Multivariate Analysis Concepts

Chapter 1

1.1 Introduction

The subject of multivariate analysis deals with the statistical analysis of the data collected on more than one (response) variable. These variables may be correlated with each other, and their statistical dependence is often taken into account when analyzing such data. In fact, this consideration of statistical dependence makes multivariate analysis somewhat different in approach and considerably more complex than the corresponding univariate analysis, when there is only one response variable under consideration.

Response variables under consideration are often described as random variables and since their dependence is one of the things to be accounted for in the analyses, these response variables are often described by their joint probability distribution. This consideration makes the modeling issue relatively manageable and provides a convenient framework for scientific analysis of the data. Multivariate normal distribution is one of the most frequently made distributional assumptions for the analysis of multivariate data. However, if possible, any such consideration should ideally be dictated by the particular context. Also, in many cases, such as when the data are collected on a nominal or ordinal scales, multivariate normality may not be an appropriate or even viable assumption.

In the real world, most data collection schemes or designed experiments will result in multivariate data. A few examples of such situations are given below.

- During a survey of households, several measurements on *each* household are taken. These measurements, being taken on the same household, will be dependent. For example, the education level of the head of the household and the annual income of the family are related.
- During a production process, a number of different measurements such as the tensile strength, brittleness, diameter, etc. are taken on the same unit. Collectively such data are viewed as multivariate data.
- On a sample of 100 cars, various measurements such as the average gas mileage, number of major repairs, noise level, etc. are taken. Also each car is followed for the first 50,000 miles and these measurements are taken after every 10,000 miles. Measurements taken on the same car at the same mileage and those taken at different mileage are going to be correlated. In fact, these data represent a very complex multivariate analysis problem.

- An engineer wishes to set up a control chart to identify the instances when the production process may have gone out of control. Since an out of control process may produce an excessively large number of out of specification items, detection at an early stage is important. In order to do so, she may wish to monitor several process characteristics on the same units. However, since these characteristics are functions of process parameters (conditions), they are likely to be correlated leading to a set of multivariate data. Thus many times, it is appropriate to set up a single (or only a few) multivariate control chart(s) to detect the occurrence of any out of control conditions. On the other hand, if several univariate control charts are separately set up and individually monitored, one may witness too many false alarms, which is clearly an undesirable situation.
- A new drug is to be compared with a control for its effectiveness. Two different groups of patients are assigned to each of the two treatments and they are observed weekly for next two months. The periodic measurements on the same patient will exhibit dependence and thus the basic problem is multivariate in nature. Additionally, if the measurements on various possible side-effects of the drugs are also considered, the subsequent analysis will have to be done under several carefully chosen models.
- In a designed experiment conducted in a research and development center, various factors are set up at desired levels and a number of response variables are measured for each of these treatment combinations. The problem is to find a combination of the levels of these factors where all the responses are at their 'optimum'. Since a treatment combination which optimizes one response variable may not result in the optimum for the other response variable, one has a problem of conflicting objectives especially when the problem is treated as collection of several univariate optimization problems. Due to dependence among responses, it may be more meaningful to analyze response variables simultaneously.
- In many situations, it is more economical to collect a large number of measurements on the same unit but such measurements are made only on a few units. Such a situation is quite common in many remote sensing data collection plans. Obviously, it is practically impossible to collectively interpret hundreds of univariate analyses to come up with some definite conclusions. A better approach may be that of data reduction by using some meaningful approach. One may eliminate some of the variables which are deemed redundant in the presence of others. Better yet, one may eliminate some of the linear combinations of all variables which contain little or no information and then concentrate only on a few important ones. Which linear combinations of the variables should be retained can be decided using certain multivariate methods such as principal component analysis. Such methods are not discussed in this book, however.

Most of the problems stated above require (at least for the convenience of modeling and for performing statistical tests) the assumption of multivariate normality. There are however, several other aspects of multivariate analysis such as factor analysis, cluster analysis, etc. which are largely distribution free in nature. In this volume, we will only consider the problems of the former class, where multivariate normality assumption may be needed. Therefore, in the next few sections, we will briefly review the theory of multivariate normal and other related distributions. This theory is essential for a proper understanding of various multivariate statistical techniques, notation, and nomenclature. The material presented here is meant to be only a refresher and is far from complete. A more complete discussion of this topic can be found in Kshirsagar (1972), Seber (1984) or Rencher (1995).

1.2 Random Vectors, Means, Variances, and Covariances

Suppose $y_1, \ldots, y_p$ are p possibly correlated random variables with respective means (expected values) $\mu_1, \ldots, \mu_p$. Let us arrange these random variables as a column vector denoted by $\mathbf{y}$, that is, let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}.$$

We do the same for $\mu_1, \mu_2, \ldots, \mu_p$ and denote the corresponding vector by $\boldsymbol{\mu}$. Then we say that the vector $\mathbf{y}$ has the mean $\boldsymbol{\mu}$ or in notation $E(\mathbf{y}) = \boldsymbol{\mu}$.

Let us denote the covariance between y_i and y_j by σ_{ij}, $i, j = 1, \ldots, p$, that is

$$\sigma_{ij} = \operatorname{cov}(y_i, y_j) = E[(y_i - \mu_i)(y_j - \mu_j)] = E[(y_i - \mu_i)y_j] = E(y_i y_j) - \mu_i \mu_j$$

and let

$$\boldsymbol{\Sigma} = (\sigma_{ij}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ & & & \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{bmatrix}.$$

Since $\operatorname{cov}(y_i, y_j) = \operatorname{cov}(y_j, y_i)$, we have $\sigma_{ij} = \sigma_{ji}$. Therefore, $\boldsymbol{\Sigma}$ is symmetric with $(i, j)^{th}$ and $(j, i)^{th}$ elements representing the covariance between y_i and y_j. Further, since $\operatorname{var}(y_i) = \operatorname{cov}(y_i, y_i) = \sigma_{ii}$, the i^{th} diagonal place of $\boldsymbol{\Sigma}$ contains the variance of y_i. The matrix $\boldsymbol{\Sigma}$ is called the dispersion or the variance-covariance matrix of $\mathbf{y}$. In notation, we write this fact as $D(\mathbf{y}) = \boldsymbol{\Sigma}$. Various books follow alternative notations for $D(\mathbf{y})$ such as $\operatorname{cov}(\mathbf{y})$ or $\operatorname{var}(\mathbf{y})$. However, we adopt the less ambiguous notation of $D(\mathbf{y})$.

Thus,

$$\boldsymbol{\Sigma} = D(\mathbf{y}) = E[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'] = E[(\mathbf{y} - \boldsymbol{\mu})\mathbf{y}'] = E(\mathbf{y}\mathbf{y}') - \boldsymbol{\mu}\boldsymbol{\mu}',$$

where for any matrix (vector) $\mathbf{A}$, the notation $\mathbf{A}'$ represents its transpose.

The quantity $tr(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \sigma_{ii}$ is called *total variance* and a determinant of $\boldsymbol{\Sigma}$, denoted by $|\boldsymbol{\Sigma}|$, is often referred to as the *generalized variance*. The two are often taken as the overall measures of the variability of the random vector $\mathbf{y}$. However, both of these two measures suffer from certain shortcomings. For example, the total variance $tr(\boldsymbol{\Sigma})$ being the sum of only diagonal elements, essentially ignores all covariance terms. On the other hand, the generalized variance $|\boldsymbol{\Sigma}|$ can be misleading since two very different variance covariance structures can sometimes result in the same value of generalized variance. Johnson and Wichern (1998) provide certain interesting illustrations of such situations.

Let $\mathbf{u}_{p\times 1}$ and $\mathbf{z}_{q\times 1}$ be two random vectors, with respective means $\boldsymbol{\mu}_u$ and $\boldsymbol{\mu}_z$. Then the covariance of $\mathbf{u}$ *with* $\mathbf{z}$ is defined as

$$\boldsymbol{\Sigma}_{uz} = \operatorname{cov}(\mathbf{u}, \mathbf{z}) = E[(\mathbf{u} - \boldsymbol{\mu}_u)(\mathbf{z} - \boldsymbol{\mu}_z)'] = E[(\mathbf{u} - \boldsymbol{\mu}_u)\mathbf{z}'] = E(\mathbf{u}\mathbf{z}') - \boldsymbol{\mu}_u \boldsymbol{\mu}_z'.$$

Note that as matrices, the p by q matrix $\boldsymbol{\Sigma}_{uz} = \operatorname{cov}(\mathbf{u}, \mathbf{z})$ is *not* the same as the q by p matrix $\boldsymbol{\Sigma}_{zu} = \operatorname{cov}(\mathbf{z}, \mathbf{u})$, the covariance of $\mathbf{z}$ *with* $\mathbf{u}$. They are, however, related in that

$$\boldsymbol{\Sigma}_{uz} = \boldsymbol{\Sigma}_{zu}'.$$

Notice that for a vector $\mathbf{y}$, $\operatorname{cov}(\mathbf{y}, \mathbf{y}) = D(\mathbf{y})$. Thus, when there is no possibility of confusion, we interchangeably use $D(\mathbf{y})$ and $\operatorname{cov}(\mathbf{y})(= \operatorname{cov}(\mathbf{y}, \mathbf{y}))$ to represent the variance-covariance matrix of $\mathbf{y}$.

A variance-covariance matrix is always positive semidefinite (that is, all its eigenvalues are nonnegative). However, in most of the discussion in this text we encounter dispersion matrices which are positive definite, a condition stronger than positive semidefiniteness in that all eigenvalues are strictly positive. Consequently, such dispersion matrices would also admit an inverse. In the subsequent discussion, we assume our dispersion matrix to be positive definite.

Let us partition the vector $\mathbf{y}$ into two subvectors as

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_{1_{p_1\times 1}} \\ \mathbf{y}_{2_{(p-p_1)\times 1}} \end{bmatrix}$$

and partition $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11_{p_1\times p_1}} & \boldsymbol{\Sigma}_{12_{p_1\times(p-p_1)}} \\ \boldsymbol{\Sigma}_{21_{(p-p_1)\times p_1}} & \boldsymbol{\Sigma}_{22_{(p-p_1)\times(p-p_1)}} \end{bmatrix}.$$

Then, $E(\mathbf{y}_1) = \boldsymbol{\mu}_1$, $E(\mathbf{y}_2) = \boldsymbol{\mu}_2$, $D(\mathbf{y}_1) = \boldsymbol{\Sigma}_{11}$, $D(\mathbf{y}_2) = \boldsymbol{\Sigma}_{22}$, $\mathrm{cov}(\mathbf{y}_1, \mathbf{y}_2) = \boldsymbol{\Sigma}_{12}$, $\mathrm{cov}(\mathbf{y}_2, \mathbf{y}_1) = \boldsymbol{\Sigma}_{21}$. We also observe that $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21}$.

The Pearson's *correlation coefficient* between y_i and y_j, denoted by ρ_{ij}, is defined by

$$\rho_{ij} = \frac{\mathrm{cov}(y_i, y_j)}{\sqrt{\mathrm{var}(y_i)\,\mathrm{var}(y_j)}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}},$$

and accordingly, we define the *correlation coefficient matrix* of $\mathbf{y}$ as

$$\mathbf{R} = \begin{bmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \dots & \rho_{2p} \\ & & & \\ \rho_{p1} & \rho_{p2} & \dots & \rho_{pp} \end{bmatrix}$$

It is easy to verify that the correlation coefficient matrix $\mathbf{R}$ is a symmetric positive definite matrix in which all the diagonal elements are unity. The matrix $\mathbf{R}$ can be written, in terms of matrix $\boldsymbol{\Sigma}$, as

$$\mathbf{R} = [\mathrm{diag}\,(\boldsymbol{\Sigma})]^{-1/2}\boldsymbol{\Sigma}[\mathrm{diag}\,(\boldsymbol{\Sigma})]^{-1/2},$$

where $\mathrm{diag}\,(\boldsymbol{\Sigma})$ is the diagonal matrix obtained by retaining the diagonal elements of $\boldsymbol{\Sigma}$ and by replacing all the nondiagonal elements by zero. Further, the *square root* of any matrix $\mathbf{A}$, denoted by $\mathbf{A}^{\frac{1}{2}}$, is a symmetric matrix satisfying the condition, $\mathbf{A} = \mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}$.

The probability distribution (density) of a vector $\mathbf{y}$, denoted by $f(\mathbf{y})$, is the same as the joint probability distribution of $y_1, \dots, y_p$. The marginal distribution $f_1(\mathbf{y}_1)$ of $\mathbf{y}_1 = (y_1, \dots, y_{p_1})'$, a subvector of $\mathbf{y}$, is obtained by integrating out $\mathbf{y}_2 = (y_{p_1+1}, \dots, y_p)'$ from the density $f(\mathbf{y})$. The conditional distribution of $\mathbf{y}_2$, when $\mathbf{y}_1$ has been held fixed, is denoted by $g(\mathbf{y}_2|\mathbf{y}_1)$ and is given by

$$g(\mathbf{y}_2|\mathbf{y}_1) = f(\mathbf{y})/f_1(\mathbf{y}_1).$$

An important concept arising from conditional distribution is the *partial correlation coefficient*. If we partition $\mathbf{y}$ as $(\mathbf{y}'_1, \mathbf{y}'_2)'$ where $\mathbf{y}_1$ is a p_1 by 1 vector and $\mathbf{y}_2$ is a $(p - p_1)$ by 1 vector, then the partial correlation coefficient between two components of $\mathbf{y}_1$, say y_i and y_j, is defined as the Pearson's correlation coefficient between y_i and y_j conditional on $\mathbf{y}_2$ (that is, for a given $\mathbf{y}_2$). If $\boldsymbol{\Sigma}_{11\cdot 2} = (a_{ij})$ is the p_1 by p_1 variance-covariance matrix of $\mathbf{y}_1$ given $\mathbf{y}_2$, then the population partial correlation coefficient between y_i and y_j, $i, j = 1, \dots, p_1$ is given by

$$\rho_{ij\cdot p_1+1,\dots,p} = a_{ij}/\sqrt{a_{ii}a_{jj}}.$$

The matrix of all partial correlation coefficients $\rho_{ij,p_1+1,\dots,p}$, $i, j = 1, \dots, p_1$ is denoted by $\mathbf{R}_{11\cdot 2}$. More simply, using the matrix notations, $\mathbf{R}_{11\cdot 2}$ can be computed as

$$[\mathrm{diag}\,(\boldsymbol{\Sigma}_{11.2})]^{-\frac{1}{2}}\boldsymbol{\Sigma}_{11.2}[\mathrm{diag}\,(\boldsymbol{\Sigma}_{11.2})]^{-\frac{1}{2}},$$

where $\mathrm{diag}\,(\boldsymbol{\Sigma}_{11.2})$ is a diagonal matrix with respective diagonal entries the same as those in $\boldsymbol{\Sigma}_{11.2}$.

Many times it is of interest to find the correlation coefficients between y_i and y_j, $i, j = 1, \ldots, p$, conditional on all $y_k, k = 1, \ldots, p, k \neq i, k \neq j$. In this case, the partial correlation between y_i and y_j can be interpreted as the strength of correlation between the two variables after eliminating the effects of all the remaining variables.

In many linear model situations, we would like to examine the overall association of a set of variables with a given variable. This is often done by finding the correlation between the variable and a particular linear combination of other variables. The *Multiple correlation coefficient* is an index measuring the association between a random variable y_1 and the set of remaining variables represented by a $(p-1)$ by 1 vector $\mathbf{y}_2$. It is defined as the maximum correlation between y_1 and $\mathbf{c}'\mathbf{y}_2$, a linear combination of $\mathbf{y}_2$, where the maximum is taken over all possible nonzero vectors $\mathbf{c}$. This maximum value, representing the multiple correlation coefficient between y_1 and $\mathbf{y}_2$, is given by

$$\left(\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\right)^{\frac{1}{2}} / \Sigma_{11}^{\frac{1}{2}}$$

where

$$D\begin{bmatrix} y_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix},$$

and the maximum is attained for the choice $\mathbf{c} = \mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$. The multiple correlation coefficient always lies between zero and one. The square of the multiple correlation coefficient, often referred to as the *population coefficient of determination*, is generally used to indicate the power of prediction or the effect of regression.

The concept of multiple correlation can be extended to the case in which the random variable y_1 is replaced by a random vector. This leads to what are called *canonical correlation coefficients*.

1.3 Multivariate Normal Distribution

A probability distribution that plays a pivotal role in multivariate analysis is *multivariate normal distribution*. We say that $\mathbf{y}$ has a multivariate normal distribution (with a mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix $\mathbf{\Sigma}$) if its density is given by

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \cdot \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right).$$

In notation, we state this fact as $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$. Observe that the above density is a straightforward extension of the univariate normal density to which it will reduce when $p = 1$.

Important properties of the multivariate normal distribution include some of the following:

- Let $\mathbf{A}_{r\times p}$ be a fixed matrix, then $\mathbf{A}\mathbf{y} \sim N_r(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}')(r \leq p)$. It may be added that $\mathbf{A}\mathbf{y}$ will admit the density if $\mathbf{A}\mathbf{\Sigma}\mathbf{A}'$ is nonsingular, which will happen if and only if all rows of $\mathbf{A}$ are linearly independent. Further, in principle, r can also be greater than p. However, in that case, the matrix $\mathbf{A}\mathbf{\Sigma}\mathbf{A}'$ will not be nonsingular. Consequently, the vector $\mathbf{A}\mathbf{y}$ will not admit a density function.
- Let $\mathbf{G}$ be such that $\mathbf{\Sigma}^{-1} = \mathbf{G}\mathbf{G}'$, then $\mathbf{G}'\mathbf{y} \sim N_p(\mathbf{G}'\boldsymbol{\mu}, I)$ and $\mathbf{G}'(\mathbf{y}-\boldsymbol{\mu}) \sim N_p(\mathbf{0}, I)$.
- Any fixed linear combination of $y_1, \ldots, y_p$, say $\mathbf{c}'\mathbf{y}, \mathbf{c}_{p\times 1} \neq \mathbf{0}$ is also normally distributed. Specifically, $\mathbf{c}'\mathbf{y} \sim N_1(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\mathbf{\Sigma}\mathbf{c})$.
- The subvectors $\mathbf{y}_1$ and $\mathbf{y}_2$ are also normally distributed, specifically, $\mathbf{y}_1 \sim N_{p_1}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_{11})$ and $\mathbf{y}_2 \sim N_{p-p_1}(\boldsymbol{\mu}_2, \mathbf{\Sigma}_{22})$.

- Individual components $y_1, \ldots, y_p$ are all normally distributed. That is, $y_i \sim N_1(\mu_i, \sigma_{ii})$, $i = 1, \ldots, p$.
- The conditional distribution of $\mathbf{y}_1$ given $\mathbf{y}_2$, written as $\mathbf{y_1}|\mathbf{y_2}$, is also normal. Specifically,

$$\mathbf{y}_1|\mathbf{y}_2 \sim N_{p_1}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2),\ \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right).$$

  Let $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2) = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{y}_2 = \mathbf{B_0} + \mathbf{B}_1\mathbf{y}_2$, and $\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. The conditional expectation of $\mathbf{y}_1$ for given values of $\mathbf{y}_2$ or the regression function of $\mathbf{y}_1$ on $\mathbf{y}_2$ is $\mathbf{B_0} + \mathbf{B}_1\mathbf{y}_2$, which is linear in $\mathbf{y}_2$. This is a key fact for multivariate multiple linear regression modeling. The matrix $\boldsymbol{\Sigma}_{11.2}$ is usually represented by the variance-covariance matrix of error components in these models. An analogous result (and the interpretation) can be stated for the conditional distribution of $\mathbf{y}_2$ given $\mathbf{y}_1$.
- Let $\boldsymbol{\delta}$ be a fixed $p \times 1$ vector, then

$$\mathbf{y} + \boldsymbol{\delta} \sim N_p(\boldsymbol{\mu} + \boldsymbol{\delta}, \boldsymbol{\Sigma}).$$

- The random components $y_1, \ldots, y_p$ are all independent if and only if $\boldsymbol{\Sigma}$ is a diagonal matrix; that is, when all the covariances (or correlations) are zero.
- Let $\mathbf{u}_1$ and $\mathbf{u}_2$ be respectively distributed as $N_p(\boldsymbol{\mu}_{u_1}, \boldsymbol{\Sigma}_{u_1})$ and $N_p(\boldsymbol{\mu}_{u_2}, \boldsymbol{\Sigma}_{u_2})$, then

$$\mathbf{u}_1 \pm \mathbf{u}_2 \sim N_p(\boldsymbol{\mu}_{u_1} \pm \boldsymbol{\mu}_{u_2}, \boldsymbol{\Sigma}_{u_1} + \boldsymbol{\Sigma}_{u_2} \pm (\mathrm{cov}(\mathbf{u}_1, \mathbf{u}_2) + \mathrm{cov}(\mathbf{u}_2, \mathbf{u}_1))).$$

  Note that if $\mathbf{u}_1$ and $\mathbf{u}_2$ were independent, the last two covariance terms would drop out.

There is a vast amount of literature available on multivariate normal distribution, its properties, and the evaluations of multivariate normal probabilities. See Kshirsagar (1972), Rao (1973), and Tong (1990) among many others for further details.

1.4 Sampling from Multivariate Normal Populations

Suppose we have a random sample of size n, say $\mathbf{y}_1, \ldots, \mathbf{y}_n$, from the p dimensional multivariate normal population $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Since $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are independently and identically distributed (*iid*), their sample mean

$$\bar{\mathbf{y}} = \frac{1}{n}[\mathbf{y}_1 + \cdots + \mathbf{y}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i \tag{1.1}$$

is also normally distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$. Thus, $\bar{\mathbf{y}}$ is an unbiased estimator of $\boldsymbol{\mu}$. Also, observe that $\bar{\mathbf{y}}$ has a dispersion matrix which is a $\frac{1}{n}$ multiple of the original population variance-covariance matrix. These results are straightforward generalizations of the corresponding well known univariate results.

The sample variance of the univariate normal theory is generalized to the sample variance-covariance matrix in the multivariate context. Accordingly, the chi-square distribution is generalized to a matrix distribution known as the *Wishart distribution*.

The p by p sample variance-covariance matrix is obtained as

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' = \frac{1}{n-1}\left\{\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i' - n\bar{\mathbf{y}}\bar{\mathbf{y}}'\right\}. \tag{1.2}$$

The matrix $\mathbf{S}$ is an unbiased estimator of $\boldsymbol{\Sigma}$. Note that $\mathbf{S}$ is a p by p symmetric matrix. Thus, it contains only $\frac{p(p+1)}{2}$ different random variables.

Let

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \vdots \\ \mathbf{y}_n' \end{bmatrix}$$

be the n by p data matrix obtained by stacking $\mathbf{y}_1', \ldots, \mathbf{y}_n'$ one atop the other. Let $\mathbf{I}_n$ stand for an n by n identity matrix and $\mathbf{1}_n$ be an n by 1 column vector with all elements as 1. Then, in terms of $\mathbf{Y}$, the sample mean $\bar{\mathbf{y}}$ can be written as

$$\bar{\mathbf{y}} = \frac{1}{n}\mathbf{Y}'\mathbf{1}_n$$

and the sample variance-covariance matrix can be written as

$$\mathbf{S} = \frac{1}{n-1}\left\{\mathbf{Y}'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right)\mathbf{Y}\right\} = \frac{1}{n-1}\left\{\mathbf{Y}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{1}_n\mathbf{1}_n'\mathbf{Y}\right\} = \frac{1}{n-1}\{\mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{y}}\bar{\mathbf{y}}'\}.$$

It is known that $(n-1)\mathbf{S}$ follows a p-(matrix) variate Wishart distribution with $(n-1)$ degrees of freedom and expectation $(n-1)\mathbf{\Sigma}$. We denote this as $(n-1)\mathbf{S} \sim W_p(n-1, \mathbf{\Sigma})$. Also, $\mathbf{S}$ is an unbiased estimator of $\mathbf{\Sigma}$ (as mentioned earlier, this is always true regardless of the underlying multivariate normality assumption and consequently, without any specific reference to the Wishart distribution).

Since $(n-1)\mathbf{S}$ has a Wishart distribution, the sample variance-covariance matrix $\mathbf{S}$ possesses certain other important properties. Many of these properties are used to obtain the distributions of various estimators and test statistics. Some of these properties are listed as follows.

- $(n-1)s_{ii}/\sigma_{ii} \sim \chi^2(n-1), i = 1, \ldots, p.$
- Let

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix},$$

  $\mathbf{S}_{11.2} = \mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}$, $\mathbf{\Sigma}_{11.2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$, $\mathbf{S}_{22.1} = \mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$ and $\mathbf{\Sigma}_{22.1} = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}$, then
  (a) $(n-1)\mathbf{S}_{11} \sim W_{p_1}((n-1), \mathbf{\Sigma}_{11})$.
  (b) $(n-1)\mathbf{S}_{22} \sim W_{p_2}((n-1), \mathbf{\Sigma}_{22})$.
  (c) $(n-1)\mathbf{S}_{11.2} \sim W_{p_1}((n-p+p_1-1), \mathbf{\Sigma}_{11.2})$.
  (d) $(n-1)\mathbf{S}_{22.1} \sim W_{p_2}((n-p_1-1), \mathbf{\Sigma}_{22.1})$.
  (e) $\mathbf{S}_{11}$ and $\mathbf{S}_{22.1}$ are independently distributed.
  (f) $\mathbf{S}_{22}$ and $\mathbf{S}_{11.2}$ are independently distributed.
- Let s^{ii} and σ^{ii} be the i^{th} diagonal elements of $\mathbf{S}^{-1}$ and $\mathbf{\Sigma}^{-1}$ respectively, then $(n-1)\sigma^{ii}/s^{ii} \sim \chi^2(n-p)$.
- Let $\mathbf{c} \neq \mathbf{0}$ be an arbitrary but fixed vector, then

$$(n-1)\frac{\mathbf{c}'\mathbf{S}\mathbf{c}}{\mathbf{c}'\mathbf{\Sigma}\mathbf{c}} \sim \chi^2(n-1),$$

$$\text{and } (n-1)\frac{\mathbf{c}'\mathbf{\Sigma}^{-1}\mathbf{c}}{\mathbf{c}'\mathbf{S}^{-1}\mathbf{c}} \sim \chi^2(n-p).$$

- Let $\mathbf{H}$ be an arbitrary but fixed $k \times p$ matrix $(k \leq p)$, then

$$(n-1)\mathbf{HSH}' \sim W_k(n-1, \mathbf{H\Sigma H}').$$

  In principle, k can also be greater than p but in such a case, the matrix $(n-1)\mathbf{HSH}'$ does not admit a probability density.

As a consequence of the above result, if we take $k = p$ and $\mathbf{H} = \mathbf{G}'$ where $\mathbf{\Sigma}^{-1} = \mathbf{GG}'$, then $(n-1)\,\mathbf{S}^* = (n-1)\mathbf{G}'\mathbf{SG} \sim W_p(n-1, \mathbf{I})$.

In the above discussion, we observed that the Wishart distribution arises naturally in the multivariate normal theory as the distribution of the sample variance-covariance matrix (of course, apart from a scaling by $(n-1)$). Another distribution which is closely related to the Wishart distribution and is useful in various associated hypothesis testing problems is the matrix variate Beta (Type 1) distribution. For example, if $\mathbf{A}_1$ and $\mathbf{A}_2$ are two independent random matrices with $\mathbf{A}_1 \sim W_p(f_1, \mathbf{\Sigma})$ and $\mathbf{A}_2 \sim W_p(f_2, \mathbf{\Sigma})$, then $\mathbf{B} = (\mathbf{A}_1 + \mathbf{A}_2)^{-\frac{1}{2}}\mathbf{A}_1(\mathbf{A}_1 + \mathbf{A}_2)^{-\frac{1}{2}}$ follows a matrix variate Beta Type 1 distribution, denoted by $B_p(\frac{n_1-1}{2}, \frac{n_2-1}{2}$, Type 1). Similarly, $\mathbf{B}^* = \mathbf{A}_2^{-\frac{1}{2}}\mathbf{A}_1\mathbf{A}_2^{-\frac{1}{2}}$ follows $B_p(\frac{n_1-1}{2}, \frac{n_2-1}{2}$, Type 2), a matrix variate Beta Type 2 (or a matrix variate F apart from a constant) distribution. The matrices $\mathbf{A}_2^{-\frac{1}{2}}$ and $(\mathbf{A}_1 + \mathbf{A}_2)^{-\frac{1}{2}}$ respectively are the symmetric "square root" matrices of $\mathbf{A}_2^{-1}$ and $(\mathbf{A}_1 + \mathbf{A}_2)^{-1}$ in the sense that $\mathbf{A}_2^{-1} = (\mathbf{A}_2)^{-\frac{1}{2}}(\mathbf{A}_2)^{-\frac{1}{2}}$ and $(\mathbf{A}_1 + \mathbf{A}_2)^{-1} = (\mathbf{A}_1 + \mathbf{A}_2)^{-\frac{1}{2}}(\mathbf{A}_1 + \mathbf{A}_2)^{-\frac{1}{2}}$. The eigenvalues of the matrices $\mathbf{B}$ and $\mathbf{B}^*$ appear in the expressions of various test statistics used in hypothesis testing problems in multivariate analysis of variance.

Another important fact about the sample mean $\bar{\mathbf{y}}$ and the sample variance-covariance matrix $\mathbf{S}$ is that they are statistically independent under the multivariate normal sampling theory. This fact plays an important role in constructing test statistics for certain statistical hypotheses. For details, see Kshirsagar (1972), Timm (1975), or Muirhead (1982).

1.5 Some Important Sample Statistics and Their Distributions

We have already encountered two important sample statistics in the previous section, namely the sample mean vector $\bar{\mathbf{y}}$ in Equation 1.1 and the sample variance-covariance matrix $\mathbf{S}$ in Equation 1.2. These quantities play a pivotal role in defining the test statistics useful in various hypothesis testing problems. The underlying assumption of multivariate normal population is crucial in obtaining the distribution of these test statistics. Therefore, we will assume that the sample $\mathbf{y}_1, \ldots, \mathbf{y}_n$ of size n is obtained from a multivariate population $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$.

As we have already indicated, $\bar{\mathbf{y}} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma}/n)$ and $(n-1)\mathbf{S} \sim W_p(n-1, \mathbf{\Sigma})$. Consequently, any linear combination of $\bar{\mathbf{y}}$, say $\mathbf{c}'\bar{\mathbf{y}}$, $\mathbf{c} \neq \mathbf{0}$, follows $N_1(\mathbf{c}'\boldsymbol{\mu}, \mathbf{c}'\mathbf{\Sigma}\mathbf{c}/n)$ and the quadratic form $(n-1)\mathbf{c}'\mathbf{Sc}/\mathbf{c}'\mathbf{\Sigma}\mathbf{c} \sim \chi^2(n-1)$. Further, as pointed out earlier, $\bar{\mathbf{y}}$ and $\mathbf{S}$ are independently distributed and hence the quantity

$$t = \sqrt{n}\mathbf{c}'(\bar{\mathbf{y}} - \boldsymbol{\mu})/\sqrt{\mathbf{c}'\mathbf{Sc}}$$

follows a t-distribution with $(n-1)$ degrees of freedom. A useful application of this fact is in testing problems for certain contrasts or in testing problems involving a given linear combination of the components of the mean vector.

Often interest may be in testing a hypothesis if the population has its mean vector equal to a given vector, say $\boldsymbol{\mu}_0$. Since $\bar{\mathbf{y}} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma}/n)$, it follows that $\mathbf{z} = \sqrt{n}\mathbf{\Sigma}^{-\frac{1}{2}}(\bar{\mathbf{y}} - \boldsymbol{\mu})$ follows $N_p(\mathbf{0}, \mathbf{I})$. This implies that the components of $\mathbf{z}$ are independent and have the standard normal distribution. As a result, if $\boldsymbol{\mu}$ is equal to $\boldsymbol{\mu}_0$ the quantity, $z_1^2 + \cdots + z_p^2 = \mathbf{z}'\mathbf{z} = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)'\mathbf{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)$ follows a chi-square distribution with p degrees of freedom. On the other hand, if μ is not equal to μ_0, then this quantity will not have a chi-square distribution. This observation provides a way of testing the hypothesis that the mean of the normal population is equal to a given vector μ_0. However, the assumption of known $\mathbf{\Sigma}$ is needed to actually perform this test. If $\mathbf{\Sigma}$ is unknown, it seems natural to replace it in $n(\bar{\mathbf{y}} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu})$ by its unbiased estimator $\mathbf{S}$, leading to Hotelling's T^2 test statistic defined as

$$T^2 = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_0),$$

where we assume that $n \geq p+1$. This assumption ensures that $\mathbf{S}$ admits an inverse. Under the hypothesis mentioned above, namely $\mu = \mu_0$, the quantity $\frac{n-p}{p(n-1)}T^2$ follows an F distribution with degrees of freedom p and $n-p$.

Assuming normality, the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are known to be

$$\hat{\boldsymbol{\mu}}_{ml} = \bar{\mathbf{y}}$$

and

$$\hat{\boldsymbol{\Sigma}}_{ml} = \mathbf{S}_n = \frac{1}{n}\mathbf{Y}'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right)\mathbf{Y} = \frac{n-1}{n}\mathbf{S}.$$

While $\hat{\boldsymbol{\mu}}_{ml} = \bar{\mathbf{y}}$ is unbiased for $\boldsymbol{\mu}$, $\hat{\boldsymbol{\Sigma}}_{ml} = \mathbf{S}_n$ is a (negatively) biased estimator of $\boldsymbol{\Sigma}$. These quantities are also needed in the process of deriving various maximum likelihood-based tests for the hypothesis testing problems. In general, to test a hypothesis H_0, the likelihood ratio test based on the maximum likelihood estimates is obtained by first maximizing the likelihood within the parameter space restricted by H_0. The next step is maximizing it over the entire parameter space (that is, by evaluating the likelihood at $\hat{\boldsymbol{\mu}}_{ml}$ and $\hat{\boldsymbol{\Sigma}}_{ml}$), and then taking the ratio of the two. Thus, the likelihood ratio test statistic can be written as

$$L = \frac{\max\limits_{H_0} f(\mathbf{Y})}{\max\limits_{\text{unrestricted}} f(\mathbf{Y})} = \frac{\max\limits_{H_0} g(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{Y})}{\max\limits_{\text{unrestricted}} g(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{Y})},$$

where for optimization purposes the function $g(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{Y}) = f(\mathbf{Y})$ is viewed as a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given data $\mathbf{Y}$. A related test statistic is the Wilks' Λ, which is the $(2/n)^{th}$ power of L. For large n, the quantity $-2\log L$ approximately follows a chi-square distribution, with degrees of freedom ν, which is a function of the sample size n, the number of parameters estimated, and the number of restrictions imposed by the parameters involved under H_0.

A detailed discussion of various likelihood ratio tests in multivariate analysis context can be found in Kshirsagar (1972), Muirhead (1982) or in Anderson (1984). A brief review of some of the relevant likelihood ratio tests is given in Chapter 6. There are certain other intuitive statistical tests which have been proposed in various contexts and used in applications instead of the likelihood ratio tests. Some of these tests have been discussed in Chapter 3.

1.6 Tests for Multivariate Normality

Often before doing any statistical modeling, it is crucial to verify if the data at hand satisfy the underlying distributional assumptions. Many times such an examination may be needed for the residuals after fitting various models. For most multivariate analyses, it is thus very important that the data indeed follow the multivariate normal, or if not exactly at least approximately. If the answer to such a query is affirmative, it can often reduce the task of searching for procedures which are robust to the departures from multivariate normality. There are many possibilities for departure from multivariate normality and no single procedure is likely to be robust with respect to all such departures from the multivariate normality assumption. Gnanadesikan (1980) and Mardia (1980) provide excellent reviews of various procedures to verify this assumption.

This assumption is often checked by individually examining the univariate normality through various Q-Q plots or some other plots and can at times be very subjective. One of the relatively simpler and mathematically tractable ways to find a support for the assumption of multivariate normality is by using the tests based on Mardia's *multivariate skewness* and *kurtosis* measures. For any general multivariate distribution we define these respectively as

$$\beta_{1,p} = E\left\{(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}^3,$$

provided that $\mathbf{x}$ is independent of $\mathbf{y}$ but has the same distribution and

$$\beta_{2,p} = E\left\{(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}^2,$$

provided that the expectations in the expressions of $\beta_{1,p}$ and $\beta_{2,p}$ exist. For the multivariate normal distribution, $\beta_{1,p} = 0$ and $\beta_{2,p} = p(p+2)$.

For a sample of size n, the estimates of $\beta_{1,p}$ and $\beta_{2,p}$ can be obtained as

$$\hat{\beta}_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}^3$$

$$\hat{\beta}_{2,p} = \frac{1}{n}\sum_{i=1}^{n} g_{ii}^2 = \frac{1}{n}\sum_{i=1}^{n} d_i^4$$

where $g_{ij} = (\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{S}_n^{-1}(\mathbf{y}_j - \bar{\mathbf{y}})$, and $d_i = \sqrt{g_{ii}}$ is the sample version of the squared *Mahalanobis distance* (Mahalanobis, 1936) between $\mathbf{y}_i$ and ($\boldsymbol{\mu}$ which is approximated by) $\bar{\mathbf{y}}$ (Mardia, 1970).

The quantity $\hat{\beta}_{1,p}$ (which is the same as the square of sample skewness coefficient when $p = 1$) as well as $\hat{\beta}_{2,p}$ (which is the same as the sample kurtosis coefficient when $p = 1$) are nonnegative. For the multivariate normal data, we would expect $\hat{\beta}_{1,p}$ to be close to zero. If there is a departure from the spherical symmetry (that is, zero correlation and equal variance), $\hat{\beta}_{2,p}$ will be large. The quantity $\hat{\beta}_{2,p}$ is also useful in indicating the extreme behavior in the squared Mahalanobis distance of the observations from the sample mean.

Thus, $\hat{\beta}_{1,p}$ and $\hat{\beta}_{2,p}$ can be utilized to detect departure from multivariate normality. Mardia (1970) has shown that for large samples, $\kappa_1 = n\hat{\beta}_{1,p}/6$ follows a chi-square distribution with degrees of freedom $p(p+1)(p+2)/6$, and $\kappa_2 = \{\hat{\beta}_{2,p} - p(p+2)\}/\{8p(p+2)/n\}^{\frac{1}{2}}$ follows a standard normal distribution. Thus, we can use the quantities κ_1 and κ_2 to test the null hypothesis of multivariate normality. For small n, see the tables for the critical values for these test statistics given by Mardia (1974). He also recommends (Mardia, Kent, and Bibby, 1979, p. 149) that if both the hypotheses are accepted, the normal theory for various tests on the mean vector or the covariance matrix can be used. However, in the presence of nonnormality, the normal theory tests on the mean are sensitive to $\beta_{1,p}$, whereas tests on the covariance matrix are influenced by $\beta_{2,p}$.

For a given data set, the multivariate kurtosis can be computed using the CALIS procedure in SAS/STAT software. Notice that the quantities reported in the corresponding SAS output are the centered quantity ($\hat{\beta}_{2,p} - p(p+2)$) (shown in Output 1.1 as Mardia's Multivariate Kurtosis) and κ_2 (shown in Output 1.1 as Normalized Multivariate Kurtosis).

EXAMPLE 1 ***Testing Multivariate Normality, Cork Data*** As an illustration, we consider the cork boring data of Rao (1948) given in Table 1.1, and test the hypothesis that this data set can be considered as a random sample from a multivariate normal population. The data set provided in Table 1.1 consists of the weights of cork borings in four directions (north, east, south, and west) for 28 trees in a block of plantations.

E. S. Pearson had pointed out to C. R. Rao, apparently without any formal statistical testing, that the data are exceedingly asymmetrically distributed. It is therefore of interest to formally test if the data can be assumed to have come from an $N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

TABLE 1.1 Weights of Cork Boring (in Centigrams) in Four Directions for 28 Trees

Tree	N	E	S	W	Tree	N	E	S	W
1	72	66	76	77	15	91	79	100	75
2	60	53	66	63	16	56	68	47	50
3	56	57	64	58	17	79	65	70	61
4	41	29	36	38	18	81	80	68	58
5	32	32	35	36	19	78	55	67	60
6	30	35	34	26	20	46	38	37	38
7	39	39	31	27	21	39	35	34	37
8	42	43	31	25	22	32	30	30	32
9	37	40	31	25	23	60	50	67	54
10	33	29	27	36	24	35	37	48	39
11	32	30	34	28	25	39	36	39	31
12	63	45	74	63	26	50	34	37	40
13	54	46	60	52	27	43	37	39	50
14	47	51	52	43	28	48	54	57	43

The SAS statements required to compute the multivariate kurtosis using PROC CALIS are given in Program 1.1. A part of the output giving the value of Mardia's multivariate kurtosis ($= -1.0431$) and normalized multivariate kurtosis ($= -0.3984$) is shown as Output 1.1. The output also indicates the observations which are most influential. Although the procedure does not provide the value of multivariate skewness, the IML procedure statements given in Program 1.2 perform all the necessary calculations to compute the multivariate skewness and kurtosis. The results are shown in Output 1.2, which also reports Mardia's test statistics κ_1 and κ_2 described above along with the corresponding p values.

In this program, for the 28 by 4 data matrix $\mathbf{Y}$, we first compute the maximum likelihood estimate of the variance-covariance matrix. This estimate is given by $\mathbf{S}_n = \frac{1}{n}\mathbf{Y}'\mathbf{Q}\mathbf{Y}$, where $\mathbf{Q} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$. Also, since the quantities $g_{ij}, i, j = 1, \ldots, n$ needed in the expressions of multivariate skewness and kurtosis are the elements of matrix $\mathbf{G} = \mathbf{Q}\mathbf{Y}\mathbf{S}_n^{-1}\mathbf{Y}'\mathbf{Q}$, we compute the matrix $\mathbf{G}$, using this formula. Their p values are then reported as PVALSKEW and PVALKURT in Output 1.2. It may be remarked that in Program 1.2 the raw data are presented as a matrix entity. One can alternatively read the raw data (as done in Program 1.1) as a data set and then convert it to a matrix. In Appendix 1, we have provided the SAS code to perform this conversion.

```
/* Program 1.1 */

options ls=64 ps=45 nodate nonumber;
data cork;
infile 'cork.dat' firstobs = 1;
input north east south west;
proc calis  data = cork kurtosis;
title1 j=l "Output 1.1";
title2 "Computation of Mardia's Kurtosis";
lineqs
north = e1,
east = e2,
south = e3,
west = e4;
std
e1=eps1, e2=eps2, e3=eps3, e4=eps4;
cov
e1=eps1, e2=eps2, e3=eps3, e4=eps4;
run ;
```

Output 1.1

```
                    Output 1.1
          Computation of Mardia's Kurtosis

   Mardia's Multivariate Kurtosis . . . . . . . .     -1.0431
   Relative Multivariate Kurtosis . . . . . . . .      0.9565
   Normalized Multivariate Kurtosis . . . . . . .     -0.3984
   Mardia Based Kappa (Browne, 1982). . . . . . .     -0.0435
   Mean Scaled Univariate Kurtosis  . . . . . . .     -0.0770
   Adjusted Mean Scaled Univariate Kurtosis . . .     -0.0770
```

```
/* Program 1.2 */

title 'Output 1.2';
options ls = 64 ps=45 nodate nonumber;

/* This program is for testing the multivariate
normality using Mardia's skewness and kurtosis measures.
Application on C. R. Rao's cork data */

proc iml ;
y ={
72 66 76 77,
60 53 66 63,
56 57 64 58,
41 29 36 38,
32 32 35 36,
30 35 34 26,
39 39 31 27,
42 43 31 25,
37 40 31 25,
33 29 27 36,
32 30 34 28,
63 45 74 63,
54 46 60 52,
47 51 52 43,
91 79 100 75,
56 68 47 50,
79 65 70 61,
81 80 68 58,
78 55 67 60,
46 38 37 38,
39 35 34 37,
32 30 30 32,
60 50 67 54,
35 37 48 39,
39 36 39 31,
50 34 37 40,
43 37 39 50,
48 54 57 43} ;
/* Matrix y can be created from a SAS data set as follows:
data cork;
infile 'cork.dat';
input y1 y2 y3 y4;
run;
proc iml;
use cork;
read all into y;
```

```
See Appendix 1 for details.
*/
/* Here we determine the number of data points and the dimension
of the vector. The variable dfchi is the degrees of freedom for
the chi square approximation of Multivariate skewness. */

n = nrow(y) ;
p = ncol(y) ;
dfchi = p*(p+1)*(p+2)/6 ;

/* q is projection matrix, s is the maximum likelihood estimate
of the variance covariance matrix, g_matrix is n by n the matrix
of g(i,j) elements, beta1hat and beta2hat are respectively the
Mardia's sample skewness and kurtosis measures, kappa1 and kappa2
are the test statistics based on skewness and kurtosis to test
for normality and pvalskew and pvalkurt are corresponding p
values. */

q = i(n) - (1/n)*j(n,n,1);
s = (1/(n))*y`*q*y ; s_inv = inv(s) ;
g_matrix = q*y*s_inv*y`*q;
beta1hat = (  sum(g_matrix#g_matrix#g_matrix)  )/(n*n);
beta2hat =trace(  g_matrix#g_matrix  )/n ;

kappa1 = n*beta1hat/6 ;
kappa2 = (beta2hat - p*(p+2) ) /sqrt(8*p*(p+2)/n) ;

pvalskew = 1 - probchi(kappa1,dfchi) ;
pvalkurt = 2*( 1 - probnorm(abs(kappa2)) );
print s ;
print s_inv ;
print 'TESTS:';
print 'Based on skewness: ' beta1hat kappa1 pvalskew ;
print 'Based on kurtosis: ' beta2hat kappa2 pvalkurt;
```

Output 1.2

```
                         Output 1.2

                   S
      280.03444 215.76148 278.13648 218.19005
      215.76148 212.07526 220.87883 165.25383
      278.13648 220.87883 337.50383 250.27168
      218.19005 165.25383 250.27168  217.9324

              S_INV
      0.0332462 -0.016361 -0.008139 -0.011533
      -0.016361 0.0228758 -0.005199 0.0050046
      -0.008139 -0.005199 0.0276698 -0.019685
      -0.011533 0.0050046 -0.019685 0.0349464

                          TESTS:

                        BETA1HAT    KAPPA1  PVALSKEW
  Based on skewness:  4.4763816 20.889781 0.4036454


                        BETA2HAT    KAPPA2  PVALKURT
  Based on kurtosis:   22.95687 -0.398352 0.6903709
```

For this particular data set with its large p values, neither skewness is significantly different from zero, nor is the value of kurtosis significantly different from that for the 4-variate multivariate normal distribution. Consequently, we may assume multivariate normality for testing the various hypotheses on the mean vector and the covariance matrix as far as the present data set is concerned. This particular data set is extensively analyzed in the later chapters under the assumption of normality.

Often we are less interested in the multivariate normality of the original data and more interested in the joint normality of contrasts or any other set of linear combinations of the variables $y_1, \ldots, y_p$. If $\mathbf{C}$ is the corresponding p by r matrix of linear transformations, then the transformed data can be obtained as $\mathbf{Z} = \mathbf{YC}$. Consequently, the only change in Program 1.2 is to replace the earlier definition of $\mathbf{G}$ by $\mathbf{QYC}(\mathbf{C}'\mathbf{S}_n\mathbf{C})^{-1}\mathbf{C}'\mathbf{Y}'\mathbf{Q}$ and replace p by r in the expressions for κ_1, κ_2 and the degrees of freedom corresponding to κ_1.

EXAMPLE 1 ***Testing for Contrasts, Cork Data (continued)*** Returning to the cork data, if the interest is in testing if the bark deposit is uniform in all four directions, an appropriate set of transformations would be

$$z_1 = y_1 - y_2 + y_3 - y_4, \quad z_2 = y_3 - y_4, \quad z_3 = y_1 - y_3,$$

where y_1, y_2, y_3, y_4 represent the deposit in four directions listed clockwise and starting from north. The 4 by 3 matrix $\mathbf{C}$ for these transformations will be

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 0 & 0 \\ 1 & 1 & -1 \\ -1 & -1 & 0 \end{bmatrix}.$$

It is easy to verify that for these contrasts the assumption of symmetry holds rather more strongly, since the p values corresponding to the skewness are relatively larger. Specifically for these contrasts

$$\hat{\beta}_1 = 1.1770, \qquad \hat{\beta}_2 = 13.5584, \qquad \kappa_1 = 5.4928, \qquad \kappa_2 = -0.6964$$

and the respective p values for skewness and kurtosis tests are 0.8559 and 0.4862. As Rao (1948) points out, this symmetry is not surprising since these are linear combinations, and the contrasts are likely to fit the multivariate normality better than the original data. Since one can easily modify Program 1.1 or Program 1.2 to perform the above analysis on the contrasts z_1, z_2, and z_3, we have not provided the corresponding SAS code or the output.

Mudholkar, McDermott and Srivastava (1992) suggest another simple test of multivariate normality. The idea is based on the facts that (i) the cube root of a chi-square random variable can be approximated by a normal random variable and (ii) the sample mean vector and the sample variance covariance matrix are independent if and only if the underlying distribution is multivariate normal. Lin and Mudholkar (1980) had earlier used these ideas to obtain a test for the univariate normality.

To test multivariate normality (of dimension say p) on the population with mean vector $\boldsymbol{\mu}$ and a variance covariance matrix $\boldsymbol{\Sigma}$, let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be a random sample of size n then the unbiased estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are respectively given by $\bar{\mathbf{y}}$ and $\mathbf{S}$. Corresponding to i^{th} observation we define,

$$\begin{aligned} D_i^2 &= (\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_i - \bar{\mathbf{y}}), \\ W_i &= (D_i^2)^h, \end{aligned}$$

and

$$U_i = \left\{ \sum_{j \neq i} W_j^2 - \left[\sum_{j \neq i} W_j \right]^2 / (n-1) \right\}^{1/3}, \; i = 1, \ldots, n,$$

where

$$h = \frac{1}{3} - \frac{0.11}{p}.$$

Let r be the sample correlation coefficient between (W_i, U_i), $i = 1, \ldots, n$. Under the null hypothesis of multivariate normality of the data, the quantity, $Z_p = tanh^{-1}(r) = \frac{1}{2} ln\{\frac{1+r}{1-r}\}$, is approximately normal with mean $\mu_{n,p} = E(Z_p) = \frac{A_1(p)}{n} - \frac{A_2(p)}{n^2}$, where $A_1(p) = \frac{-1}{p} - .52p$ and $A_2(p) = 0.8p^2$ and variance, $\sigma^2_{n,p} = var(Z_p) = \frac{B_1(p)}{n} - \frac{B_2(p)}{n^2}$, where $B_1(p) = 3 - \frac{1.67}{p} + \frac{.52}{p^2}$ and $B_2(p) = 1.8p - \frac{9.75}{p^2}$. Thus, the test based on Z_p to test the null hypothesis of multivariate normality rejects it at α level of significance if $|z_{n,p}| = \frac{|Z_p - \mu_{n,p}|}{\sigma_{n,p}} \geq z_{\frac{\alpha}{2}}$, where $z_{\frac{\alpha}{2}}$ is the right $\frac{\alpha}{2}$ cutoff point from the standard normal distribution.

EXAMPLE 1 ***Testing Multivariate Normality, Cork Data (continued)*** In Program 1.3, we reconsider the cork data of C. R. Rao (1948) and test the hypothesis of the multivariate normality of the tree population.

```
/* Program 1.3 */

options ls=64 ps=45 nodate nonumber;
title1 'Output 1.3';
title2 'Testing Multivariate Normality (Cube Root Transoformation)';

data D1;
infile 'cork.dat';
input t1 t2 t3 t4 ;
/*
t1=north, t2=east, t3=south, t4=west
n is the number of observations
p is the number of variables
*/
data D2(keep=t1 t2 t3 t4 n p);
set D1;
n=28;
p=4;
run;
data D3(keep=n p);
set D2;
if _n_ > 1 then delete;
run;
proc princomp data=D2 cov std out=D4 noprint;
var t1-t4;
data D5(keep=n1 dsq n p);
set D4;
n1=_n_;
dsq=uss(of prin1-prin4);
run;
data D6(keep=dsq1 n1 );
set  D5;
dsq1=dsq**((1.0/3.0)-(0.11/p));
run;

proc iml;
use D3;
read all var {n p};
u=j(n,1,1);
```

```
use D6;
do k=1 to n;
setin D6 point 0;
sum1=0;
sum2=0;
do data;
read next var{dsq1 n1} ;
if n1 = k then dsq1=0;
sum1=sum1+dsq1**2;
sum2=sum2+dsq1;
end;
u[k]=(sum1-((sum2**2)/(n-1)))**(1.0/3);
end;
varnames={y};
create tyy from u (|colname=varnames|);
append from u;
close tyy;
run;
quit;

data D7;
set D6; set tyy;
run;
proc corr data=D7 noprint outp=D8;
var dsq1;
with y;
run;
data D9;
set D8;
if _TYPE_ ^='CORR' then delete;
run;
data D10(keep=zp r tnp pvalue);
set D9(rename=(dsq1=r));
set D3;
zp=0.5*log((1+r)/(1-r));
b1p=3-1.67/p+0.52/(p**2);
a1p=-1.0/p-0.52*p;
a2p=0.8*p**2;
mnp=(a1p/n)-(a2p/(n**2));
b2p=1.8*p-9.75/(p**2);
ssq1=b1p/n-b2p/(n**2);
snp=ssq1**0.5;
tnp=abs(abs(zp-mnp)/snp);
pvalue=2*(1-probnorm(tnp));
run;
proc print data=D10;
run;
```

The SAS Program 1.3 (adopted from Apprey and Naik (1998)) computes the quantities, Z_p, $\mu_{n,p}$, and $\sigma_{n.p}$ using the expressions listed above. Using these, the test statistic $|z_{n,p}|$ and corresponding p value are computed. A run of the program results in a p value of 0.2216. We thus accept the hypothesis of multivariate normality. This conclusion is consistent with our earlier conclusion using the Mardia's tests for the same data set. Output corresponding to Program 1.3 is suppressed in order to save space.

1.7 Random Vector and Matrix Generation

For various simulation or power studies, it is often necessary to generate a set of random vectors or random matrices. It is therefore of interest to generate these quantities for the probability distributions which arise naturally in the multivariate normal theory. The following sections consider the most common multivariate probability distributions.

1.7.1 Random Vector Generation from $N_p(\mu, \Sigma)$

To generate a random vector from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ use the following steps:

1. Find a matrix $\mathbf{G}$ such that $\boldsymbol{\Sigma} = \mathbf{G}'\mathbf{G}$. This is obtained using the Cholesky decomposition of the symmetric matrix $\boldsymbol{\Sigma}$. The functions ROOT of Half in PROC IML can perform this decomposition.
2. Generate p independent standard univariate normal random variables $z_1, \ldots, z_p$ and let $\mathbf{z} = (z_1, \ldots, z_p)'$.
3. Let $\mathbf{y} = \boldsymbol{\mu} + \mathbf{G}'\mathbf{z}$.

The resulting vector $\mathbf{y}$ is an observation from a $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population. To obtain a sample of size n, we repeat the above-mentioned steps n times within a loop.

1.7.2 Generation of Wishart Random Matrix

To generate a matrix $\mathbf{A}_1 \sim W_p(f, \boldsymbol{\Sigma})$, use the following steps:

1. Find a matrix $\mathbf{G}$ such that $\boldsymbol{\Sigma} = \mathbf{G}'\mathbf{G}$.
2. Generate a random sample of size f, say $\mathbf{z}_1, \ldots, \mathbf{z}_f$ from $N_p(\mathbf{0}, \mathbf{I})$. Let $\mathbf{A}_2 = \sum_{i=1}^{f} \mathbf{z}_i \mathbf{z}_i'$.
3. Define $\mathbf{A}_1 = \mathbf{G}'\mathbf{A}_2\mathbf{G}$.

The generation of Beta matrices can easily be done by first generating two independent Wishart matrices with appropriate degrees of freedom and then forming the appropriate products using these matrices as defined in Section 1.4.

EXAMPLE 2 ***Random Samples from Normal and Wishart Distributions*** In the following example we will illustrate the use of PROC IML for generating samples from the multivariate normal and Wishart distributions respectively. These programs are respectively given as Program 1.4 and Program 1.5. The corresponding outputs have been omitted to save space.

As an example, suppose we want to generate four vectors from $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$\boldsymbol{\mu} = (1\ 3\ 0)'$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 5 \end{bmatrix}.$$

Then save these four vectors as the rows of 4 by 3 matrix $\mathbf{Y}$. It is easy to see that

$$E(\mathbf{Y}) = \begin{bmatrix} \boldsymbol{\mu}' \\ \boldsymbol{\mu}' \\ \boldsymbol{\mu}' \\ \boldsymbol{\mu}' \end{bmatrix} = \mathbf{M}.$$

Also, let $\mathbf{G}$ be a matrix such that $\mathbf{\Sigma} = \mathbf{G}'\mathbf{G}$. This matrix is obtained using the ROOT function which performs the Cholesky decomposition of a symmetric matrix.

```
/* Program 1.4 */

options ls = 64 ps=45 nodate nonumber;
title1 'Output 1.4';

/* Generate n random vector from a p dimensional population
with mean mu and the variance covariance matrix sigma */

proc iml ;
seed = 549065467 ;
n = 4 ;
sigma = { 4 2 1,
          2 3 1,
          1 1 5 };

mu = {1, 3, 0};
p = nrow(sigma);
m = repeat(mu`,n,1) ;
        g =root(sigma);
z =normal(repeat(seed,n,p)) ;
y = z*g + m ;
print 'Multivariate Normal Sample';
print y;
```

We first generate a 4 by 3 random matrix $\mathbf{Z}$, with all its entries distributed as $N(0, 1)$. To do this, we use the normal random number generator (NORMAL) repeated for all the entries of $\mathbf{Z}$, through the REPEAT function. Consequently, if we define $\mathbf{Y} = \mathbf{ZG}+\mathbf{M}$, then the i^{th} row of $\mathbf{Y}$, say $\mathbf{y}_i'$, can be written in terms of the i^{th} row of $\mathbf{Z}$, say $\mathbf{z}_i'$, as

$$\mathbf{y}_i' = \mathbf{z}_i'\mathbf{G} + \boldsymbol{\mu}'$$

or when written as a column vector

$$\mathbf{y}_i = \mathbf{G}'\mathbf{z}_i + \boldsymbol{\mu}.$$

Consequently, $\mathbf{y}_i, i = 1, \ldots, n\ (= 4\text{ here})$ are normally distributed with the mean $E(\mathbf{y}_i) = \mathbf{G}'E(\mathbf{z}_i) + \boldsymbol{\mu} = \boldsymbol{\mu}$ and the variance covariance matrix $D(\mathbf{y}_i) = \mathbf{G}'D(\mathbf{z}_i)\mathbf{G} + \mathbf{0} = \mathbf{G}'\mathbf{IG} = \mathbf{G}'\mathbf{G} = \mathbf{\Sigma}$.

Program 1.5 illustrates the generation of $n = 4$ Wishart matrices from $W_p(f, \mathbf{\Sigma})$ with $f = 7$, $p = 3$, and $\mathbf{\Sigma}$ as given in the previous program. After obtaining the matrix $\mathbf{G}$, as earlier, we generate a 7 by 3 matrix $\mathbf{T}$, for which all the elements are distributed as the standard normal. Consequently, the matrix $\mathbf{W} = \mathbf{G}'\mathbf{T}'\mathbf{TG}$, (written as $\mathbf{X}'\mathbf{X}$, where $\mathbf{X} = \mathbf{TG}$) follows $W_3(7, \mathbf{\Sigma})$ distribution. We have used a DO loop to repeat the process $n = 4$ times to obtain four such matrices.

```
/* Program 1.5 */

options ls=64 ps=45 nodate nonumber;
title1 'Output 1.5';
/* Generate n Wishart matrices of order p by p
with degrees of freedom f */

proc iml;
n = 4 ;
f = 7 ;
```

```
seed = 4509049 ;
sigma = {4 2 1,
         2 3 1,
         1 1 5 } ;
        g = root(sigma);
p = nrow(sigma) ;
print 'Wishart Random Matrix';
do i = 1 to n ;
t = normal(repeat(seed,f,p)) ;
x = t*g ;
w = x`*x ;
print w ;
end ;
```

These programs can be easily modified to generate the Beta matrices of either Type 1 or Type 2, as the generation of such matrices essentially amounts to generating the pairs of Wishart matrices with appropriate degrees of freedom and then combining them as per their definitions.

More efficient algorithms, especially for large values of $f - p$ are available in the literature. One such convenient method based on Bartlett's decomposition can be found in Smith and Hocking (1972). Certain other methods are briefly summarized in Kennedy and Gentle (1980, p. 231).