Chapter

1

Basic Concepts for Multivariate Statistics

1.1 Introduction

Data are information. Most crucial scientific, sociological, political, economic, and business decisions are made based on data analyis. Often data are available in abundance, but by themselves they are of little help unless they are summarized and an appropriate interpretation of the summary quantities made. However, such a summary and corresponding interpretation can rarely be made just by looking at the raw data. A careful scientific scrutiny and analysis of these data can usually provide an enormous amount of valuable information. Often such an analysis may not be obtained just by computing simple averages. Admittedly, the more complex the data and their structure, the more involved the data analysis.

The complexity in a data set may exist for a variety of reasons. For example, the data set may contain too many observations that stand out and whose presence in the data cannot be justified by any simple explanation. Such observations are often viewed as influential observations or outliers. Deciding which observation is or is not an influential one is a difficult problem. For a brief review of some graphical and formal approaches to this problem, see Khattree and Naik (1999). A good, detailed discussion of these topics can be found in Belsley, Kuh and Welsch (1980), Belsley (1991), Cook and Weisberg (1982), and Chatterjee and Hadi (1988).

Another situation in which a simple analysis based on averages alone may not suffice occurs when the data on some of the variables are correlated or when there is a trend present in the data. Such a situation often arises when data were collected over time. For example, when the data are collected on a single patient or a group of patients under a given treatment, we are rarely interested in knowing the average response over time. What we are interested in is observing any changes in the values, that is, in observing any patterns or trends.

Many times, data are collected on a number of units, and on each unit not just one, but many variables are measured. For example, in a psychological experiment, many tests are used, and each individual is subjected to all these tests. Since these are measurements on the same unit (an individual), these measurements (or variables) are correlated and, while summarizing the data on all these variables, this set of correlations (or some equivalent quantity) should be an integral part of this summary. Further, when many variables exist, in

order to obtain more definite and more easily comprehensible information, this correlation summary (and its structure) should be subjected to further analysis. There are many other possible ways in which a data set can be quite complex for analysis.

However, it is the last situation that is of interest to us in this book. Specifically, we may have n individual units and on each unit we have observed (same) p different characteristics (variables), say $x_1, x_2, \ldots, x_p$. Then these data can be presented as an n by p matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}.$$

Of course, the measurements in the i^{th} row, namely, $x_{i1}, \ldots, x_{ip}$, which are the measurements on the same unit, are correlated. If we arrange them in a column vector $\mathbf{x}_i$ defined as

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix},$$

then $\mathbf{x}_i$ can be viewed as a multivariate observation. Thus, the n rows of matrix $\mathbf{X}$ correspond to n multivariate observations (written as rows within this matrix), and the measurements within each $\mathbf{x}_i$ are usually correlated. There may or may not be a correlation between columns $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Usually, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are assumed to be uncorrelated (or statistically independent as a stronger assumption) but this may not always be so. For example, if $\mathbf{x}_i$, $i = 1, \ldots, n$ contains measurements on the height and weight of the i^{th} brother in a family with n brothers, then it is reasonable to assume that some kind of correlation may exist between the rows of $\mathbf{X}$ as well.

For much of what is considered in this book, we will not concern ourselves with the scenario in which rows of the data matrix $\mathbf{X}$ are also correlated. In other words, when rows of $\mathbf{X}$ constitute a sample, such a sample will be assumed to be statistically independent. However, before we elaborate on this, we should briefly comment on sampling issues.

1.2 Population Versus Sample

As we pointed out, the rows in the n by p data matrix $\mathbf{X}$ are viewed as multivariate observations on n units. If the set of these n units constitutes the entire (finite) set of all possible units, then we have data available on the entire reference population. An example of such a situation is the data collected on all cities in the United States that have a population of 1,000,000 or more, and on three variables, namely, cost-of-living, average annual salary, and the quality of health care facilities. Since each U.S. city that qualifies for the definition is included, any summary of these data will be the *true* summary of the population.

However, more often than not, the data are obtained through a survey in which, on each of the units, all p characteristics are measured. Such a situation represents a multivariate sample. A sample (adequately or poorly) represents the underlying population from which it is taken. As the population is now represented through only a few units taken from it, any summary derived from it merely represents the *true* population summary in the sense that we hope that, generally, it will be close to the true summary, although no assurance about an exact match between the two can be given.

How can we measure and ensure that the summary from a sample is a good representative of the population summary? To quantify it, some kinds of indexes based on probabilis-

tic ideas seem appropriate. That requires one to build some kind of probabilistic structure over these units. This is done by artificially and intentionally introducing the probabilistic structure into the sampling scheme. Of course, since we want to ensure that the sample is a good representative of the population, the probabilistic structure should be such that it treats all the population units in an equally fair way. Thus, we require that the sampling is done in such a way that each unit of (finite or infinite) population has an equal chance of being included in the sample. This requirement can be met by a simple random sampling with or without replacement. It may be pointed out that in the case of a finite population and sampling without replacement, observations are *not* independent, although the strength of dependence diminishes as the sample size increases.

Although a probabilistic structure is introduced over different units through random sampling, the same cannot be done for the p different measurements, as there is neither a reference population nor do all p measurements (such as weight, height, etc.) necessarily represent the same thing. However, there is possibly some inherent dependence between these measurements, and this dependence is often assumed and modeled as some joint probability distribution. Thus, we view each row of $\mathbf{X}$ as a multivariate observation from some p-dimensional population that is represented by some p-dimensional multivariate distribution. Thus, the rows of $\mathbf{X}$ often represent a random sample from a p-dimensional population. In much multivariate analysis work, this population is assumed to be infinite and quite frequently it is assumed to have a multivariate normal distribution. We will briefly discuss the multivariate normal distribution and its properties in Section 1.6.

1.3 Elementary Tools for Understanding Multivariate Data

To understand a large data set on several mutually dependent variables, we must somehow summarize it. For univariate data, when there is only one variable under consideration, these are usually summarized by the (population or sample) mean, variance, skewness, and kurtosis. These are the basic quantities used for data description. For multivariate data, their counterparts are defined in a similar way. However, the description is greatly simplified if matrix notations are used. Some of the matrix terminology used here is defined later in Section 1.5.

Let $\mathbf{x}$ be the p by 1 random vector corresponding to the multivariate population under consideration. If we let

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix},$$

then each x_i is a random variable, and we assume that $x_1, \ldots, x_p$ are possibly dependent. With $E(\cdot)$ representing the mathematical expectation (interpreted as the long-run average), let $\mu_i = E(x_i)$, and let $\sigma_{ii} = var(x_i)$ be the population variance. Further, let the population covariance between x_i and x_j be $\sigma_{ij} = cov(x_i, x_j)$. Then we define the *population mean vector* $E(\mathbf{x})$ as the vector of term by term expectations. That is,

$$E(\mathbf{x}) = \begin{bmatrix} E(x_1) \\ \vdots \\ E(x_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} = \boldsymbol{\mu} \text{ (say)}.$$

Additionally, the concept of population variance is generalized to the matrix with all the population variances and covariances placed appropriately within a variance-covariance matrix. Specifically, if we denote the variance-covariance matrix of $\mathbf{x}$ by $D(\mathbf{x})$, then

$$D(\mathbf{x}) = \begin{bmatrix} var(x_1) & cov(x_1, x_2) & \dots & cov(x_1, x_p) \\ cov(x_2, x_1) & var(x_2) & \dots & cov(x_2, x_p) \\ \vdots & & & \vdots \\ cov(x_p, x_1) & cov(x_p, x_2) & \dots & var(x_p) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix} = (\sigma_{ij}) = \mathbf{\Sigma} \text{ (say).}$$

That is, with the understanding that $cov(x_i, x_i) = var(x_i) = \sigma_{ii}$, the term $cov(x_i, x_j)$ appears as the $(i, j)^{th}$ entry in matrix $\mathbf{\Sigma}$. Thus, the variance of the i^{th} variable appears at the i^{th} diagonal place and all covariances are appropriately placed at the nondiagonal places. Since $cov(x_i, x_j) = cov(x_j, x_i)$, we have $\sigma_{ij} = \sigma_{ji}$ for all i, j. Thus, the matrix $D(\mathbf{x}) = \mathbf{\Sigma}$ is symmetric. The other alternative notations for $D(\mathbf{x})$ are $cov(\mathbf{x})$ and $var(\mathbf{x})$, and it is often also referred to as the dispersion matrix, the variance-covariance matrix, or simply the covariance matrix. We will use the three terms interchangeably.

The quantity $tr(\mathbf{\Sigma})$ (read as trace of $\mathbf{\Sigma}$)$= \sum_{i=1}^{p} \sigma_{ii}$ is called the *total variance* and $|\mathbf{\Sigma}|$ (the determinant of $\mathbf{\Sigma}$) is referred to as the *generalized variance.* The two are often taken as the overall measures of variability of the random vector $\mathbf{x}$. However, sometimes their use can be misleading. Specifically, the total variance $tr(\mathbf{\Sigma})$ completely ignores the nondiagonal terms of $\mathbf{\Sigma}$ that represent the covariances. At the same time, two very different matrices may yield the same value of the generalized variance.

As there exists dependence between $x_1, \dots, x_p$, it is also meaningful to at least measure the degree of linear dependence. It is often measured using the correlations. Specifically, let

$$\rho_{ij} = \frac{cov(x_i, x_j)}{\sqrt{var(x_i)\, var(x_j)}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\, \sigma_{jj}}}$$

be the Pearson's population correlation coefficient between x_i and x_j. Then we define the population correlation matrix as

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

As was the case for $\mathbf{\Sigma}$, $\boldsymbol{\rho}$ is also symmetric. Further, $\boldsymbol{\rho}$ can be expressed in terms of $\mathbf{\Sigma}$ as

$$\boldsymbol{\rho} = [\text{diag}(\mathbf{\Sigma})]^{-\frac{1}{2}}\, \mathbf{\Sigma}\, [\text{diag}(\mathbf{\Sigma})]^{-\frac{1}{2}},$$

where diag($\mathbf{\Sigma}$) is the diagonal matrix obtained by retaining the diagonal elements of $\mathbf{\Sigma}$ and by replacing all the nondiagonal elements by zero. Further, the square root of matrix $\mathbf{A}$ denoted by $\mathbf{A}^{\frac{1}{2}}$ is a matrix satisfying $\mathbf{A} = \mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}}$. It is defined in Section 1.5. Also, $\mathbf{A}^{-\frac{1}{2}}$ represents the inverse of matrix $\mathbf{A}^{\frac{1}{2}}$.

It may be mentioned that the variance-covariance and the correlation matrices are always nonnegative definite (See Section 1.5 for a discussion). For most of the discussion in this book, these matrices, however, will be assumed to be positive definite. In view of this assumption, these matrices will also admit their respective inverses.

How do we generalize (and measure) the skewness and kurtosis for a multivariate population? Mardia (1970) defines these measures as

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

where $\mathbf{x}$ and $\mathbf{y}$ are independent but have the same distribution and

$$\text{multivariate kurtosis: } \beta_{2,p} = E\left[(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^2.$$

For the univariate case, that is when $p = 1$, $\beta_{1,p}$ reduces to the *square* of the coefficient of skewness, and $\beta_{2,p}$ reduces to the coefficient of kurtosis.

The quantities $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\rho}$, $\beta_{1,p}$ and $\beta_{2,p}$ provide a basic summary of a multivariate population. What about the sample counterparts of these quantities? When we have a p-variate random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of size n, then with the n by p data matrix $\mathbf{X}$ defined as

$$\mathbf{X}_{n\times p} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix},$$

we define,

$$\text{sample mean vector: } \overline{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i = n^{-1}\mathbf{X}'\mathbf{1}_n,$$

$$\begin{aligned}\text{sample variance-covariance matrix: } \mathbf{S} &= (n-1)^{-1}\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})' \\ &= (n-1)^{-1}\left\{\sum_{i=1}^{n}\mathbf{x}_i\,\mathbf{x}_i' - n\,\overline{\mathbf{x}}\,\overline{\mathbf{x}}'\right\} \\ &= (n-1)^{-1}\left\{\mathbf{X}'(\mathbf{I}-n^{-1}\mathbf{1}_n\mathbf{1}_n')\mathbf{X}\right\} \\ &= (n-1)^{-1}\left\{\mathbf{X}'\mathbf{X}-n^{-1}\mathbf{X}'\mathbf{1}_n\mathbf{1}_n'\mathbf{X}\right\} \\ &= (n-1)^{-1}\left\{\mathbf{X}'\mathbf{X}-n\overline{\mathbf{x}}\,\overline{\mathbf{x}}'\right\}.\end{aligned}$$

It may be mentioned that often, instead of the dividing factor of $(n-1)$ in the above expressions, a dividing factor of n is used. Such a sample variance-covariance matrix is denoted by $\mathbf{S}_n$. We also have

$$\begin{aligned}\text{sample correlation matrix: } \hat{\boldsymbol{\rho}} &= \left[\text{diag}(\mathbf{S})\right]^{-\frac{1}{2}}\mathbf{S}\left[\text{diag}(\mathbf{S})\right]^{-\frac{1}{2}} \\ &= \left[\text{diag}(\mathbf{S}_n)\right]^{-\frac{1}{2}}\mathbf{S}_n\left[\text{diag}(\mathbf{S}_n)\right]^{-\frac{1}{2}},\end{aligned}$$

$$\text{sample multivariate skewness: } \hat{\beta}_{1,p} = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}g_{ij}^3,$$

and

$$\text{sample multivariate kurtosis: } \hat{\beta}_{2,p} = n^{-1}\sum_{i=1}^{n}g_{ii}^2.$$

In the above expressions, $\mathbf{1}_n$ denotes an n by 1 vector with all entries 1, $\mathbf{I}_n$ is an n by n identity matrix, and $g_{ij}, i, j = 1, \ldots p$, are defined by $g_{ij} = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{S}_n^{-1}(\mathbf{x}_j - \overline{\mathbf{x}})$. See Khattree and Naik (1999) for details and computational schemes to compute these quantities. In fact, multivariate skewness and multivariate kurtosis are computed later in Chapter 5, Section 5.2 to test the multivariate normality assumption on data. Correlation matrices also play a central role in principal components analysis (Chapter 2, Section 2.2).

1.4 Data Reduction, Description, and Estimation

In the previous section, we presented some of the basic population summary quantities and their sample counterparts commonly referred to as *descriptive statistics*. The basic idea was to summarize the population or sample data through smaller sized matrices or simply numbers. All the quantities (except correlation) defined there were straightforward generalizations of their univariate counterparts. However, the multivariate data do have some of their own unique features and needs, which do not exist in the univariate situation. Even though the idea is still the same, namely that of summarizing or describing the data, such situations call for certain unique ways of handling these, and these unique techniques form the main theme of this book. These can best be described by a few examples.

a. Based on a number of measurements such as average housing prices, cost of living, health care facilities, crime rate, etc., we would like to describe which cities in the country are most livable and also try to observe any unique similarities or differences among cities. There are several variables to be measured, and it is unlikely that attempts to order cities with respect to any one variable will result in the same ordering if another variable were used. For example, a city with a low crime rate (a desirable feature) may have a high cost of living (an undesirable feature), and thus these variables often tend to offset each other. How do we decide which cities are the best to live in? The problem here is that of data reduction. However, this problem can neither be described as that of variable selection (there is no dependent variable and no model) nor can it be viewed as a prediction problem. It is more a problem of attempting to detect and understand the unique features that the data set may contain and then to interpret them. This requires some meaningful approach for data description. The possible analyses for such a data set are principal component analysis (Chapter 2) and cluster analysis (Chapter 6).

b. As another example, suppose we have a set of independent variables which in turn have effects on a large number of dependent variables. Such a situation is quite common in the chemical industry and in economic data, where the two sets can be clearly defined as those containing input and output variables. We are not interested in individual variables, but we want to come up with a few new variables in each group. These may themselves be functions of all variables in the respective groups, so that each new variable from one group can be paired with another new variable in the other group in some meaningful sense, with the hope that these newly defined variables can be appropriately interpreted in the context. We must emphasize that analysis is not being done with any specific purpose of proving or disproving some claims. It is only an attempt to understand the data. As the information is presented in terms of new variables, which are fewer in number, it is easier to observe any striking features or associations in this latter situation. Such problems can be handled using the techniques of canonical correlation (Chapter 3) and in case of qualitative data, using correspondence analysis (Chapter 7).

c. An automobile company wants to know what determines the customer's preference for various cars. A sample of 100 randomly selected individuals were asked to give a score between 1 (low) and 10 (high) on six variables, namely, price, reliability, status symbol related to car, gas mileage, safety in an accident, and average miles driven per week. What kind of analysis can be made for these data? With the assumptions that there are some underlying hypothetical and unobservable variables on which the scores of these six observable variables depend, a natural inquiry would be to identify these hypothetical variables. Intuitively, safety consciousness and economic status of the individual may be two (perhaps of several others) traits that may influence the scores on some of these six observable variables. Thus, some or all of the observed variables can be written as a function of, say, these two unobservable traits. A question in reverse is this: can

we quantify the unobservable traits as functions of the observable ones? Such a query can be usually answered by factor analysis techniques (Chapter 4). Note, however, that the analysis provides only the functions and, their interpretations as some meaningful unobservable trait, is left to the analyst. Nonetheless, it is again a problem of data reduction and description in that many measurements are reduced to only a few traits with the objective of providing an appropriate description of the data.

As is clear from these examples, many multivariate problems involve data reduction, description and, in the process of doing so, estimation. These issues form the focus of the next six chapters. As a general theme, most of the situations either require some matrix decomposition and transformations or use a distance-based approach. Distributional assumptions such as multivariate normality are also helpful (usually but not always, in assessing the quality of estimation) but not crucial. With that in mind in the next section we provide a brief review of some important concepts from matrix theory. A review of multivariate normality is presented in Section 1.6.

1.5 Concepts from Matrix Algebra

This section is meant only as a brief review of concepts from matrix algebra. An excellent account of results on matrices with a statistical viewpoint can be found in the recent books by Schott (1996), Harville (1997) and Rao and Rao (1998). We will assume that the reader is already familiar with the addition, multiplication, and transposition of matrices. Also the working knowledge of other elementary concepts such as linear independence of vectors is assumed.

In SAS, matrix computations can be performed using the IML procedure. The first statement is

```
proc iml;
```

Matrix additions, subtractions, and multiplications are performed using the $+$, $-$, and $*$ notations. Thus, if the sum of matrices $\mathbf{A}_1$, $\mathbf{A}_2$ is to be multiplied by the difference of $\mathbf{A}_3$ and $\mathbf{A}_4$, then the final matrix, say $\mathbf{B}$, will be computed using the program

```
proc iml;
b = (a1 + a2) * (a3 - a4);
```

1.5.1 Transpose of a Matrix

For an m by n matrix $\mathbf{A}$, the transpose of $\mathbf{A}$ is obtained by interchanging its rows and columns. It is denoted by $\mathbf{A}'$. Naturally, $\mathbf{A}'$ is of order n by m. For example, if

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 4 \\ 7 & 0 & 1 \end{bmatrix}$$

then

$$\mathbf{A}' = \begin{bmatrix} 1 & 7 \\ 3 & 0 \\ 4 & 1 \end{bmatrix}.$$

Also for two matrices $\mathbf{A}$ and $\mathbf{B}$ of order m by n and n by r we have, $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.

In PROC IML, the leading quote (‘) is used for transposition. Since some keyboards may not support this key, an alternative way to obtain $\mathbf{A}'$ is to use the function T. Specifically, $\mathbf{B} = \mathbf{A}'$ is obtained as either

```
b = a‘;
```

or as

```
b = t(a);
```

1.5.2 Symmetric Matrices

An n by n matrix $\mathbf{A}$ is said to be symmetric if $\mathbf{A}' = \mathbf{A}$. For example,

$$\mathbf{A} = \begin{bmatrix} 7 & 8 & -3 \\ 8 & 0 & 1 \\ -3 & 1 & 9 \end{bmatrix}$$

is symmetric. Clearly, if a_{ij} is the $(i, j)^{th}$ element in matrix $\mathbf{A}$, then for a symmetric matrix $a_{ij} = a_{ji}$ for all i, j.

1.5.3 Diagonal Matrices

An n by n matrix $\mathbf{A}$ is diagonal if all its nondiagonal entries are zeros. A diagonal matrix is trivially symmetric. For example, $\mathbf{A} = \left[\begin{smallmatrix} 3 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right]$ is a diagonal matrix.

We will often use the notation diag($\mathbf{A}$), which stands for a matrix that retains only the diagonal entries of $\mathbf{A}$ and replaces all nondiagonal entries with zeros. Thus, for

$$\mathbf{A} = \begin{bmatrix} 3 & 4 & 5 \\ 1 & 8 & 2 \\ 4 & -1 & 0 \end{bmatrix},$$

the diag($\mathbf{A}$) will be

$$\text{diag}(\mathbf{A}) = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

In PROC IML, the function DIAG(B) requires $\mathbf{B}$ to be a vector or a square matrix. Thus, if $\mathbf{A}$ is an n by n matrix, and we want the n by n matrix $\mathbf{D} = \text{diag}(\mathbf{A})$, then the appropriate SAS statement is

```
d = diag(a);
```

1.5.4 Some Special Matrices

Here are some examples:

- An n by n diagonal matrix with all diagonal entries equal to 1 is called an *identity matrix*. It is denoted by $\mathbf{I}_n$ or simply by $\mathbf{I}$ if there is no confusion.
- An n by 1 column vector with all entries equal to 1 is denoted by $\mathbf{1}_n$ or simply by $\mathbf{1}$.
- An m by n matrix with all elements as zero is called a zero matrix. It is denoted by $\mathbf{0}_{m,n}$ or simply by $\mathbf{0}$.

In PROC IML, the respective functions are $I(n)$, $J(n, 1, 1)$, and $J(m, n, 0)$.

1.5.5 Triangular Matrices

An n by n matrix $\mathbf{A}$ is said to be *upper triangular* if all entries below the main diagonal are zero. The lower triangular matrix is similarly defined as one that has all entries above main diagonal zero. For example,

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 1 & 9 \\ 0 & 3 & 4 \\ 0 & 0 & 4 \end{bmatrix} \text{ and } \mathbf{A}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 3 & 9 \end{bmatrix}$$

are respectively upper and lower triangular.

1.5.6 Linear Independence

A set of nonzero column (or row) vectors is said to be *linearly independent* if none of them can be expressed as a linear combination of some or all of the remaining vectors. If this does not happen, then this set will be called *linearly dependent*. A set containing a zero vector will always be viewed as linearly dependent.

Given a linearly dependent set of vectors, if we discard the zero vector and we continue to discard one by one the vectors that can be expressed as a linear combination of the remaining undiscarded vectors, then we will either end with a subset that is linearly independent or with an empty set. The number of vectors that finally remain is an important concept and is formally defined for a matrix (when viewed as a set of columns or rows) in the next subsection.

1.5.7 Rank of a Matrix

The rank of a matrix $\mathbf{A}$, denoted by $R(\mathbf{A})$, is defined as the number of linearly independent rows (or columns) in the matrix. Since we can either work with only rows or with only columns, it is obvious that $R(\mathbf{A}) = R(\mathbf{A}')$. It can also be established that $R(\mathbf{AB}) \leq \min(R(\mathbf{A}), (R(\mathbf{B}))$. Further, $R(\mathbf{A}'\mathbf{A}) = R(\mathbf{A})$.

1.5.8 Nonsingular and Singular Matrices

An n by n matrix $\mathbf{A}$ is said to be nonsingular if all its rows (or columns) are linearly independent. In other words, $\mathbf{A}$ is nonsingular if $R(\mathbf{A}) = n$. If one or more rows (or columns) of $\mathbf{A}$ can be written as linear combinations of some or all of the remaining rows (or columns) of $\mathbf{A}$, then there exists some linear dependence among the rows (or columns) of $\mathbf{A}$. Consequently, $\mathbf{A}$ is said to be singular in this case. For example,

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 9 & 4 \end{bmatrix}$$

is nonsingular, as neither of the two rows can be linearly expressed in terms of the other. However,

$$\mathbf{B} = \begin{bmatrix} 1 & 3 & 4 \\ 9 & 4 & 3 \\ 11 & 10 & 11 \end{bmatrix}$$

is singular since Row 3 = 2× Row 1 + Row 2, which indicates that the third row (or any other row, for that matter) can be expressed as the linear combination of the other two.

1.5.9 Inverse of a Square Matrix

An n by n matrix $\mathbf{A}$ admits an inverse if there exists a matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$. The matrix $\mathbf{B}$ is called the inverse of $\mathbf{A}$ and is denoted by $\mathbf{A}^{-1}$. For example, for

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 9 & 4 \end{bmatrix},$$

the $\mathbf{A}^{-1}$ is given by

$$\mathbf{A}^{-1} = \begin{bmatrix} -\frac{4}{23} & \frac{3}{23} \\ \frac{9}{23} & -\frac{1}{23} \end{bmatrix} = \begin{bmatrix} -0.1739 & 0.1304 \\ 0.3913 & -0.0435 \end{bmatrix}.$$

It is obvious that the inverse of $\mathbf{A}^{-1}$, namely, $(\mathbf{A}^{-1})^{-1}$ is $\mathbf{A}$. The inverse is defined only for n by n matrices, that is, when the number of rows and the number of columns are equal. Even for such matrices, it exists if and only if $\mathbf{A}$ is nonsingular. Thus, no inverse exists for matrices that are singular or for which the number of rows is not equal to the number of columns. For such matrices, a weaker concept, known as a *generalized inverse* or simply a g-inverse can be defined. Whenever an inverse of a given matrix exists, it is unique.

In PROC IML, the inverse for a square matrix $\mathbf{A}$ can be computed by the statement

```
a_inv = inv (a);
```

Thus, A_INV is the desired inverse. It is unique.

If two matrices $\mathbf{A}$ and $\mathbf{B}$ are both of order n by n and are nonsingular, then $(\mathbf{AB})^{-1}$ and $(\mathbf{BA})^{-1}$ both exist. However, they are not equal. Specifically,

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

and

$$(\mathbf{BA})^{-1} = \mathbf{A}^{-1}\mathbf{B}^{-1}.$$

Since the product of matrices is not commutative, the right-hand sides of the above two expressions are not equal. This makes it clear why $(\mathbf{AB})^{-1}$ and $(\mathbf{BA})^{-1}$ are not the same.

1.5.10 Generalized Inverses

For an m by n matrix, $\mathbf{B}$, a generalized inverse or simply a g-inverse, say $\mathbf{G}$, is an n by m matrix such that

$$\mathbf{BGB} = \mathbf{B}.$$

In general, the g-inverse always exists. However, it is not necessarily unique. The g-inverse is unique only for nonsingular matrices and in that case, it is the same as the inverse. A g-inverse of $\mathbf{B}$ is denoted by $\mathbf{B}^{-}$.

The matrix

$$\mathbf{B} = \begin{bmatrix} 1 & 3 & 4 \\ 9 & 4 & 3 \\ 11 & 10 & 11 \end{bmatrix}$$

was earlier found to be singular. A g-inverse of $\mathbf{B}$ is given by,

$$\mathbf{B}^{-} = \begin{bmatrix} -\frac{4}{23} & \frac{3}{23} & 0 \\ \frac{9}{23} & -\frac{1}{23} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Of course, the above choice of $\mathbf{B}^{-}$ is not unique. In PROC IML, a g-inverse for matrix $\mathbf{B}$ can be computed by the statement

```
b_ginv = ginv(b);
```

Thus, B_GINV is a g-inverse. The specific generalized inverse that SAS computes using the GINV function is the Moore-Penrose g-inverse (which, in fact, has been made unique by additional restrictions (Rao, 1973)).

1.5.11 A System of Linear Equations

Consider a system of n consistent equations in m unknowns, $x_1, \ldots, x_m$, (that is, a system in which no subset of equations violates any of the remaining equations)

$$a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m = b_1$$

$$\vdots$$

$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m = b_n.$$

With $\mathbf{A} = \begin{bmatrix} a_{11} & \ldots & a_{1m} \\ \vdots & & \\ a_{n1} & \ldots & a_{nm} \end{bmatrix}$, $\mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$, the above system can be written as,

$$\mathbf{Ax} = \mathbf{b}.$$

If $m = n$ and if the matrix $\mathbf{A}$ is nonsingular, then the solution $\mathbf{x}$ is given by $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. If $m = n$ and if $\mathbf{A}$ is singular (in which case some equations may be redundant as they are implied by other equations) or if $m \neq n$, then a solution $\mathbf{x}$ is obtained as $\mathbf{x} = \mathbf{A}^{-}\mathbf{b}$, where $\mathbf{A}^{-}$ is a g-inverse of $\mathbf{A}$. Since the g-inverses are not unique, unless $\mathbf{A}$ is nonsingular, in this case there is no unique solution to the above system of linear equations. The reason for this is that changing the choice of g-inverse of $\mathbf{A}$ in the equation above yields another new solution.

In PROC IML, the solution $\mathbf{x}$ can be obtained by using the SOLVE function. Specifically, when $\mathbf{A}$ is nonsingular, we use

```
x=solve(a,b);
```

Alternatively, one can just use the INV function and get the solution by

```
x=inv(a)*b;
```

When $\mathbf{A}$ is singular, there are infinitely many solutions, all of which can be collectively expressed as $\mathbf{x} = \mathbf{A}^{-}\mathbf{b} + (\mathbf{I} - \mathbf{A}^{-}\mathbf{A})\mathbf{z}$, where $\mathbf{z}$ is any arbitrary vector and $\mathbf{A}^{-}$ is a g-inverse of $\mathbf{A}$. Of special interest is the case in which we have a system of consistent linear equations $\mathbf{Ax} = \mathbf{0}$ when $n < m$. In this case, although there are infinitely many solutions, a finite orthonormal (to be defined later) set of solutions can be obtained as a matrix $\mathbf{X}$ by using

```
x=homogen(a);
```

The columns of matrix $\mathbf{X}$ are the orthonormal solutions. The order of $\mathbf{X}$ is determined by the rank of the matrix $\mathbf{A}$.

1.5.12 Euclidean Norm of a Vector

For an n by 1 vector $\mathbf{a}$, the norm (or length) of $\mathbf{a}$ is defined as $\sqrt{\mathbf{a}'\mathbf{a}}$. Clearly, $\mathbf{b}$ defined as $\mathbf{b} = \mathbf{a}/\sqrt{\mathbf{a}'\mathbf{a}}$ has norm 1. In this case, $\mathbf{b}$ is called the normalized version of $\mathbf{a}$.

1.5.13 Euclidean Distance between Two Vectors

Visualizing the n by 1 vectors $\mathbf{a}$ and $\mathbf{b}$ as points in an n-dimensional space, we can define the distance between $\mathbf{a}$ and $\mathbf{b}$ as the norm of the vector $(\mathbf{a}-\mathbf{b})$. That is, the distance $d(\mathbf{a}, \mathbf{b})$ is defined as

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{(\mathbf{a}-\mathbf{b})'(\mathbf{a}-\mathbf{b})}$$

$$= \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2},$$

where a_i and b_i, respectively, are the i^{th} entries of vectors $\mathbf{a}$ and $\mathbf{b}$.

The Euclidean distance is the distance between the points as our eyes see it. However, sometimes distance can be defined after assigning some weights through a positive definite matrix (to be defined later). Specifically, the weighted distance with weight matrix $\mathbf{A}$ is defined as

$$d_{\mathbf{A}}(\mathbf{a}, \mathbf{b}) = \sqrt{(\mathbf{a}-\mathbf{b})'\mathbf{A}(\mathbf{a}-\mathbf{b})},$$

where $\mathbf{A}$ is positive definite. Clearly $d_{\mathbf{I}_n}(\mathbf{a}, \mathbf{b}) = d(\mathbf{a}, \mathbf{b})$. One common weighted distance that we encounter in multivariate analysis is the Mahalanobis distance (Rao, 1973).

In general, a distance function, say, $\delta(a, b)$ can be defined in many other ways. However, a distance function must satisfy the following conditions:

- $\delta(\mathbf{a}, \mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$.
- $\delta(\mathbf{a}, \mathbf{b}) = \delta(\mathbf{b}, \mathbf{a})$.
- $\delta(\mathbf{a}, \mathbf{b}) \geq 0$.
- $\delta(\mathbf{a}, \mathbf{c}) \leq \delta(\mathbf{a}, \mathbf{b}) + \delta(\mathbf{b}, \mathbf{c})$.

Clearly, $d(\mathbf{a}, \mathbf{b})$ and $d_{\mathbf{A}}(\mathbf{a}, \mathbf{b})$ satisfy all of the above conditions. It may be remarked that often in statistics, the *squared distances* are also referred to as distances. This is especially more frequent in case of certain cluster analyses. In this context, we may remark that the distance functions are often used as the measures of dissimilarity between the objects or units. However, various other dissimilarity indexes are also often applied. Many of these are not distance functions in that they do not satisfy all of the above conditions.

1.5.14 Orthogonal Vectors and Matrices

Two n by 1 vectors $\mathbf{a}$ and $\mathbf{b}$ are said to be *orthogonal* to each other if $\mathbf{a}'\mathbf{b} = 0$. Additionally, if $\mathbf{a}$ and $\mathbf{b}$ are normalized (i.e., $\mathbf{a}'\mathbf{a} = 1 = \mathbf{b}'\mathbf{b}$), then they are called *orthonormal*. For example,

$$\mathbf{a} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

are orthogonal to each other. Their normalized versions, $\mathbf{a}/\sqrt{3}$ and $\mathbf{b}/\sqrt{2}$ are orthonormal to each other.

An n by n matrix $\mathbf{A}$ is said to be an orthogonal matrix if

$$\mathbf{A}'\mathbf{A} = \mathbf{A}\mathbf{A}' = \mathbf{I}_n.$$

This necessarily is equivalent to saying that all rows (or columns) of $\mathbf{A}$ are orthonormal to one another. Since for an orthogonal matrix, $\mathbf{A}'\mathbf{A} = \mathbf{A}\mathbf{A}' = \mathbf{I}_n$, $\mathbf{A}'$ also acts as the inverse of $\mathbf{A}$. Hence, $\mathbf{A}$ is nonsingular as well. Trivially, $\mathbf{A}'$ is also orthogonal.

Let $m < n$ and let $\mathbf{A}$ be of order n by m, such that all m columns of $\mathbf{A}$ are orthonormal to each other. In that case,

$$\mathbf{A}'\mathbf{A} = \mathbf{I}_m,$$

but no such claim can be made for $\mathbf{A}\mathbf{A}'$. In this case the matrix $\mathbf{A}$ is referred to as a *sub-orthogonal* matrix.

The matrix

$$\mathbf{A} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & \frac{2}{\sqrt{6}} \end{bmatrix}$$

is orthogonal. However,

$$\mathbf{A}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \\ \frac{-1}{\sqrt{2}} & \frac{-1}{\sqrt{6}} \\ 0 & \frac{2}{\sqrt{6}} \end{bmatrix}$$

is suborthogonal because only $\mathbf{A}_1'\mathbf{A}_1 = \mathbf{I}_2$, but $\mathbf{A}_1\mathbf{A}_1'$ is not equal to $\mathbf{I}_3$.

There are many orthogonal matrices, and using PROC IML a variety of suborthogonal matrices can be generated. The premultiplication of a general matrix by an orthogonal matrix amounts to the rotation of the axes. This frequently arises in multivariate contexts such as principal components analysis and factor analysis.

1.5.15 Eigenvalues and Eigenvectors

Let $\mathbf{A}$ be an n by n matrix. The pairs $(\lambda_1, \mathbf{x}_1), \ldots, (\lambda_n, \mathbf{x}_n)$ are said to be pairs of the eigenvalues and corresponding eigenvectors if all $(\lambda_i, \mathbf{x}_i)$ satisfy the matrix equation

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}.$$

If $\mathbf{x}_i$ satisfies the above, then a constant multiple of $\mathbf{x}_i$ also satisfies the above. Thus, often we work with the eigenvector $\mathbf{x}_i$ that has norm 1. In general, λ_i as well as elements of $\mathbf{x}_i$ may be complex valued. However, if $\mathbf{A}$ is symmetric, all eigenvalues are necessarily real valued and one can find eigenvectors that are all real valued. If any eigenvalue is zero, then it implies, and is implied by, the fact that the matrix $\mathbf{A}$ is singular.

If $\mathbf{A}$ is nonsingular, then $\mathbf{A}^{-1}$ exists. The eigenvalues of $\mathbf{A}^{-1}$ are $\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}$, and the corresponding eigenvectors are the same as those of $\mathbf{A}$.

The eigenvalues may be repeated. If an eigenvalue is repeated r times, then we say that it has multiplicity r. If $\mathbf{A}$ is symmetric, then the eigenvectors corresponding to distinct eigenvalues are all orthonormal (provided they all have norm 1). Further, eigenvectors

corresponding to an eigenvalue with multiplicity r are not necessarily orthonormal, but one can always find a set of r distinct eigenvectors, corresponding to this eigenvalue, which are orthonormal to each other. Putting all these facts together suggests that we can always find a set of n orthonormal eigenvectors for a symmetric matrix. Thus, in terms of these orthonormal eigenvectors, namely, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we have n equations

$$\mathbf{A}\mathbf{x}_1 = \lambda_1 \mathbf{x}_1$$

$$\vdots$$

$$\mathbf{A}\mathbf{x}_n = \lambda_n \mathbf{x}_n.$$

Writing these n equations side by side yields the matrix equation,

$$(\mathbf{A}\mathbf{x}_1 : \cdots : \mathbf{A}\mathbf{x}_n) = (\lambda_1 \mathbf{x}_1 : \cdots : \lambda_n \mathbf{x}_n)$$

or

$$\mathbf{A}(\mathbf{x}_1 : \cdots : \mathbf{x}_n) = (\mathbf{x}_1 : \cdots : \mathbf{x}_n) \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}.$$

Let $\mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_n)$ and $\mathbf{P} = [\mathbf{x}_1 : \cdots : \mathbf{x}_n]$. Clearly, $\mathbf{\Lambda}$ is diagonal and $\mathbf{P}$ is orthogonal, since all $\mathbf{x}_i$ are orthonormal to each other. Thus, we have

$$\mathbf{AP} = \mathbf{P\Lambda}$$

or

$$\mathbf{A} = \mathbf{P\Lambda P}'.$$

The above fact results in an important decomposition of a symmetric matrix, as stated below.

1.5.16 Spectral Decomposition of a Symmetric Matrix

Let $\mathbf{A}$ be an n by n symmetric matrix. Then $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{P\Lambda P}',$$

for some orthogonal matrix $\mathbf{P}$ and a diagonal matrix $\mathbf{\Lambda}$. Of course, the choices of $\mathbf{P}$ and $\mathbf{\Lambda}$ have been indicated above.

Using PROC IML, the eigenvalues and eigenvectors of a symmetric matrix $\mathbf{A}$ can be found by using the call

```
call eigen(lambda, p, a);
```

The eigenvalues and respective eigenvectors are stored in $\mathbf{\Lambda}$ and $\mathbf{P}$. Columns of $\mathbf{P}$ are the eigenvectors. Of course, this also readily provides a choice for the spectral decomposition matrices. However, the spectral decomposition of $\mathbf{A}$ is not unique.

1.5.17 Generalized Eigenvalues and Eigenvectors

Let $\mathbf{A}$ and $\mathbf{B}$ be two n by n symmetric matrices, and let $\mathbf{B}$ be positive definite. Then $(\delta_1, \mathbf{x}_1), (\delta_2, \mathbf{x}_2), \ldots, (\delta_n, \mathbf{x}_n)$ are the pairs of eigenvalues and eigenvectors of $\mathbf{A}$ *with respect to* $\mathbf{B}$ if they all satisfy the generalized eigenequation

$$\mathbf{A}\mathbf{x} = \delta \mathbf{B}\mathbf{x},$$

for all $i = 1, \ldots, n$. With $\mathbf{Q} = (\mathbf{x}_1 : \mathbf{x}_2 : \cdots : \mathbf{x}_n)$, all of the above n equations (with $\mathbf{x} = \mathbf{x}_i$ and $\delta = \delta_i$) can be written as one matrix equation,

$$\mathbf{AQ} = \mathbf{BQ\Delta},$$

where $\mathbf{\Delta} = diag(\delta_1, \ldots, \delta_n)$.

The generalized eigenvalue problems occur naturally in many statistical contexts. One such context is the construction of the canonical discriminant functions discussed in Chapter 5, Section 5.6. Using PROC IML, and given **A** and **B**, the matrices **Q** and **Δ** can be computed by the subroutine call

```
call geneig(d,q,a,b);
```

The vector **d** obtained from the above call contains the eigenvalues of **A** with respect to **B**. Thus, **Δ** is computed as **Δ** = DIAG(**d**). The columns of **Q** are the respective eigenvectors. These eigenvectors are not necessarily orthogonal. It may be remarked that the generalized eigenvalue problem is equivalent to finding the eigenvalues and eigenvectors of a possibly nonsymmetric matrix $\mathbf{B}^{-1}\mathbf{A}$. It is known that these will necessarily be real, even though the particular matrix $\mathbf{B}^{-1}\mathbf{A}$ is possibly asymmetric.

1.5.18 Determinant of a Matrix

For our purpose, we define the determinant of an n by n matrix **A** as the product of all eigenvalues $\lambda_1, \ldots, \lambda_n$ of **A**. Thus, the determinant of **A**, denoted by $|\mathbf{A}|$, is

$$|\mathbf{A}| = \lambda_1 \ldots \lambda_n.$$

Thus, $|\mathbf{A}| = 0$ if and only if at least one eigenvalue is zero, which occurs if and only if **A** is singular. In the IML procedure, the determinant DETER, of a square matrix **A** can be computed by using the statement

```
deter = det(a);
```

1.5.19 The Trace of a Matrix

The trace of an n by n matrix **A** is defined as the sum of all its eigenvalues. Thus, the trace of **A**, denoted by $tr\,(\mathbf{A})$, is

$$tr\,(\mathbf{A}) = \lambda_1 + \cdots + \lambda_n.$$

It turns out (as already mentioned in Section 1.3) that $tr(\mathbf{A})$ is also equal to $a_{11} + \cdots + a_{nn}$, the sum of all diagonal elements of **A**. This equivalence is useful in the conceptual development of the theory for principal components. In PROC IML the trace TR of a square matrix **A**, can be computed by using the TRACE function as follows

```
tr = trace(a);
```

1.5.20 Majorization

Let $\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$ be two n by 1 vectors with $a_1 \geq a_2 \geq \cdots \geq a_n$ and $b_1 \geq b_2 \geq \cdots \geq b_n$. Then **a** is said to be majorized by **b** if

$$
\begin{aligned}
a_1 &\le b_1 \\
a_1 + a_2 &\le b_1 + b_2 \\
&\vdots \\
a_1 + \cdots + a_{n-1} &\le b_1 + \cdots + b_{n-1} \\
a_1 + \cdots + a_n &= b_1 + \cdots + b_n.
\end{aligned}
$$

One important majorization fact about a symmetric matrix is that the vector of all diagonal elements arranged in increasing order is majorized by the vector of all eigenvalues also arranged in increasing order. This result is useful in principal component analysis, and it justifies why the use of a few principal components may be superior to the use of a few individual variables in certain situations.

For two vectors **a** and **b** as defined, we can verify if **a** is majorized by **b** when we use the following SAS/IML code:

```
if all( cusum(a) <= cusum(b)) then major = 'yes';
else major = 'no';
print major;
```

1.5.21 Quadratic Forms

Let $\mathbf{A} = (a_{ij})$ be an n by n matrix and $\mathbf{x}$ be an n by 1 vector of variables. Then

$$
\begin{aligned}
\mathbf{x}'\mathbf{A}\mathbf{x} &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j \\
&= a_{11}x_1^2 + \cdots + a_{nn}x_n^2 \\
&\qquad + (a_{12} + a_{21})x_1x_2 + \cdots + (a_{n-1,n} + a_{n,n-1})x_{n-1}x_n.
\end{aligned}
$$

It is a second degree polynomial in $x_1, \ldots, x_n$, and thus it is referred to as a quadratic form in $\mathbf{x}$.

Clearly, $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{A}'\mathbf{x}$ which, by averaging, is also the same as $\mathbf{x}'\left(\frac{\mathbf{A}+\mathbf{A}'}{2}\right)\mathbf{x}$. Since $\frac{\mathbf{A}+\mathbf{A}'}{2}$ is always symmetric, without any loss of generality, the matrix $\mathbf{A}$ in the above definition of quadratic forms can be taken to be symmetric. In this case, with $\mathbf{A}$ symmetric, a quadratic form can be expanded into any one of the alternative representations:

$$
\begin{aligned}
\mathbf{x}'\mathbf{A}\mathbf{x} &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j \\
&= \sum_{i=1}^{n} a_{ii}x_i^2 + \sum_{\substack{i=1 \\ i\neq j}}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j \\
&= \sum_{i=1}^{n} a_{ii}x_i^2 + 2\sum_{\substack{i=1 \\ i<j}}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j.
\end{aligned}
$$

The equation $\mathbf{x}'\mathbf{A}\mathbf{x} = c$, where c is a constant, represents a quadratic surface in an n-dimensional space. Thus, it may be a paraboloid, a hyperboloid or an ellipsoid (or a hybrid of these). Which it is depends on the elements of matrix $\mathbf{A}$. The latter case of ellipsoid is of special interest in statistics, and it occurs if $\mathbf{A}$ is positive (semi-)definite, which is defined below.

1.5.22 Positive Definite and Semidefinite Matrices

An n by n symmetric matrix $\mathbf{A}$ is said to be *positive definite* if for any vector $\mathbf{x} \neq \mathbf{0}$, the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$. Similarly, it is *positive semidefinite* (also referred to as nonnegative definite) if $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$. Of course any positive definite matrix is also positive semidefinite. When $\mathbf{A}$ is positive definite the equation $\mathbf{x}'\mathbf{A}\mathbf{x} = c$, where c is a constant, represents an ellipsoid.

It is known that for a positive definite matrix, all eigenvalues are positive. The converse is also true. Similarly, for a positive semidefinite matrix, these are nonnegative. Since for a positive definite matrix all eigenvalues are positive, so is the determinant, being the product of these. Thus, the determinant is not equal to zero, and hence $\mathbf{A}$ is necessarily nonsingular. Thus, a positive definite matrix $\mathbf{A}$ always admits an inverse.

If $\mathbf{B}$ is an m by n matrix, then $\mathbf{B}\mathbf{B}'$ and $\mathbf{B}'\mathbf{B}$ are positive semidefinite. If $m < n$ and $R(\mathbf{B}) = m$ then $\mathbf{B}\mathbf{B}'$ is also positive definite. However $\mathbf{B}'\mathbf{B}$ is still positive semidefinite only.

1.5.23 Square Root of a Symmetric Positive Semidefinite Matrix

For a symmetric positive semidefinite matrix $\mathbf{A}$, one can find an upper triangular matrix $\mathbf{U}$ such that

$$\mathbf{A} = \mathbf{U}'\mathbf{U}.$$

This is called the Cholesky decomposition. In PROC IML, the statement

```
u=root(a);
```

performs the Cholesky decomposition. The matrix $\mathbf{U}$ in the above is upper triangular and hence not symmetric. A symmetric square root of $\mathbf{A}$ denoted by $\mathbf{A}^{1/2}$ can also be obtained. Specifically, since $\mathbf{A}$ is symmetric, we must have by its spectral decomposition, $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}' = (\mathbf{P}\mathbf{\Lambda}^{1/2}\mathbf{P}')(\mathbf{P}\mathbf{\Lambda}^{1/2}\mathbf{P}') = \mathbf{A}^{1/2}\mathbf{A}^{1/2}$, where $\mathbf{P}$ is orthogonal and $\mathbf{\Lambda}$ is diagonal. The diagonal matrix $\mathbf{\Lambda}$ contains the eigenvalues of $\mathbf{A}$ in the diagonal places, which are nonnegative since the matrix $\mathbf{A}$ is nonnegative definite. Thus, we take $\mathbf{\Lambda}^{1/2}$ as just a diagonal matrix with diagonal elements as the positive square roots of the corresponding elements of $\mathbf{\Lambda}$. Accordingly, we define $\mathbf{A}^{1/2}$ as $\mathbf{A}^{1/2} = \mathbf{P}\mathbf{\Lambda}^{1/2}\mathbf{P}'$. Thus, $\mathbf{A}^{1/2}$ is also symmetric. However, it may not be unique, since the spectral decomposition of $\mathbf{A}$ is not unique.

The SAS statements that obtain $\mathbf{A}^{1/2}$ are

```
proc iml;
a = {
10 3 9,
3 40 8,
9 8 15};
call eigen(lambda,p,a);
lam_half = root(diag(lambda));
a_half = p*lam_half*p‘;
print a, p, lam_half;
print a_half ;
```

The symmetric square root matrix $\mathbf{A}^{1/2}$ in the above program is denoted by A_HALF. It may be pointed out that $\mathbf{A}^{-1/2}$ may be computed by taking the inverse of $\mathbf{A}^{1/2}$ or by directly computing the symmetric square root of $\mathbf{A}^{-1}$ instead of $\mathbf{A}$ using the program.

1.5.24 Singular Value Decomposition

Any matrix $\mathbf{B}$ of order m by n can be presented as $\mathbf{B} = \mathbf{UQV}'$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal or suborthogonal. If m is larger than n, then $\mathbf{U}$ is suborthogonal (only $\mathbf{UU}'$ is equal to the identity matrix but $\mathbf{U}'\mathbf{U}$ is not) and $\mathbf{V}$ is orthogonal. The matrix $\mathbf{Q}$ is a diagonal matrix of order n by n. If m is smaller than n, then after ignoring the last $n-m$ zero columns of $\mathbf{U}$ this reduced matrix, say $\mathbf{U}_*$, and $\mathbf{V}$ are both orthogonal. If $\mathbf{B}$ is square, then both $\mathbf{U}$ and $\mathbf{V}$ are orthogonal. The diagonal places of matrix $\mathbf{Q}$ contain the singular values of $\mathbf{B}$. Denoting $\mathbf{U}$, $\mathbf{Q}$, and $\mathbf{V}$ by LEFT, MID, and RIGHT, the following IML subroutine call results in their computation

```
call svd(left,mid,right,b);
```

Only the diagonal elements of $\mathbf{Q}$—and not the entire matrix—are printed, and hence MID is a column vector, not a square matrix. Thus, for any further calculations involving $\mathbf{Q}$, it should be specified as DIAG(MID).

The singular value decomposition (SVD) is also written in a form when the left and right side matrices of decomposition are orthogonal and not just suborthogonal. In this case the middle matrix $\mathbf{Q}$ is of order m by n. Specifically, when $m = n$, nothing needs to be done as $\mathbf{U}$ and $\mathbf{V}$ are both orthogonal. When $m > n$, we write $\mathbf{B}$ as

$$\mathbf{B} = \left[\mathbf{U}_{m\times n} : \mathbf{U}c_{m\times(m-n)}\right] \begin{bmatrix} \mathbf{Q}_{n\times n} \\ \cdots \\ \mathbf{0}_{(m-n)\times n} \end{bmatrix} \mathbf{V}'_{n\times n}$$

$$= \mathbf{U}_*\mathbf{Q}_*\mathbf{V}'_*.$$

Here $\mathbf{V}_* = \mathbf{V}$, $\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{n\times n} \\ \cdots \\ \mathbf{0}_{(m-n)\times n} \end{bmatrix}$ and $\mathbf{U}_* = [\mathbf{U} : \mathbf{U}_c]$. The matrix $\mathbf{U}_c$ is suitably chosen such that $\mathbf{U}'_c\mathbf{U} = \mathbf{0}$. It is called the *orthogonal complement* of $\mathbf{U}$, and one choice of $\mathbf{U}_c$ (as it may not be unique) can be obtained by using the function HOMOGEN. Specifically, in PROC IML, we use the statement

```
uc=homogen(t(u));
```

to obtain the matrix $\mathbf{U}_c$.

When $m < n$, the $m \times n$ matrix $\mathbf{U}$ will necessarily have $(n - m)$ zero columns. The matrix $\mathbf{U}_*$ is obtained by eliminating these columns from $\mathbf{U}$, $\mathbf{V}_*$ is the same as $\mathbf{V}$, and $\mathbf{Q}_* = \mathbf{Q}$. Thus, we again have $\mathbf{B} = \mathbf{U}_*\mathbf{Q}_*\mathbf{V}'_*$.

It may be pointed out that the SVD of $\mathbf{B}'$ is equivalent to the SVD of $\mathbf{B}$. Thus, alternatively, the case of $m < n$ can be derived from the case of $m > n$ and vice versa.

1.5.25 Generalized Singular Value Decomposition

In the singular value decomposition of matrix $\mathbf{B}$ defined above, the matrices $\mathbf{U}_*$ and $\mathbf{V}_*$ were orthogonal. That is, we had

$$\mathbf{U}'_*\mathbf{U}_* = \mathbf{U}_*\mathbf{U}'_* = \mathbf{I}_m$$

and

$$\mathbf{V}'_*\mathbf{V}_* = \mathbf{V}_*\mathbf{V}'_* = \mathbf{I}_n.$$

While dropping the subscripts $*$, we may instead require

$$\mathbf{U}'\mathbf{C}\mathbf{U} = \mathbf{I}_m$$

and

$$\mathbf{V}'\mathbf{D}\mathbf{V} = \mathbf{I}_n,$$

where $\mathbf{C}$ and $\mathbf{D}$ are, respectively, m by m and n by n symmetric positive definite matrices. We can still have the decomposition $\mathbf{B} = \mathbf{U}\mathbf{Q}\mathbf{V}'$, where $\mathbf{U}$ and $\mathbf{V}$ satisfy the latter two requirements instead of the former two. This is known as the *generalized singular value decomposition* of matrix $\mathbf{B}$. Such a decomposition is very useful in correspondence analysis.

The generalized singular value decomposition of $\mathbf{B}$ is closely related to the singular value decomposition of $\mathbf{C}^{\frac{1}{2}}\mathbf{B}\mathbf{D}^{\frac{1}{2}}$. In fact, one can be obtained from the other. Thus, to find the generalized singular value decomposition of $\mathbf{B}$, let us call $\mathbf{C}^{\frac{1}{2}}\mathbf{B}\mathbf{D}^{\frac{1}{2}} = \mathbf{B}_*$. We can perform the singular value decomposition of $\mathbf{B}_*$, using the SVD call as shown in the previous subsection. By calling the corresponding orthogonal matrices as $\mathbf{U}_*$ and $\mathbf{V}_*$, respectively, the matrices $\mathbf{U}$ and $\mathbf{V}$, satisfying the requirements $\mathbf{U}'\mathbf{C}\mathbf{U} = \mathbf{I}_m$ and $\mathbf{V}'\mathbf{D}\mathbf{V} = \mathbf{I}_n$, are obtained as

$$\mathbf{U} = \mathbf{C}^{-\frac{1}{2}}\mathbf{U}_*$$

and

$$\mathbf{V} = \mathbf{D}^{-\frac{1}{2}}\mathbf{V}_* .$$

Of course to compute $\mathbf{C}^{\frac{1}{2}}$ and $\mathbf{D}^{\frac{1}{2}}$, the ROOT function can be used.

1.5.26 Kronecker Product

We define the Kronecker product of $\mathbf{C}$ *with* $\mathbf{D}$ (denoted by $\mathbf{C} \otimes \mathbf{D}$) by multiplying every entry of $\mathbf{C}$ by matrix $\mathbf{D}$ and then creating a matrix out of these block matrices. In notations, the Kronecker product is defined as $\mathbf{C} \otimes \mathbf{D} = (c_{ij}\mathbf{D})$. In SAS/IML software, the operator @ does this job. For example, the Kronecker product matrix $KRON\_CD$ is obtained by writing

```
kron_cd = c @ d;
```

With

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 3 & 4 \\ 0 & 4 & 1 & -1 \\ 1 & 1 & -3 & 2 \end{bmatrix},$$

and

$$\mathbf{D} = \begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix},$$

the Kronecker product $\mathbf{C} \otimes \mathbf{D}$ (using SAS syntax $\mathbf{C}@\mathbf{D}$) is equal to

$$\mathbf{C} \otimes \mathbf{D} = \begin{bmatrix} 1 & 0 & 3 & 4 \\ 3 & 0 & 9 & 12 \\ 7 & 0 & 21 & 28 \\ 0 & 4 & 1 & -1 \\ 0 & 12 & 3 & -3 \\ 0 & 28 & 7 & -7 \\ 1 & 1 & -3 & 2 \\ 3 & 3 & -9 & 6 \\ 7 & 7 & -21 & 14 \end{bmatrix}.$$

The next two subsections cover data manipulation and indicate how to create a matrix from a data set and how to convert a data set into a matrix.

1.5.27 Creating a Matrix from a SAS Data Set

Often, after running a SAS program, we may, for further calculations, need to use PROC IML. That may require converting an input or output data set to a matrix. An example follows.

Suppose we have a data set called MYDATA with three variables X1, X2, and X3 and five data points. From that we want to create a matrix called MYMATRIX. To do so, we use the following SAS statements

```
data mydata;
input x1 x2 x3;
lines;
2 4 8
3 9 1
9 4 8
1 1 1
2 7 8
;
proc iml;
use mydata;
read all into mymatrix;
quit;
print mymatrix;
```

If we want a matrix consisting of only a few variables, say in this case x_3 and x_1 (in that specific order) from the data set, then the appropriate READ statement needs to be slightly more specific:

```
read all var {x3 x1} into mymatrix;
```

1.5.28 Creating a SAS Data Set from a Matrix

Conversely, we can create a SAS data set out of a matrix. An example is presented here. Suppose we have a 5 by 3 matrix titled MYMATRIX that contains five observations from three variables for which we will use the default names COL1, COL2, and COL3. From this, we want to create a data set named NEWDATA. It is done as follows.

```
proc iml;
mymatrix = {
2 4 8,
3 9 1,
```

```
9 4 8,
1 1 1,
2 7 8};
create newdata from mymatrix;
append from mymatrix;
close newdata;
quit;
proc print data = newdata;
```

1.6 Multivariate Normal Distribution

A probability distribution that plays a pivotal role in much of the multivariate analysis is *multivariate normal distribution*. However, as this book concentrates more on the description of multivariate data, we will encounter it only occasionally. With that in mind, we give here only a very brief review of multivariate normal distribution. The material here is adopted from Khattree and Naik (1999). We say that $\mathbf{x}$ has a p-dimensional multivariate normal distribution (with a mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}$) if its probability density is given by

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

In notation, we state this fact as $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. 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 $\mathbf{Ax} \sim N_r(\mathbf{A}\boldsymbol{\mu},\ \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')(r \leq p)$. It may be added that $\mathbf{Ax}$ will admit the density if $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ 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 $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ will not be nonsingular. Consequently, the vector $\mathbf{Ax}$ will not admit a density function.
- Let $\mathbf{G}$ be such that $\boldsymbol{\Sigma}^{-1} = \mathbf{G}\mathbf{G}'$, then $\mathbf{G}'\mathbf{x} \sim N_p(\mathbf{G}'\boldsymbol{\mu},\ \mathbf{I})$ and $\mathbf{G}'(\mathbf{x}-\boldsymbol{\mu}) \sim N_p(\mathbf{0},\ \mathbf{I})$.
- Any fixed linear combination of $x_1, \ldots, x_p$, say, $\mathbf{c}'\mathbf{x}$, $\mathbf{c}_{p\times 1} \neq \mathbf{0}$ is also normally distributed. Specifically, $\mathbf{c}'\mathbf{x} \sim N_1(\mathbf{c}'\boldsymbol{\mu},\ \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})$.
- The subvectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are also normally distributed. Specifically, $\mathbf{x}_1 \sim N_{p_1}(\boldsymbol{\mu}_1,\ \boldsymbol{\Sigma}_{11})$ and $\mathbf{x}_2 \sim N_{p-p_1}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$, where with appropriate partitioning of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$,

  $$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix},$$

  and

  $$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

- Individual components $x_1, \ldots, x_p$ are all normally distributed. That is, $x_i \sim N_1(\mu_i, \sigma_{ii})$, $i = 1, \ldots, p$.
- The conditional distribution of $\mathbf{x}_1$ given $\mathbf{x}_2$, written as $\mathbf{x}_1|\mathbf{x}_2$, is also normal. Specifically,

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

Let $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\mathbf{x}_2 = \mathbf{B_0} + \mathbf{B}_1\mathbf{x}_2$, and $\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. The conditional expectation of $\mathbf{x}_1$ for given values of $\mathbf{x}_2$ or the regression function of $\mathbf{x}_1$ on $\mathbf{x}_2$ is $\mathbf{B_0} + \mathbf{B}_1\mathbf{x}_2$, which is linear in $\mathbf{x}_2$. 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 $\mathbf{x}_2$ given $\mathbf{x}_1$.

- Let $\boldsymbol{\delta}$ be a fixed $p \times 1$ vector, then

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

- The random components $x_1, \ldots, x_p$ 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

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

  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 the multivariate normal distribution, its properties, and the evaluations of the multivariate normal probabilities. See Anderson (1984), Kshirsagar (1972), Rao (1973), and Tong (1990) for further details.

1.6.1 Random Multivariate Normal Vector Generation

Oftentimes, we may want to generate random observations from a multivariate normal distribution. The following SAS/IML code, illustrated for $n = 10$ random observations from $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix},$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & .7 & .2 \\ .7 & 2 & -.8 \\ .2 & -.8 & 2.5 \end{bmatrix},$$

can be appropriately modified for this purpose. It is necessary to specify the appropriate values of $\boldsymbol{\mu}$ (MU), $\boldsymbol{\Sigma}$ (SIGMA), and the initial seed vector (SEED).

```
proc iml;
start rnorm(mu,sigma,seed);
z=normal(seed);
g=root(sigma);
x=mu+t(g)*z;
return(x);
finish;
do i=1 to 10;
x=rnorm(1,2,3,
1 .7 .2, .7 2 -.8, .2 -.8 2.5,
12345,87948,298765);
matx=matx//x‘;
end;
print matx;
```

The output, namely the ten vectors from the above trivariate normal population, are stored as a 10 by 3 matrix named MATX. Details about the steps of the generation can be found in Khattree and Naik (1999).

1.7 Concluding Remarks

This chapter is meant to be an introduction in order to prepare readers for what is covered within this book. There are many other concepts, topics, and methods that are not mentioned. However, the sections on matrix results and multivariate normality provide adequate preparation for appreciating and understanding the data analysis approaches discussed in this book. Some of the more advanced concepts are occasionally introduced in other chapters as and when their needs arise. Readers who are interested in the extensive study of matrix theory-related results as they apply in multivariate analysis should see Rao and Rao (1998).