (8.11) can never be exact, as data will always contain unexplained fluctuations or noise, and some degree of measurements error is usually present.

Explanatory variables can be random or fixed (i.e., controlled). Consider an experiment conducted to measure the yield $(y)$ of wheat at different specified levels of density planting $(x_1)$, and fertilizer application $(x_2)$. In this case, both $x_1$ and $x_2$ are fixed. If at the time of planting, soil pH $(x_3)$ was also measured on each plot, then $x_3$ would be random.

In both linear and nonlinear regression analysis, it is assumed that the mathematical form of the relationship (8.11) is known, except for some unknown constants or coefficients, called parameters, the relationship being determined by a known underlying physical process or governed by some accepted scientific laws. Therefore, mathematically, Equation 8.11 can be written as

$$y \approx f(x_1, x_2, \ldots, x_k, \theta), \tag{8.12}$$

where the function $f$ is entirely known, except for the parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$, which is unknown, and based on the observation it needs to be estimated. In both linear and nonlinear regression analysis, it is often assumed that the noise present is additive in nature. Hence, mathematically, the model can be written in the following form:

$$y = f(x_1, x_2, \ldots, x_k, \theta) + \epsilon, \tag{8.13}$$

where $\epsilon$ is the noise random variable, and $E(\epsilon) = 0$.

In case of linear regression model, the function $f$ is assumed to be linear, and the model has the following form:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \epsilon. \tag{8.14}$$

Here, $x_i$'s can include squares, cross products, higher powers, and even transformations of the original measurements. For example

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_3 + \beta_4 x_2^2 + \epsilon$$

or

$$y = \beta_0 + \beta_1 e^{x_1} + \beta_2 \ln x_2 + \beta_3 \sin x_3 + \epsilon$$

are both linear models. The important requirement is that the expression should be linear in the parameters. On the other hand, if the function $f$ is not linear in the parameters, it is called a nonlinear regression model. For example, the following models:

$$y = \beta_0 + \beta_1 e^{\beta_2 x_1} + \beta_3 e^{\beta_4 x_2} + \epsilon$$

and

$$y = \beta_0 + \beta_1 \sin(\beta_2 x_1) + \beta_3 \cos(\beta_4 x_1) + \epsilon$$

are nonlinear regression models.

As can be observed from the above examples, linear models are very flexible, and so it is often used in the absence of a theoretical model $f$. Nonlinear models tend to be used either when they are suggested by theoretical consideration to build a known nonlinear behavior into a model. Even when a linear approximation works well, a nonlinear model may still be used to retain a clear interpretation of the parameters.

The main aim of this section is to consider both the linear and nonlinear regression models and discuss different inferential issues associated in both of them. It may be mentioned that the linear regression models are very well studied in the statistical literature and they are quite well understood also. On the other hand, the nonlinear regression analysis is much less understood, although it has a huge scope of applications. There are several unanswered questions, and lots of scope for future research.

### 8.5.1 Linear Regression Analysis

It is quite convenient to represent the linear regression model in the following matrix notation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{8.15}$$

here, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ is the vector of $n$ observations, $\boldsymbol{X}$ is the known $n \times p$ design matrix, where $p < n$ and the rank of the matrix $\boldsymbol{X}$ is $p$. It will be denoted by

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{p1} & \ldots & x_{pp} \end{bmatrix},$$

and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_p)^T$ is the noise random vector. For simplicity, we will assume that $\epsilon_i$'s are independent identically distributed normal random variables with mean zero and variance $\sigma^2$. The problem is to provide the statistical inferences of the unknown parameters $\boldsymbol{\beta}$ and $\sigma^2$, based on the observation vector $\boldsymbol{y}$ and the design matrix $\boldsymbol{X}$. It is immediate from the above assumptions that $\boldsymbol{y}$ has $n$-variate normal distribution with the mean vector $\boldsymbol{X}\boldsymbol{\beta}$ and dispersion matrix $\sigma^2\boldsymbol{I}_n$. Here, $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix.

The joint probability density function of $\boldsymbol{y}$, given $\boldsymbol{\beta}$ and $\sigma^2$, is

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2}{2\sigma^2}\right\}, \end{aligned} \tag{8.16}$$

here, for the vector $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)^T$, $||\boldsymbol{a}|| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$. The likelihood function, or more simply, the likelihood $l(\boldsymbol{\beta}, \sigma|\boldsymbol{y})$, for $\boldsymbol{\beta}$ and $\sigma$ is identical in form to the joint probability density (8.16), except that $l(\boldsymbol{\beta}, \sigma|\boldsymbol{y})$ is regarded as a function of the parameters conditional of the observed data, rather than as a function of the responses conditional of the values of the parameters. Suppressing the constant $(2\pi)^{-n/2}$, the likelihood can be written as

$$l(\boldsymbol{\beta}, \sigma|\boldsymbol{y}) \propto \sigma^{-n} \exp\left\{-\frac{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2}{2\sigma^2}\right\}. \tag{8.17}$$

Therefore, the MLEs of the unknown parameters can be obtained by maximizing Equation 8.17 with respect to $\boldsymbol{\beta}$ and $\sigma$. It immediately follows that the likelihood (8.17) is maximized with respect to $\boldsymbol{\beta}$, when the residual sum of squares $S(\boldsymbol{\beta}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2$ is minimum. Thus, the MLE $\widehat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ that minimizes $S(\boldsymbol{\beta})$ can be obtained as

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T\boldsymbol{y}. \tag{8.18}$$

$\widehat{\boldsymbol{\beta}}$ is the least squares estimate of $\boldsymbol{\beta}$ also. In deriving Equation 8.18, it is assumed that the matrix $\boldsymbol{X}$ is of full rank. If the rank of the matrix is less than $p$, then clearly $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ does not exist. In this case, although $\widehat{\boldsymbol{\beta}}$ exists, it is not unique. It is not pursued any further here. Moreover, the MLE of $\sigma^2$

can be obtained as

$$\widehat{\sigma}^2 = \frac{||\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}||^2}{n}. \tag{8.19}$$

For detailed treatments on linear regression models and for their applications, the readers are referred to Arnold (1981) and Rao (2008).

Least squares estimates can also be derived using sampling theory, since the least squares estimator is the minimum variance unbiased estimator of $\boldsymbol{\beta}$, or by using a Bayesian approach with a noninformative prior density on $\boldsymbol{\beta}$ and $\sigma$. Interestingly, all three methods of inference, the likelihood approach, the sampling theory approach, and the Bayesian approach, produce the same point estimate of $\boldsymbol{\beta}$.

However, it is important to realize that the MLE of least squares estimates are appropriate when the model and the error assumptions are correct. Expressed in another way, using the least squares estimates, we assume

1. The expectation function is correct.
2. The response is expectation function plus noise.
3. The noise is independent of the expectation function.
4. Each error component has a normal distribution.
5. Each error component has mean zero.
6. The errors have equal variances.
7. The errors are independently distributed.

When these assumptions appear reasonable, we can go to make further inferences about the least squares estimates.

Least squares estimator or the MLE has a number of desirable properties. For example:

1. The least squares estimator $\widehat{\boldsymbol{\beta}}$ is normally distributed. This mainly follows as the least squares estimator is a linear function of $\boldsymbol{y}$, which in turn is a linear function of $\boldsymbol{\epsilon}$. Since $\boldsymbol{\epsilon}$ is assumed to be normally distributed, $\widehat{\boldsymbol{\beta}}$ is also normally distributed.
2. $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$, that is, $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
3. Var($\widehat{\boldsymbol{\beta}}$) = $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$; that is, the covariance matrix of the least squares estimator depends on the error variances and design matrix $\boldsymbol{X}$.
4. A $100(1-\alpha)\%$ joint confidence set for $\boldsymbol{\beta}$ is the ellipsoid

$$\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T\boldsymbol{X}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) \le ps^2 F_{p,n-p,\alpha}, \tag{8.20}$$

where

$$s^2 = \frac{S(\widehat{\boldsymbol{\beta}})}{n-p}$$

is the residual mean square or an unbiased estimator of $\sigma^2$, and $F_{p,n-p,\alpha}$ is the upper $\alpha$ quantile for Fisher's $F$ distribution with $p$ and $n-p$ degrees of freedom.

5. A $100(1-\alpha)\%$ marginal confidence interval for the parameter $\beta_j$, for $j = 1, 2, \ldots, p$, is

$$\widehat{\beta}_j \pm \text{se}(\widehat{\beta})_j t_{n-p,\alpha/2}, \tag{8.21}$$

where $t_{n-p,\alpha/2}$ is the upper $\alpha/2$ quantile for Student's $t$ distribution with $n-p$ degrees of freedom, and the standard error of the parameter estimator is

$$\text{se}(\widehat{\beta})_j = s\sqrt{\left\{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right\}_{jj}},$$

with $\left\{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right\}_{jj}$ equal to the $j$th diagonal term of the matrix $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$.

6. A $100(1-\alpha)\%$ confidence interval for the expected response at $\boldsymbol{x}_0$ is

$$\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}} \pm t_{n-p,\alpha/2}\sqrt{\boldsymbol{x}_0^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0}. \tag{8.22}$$

The least squares estimators are the most popular method to estimate the unknown parameters of a linear regression model, and it has several desirable properties. For example, it can be obtained in explicit form; in case of i.i.d. normal error distributions, the MLEs and the LSEs coincide, but it has certain disadvantages. For example, LSEs are not robust, that is, even in the presence of a very small number of outliers, the estimators can change drastically, which may not be desirable. Moreover, in the presence of heavy tail errors, the LSEs do not behave well. Owing to this reason, least absolute deviation (LAD) estimator can be used, which is more robust, and behaves very well in the presence of heavy tail error. The LAD estimator of $\boldsymbol{\beta}$ can be obtained by minimizing

$$|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}|, \tag{8.23}$$

with respect to $\boldsymbol{\beta}$, where for $\boldsymbol{a} = (a_1, \ldots, a_n)^T$ and $\boldsymbol{b} = (b_1, \ldots, b_n)^T$, $|\boldsymbol{a} - \boldsymbol{b}| = |a_1 - b_1| + \cdots + |a_n - b_n|$. The estimator $\widetilde{\boldsymbol{\beta}}$, which minimizes Equation 8.23, cannot be obtained in explicit form. It has to be obtained numerically. Several numerical methods are available to solve this problem. One important technique is to convert this problem to a linear programming problem, and then solve the linear programming problem, by some efficient linear programming problem solvers; see, for example, an excellent book by Kennedy and Gentle (1980) in this respect. Regarding the theoretical development of the LAD estimators, see Huber (1981).

Another important aspect in a linear regression problem is to estimate the number of predictors. It is a fairly difficult problem, and it can be treated as a model selection problem. Classically, the problem was solved by using stepwise regression method, but recently, different information theoretic criteria such as Akaike information criterion (AIC) or Bayesian information criterion (BIC) have been used to solve this problem. Recently, the least absolute shrinkage and selection operator (LASSO) method proposed by Tibshirani (1996) has received considerable amount of attention in the last one decade.

#### 8.5.1.1 Bayesian Inference

Before closing this section, we will briefly discuss the Bayesian inference of the linear regression model. The Bayesian marginal posterior density for $\boldsymbol{\beta}$, assuming a noninformative prior density for $\boldsymbol{\beta}$ and $\sigma$ of the form,

$$p(\boldsymbol{\beta}, \sigma) \propto \sigma^{-1} \tag{8.24}$$

is

$$p(\beta|\sigma) \propto \left\{1 + \frac{\left(\beta - \widehat{\beta}\right)^T \boldsymbol{X}^T \boldsymbol{X} \left(\beta - \widehat{\beta}\right)}{\nu s^2}\right\}^{-n/2}. \tag{8.25}$$

It is in the form of a $p$-variate Student's $t$ density with the location parameter $\widehat{\boldsymbol{\beta}}$, scaling matrix $s^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, and $\nu = n - p$ degrees of freedom.

Furthermore, the marginal posterior density for a single parameter $\beta_j$, say, is a univariate Student's $t$ density with location parameter $\widehat{\beta}_j$, scale parameter $s^2\left\{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right\}_{jj}$, and the degrees of freedom $n - p$. The marginal posterior density for the mean of $y$ at $\boldsymbol{x}_0$ is a univariate Student's $t$ density with the location parameter $\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}}$, scale parameter $s^2\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0$, and the degrees of freedom $n - p$.

A $100(1-\alpha)\%$ highest posterior density (HPD) region of content is defined as a region $R$ in the parameter space such that $P(\boldsymbol{\beta} \in R) = 1 - \alpha$, and for $\boldsymbol{\beta}_1 \in R$, and $\boldsymbol{\beta}_2 \notin R$, $P(\boldsymbol{\beta}_1 \in R) \geq P(\boldsymbol{\beta}_2 \in R)$. For linear models with a noninformative prior, an HPD region is therefore given by the ellipsoid defined in Equation 8.20. Similarly, the marginal HPD regions for $\beta_j$ and $\boldsymbol{x}_0^T\boldsymbol{\beta}$ are numerically identical to the sampling theory regions (8.21) and (8.22), respectively.

**Example 8.5**

We consider the data of the maximum January temperature (in degrees Fahrenheit) for 62 cities in the United States (from 1931 to 1960) along with their latitude (degrees), longitude (degrees), and altitude (feet). The data have been taken from Mosteller and Tukey (1977). We want to relate the maximum January temperature with the other three variables. We write the model in the following form:

$$\textit{Max Temp} = \beta_0 + \beta_1 \times \textit{Latitude} + \beta_2 \times \textit{Longitude} + \beta_3 \times \textit{Altitude} + \epsilon.$$

The following summary measures are obtained for the design matrix $\boldsymbol{X}$:

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 62.0 & 2365.0 & 5674.0 & 56{,}012.0 \\ 2365.0 & 92{,}955.0 & 217{,}285.0 & 2{,}244{,}586.0 \\ 5674.0 & 217{,}285.0 & 538{,}752.0 & 5{,}685{,}654.0 \\ 56{,}012.0 & 2{,}244{,}586.0 & 5{,}685{,}654.0 & 1.772 \times 10^8 \end{bmatrix},$$

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{bmatrix} 94{,}883.1914 & -1342.5011 & -485.0209 & 2.5756 \\ -1342.5011 & 37.8582 & -0.8276 & -0.0286 \\ -485.0209 & -0.8276 & 5.8951 & -0.0254 \\ 2.5756 & -0.0286 & -0.0254 & 0.0009 \end{bmatrix},$$

$$\boldsymbol{X}^T\boldsymbol{y} = (2739.0, 99{,}168.0, 252{,}007.0, 2{,}158{,}463.0)^T.$$

It gives

$$\widehat{\boldsymbol{\beta}} = (100.8260, -1.9315, 0.2033, -0.0017)^T \quad \text{and} \quad s = 6.05185.$$

Therefore, based on the normality assumption on the error random variables, and the priors discussed above, $\widehat{\boldsymbol{\beta}}$ can be taken as the least squares estimators and Bayes estimators of $\boldsymbol{\beta}$. Further, 100

$(1-\alpha)\%$ credible interval for $\beta$ can be obtained as

$$\left\{\beta : (\beta - \widehat{\beta})^T X^T X(\beta - \widehat{\beta}) \le ps^2 F_{p,n-p}(\alpha)\right\}.$$

### 8.5.2 Nonlinear Regression Analysis

The basic problem in the nonlinear regression analysis can be expressed as follows. Suppose that we have $n$ observations $\{(\boldsymbol{x}_i, y_i); i = 1, 2, \ldots, n\}$ from a nonlinear model with a known functional relationship $f$. Thus

$$y_i = f(\boldsymbol{x}_i, \theta^*) + \epsilon_i; \quad i = 1, 2 \ldots, n, \tag{8.26}$$

where $\epsilon_i$'s are assumed to be independent and identically distributed normal random variables with mean zero, and variance $\sigma^2$, $\boldsymbol{x}_i$ is a $k \times 1$ vector, and the true value $\theta^*$ of $\theta$ is known to belong to $\Theta$, a subset of $\mathcal{R}^p$. The problem is to estimate the unknown parameter $\theta^*$, based on the above observations $\{(\boldsymbol{x}_i, y_i); i = 1, 2, \ldots, n\}$.

Among the different methods, the most popular one is the least squares estimator, denoted by $\widehat{\theta}$, and it can be obtained by minimizing the error sum of squares $S(\theta)$, where

$$S(\theta) = \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i, \theta))^2, \tag{8.27}$$

over $\theta \in \Theta$. Unlike least squares estimator, a simple analytical solution of $\widehat{\theta}$ does not exist, and it has to be obtained numerically. Several numerical methods are available, which can be used to compute $\widehat{\theta}$, but all the methods are iterative in nature. Hence, each iterative method requires some kind of initial guesses to start the process. This is one of the major problems in computing the least squares estimator $\widehat{\theta}$. The least squares surface $S(\theta)$ may have several local minima; hence, a very careful choice of the initial guess is required to compute $\widehat{\theta}$.

For completeness purposes, we provide one numerical method that can be used to compute $\widehat{\theta}$, but it should be mentioned that it may not be the best method in all possible cases. We use the following notation: $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T$, $f_i(\theta) = \boldsymbol{f}(\boldsymbol{x}_i, \theta)$, for $i = 1, \ldots, n$, $\boldsymbol{f}(\theta) = (f_1(\theta), f_2(\theta), \ldots, f_n(\theta))^T$. The main idea of the proposed method is to approximate the nonlinear surface $\boldsymbol{f}(\theta)$ near $\widehat{\theta}$ by a linear surface as follows. Suppose $\theta^{(a)}$ is an approximation to the least squares estimate $\widehat{\theta}$. For $\theta$ close to $\widehat{\theta}$, by using Taylor series approximation, $f(\theta)$ can be written as follows:

$$\boldsymbol{f}(\theta) \approx \boldsymbol{f}(\theta^{(a)}) + \boldsymbol{F}_{\bullet}{}^{(a)}(\theta - \theta^{(a)}), \tag{8.28}$$

here, $\boldsymbol{F}_{\bullet}{}^{(a)} = \boldsymbol{F}_{\bullet}(\theta^{(a)})$, and

$$\boldsymbol{F}_{\bullet}(\theta) = \begin{bmatrix} \dfrac{\partial f_1(\theta)}{\partial \theta_1} & \cdots & \dfrac{\partial f_1(\theta)}{\partial \theta_p} \\ \cdots & \ddots & \cdots \\ \dfrac{\partial f_n(\theta)}{\partial \theta_1} & \cdots & \dfrac{\partial f_n(\theta)}{\partial \theta_p} \end{bmatrix}.$$

Applying this to the residual vector

$$\boldsymbol{r}(\theta) = \boldsymbol{y} - \boldsymbol{f}(\theta) \approx \boldsymbol{r}(\theta^{(a)}) - \boldsymbol{F}_{\bullet}{}^{(a)}(\theta - \theta^{(a)}),$$

in $S(\theta) = \boldsymbol{r}^T(\theta)\boldsymbol{r}(\theta)$, leads to

$$S(\theta) \approx \boldsymbol{r}^T(\theta^{(a)})\boldsymbol{r}(\theta^{(a)}) - 2\boldsymbol{r}^T(\theta^{(a)})\boldsymbol{F}_{\bullet}{}^{(a)}(\theta - \theta^{(a)}) + (\theta - \theta^{(a)})^T \boldsymbol{F}_{\bullet}{}^{(a)T}\boldsymbol{F}_{\bullet}{}^{(a)}(\theta - \theta^{(a)}). \tag{8.29}$$

The right-hand side of Equation 8.29 is minimized with respect to $\theta$, when

$$\theta - \theta^{(a)} = \left(\boldsymbol{F}_{\bullet}{}^{(a)T}\boldsymbol{F}_{\bullet}{}^{(a)}\right)^{-1}\boldsymbol{F}_{\bullet}{}^{(a)T}\boldsymbol{r}(\theta^{(a)}) = \delta^{(a)}.$$

This suggests that given a current approximation $\theta^{(a)}$, the next approximation can be obtained as

$$\theta^{(a+1)} = \theta^{(a)} + \delta^{(a)}.$$

This provides an iterative scheme for computing $\widehat{\theta}$. This particular iterative scheme is known as Gauss–Newton method. It forms the basis of a number of least squares algorithm used in the literature. The Gauss–Newton algorithm is convergent, that is, $\theta^{(a)} \rightarrow \widehat{\theta}$, as $a \rightarrow \infty$, provided that $\theta^{(1)}$ is close to $\theta^*$, and $n$ is large enough.

Finally, we will conclude this section to provide some basic properties of the least squares estimators without any formal proof. Under certain regularity conditions on the function $f$, if the error random variables are i.i.d. normal random variables with mean zero and variance $\sigma^2$, for large $n$, we have the following results:

1. $(\widehat{\theta} - \theta^*) \sim N_p(\boldsymbol{0}, \sigma^2\boldsymbol{C})$, where $\boldsymbol{C} = \boldsymbol{F}_{\bullet}{}^T(\theta^*)\boldsymbol{F}_{\bullet}(\theta^*)$.
2. $S(\widehat{\theta})/\sigma^2 \sim \chi^2_{n-p}$.
3. $\dfrac{(S(\theta^*) - S(\widehat{\theta}))/p}{S(\widehat{\theta})/(n-p)} \sim F_{p,n-p}$.

Here, $\chi^2_{n-p}$ and $F_{p,n-p}$ denote the chi square distribution with $n - p$ degrees of freedom, and $F$ distribution with $p$ and $n - p$ degrees of freedom, respectively. The result (1) provides the consistency and asymptotic normality properties of the least squares estimators; moreover, (2) and (3) can be used for the construction of the confidence interval or confidence set for $\sigma^2$ and $\theta$.

**Example 8.6**

The data set has been obtained from Osborne (1972), and it represents the concentration of a chemical during the different time point of an experiment. We fit the following nonlinear model:

$$y_t = \alpha_0 + \alpha_1 e^{\beta_1 t} + \alpha_2 e^{\beta_2 t} + \epsilon_t.$$

We start with the initial guesses as $\alpha_0 = 0.5$, $\alpha_1 = 1.5$, $\alpha_2 = -1.0$, $\beta_1 = -0.01$, and $\beta_2 = -0.02$. Using Newton–Raphson method, we finally obtained the least squares estimators as

$$\widehat{\alpha}_0 = 0.3754, \quad \widehat{\alpha}_1 = 1.9358, \quad \widehat{\alpha}_2 = -1.4647, \quad \widehat{\beta}_1 = -0.01287, \quad \widehat{\beta}_2 = -0.02212.$$

## 8.6 Introduction to Multivariate Analysis

Multivariate analysis (MVA) is the study based on the statistical principle of multivariate statistics, and involves the observation and analysis of more than one statistical outcome variable at a time. To motivate our readers, we present three different examples of MVA below.

**Example 8.7**

Consider as a dietician in the hospital you are interested to study the physical features and find out the relevant parameters, such as body density (BD), body mass index (BMI), etc., of patients who undergo treatment in the hospital. Your job is to decide on the right diet plan based on the data/information such as percent body fat, age (years), weight (kg), height (cm), etc. of the patients. For the study, you use the past data (http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_ BMI_Regression), which consists of a sample set of 252 patients, as this enables you to do a detailed study/analysis of the different characteristics, such as body fat index, height, and weight, using a three-dimensional (3-D) scatter plot, as illustrated in Figure 8.1.

**Example 8.8**

As the next example, assume you are the real estate agent in the state of California, USA, and your job is to forecast the median house value (MHV). To facilitate better forecasting, you have with you 20,640 data points consisting of information such as MHV, median income (MI), housing median age (HMA), total rooms (TR), total bedrooms (TB), population (P), households (H), etc. Information about the data set can be obtained in the paper by Kelley and Barry (1997). If one fits the multiple linear regression (MLR) model (as used by the authors) to this data, one obtains the ordinary least square (OLS) regression coefficient vector, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0 = 11.4939,\ \hat{\beta}_1 = 0.4790,\ \hat{\beta}_2 = -0.0166,\ \hat{\beta}_3 = -0.0002,\ \hat{\beta}_4 = 0.1570,\ \hat{\beta}_5 = -0.8582,\ \hat{\beta}_6 = 0.8043,\ \hat{\beta}_7 = -0.4077,\ \hat{\beta}_8 = 0.0477)$, using which we can forecast the 20,641th MHV as $\widehat{log_e(MHV)}_{20,641} = \hat{\beta}_0 + \hat{\beta}_1 \times MI_{20,641} + \hat{\beta}_2 \times MI^2_{20,641} + \hat{\beta}_3 \times MI^3_{20,641} + \hat{\beta}_4 \times log_e(MA_{20,641}) + \hat{\beta}_5 \times log_e(TR_{20,641}/P_{20,641}) + \hat{\beta}_6 \times log_e(TB_{20,641}/P_{20,641}) + \hat{\beta}_7 \times log_e(P_{20,641}/H_{20,641}) + \hat{\beta}_8 \times log_e(H_{20,641})$. For example, if one wants to forecast the 20,621th reading, which is 11.5129255, then the forecasted value is 12.3302108, which results in an error of 0.8172853.

**Example 8.9**

As a third example, consider that Professor Manisha Kumari, a faculty member in the Finance Group at the Indian Institute of Management, Indore, India, is interested to study the change

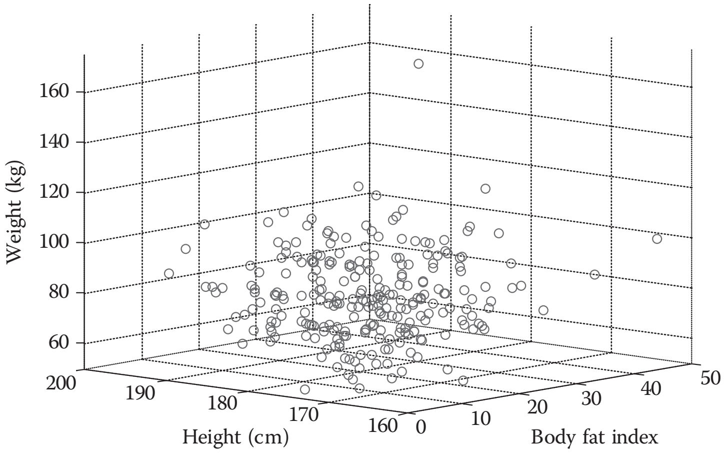


**FIGURE 8.1**
A three-dimensional scatter plot of body fat index, height, and weight of 252 patients, Example 8.7.

in the prices of seven stocks, namely, Bajaj Auto, Maruti Suzuki Indian Limited, Tata Motors, Steel Authority of India, Tata Steel, Infosys Limited, and Tata Consultancy Services Limited, for the time period January 1, 2014 to December 31, 2014. She utilizes the prices of these seven stocks from National Stock Exchange (NSE), which is available at http://in.finance.yahoo.com or http://www.nse-india.com. A closer look convinces her that the price for the first three scripts (Bajaj Auto [#1], Maruti Suzuki Indian Limited [#2], and Tata Motors [#3]), the next two (Steel Authority of India [#4] and Tata Steel [#5]), and the last two (Infosys Limited [#6] and Tata Consultancy Services Limited [#7]) moves in tandem as separate groups as they are from the automobile, steel, and information technology sectors, respectively. Her surmise is valid as the companies that are in the same sector tend to vary together as economic conditions change and this fact is also substantiated by the factor analysis (FA) performed by her (Figure 8.2).

In all these three examples, what is important to note is the fact that given a multidimensional data set, $\boldsymbol{X}_{n\times p}$, of size $(n \times p)$, the users, be it the dietician, the real estate agent, or the faculty member, are all interested to draw some meaningful conclusions from this data set, $\boldsymbol{X}_{n\times p}$. The study of multivariate statistics leads us to analyze multivariate distributions. As is apparent, such studies are of prime importance in many areas of our practical life. It is interesting to note that Francis Galton (1822–1911) may be credited as the first person who worked in the area of multivariate statistical analysis. In his work *Natural Inheritance* (1889), the author summarized the ideas of regression considering bivariate normal distribution. Other notable early researchers whose contributions in the area of multivariate statistics are worth mentioning are Theodore Wilbur Anderson (1918–), Steven F. Arnold (1944–2014), Debabrata Basu (1924–2001), Morris L. Eaton (1939–), Ronald Aylmer Fisher (1890–1962), Narayan Chandra Giri (1928–2006), Ramanathan Gnanadesikan (1932), Harold Hotelling (1895–1973), Norman Lloyd Johnson (1917–2004), Maurice George Kendall (1907–1983), C. G. Khatri (1931–1989), Samuel Kotz (1930–2010), Paruchuri R. Krishnaiah (1932–1987), Anant M. Kshirsagar (1931), Prasanta Chandra Mahalanobis (1893–1972), Calyampudi Radhakrishna Rao (1920–), Samarendra Nath Roy (1906–1964), George Arthur Frederick Seber (1938–), Samuel Stanley Wilks (1906–1964), and many others. Thus, the study of the body of methodologies to investigate simultaneous measurements on many variables is termed as MVA.

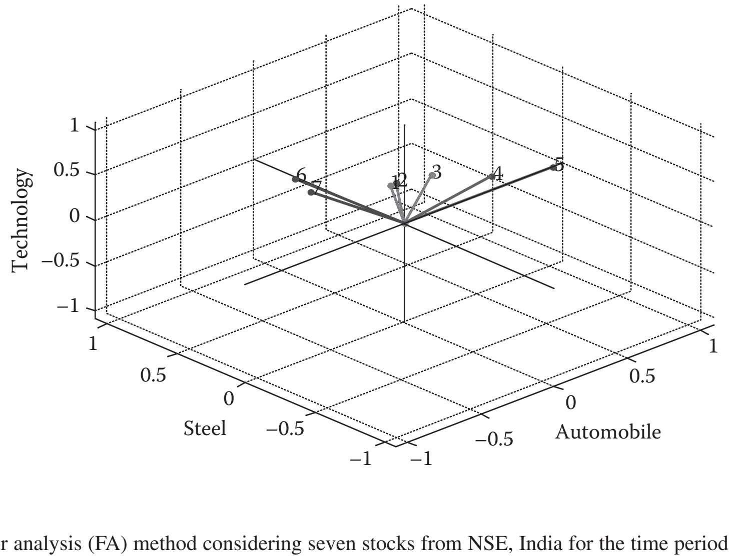


**FIGURE 8.2**
Illustration of factor analysis (FA) method considering seven stocks from NSE, India for the time period January 1, 2014 to December 31, 2014, Example 8.9.

While studying multivariate methods, the objectives that are of prime importance are: data reduction or structural simplification, sorting and grouping, investigation of the dependence among variables, prediction and hypothesis construction, and subsequent testing of the same.

To start with, let us define $\boldsymbol{X}_{n\times p} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ or $(X_{i,j})$, $i = 1, \ldots, n$ and $j = 1, \ldots, p$ as an $(n \times p)$-dimensional matrix of random variables, where $n$ signifies the number of readings and $p$ signifies the dimension, corresponding to different factors in a random variable that are of interest to us. A few important definitions that are useful to understand MVA are:

1. Mean value vector: $\boldsymbol{\mu}_{p\times 1} = (\mu_1, \ldots, \mu_p)'$, while the sample counterpart is $\bar{\boldsymbol{X}}_{p\times 1} = (\bar{X}_1, \ldots, \bar{X}_p)'$.
2. Variance–covariance matrix:

$$\boldsymbol{\Sigma}_{p\times p} = \begin{pmatrix} \sigma_{1,1} & \cdots & \sigma_{1,p} \\ \vdots & \ddots & \vdots \\ \sigma_{p,1} & \cdots & \sigma_{p,p} \end{pmatrix},$$

   while the sample counterpart is

$$\boldsymbol{S}_{p\times p} = \begin{pmatrix} s_{1,1} & \cdots & s_{1,p} \\ \vdots & \iddots & \vdots \\ s_{p,1} & \cdots & s_{p,p} \end{pmatrix}.$$

3. Correlation coefficient matrix:

$$\boldsymbol{\rho}_{p\times p} = \begin{pmatrix} 1 & \cdots & \rho_{1,p} \\ \vdots & \ddots & \vdots \\ \rho_{p,1} & \cdots & 1 \end{pmatrix},$$

   while the sample counterpart is

$$\boldsymbol{R}_{p\times p} = \begin{pmatrix} 1 & \cdots & r_{1,p} \\ \vdots & \ddots & \vdots \\ r_{p,1} & \cdots & 1 \end{pmatrix}.$$

4. Mean: $E(X_j) = \mu_j = \sum_{\forall x_j} x_j Pr(X_j = x_j)$, or $E(X_j) = \mu_j = \int_{x_{j,\min}}^{x_{j,\max}} x_j f(x_j) dx_j = \int_{x_{j,\min}}^{x_{j,\max}} x_j dF_{X_j}(x_j)$, while the sample counterpart is $\bar{X}_j = (1/n) \sum_{i=1}^{n} X_{i,j}$, for $j = 1, \ldots, p$.
5. Covariance: $Covar\left(X_{j_1}, X_{j_2}\right) = E\left[\left\{X_{j_1} - E\left(X_{j_1}\right)\right\}\left\{X_{j_2} - E\left(X_{j_2}\right)\right\}\right] = \sigma_{j_1,j_2} = \sum_{\forall x_{j_1}, x_{j_2}} \left\{X_{j_1} - E\left(X_{j_1}\right)\right\}\left\{X_{j_2} - E\left(X_{j_2}\right)\right\} Pr\left(X_{j_1} = x_{j_1}, X_{j_2} = x_{j_2}\right)$, or $Covar\left(X_{j_1}, X_{j_2}\right) = E\left[\left\{X_{j_1} - E\left(X_{j_1}\right)\right\}\left\{X_{j_2} - E\left(X_{j_2}\right)\right\}\right] = \sigma_{j_1,j_2} = \int_{x_{j_2,\min}}^{x_{j_2,\max}} \int_{x_{j_1,\min}}^{x_{j_1,\max}} \left\{X_{j_1} - E\left(X_{j_1}\right)\right\} \left\{X_{j_2} - E\left(X_{j_2}\right)\right\} f\left(x_{j_1}, x_{j_2}\right) dx_{j_1}\, dx_{j_2} = \int_{x_{j_2,\min}}^{x_{j_2,\max}} \int_{x_{j_1,\min}}^{x_{j_1,\max}} \left\{X_{j_1} - E\left(X_{j_1}\right)\right\}\left\{X_{j_2} - E\left(X_{j_2}\right)\right\} dF_{X_{j_1},X_{j_2}}\left(x_{j_1}, x_{j_2}\right)$, while the sample counterpart is $s_{j_1,j_2} = (1/(n-1)) \sum_{i=1}^{n} \left(X_{i,j_1} - \bar{X}_{j_1}\right) \left(X_{i,j_2} - \bar{X}_{j_2}\right)$, for $j_1$, $j_2 = 1, \ldots, p$.

6. Correlation coefficient: $corr\left(X_{j_1}, X_{j_2}\right) = \rho_{j_1,j_2} = Covar\left(X_{j_1}, X_{j_2}\right) / \sqrt{Var\left(X_{j_1}\right)} \sqrt{Var\left(X_{j_2}\right)}$, while the sample counterpart is $r_{j_1,j_2} = \left(s_{j_1,j_2} / \sqrt{s_{j_1,j_1}} \sqrt{s_{j_2,j_2}}\right)$, for $j_1$, $j_2 = 1, \ldots, p$.
7. Co-skewness:

$$\begin{aligned} & E\left[\left\{X_{j_1} - E\left(X_{j_1}\right)\right\}\left\{X_{j_2} - E\left(X_{j_2}\right)\right\}\left\{X_{j_3} - E\left(X_{j_3}\right)\right\}\right] \\ & \quad = \sum_{\forall x_{j_1}, x_{j_2}, x_{j_3}} \left\{X_{j_1} - E\left(X_{j_1}\right)\right\}\left\{X_{j_2} - E\left(X_{j_2}\right)\right\}\left\{X_{j_3} - E\left(X_{j_3}\right)\right\} \\ & \qquad \times Pr\left(X_{j_1} = x_{j_1}, X_{j_2} = x_{j_2}, X_{j_3} = x_{j_3}\right) \end{aligned}$$

or

$$\begin{aligned} & \int_{x_{j_3,\min}}^{x_{j_3,\max}} \int_{x_{j_2,\min}}^{x_{j_2,\max}} \int_{x_{j_1,\min}}^{x_{j_1,\max}} \left\{X_{j_1} - E\left(X_{j_1}\right)\right\}\left\{X_{j_2} - E\left(X_{j_2}\right)\right\}\left\{X_{j_3} - E\left(X_{j_3}\right)\right\} \\ & \qquad \times f\left(x_{j_1}, x_{j_2}, x_{j_3}\right) dx_{j_1}\, dx_{j_2}\, dx_{j_3}, \end{aligned}$$

for $j_1, j_2, j_3 = 1, \ldots, p$.

**Note:** Co-skewness is related to skewness as covariance is related to variance.

8. Skew relation:

$$\frac{E\left[\left\{X_{j_1} - E\left(X_{j_1}\right)\right\}\left\{X_{j_2} - E\left(X_{j_2}\right)\right\}\left\{X_{j_3} - E\left(X_{j_3}\right)\right\}\right]}{\sqrt{E\left\{X_{j_1} - E\left(X_{j_1}\right)\right\}^2} \sqrt{E\left\{X_{j_2} - E\left(X_{j_2}\right)\right\}^2} \sqrt{E\left\{X_{j_3} - E\left(X_{j_3}\right)\right\}^2}},$$

for $j_1, j_2, j_3 = 1, \ldots, p$.

9. Co-kurtosis:

$$\begin{aligned} & E\left[\left\{X_{j_1} - E\left(X_{j_1}\right)\right\}\left\{X_{j_2} - E\left(X_{j_2}\right)\right\}\left\{X_{j_3} - E\left(X_{j_3}\right)\right\}\left\{X_{j_4} - E\left(X_{j_4}\right)\right\}\right] \\ & \quad = \sum_{\forall x_{j_1}, x_{j_2}, x_{j_3}, x_{j_4}} \left\{X_{j_1} - E\left(X_{j_1}\right)\right\}\left\{X_{j_2} - E\left(X_{j_2}\right)\right\}\left\{X_{j_3} - E\left(X_{j_3}\right)\right\}\left\{X_{j_4} - E\left(X_{j_4}\right)\right\} \\ & \qquad \times \Pr\left(X_{j_1} = x_{j_1}, X_{j_2} = x_{j_2}, X_{j_3} = x_{j_3}, X_{j_4} = x_{j_4}\right) \end{aligned}$$

or

$$\begin{aligned} & \int_{x_{j_4,\min}}^{x_{j_4,\max}} \int_{x_{j_3,\min}}^{x_{j_3,\max}} \int_{x_{j_2,\min}}^{x_{j_2,\max}} \int_{x_{j_1,\min}}^{x_{j_1,\max}} \left\{X_{j_1} - E\left(X_{j_1}\right)\right\}\left\{X_{j_2} - E\left(X_{j_2}\right)\right\}\left\{X_{j_3} - E\left(X_{j_3}\right)\right\} \\ & \qquad \times \left\{X_{j_4} - E\left(X_{j_4}\right)\right\} f\left(x_{j_1}, x_{j_2}, x_{j_3}, x_{j_4}\right) dx_{j_1}\, dx_{j_2}\, dx_{j_3}\, dx_{j_4}, \end{aligned}$$

for $j_1, j_2, j_3, j_4 = 1, \ldots, p$.

**Note:** Co-kurtosis is related to kurtosis as covariance is related to variance.

10. Kurtic relation:

$$\frac{E\left[\left\{X_{j_1}-E\left(X_{j_1}\right)\right\}\left\{X_{j_2}-E\left(X_{j_2}\right)\right\}\left\{X_{j_3}-E\left(X_{j_3}\right)\right\}\left\{X_{j_4}-E\left(X_{j_4}\right)\right\}\right]}{\sqrt{E\left\{X_{j_1}-E\left(X_{j_1}\right)\right\}^2}\sqrt{E\left\{X_{j_2}-E\left(X_{j_2}\right)\right\}^2}\sqrt{E\left\{X_{j_3}-E\left(X_{j_3}\right)\right\}^2}\sqrt{E\left\{X_{j_4}-E\left(X_{j_4}\right)\right\}^2}},$$

for $j_1, j_2, j_3 j_4 = 1, \ldots, p$.

To represent the multivariate data, different graphical techniques can also be used, some of which are: scatter diagram/scatter plot/marginal dot diagram, multiple scatter plot, box plot, 3-D scatter plot, linked scatter plot, rotated plot, growth plot, Chernoff faces, stars, etc. Another important concept that is used in the study of multivariate statistics is the idea of distance measure. A few examples are: Euclidean distance, Bhattacharyya distance, Mahalanobis distance, Pitman closeness criterion, Bregman divergence, Kullback–Leibler distance, Hellinger distance, Chernoff bound, Rényi entropy, and Cook's distance. An interested reader can refer many good references for a better understanding of MVA, a few examples of which are: Anderson (2003), Arnold (1981), Bock (1975), Cooley and Lohnes (1971), Dillon and Goldstein (1984), Eaton (1983), Everitt and Dunn (2001), Giri (2004), Gnanadesikan (2011), Hair et al. (2005), Härdle and Simar (2007), Jobson (1991), Johnson and Wichern (2002), Kendall (1980), Kotz et al. (2000), Kshirsagar (1972), Mardia et al. (1979), Morrison (1990), Muirhead (2005), Press (1982), Rao (2008), Roy (1957), Roy et al. (1971), Seber (2004), Srivastava and Khatri (1979), Takeuchi et al. (1982), and Tatsuoka (1988). Sen (1986) gives a good review of textbooks, papers, monographs, and other related materials in the area of multivariate statistics.

For the interest of the readers, we consider a few multivariate distributions such as multinomial distribution, multivariate normal distribution (MND), multivariate Student $t$-distribution, Wishart distribution, and multivariate extreme value distribution (MEVD) before discussing copula theory. After that, we cover different multivariate techniques that are widely used. One may note that other multivariate distributions such as Dirichlet distribution, Hotelling distribution, multivariate gamma distribution, multivariate beta distribution, multivariate exponential distribution, etc. are not considered due to paucity of space. Moreover, the distributions discussed here are based on their general relevance and practicality.

## 8.7 Joint and Marginal Distribution

The joint distribution of $(X_1, \ldots, X_p)$ may be expressed as $F_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = Pr(X_1 \leq x_1, \ldots, X_p \leq x_p)$. If one thinks from the marginal distribution point of view, then $F_{X_1,\ldots,X_p}(x_1, \ldots, x_p)$ consists of $(2^p - 2)$ number of marginal distributions of which $\binom{p}{1}$ are univariate, $\binom{p}{2}$ are bivariate,..., and finally $\binom{p}{p-1}$ are $(p-1)$ variate. If $X_1, \ldots, X_p$ are pairwise independent, then the joint distribution function $F_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = F_{X_1}(x_1) \times \cdots \times F_{X_p}(x_p)$, where $F_{X_j}(x_j)$ is the corresponding marginal distribution of $X_j$, $j = 1, \ldots, p$. In case $F_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = F_{X_1,\ldots,X_{j_1}}\left(x_1, \ldots, x_{j_1}\right) \times F_{X_{j_1+1},\ldots,X_{j_2}}\left(x_{j_1+1}, \ldots, x_{j_2}\right) \times F_{X_{j_2+1},\ldots,X_p}\left(x_{j_2+1}, \ldots, x_p\right)$, then one can similarly add that $\left(X_1, \ldots, X_{j_1}\right), \left(X_{j_1+1}, \ldots, X_{j_2}\right)$, and $\left(X_{j_2+1}, \ldots, X_p\right)$ are independent and $F_{X_1,\ldots,X_{j_1}}\left(x_1, \ldots, x_{j_1}\right)$, $F_{X_{j_1+1},\ldots,X_{j_2}}\left(x_{j_1+1}, \ldots, x_{j_2}\right)$, and $F_{X_{j_2+1},\ldots,X_p}\left(x_{j_2+1}, \ldots, x_p\right)$ are the joint distributions of $\left(X_1, \ldots, X_{j_1}\right)$, $\left(X_{j_1+1}, \ldots, X_{j_2}\right)$, and $\left(X_{j_2+1}, \ldots, X_p\right)$, respectively.

If the distribution is of discrete type, then the total mass of the distribution of $(X_1, \ldots, X_p)$ is concentrated at the points in a way such that $\sum_{\forall j_1} \cdots \sum_{\forall j_p} Pr\left\{X_1 = x_{1,j_1}, \ldots, X_p = x_{p,j_p}\right\} = 1$, while for the continuous case, we have $\int_{x_p=-\infty}^{x_p=+\infty} \cdots \int_{x_1=-\infty}^{x_1=+\infty} f(x_1, \ldots, x_p)\, dx_1 \cdots dx_p = 1$. The corresponding joint distribution functions would be given as $F_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = \sum_{X_1 \le x_1} \cdots \sum_{X_p \le x_p} Pr\{X_1 \le x_1, \ldots, X_p \le x_p\}$ or $F_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = \int_{x_p=-\infty}^{x_p} \cdots \int_{x_1=-\infty}^{x_1} f(x_1, \ldots, x_p) dx_1 \cdots dx_p$ as the case may be.

Considering $\boldsymbol{X}_{n\times p} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, we may be interested to measure the degree to which one of the variable, say $X_j$, is dependent on the remaining $(p-1)$ number of variables, that is, $(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$ taken jointly. This is called *multiple correlation* and the measure is given by *multiple correlation coefficient*

$$\rho_{j,(1,\ldots,j-1,j+1,\ldots,p)} = \frac{Covar\left(X_j, X_{1,\ldots,j-1,j+1,\ldots,p}\right)}{\sqrt{Var(X_j)} \times \sqrt{Var\left(X_{j,(1,\ldots,j-1,j+1,\ldots,p)}\right)}} = \left(1 - \frac{R}{R_{j_1,j_2}}\right)^{1/2},$$

where $R_{j_1,j_2}$ is the cofactor of $\rho_{j_1,j_2}$ in the determinant $R$ of the correlation matrix

$$\boldsymbol{R} = \begin{pmatrix} \rho_{1,1} & \cdots & \rho_{1,p} \\ \vdots & \ddots & \vdots \\ \rho_{p,1} & \cdots & \rho_{p,p} \end{pmatrix}.$$

Another way of representing the multiple correlation coefficient is

$$\rho^2_{j,(1,\ldots,j-1,j+1,\ldots,p)} = 1 - \left\{\frac{Var\left(\varepsilon_{j,(1,\ldots,j-1,j+1,\ldots,p)}\right)}{Var(X_j)}\right\},$$

where $\varepsilon_{j,(1,\ldots,j-1,j+1,\ldots,p)}$ is the residual of $X_j$ corresponding to its multiple regression on $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$. This multiple correlation coefficient may be interpreted as the *maximum* correlation between $X_j$ and a linear function of $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$, say $X_j = \alpha + \beta_1 X_1 + \cdots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \cdots + \beta_p X_p + \varepsilon$. On the other hand, *partial correlation coefficient* between $X_{j_1}$ and $X_{j_2}$ is denoted by $\rho_{(j_1,j_2|1,\ldots,j_1-1,j_1+1,\ldots,j_2-1,j_2+1,\ldots,p)} = (-1)^{j_1+j_2}\left(R_{j_1,j_2}/\left(R_{j_1,j_1} R_{j_2,j_2}\right)^{1/2}\right)$, where $R_{j_1,j_1}$, $R_{j_2,j_2}$, and $R_{j_1,j_2}$ have the usual definition as already mentioned while discussing multiple correlation coefficient, $\rho_{j,(1,\ldots,j-1,j+1,\ldots,p)}$. Remember that partial correlation coefficients are related to the partial regression coefficients by the formula

$$\beta_{(j_1,j_2|1,\ldots,j_1-1,j_1+1,\ldots,j_2-1,j_2+1,\ldots,p)} = \rho_{(j_1,j_2|1,\ldots,j_1-1,j_1+1,\ldots,j_2-1,j_2+1,\ldots,p)} \times \frac{\sigma_{(j_1|1,\ldots,j_1-1,j_1+1,\ldots,j_2-1,j_2+1,\ldots,p)}}{\sigma_{(j_2|1,\ldots,j_1-1,j_1+1,\ldots,j_2-1,j_2+1,\ldots,p)}}.$$

**Example 8.10**

Consider the data related to cigarettes given in Mendenhall and Sincich (2006). The data set contains measurements related to brand, tar content (mg), nicotine content (mg), weight (g), and carbon monoxide content (mg) for $n = 25$ brands of cigarettes. The data set can be accessed at http://www.amstat.org/publications/jse/datasets/cigarettes. dat.txt. Considering the variables,

```
DEFINE: $n$, $p$, $\boldsymbol{A}$, $\boldsymbol{R}$, determinant of $\boldsymbol{R}$ (i.e., det(R)), cofactors of $(j_1,j_2)$ (i.e.,$R_{j_1,j_2}$) from $\boldsymbol{R}$
INPUT: $n$, $p$, $\boldsymbol{A}$
START If: $j_1 = 1{:}p$
   START If: $j_2 = 1{:}p$
   CALCULATE: $\boldsymbol{R}$, determinant of $\boldsymbol{R}$ (i.e., det(R)), cofactors of $(j_1,j_2)$ (i.e.,$R_{j_1,j_2}$) from $\boldsymbol{R}$
   END if
END if
CALCULATE: $\left(1-\frac{det\,(R)}{R_{j_1,j_2}}\right)^{\frac{1}{2}}$
REPORT: $\left(1-\frac{det\,(R)}{R_{j_1,j_2}}\right)^{\frac{1}{2}}$
END
```

**FIGURE 8.3**
Pseudo-code used for calculating multiple correlation coefficient vector.

$p = 4$, that is, $X_1$, $X_2$, $X_3$, and $X_4$ as the tar content (mg), nicotine content (mg), weight (g), and carbon monoxide content (mg), respectively, one obtains

$$\boldsymbol{R} = \begin{bmatrix} 1 & 0.9766 & 0.4908 & 0.9575 \\ 0.9766 & 1 & 0.5002 & 0.9259 \\ 0.4908 & 0.5002 & 1 & 0.4640 \\ 0.9575 & 0.9259 & 0.4640 & 1 \end{bmatrix}.$$

Furthermore, utilizing $\boldsymbol{R}$, we get the multiple correlation coefficient vector as

$$\begin{pmatrix} 0.9867 \\ 0.9774 \\ 0.5001 \\ 0.9584 \end{pmatrix},$$

while the corresponding $R^2$ values are 0.9720, 0.9554, 0.5366, and 0.9174. Simple calculations would also yield the partial correlation coefficient matrix as

$$\begin{pmatrix} 1 & -0.8199 & -0.0141 & -0.6556 \\ -0.8199 & 1 & -0.1092 & 0.1465 \\ -0.0141 & -0.1092 & 1 & 0.0072 \\ -0.6556 & 0.1465 & 0.0072 & 1 \end{pmatrix}.$$

To double verify the calculation, we may also calculate the partial regression coefficients. The pseudo-codes for calculating the multiple correlation coefficient vector and the partial correlation coefficient matrix are given in Figures 8.3 and 8.4, respectively.

```
DEFINE: $n$, $p$, $\boldsymbol{A}$, $\boldsymbol{R}$, cofactors of $(j_1,j_2)$ (i.e.,$R_{j_1,j_2}$) from $\boldsymbol{R}$
INPUT: $n$, $p$, $\boldsymbol{A}$
START If: $j_1 = 1{:}p$
   START If: $j_2 = 1{:}p$
   CALCULATE: $\boldsymbol{R}$, cofactors of $(j_1,j_2)$ (i.e.,$R_{j_1,j_2}$) from $\boldsymbol{R}$
   END if
END if
CALCULATE: $(-1)^{j_1+j_2}\frac{R_{j_1,j_2}}{(R_{j_1,j_1}R_{j_2,j_2})^{\frac{1}{2}}}$
REPORT: $(-1)^{j_1+j_2}\frac{R_{j_1,j_2}}{(R_{j_1,j_1}R_{j_2,j_2})^{\frac{1}{2}}}$
END
```

**FIGURE 8.4**
Pseudo-code used for calculating partial correlation coefficient matrix.

## 8.8 Multinomial Distribution

Suppose $X_1, \ldots, X_p$ be $p$ jointly distributed random variables each of which is discrete, nonnegative, and integer valued. Then, the joint probability mass function of $X_1, \ldots, X_p$ is called the *multinomial* distribution and is of the form

$$\binom{n}{x_1, \ldots, x_p} \times p_1^{x_1} \times \cdots \times p_p^{x_p},$$

where $x_i \in (0, 1, \ldots n-1, n)$, $\sum_{j=1}^{p} x_j = n$, and $\sum_{j=1}^{p} p_j = 1$.

For the multinomial distribution, it can be easily proved that

1. $E(X_j) = np_j, j = 1, \ldots, p.$
2. $Var(X_j) = np_j(1 - p_j), j = 1, \ldots, p.$
3. $Covar\left(X_{j_1}, X_{j_2}\right) = -np_{j_1}p_{j_2}, j_1 \neq j_2 = 1, \ldots, p.$
4. $$corr\left(X_{j_1}, X_{j_2}\right) = -\left\{\frac{p_{j_1}p_{j_2}}{\left(1 - p_{j_1}\right)\left(1 - p_{j_2}\right)}\right\}^{1/2}, \quad j_1 \neq j_2 = 1, \ldots, p.$$
5. Moment generating function (MGF) $= \left\{\sum_{j=1}^{p} p_j e^{t_j}\right\}^n$.
6. Characteristic function (CF) $= \left\{\sum_{j=1}^{p} p_j e^{it_j}\right\}^n$, where $i^2 = -1$.
7. Probability generating function (PGF) $= \left\{\sum_{j=1}^{p} p_j z_j\right\}^n$ for $(z_1, \ldots, z_p) \in C^p$.

For a better appreciation of the multinomial distribution, consider the polynomial coefficients of the expansion of the multinomial expansion, $\{p_1x_1 + \cdots + p_px_p\}^n$. A closer look at this expansion will make it immediately evident how the polynomial coefficients of the multinomial expansion correspond to the multinomial distribution discussed above. Another interesting analogy for multinomial distribution can be made from the *Pascal pyramid* (Figure 8.5).

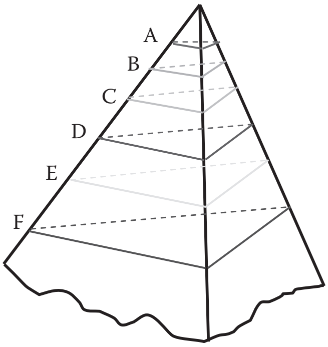


**FIGURE 8.5**
Pascal pyramid considered to depict the coefficient of the multinomial distribution.

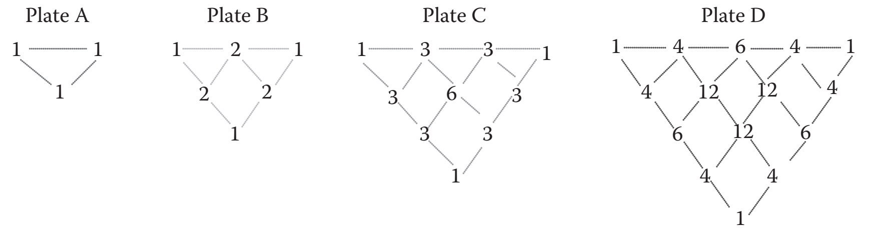


**FIGURE 8.6**
Depiction of the flat plates of Pascal pyramid to signify the concept of multinomial distribution.

If we view the slices (as represented by A, B, C, D, E, F, and so on) of the *Pascal pyramid* as flat triangular plates, then the numbers depicted on them are as shown in Figure 8.6.

When $p = 2$, we have the binomial distribution, while for $p = 3$, one obtains the trinomial distribution, and so on. It can be shown that the *marginal distribution* of $X_j$, $j = 1, \ldots, p$ is binomially distributed with parameters $n$ and $p_j$, and is given by

$$Pr(X_j = x_j) = \binom{n}{x_j} p_j^{x_j}(1 - p_j)^{n - x_j},$$

where $(p_1 + \cdots + p_{j-1}) + p_j + (p_{j+1} + \cdots + p_p) = 1$ and $\{(x_1 + \cdots + x_{j-1}) + x_j + (x_{j+1} + \cdots + x_p)\} = n$. If one considers the *conditional distribution* of $X_j$, then given $X_1 = x_1, \ldots, X_{j-1} = x_{j-1}, X_{j+1} = x_{j+1}, \ldots, X_p = x_p$, the *conditional distribution* is

$$Pr(X_j = x_j | X_1 = x_1, \ldots, X_{j-1} = x_{j-1}, X_{j+1} = x_{j+1}, \ldots, X_p = x_p)$$
$$= \left( n / \left( j!\{n - \sum_{i \in (J-j)} x_i\}! \right) \right) p_j^{x_j}(1 - p_j)^{n - x_j},$$

**Example 8.11**

Consider the use of contraceptive among married women in El Salvador in 1985, and the data for the same for a sample of 3165 respondents is shown in Table 8.1.

**TABLE 8.1**
Data Related to Use of Contraceptive among Married Women in El Salvador, 1985

| | **Contraceptive Method** | | | |
|---|---|---|---|---|
| **Age (Years)** | **Sterilization** | **Other Method** | **None** | **All** |
| 15–19 | 3 | 61 | 232 | 296 |
| 20–24 | 80 | 137 | 400 | 617 |
| 25–29 | 216 | 131 | 301 | 648 |
| 30–34 | 268 | 76 | 203 | 547 |
| 35–39 | 197 | 50 | 188 | 435 |
| 40–44 | 150 | 24 | 164 | 338 |
| 45–49 | 91 | 19 | 183 | 284 |
| All | 1005 | 489 | 1671 | 3165 |

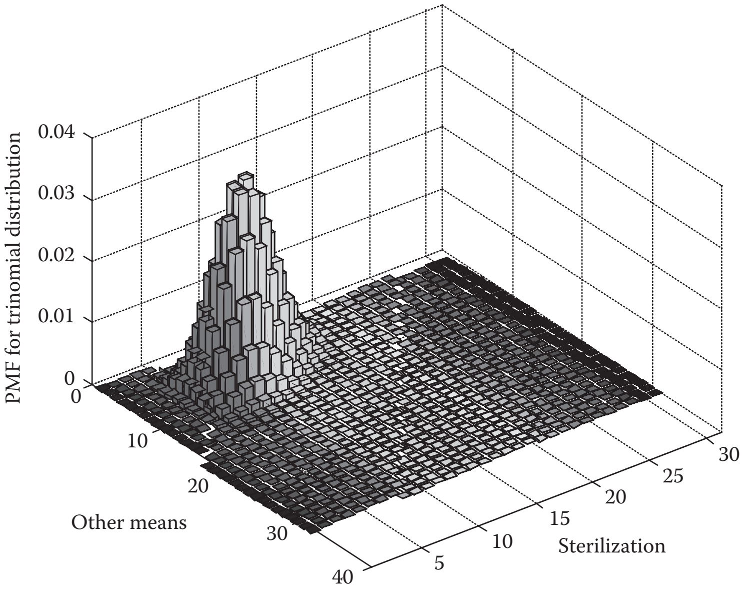


**FIGURE 8.7**
PMF for the trinomial distribution utilizing the data for the contraceptive use by married women in El Salvadore, 1985, Example 8.11.

Consider $X_1$, $X_2$, and $X_3$ as the random variable signifying the case of sterilization, other method, and none. Then, the joint multinomial distribution may be written as

$$\begin{aligned} f_{X_1,X_2,X_3}(x_1,x_2,x_3) &= Pr(X_1=x_1,X_2=x_2,X_3=x_3) \\ &= \binom{n}{x_1,x_2,x_3}\times\left(\frac{1005}{3165}\right)^{x_1}\times\left(\frac{489}{3165}\right)^{x_2}\times\left(\frac{1671}{3165}\right)^{x_3}, \end{aligned}$$

where $(x_1+x_2+x_3)=n$. If we assume $n=30$, then the PMF of the trinomial distribution and the general pseudo-code used for generating the same are given in Figures 8.7 and 8.8, respectively.

In case one is interested to generate the multivariate multinomial random variable, then the MATLAB® code is `mnrnd(n,`$\boldsymbol{p}_p$`,m)`, where $n$ is the dimension of the multinomial distribution, $\boldsymbol{p}_p=(p_1,\ldots,p_p)$, and $m$ is the number of observations, that is, the sample size we wish to generate. Hence, if $p=(0.1,0.2,0.3,0.4)$, $n=100$ then $E(\boldsymbol{X})=(\mu_1,\mu_2,\mu_3,\mu_4)=(10,20,30,40)$,

```
DEFINE: $\boldsymbol{p}$, $p_j$, $x_j$, $j=1,\ldots,p$, $f(x)=\binom{n}{x_1,\ldots,x_p}\times p_1^{x_1}\times\ldots..\times p_p^{x_p}$, x-axis, y-axis, z-axis
INPUT: $\boldsymbol{p}$, $p_j$, $j=1,\ldots,p$
START If: $i=1{:}n$
    CALCULATE: $f(x)$
END if
PLOT: (x-axis, y-axis, z-axis)
REPORT: $f(x)$
END
```

**FIGURE 8.8**
Pseudo-code used for Example 8.11.

$Var(\boldsymbol{X}) = (9, 16, 21, 24)$. Other values such as $Covar\left(X_{i_1}, X_{i_2}\right)$, $corr\left(X_{i_1}, X_{i_2}\right)$, PGF, MGF, and CF can be calculated accordingly.

## 8.9 Multivariate Normal Distribution

We say $\boldsymbol{X}_p \sim N_p(\mu, \boldsymbol{\Sigma})$ is a nonsingular MND when its density function is given by

$$f_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = \frac{1}{(2\pi)^{p/2}} |\boldsymbol{\Sigma}|^{-1/2} e^{\{-(1/2)(\boldsymbol{X}-\mu)'\Sigma^{-1}(\boldsymbol{X}-\mu)\}},$$

where $-\infty < X_i < +\infty$, $E(\boldsymbol{X}) = \mu$, $Covar(\boldsymbol{X}) = \boldsymbol{\Sigma}$ along with the fact that

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix} > \mathbf{0}.$$

The study of MND is useful as many other multivariate statistics are approximately normal regardless of the parent distribution because of the CLT effect.

For the MND, it can be easily proved that

1. $E(X_j) = \mu_j$, $j = 1, \ldots, p$.
2. $Var(X_j) = \sigma^2$, $j = 1, \ldots, p$.
3. $Covar\left(X_{j_1}, X_{j_2}\right) = \sigma_{j_1,j_2}$, $j_1, j_2 = 1, \ldots, p$.
4. $corr\left(X_{j_1}, X_{j_2}\right) = \rho_{j_1,j_2}$, $j_1, j_2 = 1, \ldots, p$.
5. $\text{MGF} = e^{(\mu' \boldsymbol{t} + (1/2)\boldsymbol{t}'\boldsymbol{\Sigma t})}$, $t \in \mathbb{R}$.
6. $\text{CF} = e^{(i_m \mu' \boldsymbol{t} + (1/2)\boldsymbol{t}'\boldsymbol{\Sigma t})}$, where $i_m^2 = -1$.

Furthermore, we state a few results without proofs. Given $\boldsymbol{X}_p \sim N_p(\mu, \boldsymbol{\Sigma})$, we have the following:

1. Contours of consistent density for the $p$-dimensional normal distribution are ellipsoids defined by $\boldsymbol{X}$, such that $(\boldsymbol{X} - \mu)', \boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \mu) = c^2$, where $c$ is a constant. These ellipsoids are centered at $\mu$ and have axes $\pm c\sqrt{\lambda_j} e_j$, where $\boldsymbol{\Sigma} e_j = \lambda_j e_j$, $j = 1, \ldots, p$. Here, $(\lambda, \boldsymbol{e})$ is the eigen value–eigen vector for $\boldsymbol{\Sigma}$ corresponding to the pair $(1/\lambda, \boldsymbol{e})$ for $\boldsymbol{\Sigma}^{-1}$. Remember $\boldsymbol{\Sigma}^{-1}$ is positive definite.
2. If $\boldsymbol{\Sigma}$ is positive definite, so that $\boldsymbol{\Sigma}^{-1}$ exists, then $\boldsymbol{\Sigma e} = \lambda \boldsymbol{e}$ implies $\boldsymbol{\Sigma}^{-1}\boldsymbol{e} = (1/\lambda)\boldsymbol{e}$.
3. All subsets of the components of $\boldsymbol{X}$ have MNDs.
4. Zero covariance implies that the corresponding components are independently distributed.
5. The conditional distributions of the components are multivariate normal.
6. The $q$ linear combinations of the components of $\boldsymbol{X}$ are also normally distributed. Thus, if $\boldsymbol{X}$ is distributed as $N_p(\mu, \boldsymbol{\Sigma})$, then $q$ linear combinations $\boldsymbol{A}_{q\times p}\boldsymbol{X}_{p\times 1} \sim N_q(\boldsymbol{A}\mu, \boldsymbol{A}'\boldsymbol{\Sigma}\mathbf{A})$. Also, $\boldsymbol{X}_{p\times 1} + \boldsymbol{d}_{p\times 1}$, where $\boldsymbol{d}_{p\times 1}$ is a vector of constants, is distributed as $N_p(\mu + \boldsymbol{d}\boldsymbol{\Sigma})$.
7. $(\boldsymbol{X}_p - \mu) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$.

8. $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, provided $X_j \sim N(\mu_j,\ \sigma^2),\ j = 1, \ldots, p$, are mutually independent univariate normal distributions.
9. The solid ellipsoid of $\boldsymbol{x}$ values satisfying $(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \le \chi^2_p(\alpha)$ has probability $(1 - \alpha)$, that is, $Pr\left\{(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \le \chi^2_p(\alpha)\right\} = (1 - \alpha)$.
10. If $\boldsymbol{X}$ is distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then any linear combination of variables $\boldsymbol{a}'\boldsymbol{X} = \sum_{j=1}^{p} a_j X_j$ is distributed as $N_p(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$. Conversely, if $\boldsymbol{a}'\boldsymbol{X}$ is distributed as $N_p(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a})$ for every $\boldsymbol{a}$, then $\boldsymbol{X}$ must be $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
11. Suppose $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and

$$\boldsymbol{X}_p = \begin{bmatrix} \boldsymbol{X}_{k\times 1} \\ \cdots \\ \boldsymbol{X}_{(p-k)\times 1} \end{bmatrix}, \quad \boldsymbol{\mu}_p = \begin{bmatrix} \boldsymbol{\mu}_{k\times 1} \\ \cdots \\ \boldsymbol{\mu}_{(p-k)\times 1} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{p\times p} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \vdots & \boldsymbol{\Sigma}_{12} \\ \cdots & \cdots & \cdots \\ \boldsymbol{\Sigma}_{21} & \vdots & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

then $\boldsymbol{X}_k \sim N_k(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{X}_{p-k} \sim N_{p-k}(\boldsymbol{\mu}_{p-k}, \boldsymbol{\Sigma}_{22})$. The converse of this is also true.

An interesting and important concept in MND is something to do with circles and ellipses. Consider $p = 2$, then

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}\left(1 - \rho_{12}^2\right)}} \times \exp\left[-\frac{1}{2\left(1 - \rho_{12}^2\right)}\left\{\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}}\right)^2 \right.\right.$$
$$\left.\left. -2\rho_{12}\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}}\right)\right\}\right].$$

If we consider three different values of $\rho_{12}$, that is, negative, zero, and positive, and consider, $\sigma_{11} = \sigma_{22}$, then we obtain the contours as shown in Figure 8.9.

**Example 8.12**

Let us consider the data presented by Galton (1886), which shows a cross-tabulation of 963 adult children (486 sons and 476 daughters) born to 205 families, by their height and their midparent's

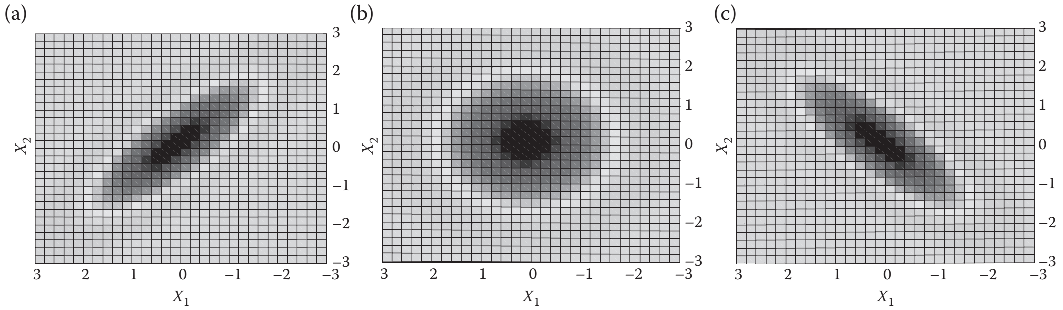


**FIGURE 8.9**
Contour plots for bivariate normal distribution considering $\rho_{12}$ as (a) negative, (b) zero, and (c) positive.

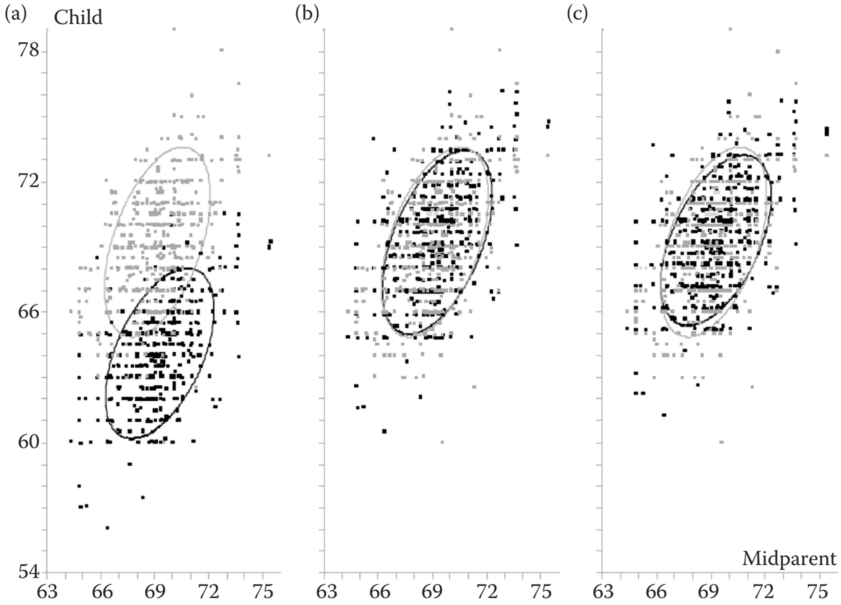


**FIGURE 8.10**
Ellipsoidal plots for Example 8.12 (refer Hanley, 2004).

height. The author visually smoothed the bivariate frequency distribution and showed that the contours formed concentric and similar ellipses, thus setting the stage for correlation, regression, and the bivariate normal distribution. The data is recorded in class intervals of width 1.0″. Furthermore, he used noninteger values for the center of each class interval because of the strong bias toward integral inches. All of the heights of female children were multiplied by 1.08 before tabulation to compensate for sex differences. One can also refer to Hanley (2004), along with the source materials at http://www.medicine.mcgill.ca/epidemiology/hanley/galton/ to have a better understanding about this study and the corresponding data analysis. The related ellipsoidal plots for this problem are shown in Figure 8.10.

The use of the basic concept of ellipsoids may be found in the area of reliability-based design optimization (RBDO), where this concept is used to depict the most reliable area (search space) within which the optimization solution is feasible, depending on the level of confidence. Though not exhaustive, a few good references in the area of RBDO are: Ben-Tal et al. (2009), Ben-Tal and Nemirovski (1998, 1999, 2002), and Bertsimas and Sim (2003, 2004, 2006).

## 8.10 Multivariate Student $t$-Distribution

The joint probability distribution function for the multivariate Student $t$-distribution (standard form) is

$$f_{Y_1,\ldots,Y_p}(y_1,\ldots,y_p) = \frac{\Gamma\{(v+p)/2\}}{(\pi v)^{p/2}\Gamma(v/2)|\mathbf{R}|^{1/2}}\left\{1+\frac{\boldsymbol{y}'\mathbf{R}^{-1}\boldsymbol{y}}{v}\right\}^{-(v+p)/2},$$

where $\boldsymbol{Y} = (Y_1,\ldots,Y_p)$, $v$ is the degree of freedom for univariate $t$-distribution, and $Y_j = X_j/(S_j/\sqrt{v})$. Remember $X_1,\ldots,X_p$ have a joint standard multinormal distribution with

$E(\boldsymbol{X}) = \mathbf{0}, Covar(\boldsymbol{X}) = \mathbf{R}$ (the correlation matrix), and $S_j = (1/(n-1))\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2$. Given this, the following results for this distribution can be easily derived:

1. $E(\boldsymbol{Y}) = \mathbf{0}$.
2. $Median(\boldsymbol{Y}) = \mathbf{0}$.
3. $Mode(\boldsymbol{Y}) = \mathbf{0}$.
4. $Covar(\boldsymbol{Y}) = (v/(v-2))\boldsymbol{R}$.

In its nonstandard form, it can be expressed as

$$f_{X_1,\ldots,X_p}\left(x_1,\ldots,x_p\right) = \frac{\Gamma\{(v+p)/2\}}{(\pi v)^{p/2}\Gamma(v/2)|\boldsymbol{\Sigma}|^{1/2}}\left\{1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{v}\right\}^{-(v+p)/2},$$

where $E(\boldsymbol{X}) = \boldsymbol{\mu}(v > 1)Covar(\boldsymbol{X}) = (v/(v-2))\boldsymbol{\Sigma}(v > 2)$, mode is $\boldsymbol{\mu}$, and finally median is also $\boldsymbol{\mu}$. When the $p$ random variables are independent then $Y_1^2 + \cdots + Y_k^2$ has a joint multivariate $F$-distribution with parameters $(k-1)$ and $(v+k)$, and for the case when $v = 1$ one obtains the multivariate Cauchy distribution. One should remember that as $v$ tends to infinity the joint distribution of $Y_1,\ldots,Y_p$ tends to the multinormal distribution with $E(\boldsymbol{Y}) = \mathbf{0}$ and $Covar(\boldsymbol{Y}) = \boldsymbol{R}$. Furthermore, the conditional probability distribution for the independent case is given by

$$f_{Y_{k+1},\ldots,Y_p|Y_1,\ldots,Y_k}(Y_{k+1},\ldots,Y_p|Y_1,\ldots,Y_k) = \frac{\Gamma\left(\dfrac{v+p}{2}\right)\left\{\dfrac{\left(1+v^{-1}\sum_{j=1}^{k}y_j^2\right)(v+k)}{v}\right\}^{(p-k)/2}}{\{\pi(v+k)\}^{(p-k)/2}\Gamma\left(\dfrac{v+k}{2}\right)}$$
$$\times\left[1 + \frac{1}{v\left(1+v^{-1}\sum_{j=1}^{k}y_j^2\right)}\sum_{j=k+1}^{p}y_j^2\right]^{-(v+p)/2}.$$

Provided all the marginals have the same degrees of freedom, $v$, the marginal probability distribution is of the form

$$\frac{\Gamma((v+p)/2)|\boldsymbol{\Sigma}^{-1}|^{1/2}}{\sqrt{v^p\pi^p}\,\Gamma(v/2)}\left\{1 + \frac{1}{v}\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}\sum_{j_1,j_2}^{-1} y_{j_1,j_2}\right\},$$

where

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \Sigma_{1,1}^{-1} & \cdots & \Sigma_{1,p}^{-1} \\ \vdots & \ddots & \vdots \\ \Sigma_{p,1}^{-1} & \cdots & \Sigma_{p,p}^{-1} \end{bmatrix}.$$

In case one is interested to obtain the noncentral multivariate $t$-distribution of

$$Y_j = \left(\frac{U_j + \delta_j}{S_j/\sqrt{v}}\right),$$

where $\delta_j$ is the noncentrality parameters for $t_j$ distribution (univariate $t$-distribution with parameter $v$), $\boldsymbol{U} \sim MVN(\boldsymbol{0}, \boldsymbol{R})$ and $\boldsymbol{U} = (U_1, \ldots, U_p)$, then it is given by

$$f_{Y_1,\ldots,Y_p}(y_1, \ldots, y_p) = \frac{e^{-(1/2)\boldsymbol{\delta}'\boldsymbol{R}^{-1}\boldsymbol{\delta}}}{(\pi v)^{p/2}\Gamma(v/2)|\boldsymbol{R}|^{1/2}}(1 + v^{-1}\boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{y})^{-(v+p)/2}$$
$$\times \sum_{j=0}^{\infty} \frac{\Gamma((v+p+j)/2)}{j!} \left\{ \frac{2(\boldsymbol{\delta}'\boldsymbol{R}^{-1}\boldsymbol{y})^2}{v(1 + v^{-1}\boldsymbol{y}'\boldsymbol{R}\boldsymbol{y})} \right\}^{(1/2)j}$$

Here, $\boldsymbol{R}$ is the correlation matrix for the standardized multinormal variables $\boldsymbol{U} = (U_1, \ldots, U_p)$. The use of multivariate Student $t$-distribution can be found in areas ranging from constructing simultaneous confidence intervals for the expected values of a number of normal populations to the study of stepwise linear multiple regression analysis. The multivariate Behrens–Fisher distribution may also be created using multivariate Student $t$-distribution. The use of multivariate Student $t$-distribution is nowadays utilized in finance whereby one uses the Student $t$-copula to find the dependence structure of $p$ number of financial scripts (Cherubini et al. 2004). Other areas where multivariate Student $t$-distribution is used are multiple decision problems, discriminant and cluster analysis, speech recognition, etc. One may refer to the following texts, viz., Johnson and Kotz (1972) and Kotz and Nadarajah (2004) to get a good idea about multivariate Student $t$-distribution.

**Example 8.13***

A copula, $C(u_1, \ldots, u_p)$, is a multivariate probability distribution for which the marginal probability distribution of each variable, $u_j$, $j = 1, \ldots, p$ is uniform, that is, [0,1]. They are used to describe the dependence between random variables, $X_1, \ldots, X_p$. As per the fundamental theorem of Sklar, every distribution $F_{X_1,\ldots,X_p}(x_1, \ldots, x_p)$ with marginals $F_{X_1}(x_1), \ldots, F_{X_p}(x_p)$ may be written using the copula function as $F_{X_1,\ldots,X_p}\left(x_1, \ldots, x_p\right) = C\left\{F_{X_1}(x_1), \ldots, F_{X_p}(x_p)\right\}$. Alternatively, $C(u_1, \ldots, u_p) = F_{U_1,\ldots,U_p}\left\{F_{u_1}^{-1}(u_1), \ldots, F_{u_p}^{-1}(u_p)\right\}$.

To illustrate the application of multivariate $t$-distribution, let us consider the following two scripts, namely, TATA STEEL and SBI from NSE, India (http://www.nse-india.com/) for the time period January 1, 2015 to May 29, 2015. If we draw the two-dimensional (2-D) copula (Figure 8.11) considering bivariate $t$-distributions between the returns, $r$, of the pair of stocks, then one obtains the PDF graphs. The Pearson correlation coefficient between the two scripts is found out to be

$$\begin{pmatrix} 1.0000 & 0.3513 \\ 0.3513 & 1.0000 \end{pmatrix}.$$

One should use the closing price, $P_{it}$, of each day, say $t$, to find

$$r = log_e\left(\frac{P_{i,t+1}}{P_{i,t}}\right),$$

and we use this for our calculations.

[*] The results of Examples 8.13 and 8.15 are part of different unpublished master's theses of students in Industrial and Management Engineering, Indian Institute of Technology, Kanpur, India, who have worked under the guidance of the first author, Raghu Nandan Sengupta.

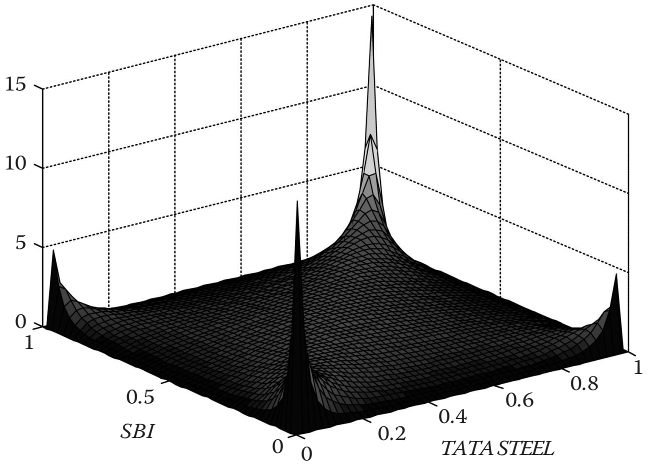


**FIGURE 8.11**
PDF of $C$(*TATA STEEL*,*SBI*) utilizing the concept of bivariate $t$-distribution, Example 8.13.

## 8.11 Wishart Distribution

In statistics, the Wishart distribution is a generalization of the chi-squared distribution in multiple dimensions. It was first formulated by John Wishart (1898–1956) (Wishart, 1928). Suppose $\boldsymbol{X}_k \sim N_p(\mu_k \boldsymbol{\Sigma})$, $k = 1, \ldots, \upsilon$ be independently distributed, then $\boldsymbol{W} = \sum_{k=1}^{\upsilon} X_k X_k' \sim W_p(\upsilon, \boldsymbol{\Sigma})$, where the parameters $\Sigma > 0$ is of size $(p \times p)$ and is positive definite, while $\upsilon > (p-1)$ is the degree of freedom. The Wishart distribution arises as the distribution of the sample covariance matrix for a sample from an MND. If $\boldsymbol{X}_{n \times p}$ is an $(n \times p)$ matrix of random variables, then the PDF is given by

$$f_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = \frac{1}{2^{\upsilon p/2} |\Sigma|^{\upsilon/2} \Gamma_p(\upsilon/2)} |\boldsymbol{X}|^{(\upsilon-p-1)/2} e^{-(1/2)tr(\Sigma^{-1}\boldsymbol{X})}.$$

We now state a few relevant properties of the Wishart distribution:

1. $E(\boldsymbol{W}) = \upsilon\boldsymbol{\Sigma} + \boldsymbol{M}'\boldsymbol{M}$, where $\boldsymbol{M}' = (\mu_1, \ldots, \mu_\upsilon)$.
2. Rank of $\boldsymbol{W} = \min(\upsilon, p)$.
3. If $\boldsymbol{W}_k \sim W_p(k, \boldsymbol{\Sigma}), k = 1, \ldots, \upsilon,$ then $\sum_{k=1}^{\upsilon} \boldsymbol{W}_k \sim W_p\left(\sum_{k=1}^{\upsilon} k, \boldsymbol{\Sigma}, \mathbf{M}\right)$, where $\boldsymbol{M}' = \left[\boldsymbol{M}_1 \vdots \cdots \vdots \boldsymbol{M}_\upsilon\right]$.
4. If $\boldsymbol{W} \sim W_p(\upsilon, \boldsymbol{\Sigma})$ and $\boldsymbol{C}$ is any $(p \times q)$ matrix of constants, then $\boldsymbol{C}'\boldsymbol{W}\boldsymbol{C} \sim W_q(\upsilon, \boldsymbol{C}'\boldsymbol{\Sigma}\mathbf{C}, \mathbf{MC})$, where $\boldsymbol{M}' = (\mu_1, \ldots, \mu_\upsilon)$.
5. $E(\boldsymbol{W}) = \upsilon\boldsymbol{\Sigma}$.
6. Mode$(\boldsymbol{W}) = (\upsilon - p - 1)$.
7. $\mathrm{Var}(\boldsymbol{W}_{i,j}) = \upsilon\left(\sigma_{i,j}^2 + \sigma_{i,i}\sigma_{j,j}\right)$, where $\sigma_{i,j}$ is the $i$th row and $j$th column element of $\boldsymbol{\Sigma}$.

Let us now consider the inverse Wishart distribution, denoted by $IW_p(.)$. It is the multivariate extension of the inverse gamma distribution. If one considers that the Wishart distribution generates

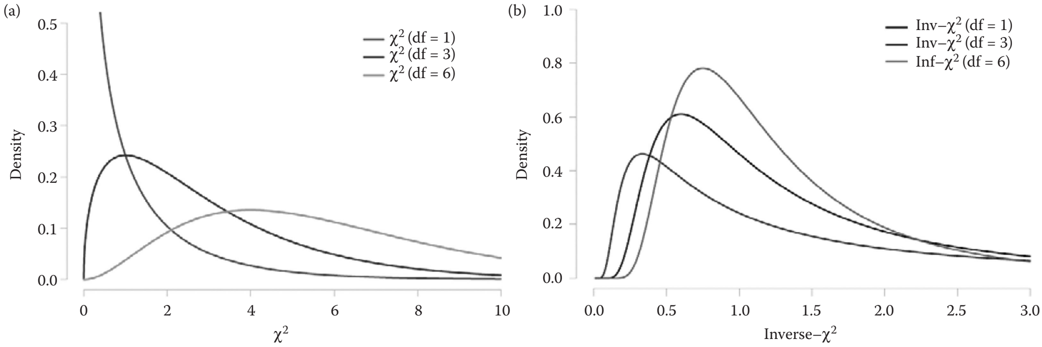


**FIGURE 8.12**
PDF for (a) $\chi^2_\upsilon$ and (b) inverse-$\chi^2_\upsilon$ considering $\upsilon = 1, 3, 6$, Example 8.14.

the sum of squares matrices, then the inverse Wishart distribution can be imagined as that which generates random covariance matrices. Hence, if $\boldsymbol{W} \sim W_p(\upsilon, \boldsymbol{\Sigma})$, then $\boldsymbol{W}^{-1} \sim IW_p(\upsilon, \boldsymbol{\Sigma}^{-1})$. The use of inverse Wishart distribution can be found in Bayesian statistics where it is used as a prior on the variance/covariance matrix, $\boldsymbol{\Sigma}$, of an MND. If we consider the inverse gamma distribution as the conjugate prior of the variance parameter, $\sigma^2$, for the univariate normal distribution, then the inverse Wishart distribution can be said to extend this conjugacy to the MND case. Another important point worth mentioning is the fact that using Helmert transformation, the Wishart distribution can be expressed as two distributions, one for the sample means and another for the sample variances–covariances. These transformations are orthogonal in nature, which makes it intuitive to understand that the sample means and the sample variances–covariances are independent of each other. Thus, as in univariate theory, the sample mean vector and the sample variance–covariance are also independently distributed for the multidimensional case.

A few good references for Wishart distribution are: Anderson (2003), Chatfield and Collins (1980), Cuadras and Rao (1993), Dempster (1969), and Eaton (1983).

**Example 8.14**

Let us illustrate the Wishart and inverse Wishart distributions in the simple case where we consider their univariate counterpart, which are $\chi^2_\upsilon$ and inverse-$\chi^2_\upsilon$ (Figure 8.12). In Figure 8.12, the degrees of freedom considered are 1, 3, and 6. Remember that for the multivariate case, one can make deductions about the Wishart and inverse Wishart distributions in a similar manner as we can do for $\chi^2_\upsilon$ and inverse-$\chi^2_\upsilon$ cases in the univariate setup.

## 8.12 Multivariate Extreme Value Distribution

Multivariate extreme gives us the picture of the asymptotic behavior of componentwise maxima of $i.i.d.$ observations. The main problem one faces is how to define MEVD. This problem arises due to the fact that there does not exist any strict ordering principle for multivariate observations. Though we use concepts of ordering such as marginal ordering (M-ordering), reduced (aggregate) ordering (R-ordering), partial ordering (P-ordering), conditional (sequential) ordering (C-ordering), etc. to accomplish this task, yet the ordering problem does occur in many cases.

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})$ be $i.i.d.$ such that $M_{\max,j} = \max\{Y_{1,j}, \ldots, Y_{n,j}\}$, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$. As per definition, $n$ is the number of observations, while $p$ is the dimension. Given this, we are interested to find the normalizing scaling constants $a_{n,j}$ and $b_{n,j}$ such that $Pr\{((\boldsymbol{M}_{\max} - \boldsymbol{b}_n)/\boldsymbol{a}_n) \leq \boldsymbol{x}\} \rightarrow G(\boldsymbol{x})$ as $n \rightarrow \infty$. Here, three important points about $G(\boldsymbol{x})$ should be mentioned:

1. If $G(\boldsymbol{x})$ is MEVD, then each of its marginal must be one of the univariate extreme value distributions (EVDs) and hence can be represented in general extreme value (GEV) form.
2. The form of the limiting distribution is invariant under monotonic transformation for each of the component.
3. Each marginal distribution can be transformed into specified forms, and one of these transformed specified forms is the Fréchet form given by $Pr\{X_j \leq x\} = e^{-x^{-\alpha}}, x > 0, j = 1, \ldots, p$ and $\alpha > 0$. When $\alpha = 1$, we have the unit Fréchet.

In general, we are interested to find the scalar transformations of each, $X_j,\ j = 1, \ldots, p$. Using the same concept as used in the univariate case, we intend to find $M_{\max,1}, \ldots, M_{\max,p} = \{\max_{1 \leq i \leq n}, X_{i,1}, \ldots, \max_{1 \leq i \leq n} X_{i,p}\}$ as $n \rightarrow \infty$. Utilizing simple scalar transformation, one needs to find

$$Pr\left\{\frac{(M_{n,1} - b_{n,1})}{a_{n,1}} \leq x_1, \ldots, \frac{(M_{n,p} - b_{n,p})}{a_{n,p}} \leq x_p\right\}$$
$$= F^n\{(a_{n,1}x_1 + b_1), \ldots, (a_{n,p}x_p + b_p)\} \rightarrow G(x_1, \ldots, x_p),$$

when $n \rightarrow \infty$. In case if

$$Pr\left\{\frac{(M_{n,1} - b_{n,1})}{a_{n,1}} \leq x_1, \ldots, \frac{(M_{n,p} - b_{n,p})}{a_{n,p}} \leq x_p\right\}$$
$$= F^n\{(a_{n,1}x_1 + b_1), \ldots, (a_{n,p}x_p + b_p)\} \rightarrow G(x_1, \ldots, x_p)$$

holds, for some suitable choices of $a_{n,j}$ and $b_{n,j},\ j = 1, \ldots, p$, then we say that $G(x_1, \ldots, x_p)$ is a MEVD and $F$ is in the domain of attraction of $G$. A question that automatically arises is what are the normalizing scaling constants, $a_{n,j}$ and $b_{n,j},\ j = 1, \ldots, p$, in their general form. For the convenience of the readers, we state below the scaling constants, $a_{n,j}$ and $b_{n,j}$ for the case when $n \rightarrow \infty$ and $j = 1, \ldots, p$.

1. For Type I distribution: $a_{n,j} = F_{X_j}^{-1}(1 - 1/n)$ and $b_{n,j} = F_{X_j}^{-1}(1 - 1/ne) - F_{X_j}^{-1}(1 - 1/n)$.
2. For Type II distribution: $a_{n,j} = 0$ and $b_{n,j} = F_{X_j}^{-1}(1 - 1/n)$.
3. For Type III distribution: $a_{n,j} = F_{X_j}^{-1}(1)$ and $b_{n,j} = F_{X_j}^{-1}(1) - F_{X_j}^{-1}(1 - 1/n)$.

The two extreme forms of the limiting multivariate distribution correspond to (i) the case of the asymptotic total independence between componentwise maxima for which $G(x_1, \ldots, x_p) = G_1(x_1) \cdots G_p(x_p)$ and (ii) the case of asymptotic total dependence between componentwise maxima for which $G(x_1, \ldots, x_p) = \min\{G_1(x_1), \ldots, G_p(x_p)\}$.

Remember that $G(x_1, \ldots, x_p) = G_1(x_1) \cdots\ G_p(x_p)$ holds true if and only if

1. $G(0, \ldots, 0) = G_1(0) \cdots G_p(0) = e^{-p}$, provided $G_j's$ are Gumble type with $G_j(x_j) = exp\{-exp(-x_j)\}$ for $j = 1, \ldots, p$.
2. $G(1, \ldots, 1) = G_1(1) \cdots G_p(1) = e^{-p}$, provided $G_j's$ are Fréchet type with $G_j(x_j) = exp\left\{-x_j^{-\alpha_j}\right\}$ and $\alpha_j > 0$ for $j = 1, \ldots, p$.

3. $G(-1,\ldots,-1) = G_1(-1)\cdots G_p(-1) = e^{-p}$, provided $G'_j s$ are Weibull type with $G_j(x_j) = exp\{-(-x_j)^{\alpha_j}\}$ and $\alpha_j > 0$ for $j = 1,\ldots,p$.

Before we end the discussion regarding MEVD, we give a few examples of MEVD considering $p = 2$.

1. Logistic MEVD:

$$exp\left[-\left(\frac{1-\psi_1}{x_1}\right)-\left(\frac{1-\psi_2}{x_2}\right)-\left\{\left(\frac{\psi_1}{x_1}\right)^q+\left(\frac{\psi_2}{x_2}\right)^q\right\}^{1/q}\right],$$

   where $0 \le \psi_1, \psi_2 \le 1$, and $q > 1$ have three usual meanings. In case $\psi_1 = 1$ and $\psi_2 = \alpha$, then we obtain the biextremal distribution of the form

$$exp\left[-\left(\frac{1-\alpha}{x_2}\right)-\left\{\left(\frac{1}{x_1}\right)^q+\left(\frac{\alpha}{x_2}\right)^q\right\}^{1/q}\right],$$

   while for $\psi_1 = \psi_2 = \alpha$, one obtains the Gumble distribution.

2. Negative logistic MEVD:

$$exp\left[-\left(\frac{1}{x_1}\right)-\left(\frac{1}{x_2}\right)-\alpha\left\{\left(\frac{1}{x_1}\right)^q+\left(\frac{1}{x_2}\right)^q\right\}^{1/q}\right],$$

   where $0 \le \psi_1, \psi_2 \le 1, q < 0$, and $\alpha$ have three usual meanings.

3. Bilogistic MEVD:

$$exp\left[-\int_0^1 \max\left\{\frac{(q_1-1)s^{-1/q_1}}{q_1x_1}, \frac{(q_2-1)s^{-1/q_2}}{q_2x_2}\right\}ds\right],$$

   where $q_1, q_2 > 1$.

   **Note:** A general multivariate case can be thought as

$$exp\left[-\int_0^1 \max\left\{\frac{(q_1-1)s^{-1/q_1}}{q_1x_1},\ldots, \frac{(q_p-1)s^{-1/q_p}}{q_px_p}\right\}ds\right],$$

   where $q_j > 1, j = 1,\ldots,p$.

4. Negative bilogistic MEVD:

$$exp\left[-\int_0^1 \max\left\{\frac{(q_1-1)s^{-1/q_1}}{q_1x_1}, \frac{(q_2-1)s^{-1/q_2}}{q_2x_2}\right\}ds\right],$$

   where $q_1, q_2 < 1$.

5. Gaussian MEVD: If one considers the bivariate extremes for the normal distribution, then one obtains

$$exp\left[-\left(\frac{1}{x_1}\right)\Phi\left\{a-s\left(\frac{x_1}{x_1+x_2}\right)\right\}-\left(\frac{1}{x_2}\right)\Phi\left\{s\left(\frac{x_1}{x_1+x_{12}}\right)\right\}\right],$$

where $s(w) = (a^2 + 2log_e w - 2log_e(1 - w))/2a$ and $a = ((x_1 - x_2)/\sigma)^2$. Here, $\Phi$ and $\sigma$ imply the standard normal cumulative deviate and the standard deviation, respectively, for the normal distribution, based on which the Gaussian MEVD is formulated.

A few good references for MEVD are: Coles (2001), de Haan and Resnick (1977), Kotz and Nadarajah (2000), Marshall and Olkin (1967, 1983), Pickards (1981), Sibuya (1960), and Tiago de Oliveira (1958, 1975).

**Example 8.15**

Figure 8.13 illustrates the EVD (for ease of illustration, we show the univariate EVD case only) using the *positive* values of returns, $r = log_e(P_{i,t+1}/P_{i,t})$, of the indices of four countries, namely, Nikkei (Japan), Nifty (India), FTSE (the United Kingdom), and KOSPI (Korea), for a time period of 10 years from 2003 to 2012. The values of *shape* (μ), *scale* (σ), and *location* (ξ) parameters for the EVD for the four indices are (i) 0.092, 0.0075, 0.0158; (ii) 0.1745, 0.0084, 0.0169; (iii) 0.1475, 0.0068, 0.0118; and (iv) 0.1571, 0.0074, 0.0163, respectively. Considering returns as *negative*,, one can also calculate the values of μ, σ, and ξ and draw similar EVD graphs for these four indices. We leave that to the readers to work on them so that they get a better understanding of the concepts about which we have discussed. Finally, considering the overall

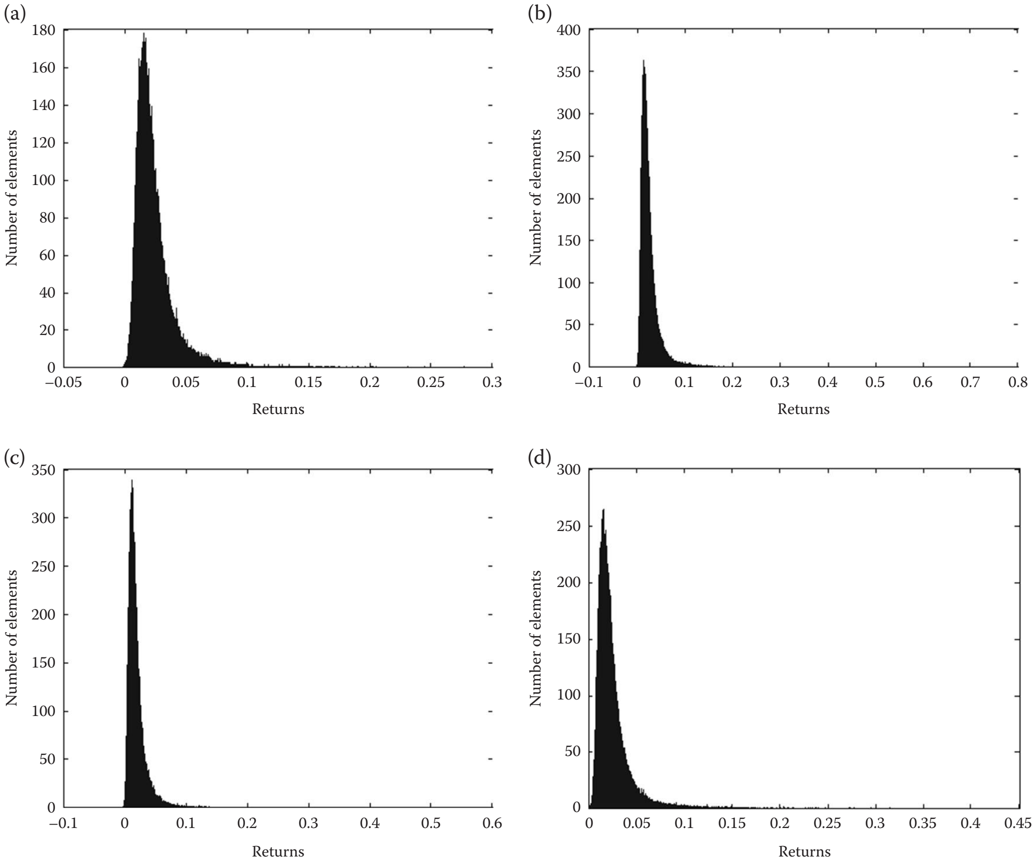


**FIGURE 8.13**
PDF for (a) Nikkei, (b) Nifty, (c) FTSE, and (d) KOSPI considering positive returns, Example 8.15.

returns (both positive and negative), we calculate the mean and standard deviation values for the four indices, which are: (i) $E(X_{NIKKEI}) = -0.022066$, $Var(X_{NIKKEI}) = 0.0153958$; (ii) $E(X_{NIFTY}) = -0.022211493$, $Var(X_{NIFTY}) = 0.000330646$; (iii) $E(X_{FTSE}) = -0.01670011$, $Var(X_{FTSE}) = 0.000155784$; and (iv) $E(X_{KOSPI}) = -0.021793$, $Var(X_{KOSPI}) = 0.0002269$.

## 8.13 MLE Estimates of Parameters (Related to MND Only)

For

$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} X_{1,1} & \cdots & X_{1,p} \\ \vdots & \ddots & \vdots \\ X_{n,1} & \cdots & X_{n,p} \end{pmatrix}$$

if one needs to estimate the parameters, then the total number of estimates required is $(1/2)p(p+1)$. Let us consider the case of MND, such that its log likelihood equation may be written as $log_e L(\mu, \boldsymbol{\Sigma}) = -(np/2)log_e 2\pi - (n/2)log_e|\boldsymbol{\Sigma}| - (1/2)\sum_{i=1}^{n}(\boldsymbol{X}-\mu)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\mu)$. Solving $\partial log_e L(\mu, \boldsymbol{\Sigma})/\partial\mu_j$ and $\partial log_e L(\mu, \boldsymbol{\Sigma})/\partial\sigma_{j_1,j_2}$, $j_1 < j_2$, where $j = 1, \ldots, p$ and $j_1, j_2 = 1, \ldots, p$ we obtain

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_p \end{pmatrix} = \bar{\boldsymbol{X}} = \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{pmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\sigma}_{1,1} & \cdots & \hat{\sigma}_{1,p} \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{p,1} & \cdots & \hat{\sigma}_{p,p} \end{pmatrix} = \mathbf{S} = \begin{pmatrix} s_{1,1} & \cdots & s_{1,p} \\ \vdots & \ddots & \vdots \\ s_{p,1} & \cdots & s_{p,p} \end{pmatrix},$$

where $\bar{x}_j = (1/n)\sum_{i=1}^{n} x_{ij}$ and $s_{j_1,j_2} = (1/(n-1))\sum_{i=1}^{n}\left(x_{i,j_1} - \bar{x}_{j_1}\right)\left(x_{i,j_2} - \bar{x}_{j_2}\right)$, for $i = 1, \ldots, n$, $j = 1, \ldots, p$ and $j_1, j_2 = 1, \ldots, p$.

Let us now suppose the case of hypothesis testing. A few relevant results without proofs for the same can be stated as follows:

1. The test statistics for $H_o : \boldsymbol{a}'\mu = \boldsymbol{a}'\mu_o$ against $H_A : \boldsymbol{a}'\mu >$ or $\neq$ or $< \boldsymbol{a}'\mu_o$, given $\boldsymbol{\Sigma}$ is *known*, is $n(\bar{\boldsymbol{X}} - \mu_o)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \mu_o)$. The distribution $n(\bar{\boldsymbol{X}} - \mu_o)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \mu_o) \sim \chi^2_{\mathrm{p}}$ and the value of $\alpha$, that is, the level of confidence, is assumed, based on the problem formulation and practical requirements.
2. The test statistics for $H_o : \boldsymbol{a}'\mu = \boldsymbol{a}'\mu_o$ against $H_A : \boldsymbol{a}'\mu >$ or $\neq$ or $< \boldsymbol{a}'\mu_o$, given $\boldsymbol{\Sigma}$ is *unknown*, is $n(\bar{\boldsymbol{X}} - \mu_o)'\mathbf{S}^{-1}(\bar{\boldsymbol{X}} - \mu_o)$. The distribution $n(\bar{\boldsymbol{X}} - \mu_o)'\mathbf{S}^{-1}(\boldsymbol{X} - \mu_o) \sim T^2_p$ and the value of $\alpha$, that is, the level of confidence, is assumed, based on the problem formulation and practical requirements.
3. The characteristics form of $T^2$ statistic is $T^2 = (\bar{\boldsymbol{X}} - \mu_o)'(\mathbf{S}/n)^{-1}(\bar{\boldsymbol{X}} - \mu_o)$. A few important points about the statistics are: (i) $\boldsymbol{S}/n$ is the sample covariance matrix of $\bar{\boldsymbol{X}}$, (ii) $\bar{\boldsymbol{X}} \sim N_p(\mu, (1/n)\boldsymbol{\Sigma})$, (iii) $(n-1)\boldsymbol{S} \sim W(n-1, \boldsymbol{\Sigma})$, and (iv) $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ are independent.
4. One should always have $n - 1 > p$, otherwise, $\boldsymbol{S}$ is singular and hence $T^2$ cannot be calculated.

Using the property of sufficiency and the concept of factorization, we obtain the following results, which we state, again without any proofs:

1. If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are the observations from $N_p(\mu, \boldsymbol{\Sigma})$, then $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ are sufficient for $\mu$ and $\boldsymbol{\Sigma}$.

2. The sufficient set of statistics $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ is complete for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, where the sample is drawn from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
3. Let the $m$th component $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdot s$ be $i.i.d.$, with means $E(\boldsymbol{Y}_i) = \boldsymbol{\upsilon}$ (do not confuse with the degree of freedom) and covariance matrices $E(\boldsymbol{Y}_i - \boldsymbol{\upsilon})(\boldsymbol{Y}_i - \boldsymbol{\upsilon})' = \boldsymbol{T}$, then $(1/\sqrt{n})\sum_{i=1}^{n}(Y_i - \upsilon) \rightarrow N(\boldsymbol{0}, \boldsymbol{T})$ as $n \rightarrow \infty$, for $i = 1, \ldots, n$.

Without going into the detailed discussion and proofs, we would like to mention that for different loss functions (even considering the ubiquitous squared error loss), one should be careful to understand what are the best estimates for the mean and the standard deviation in the multivariate case, as it may not always be the sample means or the sample variances we all know so well in the univariate case. Some seminal work in this respect has been done by James and Stein (1961).

## 8.14 Copula Theory

When we talk about correlation coefficient, $\rho(X, Y)$, we generally refer to one of the following: Pearson product–moment correlation coefficient, intraclass correlation, rank correlation, Spearman's rank correlation coefficient, Kendall tau rank correlation coefficient, and Goodman and Kruskal's gamma. For the above definitions, the idea of linear correlation coefficient between two vectors of random variables $X$ and $Y$ is always assumed to be true. Furthermore, the following properties for $\rho(X, Y)$ are also important:

1. $-1 < \rho(X, Y) < 1$, for any range of $X$ and $Y$.
2. If $X$ and $Y$ are independent, then $\rho(X, Y) = 0$.
3. $\rho(\alpha X + \beta, \gamma Y + \delta) = sgn(\alpha\gamma)\rho(X, Y)$, for any range of $X$ and $Y$.

But in general, most random variables are not jointly elliptically distributed (normal is a class of elliptical distributions) and using linear correlation as a measure of dependence in such situations might prove very misleading. An example to illustrate how linear correlation is misused is as follows. Let $X \sim N(0, \sigma^2)$ and let $Y = X^2$. Then it is expected that both $X$ and $Y$ should be correlated, though on calculation we find that $Cov(X, Y) = 0$. Hence, from the discussion, it is obvious that linear correlation coefficient has some shortcomings and here is where copula comes into play. Before we define a copula function, we state its few properties which we think will benefit the readers so that he/she is in a much better position to appreciate the relevance of copula function later on. The properties are:

1. The variances of $X$ and $Y$ need to be finite.
2. Independence of two random variables implies that they are uncorrelated. The reverse is true only in case of MND.
3. Linear correlation is not invariant under nonlinear strictly increasing transformations, $T : \mathbb{R} \rightarrow \mathbb{R}$, since in general, for two real-valued random variables $\rho(T(X), T(Y)) \neq \rho(X, Y)$. Also, the value of linear correlation may change due to the presence of outliers.

One should remember that the linear relationship between two random variables breaks down at the tail. Hence, the concept of tail dependence is very important to understand why $\rho(X, Y)$ does not work at the extremes. This is where the concept of copula comes into play.

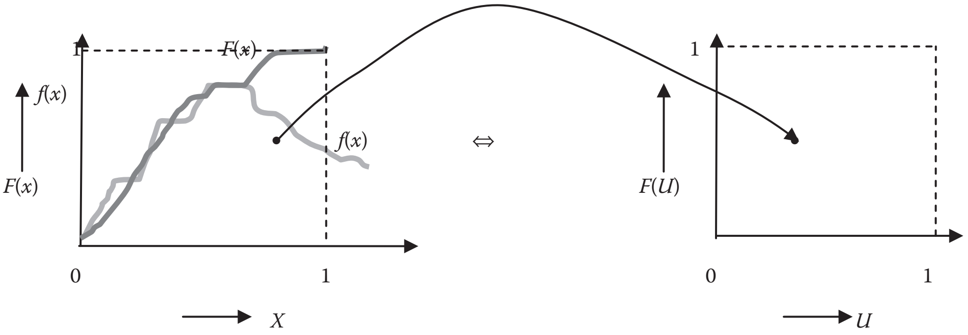


**FIGURE 8.14**
Copula concept using the mapping idea from $\boldsymbol{X}$ space to $\boldsymbol{U}$ space.

Example 8.13 has already given us the basic definition of copula, so rather than repeat the same, we proceed further to give an overview of copula theory and its use in general. The readers should understand that when we use a copula function, we are in a way trying to map from $F_{\boldsymbol{X}}(\boldsymbol{x})$ to $F_{\boldsymbol{U}}(\boldsymbol{u})$, that is, we are mapping from the $\boldsymbol{X}$ space to the unit vector, $\boldsymbol{U}$, space (which is a hypercube of unit dimension on all sides). This may be illustrated for the case when $p = 2$ and is shown in Figure 8.14.

Thus, in general, a copula $C : [0, 1]^p \rightarrow [0, 1]$ has the following properties:

1. $C(U_1, \ldots, U_p)$ is a nondecreasing distribution function in $u_j$, $j = 1, \ldots, p$.
2. $C(1, 1, \ldots, u_j, 1, \ldots, 1) = u_j$ for $j = 1, \ldots, p$ and $u_j \in [0, 1]$ since all marginal distributions of copula are uniformly distributed.
3. For all $(a_1, a_2), (b_1, b_2) \in [0, 1]^2$ with $a_1 \leq b_1$ and $a_2 \leq b_2$, we have $Pr(0 \leq x_1 \leq a_1, 0 \leq x_2 \leq a_2) - Pr(0 \leq x_1 \leq a_1, 0 \leq x_2 \leq b_2) - Pr(0 \leq x_1 \leq b_1, 0 \leq x_2 \leq a_2) + Pr(0 \leq x_1 \leq b_1, 0 \leq x_2 \leq b_2) \geq 0$.

One important concept in copula theory is the Sklar's theorem, which states that if $H_{X_1,\ldots,X_p}(x_1, \ldots, x_p)$ be the joint distribution of $(X_1, \ldots, X_p)$, and $F_{X_1}(x_1), \ldots, F_{X_p}(x_p)$ be the continuous marginal distributions of $(X_1, \ldots, X_p)$, then there exists a copula function $C\left\{F_{X_1}(x_1), \ldots, F_{X_p}(x_p)\right\}$ such that $H_{X_1,\ldots,X_p}\left(x_1, \ldots, x_p\right) = C\left\{F_{X_1}\left(x_1\right), \ldots, F_{X_p}\left(x_p\right)\right\}$. Thus, the distribution function $C(.)$ is in a way a mapping between the marginals and the joint distributions. We must remember that if $F_{X_1}(x_1), \ldots, F_{X_p}(x_p)$ are all continuous, then $C\left\{F_{X_1}(x_1), \ldots, F_{X_p}(x_p)\right\}$ is unique, else it may not be so.

Without going into the detailed concepts, we state a few important properties of copula, which are (i) invariance; (ii) comonotonicity and countermonotonicity; (iii) tail dependence; (iv) upper tail dependence; and (v) lower tail dependence.

To end this discussion about copula theory, we give a few examples of multivariate copula, which are the Gaussian and Student $t$-copula.

*Gaussian copula*: The copula of the $p$-variate normal distribution with linear correlation matrix $\boldsymbol{R}$ is of the following form: $C_R^{Ga}(\boldsymbol{U}) = \phi_R^p\{\phi^{-1}(u_1), \ldots, \phi^{-1}(u_p)\}$, where, $\phi_R^p$ denotes the joint distribution function of $p$-variate standard normal distribution function with linear correlation matrix $\boldsymbol{R}$, while $\phi^{-1}$ denotes the inverse of the distribution function of the univariate standard normal distribution. Another way of writing the Gaussian copula is $C_G(\boldsymbol{U}) = |\boldsymbol{\Sigma}|^{-(1/2)}\ exp(-(1/2)\boldsymbol{q}'\boldsymbol{\Sigma}^{-1}\boldsymbol{q} +$

$(1/2)\boldsymbol{q}'\boldsymbol{q})$. In the bivariate case, the expression takes the form of

$$C_R^{Ga}(\boldsymbol{U}) = \int\limits_{-\infty}^{\phi^{-1}(u)} \int\limits_{-\infty}^{\phi^{-1}(\upsilon)} \frac{1}{2\pi\sqrt{1-R_{12}^2}} exp\left\{-\frac{s^2 - 2R_{12}st + t^2}{2\left(1-R_{12}^2\right)}\right\} ds\, dt.$$

Here, $R_{12}$ is a linear correlation coefficient of the corresponding bivariate normal distribution. Since elliptical distribution is radially symmetric, Gaussian copula does not have either upper or lower tail dependence.

*Student t-copula*: A $p$-dimensional $t$-copula is generally of the form

$$C_t(\boldsymbol{U}) = \int\limits_{-\infty}^{t_\upsilon^{-1}(u_1)} \cdots \int\limits_{-\infty}^{t_\upsilon^{-1}(u_p)} \frac{\Gamma((\upsilon+p)/2)}{\Gamma(\upsilon/2)\sqrt{(\pi\upsilon)^p|\boldsymbol{R}|}} \left(1 + \frac{\boldsymbol{x}'\boldsymbol{R}^{-1}\bar{\boldsymbol{x}}}{\upsilon}\right)^{-((\upsilon+p)/2)} d\boldsymbol{x}.$$

The bivariate $t$-copula is characterized by univariate Student $t$-distribution and is given as

$$C_{\upsilon,R}^{t}(u_1, u_2) = \int\limits_{-\infty}^{t_\upsilon^{-1}(u_1)} \int\limits_{-\infty}^{t_\upsilon^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-R_{12}^2}} exp\left\{1 + \frac{s^2 - 2R_{12}st + t^2}{\upsilon\left(1-R_{12}^2\right)}\right\}^{-((\upsilon+2)/2)} ds\, dt,$$

where $R_{12}$ is a linear correlation coefficient of the corresponding bivariate $t_\upsilon$ distribution if $\upsilon > 2$. A $t$-copula has both upper and lower tail dependences. Before we wind up this section, we mention the names of Cherubini et al. (2004) and Nelsen (2006), which are a few of the good references one may refer to understand copula theory.

## 8.15 Principal Component Analysis

Principal component analysis (PCA) is a multivariate ordination technique used to display patterns in multivariate data. It aims to graphically display the relative positions of data points in fewer dimensions while retaining as much information as possible, and also explore relationships between dependent variables. It is a hypothesis-generating technique that is intended to describe patterns in a data table, rather than test formal statistical hypotheses. PCA assumes linear responses of variables and has a range of applications other than data display, including multiple regression and variable reduction.

As mentioned, the main purpose of PCA is to reduce the dimensionality of multivariate data to make its structure clearer. It does this by looking for the linear combination of the variables, which accounts for as much as possible of the total variation in the data. It then goes on to look for a second combination, uncorrelated with the first, which accounts for as much of the remaining variation as possible and so on. If the greater part of the variation is accounted for by a small number of components, then they may be used in place of the original variables.

The principal idea of PCA is to reduce the dimension of $\boldsymbol{X}_{n\times p} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, in order to find the best combination of $(X_1, \ldots, X_p)$, which is able to give us the maximum information as required. This reduction in dimension may be achieved using linear combinations. Thus, in PCA, one looks for linear combination aimed at creating the so-called largest spread among the variables, $X_1, \ldots, X_p$. This concept of largest spread invariably leads us to look into linear combinations, which have the largest variances. As the reader may be aware that PCA is performed on the covariance matrix, it is not scale invariant, as the units of measurement of $X_1$ or $X_2$ or, $\ldots$, or $X_p$ may be different. Hence, we generally try to use the normalized version of PCA.

The main objective of PCA as mentioned above is to reduce the dimension of the observations, and the simplest way to do that would be to retain one of the variable, say $X_j$, and discard the rest, that is, $X_1, \ldots, X_{j-1}$, $X_{j+1}, \ldots, X_p$. Though the idea may seem plausible, but it is definitely not a reasonable approach as the strength or the ability of explanation is definitely not possible using any arbitrary $X_j$. An alternative plan may be to consider the simple average, that is, $(1/p)\sum_{j=1}^{p} X_j$ of all the elements of $\boldsymbol{X}_{n\times p} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, but this again is not without its drawback as all the elements of $\boldsymbol{X}_{n\times p}$ are considered of equal importance. A more logical intuitive method would be to consider the weighted average $\sum_{j=1}^{p} \delta_j X_j$, given $\sum_{j=1}^{p} \delta_j^2 = 1$, where $\delta = (\delta_1, \ldots, \delta_p)$ is the weighting vector, which needs to be optimized.

Thus, the standard linear combination (SLC), that is, $\sum_{j=1}^{p} \delta_j X_j$, so that $\sum_{j=1}^{p} \delta_j^2 = 1$ should be chosen to *maximize* the variance of the projection of $\sum_{j=1}^{p} \delta_j X_j$.

Hence, we consider the following:

$$\max\left\{Var\left(\sum_{j=1}^{p}\delta_j X_j\right)\right\}$$

$$\text{s.t.}: \sum_{j=1}^{p}\delta_j^2 = 1$$

$$-1 \le \delta_j \le 1, \quad \forall j = 1, \ldots, p.$$

Here, one may easily deduce that the *required direction* of $\delta$ may be found using spectral decomposition of the covariance matrix of $\boldsymbol{X}_{n\times p}$, that is, $\boldsymbol{\Sigma}$. Using basic rules of matrix algebra, we know that the first direction of $\delta$ is given by the eigen vector, $\gamma_1$, corresponding to the largest eigen vector value $\lambda_1$ of the covariance matrix, $\boldsymbol{\Sigma}$. Hence, the first SLC is the one with the highest variance, obtained from the optimization model and is termed as the first *principal component* (PC), that is, $Y_1 = \gamma_1'\boldsymbol{X}$. Once $Y_1$ is found, we proceed to find the second SLC with the second highest variance, that is, the second PC, which is given by $Y_2 = \gamma_2'\boldsymbol{X}$. Diagrammatically, it may be represented as shown in Figure 8.15. In Figure 8.15, let us consider three variables, $X_1$, $X_2$, and $X_3$. Thus, if one uses PCA, then the method would choose the first PCA axis as that line (marked here as PC # 1) that goes through the centroid, but at the same time it also minimizes the square of the distance of each point to that line. PC # 1 thus goes through the maximum variation in the data. If now one finds the second PCA axis (i.e., PC # 2), then this will also pass through the centroid, and would also go through the maximum variation in the data, but with a certain constraint, that it must be completely *uncorrelated* (i.e., at right angles, or *orthogonal*) to PC # 1. Hence, as shown, the angle between PC # 1 and PC # 2 is 90°. In a similar way, one can proceed and obtain PC # 3, such that the angle between PC # 2 and PC # 3 is also 90°.

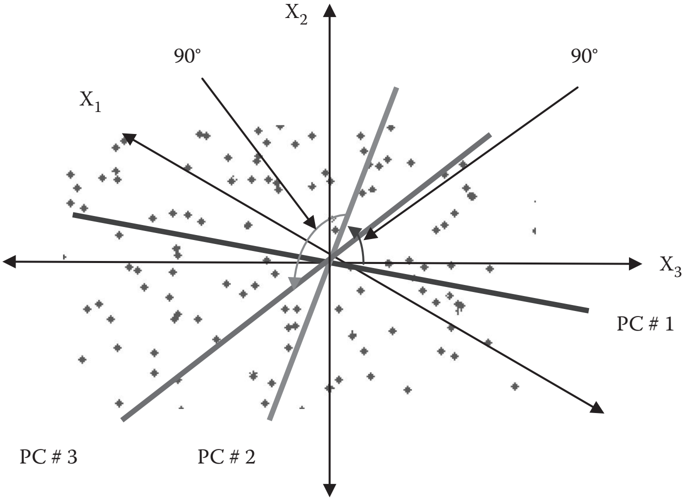


**FIGURE 8.15**
A hypothetical example illustrating the concept of PCA using orthogonality.

**Example 8.16**

Consider the MND $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (2.0, 3.0, 2.5)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 4.00 & -2.00 & 4.00 \\ -2.00 & 9.00 & 3.00 \\ 4.00 & 3.00 & 16.00 \end{pmatrix}.$$

Then the eigen values are $\lambda_1 = 1.6793$, $\lambda_2 = 9.4789$, and $\lambda_3 = 17.8418$, while the corresponding eigen vectors are $\gamma_1 = (0.8719, 0.3697, -0.3210)'$, $\gamma_2 = (0.4311, -0.8905, 0.1452)'$, and $\gamma_3 = (0.2322, 0.2650, 0.9359)'$, respectively. Thus, the PC transformation is given by

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} 0.8719 & 0.3697 & -0.3210 \\ 0.4311 & -0.8905 & 0.1452 \\ 0.2322 & 0.2650 & 0.9359 \end{pmatrix} \begin{pmatrix} X_1 - 2.0 \\ X_2 - 3.0 \\ X_3 - 2.5 \end{pmatrix}.$$

Hence, the PCA axes are:

$$Y_1 = 0.8719X_1 + 0.3697X_2 - 0.3210X_3 - 2.0 \times 0.8719 - 3.0 \times 0.3697 - 2.5 \times (-0.3210)$$

$$Y_2 = 0.4311X_1 - 0.8905X_2 + 0.1452X_3 - 2.0 \times 0.4311 - 3.0 \times (-0.8905) - 2.5 \times 0.1452$$

$$Y_3 = 0.2322X_1 + 0.2650X_2 + 0.9359X_3 - 2.0 \times 0.2322 - 3.0 \times 0.2650 - 2.5 \times 0.9359$$

A way of double checking whether the PC transformations are correct is to calculate $\sum_{j=1}^{3} \delta_j^2$ for each of the eigen values. It is very intuitive to note that each of these values, that is, $\{0.8719^2 + 0.3697^2 + (-0.3210)^2\}$, $\{0.4311^2 + (-0.8905)^2 + 0.1452^2\}$, and $\{0.2322^2 + 0.2650^2 + 0.9359^2\}$ are equal to 1 as the case should be. Another method to double check is to find the variances of $\boldsymbol{Y}$.

Thus, we have $Var(Y_1) = 0.8719^2 \times Var(X_1) + 0.36976^2\, Var(X_2) + (-0.3210)^2\, Var(X_3) + 2 \times 0.8719 \times 0.36976 \times Covar(X_1, X_2) + 2 \times 0.36976 \times (-0.3210) \times Covar(X_2, X_3) + 2 \times$

$0.8719 \times (-0.3210) \times Covar(X_1,\ X_3) = 1.6792$, which is the value of the first eigen vector as calculated above. Similarly, $Var(Y_2) = 9.4789$ and $Var(Y_3) = 17.8418$.

Though not exhaustive, yet we state a few important results for PCA:

1. For a given $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, let $\boldsymbol{Y} = \Gamma'(\boldsymbol{X} - \boldsymbol{\mu})$ be the PC transformation, then
   a. $E(Y_j) = 0, j = 1, \ldots, p$.
   b. $Var(Y_j) = \lambda_j, j = 1, \ldots, p$.
   c. $Covar\left(Y_{j_1}, Y_{j_2}\right) = 0, j_1 \neq j_2 = 1, \ldots, p$.
   d. $\sum_{j=1}^{p} Var(Y_j) = tr(\boldsymbol{\Sigma})$.
   e. $\prod_{j=1}^{p} Var(Y_j) = |\boldsymbol{\Sigma}|$.
2. There exists no SLC which has larger variance than $\lambda_1 = Var(Y_1)$.
3. For the practical implementation of PCA, we replace $\boldsymbol{\mu}$ by $\bar{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}$ by $\boldsymbol{S}$, and we evaluate the eigen values and the eigen vectors of $\boldsymbol{S}$.
4. The components of the eigen vectors are the weights of the original variables in the PC.
5. PCs are not scale invariant.

A few good references for PCA are Hastie et al. (2011), Jackson (2003), and Jolliffe (2002).

## 8.16 Factor Analysis

The origins of FA may be traced back to the work of Pearson (1901) and Spearman (1904). The term “FA” as we know today was first introduced by Thurstone (1931). It is a multivariate statistical method based on a model when the observed vector is partitioned into an *unobserved systematic* part and an *unobserved error* part. The components of the error vector are considered as uncorrelated or independent, while the systematic part is taken as a linear combination of a relatively smaller number of unobserved factor variables. Using FA, one can separate the effects of the factors (which are of primary interest to us) from the errors. Stated explicitly using FA, we intend to partition variables into particular groups such that within a particular group they are highly correlated among themselves. Moreover, these variables have relatively small correlations with variables in a different group. Thus, each group of variables represents a single underlying construct/factor that is responsible to provide information about the observed correlations. Before we go in to the mathematical discussion of FA and how it is used, we state here a few good references for FA: Anderson (2003), Basilevsky (1994), Child (2006), Gorsuch (1983), Harman (1976), Johnson and Wichern (2002), Lawley and Maxwell (1971), Mulaik (2009), Thompson (2004), and Thurstone (1931, 1947).

### 8.16.1 Mathematical Formulation of Factor Analysis

Suppose $\boldsymbol{X}_{(p\times 1)}$ be $p$ number of variables, such that $E(\boldsymbol{X}) = \boldsymbol{\mu}_{(p\times 1)}$ and $Covar(\boldsymbol{X}) = \boldsymbol{\Sigma}_{(p\times p)}$, then using FA, one can express $\boldsymbol{X}_{(p\times 1)}$ as being dependent on

$$\boldsymbol{F}_{(m\times 1)} = \begin{pmatrix} F_1 \\ \vdots \\ F_m \end{pmatrix}$$

common factors and $p$ additional specific factors denoted by $\varepsilon_{(p\times 1)}$. Mathematically, it can be expressed as $\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{LF} + \boldsymbol{\varepsilon}$, that is,

$$\begin{bmatrix} X_1 - \mu_1 \\ \cdot \\ \cdot \\ X_p - \mu_p \end{bmatrix} = \begin{bmatrix} l_{1,1}F_1 + \cdots + l_{1,m}F_m + \varepsilon_1 \\ \cdot \\ \cdot \\ l_{p,1}F_1 + \cdots + l_{p,m}F_m + \varepsilon_p \end{bmatrix},$$

where

$$\boldsymbol{L}_{(p\times m)} = \begin{pmatrix} l_{1,1} & \cdots & l_{1,m} \\ \vdots & \ddots & \vdots \\ l_{p,1} & \cdots & l_{p,m} \end{pmatrix}$$

is the matrix of *factor* loading.

A careful look at this mathematical formulation, $\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{LF} + \boldsymbol{\varepsilon}$, would distinguish this from MLR due to the fact that in MLR the independent variables can be observed while in FA it is not so. One also needs to make a distinction between FA and PCA. In the PCA method, the PCs are just the linear transformation arranged in the sense that the variance corresponding to the PCs *decreases* as one goes from the first PC to the second and so on. In doing so, the dimension of the data set is reduced, and this as we know is the main idea of PCA. On the other hand, in FA, one aims to model the variations using a linear transformation of a *fixed* number of variables, called the *factor* or the *latent* variables.

A few important properties/assumptions for FA, $(m < p)$, are as follows:

1.

$$E(\boldsymbol{F}) = \boldsymbol{0}_{(m\times 1)} = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}_{(m\times 1)}.$$

2.

$$Covar(\boldsymbol{F}) = \boldsymbol{I}_{(m\times m)} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix}_{(\boldsymbol{m}\times m)}.$$

3.

$$E(\boldsymbol{\varepsilon}) = \boldsymbol{0}_{(p\times 1)} = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}_{(p\times 1)}.$$

4.

$$Covar(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}_{(p\times p)} = \begin{bmatrix} \Psi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Psi_p \end{bmatrix}_{(p\times p)}.$$

5.

$$Covar(\varepsilon, \boldsymbol{F}) = \boldsymbol{0}_{(p\times m)} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}_{(p\times m)}.$$

6.

$$\boldsymbol{\Sigma} = \boldsymbol{L}_{(p\times m)}\boldsymbol{L}'_{(m\times p)} + \boldsymbol{\Psi}_{(p\times p)} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{p1} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix}_{(p\times p)},$$

where $\sigma_{jj} = \left(l_{j,1}^2 + \cdots + l_{j,m}^2\right) + \Psi_j$, $j = 1, \ldots, p$. Here, $\left(l_{j,1}^2 + \cdots + l_{j,m}^2\right)$ is called the communality, while $\Psi_j$ is the specific variance.

7. The eigen value and eigen vector for $\sum$ are $(\lambda_j, \boldsymbol{e}_j)$, $j = 1, \ldots, p$ such that $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$.

8.

$$Covar(\boldsymbol{X}, \boldsymbol{F}) = \boldsymbol{L}_{(p\times m)} = \begin{pmatrix} l_{1,1} & \cdots & l_{1,m} \\ \vdots & \ddots & \vdots \\ l_{p,1} & \cdots & l_{p,m} \end{pmatrix}_{(p\times m)}.$$

9. Factor loadings, that is, $\boldsymbol{L}_{(p\times m)}$, are determined only up to an orthogonal matrix $\boldsymbol{T}_{(m\times m)}$. Thus, the loadings $\boldsymbol{L}^*_{(p\times m)} = \boldsymbol{L}_{(p\times m)}\boldsymbol{T}_{(m\times m)}$ and $\boldsymbol{L}$ both give the same representations. Furthermore, the communalities given by the diagonal elements of $\boldsymbol{L}_{(p\times m)}\boldsymbol{L}'_{(m\times p)}$ and $\boldsymbol{L}^*_{(p\times m)}\boldsymbol{L}^{*\prime}_{(m\times p)}$ are also unaffected by the choice of $\boldsymbol{T}_{(m\times m)}$.

The properties/assumptions stated above constitute the orthogonal factor model. When $m \ll p$, then FA as a method is very useful. On the other hand, if the off-diagonal elements of $\boldsymbol{S}(\boldsymbol{R})$ are small (zero), then the variables are not related and FA as a multivariate statistical technique is not useful. If we allow the $\boldsymbol{F}_{(m\times 1)}$ common factors to be correlated such that $Covar(\boldsymbol{F}) \neq \boldsymbol{I}_{(m\times m)}$, then we obtain the oblique factor model.

### 8.16.2 Estimation in Factor Analysis

In statistical literature, we have two methods for estimating the parameters in FA: (i) principal component method and (ii) maximum likelihood method.

### 8.16.3 Principal Component Method

1. For the population, we know

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & 0 & 0 & 0 \\ 0 & \sigma_{22} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_{pp} \end{bmatrix}_{(p\times p)} = \boldsymbol{L}_{(p\times m)}\boldsymbol{L}'_{(m\times p)} + \boldsymbol{\Psi}_{(p\times p)}$$

$$= \left[\sqrt{\lambda_1}\boldsymbol{e}_1\sqrt{\lambda_2}\boldsymbol{e}_2\cdots\sqrt{\lambda_m}\boldsymbol{e}_m\right]_{(p\times m)} \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e}'_1 \\ \sqrt{\lambda_2}\boldsymbol{e}'_2 \\ \vdots \\ \sqrt{\lambda_m}\boldsymbol{e}'_m \end{bmatrix}_{(m\times p)} + \begin{bmatrix} \Psi_1 & 0 & 0 & 0 \\ 0 & \Psi_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \Psi_p \end{bmatrix}_{(p\times p)}$$

$$= \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_m} \\ \vdots & \vdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 \end{bmatrix}_{(p\times m)} \times \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \sqrt{\lambda_m} & 0 & 0 \end{bmatrix}_{(m\times p)}$$

$$+ \begin{bmatrix} \Psi_1 & 0 & 0 & 0 \\ 0 & \Psi_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \Psi_p \end{bmatrix}_{(p\times p)}$$

$$= \begin{bmatrix} (\lambda_1+\Psi_1) & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & (\lambda_2+\Psi_2) & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & (\lambda_m+\Psi_m) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \Psi_p \end{bmatrix}_{(p\times p)},$$

such that $\sigma_{jj} = \sum_{i=1}^{m} l_{ji}^2 + \Psi_j$, $j = 1, \ldots, p$. We assume the contribution of $(\lambda_{m+1}\boldsymbol{e}_{m+1}\boldsymbol{e}'_{m+1} + \cdots + \lambda_p\,\boldsymbol{e}_p\boldsymbol{e}'_p)$ is negligible.

2. We can also use

$$\boldsymbol{x}_i - \bar{\boldsymbol{x}} = \begin{bmatrix} x_{i,1} - \bar{x}_1 \\ \vdots \\ x_{i,p} - \bar{x}_p \end{bmatrix}, \quad \text{or} \quad \boldsymbol{z}_i = \begin{bmatrix} \dfrac{(x_{i,1} - \bar{x}_1)}{\sqrt{s_{11}}} \\ \vdots \\ \dfrac{(x_{i,p} - \bar{x}_p)}{\sqrt{s_{pp}}} \end{bmatrix},$$

for $i = 1, \ldots, n$, the latter being used for the case to avoid problems of having one variable with large variance unduly affecting the factor loading.

3. In case we have a sample, then the eigen vector and eigen pair for $\mathbf{S}$ will be $(\hat{\lambda}_j, \hat{\boldsymbol{e}}_j), \ldots, p$ such that $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$. In case $m < p$, then

$$\mathbf{S} = \begin{bmatrix} s_{11} & 0 & 0 & 0 \\ 0 & s_{22} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & s_{pp} \end{bmatrix}_{(p \times p)} = \hat{\boldsymbol{L}}_{(p \times m)} \hat{\boldsymbol{L}}'_{(m \times p)} + \hat{\boldsymbol{\Psi}}_{(p \times p)}$$

   such that $s_{jj} = \sum_{i=1}^{m} \hat{l}_{ji}^2 + \hat{\Psi}_j$, $j = 1, \ldots, p$.
4. If the number of common factors is not known beforehand or is not determined *a priori*, one can use the knowledge of previous researchers. A thumb rule is to find the value of the residual matrix, that is, $\boldsymbol{S} - \left(\hat{\boldsymbol{L}}_{(p \times m)} \hat{\boldsymbol{L}}'_{(m \times p)} + \hat{\boldsymbol{\Psi}}_{(p \times p)}\right) \leq \hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2$. A small value of the sum of squares of the neglected eigen values means a small value for the sum of square errors of approximation.
5. The contribution to the total sample variance, that is, $s_{11} + \cdots + s_{pp}$ from the $i$th common factor is given by $\hat{l}_{i,1}^2 + \cdots + \hat{l}_{i,p}^2 = \left(\sqrt{\hat{\lambda}_i} \hat{\boldsymbol{e}}_i\right)' \left(\sqrt{\hat{\lambda}_i} \hat{\boldsymbol{e}}_i\right) = \hat{\lambda}_i$. Thus, the proportion of the total sample variance due to the $j$th factor is given by $\hat{\lambda}_j / (s_{11} + \cdots + s_{pp})$ or $(\sum_{j=1}^{i} \hat{\lambda}_j)/p$ depending on whether it is to do with $\boldsymbol{S}$ or $\boldsymbol{R}$.

   A modified approach called the principal factor method works in a similar method as stated above.

### 8.16.4 Maximum Likelihood Method

For the maximum likelihood methodology to work, a few important assumptions should hold:

1. $\boldsymbol{F}$ and $\varepsilon$ are normally distributed
2. If the first holds true, then $\boldsymbol{X}_j - \mu = \boldsymbol{LF}_j + \varepsilon_j$ is also normally distributed
3. The likelihood function is of the form

$$L(\mu, \boldsymbol{\Sigma}) = (2\pi)^{-((n-1)p)/2} |\boldsymbol{\Sigma}|^{-(n-1)/2} e^{-(1/2)tr\left[\boldsymbol{\Sigma}^{-1}\{\sum_{i=1}^{n}(\mathbf{x}_{\mathrm{i}}-\bar{\mathbf{x}})(\mathbf{x}_{\mathrm{i}}-\bar{\mathbf{x}})'\}\right]}.$$

### 8.16.5 General Working Principle for FA

The general plan based on which one can use FA is as follows:

1. Generate a variance–covariance matrix of the observed variables, that is, $\boldsymbol{S}$, which is an estimate of $\boldsymbol{\Sigma}$.
2. Select the number of factors, that is, $m$, by first finding $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{e}}$, which are the respective estimates from the sample of size $n$. In general, find $\hat{\lambda}_j / (s_{11} + \cdots + s_{pp})$ or $(\sum_{j=1}^{i} \hat{\lambda}_j)/p$ for those $\hat{\lambda}_j s$ which are greater than 1.
3. Extract your initial set of factors, that is, find $F_1, \ldots, F_m, \hat{h}_i^2, \hat{\Psi}_i^2$.
4. Perform factor rotation to a terminal solution.
5. Interpret the factor structure, that is, $\boldsymbol{S} - (\hat{\boldsymbol{L}}_{(p \times m)} \hat{\boldsymbol{L}}_{(m \times p)} + \hat{\boldsymbol{\Psi}}_{(p \times p)})$.

```
DEFINE: m, $\hat{\boldsymbol{\lambda}}$, $\hat{\boldsymbol{e}}$, $\boldsymbol{F}_{(m\times 1)} = \begin{pmatrix} F_1 \\ \vdots \\ F_m \end{pmatrix}$, $\boldsymbol{L}_{(p\times m)} = \begin{pmatrix} l_{1,1} & \cdots & l_{1,m} \\ \vdots & \ddots & \vdots \\ l_{p,1} & \cdots & l_{p,m} \end{pmatrix}$
INPUT: $\boldsymbol{X}_{n\times p}$
CALCULATE: $E(\boldsymbol{X}) = \hat{\boldsymbol{\mu}}_{(\mathrm{p}\times 1)}$, $Covar(\boldsymbol{X}) = \hat{\boldsymbol{\Sigma}}_{(\mathrm{p}\times\mathrm{p})} = \boldsymbol{S}_{(\mathrm{p}\times\mathrm{p})}$, $\hat{\boldsymbol{\lambda}}$, $\hat{\boldsymbol{e}}$, $\frac{\hat{\lambda}_j}{s_{11}+\cdots+s_{pp}}$, $\frac{\sum_{j=1}^{i}\hat{\lambda}_j}{p}$ and also max value of $\frac{\sum_{j=1}^{i}\hat{\lambda}_j}{p}$ for some fixed $j$
START if: $i = 1, \ldots, m$
    CALCULATE: (i) $\hat{\lambda}_i$ which is the $i^{th}$ eigen value of $\boldsymbol{S}_{(\mathrm{p}\times\mathrm{p})}$ and $\hat{\boldsymbol{e}}_i$ which is its corresponding $i^{th}$ eigen vector, (ii) in calculating these $i^{th}$ values, also find out $\hat{l}_{i,j} = \sqrt{\hat{\lambda}_{i,j}}\hat{\boldsymbol{e}}_{i,j}$, $j = 1,\cdots,p$, $i = 1,\cdots,m$, $\hat{h}_i^2 = \hat{l}_{i,1}^2 + \cdots + \hat{l}_{i,m}^2$ and $\hat{\Psi}_i^2 = 1 - \hat{h}_i^2$, $i = 1,\cdots,p$
END if
REPORT: $m$, $\boldsymbol{F}_{(m\times 1)}$, $\boldsymbol{L}_{(p\times m)}$ based on max value of $\frac{\sum_{j=1}^{i}\hat{\lambda}_j}{p}$
END
```

**FIGURE 8.16**
Pseudo-code to implement FA method.

6. Construct factor scores to use it in further analyses, that is,

$$\begin{bmatrix} l_{1,1}F_1 + \cdots + l_{1,m}F_m + \varepsilon_1 \\ \cdot \\ \cdot \\ l_{p,1}F_1 + \cdots + l_{p,m}F_m + \varepsilon_p \end{bmatrix}.$$

For the convenience of the reader, the general pseudo-code based on which one can work on a data set from the point of view of analyzing its use using FA is given in Figure 8.16.

**Example 8.17**

As an example, let us consider the data set taken from Holzinger and Swineford (1939). The brief background of the study is as follows. Twenty-six tests, intended to measure a general factor and five specific factors, were administered to seventh and eighth grade students in two schools, namely, Grant-White School ($n = 145$) and Pasteur School ($n = 156$). Students from the Grant-White School came from homes where the parents were American-born, while those from the Pasteur School were from homes where the parents were foreign-born. Data for the analysis include 19 tests intended to measure four domains, namely, (i) spatial ability (visual perception test, cubes, paper form board, lozenges), (ii) verbal ability (general information, paragraph comprehension, sentence completion, word classification, word meaning), (iii) speed (add, code, counting groups of dots, straight and curved capitals), and (iv) memory (word recognition, number recognition, figure recognition, object-number, number-figure, figure-word). For the FA study, consider 24, that is, $p = 24$, psychological tests are administered to the first group of students, that is, $n = 145$. Let us start with $m = 5$, such that we obtain Table 8.2, which has all the relevant information such as estimated factor loading, communalities, specific variances, etc. The method used is the principal component method.

Now, $\boldsymbol{S} - (\hat{\boldsymbol{L}}_{(p\times m)}\hat{\boldsymbol{L}}_{(m\times p)} + \hat{\boldsymbol{\Psi}}_{(p\times p)})$ will give the variability in the sample variance one is not able to explain using $m = 5$. In case $m = 15$, then the cumulative proportion goes up to 0.8879 from a value of 0.6021. One can also use the maximum likelihood method to get the solution, and we request the readers to solve this problem on their own to get a good understanding of the maximum likelihood method used in FA.

**TABLE 8.2**

FA Solution Using Data from Holzinger and Swineford (1939), Considering $m = 5$, Example 8.17

| | **Estimated Factor Loading $\hat{l}_{i,j} = \sqrt{\hat{\lambda}_{i,j}} e_{i,j}$ for** | | | | | | |
|---|---|---|---|---|---|---|---|
| **Variables** | $F_1$ | $F_2$ | $F_3$ | $F_4$ | $F_5$ | $\hat{h}_i^2$ | $\hat{\Psi}_i^2 = 1 - \hat{h}_i^2$ |
| Visual perception | 0.616 | 0.005 | 0.428 | −0.205 | 0.009 | 0.6048 | 0.3952 |
| Cubes | 0.400 | 0.079 | 0.400 | −0.202 | −0.348 | 0.4881 | 0.5119 |
| Paper form board | 0.445 | 0.191 | 0.476 | −0.106 | 0.375 | 0.6129 | 0.3871 |
| Flags | 0.511 | 0.178 | 0.335 | −0.216 | 0.010 | 0.4518 | 0.5482 |
| General information | 0.695 | 0.321 | −0.335 | −0.053 | −0.079 | 0.7073 | 0.2927 |
| Paragraph comprehension | 0.690 | 0.418 | −0.265 | 0.081 | 0.008 | 0.7277 | 0.2723 |
| Sentence completion | 0.677 | 0.425 | −0.355 | −0.073 | 0.041 | 0.7720 | 0.228 |
| Word classification | 0.694 | 0.243 | −0.144 | −0.116 | 0.141 | 0.5948 | 0.4052 |
| Word meaning | 0.694 | 0.451 | −0.291 | 0.080 | 0.005 | 0.7761 | 0.2239 |
| Addition | 0.474 | −0.542 | −0.446 | −0.202 | −0.079 | 0.7644 | 0.2356 |
| Code | 0.576 | −0.434 | −0.210 | 0.034 | −0.003 | 0.5654 | 0.4346 |
| Counting dots | 0.482 | −0.549 | −0.127 | −0.340 | −0.099 | 0.6753 | 0.3247 |
| Straight curved capitals | 0.618 | −0.279 | 0.035 | −0.366 | 0.075 | 0.6006 | 0.3994 |
| Word recognition | 0.448 | −0.093 | −0.055 | 0.555 | −0.156 | 0.5447 | 0.4553 |
| Number recognition | 0.416 | −0.142 | 0.078 | 0.526 | −0.306 | 0.5696 | 0.4304 |
| Figure recognition | 0.534 | −0.091 | 0.392 | 0.327 | −0.171 | 0.5833 | 0.4167 |
| Object-number | 0.488 | −0.276 | −0.052 | 0.469 | 0.255 | 0.6020 | 0.3980 |
| Number-figure | 0.544 | −0.386 | 0.198 | 0.152 | 0.104 | 0.5181 | 0.4819 |
| Figure-word | 0.476 | −0.138 | 0.122 | 0.193 | 0.605 | 0.6638 | 0.3362 |
| Deduction | 0.643 | 0.186 | 0.132 | 0.070 | −0.285 | 0.5516 | 0.4484 |
| Numerical puzzles | 0.622 | −0.232 | 0.100 | −0.202 | −0.174 | 0.5218 | 0.4782 |
| Problem reasoning | 0.640 | 0.146 | 0.110 | 0.056 | 0.023 | 0.4467 | 0.5533 |
| Series completion | 0.712 | 0.105 | 0.150 | −0.103 | −0.064 | 0.5552 | 0.4448 |
| Arithmetic problems | 0.673 | −0.196 | −0.233 | −0.062 | 0.097 | 0.5589 | 0.4411 |
| Eigen values | 8.1354 | 2.096 | 1.6926 | 1.5018 | 1.0252 | | |
| Cumulative proportion of total sample variance | 0.3390 | 0.4263 | 0.4968 | 0.5594 | 0.6021 | | |

*Data set # 1 for FA:* Consider the data from the stock market, which one can obtain from http://in.finance.yahoo.com/. The data is related to the daily closing prices of Hang Seng Index of Hong Kong stock market (http://www.hkex.com.hk/eng/ index.htm). The number of stocks in Hang Seng is 48 and the time frame of our analysis is 2013–2015 or any appropriate time frame as appropriate. Our aim is to study the effect of a few fundamental financial ratios such as current ratio, cash ratio, return of assets, debt ratio, sales to revenue per employee, dividend payout ratio, price-to-book-value ratio, and price-to-sales ratio on the performance of the company and hence on the stock market index of that particular company. If we consider the returns, $r = log_e(P_t/P_{t-1})$, then one can use the concept of FA to study the effects of these ratios of a particular company on the stock market index of that company itself in more details.

*Data set # 2 for FA:* Let us consider the study performed by Linden (1977), which consists of the performance of $n = 160$ athletes. The $p = 10$ variables are: 100-m run, long jump, shot put, high jump, 400-m run, 110-m run, discus, pole vault, javelin, and 1500-m run.

One can access the data and study the analysis, which can be found in Basilevsky (1994). We urge the readers to work with this data set to gain a better appreciation of FA, which is an interesting multivariate statistical method.

*Data set # 3 for FA:* The information and background for the third data set can be found in Davis et al. (1997). The study deals with eating disorders and pertains to $n = 191$ individuals with respect to $p = 7$ variables. The data can be found in http://www.unt.edu/rss/class/mike/data/DavisThin.txt. The reader can study the analysis to appreciate the use of FA in a variety of fields, be it sociology or finance.

## 8.17 Multiple Analysis of Variance and Multiple Analysis of Covariance

### 8.17.1 Introduction to Analysis of Variance

Consider as a doctor you are interested to analyze the effect of both food habits as well as exercise regime/physical activity on a group of people who are your patients. The weights of these patients are known before the start of this experiment. Your emphasis is to study the reduction of individual weights, which is considered as the response variable. Each of the group of patients undergoes a particular food habit as well as an exercise regime/physical activity. Consider you have $I(i = 1, \ldots, I)$ number of patients in each group, $J(j = 1, \ldots, J)$ as the number of different food habits such as vegetarian, nonvegetarian, vegan, etc., and $K(k = 1, \ldots, K)$ as the number of different exercise regime/physical activities, such as weightlifting, yoga, aerobics, etc. In order to study the effects of food habit, exercise regime/physical activity on the reduction of individual weights, we may use the concept of analysis of variance (ANOVA). Formally, this method studies the total variation present in a set of observations, which is measured by the sum of squares of deviations of the observations from the mean, that is, *SS*. This deviation is partitioned into components associated with assignable effect due to fixed and/or random effects/unassignable effect due to residual random effect. The technique of ANOVA also provides the means for the systematic study of regression analysis and correlation coefficients. A few assumptions/properties that are relevant to ANOVA are:

1. The responses are independently and normally distributed, that is, $X_{i,j,k} \sim N(\mu_{j,k}, \sigma^2), i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$, with *constant* variances (property of homoscedasticity) so that the only difference between the distributions of observations is the means, which is denoted by $\mu_{ij}$.
2. If the number of observations in each group is *equal*, then the ANOVA model is termed as the *balanced* model, else it is an *unbalanced* model.
3. If the *assignable* effects are all fixed, then we have the *fixed-effects model*; otherwise, it is the *random-effects model* where the effects are random except for the additive constants. Note that a *mixed-effects model* contains effects of both fixed- and random-effects types.
4. We say we have a *two-way crossed* ANOVA model if we can categorize the observations in two ways, that is, categorizing in every possible food habit and exercise regime/physical activity pair. In case we have only $I$ and $J$, then it is the *one-way* ANOVA model.
5. An ANOVA model is *additive* if one can express $\mu_{ij} = \gamma_i + \tau_j$, that is, the effect due to particular food habit and exercise regime/physical activity is the sum of the effect due to the food habit and an effect due to exercise regime/physical activity taken separately. In the additive model, the general hypotheses one is interested to study are that $\mu_{ij}s$are neither

dependent on $i$ nor $j$. In a *nonadditive* model, apart from the two hypotheses, we have a third one, which is that the model is additive (i.e., $\mu_{ij} = \gamma_i + \tau_j$).

6. When $\mu_{ij} = \mu_{ji}$, then we have the *symmetric* model; else it is the *asymmetric* model.
7. In case any exercise regime/physical activity occurs in only one type of exercise regime/physical activity, then we have the *twofold nested* ANOVA model; else it is the *twofold nonnested* model, an example of which is the one described above.

To continue with our discussion further, consider the general linear model $\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$, such that $E(\epsilon) = \boldsymbol{0}$ and $Covar(\epsilon) = \sigma^2\boldsymbol{I}$, which implies that $E(\boldsymbol{Y}) = \boldsymbol{X}\beta$ and $Covar(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$. In case this is true, then a few relevant results for ANOVA are as stated below:

1. A linear estimator $\boldsymbol{c} + \boldsymbol{a}'\boldsymbol{Y}$ is unbiased for $\lambda'\beta$ iff $E(\boldsymbol{c} + \boldsymbol{a}'\boldsymbol{Y}) = \lambda'\beta$ for all $\beta$.
2. When $\lambda'\beta$ is estimable, then it is possible to find several estimators that are unbiased for $\lambda'\beta$ and the OLS estimator $\lambda'\hat{\beta}$ is also the best linear unbiased estimator (BLUE).
3. When $\hat{\beta}$ is any solution to the normal equations $\boldsymbol{X}'\boldsymbol{X}\beta = \boldsymbol{X}\boldsymbol{Y}$, then it is unbiased for $\lambda'\beta$.
4. Consider $\boldsymbol{\Lambda}'\beta$ is any $d$-dimensional estimable vector and $\boldsymbol{c} = \boldsymbol{A}'\boldsymbol{Y}$ is any vector of linear unbiased estimators of $\boldsymbol{\Lambda}'\beta$, then if $\hat{\beta}$ denotes any solution to the normal equations, then the matrix $Covar(\boldsymbol{c} + \boldsymbol{A}'\boldsymbol{Y}) - Covar(\lambda'\hat{\beta})$ is nonnegative definite.

It is generally acknowledged that Fisher (1921) developed the technique of ANOVA. A few relevant references in this area (along with scope of applications in social sciences and other areas) are: Hoaglin et al. (1991), Iversen and Norpoth (1987), Krishnaiah (1984), Lewis (1971), Rutherford (2001), Sahai and Ageel (2000), Scheffé (1999), Searle et al. (2009), Stuart et al. (1999), and Turner and Thayer (2001).

### 8.17.2 Multiple Analysis of Variance

With this short background about ANOVA, we come to the area of multianalysis of variance (MANOVA), which as a technique determines the effects of independent categorical variables on multiple continuous-dependent variables. It is usually used to compare several groups with respect to multiple continuous variables, as it tests for differences between centroids, that is, the vectors of the mean values of the dependent variables. On the other hand, one should remember that ANOVA tests for intergroup differences between the mean values of dependent variables. Before we discuss the methodology, we would like to stress the advantages of MANOVA over its univariate counterpart, which is ANOVA:

1. MANOVA can protect against Type I error that occurs if multiple ANOVAs are carried.
2. By measuring several dependent variables in a single experiment, there is a better chance of finding out which variables are important.
3. MANOVA is sensitive not only to mean difference but also to the direction and size of correlations among the dependent variables.

On the other hand, there are also some disadvantages of MANOVA when compared with ANOVA:

1. To do with loss of degrees of freedom, as we know that one degree of freedom is lost for each dependent variable. Hence, the gain of power obtained from the decrease in the sum of squares may be offset due to this loss in the degrees of freedom.

2. High level of dependence of variables does not give us a good picture of the data, so it may be prudent to use the ANOVA model instead.

Before we go into the methodology, we state the general assumptions for MANOVA, which though intuitive is important to understand as it sets the tone about the efficacy of MANOVA as a statistical method when used as a data analysis tool. The assumptions are:

1. The dependent variable should be normally distributed within groups.
2. There should be linear relationships among all pairs (i) of dependent variables, (ii) of covariates, and (iii) of dependent variables–covariables in each cell.
3. The dependent variables should exhibit equal levels of variance across the range of predictor variables. This property is termed as homogeneity.
4. Intercorrelations (covariance) should be homogeneous across the cells of the design.

*MANOVA model*: Let us consider the two-way MANOVA model of the form

$$\begin{pmatrix} Y_{ij1} \\ \vdots \\ Y_{ijp} \end{pmatrix}_{p\times 1} = \begin{pmatrix} v_1 \\ \vdots \\ v_p \end{pmatrix}_{p\times 1} + \begin{pmatrix} \alpha_{i1} \\ \vdots \\ \alpha_{ip} \end{pmatrix}_{p\times 1} + \begin{pmatrix} \beta_{j1} \\ \vdots \\ \beta_{jp} \end{pmatrix}_{p\times 1} + \begin{pmatrix} \varepsilon_{ij1} \\ \vdots \\ \varepsilon_{ijp} \end{pmatrix}_{p\times 1}.$$

Here, $Y_{ijk}$ is the observation corresponding to the $i$th treatment, $j$th block, and $k$th variable; $v_k$ is the overall mean for the $k$th variable; $\alpha_{ik}$ is the effect of the $i$th treatment on the $k$th variable; $\beta_{jk}$ is the effect of the $j$th block on the $k$th variable; and finally, $\varepsilon_{ijk}$ is the experimental error for the $i$th treatment, $j$th block, and $k$th variable. The relevant assumptions for the MANOVA model are the ones which one can refer in any good book in multivariate statistics.

Since this model assumes no interaction, we use this error to test the block and treatment effects. Thus, one can define the mean vector for a treatment $i$ as $\boldsymbol{\mu}_i = \boldsymbol{v} + \boldsymbol{\alpha}_i$. In case the null hypothesis states that all of the treatment mean vectors are identical, then we have $H_O{:}\mu_1 = \cdots = \mu_g$, or equivalently $\boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_g$, where $i = 1, \ldots, g$ are the number of treatments. The alternative hypothesis is $H_A{:}\mu_{ik} \neq \mu_{jk}$ for at least one $i \neq j$ and at least one variable $k$.

We now define the sample mean vector for treatment $i$ and block $j$ as

$$\begin{pmatrix} Y_{i\cdot 1} = \frac{1}{b}\sum_{j=1}^{b} Y_{ij1} \\ \vdots \\ Y_{i\cdot p} = \frac{1}{b}\sum_{j=1}^{b} Y_{ijp} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} Y_{\cdot j1} = \frac{1}{a}\sum_{i=1}^{a} Y_{ij1} \\ \vdots \\ Y_{\cdot jp} = \frac{1}{a}\sum_{i=1}^{a} Y_{ijp} \end{pmatrix},$$

respectively, while the grand mean vector is

$$\begin{pmatrix} Y_{i\cdot 1} = \frac{1}{a\times b}\sum_{j=1}^{b}\sum_{i=1}^{a} Y_{ij1} \\ \vdots \\ Y_{i\cdot p} = \frac{1}{a\times b}\sum_{j=1}^{b}\sum_{i=1}^{a} Y_{ijp} \end{pmatrix}.$$

Furthermore, let us also define the total sum of squares and cross products matrix as $T = b\sum_{i=1}^{a}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})' + a\sum_{j=1}^{b}(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})' + \sum_{i=1}^{a}\sum_{j=1}^{b}(Y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})$ $(Y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})'$, where the first, second, and the third terms are the *treatment* sum of squares and cross products matrix, *block* sum of squares and cross products matrix, and finally *error* sum of squares and cross products matrix, respectively. The individual element of the *treatment*/*block*/*error* sum of squares and cross products matrices can be expressed accordingly.

Finally, to close this section, we give a few of the statistics that are used in the study of MANOVA, but before that, let us define $A = SS_{Hypothesis}/SS_{Error}$, where SS means the sum of square. Now, based on $\boldsymbol{A}$, and the fact that $\lambda_i$ denotes the $i$th eigen value of the matrix $\boldsymbol{A}$, the relevant statistics for MANOVA are as follows:

1. Wilk's lambda: $\prod_{i=1}^{q} 1/(1+\lambda_i)$, where $q$ denotes the dependent variables in MANOVA study.
2. Pillai's trace: $\sum_{i=1}^{q} \lambda_i/(1+\lambda_i)$, where $q$ denotes the dependent variables in MANOVA study.
3. Lawley–Hotelling trace: $\sum_{i=1}^{q} \lambda_i$, where $q$ denotes the dependent variables in MANOVA study.
4. Roy's largest root: $\max_{i=1,\ldots,q}\lambda_i$, where $q$ denotes the dependent variables in MANOVA study.

**Example 8.18**

A researcher randomly assigns 33 subjects to one of three groups. The first group receives technical dietary information interactively from an online website. Group 2 receives the same information from a nurse practitioner, while group 3 receives the information from a video tape made by the same nurse practitioner. The researcher looks at three different ratings of the presentation, which are to do with *difficulty*, *usefulness*, and *importance*, to determine if there is a difference in the modes of presentation. In particular, the researcher is interested to know whether the interactive website is superior because that is the most cost-effective way of delivering the information. Furthermore, the reader should note that (i) level 1 of the group variable is the treatment group; (ii) level 2 is control group 1, and finally (iii) level 3 is control group 2. The information about the data can be accessed at http://www.ats.ucla.edu/stat/stata/ado/analysis/. Without going through each and every step of the MANOVA calculation, we give the MANOVA test criteria and exact $F$ statistics for the hypothesis of no overall group effect and the results are shown in Table 8.3.

**Note:** Multivariate analysis of covariance (MANCOVA) is an extension of analysis of covariance (ANCOVA) methods to cover cases where there is more than one dependent variable and where the

**TABLE 8.3**

MANOVA Test Criteria and Exact $F$ Statistics for the Hypothesis of No Overall Group Effect, Example 8.18

| Statistic | Value | $F$ Value | Num dof | Den dof | $Pr > F$ |
|---|---|---|---|---|---|
| Wilks' lambda | 0.53598494 | 12.99 | 2 | 30 | <0.0001 |
| Pillai's trace | 0.46401506 | 12.99 | 2 | 30 | <0.0001 |
| Hotelling–Lawley trace | 0.86572405 | 12.99 | 2 | 30 | <0.0001 |
| Roy's greatest root | 0.86572405 | 12.99 | 2 | 30 | <0.0001 |

control of concomitant continuous independent variables—covariates—is required. The significant benefit of MANCOVA over MANOVA is the *factoring out* of noise or error that has been introduced by the covariant. The analysis one uses to solve problem using ANOVA versus MANOVA can be extended to the case when one is solving ANCOVA versus MANCOVA. We leave this for the reader to study and close this section with a few good references in this area such as Cooley and Lohnes (1971), Huberty and Olejnik (2006), and Morrison (1990).

## 8.18 Conjoint Analysis

When confronted with any decision process with different alternatives, human beings accept alternative(s) or reject alternative(s) or are ambivalent/indifferent (to different levels of degree) to alternative(s). Their choices are influenced by their likings, experiences, habits, role of advertisements, peer pressures, environmental effects, societal or family constraints, etc. Here is where conjoint analysis (CA) and discrete choice experimentation (DCE) may be used as tools for understanding how individuals develop preferences for alternatives. To study CA, discrete choice models, multiattribute utility theory (MAUT) and random utility theory (RUT) are used, but for the sake of brevity, we skip these discussions and concentrate on the general formulation, models, and applications of CA. Before we start our discussion, we state a few good texts in the area of CA, such as Louviere (1988), Orme and King (2006), Raghavarao et al. (2011), etc.

As a method, CA simultaneously finds a monotonic scoring of the dependent variable and numerical value for each level of each independent variable. This method is based on the main effects of ANOVA model. We state here the conjoint model in its simplest form for the ease of understanding of the readers. Consider $y_{i_1 \cdots i_p} = \mu + \beta_{1i_1} + \cdots + \beta_{pi_p} + \varepsilon_{i_1 \cdots i_p}$, such that $\sum \beta_{1i_1} = \cdots = \sum \beta_{pi_p} = 0$. Consider an example where you are an executive for a marketing firm and you are analyzing the factors that affect the decision of your target customer to buy a car. You want to investigate the preferences for the cars based on $p$ attributes say, for example, mileage, price, safety, resale value, style, passenger space, luggage space, etc. Thus, $y_{i_1 \cdots i_p}$ denotes the a buyer's stated preference for a car with respect to $i_1^{th}$ level of mileage, $i_2^{th}$ level of price, and so on. The nonmetric CA model for the above model can be expressed as $\Phi\left(y_{i_1 \cdots i_p}\right) = \mu + \beta_{1i_1} + \cdots + \beta_{pi_p} + \varepsilon_{i_1 \cdots i_p}$, where $\Phi(\cdot)$ implies a monotonic transformation of the variable $y$. CA can be solved by the method of ANOVA. An important assumption is that the distance between any two adjacent preference ordering corresponds to the same difference in utility. Thus, we treat the ranking, which is a cardinal variable as if it were metric variable.

**Example 8.19: (Härdle and Simar, 2007)**

A manufacturer of food items intends to make a new margarine and varies the *product characteristics* as well as its *packaging*. The four different products made by the food manufacturer are ordered as shown in Table 8.4. The information about the data set can be found in Härdle and Simar (2007).

Let us consider the part worth $X_1$ as usage, and suppose a person ranks the six different products as shown in Table 8.5.

Solving, one obtains $\beta_{11} = -2, \beta_{12} = 0$, and $\beta_{13} = 2$, while on the other hand, we get $\beta_{21} = 0.16$ and $\beta_{22} = -0.16$, and $\mu = 3.5$. Using these values, we can easily obtain $\hat{Y}_1 = \beta_{11} + \beta_{21} + \mu = 1.66$, $\hat{Y}_2 = \beta_{11} + \beta_{22} + \mu = 1.34$, $\hat{Y}_3 = \beta_{12} + \beta_{21} + \mu = 3.66$, $\hat{Y}_4 = \beta_{12} + \beta_{22} + \mu = 3.34$, $\hat{Y}_5 = \beta_{13} + \beta_{21} + \mu = 5.66$, and $\hat{Y}_6 = \beta_{13} + \beta_{22} + \mu = 5.34$.

**TABLE 8.4**

Ranking of the Four Products, Example 8.19

| Product Type | Product Characteristics | Packaging | Ranking |
|---|---|---|---|
| 1 | Low calories | Plastic pack | 3 |
| 2 | Low calories | Paper pack | 4 |
| 3 | High calories | Plastic pack | 1 |
| 4 | High calories | Paper pack | 2 |

**TABLE 8.5**

Ranked Products, Example 8.19

| | | $X_2$ Calories | |
|---|---|---|---|
| | | Low | High |
| $X_1$ Usage | | 1 | 2 |
| Bread | 1 | 2 | 1 |
| Cooking | 2 | 3 | 4 |
| Universal | 3 | 6 | 5 |

## 8.19 Canonical Correlation Analysis

Hotelling (1935, 1936) may be credited for developing the technique of canonical correlation analysis (CCA), where the author studied how arithmetic speed and arithmetic power are related to reading speed and reading power. Mathematically, CCA may be stated as follows. Suppose we are given $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Y} \in \mathbb{R}^q$, $(p \leq q)$, then the idea is to find an index describing a possible link between $\boldsymbol{X}$ and $\boldsymbol{Y}$. Using CCA, we are interested to find vectors $\boldsymbol{a}$ of size $(p \times 1)$ and $\boldsymbol{b}$ of size $(q \times 1)$, such that the correlation coefficient between $U = \boldsymbol{a}'\boldsymbol{X}$ and $V = \boldsymbol{b}'\boldsymbol{Y}$, given by $\rho(U, V) = \rho(\boldsymbol{a}, \boldsymbol{b})$, is *maximized*. Remember, CCA is based on linear indices or linear combination as both $U = \boldsymbol{a}'\boldsymbol{X}$ and $V = \boldsymbol{b}^T\boldsymbol{Y}$ are linear combinations and may be expressed as $a_1 X_1 + \cdots + a_p X_p$ and $b_1 Y_1 + \cdots + b_q Y_q$, respectively.

The idea of CCA is to go stepwise, where in the *first* step we determine the first pair of linear combinations, $(U_1, V_1)$, which results in the largest value of correlation, $\rho_1^{*2}$. In the next step, one then determines the *second* pair of linear combinations, $(U_2, V_2)$, which has the largest correlation, given by $\rho_2^{*2}$, among all the pairs such that they are *uncorrelated* with the initially selected pairs. We continue doing this till $p$ stages, so that we obtain $(U_1, V_1), \ldots, (U_p, V_p)$ and $\left(\rho_1^{*2}, \rho_2^{*2}, \ldots, \rho_p^{*2}\right)$. In case the condition $(p \leq q)$ does not hold true, then we continue doing this till $\min(p, q)$. CCA is a simple and useful method to describe the correlation structure between two sets of variables. It is a generalization of the concept of multiple correlation and successively maximizes the correlation between appropriate pairs of linear combinations of the variables of the two sets. The method can be viewed as a dimension reduction technique in that it represents the correlation structure between two sets of variables in terms of a smaller number of canonical correlations. The pairs of linear combinations $\{(U_1, V_1), (U_2, V_2), \ldots, (U_p, V_p)\}$ thus obtained are termed as the *canonical variables*, while the corresponding correlations $\left(\rho_1^{*2}, \rho_2^{*2}, \ldots., \rho_p^{*2}\right)$ are known as the *canonical correlations* values.

Some interesting application areas of CCA are:

1. Study of how government policy variables such as interest rates, expenditures in various economic sectors, etc. are related to other economic goal variables such as foreign currency rates, inflation rates, etc.
2. Study of college performance variables of students with respect to their scholastic achievements before joining the college.

### 8.19.1 Formulation of Canonical Correlation Analysis

Suppose $\boldsymbol{X}$ is distributed with $E(\boldsymbol{X}) = \boldsymbol{\mu}_X$, $Covar(\boldsymbol{X}) = \boldsymbol{\Sigma}_{XX}$, while $\boldsymbol{Y}$ is distributed with $E(\boldsymbol{Y}) = \boldsymbol{\mu}_Y$, $Covar(\boldsymbol{Y}) = \boldsymbol{\Sigma}_{YY}$. Moreover, $Covar(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{\Sigma}_{XY} = \boldsymbol{\Sigma}'_{YX}$. Then one can easily verify that

$$\rho(U, V) = \rho(\boldsymbol{a}, \boldsymbol{b}) = \left\{ \frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{XY}\mathbf{b}}{(\boldsymbol{a}'\boldsymbol{\Sigma}_{XX}\boldsymbol{a})^{1/2}(\boldsymbol{b}'\boldsymbol{\Sigma}_{YY}\mathbf{b})^{1/2}} \right\}.$$

One may also note that $\rho(c\mathbf{a}, \boldsymbol{b}) = \rho(\boldsymbol{a}, \boldsymbol{b}) = \rho(\boldsymbol{a}, c\boldsymbol{b})$, where $c \in \mathbb{R}^+$.

From the optimization point of view, CCA can be stated simply as follows:

$$\max \left(\boldsymbol{a}'\boldsymbol{\Sigma}_{XY}\mathbf{b}\right)$$
$$\text{s.t.} : \boldsymbol{a}'\boldsymbol{\Sigma}_{XX}\boldsymbol{a} = 1$$
$$\boldsymbol{b}'\boldsymbol{\Sigma}_{YY}\mathbf{b} = 1$$

A closer look at the above optimization makes it obvious that in maximizing the ratio, which is the correlation coefficient, $\rho(U, V)$, we maximize the numerate, that is, $Covar(\boldsymbol{X}, \boldsymbol{Y})$ with the restrictions that both $Covar(\boldsymbol{a}'\boldsymbol{X})$ and $Covar(\boldsymbol{b}'\boldsymbol{Y})$ are equal to 1. Standard nonlinear optimization algorithms are available, which solves this problem. For a better explanation of how we go about in achieving this, one can refer to the algorithm described as a pseudo-code in Figure 8.17.

Before discussing the standardized form of CCA, we state a few important results relevant to CCA:

```
DEFINE: $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{X}_{n\times p}$, $\boldsymbol{Y}_{n\times q}$ where $p < q, U = \boldsymbol{a}'\boldsymbol{X}$, $V = \boldsymbol{b}'\boldsymbol{Y}$
INPUT: $\boldsymbol{X}_{n\times p}$, $\boldsymbol{Y}_{n\times q}$ where $p < q$
CALCULATE: $E(\boldsymbol{X}) = \boldsymbol{\mu}_X$, $E(\boldsymbol{Y}) = \boldsymbol{\mu}_Y$, $Covar(\boldsymbol{X}) = \boldsymbol{\Sigma}_{XX}$, $Covar(\boldsymbol{Y}) = \boldsymbol{\Sigma}_{YY}$, $Covar(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{\Sigma}_{XY}$, $\boldsymbol{a}$, $\boldsymbol{b}$ such that $corr(U,V) = \frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{XY}\mathbf{b}}{(\boldsymbol{a}'\boldsymbol{\Sigma}_{XX}\boldsymbol{a})^{\frac{1}{2}}(\boldsymbol{b}'\boldsymbol{\Sigma}_{YY}\mathbf{b})^{\frac{1}{2}}}$ is maximized
START if: $i = 1{:}p$
    Maximize: $Covar(U_i,V_i)$
    s.t.: $Covar(U_i) = 1$ and $Covar(V_i) = 1$
    CALCULATE: (i) $\rho_i^{*2}$ which is the $i^{th}$ eigen value of $\boldsymbol{\Sigma}_{XX}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-\frac{1}{2}}$ and $\boldsymbol{e}_i$ which is its corresponding $i^{th}$ eigen vector, (ii) $\rho_i^{*2}$ which is the $i^{th}$largesteigen value of $\boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-\frac{1}{2}}$ and $f_i$which is its corresponding $i^{th}$ eigen vector, (iii) in calculating these $i^{th}$ values, ensure that we find those linear combinations which are uncorrelated with the preceding $1,2,..,i-1$ number of canonical variables, (iv) $\boldsymbol{a}_i = \boldsymbol{e}_i'\boldsymbol{\Sigma}_{XX}^{-\frac{1}{2}}$ and $\boldsymbol{b}_i = \boldsymbol{f}_i'\boldsymbol{\Sigma}_{XX}^{-\frac{1}{2}}$, (v) $U_i = \boldsymbol{a}_i'\boldsymbol{X}$ and $V_i = \boldsymbol{b}_i'\boldsymbol{Y}$
END if
REPORT: $\left(\rho_1^{*2}, \rho_2^{*2}, \ldots, \rho_p^{*2}\right)$ and $\left\{(U_1,V_1), \ldots.., \left(U_p,V_p\right)\right\}$
END
```

**FIGURE 8.17**
Pseudo-code to implement CCA method.

1. For any given $r, 1 \leq j \leq p$, the maximum value of $\boldsymbol{a}'\boldsymbol{\Sigma}_{XY}\mathbf{b}$ subject to (i) $\boldsymbol{a}'\boldsymbol{\Sigma}_{XX}\boldsymbol{a} = 1$, (ii) $\boldsymbol{b}'\boldsymbol{\Sigma}_{YY}\mathbf{b} = 1$, and (iii) $\boldsymbol{a}'_j\ \boldsymbol{\Sigma}_{XX}\boldsymbol{a} = 0$ for $j = 1, \ldots, r-1$ is given by $\rho_j = \sqrt{\lambda_j}$, where $e_j$ is the $j$th eigen value of $\left(\boldsymbol{\Sigma}_{\mathrm{XX}}^{-1/2}\boldsymbol{\Sigma}_{\mathrm{XY}}\boldsymbol{\Sigma}_{\mathrm{YY}}^{-1/2}\right)\left(\boldsymbol{\Sigma}_{\mathrm{XX}}^{-1/2}\boldsymbol{\Sigma}_{\mathrm{XY}}\boldsymbol{\Sigma}_{\mathrm{YY}}^{-1/2}\right)'$, and the maximum value is obtained when $\boldsymbol{a} = \boldsymbol{a}_j$ and $\boldsymbol{b} = \boldsymbol{b}_j$.
2. Let $U_j = \boldsymbol{a}'_j\boldsymbol{X}$ and $V_j = \boldsymbol{b}'_j\boldsymbol{Y}$ be the $j$th canonical correlation variables, $j = 1, \ldots, p$. Then

$$Var\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_p & \Lambda \\ \Lambda & \boldsymbol{I}_p \end{pmatrix},$$

where $\boldsymbol{U} = (U_1, \ldots, U_p), \boldsymbol{V} = (V_1, \ldots, V_p)$ and $\Lambda = \text{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p}\right)$.

**Note:** Thus, the canonical correlation coefficients, $\rho_j = \sqrt{\lambda_j}$, are the covariances between $U_j$ and $V_j$, where $j = 1, \ldots, p$. Moreover, $\boldsymbol{a}'_1\boldsymbol{X}$ and $\boldsymbol{b}'_1\boldsymbol{Y}$ have the maximum covariance of value $\rho_1 = \sqrt{\lambda_1}$.

One should also remember a few important things (considering, $p \leq q$):

1. For the matrix $\left(\boldsymbol{\Sigma}_{XX}^{-1/2}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2}\right)\left(\boldsymbol{\Sigma}_{XX}^{-1/2}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2}\right)'$ which is of size $(p \times p)$, we have $\rho_1^2 \geq \cdots \geq \rho_p^2$ as the eigen values for which the associated eigen vectors are $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$.
2. For the matrix $\left(\boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1/2}\right)\left(\boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1/2}\right)'$, which is of size $(q \times q)$, we have $\rho_1^2 \geq \cdots \geq \rho_p^2$ as the *largest* $p$ eigen values for which the associated eigen vectors are $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_p$.
3. $f_j \propto \left(\boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1/2}\right)\boldsymbol{e}_i, \quad j = 1, \ldots, p.$
4. $\left(U_1 = \boldsymbol{a}'_1\boldsymbol{X} = \boldsymbol{e}'_1\boldsymbol{\Sigma}_{XX}^{-1/2}\boldsymbol{X}, V_1 = \boldsymbol{b}'_1\boldsymbol{Y} = \boldsymbol{f}'_1\boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{Y}\right), \ldots,$
$\left(U_p = \boldsymbol{b}'_p\boldsymbol{X} = \boldsymbol{e}'_p\boldsymbol{\Sigma}_{XX}^{-1/2}\boldsymbol{X}, V_p = \boldsymbol{b}'_p\boldsymbol{Y} = \boldsymbol{f}'_p\boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{Y}\right).$
5. $Var(U_j) = Var(V_j) = 1, j = 1, \ldots, p.$
6. $Covar\left(U_{j_1}, U_{j_2}\right) = corr\left(U_{j_1}, U_{j_2}\right) = 0, \quad j_1 \neq j_2 = 1, \ldots, p.$
7. $Covar\left(V_{j_1}, V_{j_2}\right) = corr\left(V_{j_1}, V_{j_2}\right) = 0, \quad j_1 \neq j_2 = 1, \ldots, p.$
8. $Covar\left(U_{j_1}, V_{j_2}\right) = corr\left(U_{j_1}, V_{j_2}\right) = 0, \quad j_1 \neq j_2 = 1, \ldots, p.$

### 8.19.2 Standardized Form of CCA

In case

$$Z_{\boldsymbol{X}_i} = \frac{X_i - \mu_{Y_i}}{\sqrt{\sigma_{X_iX_i}}},$$

where

$$\rho_{XX} = \begin{pmatrix} \rho_{X_1X_1} & \cdots & \rho_{X_pX_1} \\ \vdots & \ddots & \vdots \\ \rho_{X_1X_p} & \cdots & \rho_{X_pX_p} \end{pmatrix}, \quad \boldsymbol{\Sigma}_{XX} = \begin{pmatrix} \sigma_{X_1X_1} & \cdots & \sigma_{X_pX_1} \\ \vdots & \ddots & \vdots \\ \sigma_{X_1X_p} & \cdots & \sigma_{X_pX_p} \end{pmatrix} \quad \text{and} \quad Z_{\boldsymbol{Y}_j} = \frac{Y_j - \mu_{Y_j}}{\sqrt{\sigma_{Y_jY_j}}},$$

where

$$\boldsymbol{\rho}_{YY} = \begin{pmatrix} \rho_{Y_1Y_1} & \cdots & \rho_{Y_qY_1} \\ \vdots & \ddots & \vdots \\ \rho_{Y_1Y_q} & \cdots & \rho_{Y_qY_q} \end{pmatrix}, \quad \boldsymbol{\Sigma}_{YY} = \begin{pmatrix} \sigma_{Y_1Y_1} & \cdots & \sigma_{Y_qY_1} \\ \vdots & \ddots & \vdots \\ \sigma_{Y_1Y_q} & \cdots & \sigma_{Y_qY_q} \end{pmatrix},$$

then $U_k = \boldsymbol{a}'_k\boldsymbol{Z_X} = \boldsymbol{e}'_k\boldsymbol{\rho}_{XX}^{-1/2}\boldsymbol{Z_X}$ and $V_k = \boldsymbol{b}'_k\boldsymbol{Z_Y} = \boldsymbol{f}'_k\boldsymbol{\rho}_{YY}^{-1/2}\boldsymbol{Z_Y}, k = 1, \dots, p$.

One should also remember the following:

1. $Covar(Z_{\boldsymbol{X}}) = \boldsymbol{\rho}_{XX}$.
2. $Covar(Z_{\boldsymbol{Y}}) = \boldsymbol{\rho}_{YY}$.
3. $Covar(Z_X, Z_Y) = \boldsymbol{\rho}_{XY} = \boldsymbol{\rho}'_{YX}$.
4. $\rho_1^{*2} \geq \cdots \geq \rho_p^{*2}$ are the nonzero eigen values of $\boldsymbol{\rho}_{XX}^{-1/2}\boldsymbol{\rho}_{XY}\boldsymbol{\rho}_{YY}^{-1}\boldsymbol{\rho}_{YX}\boldsymbol{\rho}_{XX}^{-1/2}$ for which $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_p)$ is the set of corresponding eigen vectors.
5. $\rho_1^{*2} \geq \cdots \geq \rho_p^{*2}$ are the nonzero *largest p* set of eigen values, from among $q$ of them, of $\boldsymbol{\rho}_{YY}^{-1/2}\boldsymbol{\rho}_{YX}\boldsymbol{\rho}_{XX}^{-1}\boldsymbol{\rho}_{XY}\boldsymbol{\rho}_{YY}^{-1/2}$ for which $(\boldsymbol{f}_1, \dots, \boldsymbol{f}_p)$ is the set of corresponding eigen vectors.

### 8.19.3 Correlation between Canonical Variates and Their Component Variables

In case one is interested to find the correlation between the original variables and their respective transformed variables, then a few interesting results can be stated. But before that, consider $\boldsymbol{A}_{p\times p} = [\boldsymbol{a}_1, \dots, \boldsymbol{a}_p]', \boldsymbol{B}_{q\times q} = [\boldsymbol{b}_1, \dots, \boldsymbol{b}_q]'$, so that $\boldsymbol{U}_p \times 1 = \boldsymbol{A}_{p\times p}\boldsymbol{X}_{p\times 1}$ and $\boldsymbol{V}_{q\times 1} = \boldsymbol{B}_{q\times q}\boldsymbol{Y}_{q\times 1}$ and $p \leq q$. With these, the following results stated below hold:

1. $Covar(\boldsymbol{U}, \boldsymbol{X}) = \boldsymbol{A}'\boldsymbol{\Sigma}_{XX}$.
2. $Covar(\boldsymbol{V}, \boldsymbol{Y}) = \boldsymbol{B}'\boldsymbol{\Sigma}_{YY}$.
3. $$\boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{X}_{p\times p}} = \boldsymbol{A}'_{p\times p}\boldsymbol{\Sigma}_{XX\,p\times p}\begin{bmatrix} \sqrt{Var(X_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{Var(X_p)} \end{bmatrix}_{p\times p}.$$
4. $$\boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{Y}_{p\times q}} = \boldsymbol{A}'_{p\times p}\boldsymbol{\Sigma}_{XY\,p\times q}\begin{bmatrix} \sqrt{Var(Y_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{Var(Y_q)} \end{bmatrix}_{q\times q}.$$
5. $$\boldsymbol{\rho}_{\boldsymbol{V},\boldsymbol{X}_{q\times p}} = \boldsymbol{B}_{q\times q}\boldsymbol{\Sigma}_{YX\,q\times p}\begin{bmatrix} \sqrt{Var(X_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{Var(X_p)} \end{bmatrix}_{p\times p}.$$
6. $$\boldsymbol{\rho}_{\boldsymbol{V},\boldsymbol{Y}_{q\times q}} = \boldsymbol{B}_{q\times q}\boldsymbol{\Sigma}_{YY\,q\times q}\begin{bmatrix} \sqrt{Var(Y_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{Var(Y_q)} \end{bmatrix}_{q\times q}.$$

While in the standardized variable case, we have

1. $Covar(\boldsymbol{U}, \boldsymbol{Z}_X) = \boldsymbol{A}\rho_{XX}$.
2. $Covar(\boldsymbol{V}, \boldsymbol{Z}_Y) = \boldsymbol{B}\rho_{YY}$.
3. 
$$\rho_{\boldsymbol{U},\boldsymbol{Z}_{X_{p\times p}}} = \boldsymbol{A}_{Z_{X_{p\times p}}}\rho_{XX_{p\times p}}\begin{bmatrix}\sqrt{Var\left(Z_{X_1}\right)} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{Var\left(Z_{X_p}\right)}\end{bmatrix}_{p\times p}.$$

4. 
$$\rho_{\boldsymbol{U},\boldsymbol{Z}_{Y_{p\times q}}} = \boldsymbol{A}_{Z_{X_{p\times p}}}\rho_{XY_{p\times q}}\begin{bmatrix}\sqrt{Var\left(Z_{Y_1}\right)} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{Var\left(Z_{Y_q}\right)}\end{bmatrix}_{q\times q}.$$

5. 
$$\rho_{\boldsymbol{V},\boldsymbol{Z}_{X_{q\times p}}} = \boldsymbol{B}_{Z_{Y_{q\times q}}}\rho_{YX_{q\times p}}\begin{bmatrix}\sqrt{Var\left(Z_{X_1}\right)} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{Var\left(Z_{X_p}\right)}\end{bmatrix}_{p\times p}.$$

6. 
$$\rho_{\boldsymbol{V},\boldsymbol{Z}_{Y_{q\times q}}} = \boldsymbol{B}_{Z_{Y_{q\times q}}}\rho_{YY_{q\times q}}\begin{bmatrix}\sqrt{Var\left(Z_{Y_1}\right)} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{Var\left(Z_{Y_q}\right)}\end{bmatrix}_{q\times q}.$$

**Note:** A different perspective when analyzing CCA from the optimization point of view is stated here, with the idea that it acts as a good motivation for the interested readers. Let us consider the optimization problem stated before. In its Lagrangian form, considering the Lagrangian multipliers as $\lambda_1$ and $\lambda_2$, the expression to be differentiated is $g(\lambda_1, \lambda_2) = \boldsymbol{a}'\boldsymbol{\Sigma}_{XY}\mathbf{b} - \lambda_1\boldsymbol{a}'\boldsymbol{\Sigma}_{XX}\boldsymbol{a} - \lambda_2\boldsymbol{b}'\boldsymbol{\Sigma}_{YY}\mathbf{b}$. Now, putting $\partial g(\lambda_1, \lambda_2)/\partial\lambda_1 = \partial g(\lambda_1, \lambda_2)/\partial\lambda_2 = 0$, one obtains $\lambda_1 = \boldsymbol{a}'\boldsymbol{\Sigma}_{XX}\boldsymbol{a}$. With a few relevant mathematical changes, we obtain $\lambda^2$ (here, $\lambda_1 = \lambda_1 = \lambda$) and $\boldsymbol{b}$ as the eigen root and eigen vector corresponding to the determinantal equation $|\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY} - \lambda^2\boldsymbol{\Sigma}_{YY}| = 0$. On a similar line, one can also solve and get the other set of eigen root and eigen vector.

### 8.19.4 Testing the Test Statistics in CCA

A main concern in CCA is to find whether there is some significant relationship or dependence between the variables $\boldsymbol{X}$ and $\boldsymbol{Y}$. Gittins (1985) suggested the following test to verify the dependence between the variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ using Wilk's likelihood ratio statistics, which is given by $T^{2/n} = |\boldsymbol{I} - \boldsymbol{S}_{YY}^{-1}\boldsymbol{S}_{YX}\boldsymbol{S}_{XX}^{-1}\boldsymbol{S}_{XY}| = \prod_{i=1}^{k}(1 - l_i)$. As this statistic has a complicated distribution, hence it is denoted by $-\{n - (p + q + 3)/2\}log\prod_{i=1}^{k}(1 - l_i) \sim \chi^2_{p\times q}$, provided $n \to \infty$ (Barlett, 1954). In case one is interested to find if only $s$ of the total number of canonical correlations are nonzero, then the statistic is of the form $-\{n - (p + q + 3)/2\}log\prod_{i=s+1}^{k}(1 - l_i) \sim \chi^2_{(p-s)\times(q-s)}$, and this holds true in the approximate sense as $n \to \infty$.

**Example 8.20**

Consider we have a theoretical set of data where the following is given: $\boldsymbol{X} = (X_1,\ X_2,\ X_3)$, that is, $p = 3$, $\boldsymbol{Y} = (Y_1,\ Y_2,\ Y_3,\ Y_4,\ Y_5)$, that is, $q = 5$, such that $E(\boldsymbol{X}) = (2, 3, 6)$, $E(\boldsymbol{Y}) = (45, 44, 34, 32, 40)$,

$$Covar(\boldsymbol{X}) = \boldsymbol{\Sigma}_{XX} = \begin{pmatrix} 0.4 & 0.2449 & 0.45 \\ 0.2449 & 0.6 & 0.1837 \\ 0.45 & 0.1837 & 0.9 \end{pmatrix},$$

$$Covar(\boldsymbol{Y}) = \boldsymbol{\Sigma}_{YY} = \begin{pmatrix} 4 & 1.3416 & 1.4697 & 0 & 2.2627 \\ 1.3416 & 5 & 2.1909 & 5.3666 & 0.6325 \\ 1.4697 & 2.1909 & 6 & 0.7348 & 4.1569 \\ 0 & 5.3666 & 0.7348 & 9 & 5.9397 \\ 2.2627 & 0.6325 & 4.1569 & 5.9397 & 8 \end{pmatrix},$$

and

$$Covar(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{\Sigma}_{XY} = \begin{pmatrix} 0.1265 & 0.4243 & 1.3943 & 1.8974 & 0.7155 \\ 1.2394 & 0 & 0.3795 & 2.3238 & 0.6573 \\ 0.3795 & 0.8485 & 1.1619 & 1.7076 & 2.1466 \end{pmatrix}.$$

Given this set of information, let us calculate the following, the values of which are given alongside the formulae. We urge the reader to recalculate the values to get a good idea about the steps involved in CCA calculations:

1.

$$\boldsymbol{\Sigma}_{XX}^{-1/2}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1/2} = \begin{pmatrix} -5.3437 & 13.4544 & -7.6743 \\ 13.4544 & -32.9776 & 19.9303 \\ -7.6743 & 19.9303 & -13.6489 \end{pmatrix}.$$

2.

$$\boldsymbol{e} = \begin{pmatrix} -50.7400 & 0 & 0 \\ 0 & 0.2008 & 0 \\ 0 & 0 & -1.4311 \end{pmatrix}.$$

3.

$$\boldsymbol{\rho}^* = \begin{pmatrix} 0.3227 & -0.8657 & -0.3827 \\ -0.8043 & -0.4640 & 0.3713 \\ 0.4990 & -0.1880 & 0.8460 \end{pmatrix}.$$

4.

$$\boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2}$$

$$= \begin{pmatrix} -0.2219 - 0.1508i & 0.9157 - 2.1750i & -0.7086 + 0.4421i & 0.9504 - 2.1756i & 0.4862 + 1.8497i \\ 0.9157 - 2.1750i & -22.7845 + 1.1740i & 2.2736 - 0.9611i & -22.2198 - 2.8159i & 12.9425 - 4.9868i \\ -0.7086 + 0.4421i & 2.2736 - 0.9611i & -0.3291 + 0.8951i & 1.8779 - 0.3799i & -0.2866 + 0.6717i \\ 0.9504 - 2.1756i & -22.2198 - 2.8159i & 1.8779 - 0.3799i & -21.2278 - 6.7355i & 13.8364 - 2.2843i \\ 0.4862 + 1.8497i & 12.9425 - 4.9868i & -0.2866 + 0.6717i & 13.8364 - 2.2843i & -7.4071 + 4.8171i \end{pmatrix}.$$

5.

$$\boldsymbol{f} = \begin{pmatrix} -50.7400 & 0 & 0 \\ 0 & 0.2008 & 0 \\ 0 & 0 & -1.4311 \end{pmatrix}.$$

6.

$$\rho^* = \begin{pmatrix} 0.3227 & -0.8657 & -0.3827 \\ -0.8043 & -0.4640 & 0.3713 \\ 0.4990 & -0.1880 & 0.8460 \end{pmatrix}.$$

**Example 8.21**

As the next example, consider the data given in http://www.ats.ucla.edu/stat/r/dae/canonical.htm, which consists of 600 observations on eight variables. The psychological variables are (i) locus of control, $X_1$, (ii) self-concept, $X_2$, and (iii) motivation, $X_3$, such that $\boldsymbol{X}_{600\times 3}$ is the first set of variable, while the academic variables are standardized tests in (i) reading, $Y_1$ (ii) writing, $Y_2$, (iii) mathematics, $Y_3$, and (iv) science, $Y_4$ are such that $\boldsymbol{Y}_{600\times 4}$ is the second set of variables. Additionally, the variable female is a zero-one indicator variable with one indicating a female student, while a zero denotes a male student.

Solving the CCA problems yields

$$\boldsymbol{A} = (\boldsymbol{a}_1\boldsymbol{a}_2\boldsymbol{a}_3) = \begin{pmatrix} -1.2501 & 0.7660 & -0.4967 \\ 0.2367 & 0.8421 & 1.2051 \\ -1.2491 & -2.6360 & 1.0935 \end{pmatrix},$$

$$\boldsymbol{B} = (\boldsymbol{b}_1\boldsymbol{b}_2\boldsymbol{b}_3\boldsymbol{b}_4) = \begin{pmatrix} -0.0440 & -0.0016 & 0.0883 \\ -0.0551 & -0.0904 & -0.0961 \\ -0.0194 & -0.0030 & 0.0878 \\ 0.0038 & 0.1242 & -0.0885 \end{pmatrix},$$

$$\rho^{*^2} = (0.4464 \quad 0.1534 \quad 0.0225).$$

This means that the set of linear combinations of the variables are:

1.

$$U_1 = \boldsymbol{a}_1'\boldsymbol{X} = (-1.2501\ 0.2367\ -1.2491)\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = -1.2501X_1 + 0.2367X_2 - 1.2491X_3,$$

$$V_1 = \boldsymbol{b}_1'\boldsymbol{Y} = (-0.0440\ -0.0551\ -0.0194\ 0.0038)\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = -0.0440Y_1 - 0.0551Y_2$$

$$- 0.0194Y_3 + 0.0038Y_4,$$

2.

$$U_2 = \boldsymbol{a}_2'\boldsymbol{X} = (0.7660\ 0.8421 - 2.6360)\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = 0.7660X_1 + 0.8421X_2 - 2.6360X_3,$$

$$V_2 = \boldsymbol{b}_2'\boldsymbol{Y} = (-0.0016 - 0.0904 - 0.0030\ 0.1242)\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = -0.0016Y_1 - 0.0904Y_2$$

$$- 0.0030Y_3 + 0.1242Y_4,$$

3.

$$U_3 = \boldsymbol{a}_3'\boldsymbol{X} = (-0.4967\ 1.2051\ 1.0935)\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = -0.4967X_1 + 1.2051X_2 + 1.0935X_3,$$

$$V_3 = \boldsymbol{b}_3'\boldsymbol{Y} = (0.0883\ \ -0.0961\ \ -0.0878\ 0.0885)\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = 0.0883Y_1\ -0.0961Y_2$$

$$-\,0.0878Y_3\ -0.0885Y_4,$$

respectively. The corresponding linear combination graphs for $(U_1, V_1)$, $(U_2, V_2)$ and $(U_3, V_3)$ are shown in Figure 8.18. Though not apparent but one can easily discern that the value of correlation coefficient or the slope of the set of $(U_1, V_1)$ are the maximum, followed by $(U_2, V_2)$, and then $(U_3, V_3)$. This fact is also corroborated by the values of $\rho_1^{*^2} = 0.4464$, $\rho_2^{*^2} = 0.1534$, and $\rho_3^{*^2} = 0.0225$. Another way of verifying the values of $\rho_1^{*^2}, \rho_2^{*^2}, \rho_3^{*^2}$ is to have a look at

$$Covar(\boldsymbol{U}, \boldsymbol{V}) = \begin{bmatrix} 1 & 0.4464 & 0 & 0 & 0 & 0 \\ 0.4464 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.1543 & 0 & 0 \\ 0 & 0 & 0.1534 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.0225 \\ 0 & 0 & 0 & 0 & 0.0225 & 1 \end{bmatrix}.$$

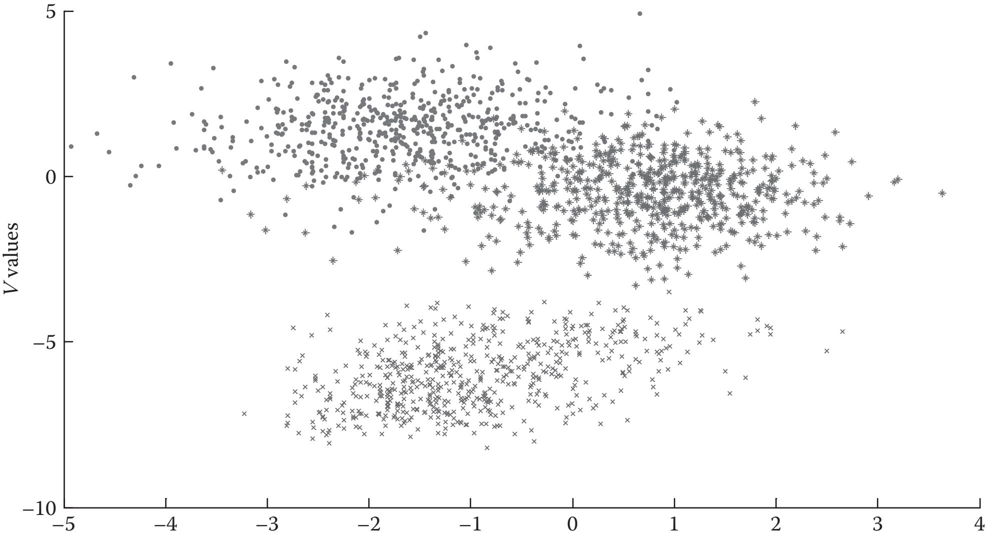


**FIGURE 8.18**
Graphs showing linear relationship between the set of variables using CCA method, Example 8.20.

For this problem, let us also find the following detailed calculations, which are shown for ease of understanding:

- $$Covar(\boldsymbol{U}, \boldsymbol{X}) = \boldsymbol{A}'\boldsymbol{\Sigma}_{XX} = \begin{pmatrix} -0.6128 & -0.0705 & -0.2006 \\ 0.2639 & 0.2972 & -0.2077 \\ -0.0640 & 0.6359 & 0.1846 \end{pmatrix}$$

- $$Covar(\boldsymbol{V}, \boldsymbol{Y}) = \boldsymbol{B}'\boldsymbol{\Sigma}_{YY} = \begin{pmatrix} -8.8950 & -8.8523 & -7.5317 & -6.7371 \\ 2.4743 & -2.1492 & 1.7693 & 6.5603 \\ 2.7587 & -3.3050 & 2.6697 & -2.3069 \end{pmatrix}$$

- $$\boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{X}_{p\times p}} = \boldsymbol{A}'_{p\times p}\boldsymbol{\Sigma}_{XX_{p\times p}} \begin{bmatrix} \sqrt{Var(X_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{Var(X_p)} \end{bmatrix}_{p\times p}$$
$$= \begin{pmatrix} 0.6703 & 0 & 0 \\ 0 & 0.7055 & 0 \\ 0 & 0 & 0.3427 \end{pmatrix}.$$

- $$\boldsymbol{\rho}_{\boldsymbol{U},\boldsymbol{Y}_{p\times q}} = \boldsymbol{A}'_{p\times p}\boldsymbol{\Sigma}_{XY_{p\times q}} \begin{bmatrix} \sqrt{Var(Y_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{Var(Y_q)} \end{bmatrix}_{q\times q}$$
$$= \begin{pmatrix} -5.6368 & -5.4272 & -5.2782 & -5.4888 \\ 4.3703 & 4.2584 & 4.0730 & 4.2295 \\ 5.0434 & 4.9145 & 4.7151 & 4.8921 \end{pmatrix}.$$

Furthermore, the value of

1. $$\boldsymbol{\rho}_{\boldsymbol{V},\boldsymbol{X}_{q\times p}} = \boldsymbol{B}'_{q\times q}\boldsymbol{B}_{q\times q}\boldsymbol{\Sigma}_{YX_{q\times p}} \begin{bmatrix} \sqrt{Var(X_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{Var(X_p)} \end{bmatrix}_{p\times p}$$

   and

2. $$\boldsymbol{\rho}_{\boldsymbol{V},\boldsymbol{Y}_{q\times q}} = \boldsymbol{B}_{q\times q}\boldsymbol{\Sigma}_{YY_{q\times q}} \begin{bmatrix} \sqrt{Var(Y_1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{Var(Y_q)} \end{bmatrix}_{q\times q}$$

are left for the readers to calculate, which we are sure will give them a certain level of confidence to appreciate this multivariate statistical method in a better manner. Similarly, one can also find the standardized variables:

1. $Covar(\boldsymbol{U}, \boldsymbol{Z}_X) = \boldsymbol{A}\rho_{XX}$,
2. $Covar(\boldsymbol{V}, \boldsymbol{Z}_Y) = \boldsymbol{B}\rho_{YY}$,
3. $$\rho_{\boldsymbol{U},\boldsymbol{Z}_{X\,p\times p}} = \boldsymbol{A}_{Z_{X_{p\times p}}}\rho_{XX_{p\times p}}\begin{bmatrix}\sqrt{Var\left(Z_{X_1}\right)} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{Var\left(Z_{X_p}\right)}\end{bmatrix}_{p\times p},$$
4. $$\rho_{\boldsymbol{U},\boldsymbol{Z}_{Y\,p\times q}} = \boldsymbol{A}_{Z_{X_{p\times p}}}\rho_{XY_{p\times q}}\begin{bmatrix}\sqrt{Var\left(Z_{Y_1}\right)} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{Var\left(Z_{Y_q}\right)}\end{bmatrix}_{q\times q},$$
5. $$\rho_{\boldsymbol{V},\boldsymbol{Z}_{X_{q\times p}}} = \boldsymbol{B}_{Z_{Y_{q\times q}}}\rho_{YX_{q\times p}}\begin{bmatrix}\sqrt{Var\left(Z_{X_1}\right)} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{Var\left(Z_{X_p}\right)}\end{bmatrix}_{p\times p},$$
6. $$\rho_{\boldsymbol{V},\boldsymbol{Z}_{Y\,q\times q}} = \boldsymbol{B}_{Z_{Y_{q\times q}}}\rho_{YY_{q\times q}}\begin{bmatrix}\sqrt{Var\left(Z_{Y_1}\right)} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \sqrt{Var\left(Z_{Y_q}\right)}\end{bmatrix}_{q\times q}.$$

A few other sources for CCA are those provided by Karamouz et al. (2010) and Szakács et al. (2004). The interested readers can definitely have a look at the data sets in these references and work on them to hone their skills and understanding in CCA.

### 8.19.5 Geometric and Graphical Interpretation of CCA

It would not be out of context to say that a better appreciation of the CCA method can be obtained if one looks at the geometrical as well as graphical interpretation of CCA. A relevant reference for this is González et al. (2008). It is interesting to note that neural network (Hsieh, 2000) and kernel-based methods (Bach and Jordan, 2002, Lai and Fyfe, 2000, Melzer et al., 2001) have been used in the area of nonlinear CCA.

### 8.19.6 Conclusions about CCA

Before we wind up this topic, there are a few important points that should be remembered:

1. The classical CCA method may be used only for the condition when $n \geq (p+q+1)$ (Eaton and Perlman, 1973).

2. In case $\boldsymbol{X}$ and $\boldsymbol{Y}$ are highly correlated, then the matrices $\boldsymbol{\Sigma}_{XX}$ and $\boldsymbol{\Sigma}_{XX}$ are ill conditioned and their respective inverses are unreliable.
3. In case $n < (p+q)$, then CCA cannot be utilized and for such cases, that is, $n < (p+q)$, partial least square (PLS) regression may be used. The advantage of PLS over CCA is the fact that in the former the asymmetry relationship between the predictors and dependent variables is preserved, while CCA treats them symmetrically.
4. Vinod (1976) and Leurgans et al. (1993) have shown the extension of ridge regression in the area of CCA.

## 8.20 Cluster Analysis

Cluster analysis (CA) is a statistical technique whereby we form clusters/groups of *similar* individuals/objects using data/information from individuals/objects. This statistical method develops tools and methods, where given a data matrix, $\boldsymbol{X}_{(n \times p)}$, consisting of $n$ number of individuals/objects where each of these $n$ individuals/objectives are of dimension $p$, our aim is to build some natural subgroups or clusters of these individuals/objects. Using CA, we try to find some similarity or patterns in the data, for example, classification of plants/animals using taxonomy, diseases using epidemiology, etc. From a historical perspective, the origin of CA may be traced back to the work of Driver and Kroeber (1932) in anthropology. Later on, it was used in psychology (Cattell, 1943, Tryon, 1939, Zubin, 1938). Cluster analysis has been used in a variety of fields ranging from anthropology, agriculture, economics, psychology, geophysics, psychiatry, sociology, marketing, finance, behavioral sciences, different fields of engineering, etc. Even though old, good references with interesting applications can be found in Gordon (1981) and Hartigan (1975). Other good references from a theoretical points of view are Anderberg (1973), Duda et al. (2001), Duran and Odell (1974), Everitt and Dunn (2001), Gordon (1981), Hartigan (1975), Jain and Dubes (1988), Kaufman and Rousseeuw (2005), Späth (1980), etc. Another good book in the area of CA is by Xu and Wunsch (2008). Some other mathematical technique methods similar to cluster analysis are pattern recognition, numerical taxonomy, morphometrics, etc.

For a better understanding of clustering analysis as a technique, one should understand the basic four steps involved in cluster analysis:

- *Feature selection or extraction*: In the feature selection, step/stage one chooses the distinguishing features from a set of candidates, while on the other hand, in feature extraction step/stage, we utilize some transformations to generate useful and novel features from the original ones.
- *Clustering algorithm design and selection*: Depending on the proximity measure $d(P, Q)$, one constructs clustering criterion function so that the clustering algorithms may be developed. The main focus of the clustering algorithms is to cluster the objectives in groups based on some predefined criterion.
- *Cluster validation*: Effective validation standards and criteria are important to provide the degree of confidence for the clustering results derived from the used algorithms. This is what is done in the third stage step, which is the clustering validation step/stage.
- *Result interpretation*: The ultimate goal of clustering analysis step/stage is to provide the user with meaningful insights from the original data, so that they can effectively

solve the problems encountered, and this is what the result interpretation step/stage does.

The two fundamental steps in CA, which would be discussed by us here are: (i) choice of proximity (closeness) measure and (ii) choice of group-building algorithm.

Consider

$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}_{n\times p},$$

such that the proximity/distance matrix is given by

$$\boldsymbol{D}_{n\times n} = \begin{pmatrix} d_{1,1} & \cdots & d_{1,n} \\ \vdots & \ddots & \vdots \\ d_{n,1} & \cdots & d_{n,n} \end{pmatrix}_{n\times n},$$

where $d_{i,j}$ gives the measure of proximity/distance and is denoted by $\|x_i - x_j\|_2$ or $\{\max_{i,j}(d_{i,j}) - d_{i,j}\}$ as the case may be. In case, we have a *binary* structure pertaining to $\boldsymbol{X}$, that is, $x_{i,k} \in \{0, 1\}$, $i = 1, \ldots, n$ and $k = 1, \ldots, p$, then $d_{i,j} = (a_1 + \delta a_4)/(a_1 + \delta a_4 + \lambda(a_2 + a_3))$, where $\delta$ and $\lambda$ are the weighting factors. Here, $a_1 = \sum_{k=1}^{p} I(x_{i,k} = x_{j,k} = 1)$, $a_2 = \sum_{k=1}^{p} I(x_{i,k} = 0, x_{j,k} = 1)$, $a_3 = \sum_{k=1}^{p} I(x_{i,k} = 1, x_{j,k} = 0)$, and $a_4 = \sum_{k=1}^{p} I(x_{i,k} = x_{j,k} = 0)$. A few examples of weighting factors are $(\delta = 0, \lambda = 1)$ (Jaccard, 1901), $(\delta = 1, \lambda = 2)$ (Tanimoto, 1957), and $(\delta = 0, \lambda = 0.5)$ (Dice,[*] 1945). On the other hand, when we have the *continuous* variable, then $d_{i,j} = \|x_i - x_j\|_r = \left\{\sum_{k=1}^{p} |x_{i,k} - x_{j,k}|^r\right\}^{1/r}$ (Minkowski metric), when expressed in the nonstandardized form, while $d^2_{i,j} = \sum_{k=1}^{p}((x_{i,k} - x_{j,k})^2/s_{X_k,X_k})$ is the standardized version of this distance measure. A few other distance measures that have found use in CA are: Hamming distance, Euclidean distance: $d^2_{i,j} = \sum_{k=1}^{p}(x_{i,k} - x_{j,k})^2$, Soergel distance: $d_{i,j} = \sum_{k=1}^{p} |x_{i,k} - x_{j,k}| / \sum_{k=1}^{p} \max(x_{i,k}, x_{j,k})$, Canberra metric: $d_{i,j} = \sum_{k=1}^{p}\{|x_{i,k} - x_{j,k}|/(|x_{i,k}| + |x_{j,k}|)\}$, Czekanowski metric:

$$d_{i,j} = \left\{1 - \frac{2\sum_{k=1}^{p} \min(x_{i,k}, x_{j,k})}{\sum_{k=1}^{p}(x_{i,k} + x_{j,k})}\right\},$$

etc. One should remember that both Canberra and Czekanowski measures are defined for nonnegative variables only. Even without the precise notion of a natural grouping, one is often able to cluster/group individuals/objects in 2-D or 3-D plots using eye, stars, and Chernoff faces.

When items (units or cases) are clustered, proximity is usually indicated by some sort of distance. For variables, the grouping is on the basis of correlation coefficients or such similar measure. Central to the goal of cluster analysis is the idea of the degree of similarity ($S(.,.)$) (or dissimilarity, $d(.,.)$) between the individual objects that are being clustered. It is important that the following properties are satisfied for the *distance* or *dissimilarity function* being used in CA.

1. $d(x_{i,k}, x_{j,k}) = d(x_{j,k}, x_{i,k})$, which is the property of symmetry.
2. $d(x_{i,k}, x_{j,k}) > 0$, if $x_{i,k} \neq x_{j,k}$, which is the property of positivity.

[*] Also independently developed by Sørensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, *Kongelige Danske Videnskabernes Selskab*, **5**, 1–34, 1957.

3. $d(x_{i,k}, x_{j,k}) = 0$, if $x_{i,k} = x_{j,k}$, which is the property of reflexivity.
4. $d(x_{i,k}, x_{j,k}) \leq d(x_{i,k}, x_{l,k}) + d(x_{l,k}, x_{j,k})$, which is generally called the triangle law.

Where $x_{i,k}$, $x_{j,k}$, and $x_{l,k}$ are some points in space. If along with the first two, the third and fourth property also holds for $d(.,.)$, then $d(.,.)$ is a *metric*. In line with *distance* or *dissimilarity* function, a *similarity* function, $S(.,.)$, can also be defined with the following properties, which are on similar lines as mentioned for $d(.,.)$:

1. $S(x_{i,k}, x_{j,k}) = S(x_{j,k}, x_{i,k})$, which is the property of symmetry.
2. $0 \leq S(x_{i,k}, x_{j,k}) \leq 1$, which is the property of positivity.
3. $S(x_{i,k}, x_{j,k})S(x_{j,k}, x_{l,k}) \leq \{S(x_{i,k}, x_{j,k}) + S(x_{j,k}, x_{l,k})\}S(x_{i,k}, x_{l,k})$, which is the property of reflexivity.
4. $S(x_{i,k}, x_{j,k}) = 1$, iff $x_{i,k} = x_{j,k}$.

Remember, $S(.,.)$ is called *similarity metric* if all the above four properties hold. If the original data was collected as similarities, then a suitable monotone decreasing function may be used to convert them to dissimilarities.

Typically, distance/dissimilarity functions are used to measure continuous features, whereas similarity functions are more appropriate for qualitative variables. Table 8.6 gives the similarity as well as dissimilarity measure for quantitative features/characteristics.

### 8.20.1 Clustering Algorithms

A widely agreed framework is to classify clustering as hierarchical clustering and partitioning clustering, based on the properties of the clusters generated. While generating the clusters, the concept of distance as a measure which groups objects into clusters with certain properties with respect to the idea of distance and its functional form comes in play. Most of the algorithms assume symmetric dissimilarity matrices. In case the original matrix $\boldsymbol{D}_{n\times n}$ is not symmetric, then one can replace the matrix by $(1/2)\left(\boldsymbol{D}_{n\times n} + \boldsymbol{D}^T_{n\times n}\right)$. The reader should remember that clustering algorithms may be classified as (i) exclusive clustering, (ii) overlapping clustering, (iii) hierarchical clustering, and (iv) probabilistic clustering. Without going into detailed analysis, we give here the pseudo-codes of a few of the clustering algorithms, so that it motivates the reader to understand them and do a thorough search of such algorithms which may be found in good references, a few of which have already been stated in due course of our discussion of CA.

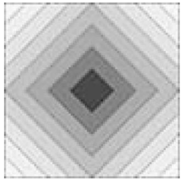
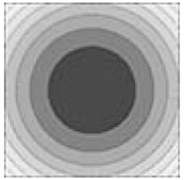
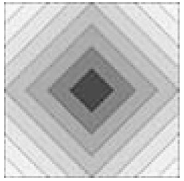
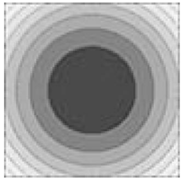
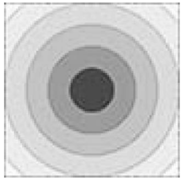
*Basic K-mean algorithm:* The $K$-mean clustering algorithm works on the premise that centroids of a group of objects best depict the characteristics of that group/cluster. The pseudo-code for the $K$-mean clustering algorithm is as follows and is shown in Figure 8.19.

*Bisecting K-mean algorithm:* The bisecting $K$-mean algorithm is a simple extension of the basis $K$-mean algorithm. The pseudo-code for this algorithm is illustrated in Figure 8.20. The idea is to obtain $K$ clusters and split the set of points into two clusters and then select one of them to split it again. We continue doing this until $K$ clusters are obtained.

*Basic agglomerative hierarchical clustering algorithm:* It is a hierarchical clustering algorithm, whereby we start with points as individual clusters and at each step merge the closet pairs of clusters. Hence, a cluster proximity function is important, which needs to be defined before one ventures to use this clustering algorithm. For the benefit of the reader, the pseudo-code for this third algorithm is given in Figure 8.21.

**TABLE 8.6**

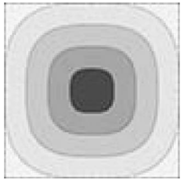
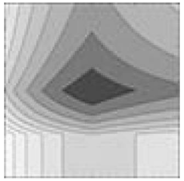
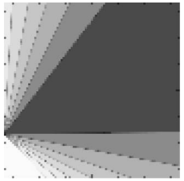
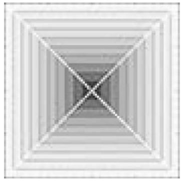
Similarity as Well as Dissimilarity Measure for Quantitative Features/Characteristics

| Measure | Mathematical Expression | Comments | Examples/ Applications | Pictorial Representation ($p = 2$ Where Vertical Axis Is $x_{.,1}$ and Horizontal Axis Is $x_{.,2}$) |
|---|---|---|---|---|
| Manhattan distance/taxi cab norm/sum of absolute difference (SAD) of the difference | $d(x_i, x_j) = \sum_{k=1}^{p} \lvert x_{i,k} - x_{j,k} \rvert$ | • Equivalent to the $L_1$ norm Tends to form hyperrectangular cluster<br>• Special case of Minkowski distance when $r = 1$ | Fuzzy adaptive resonance theory |  |
| Sum of square distance (SSD) | $d(x_i, x_j) = \sum_{k=1}^{p} \lvert x_{i,k} - x_{j,k} \rvert^2$ | • Equivalent to the $L_2$ norm Tends to form hyperspherical cluster<br>• Special case of Minkowski distance when $r = 2$ | $K$-mean algorithm |  |
| Mean absolute error (MAE), that is, normalized version of SAD | $d(x_i, x_j) = \frac{1}{p} \sum_{k=1}^{p} \lvert x_{i,k} - x_{j,k} \rvert$ | • Tends to form hyperrectangular cluster | – |  |
| Mean squared error (MSE), that is, normalized version of SSD | $d(x_i, x_j) = \frac{1}{p} \sum_{k=1}^{p} \lvert x_{i,k} - x_{j,k} \rvert^2$ | • Tends to form hyperspherical cluster | – |  |
| Euclidean distance | $d(x_i, x_j) = \left( \sum_{k=1}^{p} \lvert x_{i,k} - x_{j,k} \rvert^2 \right)^{1/2}$ | • Equivalent to the $L_2$ norm Tends to form hyperspherical cluster | $K$-mean algorithm |  |

(*Continued*)

**TABLE 8.6 (*Continued*)**

Similarity as Well as Dissimilarity Measure for Quantitative Features/Characteristics

| Measure | Mathematical Expression | Comments | Examples/ Applications | Pictorial Representation ($p = 2$ Where Vertical Axis Is $x_{.,1}$ and Horizontal Axis Is $x_{.,2}$) |
|---|---|---|---|---|
| Minkowski distance | $d(x_i, x_j) = \left(\sum_{k=1}^{p} \|x_{i,k} - x_{j,k}\|^r\right)^{1/r}$ | • Invariant to translation/rotation<br>• Features with large values/variances tend to dominate | Fuzzy $c$-means |  |
| Canberra distance | $d(x_i, x_j) = \sum_{k=1}^{p} \frac{\|x_{i,k} - x_{j,k}\|}{\|x_{i,k}\| + \|x_{j,k}\|}$ | – | – |  |
| Cosine distance | $S(x_i, x_j) = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}$ | • Independent of vector length<br>• Invariant to rotation<br>• Noninvariant to linear transformation | Document clustering |  |
| Maximum distance/ chessboard distance/ Chebyshev distance | $d(x_i, x_j) = \max_{1 \le k \le d} \|x_{i,k} - x_{j,k}\|$ | • Equivalent to the $L_\infty$ norm<br>• Special case of Minkowski distance when $r = \infty$ | Fuzzy $c$-means with sup norm |  |

(*Continued*)

**TABLE 8.6 (*Continued*)**

Similarity as Well as Dissimilarity Measure for Quantitative Features/Characteristics

| Measure | Mathematical Expression | Comments | Examples/ Applications | Pictorial Representation ($p = 2$ Where Vertical Axis Is $\boldsymbol{x}_{.,1}$ and Horizontal Axis Is $\boldsymbol{x}_{.,2}$) |
|---|---|---|---|---|
| Pearson's correlation coefficient | $d(x_i, x_j) = \frac{1}{2}\left\{1 - \frac{\sum_{k=1}^{p}(x_{i,k} - \bar{x}_i)(x_{j,k} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{p}(x_{i,k} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{p}(x_{j,k} - \bar{x}_j)^2}}\right\}$ | • Not a metric<br>• Derived from correlation coefficient<br>• Magnitude of differences of variables not considered | Analysis of gene expression data | |
| Spearman's correlation coefficient | $d(x_i, x_j) = 1 - 6\left(\sum_{k=1}^{p}(x_{i,k} - x_{j,k})^2/n(n^2 - 1)\right)$ | • Not a metric<br>• Derived from correlation coefficient<br>• Magnitude of differences of variables not considered | – | – |
| Mahalanobis distance | $d(x_i, x_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{S}(\boldsymbol{x}_j - \boldsymbol{x}_i)$, where $\boldsymbol{S}$ is within covariance matrix | • Invariant to any nonsingular linear transformation<br>• $\boldsymbol{S}$ is based on all the objects<br>• Tend to form hyperellipsoid clusters<br>• When correlation is zero, then squared Mahalanobis distance is equal to squared Euclidean distance | Ellipsoidal adaptive resonance theory<br>Hyperellipsoidal clustering algorithm | – |

```
DEFINE: $K$ points
INPUT: $K$ points as initial centroids
REPEAT
    Form $K$ clusters by assigning each point to its closest centroid
    Recompute the centroid of each cluster
UNTIL:Cenroids do not change
```

**FIGURE 8.19**
Pseudo-code for $K$-mean algorithm.

```
DEFINE: $K$ points
INPUT: $K$ points and initialize the list of clusters so that they consists of
cluster containing all points
REPEAT:
  Remove a cluster from the list of clusters and perform several bisection of the cluster
  choosen
  FOR: i:number of trials do
    Bisect the selected cluster using basic $K$-means
  END FOR
  Select two clusters from the bisection based on some criteria (e.g., SSE, etc.)
  Add these two clusters to the list of clusters
UNTIL:List of clusters contains $K$ clusters
```

**FIGURE 8.20**
Pseudo-code for bisecting $K$-mean algorithm.

```
DEFINE: $K$ points, $\boldsymbol{D}$ proximity matrix
INPUT: $K$ points and $\boldsymbol{D}$ proximity matrix
REPEAT
  Merge the closest two clusters based on the proximity property defined
  Update the proximity matrix
  Bisect the selected cluster using basic $K$-means
UNTIL: Only one cluster remains
```

**FIGURE 8.21**
Pseudo-code for basic agglomerative hierarchical clustering algorithm.

This chapter is neither an exclusive discussion about clustering algorithms, nor about cluster analysis; hence, we desist ourselves from analyzing other algorithms along with their merits and demerits. We request the readers to check any good book in the area of clustering algorithm, a few examples of which are: Abonyi and Feil (2007), Everitt et al. (2011), and Höppner et al. (1999), to name a few.

*Data set # 1:* One can consider the data set given in Chapter 13 of Hartigan (1975), which relates to 13 different Indo-European languages equivalent of the names associated with certain common words. The data set has 13 rows and 17 columns (16 different words, e.g., black, eat, drink, fish, five, etc.). The interested reader can use any clustering technique to find the groups accordingly to which the languages can be clustered depending on these 16 different words, their pronunciation, diction, speech, etc.).

*Data set # 2:* Another interesting application can be the study of mutation sequence of amino acids in different species such as man, monkey, chicken, duck, kangaroo, and rattlesnake. Again, the data for the same can be referred to in Hartigan (1975). For this problem, we would try to group the species depending on the different characteristics one thinks are important and are found in the gene sequences related to amino acid.

Thus, in a nutshell, CA is ideally suited for defining groups of individuals/objects with maximal homogeneity within the groups, while also having maximum heterogeneity between groups, that is, determining the most similar groups that are also most different from each other. One difference between discriminant analysis and logistic regression and classification tree analysis is the fact that discriminant analysis and logistic regression make a number of assumptions about the underlying data, whereas classification tree analysis is a nonparametric technique. Another difference is that discriminant analysis and logistic regression can be used to derive probabilities of group membership for individuals, whereas classification tree analysis only produces average probabilities for the different groups.

## 8.21 Multiple Discriminant and Classification Analysis

Before discussing what multiple discriminant and classification analysis (MDCA) is all about, let us consider a few relevant examples to aid a better understanding of this multivariate statistical method, which consists of (i) a separation/discrimination rule along with (ii) an allocation/classification rule.

For a good motivation, let us consider a few practical applications of the MCDA technique. First, consider a pediatrician has with her the data of *height*, *age*, *sex*, and *age* of children (in the age group of 2–5 years) from the past. Based on this information, a child is categorized as being obese, normal, or malnourished. When a new patient (in this case, a child) visits the pediatrician, then she, that is, the pediatrician, has to categorize the child as being in any one of the above-mentioned *three* categories, such that she can suggest medical care for the child as deemed appropriate. As a second example of application, take into account a musicologist who is studying the composition written by composers between 1750 and 1820 AD (which is termed as the classical period of Western music). The musicologist is aware that a classification of the composer in that period may be made depending on the *melody*, *rhythm*, *dynamics*, *mood*, *timbre*, etc. When the musicologist is given a new piece or an unknown piece, then, depending on the characteristics of the musical piece, he/she may classify the musical piece as that belonging to Joseph Hayden, Wolfgang Amadeus Mozart, Ludwing van Beethoven, etc. Continuing our discussion further, next, think of a geologist who classifies rock as igneous, sedimentary, or metamorphic, depending on its *chemical composition*, *physical properties*, *texture* of its constituent particles, and *permeability*. Suppose the geologist is given a new sample of rock, and is told to classify the given sample, then he/she may do so as desired using MDCA. Finally, let us illustrate a fourth application where you as a credit risk analyst are first interested to discriminate a company as *good*, *average*, or *bankrupt*, depending on *price to earning ration* $\left(\frac{P}{E}\right)$, *amount of liability*, and *price of stock*. After having done that, you study the credentials of different new companies and classify them as belonging to any one of the categories as already decided.

In all these four application areas of MCDA, the essence of what one intends to achieve may be summarized as follows. Given observations/objects, as a *first* step, we separate/discriminate the observations/objects into clusters/groups, which are known *a priori*. Afterward, when a new set of observations/objects arrive, we intend to classify them into these known groups. Hence, the immediate goal of MCDA is as follows:

*Step # 1: Separation/discrimination rule:* In this step/rule, describe an algebraic or graphical rule such that one is able to differentiate observations/objects into different classes, depending on different characteristics/features, which are inherent in the observations/objects.

*Step # 2: Allocation/classification rule:* Once the first step is over, our next job is to sort out the new observations/objects into one of these classes depending on some logic/rule.

Thus, in MDCA, we first define $\Pi_j$, $j = 1, \ldots, J$, number of clusters/classes each with probability density function $f_j(x)$ such that all the observations *a priori* may be classified in any one of the $J$ cluster/classes. Later, when a new set of observation/object arrive, we categorize that observations/objects, say $x$, into any one of the cluster/group. One should remember that the allocation/classification rules are developed based on the fact that the measured characteristics/features of randomly selected observations/objects are such that the possible sample space ($\Omega = \mathbb{R}^p$) is divided into $R_j$ *disjoint* sets/regions, such that $\cup_{j=1}^{J} R_j = \mathbb{R}^p$ holds (Figure 8.22).

Now, if the new observation/object falls in $R_j$, then it is characterized as belonging to population, $\Pi_j$. For the classification/allocation rule, there are several situations that make it difficult to complete Step # 2 with minimum error: (i) incomplete information or knowledge of future performance; (ii) perfect information, and (iii) unavailable or expensive information. However one may try, there are always some misclassification/misallocation errors, and hence our main job is to have a good classification/allocation rule that results in a few of these errors. Apart from that, an optimal classification/allocation rule should also take into account both (i) *prior probabilities of occurrence* as well as (ii) *cost of classification/allocation*. Thus, the conditional probability of allocating/classifying an object/observation as $\Pi_i$ when in fact it belongs to $\Pi_j$ is $p_{i,j} = Pr(X \in R_i|\Pi_j) = \int I_{x\in R_i} f_j(x)dx, i \neq j = 1, \ldots, J$. Moreover, if the prior distributions are $\pi_j(x)$, then we have different cases which would dictate how the probability values would be calculated. One should remember that the concept of prior is practical and also logical as it gives the most likelihood case, that when a new observation is chosen, what is the probability that it will belong to a particular cluster/group? Considering there are $J$ number of clusters/groups, classifying $x$ in cluster/group $i$, although it is from cluster/group population $\Pi_j$, may be given by the conditional distribution $p_{i,j} = Pr(X \in R_i|\Pi_j) = \int I_{x\in R_i} f_j(x)dx, i \neq j = 1, \ldots, J$. Now, these $R_j$ correspond to the $\Pi_j$ population and one is interested to categorize an observation, say $x$, into any one of the group. In doing so, we entail a cost that is written as $C(i|j), i, j = 1, \ldots, J$ such that it signifies the cost of assigning the observation in the $i$th population, that is, $\Pi_i$, when actually it should belong to the $j$th, that is, $\Pi_j$ population. Hence, the cost structure matrix may be given as shown in Figure 8.23, where $C(j|j) = 0, \forall j = 1, \ldots, J$.

Hence, the total gain/benefit given that the observation chosen belongs to population $\Pi_j$ may be expressed as $TG(R_j) = -C(j|1)\pi_1 \int I_{x\in R_j} f_1(x)dx - \cdots - C(j|j-1)$ $\pi_{j-1} \int I_{x\in R_j} f_{j-1}(x)dx - C(j|j+1)\pi_{j+1} \int I_{x\in R_j} f_{j+1}(x)dx - \cdots - C(j|J)\pi_J \int I_{x\in R_j} f_J(x)dx$. Let us further define the $\Pi_j$ population $\left(\mathbf{x}_{j1}, \ldots, \mathbf{x}_{jn_j}\right) \sim f_j(x)$, where $f_j(x)$ is the probability density function of the $j$th cluster/group. Then, the maximum likelihood discriminate rule would intuitively assign $x$ to $\Pi_j$ such that the likelihood function $L_j(x) = f_j(x)$ is maximized. Mathematically, the sets/regions, $R_j$, would be given by $R_j = \{x : L_j(x) > L_i(x), i = 1, \ldots, J, i \neq j\}$. Obviously, there is a *misclassification penalty* for assigning an observation in the wrong cluster/group, and hence in the same spirit there would be a *nonnegative penalty* for right classification of

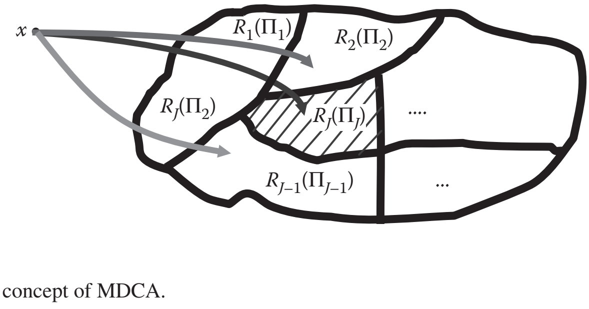


**FIGURE 8.22**
Pictorial illustration of the concept of MDCA.

$$
\begin{array}{c|ccccc|}
 & \multicolumn{5}{c}{\text{Classified population}} \\
 & C(1|1) & C(2|1) & \cdots & C(J-1|1) & C(J|1) \\
 & C(1|2) & C(2|2) & \cdots & C(J-1|2) & C(J|2) \\
\text{True population} & \vdots & \cdots & \ddots & \cdots & \vdots \\
 & C(1|J-1) & C(2|J-1) & \cdots & C(J-1|J-1) & C(J|J-1) \\
 & C(1|J) & C(2|J) & \cdots & C(J-1|J) & C(J|J)
\end{array}
$$

**FIGURE 8.23**
Cost structure matrix for MDCA.

the same. To design a rule, let us assume $\gamma_j$ and $\delta_j$ as the gain or loss in correct or erroneous classification of the observation, $x$, into the $j$th cluster/group. Furthermore, let the cost function be denoted as $C(i|j), i, j = 1, \ldots, J$. Intuitively one can easily comment that $C(j|j) = 0$ as we classify the observation in its right cluster/group and hence there is no error. Thus, the total gain/loss is given by the following equation: $TC(R_j) = -C(j|1)\pi_1 \int I_{x \in R_j} \times f_1(x)dx$. If one wants to diagrammatically represent the classification/misclassification probabilities, then one can refer to Figure 8.24 (refer Johnson and Wichern, 2002).

Before we discuss an example, let us highlight a few important points for the MCDA method, which may be useful for the readers to appreciate this multivariate statistical method in a much better way:

1. The MDCA method is appropriate when the dependent variable is categorical, while the independent variables are metric.
2. MDCA derives the variate that best distinguishes between *a priori* groups.
3. MDCA sets variate's weights to maximize between-group variance relative to within-group variance.

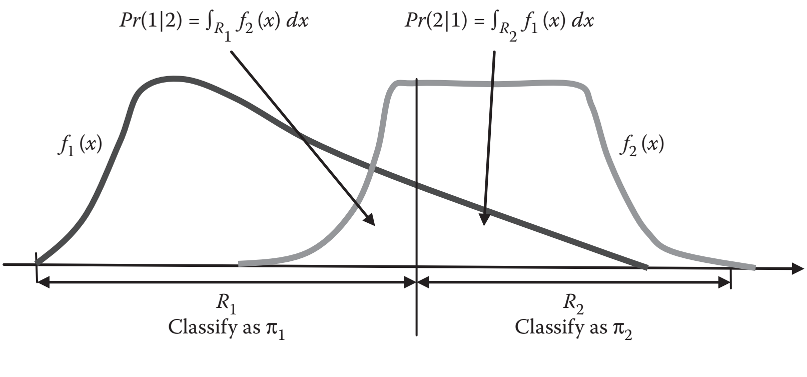

$Pr(1|2) = \int_{R_1} f_2(x)\,dx$
$Pr(2|1) = \int_{R_2} f_1(x)\,dx$
$f_1(x)$
$f_2(x)$
$R_1$
Classify as $\pi_1$
$R_2$
Classify as $\pi_2$

**FIGURE 8.24**
Misclassification probability for two regions marked as $\pi_1$ and $\pi_2$. (Adapted from Johnson, R.A. and Wichern, D.W. 2002. *Applied Multivariate Statistical Analysis*, Pearson Education, ISBN: 8178086867.)

Without being repetitive, we would like to mention the following important points for the MDCA method. It should be remembered that for each observation, we can obtain a discriminant $Z$-score, such that the average $Z$-score for a group gives the centroid for the group. Furthermore, the classification done using cutting scores which are derived from group centroids and finally the statistical significance of discriminant function is accomplished using distance between group centroids. Finally, to conclude this section, before discussing an example, we state the general steps one should remember for the MDCA method:

1. *Step # 1: Research problem/objectives*: In the research problem and objective formulation stage, one needs to do the following: (i) evaluate the differences between average scores for *a priori* groups on a set of variables; (ii) determine which set of independent variables account for most of the differences between groups; and finally (iii) classify the observations into groups.
2. *Step # 2: Research design*: The second stage is the research design stage in which the important things to remember are (i) there should be a proper selection of dependent as well as independent variables; (ii) the sample size considerations should be done appropriately, and finally (iii) the division of sample into analysis and holdout sample should be done rightly.
3. *Step # 3: Assumptions of MDCA*: As for any statistical method, MCDA also has some inherent assumptions, which are (i) multivariate normality for the independent variables; (ii) equal covariance matrices for the groups; (iii) low correlation among independent variables; and finally (iv) linear nature of the discriminant function.
4. *Step # 4: Estimation of MDCA and assessing fit*: In the fourth stage, that is, the estimation stage, one should remember that the estimation process can be either simultaneous or stepwise and to test the statistical significance of the discriminant function, one should use existing statistics such as Wilk's lambda, Hotelling's trace, Pillai's criterion, Roy's greatest root, and Mahalanobis distance function to test the efficacy of the data set as well as the MCDA method. It is important to note that the test statistic signifies the overall discrimination between groups and of each discriminant function. Moreover, to assess the overall fit, one should calculate the discriminant $Z$-score for each observation and then evaluate the group differences on $Z$-scores and predict the group membership accurately. To do this, we need to address the following rationale for classification matrices: (i) cutting score determination; (ii) considering costs of misclassification; (iii) constructing classification matrices; (iv) assessing classification accuracy; and (v) proper casewise diagnosis.
5. *Step # 5: Interpretation of results*: For this penultimate stage, remember the following and they are related to the interpretation of results, which can be further broken into those related to (i) methods for single discriminant function; (ii) discriminant weights; (iii) discriminant loadings; and (iv) partial $F$-values. Additional methods for more than two functions may be required to be used, and they are: (i) rotation of discriminant functions; (ii) potency index; and (iii) stretched attribute vectors.
6. *Step # 6: Validation of results*: Finally, for the validation stage, one needs to analyze results with what is practical/feasible, and then based on the feedback, conduct further tests if required.

A few relevant references are: Duda et al. (2001), Härdle and Simar (2007), and Lachenbruch (1975).

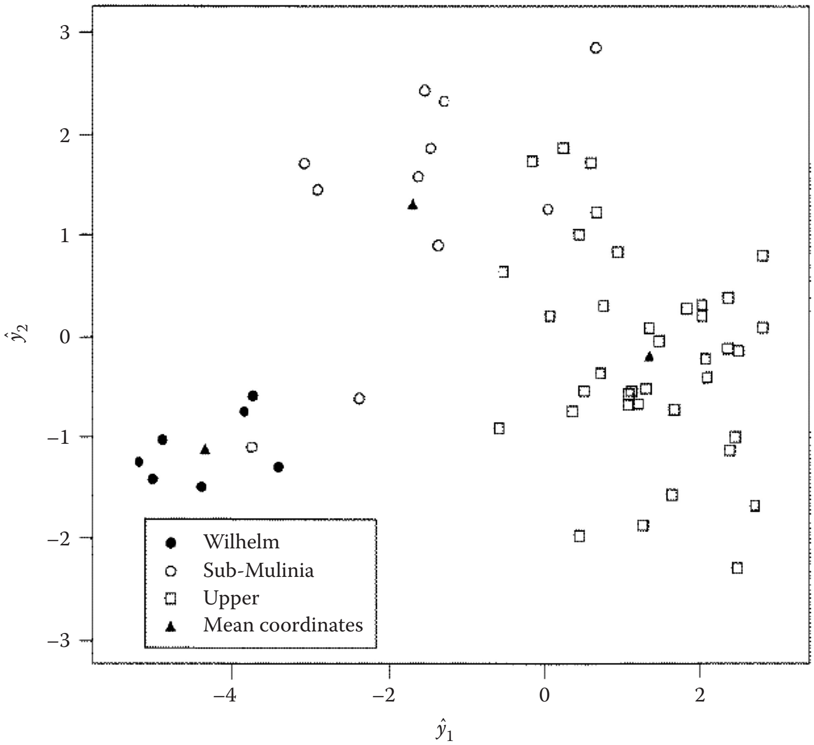


**FIGURE 8.25**
Crude oil sample in the discriminant space, Example 8.22. (Adapted from Johnson, R.A. and Wichern, D.W. 2002. *Applied Multivariate Statistical Analysis*, Pearson Education, ISBN: 8178086867.)

**Example 8.22: (Johnson and Wichern, 2002)**

As an example for the MDCA, let us consider the data set related to crude oil samples study by Gerrild and Lantz (1969). The crude obtained may be assigned to any one of the three populations, namely, $\pi_1$, $\pi_2$, and $\pi_3$, which are Wilhelm stone, sub-Mulinia sandstone, and upper sandstone, respectively. We are interested to study the characteristics, namely, vanadium, iron, beryllium, saturated hydrocarbon, and aromatic hydrocarbon, which may be denoted by $X_1$, $X_2$, $X_3$, $X_4$, and $X_5$. One can calculate the eigen values as 4.354 and 0.559. Finally, the Fisher linear discriminants are given by (i) $\hat{y}_1 = 0.312(x_1 - 6.180) - 0.710(x_2 - 5.081) + 2.764(x_3 - 0.511) + 11.809(x_4 - 0.201) - 0.235(x_5 - 6.434)$ and (ii) $\hat{y}_2 = 0.169(x_1 - 6.180) - 0.245(x_2 - 5.081) - 2.046(x_3 - 0.511) - 24.453(x_4 - 0.201) - 0.378(x_5 - 6.434)$. The crude oil sample in the discriminant space is illustrated in Figure 8.25. One can also find the Fisher values of discriminants which we omit and request the readers to study and do the necessary calculation as required.

## 8.22 Multidimensional Scaling

Like PCA, multidimensional scaling (MDS) is also a dimension reduction technique. Classic Torgerson metric MDS is actually done by transforming distances into similarities and performing PCA. Thus, PCA might be called the algorithm of the simplest MDS. MDS and PCA are not at the same level to be in line or opposite to each other. PCA is just a method while MDS is a class of analysis. As mapping, PCA is a particular case of MDS. On the other hand, PCA is a particular case of FA, which, being a data reduction, is more than only mapping, while MDS is only a mapping. Furthermore, PCA as a technique projects a multidimensional space onto direction of maximum variability, whereas, in MDS, the multidimensional space is projected while at the same time maintaining the interpoint distances.

In PCA, we use the concept of covariance matrix to study the correlation between design variables, and this is summarized using *dot* products, while, in MDS, one uses distance and loss function in order to study the similarity/dissimilarity, which is summarized using *cross* product.

Consider $\boldsymbol{X}_{n\times p} = (X_{i,j})$, with $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$ such that $\boldsymbol{X}_{n\times p}$ signify a matrix corresponding to $n$ number of readings where each of the readings are of dimension $p$. Our aim using MDS is to calculate interpoint distances $\delta(X)_{i_1,i_2} = \|\boldsymbol{X}_{i_1} - \boldsymbol{X}_{i_2}\|$, $i_1$, $i_2 = 1, 2, \ldots, n$ and then try to find $k(k \leq p)$-dimensional vector $\boldsymbol{Y}_i$, $i = 1, 2, \ldots, n$ with $\delta(Y)_{i_1,i_2} = \|\boldsymbol{Y}_{i_1} - \boldsymbol{Y}_{i_2}\|$, such that $\delta(X)_{i_1,i_2} \approx \delta(Y)_{i_1,i_2}$ for all $i_1$, $i_2 = 1, 2, \ldots, n$. The proximity measure, $\delta(X)_{i_1,i_2}$, need not be Euclidean or any distance measure as such. It can also be an error and in the general sense, the proximity measure is described as similarity or dissimilarity.

Before we solve a simple problem, one should be aware that the classical scaling concepts used in the literature are either ordinal or metrical. Furthermore, there can be many dimensions based on which distance measure has to be calculated and here is where the weighting function can be utilized to scale the distances and also to draw them in a 2-D scale.

**Example 8.23: (Johnson and Wichern, 2002)**

As an example for MDS, consider the distance matrix between 10 cities, namely, London, Berlin, Oslo, Moscow, Paris, Rome, Beijing, Istanbul, Gibraltar, and Reykjavik. The 2-D plot of the cities based on the distances is shown in Figure 8.26. There are a few important points that one needs to mention here related to multidimensional scaling: (i) the points can be reflected without changing the interpoint distances; (ii) the interpoint distances are not affected if one changes the origin by adding/subtracting a constant from the rows/columns; and (iii) the set of points can be rotated without affecting the interpoint distances.

## 8.23 Structural Equation Modeling

Many of the multivariate statistical methods such as multiple regression, FA, MANOVA, etc. suffer from one common limitation, which is to do with the fact that they can examine only a single relationship at a time. One method that is able to overcome this lacuna is structural equation modeling (SEM). SEM is an extension of several multivariate techniques notably multiple regression and FA, and is distinguished by two characteristics:

1. Estimation of multiple and interrelated dependence relationship.
2. Ability to represent unobserved concepts in these relationships and also the amount of measurement error in the estimation process.

Another interesting fact is that this topic under multivariate statistical analysis is the only one to have an exclusive peer-reviewed journal by the same name, which is *Structural Equation Modeling: A Multidisciplinary Journal*, Taylor & Francis, ISSN: 1070-5511 (Print) and 1532-8007 (Online). The origins of modern SEM is usually traced to biologist Sewall Wright's development of path analysis

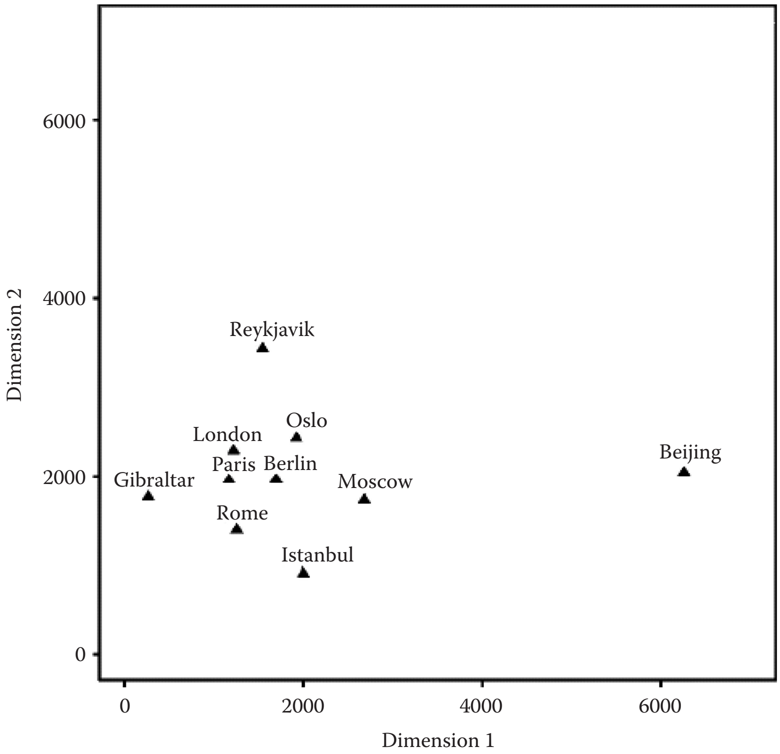


**FIGURE 8.26**
2-D plot of distances for the 10 cities, Example 8.23. (Adapted from Johnson, R.A. and Wichern, D.W. 2002. *Applied Multivariate Statistical Analysis*, Pearson Education, ISBN: 8178086867.)

(e.g., Wright, 1921, 1934). Another precursor of SEM is the path model (Duncan, 1966). If one follows the history of SEM, then a few good references that may be cited to trace the development of this statistical method are Bentler and Chou (1987), Bielby and Hauser (1977), Bollen (1989), and Bollen and Long (1993).

SEM as a statistical technique refers to the body of a comprehensive statistical methodology used to test and estimate the causal relations. The method uses a combination of cross-sectional statistical data and qualitative causal assumptions. It is different from another statistical technique, namely, multivariate linear regression model in the sense that the response variable in one regression equation in SEM may appear as a predictor in another equation. The variables in SEM may influence another reciprocally, either directly or through other variables. The proliferation of the use of SEM in social sciences, psychology, and related areas is due to the fact that implementation and the thinking process needed to actually theorize and practically apply this technique is akin to the informal thinking about causal relation that is common in the areas just mentioned.

The common aspects/concepts in SEM are: (i) model specification, (ii) estimation of free parameters, (iii) assessment of model and model fit, (iv) model modification, (v) sample size and power, and (vi) interpretation and communication.

We state a simple algorithm used for SEM, whereby we assume that latent variables are not present. First, let us consider $x_i$, $i = 1, \ldots, n$ points. Also assume the dependent variables are $x_j$, $j \in D$, while the predictor variables are $x_k$, $k \in P$ and $D \cap P \neq \emptyset$. Furthermore, consider there is a set $T$ such that $p = P - T$ are the actual predictor variables that are used. Our task is to find the

```
DEFINE: $x_i$, $i = 1,\cdots,n$, variables, $D$ as predictees set, $P$ as predictors set, $Q$ as queue for predictees and $M$ as model
INPUT: $x_i$, $i = 1,\cdots,n$, set of variables;$x_j$, $j \in D$, set of predictees and $x_k;k \in P$, set of predictors
SELECT:$x_0\,(x_0 \in D)$ and put it in set $Q$
CREATE: Empty model, $M$, of $n$ nodes with no links, $Q$ as queue for predictees which is initially empty
REPEAT
  Select $x_i \in D$
  Find $x_j \in P$, such that $P = \{x_j : x_j \neq x_i \; and \; x_j \rightarrow x_i\}$, i.e., $x_j$ is a predictor of $x_i$
  For all $x_j \in P$, add $x_j \rightarrow x_i$ into set $M$
  For all $x_j \in P$, add $x_j$ into set $Q$
UNTIL: Set $Q$ is empty
```

**FIGURE 8.27**
Pseudo-code for structural equation modeling.

set of variables in $p$ and $T$ for $x_i$, $i \in D$. The algorithms (Figure 8.27) use covariance information, in the form of estimated standardized regression coefficients, to direct the construction of SEMs and to estimate the parameters of the models. Latent variables can result in biased estimates; hence, the algorithm might give erroneous results when latent variables are considered.

In the literature, one finds differing approaches to assess the best fit. Traditional approaches to modeling start from a null hypothesis, rewarding more parsimonious models, to others such as AIC that focus on how little the fitted values deviate from a comparison model, taking into account the number of free parameters used. Because different measures of fit capture different elements of the fit of the model, it is appropriate to report a selection of different fit measures. A few commonly used measures of fit are: (i) chi-squared, (ii) AIC; (iii) root mean square error (RMSE); (iv) standardized root mean residual (SRMR); and (v) comparative fit index. Though not exhaustive, but a few useful references for SEM are: Bagozzi (1982), Grace (2006), Hancock and Mueller (2006), Kaplan (2000), Kline (1998), Muthén (1983), Rabe-Hesketh et al. (2004), Raykov and Marcoulides (2006), McDonald and Ho (2002), and Skrondal and Rabe-Hesketh (2004).

*Data set # 1*: SEM process can be thought of as a four-stage process, namely, (i) model specification, (ii) model estimation, (iii) model evaluation, and finally (iv) model modification. One can use the data from Stein et al. (2003). The main idea of the study is to model two separate paths for alcohol and drugs and then test them in which psychosocial, environmental, and sociodemographic variables is the prediction the best. The behavioral and substance abuse-related factors as well as the key outcome of positive attitudes about quitting drugs ($N = 620$) or alcohol ($N = 526$) in a sample of 709 homeless women are used. A positive attitude about quitting alcohol was predicted by more addiction symptoms, fewer positive effects from using alcohol, and not having a partner who uses alcohol. A positive attitude about quitting drugs was predicted by more drug problems, greater drug use in the past 6 months, more active coping, more education, less emotional distress, not having a partner who uses drugs, and fewer addiction symptoms. As one understands that the primary goal of the study is to determine if a set of items that query both alcohol and drug problems are adequate indicators for the two underlying constructs, namely, alcohol use problem and drug use problem. The reader is urged to study the data set and utilize the concepts of SEM to solve this problem.

To end the discussion of MVA, we give a few of the necessary information related to journals, data sets, etc., which we think would make this chapter more interesting both from the point of view of theory as well as application. Finally, a few references that we definitely think are important with respect to this chapter and related topics are: Agresti (2007), Aitchison (1986), Atchley and Bryant

(1975), Bishop et al. (1975), Carroll and Green (1997), Cheung (2015), Gifi (1990), Harris (1975), Kachigan (1991), Karson (1982), Krzanowski (1988, 1995), McCullagh and Nelder (1989), Van de Geer (1993), Whittaker (1990), and others. This section has been an honest attempt to look into MVA and multivariate statistics from a fresh perspective and try to deal with the theoretical as well as practical aspects of different methods from the point of view of decision analysis. To end this chapter, we discuss very briefly the future trends in statistics and big data analysis, which we think would be an ever-burgeoning field considering the ever-increasing applications of statistics and its tools in our everyday life.

## 8.24 Future Areas of Research

Today, science is passing through an era of transformation. Any decision-making process today is based on the efficient analysis of data available at hand. Science is driven by the data and it is being termed as data science. In the coming few decades, the most important areas of research will be on the analysis of big data. Big data usually indicate data sets whose sizes are beyond the ability of commonly used software tools to manage and analyze within a tolerable time limit. The term “big data” is a constantly moving target. As of today, the size of “big data set” ranges from few dozen terabytes to many petabytes of data. Interestingly, big data are available today in every sphere of life. Starting from industry, environment, health care, and government security, big data are being collected and stored everyday. This large complex, structured or unstructured, and heterogeneous data in the form of big data has gained significant attention. The velocity of the expansion of the amount of data gives rise to a complete paradigm shift in how new-age data are processed.

The age of data science is in its infancy and is experiencing a tactical evolution by leaps and bounds in all dimensions of science. Even though over the past few years a few robust big data models have come into existence, there is still a need for the pool to expand at a faster pace to meet the challenges of data proliferation. The concept of big data is relatively new and needs further research. Big data sets cannot be practically analyzed on a single commodity computer because their sizes are too large to fit in memory or it is too time consuming to process when the current statistical methods are used. To circumvent this obstacle, one may have to resort to parallel and distributed architectures, with multicore and cloud computing platforms providing access to hundreds or thousands of processors. While the parallel and distributed architectures present new capabilities for storage and manipulation of data, from an inferential point of view, it is unclear how the current statistical methodologies can be transported to the paradigm of big data. Big data have put a great challenge on the current statistical methodology.

There are several algorithms that are recently developed and feasible for statistical inference of big data and workable on parallel machines, including the bag of little bootstraps by Kleiner et al. (2014), aggregated estimating equation of Xi et al. (2009), split and conquer algorithms of Chen and Xie (2014), and the subsampling-based stochastic approximation algorithm by Liang et al. (2013). On the other hand, iterative algorithms have been widely used in current society of scientific computing, and it mainly includes Markov chain Monte Carlo (MCMC) algorithms and the EM algorithm, which typically requires a large number of iterations and a complete scan of the full data set for each iterations. Given the success of the iterative algorithms in modern scientific computing, it would be of great interest to develop some innovative iterative algorithms that are feasible for big data.

There have been significant advances made by the statistical community on big data research in the past few years. One of the open problems is how to generalize and scale up such proposed techniques to the true big data setting. One of the key features of big data is that the statistical methods, which work well on small-scale data set, usually perform poorly in big data setting. Some of the other open problems include (i) to have better understanding of big data and associated statistical issues, (ii) to think more carefully about how to solve big data issues, and (iii) to have a more concrete focus on big data problems. There are lots of challenges in this exciting field of research. More people need to come and join this active area of research for further development.

## References

Abonyi, J. and Feil, B. 2007. Cluster analysis for data mining and system identification, *Birkhäuser*, 978-3764379872.

Agresti, A. 2007. *An Introduction to Categorical Data Analysis*, Wiley, ISBN: 9780471226185.

Aitchison, J. 1986. *The Statistical Analysis of Compositional Data*, Chapman & Hall, ISBN: 978-94-010-8324-9.

Anderberg, M. R. 1973. *Cluster Analysis for Applications*, Academic Press, New York, ISBN: 0120576503.

Anderson, T. W. 2003. *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, ISBN: 9780471360919.

Arnold, S. F. 1981. *The Theory of Linear Models and Multivariate Analysis*, John Wiley & Sons, ISBN: 9780471050650.

Atchley, W. R. and Bryant, E. H. (Editors). 1975. *Multivariate Statistical Methods: Within Group Covariation*, Dowden Hutchinson & Ross, ISBN: 9780470035955.

Athreya, K. B. and Lahiri, S. N. 2006. *Measure Theory*, Hindustan Book Agency, New Delhi.

Bach, F. R. and Jordan, M. I. 2002. Kernel independent component analysis, *Journal of Machine Learning Research*, **3**, 1–48.

Bagozzi, R. P. (Editor). 1982. Special issue on causal modeling, *Journal of Marketing*, **19**, 403–584.

Bandyopadhyay, S., Rao, A. R., and Sinha, B. K. 2011. *Models for Social Networks with Statistical Applications*, SAGE Publications, Thousand Oaks, CA.

Barlett, M. S. 1954. A note on multiplying factors for various chi-squared approximations, *Journal of the Royal Statistical Society (Series B)*, **16**, 296–298.

Basilevsky, A. T. 1994. *Statistical Factor Analysis and Related Methods: Theory and Applications*, John Wiley & Sons, ISBN: 978-0-471-57082-0.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. 2009. *Robust Optimization*, Princeton University Press, ISBN: 9781400831050.

Ben-Tal, A. and Nemirovski, A. 1998. Robust convex optimization, *Mathematics of Operations Research*, **23**, 769–805.

Ben-Tal, A. and Nemirovski, A. 1999. Robust solutions to uncertain linear programs, *Operations Research Letters*, **25**, 1–13.

Ben-Tal, A. and Nemirovski, A. 2002. Robust optimization—Methodology and applications, *Mathematical Programming, Series B*, **92**, 453–480.

Bentler, P. M. and Chou, C. P. 1987. Practical issues in structural modeling, *Sociological Methods & Research*, **16**, 78–117.

Bertsimas, D. and Sim, M. 2003. Robust discrete optimization and network flows, *Mathematical Programming*, **98**, 49–71.

Bertsimas, D. and Sim, M. 2004. The price of robustness, *Operations Research*, **52**, 35–53.

Bertsimas, D. and Sim, M. 2006. Tractable approximations to robust conic optimization problems, *Mathematical Programming*, **107**, 5–36.

Bielby, W. T. and Hauser, R. M. 1977. Structural equation models, *Annual Review of Sociology*, **3**, 137–161.

Billingsley, P. 1995. *Probability and Measures*, 3rd edition, John Wiley, New York.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. 1975. *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, ISBN: 9780262520409.

Bock, R. D. 1975. *Multivariate Statistical Methods in Behavioral Research*, McGraw-Hill, ISBN: 9780894980145.

Bollen, K. A. 1989. *Structural Equations with Latent Variables*, Wiley, ISBN: 978-0-471-01171-2.

Bollen, K. A. and Long, J. S. (Editors). 1993. *Testing Structural Equation Models*, Sage Publications, ISBN: 978-0803945074.

Carroll, J. D. and Green, P. E. 1997. *Mathematical Tools for Applied Multivariate Analysis*, Academic Press, ISBN: 978-0121609559.

Cattell, R. B. 1943. The description of personality: Basic traits resolved into clusters, *Journal of Abnormal and Social Psychology*, **38**, 476–506.

Chatfield, C. and Collins, A. J. 1980. *Introduction to Multivariate Analysis*, Chapman & Hall, ISBN: 978-0-412-16030-1.

Chen, X. and Xie, M. 2014. A split and conquer approach for extraordinary large data analysis, *Statistica Sinica*, **24**, 1655–1684.

Chen, Z., Aiyi, L., Qu, Y., Tang, L., Ting, N., and Tsong, Y. 2015. *Applied Statistics in Biomedicine and Clinical Trials Design*, Springer, New York.

Cherubini, U., Luciano, E., and Vecchiato, W. 2004. *Copula Methods in Finance*, John Wiley & Sons, ISBN 0470863447.

Cheung, M. W.-L. 2015. *Meta-Analysis: A Structural Equation Modeling Approach*, Wiley, ISBN: 978-1-119-99343-8.

Child, D. 2006. *The Essentials of Factor Analysis*, Continuum International, ISBN: 978-0826480002.

Chung, K. L. 1974. *A Course in Probability Theory*, Academic Press, New York.

Coles, S. G. 2001. *An Introduction to Statistical Modeling of Extreme Values*, Springer, ISBN: 978-1849968744.

Cooley, W. W. and Lohnes, P. R. 1971. *Multivariate Data Analysis*, John Wiley & Sons, ISBN: 978-0471170600.

Cuadras, C. M. and Rao, C. R. (Editors). 1993. *Multivariate Analysis: Future Directions 2*, North-Holland, ISBN: 978-0-444-81531-6.

Davis, C., Claridge, G., and Cerullo, D. 1997. Personality factors predisposing to weight preoccupation: A continuum approach to the association between eating disorders and personality disorders, *Journal of Psychiatric Research*, **31**, 467–480.

Davison, A. C. and Hinkley, D. V. 1997. *Bootstrap Methods and Their Applications*, Cambridge University Press, Cambridge, MA.

de Haan, L. and Resnick, S. I. 1977. Limit theory for multidimensional sample extremes, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **40**, 317–337.

Dempster, A. P. 1969. *Elements of Continuous Multivariate Analysis*, Addison-Wesley, ISBN: 978-0201014853.

Dempster, A. P., Laird, M. N., and Rubin, D. B. 1977. Maximum likelihood estimation from incomplete data via EM algorithm (with discussion), *Journal of the Royal Statistical Society*, Series B, **39**, 1–38.

Dice, L. R. 1945. Measures of the amount of ecologic association between species, *Ecology*, **26**, 297–302.

Dillon, W. R. and Goldstein, M. 1984. *Multivariate Analysis, Methods and Applications*, Wiley, ISBN: 0-471-08317-8.

Doob, J. L. 1953. *Stochastic Processes*, John Wiley, New York.

Driver, H. E. and Kroeber, A. L. 1932. Quantitative expression of cultural relationships, *University of California Publications in American Archeology and Ethnology*, **31**, 211–256.

Duda, R. O., Hart, P. E., and Stork, D. G. 2001, *Pattern Classification*, Wiley, New York, ISBN: 978-0-471-05669-0.

Duncan, O. D. 1966. Path analysis: Sociological examples, *The American Journal of Sociology*, **72**, 1–16.

Duran, B. S. and Odell, P. L. 1974. *Cluster Analysis: A Survey*, Springer-Verlag, New York, ISBN 978-3-642-46309-9.

Eaton, M. L. 1983. *Multivariate Statistics: A Vector Space Approach*, John Wiley & Sons, ISBN: 9780471027768.

Eaton, M. L. and Perlman, M. D. 1973. The non-singularity of generalized sample covariance matrices, *Annals of Statistics*, **1**, 710–717.

Everitt, B. S. and Dunn, G. 2001. *Applied Multivariate Data Analysis*, Hodder Arnold Publication, ISBN: 9780340741221.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. 2011. *Cluster Analysis*, John Wiley & Sons, ISBN: 978-0-470-74991-3.

Fisher, R. A. 1921. Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk, *Journal of Agricultural Science*, **11**, 107–135.

Galton, F. 1886, Regression towards mediocrity in hereditary stature, *Journal of the Anthropological Institute*, **15**, 246–263

Galton, F. 1889. *Natural Inheritance*, MacMillan and Company, London and New York. https://archive.org/details/naturalinherita03galtgoog>

Gerrild, P. M. and Lantz, R. J. 1969. *Chemical Analysis of 75 Crude Oil Samples from Pliocence Sand Units, Elk Hills Oil Field, California*, U.S. Geological Survey Open File Report.

Ghosh, J. K., Delampady, M., and Samanta, T. 2006. *An Introduction to Bayesian Analysis: Theory and Methods*, Springer, New Delhi.

Gifi, A. 1990. *Nonlinear Multivariate Analysis*, Wiley, ISBN: 0-471-92620-5.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. 1996. *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC Press, New York.

Giri, N. C. 2004. *Multivariate Statistical Analysis*, Marcel Dekker, ISBN: 0824747135.

Gittins, R. 1985. *Canonical Analysis: A Review with Application in Ecology*, Springer-Verlag, ISBN: 3-540-13617-7.

Gnanadesikan, R. 2011. *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley & Sons, ISBN: 0471161195.

Goldberg, D. 1989. *Genetic Algorithm in Search, Optimization and Machine Learning*, Addison-Wesley Professional, Reading, MA.

González, I., Déjean, S., Martin, P. G. P., and Baccini, A. 2008. CCA: An R package to extend canonical correlation analysis, *Journal of Statistical Software*, **23**, 1–14.

Gordon, A. E. 1981. *Classification: Methods for the Exploratory Analysis of Multivariate Data*, Chapman & Hall, New York, ISBN: 9780412228506.

Gorsuch, R. L. 1983. *Factor Analysis*, Lawrence Erlbaum Associates, ISBN: 978-0898592023.

Grace, J. B. 2006. *Structural Equation Modeling and Natural Systems*, Cambridge University Press, ISBN: 9780521546539.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. 2005. *Multivariate Data Analysis*, Pearson Education, ISBN: 9780130329295.

Hancock, G. R. and Mueller, R. O. 2006. *Structural Equation Modeling: A Second Course*, Information Age Publishing, ISBN: 978-1593110154.

Hanley, J. A. 2004, Transmuting women into men: Galton's family data on human stature, *The American Statistician*, **58**, 237–243

Härdle, W. K. and Simar, L. 2007. *Applied Multivariate Statistical Analysis*, Springer-Verlag, ISBN: 9783540722434.

Harman, H. H. 1976. *Modern Factor Analysis*, University of Chicago Press, ISBN: 978-0226316529.

Harris, R. J. 1975. *A Primer of Multivariate Statistics*, Academic Press, ISBN: 9780123272508.

Hartigan, J. A. 1975. *Clustering Algorithms*, Wiley, ISBN: 047135645X.

Hastie, T., Tibshirani, R., and Friedman, J. 2011. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, ISBN: 978-0387848570.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (Editors). 1991, *Fundamentals of Exploratory Analysis of Variance*, John Wiley & Sons, ISBN: 0-471-52735-1.

Holzinger, K. J. and Swineford, F. 1939. *A Study in Factor Analysis: The Stability of Bi-Factor Solution*, University of Chicago: Supplementary Educational Monographs, **48**.

Höppner, F., Klawonn, F., Kruse, R., and Runkler T. 1999. *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*, Wiley-Blackwell, 978-0471988649.

Hotelling, H. 1935. The most predictable criterion, *Journal of Educational Psychology*, **26**, 139–142.

Hotelling, H. 1936. Relationship between two set of variables, *Biometrika*, **28**, 321–377.

Huber, P. J. 1981. *Robust Statistics*, JohnWiley and Sons, New York.

Hsieh, W. W. 2000. Nonlinear canonical correlation analysis by neural networks, *Neural Networks*, **13**, 1095–1105.

Huberty, C. J. and Olejnik, S. 2006. *Applied MANOVA and Discriminant Analysis*, Wiley, ISBN: 978-0-471-46815-8.

Iversen, G. R. and Norpoth, H. 1987. *Analysis of Variance*, Sage Publications, ISBN: 0-8039-3001-1.

Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura, *Bulletin de la Société Vaudoise des Sciences Naturelles*, **37**, 547–579.

Jain, A. K. and Dubes, R. C. 1988. *Algorithms for Clustering Data*, Prentice-Hall, ISBN: 0-13-022278-X.

James, W. and Stein, C. 1961. Estimation with Quadratic Loss, Berkeley Symposium on Mathematical Statistics and Probability, *Proceeding of the fourth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, **1**, 361–379.

Jobson, J. D. 1991. *Applied Multivariate Data Analysis*, Springer-Verlag, ISBN: 978-1-4612-0955-3.

Jackson, J. E. 2003. *User's Guide to Principal Components*, Wiley-Interscience, ISBN: 978-0471471349.

Johnson, N. L. and Kotz, S. 1972. *Distributions in Statistics: Continuous Multivariate Distributions*, John Wiley & Sons, ISBN: 0471521620.

Johnson, R. A. and Wichern, D. W. 2002. *Applied Multivariate Statistical Analysis*, Pearson Education, ISBN: 8178086867.

Jolliffe, I. T. 2002. *Principal Component Analysis*, Springer-Verlag, ISBN: 978-0-387-22440-4.

Kachigan, S. K. 1991. *Multivariate Statistical Analysis: A Conceptual Introduction*, Radiu Press, ISBN: 978-0942154917.

Kaplan, D. 2000. *Structural Equation Modeling: Foundations and Extensions*, Sage Publications, ISBN: 978-1412916240.

Karamouz, M., Nazif, S., and Fallahi, M. 2010. Rainfall downscaling using statistical downscaling model and canonical correlation analysis: A case study. In: Palmer, R.N. (Ed.), *World Environmental and Water Resources Congress*, American Society of Civil Engineers, Reston, pp. 4579–4587.

Karson, M. J. 1982. *Multivariate Statistical Methods: An Introduction*, Iowa State University Press, ISBN: 9780813818450.

Kaufman, L. and Rousseeuw, P. J. 2005. *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, ISBN: 978-0-471-73578-6.

Kelley, P. R. and Barry, R. 1997. Sparse spatial autoregressions. *Statistics and Probability Letters*, **33**, 291–297.

Kempton, R. A. and Fox, P. N. 2012. *Statistical Methods for Plant Variety Evaluation*, Chapman & Hall, London.

Kendall, M. G. 1980. *Multivariate Analysis*, Hodder Arnold, ISBN: 978-085264264.

Kennedy, W. J. and Gentle, J. E. 1980. *Statistical Computing*, Marcel Dekker, New York.

Kirkpatrick, S., Gelatt, Jr, C. D., and Vecchi, M. P. 1983. Optimization by simulated annealing, *Science*, **220**, 671–680.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. 2014. A scalable bootstrap for massive data, *Journal of the Royal Statistical Society*, Series B, doi:10.1111.rssb.12050.

Kline, R. B. 1998. *Principles and Practice of Structural Equation Modeling*, The Guilford Press, ISBN: 978-1606238769.

Kotz, S., Balakrishnan, N., and Johnson, N. L. 2000. *Continuous Multivariate Distributions: Volume 1—Models & Applications*, Wiley-Blackwell, ISBN: 978-0471183877.

Kotz, S. and Nadarajah, S. 2000. *Extreme Vale Distributions*, Imperial College Press, ISBN: 978-1-86094-224-2.

Kotz, S. and Nadarajah, S. 2004. *Multivariate Distributions and Their Applications*, Cambridge University Press, ISBN: 0521826543.

Krishnaiah, P. R. (Editor). 1984. *Handbook of Statistics 1: Analysis of Variance*, Elsevier, ISBN: 978-0-444-85335-6.

Krzanowski, W. J. 1988. *Principles of Multivariate Analysis: A User's Perspective*, Oxford University Press, ISBN: 0-198-52211-8.

Krzanowski, W. J. (Editor). 1995. *Recent Advances in Descriptive Multivariate Analysis*, Oxford University Press, ISBN: 9780198522850.

Kshirsagar, A. M. 1972. *Multivariate Analysis*, Mercel Dekker, ISBN: 9780824713867.

Lachenbruch, P. A. 1975. *Discriminant Analysis*, Macmillan Publishers, ISBN: 9780028482507.

Lai, P. L. and Fyfe, C. 2000. Kernel and nonlinear canonical correlation analysis, *International Journal of Neural Systems*, **10**, 365–377.

Lawley, D. N. and Maxwell, A. 1971. *Factor Analysis as a Statistical Method*, Macmillan, ISBN: 978-0408701525.

Lehmann, E. H. and Casella, G. 1998. *Theory of Point Estimation*, 2nd edition, Springer, New York.

Leurgans, S. E., Moyeed, R. A., and Silverman, B. W. 1993. Canonical correlation analysis when the data are curves, *Journal of Royal Statistical Society: Series B*, **55**, 725–740.

Lewis, D. G. 1971. *The Analysis of Variance*, Manchester University Press, ISBN: 0-7190-0467-5.

Liang, F., Cheng, Y., Song, Q., Park, J., and Yang, P. 2013. Aresampling-based stochastic approximation method for analysis of large geostatistical data, *Journal of the American Statistical Association*, **108**, 325Ű339.

Linden, M. 1977. A factor analytic study of Olympic decathlon data, *Research Quarterly*, **48**, 562–568.

Louviere, J. 1988. *Analyzing Decision Making: Metric Conjoint Analysis*, SAGE Publications, ISBN: 9780803927575.

Mardia, K. V., Kent, J. T., and Bibby, J. M. 1979. *Multivariate Analysis*, Academic Press, ISBN: 012471252.

Marshall, A. W. and Olkin I. 1967. A multivariate exponential distribution, *Journal of American Statistical Association*, **61**, 30–44.

Marshall, A. W. and Olkin I. 1983. Domains of attraction of multivariate extreme value distributions, *Annals of Probability*, **11**, 168–177.

McCullagh, P. and Nelder, J. A. 1989. *Generalized Linear Models*, Chapman & Hall, ISBN: 978-0412317606.

McDonald, R. P. and Ringo Ho, M.-H. 2002. Principles and practice in reporting structural equation analyses, *Psychological Methods*, **7**, 64–82.

McLachlan, G. J. and Krishnan, T. 1997. *The EM Algorithm and Extensions*, John Wiley and Sons, New York.

Melzer, T., Reiter, M., and Bischof, H. 2001. Nonlinear feature extraction using generalized canonical correlation analysis canonical correlation analysis, *Artificial Neural Networks-ICANN 2001* (Editors G. Dorner, H. Bischof, and K. Hornik), **2130** of LNCS, 353–360, Springer.

Mendenhall, W. and Sincich, T. 2006. *Statistics for Engineering and the Sciences*, Pearson Publication, ISBN: 978-0131877061.

Morrison, D. F. 1990. *Multivariate Statistical Methods*, McGraw-Hill Ryerson, ISBN: 9780070431867.

Mosteller, F. and Tukey, J. W. 1977. *Data Analysis and Regression*, Addison-Wesley, Reading, MA.

Muirhead. R. J. 2005. *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, ISBN: 978-0471769859.

Mulaik, S. A. 2009. *The Foundations of Factor Analysis*, Chapman & Hall, ISBN: 978-1420099614.

Muthén, B. 1983. Latent variable structural equation modeling with categorical data, *Journal of Econometrics*, **22**, 48–65.

Nelder, J. A. and Mead, R. 1965. A simplex method for function minimization, *Computer Journal*, **7**, 308–313.

Nelsen, R. B. 2006. *An Introduction to Copulas*, Springer, ISBN 978-0-387-28678-5.

Orme, B. K. and King, W. C. 2006. *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research*. Research Publishers, Madison, WI, ISBN: 978-0972729741.

Osborne, M. R. 1972. Some aspects of non-linear least squares calculations, *Numerical Methods for Nonlinear Optimization* (Editor F. A. Lootsma), Academic Press, New York.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space, *Philosophical Magazine*, **2**, 559–572.

Pickards, J. 1981. *Multivariate Extreme Value Distributions*, *Proceedings 43rd Session International Statistical Institute*, Buenos Aires, 859–878.

Press, S. J. 1982. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, R. E. Krieger Publishing Company, ISBN: 9780882759760.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. 2004. Generalized multilevel structural equation modeling, *Psychometrika*, **69**, 167–190.

Raghavarao, D., Wiley, J. B., and Chitturi, P. 2011. *Choice Based Conjoint Analysis: Models and Designs*, CRC Press, Taylor & Francis, ISBN: 978-1-4200-9996-6.

Rao, C. R. 2008. *Linear Statistical Inference and Its Applications*, John Wiley & Sons, ISBN: 9780470316436.

Raykov, T. and Marcoulides, G. A. 2006. *A First Course in Structural Equation Modeling*, Lawrence Earlbaum Associates, ISBN: 0-8058-3569-5.

Roy, S. N. 1957. *Some Aspects of Multivariate Analysis*, Wiley, ISBN: 978-1124038551.

Roy, S. N., Gnanadesikan, R., and Srivastava, J. N. 1971. *Analysis and Design of Certain Quantitative Multiresponse Experiments*, Pergamon Press, ISBN: 9780080069173.

Rutherford, A. 2001. *Introducing ANOVA and ANCOVA: A GLM Approach*, Sage Publications, ISBN: 0-7619-5160-1.

Sahai, H. and Ageel, M. I. 2000. *The Analysis of Variance: Fixed, Random, and Mixed Models*, Springer, ISBN 978-1-4612-1344-4.

Scheffé, H. 1999. *The Analysis of Variance*, John Wiley & Sons, ISBN: 0-471-75834-5.

Searle, S. R., Casella, G., and McCulloch, C. E. 2009. *Variance Components*, John Wiley & Sons, ISBN: 0-470-00959-4.

Seber, G. A. F. 2004. *Multivariate Observations*, John Wiley & Sons, ISBN: 9780471691211.

Sen, P. K. 1986. Contemporary textbooks on multivariate statistical analysis: A panoramic appraisal and critique, *Journal of the American Statistical Association*, **81** (394), 560–564.

Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics*, JohnWiley and Sons, New York.

Sibuya, M. 1960. Bivariate extreme statistics, *Annals of Institute of Mathematical Statistics*, **11**, 195–210.

Skrondal, A. and Rabe-Hesketh, S. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Chapman & Hall, ISBN: 978-1584880004.

Späth, H. 1980. *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*, Ellis Horwood, ISBN: 978-0853121411.

Spearman, C. 1904. General intelligence, objectively determined and measured, *The American Journal of Psychology*, **15**, 201–292.

Srivastava, M. S. and Khatri, C. G. 1979. *An Introduction to Multivariate Statistics*, North Holland, ISBN: 9780444003027.

Stein, J. A., Dixon, E., Longshore, D., and Galaif, E. 2003. Predicting positive attitudes about quitting drug- and alcohol-use among homeless women, *Psychology of Addictive Behaviors*, **17**, 32–41.

Stuart, A., Ord, J. K., and Arnold, S. 1999. *Kendall's Advanced Theory of Statistics, Volume 2A: Classical Inference and the Linear Model*, Wiley Publications, ISBN: 978-0-470-68924-0.

Szakács, G., Annereau, J. P., Lababidi, S., Shankavaram, U., Arciello, A., Bussey, K. J., Reinhold, W. et al. 2004. Predicting drug sensitivity and resistance: Profiling ABC transporter genes in cancer cells, *Cancer Cell*, **6**, 129–137.

Takeuchi, K., Yanai, H., and Mukherjee, B. N. 1982. *The Foundations of Multivariate Analysis*, Halstead Press, ISBN: 0852269641.

Tanimoto, T. 1957. *An Elementary Mathematical Theory of Classification and Prediction*, IBM Technical Report.

Tatsuoka, M. M. 1988. *Multivariate Analysis: Techniques for Educational and Psychological Research*, Macmillan, ISBN: 978-0024191205.

Theus, M. and Urbanek, S. 2008. *Interactive Graphics for Data Analysis: Principles and Examples*, CRC Press, Boca Raton, FL.

Thompson, B. 2004. *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*, American Psychological Association, Washington DC.

Thurstone, L. L. 1931. Multiple factor analysis, *Psychological Review*, **38**, 406–427.

Thurstone, L. L. 1947. *Multiple-Factor Analysis*, The University of Chicago Press, Chicago, ISBN: 978-0226801094.

Tiago de Oliveira, J. 1958. Extremal distributions, *Revista der Faculdade de Ciencias de Lisboa: Series A*, **7**, 219–227.

Tiago de Oliveira, J. 1975. Bivariate and multivariate extreme distributions. In: G. P. Patil, S. Kotz, and J. K. Ord (eds), *A Modern Course on Statistical Distributions in Scientific Work, NATO Advanced Study Institutes Series*, Springer, Netherlands, **17**, 355–361.

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

Tryon, R. C. 1939. *Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality*, Edwards Brothers.

Tukey, J.W. 1977. *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

Turner, J. R. and Thayer, J. F. 2001. *Introduction to Analysis of Variance: Design, Analysis & Interpretation*, Sage Publications, ISBN: 0-8039-7074-9.

Vinod, H. D. 1976. Canonical ridge and econometrics of joint production, *Journal of Econometrics*, **4**, 147–166.

Van de Geer, J. P. 1993. *Multivariate Analysis of Categorical Data: Applications*, SAGE Publications, ISBN: 978-0803945647.

Whittaker, J. 1990. *Graphical Models in Applied Mathematical Multivariate Statistics*, Wiley, ISBN: 978-0-471-91750-2.

Wishart, J. 1928. The generalized product moment distribution in samples from a normal multivariate population, *Biometrika*, **20**, 35–52.

Wright, S. 1921. Correlation and causation, *Journal of Agricultural Research*, **20**, 557–585.

Wright, S. 1934. The method of path coefficients, *Annals of Mathematical Statistics*, **5**, 161–215.

Xi, R., Lin, N., and Chen, Y. 2009. Compression and aggregation for logistic regression analysis in data cubes, *IEEE Transactions on Knowledge and Data Engineering*, **21**, 479–492.

Xu, R. and Wunsch, D. 2008. *Clustering*, Wiley, ISBN: 978-0-470-27680-8.

Zubin, T. 1938. A technique for measuring like-mindedness, *Journal of Abnormal Social Psychology*, **33**, 508–516.

Few Relevant URLs

| S. No. | Details of Relevant URLs | URL |
|---|---|---|
| 1 | ACM Special Interest Group for Genetic and Evolutionary Computation | http://www.sigevo.org/ |
| 2 | Ant Colony Optimization | http://www.aco-metaheuristic.org/ |
| 3 | Artificial Immune System Web | http://www.artificial-immune-systems.org/ |
| 4 | AURORA—Advanced Models, Applications and Software for High-Performance Computing in Finance, University of Vienna, Switzerland | http://www.univie.ac.at/sor/aurora6/index.html |
| 5 | Center of Banking and Financial Research, University of Cyprus, Cyprus | https://www.ucy.ac.cy/hermes/en/ |
| 6 | Centre for Computational Finance and Economic Agents, University of Essex, UK | http://www.essex.ac.uk/ccfea/ |
| 7 | Center for Machine Learning and Intelligent Systems, University of California Irwin, USA | https://archive.ics.uci.edu/ml/datasets.html |
| 8 | Centre for Optimization and Statistical Learning, Northwestern University, USA | http://www.mccormick.northwestern.edu/research/optimization-machine-learning-center/ index.html |
| 9 | CiteSeerX (beta version) | http://citeseerx.ist.psu.edu/index |
| 10 | Convex Optimization | http://www.convexoptimization.com/ |
| 11 | Computational Infrastructure for Operations Research | http://www.coin-or.org/ |
| 12 | Data Publisher for Earth & Environmental Science (Pangaea) | http://www.pangaea.de/ |
| 13 | Data Sets | http://www.statsci.org/datasets.html |
| 14 | Decision Tree for Optimization Software | http://plato.asu.edu/guide.html |
| 15 | Economic Papers | http://econpapers.repec.org/ |

*Continued*

| S. No. | Details of Relevant URLs | URL |
|---|---|---|
| 16 | EMOO Repository | http://delta.cs.cinvestav.mx/~ccoello/EMOO |
| 17 | Evolutionary Algorithms for Solving Multi-Objective Problems | http://www.cs.cinvestav.mx/~emoobook/ |
| 18 | IBM Research | http://researchweb.watson.ibm.com/ |
| 19 | ICER—International Centre for Economic Studies | http://nf.vse.cz/english/science-and-research/icer-international-centre-for-economic-studies/ |
| 20 | International Neural Network Society | http://www.inns.org/ |
| 21 | International Society for Genetic and Evolutionary Computation | http://www.isgec.org/ |
| 22 | International Society on Multiple Criteria Decision Making | http://www.mcdmsociety.org |
| 23 | International Statistical Institute | http://www.isi-web.org/ |
| 24 | Kanpur Genetic Algorithm Lab | http://www.iitk.ac.in/kangal/index.shtml |
| 25 | Metaheuristic Network | http://www.metaheuristics.org/ |
| 26 | Modeling & Optimization Research & Education, The University of Arizona | http://www.sie.arizona.edu/MORE/index.html |
| 27 | NEOS Guide: Companion Site to the NEOS Serve | http://www.neos-guide.org/ |
| 28 | NEOS Server: State-of-the-Art Solvers for Numerical Optimization | http://www.neos-server.org/ neos/ |
| 29 | Netlib Repository | http://www.netlib.org/ |
| 30 | Network for Artificial Immune Systems | http://www.elec.york.ac.uk/ARTIST/ |
| 31 | NIST, Information Technology Library | http://www.itl.nist.gov |
| 32 | Numerical Algorithm Group | http://www.nag.com/ |
| 33 | Oak Ridge National Laboratory | http://www.ornl.gov/ |
| 34 | Optimization Online | http://optimization-online.org/ |
| 35 | Particle Swarm Intelligence | http://www.swarmintelligence.org/ |
| 36 | Risk Management and Financial Engineering Lab, University of Florida, USA | http://www.ise.ufl.edu/rmfe/ |
| 37 | Social Science Research Network | http://www.ssrn.com/ |
| 38 | Stochastic Programming Society | http://stoprog.org/ |
| 39 | Stochastic Optimization Research Group, Georgia Institute of Technology, USA | http://www2.isye.gatech.edu/so/ |
| 40 | The Probability Web | http://probweb.berkeley.edu/ |

Few Relevant Softwares

| S. No. | Details of Softwares | URL |
|---|---|---|
| 1 | A Quadratic Assignment Problem Library (QAPLIB) | http://www.seas.upenn.edu/qaplib/ |
| 2 | AMPL Modeling Language for Mathematical Programming | http://www.ampl.com/ |
| 3 | IBM ILOG CPLEX Optimizer | http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/ |
| 4 | Data Envelopment Analysis | http://www.dea-analysis.com/ and<br>http://www.deazone.com/ |

*Continued*

| S. No. | Details of Softwares | URL |
|---|---|---|
| 5 | Data Envelopment Analysis Online Software | https://www.deaos.com |
| 6 | EVIEWS | http://www.eviews.com/ |
| 7 | GAUSS | http://www.aptech.com/products/gauss-mathematical-and-statistical-system/ |
| 8 | GNU Octave | http://www.gnu.org/software/octave/ |
| 9 | Graph Visualization Software | http://gephi.org/ |
| 10 | High Performance 3D Visualization Software | http://www.vsg3d.com/ |
| 11 | Indirect Optimization on the Basis of Self-Organization (IOSO) | http://iosotech.com/ |
| 12 | jMetal (Metaheuristic Algorithms in Java) | http://jmetal.sourceforge.net/index.html |
| 13 | Julia Software | http://julialang.org/ |
| 14 | LINDO | http://www.lindo.com/ |
| 15 | Machine Learning Open Source Software | http://mloss.org/software/ |
| 16 | Maple Soft | http://www.maplesoft.com/ |
| 17 | MATLAB (MathWorks) | http://www.mathworks.com/ |
| 18 | Mathematica (Wolfram Research) | http://www.wolfram.com/ |
| 19 | MINITAB | http://www.minitab.com/en-us/ |
| 20 | ModeFrontier | http://www.esteco.com/home.html |
| 21 | MOSEK | http://www.mosek.com/ |
| 22 | MTC Software | http://www.hyperthermcam.com/en-us/ |
| 23 | NetLearn (Interactive Demonstrations of Network Concepts) | http://www.ladamic.com/netlearn/ |
| 24 | Neural Network Software | http://www.alyuda.com/<br>and<br>http://www.alyuda.com/neural-networks-software.htm |
| 25 | NeuroSolutions | http://www.neurosolutions.com/ |
| 26 | Numerical Algorithms Group (NAG) | http://www.nag.co.uk |
| 27 | Optimization Software | http://www.optimalon.com/ |
| 28 | OR Softwares | http://www.mccormick.northwestern.edu/research/optimization-machine-learning-center/software-downloads/index.html |
| 29 | Paradiseo (A Software Framework for Metaheuristics) | http://paradiseo.gforge.inria.fr/ |
| 30 | PISA (A Platform and Programming Language Independent Interface for Search Algorithms) | http://www.tik.ee.ethz.ch/pisa/ |
| 31 | Robust Optimization Made Easy (ROME) | http://robustopt.com/ |
| 32 | The Comprehensive R Archive Network | http://cran.r-project.org/ |
| 33 | R-Studio | http://www.rstudio.com/ |
| 34 | R Software | http://www.r-project.org |
| 35 | Reactive System, Inc. | http://www.reactive-systems.com/ |
| 36 | Reflector CAD | http://www.breault.com/software/reflectorCAD.php |
| 37 | SAS | http://www.sas.com/ |
| 38 | Scientific Software International, Inc. | http://www.ssicentral.com/ |
| 39 | Scilab | http://www.scilab.org/ |

*Continued*

| S. No. | Details of Softwares | URL |
|---|---|---|
| 40 | Simulation | http://www.simulations-plus.com/ |
| 41 | Spotfire Analytical Tool | http://spotfire.tibco.com/discover-spotfire |
| 42 | SPSS Data Collection Data Model | http://www-01.ibm.com/software/analytics/spss/products/data-collection/data-model/ |
| 43 | SPSS Modeler | http://www-01.ibm.com/ software/analytics/spss/products/modeler/ |
| 44 | STATA Data Analysis and Statistical Software | http://www.stata.com/ |
| 45 | Stuttgart Neural Network Simulator | http://www.ra.cs.uni-tuebingen.de/SNNS/ |
| 46 | Statistica | http://www.statsoft.com/# |
| 47 | Systat | http://www.systat.com/ |

Few Relevant Societies/Department/Schools/Institutes, etc.

| S. No. | Details of Societies/Department/ Schools/Institutes, etc. | URL |
|---|---|---|
| 1 | American Statistical Association | http://www.amstat.org/ |
| 2 | The Royal Statistical Society | http://www.rss.org.uk/ |
| 3 | Royal Economic Society | http://www.res.org.uk/ |
| 4 | Institute of Operations Research and Management Science (INFORMS) | http://www.informs.org/ |
| 5 | Society for Industrial and Applied Mathematics (SIAM) | http://www.siam.org/ |
| 6 | Mathematical Optimization Society | http://www.mathopt.org/ |
| 7 | The Association of European Operational Research Societies (EURO) | http://www.euro-online.org/web/pages/ 1/home |
| 8 | Canadian Operations Research Society | http://www.cors.ca/ |
| 9 | French Society for Operations Research and Decisions | http://www.roadef.org/content/index.htm |
| 10 | The Operational Research Society | http://www.theorsociety.com/ |
| 11 | Operational Research Society, Turkey | http://www.yad.org.tr/ |
| 12 | Operations Research Society of China | http://www.orsc.org.cn/ and http://www.orsc.org.cn/engindex.html |
| 13 | The Operations Research Society of Japan | http://www.orsj.or.jp/ |
| 14 | Operational Research Society of India | http://www.orsi.in/ |
| 15 | Brazilian Society of Operations Research | http://www.sobrapo.org.br/ |
| 16 | Stochastic Programming Community Home Page | http://www.stoprog.org/ |
| 17 | International Society on Multiple Criteria Decision Making | http://www.mcdmsociety.org/ |
| 18 | Society for Judgment and Decision Making | http://www.sjdm.org/ |
| 19 | International Institute for Applied Systems Analysis | http://www.iiasa.ac.at/ |
| 20 | The European Association for Decision Making | http://www.eadm.eu/ |
| 21 | ESIGMA (European Summer Institute Group on Multicriteria Analysis):<br>European Working Group on Multiple Criteria Decision Aiding | http://www.cs.put.poznan.pl/ewgmcda/ |
| 22 | European Mathematical Information Service | http://www.emis.de/ |
| 23 | Algorithms, Combinatorics, and Optimization | http://www.aco.gatech.edu/ |

*Continued*

| S. No. | Details of Societies/Department/ Schools/Institutes, etc. | URL |
|---|---|---|
| 24 | University of Waterloo (Combinatorics & Optimization) | http://www.math.uwaterloo.ca/co/ |
| 25 | Mathematical Science Research Institute | http://www.msri.org/web/msri |
| 26 | Institute of Quantum Information | http://www.iqi.caltech.edu/ |
| 27 | Perimeter Institute of Theoretical Physics | http://www.perimeterinstitute.ca/ |
| 28 | Centre for Discrete Mathematics & Theoretical Computer Science | http://www.dimacs.rutgers.edu/ |
| 29 | Canadian Institute for Advanced Research (CIFAR) | http://www.cifar.ca/ |
| 30 | QuantumWorks | http://www.quantumworks.ca/section/view |
| 31 | Center for the Mathematics of Information (CMI) | http://www.cmi.caltech.edu/index.shtml |
| 32 | Optimization Online | http://www.optimization-online.org/ |
| 33 | The Fields Institute for Research in Mathematical Sciences | http://www.fields.utoronto.ca/ |
| 34 | American Mathematical Society | http://www.ams.org/home/page |
| 35 | Canadian Mathematical Society | http://cms.math.ca/ |
| 36 | Good e-books | http://sites.stat.psu.edu/~zuz13/resources.html |
| 37 | European Centre for Advanced Research in Economics and Statistics | http://www.ecares.org/ |
| 38 | Cornell University Library | http://arxiv.org/ |
| 39 | DEA Data Repository | http://www.etm.pdx.edu/dea/dataset/default.htm |
| 40 | Virtual Library For Economics and Business Studies | http://www.econbiz.de/en/ |
| 41 | Data Envelopment Analysis: Applications for Measuring Efficiency | http://www.dea-analysis.com/ |
| 42 | Decision Sciences Institute | http://www.decisionsciences.org/ |
| 43 | DSpace MIT, USA | http://dspace.mit.edu/ |
| 44 | Mathematical Programming Glossary | http://glossary.computing.society.informs.org/ |
| 45 | Indian Statistical Institute | http://www.isical.ac.in/ |
| 46 | Department of Statistics, Stanford University, USA | https://statistics.stanford.edu/ |
| 47 | Department of Statistical Sciences (DSS), Cornell University, USA | http://stat.cornell.edu/ |
| 48 | Department of Statistics, Harvard University, USA | http://statistics.fas.harvard.edu/ |
| 49 | Department of Statistics, LSE, UK | http://www.lse.ac.uk/statistics/home.aspx |
| 50 | Department of Statistical Sciences, Duke University, USA | https://stat.duke.edu/ |
| 51 | Department of Statistics, Oxford University, UK | https://www.stats.ox.ac.uk/ |
| 52 | Department of Statistics, University of Washington, USA | https://www.stat.washington.edu/ |
| 53 | Department of Statistics, University of California Berkeley, USA | http://statistics.berkeley.edu/ |
| 54 | Department of Statistics, Columbia University, USA | http://stat.columbia.edu/ |
| 55 | Department of Statistics, North Carolina State University, USA | http://www.stat.ncsu.edu/ |
| 56 | Statistics Department, Wharton, University of Pennsylvania, USA | https://statistics.wharton.upenn.edu/ |

*Continued*

| S. No. | Details of Societies/Department/ Schools/Institutes, etc. | URL |
|---|---|---|
| 57 | Department of Statistics, Yale University, USA | http://statistics.yale.edu/ |
| 58 | Department of Statistics, University of Michigan Ann Arbor, USA | http://lsa.umich.edu/stats/ |
| 59 | Department of Statistics, University of California Los Angeles, USA | http://statistics.ucla.edu/ |
| 60 | Department of Statistics, Carnegie Mellon University, USA | http://www.stat.cmu.edu/ |
| 61 | Department of Statistics, University of Wisconsin Madison, USA | https://www.stat.wisc.edu/ |
| 62 | Department of Statistics, University of Florida, USA | http://www.stat.ufl.edu/ |
| 63 | Department of Mathematics and Statistics, Indian Institute of Technology Kanpur, India | http://www.iitk.ac.in/math/ |
| 64 | Department of Mathematics, Indian Institute of Technology Bombay, India | http://www.math.iitb.ac.in/ |
| 65 | Department of Mathematics, Indian Institute of Technology Kharagpur, India | http://www.iitkgp.ac.in/academics/?page=acadunits&&dept=MM |
| 66 | Department of Mathematics, Indian Institute of Technology Madras, India | https://mat.iitm.ac.in/ |
| 67 | Department of Statistics, Pune University, India | http://stats.unipune.ernet.in/ |
| 68 | Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland | http://www.mimuw.edu.pl/ |
| 69 | Department of Statistics and Applied Probability, National University of Singapore, Singapore | http://www.stat.nus.edu.sg/opencms/ |
| 70 | Department of Mathematics, Hong Kong University of Science and Technology, China | http://www.math.ust.hk/welcome.php |
| 71 | List of Department in Statistics and Mathematics, Academia Sinica, Taiwan | http://www.math.sinica.edu.tw/addbook/default_e.jsp |
| 72 | Department of Mathematics, ETH Zurich, Switzerland | https://www.math.ethz.ch/ |
| 73 | Institute for Operations Research, Department of Mathematics ETH Zurich, Switzerland | http://www.ifor.math.ethz.ch/ |
| 74 | Graduate School of Mathematical Sciences, University of Tokyo, Japan | http://www.ms.u-tokyo.ac.jp/ |
| 75 | Department of Mathematical Sciences, Tsinghua University, China | http://www.tsinghua.edu.cn/publish/mathen/ 2780/ |
| 76 | Department of Mathematics, Katholieke Universiteit Leuven, Belgium | https://wis.kuleuven.be/english |
| 77 | Doctoral Program in Mathematics, École Normale Superieure, Paris (ENS Paris), France | http://www.math.u-psud.fr/~ecdoct/ecdoct/index.php?l=ANG |
| 78 | Department of Pure Mathematics, École Polytechnique (ParisTech), France | https://www.polytechnique.edu/en/department-of-pure-mathematics |
| 79 | Master de Sciences et Technologies, Université Pierre et Marie Curie (UPMC), France | http://www.master.ufrmath.upmc.fr/ |
| 80 | Department of Statistics, Purdue University, USA | http://www.stat.purdue.edu/ |
| 81 | Department of Statistics and Operations Research, University of North Carolina, USA | http://stat-or.unc.edu/ |

*Continued*

| S. No. | Details of Societies/Department/ Schools/Institutes, etc. | URL |
|---|---|---|
| 82 | Department of Statistics, University of Chicago, USA | https://galton.uchicago.edu/ |
| 83 | Department of Probability and Statistics, Peking University, China | http://www.stat.pku.edu.cn/en/ |
| 84 | Institute of Mathematics and Statistics, University of Sao Paulo, Brazil | https://www.ime.usp.br/en |
| 85 | Leuven Statistics Research Centre, UCL Leuven | https://lstat.kuleuven.be/ |
| 86 | Department of Statistics, Seoul National University, South Korea | http://gsis.snu.ac.kr/career/statistics |
| 87 | Institute of Operations Research and Statistics, National Tsing Hua University, China | http://www.ie.tsinghua.edu.cn/eng/content.php?pid=171&ty=173# |
| 88 | Department of Statistics and Operations Research, University of Vienna, Switzerland | https://isor.univie.ac.at/ |
| 89 | Institute of Stochastics, University of Ulm, Germany | http://www.uni-ulm.de/index.php?id=3835&L=1 |
| 90 | Department of Statistics, Ludwig-Maximilians-University Munich, Germany | http://www.statistik.lmu.de/index_e.html |
| 91 | Department of Mathematics, Humboldt-Universität zu Berlin, Germany | https://www.mathematik.hu-berlin.de/en |
| 92 | Department of Mathematics, Karlsruhe Institute of Technology, Germany | http://www.math.kit.edu/en |
| 93 | Department of Mathematics, RWTH Aachen University, Germany | http://www.mathematik.rwth-aachen.de/cms/~mxy/Mathematik/?lidx=1 |
| 94 | Department of Mathematics, Technische Universität Dresden, Germany | http://tu-dresden.de/die_tu_dresden fakultaeten/fakultaet_mathematik_und_ naturwissenschaften/fachrichtung_mathematik |
| 95 | National Institute of Statistical Sciences | http://www.niss.org/ |
| 96 | Statistical and Applied Mathematical Sciences Institute | http://www.samsi.info |
| 97 | List of Statistics Departments | http://www.stat.ufl.edu/vlib/statistics.html |

Few Relevant Publication House/Publishers/Book and Journal Providers

| S. No. | Details of Publication House/ Publishers/Book and Journal Providers | URL |
|---|---|---|
| 1 | ACM Publications | http://www.acm.org/publications |
| 2 | Addison-Wesley | http://www.pearsoned.co.uk/imprints/addison-wesley/ |
| 3 | American Mathematical Society (AMS) | http://www.ams.org/home/page |
| 4 | Association of American University Presses (AAUP) | http://www.aaupnet.org/ |
| 5 | Baltzer Science Publishers | http://www.baltzersciencepublishers.com/en/ |
| 6 | Birkhäuser Basel | http://www.springer.com/birkhauser?SGWID=0-40290-0-0-0 |
| 7 | Blackwell | http://as.wiley.com/WileyCDA/Section/ index.html |

*Continued*

| S. No. | Details of Publication House/ Publishers/Book and Journal Providers | URL |
|---|---|---|
| 8 | Cambridge University Press | http://www.cambridge.org |
| 9 | Centre de Recherches Mathématiques (University of Montréal, PQ) | http://www.crm.umontreal.ca/en/index.shtml |
| 10 | CRC Press | http://www.crcpress.com/ |
| 11 | Dover Publications, Inc. | http://store.doverpublications.com/ |
| 12 | Duke University Press | https://www.dukeupress.edu/ |
| 13 | Elsevier | https://www.elsevier.com/ |
| 14 | Gale Research, Cengage | http://www.cengage.com/search/showresults.do?N=197 |
| 15 | Hindawi Publishing Corporation | http://www.hindawi.com/ |
| 16 | IEEE | https://www.ieee.org/index.html |
| 17 | Institute of Mathematical Statistics Publications | http://imstat.org/publications/ |
| 18 | Indian Institute of Science (IISc) Press | http://www.iiscpress.iisc.in/ |
| 19 | Indiana University Press | http://www.iupress.indiana.edu/ |
| 20 | JSTOR | http://www.jstor.org/ |
| 21 | Mathematical Association of America | http://www.maa.org/ |
| 22 | McGraw-Hill | http://www.mheducation.com/ |
| 23 | Marcel Dekker, Inc. | http://www.dekker.com |
| 24 | MIT Press | https://mitpress.mit.edu/ |
| 25 | Oxford University Press | http://global.oup.com |
| 26 | Pearson | http://www.pearsoned.co.uk/ |
| 27 | Prentice-Hall | http://www.prenticehall.com/ |
| 28 | Princeton University Press | http://press.princeton.edu/ |
| 29 | Scientific & Academic Publishing | http://www.sapub.org/journal/index.aspx |
| 30 | SIAM | http://www.siam.org/ |
| 31 | Springer | http://www.springer.com/gp/ |
| 32 | Taylor & Francis | http://www.taylorandfrancis.com/ |
| 33 | University of Chicago Press | http://www.press.uchicago.edu/index.html |
| 34 | Walter de Gruyter, Inc. | http://www.degruyter.com/ |
| 35 | Wiley | http://www.wiley.com |
| 36 | Wolters Kluwer Group | http://wolterskluwer.com/ |

List of Few Relevant Journals

| S. No. | Details of Relevant List of Few Journals | URL |
|---|---|---|
| 1 | *American Review of Mathematics and Statistics* | http://armsnet.info/ |
| 2 | *Applied Econometrics and International Development* | http://www.usc.es/economet/eaa.htm |
| 3 | *The Annals of Probability* | http://www.imstat.org/aop/ |
| 4 | *The Annals of Applied Probability* | http://www.imstat.org/aap/ |
| 5 | *The Annals of Statistics* | http://www.imstat.org/aos/ |
| 6 | *The Annals of Applied Statistics* | http://imstat.org/aoas/ |

*Continued*

| S. No. | Details of Relevant List of Few Journals | URL |
|---|---|---|
| 7 | *Applied Stochastic Models in Business and Industry* | http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291526-4025 |
| 8 | *Australian & New Zealand Journal of Statistics* | http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291467-842X |
| 9 | *Bayesian Analysis* | http://projecteuclid.org/euclid.ba |
| 10 | *Biometrika* | http://biomet.oxfordjournals.org/ |
| 11 | *Biometrical Journal* | http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291521-4036 |
| 12 | *Biometrics* | http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291541-0420 |
| 13 | *Biostatistics* | http://biostatistics.oxfordjournals.org/ |
| 14 | *Canadian Journal of Statistics* | http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291708-945X |
| 15 | *Communications in Statistics—Simulation and Computation* | http://www.tandfonline.com/loi/lssp20#. VipG_iv2Qnk |
| 16 | *Communications in Statistics—Theory and Methods* | http://www.tandfonline.com/loi/lsta20#. VipG2yv2Qnk |
| 17 | *Computational Statistics* | http://link.springer.com/journal/180 |
| 18 | *Computational Statistics and Data Analysis* | http://www.journals.elsevier.com/computational-statistics-and-data-analysis/ |
| 19 | *Econometrica* | http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291468-0262 |
| 20 | *Econometrics* | http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291368-423X |
| 21 | *Econometric Reviews* | http://www.tandfonline.com/action/journalInformation?journalCode=lecr20#. Vi3_Yyus0nk |
| 22 | *Environmetrics* | http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291099-095X |
| 23 | *International Journal of Forecasting* | http://www.journals.elsevier.com/international-journal-of-forecasting/ |
| 24 | *International Statistical Review* | http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291751-5823 |
| 25 | *Journal of Agricultural, Biological, and Environmental Statistics* | http://link.springer.com/journal/13253 |
| 26 | *Journal of the American Statistical Association* | http://www.tandfonline.com/loi/uasa20#. Vii_Qiv2Qnk |
| 27 | *Journal of Applied Econometrics* | http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291099-1255 |
| 28 | *Journal of Applied Statistics* | http://www.tandfonline.com/action/journalInformation?journalCode=cjas20#. Vi4CQCus0nk |
| 29 | *Journal of Business & Economic Statistics* | http://amstat.tandfonline.com/action/journalInformation?journalCode=ubes20#.Vi4CbCus0nk |
| 30 | *Journal of Chemometrics* | http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291099-128X/issues |
| 31 | *Journal of Computational and Graphical Statistics* | http://amstat.tandfonline.com/action/journalInformation?journalCode=ucgs20#.Vi4Coyus0nk |
| 32 | *Journal of Econometrics* | http://www.journals.elsevier.com/journal-of-econometrics/ |

*Continued*

| S. No. | Details of Relevant List of Few Journals | URL |
|---|---|---|
| 33 | *Journal of Economic and Social Measurement* | http://www.iospress.nl/journal/journal-of-economic-and-social-measurement/ |
| 34 | *Journal of Environmental Statistics* | http://www.jenvstat.org/ |
| 35 | *Journal of Japanese Society of Computational Statistics* | http://jscs.jp/oubun/ |
| 36 | *Journal of the Japanese Statistical Association* | http://www.jss.gr.jp/en/journal/index.html |
| 37 | *Journal of Machine Learning Research* | http://www.jmlr.org/ |
| 38 | *Journal of Modern Applied Statistical Methods* | http://www.jmasm.com/ |
| 39 | *Journal of Multivariate Analysis* | http://www.journals.elsevier.com/journal-of-multivariate-analysis |
| 40 | *Journal of the Royal Statistical Society—Series A: Statistics in Society* | http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291467-985X/ |
| 41 | *Journal of the Royal Statistical Society—Series B: Statistical Methodology* | http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291467-9868/ |
| 42 | *Journal of the Royal Statistical Society—Series C: Applied Statistics* | http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291467-9876 |
| 43 | *Journal of Statistical Computation and Simulation* | http://www.tandfonline.com/action/journalInformation?journalCode=gscs20#.Vi4EICus0nk |
| 44 | *Journal of Statistics Education* | https://www.amstat.org/publications/jse/ |
| 45 | *Journal of Statistical Physics* | http://www.springer.com/physics/complexity/ journal/10955 |
| 46 | *Journal of Statistical Software* | http://www.jstatsoft.org/index |
| 47 | *Journal of Time Series Analysis* | http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291467-9892 |
| 48 | *Pharmaceutical Statistics* | http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291539-1612 |
| 49 | *Physica A: Statistical Mechanics and Its Applications* | http://www.journals.elsevier.com/physica-a-statistical-mechanics-and-its-applications/ |
| 50 | *Psychometrika* | http://www.springer.com/psychology/journal/11336 |
| 51 | *Quality and Reliability Engineering International* | http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291099-1638/ |
| 52 | *Sankhyā A: The Indian Journal of Statistics* | http://www.springer.com/statistics/journal/13171 |
| 53 | *Sankhyā B: The Indian Journal of Statistics* | http://www.springer.com/statistics/journal/13571 |
| 54 | *Scandinavian Journal of Statistics* | http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291467-9469 |
| 55 | *Statistics and Computing* | http://www.springer.com/statistics/computational+statistics/journal/11222 |
| 56 | *Statistical Methods in Medical Research* | http://smm.sagepub.com/ |
| 57 | *Statistics in Biopharmaceutical Research* | http://www.tandfonline.com/action/journalInformation?journalCode=usbr20#. Vi4G4Cus0nk |
| 58 | *Statistics in Medicine* | http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291097-0258 |

*Continued*

| S. No. | Details of Relevant List of Few Journals | URL |
|---|---|---|
| 59 | *Statistics and Probability Letters* | http://www.journals.elsevier.com/statistics-and-probability-letters/ |
| 60 | *Statistical Analysis and Data Mining* | http://onlinelibrary.wiley.com/journal/10.1002/%28ISSN%291932-1872/ |
| 61 | *Statistical Applications in Genetics and Molecular Biology* | http://www.degruyter.com/view/j/sagmb |
| 62 | *Statistical Modelling* | http://smj.sagepub.com/ |
| 63 | *Statistica Neerlandica* | http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291467-9574 |
| 64 | *Statistica Sinica* | http://www3.stat.sinica.edu.tw/statistica/ |
| 65 | *Statistical Science* | http://www.imstat.org/sts/ |
| 66 | *Statistics and Risk Modeling* | http://www.degruyter.com/view/j/strm |
| 67 | *Statistics Surveys* | http://imstat.org/ss/ |
| 68 | *Stochastic Environmental Research and Risk Assessment* | http://www.springer.com/environment/journal/477 |
| 69 | *Stochastic Processes and Their Applications* | http://www.journals.elsevier.com/stochastic-processes-and-their-applications/ |
| 70 | *Statistics and Probability Letters* | http://www.journals.elsevier.com/statistics-and-probability-letters/ |
| 71 | *Structural Equation Modeling: A Multidisciplinary Journal* | http://www.tandfonline.com/toc/hsem20/current |
| 72 | *Technometrics* | http://amstat.tandfonline.com/action/journalInformation?show=aimsScope&journalCode=utch20#.Vi4I3yus0nk |
| 73 | *The Review of Economics and Statistics* | http://www.mitpressjournals.org/loi/rest |

Few Relevant Data Sets

| S. No. | Details of Data Sets | URL |
|---|---|---|
| 1 | Amazon Public Data | http://aws.amazon.com/public-data-sets/ |
| 2 | Bernoulli Society | http://www.bernoulli-society.org/ |
| 3 | Biologic Specimen and Data Repository Information Coordinating Center | https://biolincc.nhlbi.nih.gov/home/ |
| 4 | Biostatistics Data, Vandebilt University, USA | http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/DataSets |
| 5 | Chemical Informatics | http://www.cheminformatics.org/datasets/index.shtml |
| 6 | Critical Assessment of Microarray Data Analysis | http://www.camda.duke.edu/camda03/datasets/ |
| 7 | DataCite | http://www.datacite.org/ |
| 8 | Data Sets | http://www.grappa.univ-lille3.fr/~torre/Recherche/Experiments/Datasets/ |
| 9 | Data Sets for "The Elements of Statistical Learning" | http://statweb.stanford.edu/~tibs/ElemStatLearn/ data.html |
| 10 | Earthquake Data | http://earthquake.usgs.gov/data/ |

*Continued*

| S. No. | Details of Data Sets | URL |
|---|---|---|
| 11 | European Cities 1M Data Sets | http://image.ntua.gr/iva/datasets/ec1 m/ |
| 12 | Geo Data Centre, Arizona State University, USA | http://geodacenter.asu.edu/sdata |
| 13 | Hubble Space Data | http://hla.stsci.edu/ |
| 14 | International Statistical Institute | http://isi-web.org/ |
| 15 | MIT Airline Data Project | http://web.mit.edu/airlinedata/www/Revenue&Related.html |
| 16 | NCBI GenBank | http://www.ncbi.nlm.nih.gov/genbank/ |
| 17 | NHLBI Data Repository and Biospecimen Repository Information Coordination Center | https://biolincc.nhlbi.nih.gov/redirect/ |
| 18 | National Flight Data Center (NFDC) | https://nfdc.faa.gov/xwiki/bin/view/NFDC/WebHome |
| 19 | Network Data | http://www-personal.umich.edu/~mejn/netdata/ |
| 20 | NOAA Data | http://www.ncdc.noaa.gov/ |
| 21 | Open Sports Data/API | http://www.openligadb.de/ |
| 22 | Online Glossary of Research Economics | http://www.econterms.com/ |
| 23 | Online for Time Series Data | http://datamarket.com/ |
| 24 | Online Statistical Software for Astronomy and Related Physical Sciences | http://astrostatistics.psu.edu/statcodes/ |
| 25 | Precipitation Measurement Data | http://pmm.nasa.gov/data-access/google-earth |
| 26 | Public Data, University of Utah, USA | http://www.utah.gov/data/ |
| 27 | Public Government Data Sets | http://catalog.data.gov/dataset |
| 28 | Quandl—Intelligent Search for Numerical Data | http://www.quandl.com/ |
| 29 | Real-Time Space Weather Data Sources | http://space.rice.edu/ISTP/#RT |
| 30 | Stat Lib, Statistical Library, Carnegie Mellon University, USA | http://lib.stat.cmu.edu/ |
| 31 | Statistical Data Set, University of Massachusetts Amherst, USA | http://www.umass.edu/statdata/statdata/ |
| 32 | Statistical Data Set, University of Vienna, Switzerland | http://www.mat.univie.ac.at/~neum/statdat.html |
| 33 | Statistical Reference Data Set | http://www.itl.nist.gov/div898/strd/ |
| 34 | Statistical Science Web (Data Sets) | http://www.statsci.org/datasets.html |
| 35 | Survival Analysis, Including Penalised Likelihood | http://sites.stat.psu.edu/~dhunter/R/html/survival/html/00Index.html |
| 36 | The Center for Innovation in Engineering and Science Education Real-Time Data Sites | http://www.k12science.org/materials/resources/realtimedata/ |
| 37 | The Data and Story Library, Carnegie Mellon University, USA | http://lib.stat.cmu.edu/DASL/ |
| 38 | The World Wide Web Virtual Library: Statistics | http://www.stat.ufl.edu/vlib/statistics.html |
| 39 | U.S. Department of Homeland Security Data | http://www.dhs.gov/topic/data |

*Continued*

| S. No. | Details of Data Sets | URL |
|---|---|---|
| 40 | University of Edinburgh School of Informatics Data Sets for Data Mining | http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html |
| 41 | Weibull.com, Reliability Engineering Resource Website | http://www.weibull.com/ |
| 42 | U.S. Government Data | https://www.usa.gov/statistics |
| 43 | U.S. Web Traffic | https://analytics.usa.gov/ |
| 44 | Ministry of Statistics and Programme Implementation (GoI) | http://mospi.nic.in/ |
| 45 | U.K. Government Data | https://www.data.gov.uk/ |
| 46 | German Government data | https://www.govdata.de/ |
| 47 | National Institute of Statistics and Economic Studies (France) | http://www.insee.fr/en/ |
| 48 | National Bureau of Statistics of China | http://chinadatacenter.org/AboutCDC/PartnersContent.aspx?id=23 |
| 49 | Canadian Government Data | http://open.canada.ca/en |
| 50 | Brazilian Government Data | http://www.ibge.gov.br/english/ |
| 51 | Statistics Bureau of Japan | http://www.stat.go.jp/english/ |
| 52 | Statistics Sweden | http://www.scb.se/ |
| 53 | Statistics Korea | http://kostat.go.kr/portal/english/index.action |
| 54 | World Bank Data | http://data.worldbank.org/ |