Methods for Statistical Data Analysis of Multivariate Observations

R. GNANADESIKAN
Bell Telephone Laboratories
Murray Hill, New Jersey

John Wiley & Sons
New York · London · Sydney · Toronto

Published by John Wiley & Sons, Inc.

Library of Congress Cataloging in Publication Data:

Gnanadesikan, R
Methods for statistical data analysis of
multivariate observations.

(Wiley series in probability and mathematical
statistics)
Bibliography: p.
Includes index.
1. Multivariate analysis. I. Title.
QA278.G6 519.5'35 76-14994
ISBN 0-471-30845-5

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Preface

This book had its origins in a General Methodology Lecture presented at the annual meetings of the American Statistical Association at Los Angeles in 1966. A more concrete format for the book emerged from a paper (see Gnanadesikan & Wilk, 1969) presented at the Second International Symposium on Multivariate Analysis held at Dayton in June, 1968. That paper provided an outline of objectives for organizing the material in the present book, although the coverage here is more up to date, extensive, and detailed than the one in the paper. Specifically, the book is concerned with the description and discussion of multivariate statistical techniques and concepts, structured according to five general objectives in analyzing multiresponse data. The methods and underlying concepts are grouped according to these five objectives, and a chapter of the book is devoted to each objective.

The book is intended to emphasize methodology and data-based interpretations relevant to the needs of data analysis. As such, it is directed primarily toward applied statisticians and users of statistical ideas and procedures in various scientific and technological disciplines. However, some issues, arising especially out of the newer techniques described in the book, may be of interest to theoretical statisticians. Also, there are algorithmic aspects of the procedures which numerical analysts may find interesting.

Portions of the material in this book have been used by the author as the basis for a graduate-level series of lectures presented at Imperial College of Science & Technology of the University of London in 1969 and at Princeton University in 1971. Although the book can thus serve as a text, it differs from standard textbooks in not containing exercises. In view of the orientation of the book, the natural exercises would be to analyze specific sets of data by using the methods described in the text. However, rather than setting such exercises, which often tend to be artificial, it would seem to be far more useful to expect the students to use the relevant techniques on any real problems which they encounter either in their own work or in the course of their being consulted for statistical advice on the problems of others. Also, for making the purpose and usefulness of a technique more apparent, illustrative examples are used.

Such examples appear throughout the book and constitute an important facet of the presentation.

The coverage in this book is mainly of relatively recent (i.e., within the last decade) developments in multivariate methodology. When more classical techniques are described, the intention is either to provide a more natural motivation for a recent concept or method or to attempt a more complete discussion. A thorough review of all multivariate techniques is not a goal of the book. Specifically, for instance, no attention is given here to the analysis of multiple time series.

Despite the intention to emphasize relatively recent developments, the book inevitably reflects the fact that it was written over a period of six or seven years that have seen a spate of publications on multivariate topics. For instance, whereas material on cluster analysis written from a statistical viewpoint was relatively sparse when Chapter 4 of this book was conceived, there have been several recent articles and even whole books (e.g., Everitt, 1974; Hartigan, 1975) on this topic.

I am grateful to Bell Telephone Laboratories for its support of my efforts in writing this book and for providing so many important facilities without which the task could not have been undertaken. I also thank Imperial College and Princeton University for providing me with the stimulus and opportunity to organize the material for this book. It is a particular pleasure to acknowledge the many valuable comments of Professor D. R. Cox at the time of these lectures at Imperial College. Thanks are due also to my colleague Dr. J. R. Kettenring for his willingness to use parts of this material in a course that he taught and for his several helpful comments, based partly on the experience. I am deeply indebted to many past and present colleagues for their collaborative research efforts with me, which are reflected in various parts of this book. I wish also to acknowledge the kind permissions of other authors, several journals and publishers (including Academic Press, the American Statistical Association, Biometrics, the Biometrika Trustees, the Institute of Mathematical Statistics, Methuen & Co., Pergamon Press, Psychometrika, Statistica Neerlandica, and Technometrics) to incorporate material published elsewhere.

I am grateful to Mrs. M. L. Culp, Miss D. A. Williams, and Mrs. J. Charles for their careful typing of different parts of the manuscript, and the assistance of Messrs. I. L. Patterson and J. L. Warner in collating the exhibits is gratefully acknowledged.

Finally, I express my deepest sense of gratitude to my wife not only for her valuable comments but also for her constant encouragement during the writing of the book.

R. GNANADESIKAN

Murray Hill, New Jersey
January 1977

Contents

CHAPTER 1

Introduction

Most bodies of data involve observations associated with various facets of a particular background, environment, or experiment. Therefore, in a general sense, data are always multivariate in character. Even in a narrow sense, when observations on only a single response variable are to be analyzed, the analysis often leads to a multivariate situation. For example, in multiple linear regression, or in fitting nonlinear models, even with a single dependent variable, one often is faced with correlations among the estimated coefficients, and analyzing the correlation structure for possible reparametrizations of the problem is not an uncommon venture.

For the purposes of the present book, a more limited definition of a multivariate situation is used: multiresponse (or multivariate) problems are those that are concerned with the analysis of n points in p-space, i.e., when each of n persons, objects, or experimental units has associated with it a p-dimensional vector of responses. The experimental units need not necessarily constitute an unstructured sample but, in fact, may have a superimposed design structure, i.e., they may be classified or identified by various extraneous variables. One essential aspect of a multivariate approach to the analysis of such multiresponse problems is that, although one may choose to consider the p-dimensional observations from object to object as being statistically independent, the observed components within each vector will usually be statistically related. Exploitation of the latter feature to advantage in developing more sensitive statistical analyses of the observations is the pragmatic concern and value of a multivariate approach.

Most experimenters probably realize the importance of a multivariate approach, and most applied statisticians are equally well aware that multivariate analysis of data can be a difficult and frustrating problem. Some users of multivariate statistical techniques have, with some justification, even asserted that the methods may be unnecessary, unproductive, or misguided. Reasons for the frustrations and difficulties characteristic of multivariate data analysis, which often far exceed those encountered in univariate circumstances, appear to include the following:

1. It seems very difficult to know or to develop an understanding of what one really wants to do. Much iteration and interaction is required. This is also true in the uniresponse case in real problems. Perhaps in the multiresponse case one is simply raising this difficulty to the *p*th power!
2. Once a multiresponse view is adopted, there is no obvious "natural" value of *p*, the dimensionality of response. For any experimental unit it is always possible to record an almost indefinitely large list of attributes. Any selection of responses for actual observation and analysis is usually accomplished by using background information, preliminary analysis, informal criteria, and experimental insight. On the other hand, the number of objects or replications, *n*, will always have some upper bound. Hence *n* may at times be less than *p*, and quite often it may not be much greater. These dimensionality considerations can become crucial in determining what analyses or insights can be attained.
3. Multivariate data analysis involves prodigious arithmetic and considerable data manipulation. Even with modern high-speed computing, many multivariate techniques are severely limited in practice as to number of dimensions, *p*, number of observations, *n*, or both.
4. Pictures and graphs play a key role in data analysis, but with multiresponse data elementary plots of the raw data cannot easily be made. This limitation keeps one from obtaining the realistic primitive stimuli, which often motivate uniresponse analyses as to what to do or what models to try.
5. Last, but of great importance and consequence, points in *p*-space, unlike those on a line, do not have a unique linear ordering, which sometimes seems to be almost a basic human requirement. Most formal models and their motivations seem to grasp at optimization or things to order. There is no great harm in this unless, in desperation to achieve the comfort of linear ordering, one closes one's mind to the nature of the problem and the guidance which the data may contain.

Much of the theoretical work in multivariate analysis has dealt with formal inferential procedures, and with the associated statistical distribution theory, developed as extensions of and by analogy with quite specific univariate methods, such as tests of hypotheses concerning location and/or dispersion parameters. The resulting methods have often turned out to be of very limited value for multivariate data analysis.

The general orientation of the present book is that of statistical data analysis, concerned mainly with providing descriptions of the informational content of the data. The emphasis is on *methodology*—on underlying or motivating concepts and on data-based interpretations of the methods. Little or no coverage is given to distribution theory results, optimality properties, or formal or detailed mathematical proofs, or, in fact, to fitting the methods discussed into the framework of any currently known formal theory of statistical inference, such as decision theory or Bayesian analysis.

The framework for the discussion of multivariate methods in this book is provided by the following five general objectives of analyzing multiresponse data:

1. Reduction of dimensionality (Chapter 2).
2. Development and study of multivariate dependencies (Chapter 3).
3. Multidimensional classification (Chapter 4).
4. Assessment of statistical models (Chapter 5).
5. Summarization and exposure (Chapter 6).

The classification of multivariate methods provided by these five objectives is not intended to be in terms of mutually exclusive categories, and some techniques described in this book may be used for achieving more than one of the objectives. Thus, for example, a technique for reducing dimensionality may also prove to be useful for studying the possible internal relationships among a group of response variables.

With regard to the technology of data analysis, although it is perhaps true that this is still in a very primitive state, some important aids either are available or are under development. Raw computing power has grown astronomically in recent years, and graphical display devices are now relatively cheap and widely available. Much more data-analytic software is to be expected in the near future. Hardware-software configurations are being designed and developed, for both passive and interactive graphics, as related to the needs of statistical data analysis. Graphical presentation and pictorialization are important and integral tools of data analysis. (See Gnanadesikan, 1973, for a discussion of graphical aids for multiresponse data analysis.) A feature common to most of the methods discussed in the subsequent chapters of this book is their graphical nature. either implicit in their motivating ideas or explicit in their actual output and use.

In general, the mathematical notation used conforms to familiar conventions. Thus, for instance, $\mathbf{a}$, $\mathbf{x}$, . . . denote column vectors; $\mathbf{a}'$, $\mathbf{x}'$, . . . , row vectors; and $\mathbf{A}$, $\mathbf{Y}$, . . . , matrices. Whenever it is feasible and not unnatural, a distinction is made between parameters and random variables by using the familiar convention that the former are denoted by Greek letters and the latter by letters of the English alphabet. Most of the concepts and methods discussed are, however, introduced in terms of observed or sample statistics, i.e., quantities calculated from a body of data. Statistics that are estimates of parameters are often denoted by the usual convention of placing a hat (ˆ) over the parameter symbol.

Equations, figures, and tables that occur as part of the main text are numbered sequentially throughout the book. However, no distinction is made between figures and tables when they occur in the context of an example, and both are referred to as "exhibits." Thus Exhibit 5*a* is a table of numbers that appears in Example 5, whereas Exhibits 5*b* and *c* both are figures that are part of the same example.

A bibliography is included at the end of the book, and specific items of it that are directly relevant to a particular chapter are listed at the end of the chapter. An item in the bibliography is always cited by the name(s) of the author(s) and the year of publication. Thus Gnanadesikan (1973), Gnanadesikan & Wilk (1969), Kempthorne (1966), Tukey (1962), and Tukey & Wilk (1966) are specifically relevant references for the present chapter.

CHAPTER 2

Reduction of Dimensionality

2.1. GENERAL

The issue in reduction of dimensionality in analyzing multiresponse data is between attainment of simplicity for understanding, visualization, and interpretation, on the one hand, and retention of sufficient detail for adequate representation on the other hand.

Reduction of dimensionality can lead to parsimony of description, of measurement, or of both. It may also encourage consideration of meaningful physical relationships between the variables, for example, summarizing bivariate mass-volume data in terms of the ratio density = mass/volume.

As mentioned in Chapter 1, in many problems the dimensionality of response, p, is conceptually unlimited, whereas the number, n, of experimental units available is generally limited in practice. By some criteria of relevance, the experimenter always drastically reduces the dimensionality of the observations to be made. Such reduction may be based on (i) exclusion before the experiment; (ii) exclusion of features by specific experimental judgment; (iii) general statistical techniques, such as principal components analysis (see Section 2.2), use of distance functions of general utility, and methods for recognizing and handling nonlinear singularities (see Section 2.3); and/or (iv) specific properties of the problem which indicate the choice of a particular (unidimensional) real-valued function for analysis, e.g., relative weights for assigning an overall grade in matriculation examinations.

The first two of these approaches lead to a reduction of measurement in that the number of variables to be observed is diminished. The last two will not, in general, result in reducing current measurements but may reduce future measurements by showing that a subset of the variables is "adequate" for certain specifiable purposes of analysis. The major concern of the present chapter is the discussion of some specific examples of the third approach in the list above.

From the point of view of description, too severe a reduction may be undesirable. Meaningful statistical analysis is possible only when there has not been excessive elimination. Clearly a dominant consideration in the use of statistical procedures for the reduction of dimensionality is the interpretability of the lower dimensional representations. For instance, the use of principal components per se does not necessarily yield directly interpretable measures, whereas a reasonable choice of a distance function will sometimes permit interpretation.

Circumstances under which one may be interested in reducing the dimensionality of multiple response data include the following:

1. Exploratory situations in data analysis, for example, in psychological testing results or survey questionnaire data, especially when there is ignorance of what is important in the measurement planning. Here one may want to screen out redundant coordinates or to find more insightful ones as a preliminary step to further analysis or data collection.
2. Cases in which one hopes to stabilize "scales" of measurement when a similar property is described by each of several coordinates, for example, several measures of size of a biological organism. Here the aim is to compound the various measurements into a fewer number which may exhibit more stable statistical properties.
3. The compounding of multiple information as an aid in significance assessment. Specifically, one may hope that small departures from null conditions may be evidenced on each of several jointly observed responses. Then one might try to integrate these noncentralities into a smaller-dimensional space wherein their existence might be more sensitively indicated. One particular technique that has received some usage is the univariate analysis of variance applied to principal components.
4. The preliminary specification of a space that is to be used as a basis for eventual discrimination or classification procedures. For example, the raw information per object available as a basis for identifying people from their speech consists, in one version of the problem, of a 15,000-dimensional vector which characterizes each utterance! This array must be condensed as a preliminary to further classification analysis.
5. Situations in which one is interested in the detection of possible functional dependencies among observations in high-dimensional space. This purpose is perhaps the least well defined but nevertheless is prevalent, interesting, and important.

Many problems and issues exist in this general area of transformation of coordinates and reduction of dimensionality. These are problems of concept as to what one hopes to achieve, of techniques or methods to exhibit information that may be in the data, of interpretations of the results of applying available techniques, and of mathematical or algorithmic questions related to implementation. Specifically, if one develops a transformed or derived set of (reduced) coordinates, there is the question of whether these can be given some meaning or interpretation that will facilitate understanding of the actual problem. Similarly, it may or may not be true that derived coordinates, or approximations to these, will be directly observable. Sometimes such observability may occur with gains in efficiency and simplicity of both experiment and analysis.

Another problem in this area is that of the commensurability of the original coordinates and of the effect of this issue on a derived set of coordinates. This is not, apparently, a problem in principle, since there is no difficulty in dealing with functions of variables having different units. However, if the functions are themselves to be determined or influenced by the data, as in principal components analysis, some confusion may exist. An example of the issue involved here is presented in Section 2.2.1.

In looking for a reduced set of coordinates, classical statistical methodology has been largely concerned with derived coordinates that are just linear transforms of the original coordinates. This limitation of concern to linearity is perhaps due at least in part to the orientation of many of the techniques toward multivariate normal distribution theory. More recently, however, techniques have been suggested (Shepard, 1962a, b; Shepard & Carroll, 1966; Gnanadesikan & Wilk, 1966, 1969) for nonlinear reduction of dimensionality.

2.2. LINEAR REDUCTION TECHNIQUES

This section reviews briefly the classical linear reduction methods. First, discussion is provided of principal components analysis, a technique initially described by Karl Pearson (1901) and further developed by Hotelling (1933), which is perhaps the most widely used multivariate method. Second, concepts and techniques associated with linear factor analysis are outlined. Both the principal factor method due to Thurstone (1931) and the maximum likelihood approach due to Lawley (1940) are considered.

2.2.1. Principal Components Analysis

The basic idea of principal components analysis is to describe the dispersion of an array of n points in p-dimensional space by introducing a new set of orthogonal linear coordinates so that the sample variances of the given points with respect to these derived coordinates are in decreasing order of magnitude. Thus the first principal component is such that the projections of the given points onto it have maximum variance among all possible linear coordinates; the second principal component has maximum variance subject to being orthogonal to the first; and so on.

If the elements of $\mathbf{y}' = (y_1, y_2, \ldots, y_p)$ denote the p coordinates of observation, and the rows of the $n \times p$ matrix, $\mathbf{Y}'$, constitute the n p-dimensional observations, the sample mean vector and covariance matrix may be obtained, respectively, from the definitions

$$\bar{\mathbf{y}}' = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_p) = \frac{1}{n}\mathbf{1}'\mathbf{Y}', \tag{1}$$

$$\mathbf{S} = ((s_{ij})) = \frac{1}{n-1}(\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})', \tag{2}$$

where $\mathbf{1}'$ is a row vector all of whose elements are equal to 1, and $\bar{\mathbf{Y}}'$ is an $n \times p$ matrix each of whose rows is equal to $\bar{\mathbf{y}}'$. The $p \times p$ sample correlation matrix, $\mathbf{R}$, is related to $\mathbf{S}$ by

$$\mathbf{R} = \mathbf{D}_{1/\sqrt{s_{ii}}} \cdot \mathbf{S} \cdot \mathbf{D}_{1/\sqrt{s_{ii}}}, \tag{3}$$

where $\mathbf{D}_{1/\sqrt{s_{ii}}}$ is a $p \times p$ diagonal matrix whose ith diagonal element is $1/\sqrt{s_{ii}}$ for $i = 1, 2, \ldots, p$.

A geometric interpretation of principal components analysis is as follows: The inverse of the sample covariance matrix may be employed as the matrix of a quadratic form which defines a family of concentric ellipsoids centered on the sample center of gravity; i.e., the equations

$$(\mathbf{y} - \bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y} - \bar{\mathbf{y}}) = c, \tag{4}$$

for a range of nonnegative values of c, define a family of concentric ellipsoids in the p-dimensional space of $\mathbf{y}$. The principal components transformation of the data is just the projections of the observations onto the principal axes of this family. The basic idea is illustrated, for the two-dimensional case, in Figure 1. The original coordinates, (y_1, y_2), are transformed by a shift of origin to the sample mean, $(\bar{y}_1, \bar{y}_2)$, followed by a rigid rotation about this origin that yields the principal component coordinates, z_1 and z_2.

Algebraically, the principal components analysis involves finding the eigenvalues and eigenvectors of the sample covariance matrix. Specifically, for obtaining the first principal component, z_1, what is sought is the vector of coefficients, $\mathbf{a}' = (a_1, a_2, \ldots, a_p)$, such that the linear combination, $\mathbf{a}'\mathbf{y}$, has maximum sample variance in the class of all linear combinations, subject to the normalizing constraint, $\mathbf{a}'\mathbf{a} = 1$. For a given $\mathbf{a}$, since the sample variance of $\mathbf{a}'\mathbf{y}$ is $\mathbf{a}'\mathbf{S}\mathbf{a}$, the problem of finding $\mathbf{a}$ turns out to be equivalent to determining a nonnull $\mathbf{a}$ such that the ratio $\mathbf{a}'\mathbf{S}\mathbf{a}/\mathbf{a}'\mathbf{a}$ is

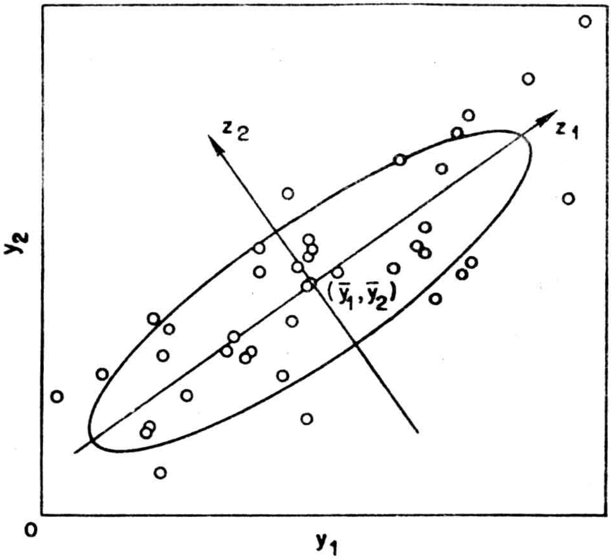

Fig. 1. Illustration of principal components with bivariate data.

maximized. It is well known that the maximum value of this ratio is the largest eigenvalue, c_1, of $\mathbf{S}$, and the required solution for $\mathbf{a}$ is the eigenvector, $\mathbf{a}_1$, of $\mathbf{S}$ corresponding to c_1.

After the first principal component has been determined, the next problem is to determine a second normalized linear combination orthogonal to the first and such that, in the class of all normalized linear functions of $\mathbf{y}$ orthogonal to $\mathbf{a}_1'\mathbf{y}$, the second principal component has largest variance. At the next stage, one would determine a third normalized linear combination with maximum variance in the class of all normalized linear combinations orthogonal to the first two principal components. The process may be repeated until p principal components have been determined. The problem of determining the p principal components is equivalent to determining the stationary values of the ratio $\mathbf{a}'\mathbf{S}\mathbf{a}/\mathbf{a}'\mathbf{a}$ for variation over all nonnull vectors, $\mathbf{a}$. These stationary values are known to be the eigenvalues, $c_1 \geq c_2 \geq \cdots \geq c_p \geq 0$, of $\mathbf{S}$, and the required principal components are provided by $\mathbf{a}_1'\mathbf{y}$, $\mathbf{a}_2'\mathbf{y}$, $\ldots$, and $\mathbf{a}_p'\mathbf{y}$, where $\mathbf{a}_i'$ is the normalized eigenvector of $\mathbf{S}$ corresponding to the eigenvalue, c_i, for $i = 1, 2, \ldots, p$. The ranked eigenvalues are in fact just the sample variances of the linear combinations of the original variables specified by the eigenvectors.

The above results can also be related to the so-called spectral decomposition (see, e.g., Rao, 1965, p. 36) of the matrix $\mathbf{S}$: there exists an orthogonal matrix, $\mathbf{A}$, such that $\mathbf{S}=\mathbf{A}\mathbf{D}_c\mathbf{A}'$, where $\mathbf{D}_c$ is a diagonal matrix with diagonal elements $c_1, c_2, \ldots, c_p$. The columns of $\mathbf{A}$ are the eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$. The principal component coordinates, which for convenience are defined to include a shift of origin to the sample mean, are then specified by the transformation

$$\mathbf{z}=\mathbf{A}'(\mathbf{y}-\bar{\mathbf{y}}), \tag{5}$$

and the principal components transformation of the data is

$$\mathbf{Z}=\mathbf{A}'(\mathbf{Y}-\bar{\mathbf{Y}}). \tag{6}$$

When transformed to the principal component coordinate system, the observations have certain desirable statistical properties. For instance, the sample variance of the observations with respect to the ith principal component is $\mathbf{a}_i'\mathbf{S}\mathbf{a}_i = c_i$, the ith largest eigenvalue of $\mathbf{S}$, for $i = 1, 2, \ldots, p$, and the sum of the sample variances with respect to the derived coordinates $=\sum_{i=1}^p c_i = \text{tr}(\mathbf{S}) = \sum_{i=1}^p s_{ii}$ = sum of the variances with respect to the original coordinates. Furthermore, because of the mutual orthogonality of the representations of the original observations in terms of the principal component coordinates, the sample covariances (and hence the sample correlations) between pairs of the derived variables are all 0. This follows geometrically from the "orthogonal" nature of the two-dimensional configuration of the projections of the observations onto each member of every pair of principal component coordinates. Equivalently, it follows algebraically from the relationship that the sample covariance between the ith and jth principal components coordinates $=$ $\mathbf{a}_i'\mathbf{S}\mathbf{a}_j = c_j\mathbf{a}_i'\mathbf{a}_j = 0$ since $\mathbf{a}_i$ and $\mathbf{a}_j$ (for $i \neq j$) are orthogonal.

The above geometrical, algebraic, and algorithmic descriptions have been presented in terms of the covariance matrix. Clearly, if one standardizes each coordinate by dividing by its sample standard deviation, then the covariance matrix of the standardized variables is just the correlation matrix of the original variables.. Thus the above discussion applies to principal components analysis of the correlation matrix.

In light of the current state of the knowledge on numerically stable computational methods, the recommended algorithm for performing the eigenanalysis involved in obtaining the principal components is either the so-called QR method applied to $\mathbf{S}$ or $\mathbf{R}$ (Businger, 1965), or the so-called singular value decomposition technique performed on $(\mathbf{Y}-\bar{\mathbf{Y}})$ or on the standardized form, $\mathbf{D}_{1/\sqrt{s_{ii}}}(\mathbf{Y}-\bar{\mathbf{Y}})$ (Businger & Golub, 1969; Golub, 1968).

If the sample size, n, is not greater than the dimensionality, p, the sample covariance matrix will be singular, corresponding to the fact that all n points will lie on a hyperplane of dimension less than p. Within that linear subspace one can define a dispersion matrix and find its principal components. This will be reflected in the eigenvalue analysis of the singular covariance matrix, in that some of the eigenvalues will be 0. The eigenvectors corresponding to the nonzero eigenvalues will give the projections of the observations onto orthogonal coordinates within the linear subspace containing the observations.

One hope in the case of principal components analysis is that the bulk of the observations will be near a linear subspace and hence that one can employ a new coordinate system of reduced dimension. Generally, interest will lie in the coordinates along which the data show their greatest variability. However, although the eigenvector corresponding to the largest eigenvalue, for example, provides the projection of each point onto the first principal component, the equation of the first principal component coordinate is given by the conjunction of the equations of planes defined by the remaining eigenvectors. More generally, if most of the variability of a p-dimensional sample is confined to a q-dimensional linear subspace, that subspace is described by the $(p-q)$ eigenvectors which correspond to the $(p-q)$ "small" eigenvalues. For purposes of interpretation—detection or specification of constraints on, or redundancy of, the observed variables—it may often be the relations which define near constancy (i.e., those specified by the smallest eigenvalues) that are of greatest interest.

An important practical issue in eigenanalyses is that of judging the relative magnitudes of the eigenvalues, both for isolating "negligibly small" ones and for inferring groupings, if any, among the others. The issue involves not only computational questions, such as the specification of what constitutes a zero eigenvalue, but also questions of statistical inference and useful insight. The interpretation of magnitude and separation of eigenvalues from a sample covariance matrix is considerably complicated by the sampling variation and statistical interdependence, as exhibited even by the eigenvalues of a covariance matrix calculated from observations from a spherical normal distribution. Although there are some tests of significance, which have been proposed as formal inferential aids, a real need exists for data-analytic procedures for studying the configuration of a collection of sample eigenvalues as a whole (see Section 6.2 for further discussion).

Clearly, principal components are *not* invariant under linear transformation, including separate scaling, of the original coordinates. Thus the principal components of the covariance matrix are not the same as those of the correlation matrix or of some other scaling according to measures of "importance." Note, however, that the principal components of the correlation matrix are invariant under separate scaling of the original