Methods for Statistical Data Analysis of Multivariate Observations

Methods for Statistical Data Analysis of Multivariate Observations

Second Edition

R. GNANADESIKAN
Rutgers University
New Brunswick, New Jersey

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York · Chichester · Weinheim · Brisbane · Singapore · Toronto

This text is printed on acid-free paper.

Published by John Wiley & Sons, Inc.

Library of Congress Cataloging in Publication Data:

Gnanadesikan, Ramanathan, 1932–
Methods for statistical data analysis of multivariate observations. -- 2nd ed. / R. Gnanadesikan.
p. cm. -- (Wiley series in probability and statistics. Applied probability and statistics)
"A Wiley-Interscience publication."
Includes index.
ISBN 0-471-16119-5 (cloth : alk. paper)
1. Multivariate analysis. I. Title. II. Series.
QA278.G6 1997 96-30766
519.5'35--dc20 CIP

10 9 8 7 6 5 4 3 2 1

To the family of my childhood
and
the family of my parenthood

स्वस्ति नो बृहस्पतिर्दधातु | सहायकानां च स्वस्ति |

नमः पूर्वपुरूषेभ्यः | ॐ शान्तिः शान्तिः शान्तिः |

*In the manner of the Upanishads, I invoke the
blessings of the Vedic Lord of Prayers upon you,
my good reader, upon myself and upon my collaborators.
I also take this opportunity to pay my obeisance to
my ancestors and my predecessors in my field.*

Contents

Preface to the Second Edition

Since the publication of the first edition a number of developments have had major effects on the current state of the art in multiresponse data analysis. These include not only significant augmentation of the technology, such as enhanced computing power including numerics and graphics, but also major statistical methodological developments stimulated in part by real-world problems and needs. Diagnostic aids which tend to be mostly graphical, robust/resistant methods whose results are not sensitive to deviant behavior of real-world data, and cluster analysis techniques for pattern recognition, are just a few examples of such methodological developments over the past two decades. The scope, structure and applied emphasis of the first edition provide a natural setting for many of the new methods. The main objective of the second edition is to expand the coverage of methods while retaining the framework of the first edition.

The recent decade has also seen a fundamental change in the paradigm of data analysis. While there are differences in the details depending on the specific problem at hand, there are some general features of the new paradigm that can be contrasted with the more classical approach. For example, the newer developments tend to cast solutions in terms of "fitting" functions that are not globally parametrized (such as planes fitted to points in p-dimensional space) but instead are more "locally" focused and then "pieced together" in some fashion (e.g., low-order splines, or even planes), with some form of trade-off between "accuracy" and "smoothness" involved in the fitting algorithms. The flexibility gained in the richness of the relationships that can be handled, including the ability to accommodate nonlinear ones, is often at the expense of iterative fitting of several local relationships, and the lack of succinct or parsimonious descriptions that are features of the more classical approaches to the same problems. Also, distributional models that play a role in statistical assessment and inferences in the more classical approaches, tend to be deemphasized in the new paradigm. The reliance is on more data-dependent and computer-intensive tools, such as resampling (e.g., jackknife, bootstrap, cross validation), to provide the basis for inferences and assessments of performance.

The methods based on the new paradigm have a great deal of appeal but yet, like most things in the context of the complexities of the real world, they are not a panacea. With a considerably broadened base of experience, and inevitable modifications and adaptations of them, the newer methods will eventually perhaps replace the classical techniques. However, for both pedagogy and practice, the classical methods will probably be around for quite a while. Widely accessible software implementations of the classical techniques, as well as the comfort of the familiarity of their conceptual underpinnings, suggest that this will be so.

For the immediate purposes of this second edition, it was tempting to incorporate all of the newer developments and integrate them with the more classical methods. However, although many of the methods developed since the first edition may adopt the classical paradigm rather than the new one, because of their number and the wide relevance of their conceptual underpinnings, a decision was made to include the details of just some of the newer methods that adopt the classical paradigm, and only briefly mention (with appropriate references) specific approaches which fall under the new paradigm. Among other things, this has enabled a more manageable increase in the volume of material to be included in the second edition. Currently, there are a few books available that are concerned with methods based on the new paradigm and addressed to specific topics of multivariate analysis. Hopefully, one or more of the people who have played a central role in the development of multivariate data analysis techniques with the new framework, will soon write a comprehensive book on these methods.

New material appears in virtually every chapter of this edition. However, there are heavier concentrations in some more than in others. A major expansion, reflecting the vigorous development of methods as well as applications in the field of pattern recognition, is the material on cluster analysis. New sections, focused on issues of inputs to clustering algorithms and on the critical need for aids in interpreting the results of cluster analysis, have been added. Other new material in this edition pertains to useful summarization and exposure techniques in Chapter 6 that did not exist at the time of the first edition. For instance, descriptions have been added of new graphical methods for assessing the separations amongst the eigenvalues of a correlation matrix and for comparing sets of eigenvectors. Topics that have been enlarged on, largely due to the increased experience with some of the techniques that were relatively new at the time of publication of the first edition, include robust estimation and a class of distributional models that is slightly broader than the multivariate normal in Chapter 5. A new appendix on software, with particular reference to the functions available in two widely-used systems, S (or Splus) and SAS, is included for help with statistical computing aspects.

In the light of the decision regarding the second edition, the intended audience for it is the same one identified in the preface to the first edition. In the years since the first edition was published, the author has been gratified to hear from many people in fields of application of multivariate statistical methods that they have found the book useful in their work. These fields have ranged from industry, management science and engineering and physical sciences to anthropology, biology, behavioral and social sciences, and information science. As for pedagogy, the author has used the book as the basis of a graduate course in applied multivariate analysis taught at Rutgers University with students drawn from various disciplines in addition to statistics. The material, combined with projects that involve using the techniques discussed in the book for analyzing data from the students' own discipline or job interests, has proven to be highly effective.

I wish to thank the many students who have provided valuable feedback about the contents and clarity of the book. Bellcore as an organization and, in particular, my research collaborator and friend across the years, Dr. Jon Kettenring, deserve special thanks for their support of my work on this edition. I am grateful to Suzanne Merten for her patient and cheerful word processing help at Bellcore.

R. GNANADESIKAN

New Brunswick, New Jersey
December 1996

Preface to the First Edition

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

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, that is, 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, that is, they may be classified or identified by various extraneous variables. One essential aspect of a multivariate approach to the analysis of such multireponse 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 pth 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); and
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, $\mathbf{a}, \mathbf{x}, \ldots$ denote column vectors; $\mathbf{a}', \mathbf{x}', \ldots$, row vectors; and $\mathbf{A}, \mathbf{Y}, \ldots$, 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, that is, quantities calculated from a body of data. Statistics that are estimates of parameters are often denoted by the usual convention of placing a hat ($\hat{\ }$) 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 variable selection procedures for choosing a subset of the variables that is particularly appropriate for the analysis at hand, 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, for example, 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 np-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; that is, 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 analyis 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 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

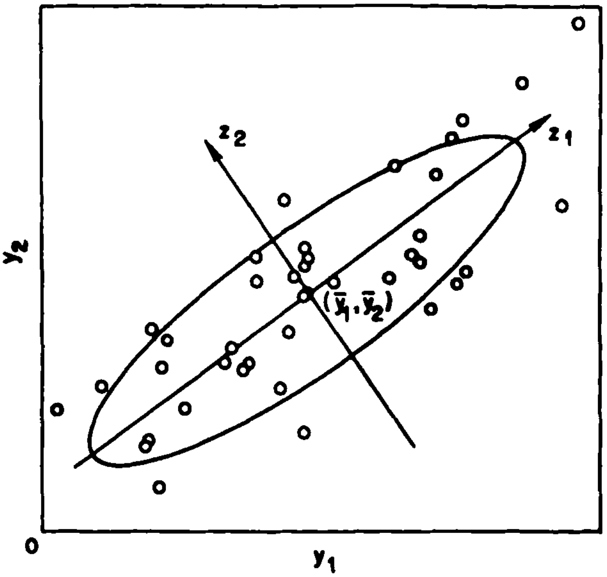


Fig. 1. Illustration of principal components with bivariate data.

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, $c_1 \geqslant c_2 \geqslant \cdots \geqslant c_p \geqslant 0$, of 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 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, for example, Rao, 1965, p. 36) of the matrix S: there exists an orthogonal matrix, 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 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 S, for $i = 1, 2, \ldots, p$, and the sum of the sample variances with respect to the derived coordinates $= \Sigma_{i=1}^{p} c_i = \mathrm{tr}(\mathbf{S}) = \Sigma_{i=1}^{p} s_{ii}$ = 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). For example, the singular value decomposition of the $p \times n$ matrix, $(\mathbf{Y} - \bar{\mathbf{Y}})$, is the matrix product, $\mathbf{ADQ}'$, where both $\mathbf{A}$ and $\mathbf{Q}'$ are orthogonal matrices. The columns of the $p \times p$ matrix, $\mathbf{A}$, are the eigenvectors of $\mathbf{S}$, while the columns of $\mathbf{Q}$ are the eigenvectors of $(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}})$. The $p \times n$ matrix, $\mathbf{D}$ is defined by $\mathbf{D} = [\mathbf{D}_d \mid \mathbf{0}]$, where $\mathbf{D}_d$ is a $p \times p$ diagonal matrix with diagonal elements $d_i = \sqrt{(n-1)c_i}, i = 1, \ldots, p$, where $c_1, c_2, \ldots, c_p$ are the eigenvalues of S. In terms of the singular value decomposition, the principal components transformation of the data defined in Eq. 6 may be calculated as $\mathbf{Z} = \mathbf{DQ}'$. The singular value decomposition of the standardized form of the data is related to the principal components of $\mathbf{R}$ in an analogous manner.

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 variables. For this reason, as well as for numerical computational ones, some have urged that principal components analysis always be performed on the correlation matrix. However, for other reasons of a statistical nature, such as interpretation, formal statistical inference, and distribution theory, it often is preferable to work with the covariance matrix. There does not seem to be any *general* elementary rationale to motivate the choice of scaling of the variables as a preliminary to principal components analysis on the resulting covariance matrix.

An important exception regarding invariance occurs when the observations are confined to a linear subspace. In this case, the specification of the singularities is unique under nonsingular linear transformation of the variables. One might expect that, loosely speaking, similar near uniqueness would hold when the data have a "nearly singular" structure.

A different issue, which arises specifically in the context of interpreting the principal components, is the tendency to interpret the relative weights assigned to the different variables in a given principal component. By scanning the numerical values of the coefficients, one may wish to simplify the pattern by setting some of them to 0 or to ± 1 (keeping in mind the need to normalize the vector to unit length). This is a natural step in the analysis (as illustrated in Example 1 below) and, from a data analysis viewpoint, rather than decreeing that it should not be done, a more useful exercise would be to provide a measure of how good an approximation the modified (and more easily interpretable) principal component is to the unmodified one. For example, suppose the eigenvector, $\mathbf{a}_1$, defining the first principal component is modified in the above manner to obtain the vector, $\mathbf{a}_1^*$. Then, the variance of the linear combination that results from using the elements of $\mathbf{a}_1^*$ as the coefficients of the variables would be $\mathbf{a}_1^{*\prime}\mathbf{S}\mathbf{a}_1^*$, which would necessarily be smaller than $\mathbf{a}_1'\mathbf{S}\mathbf{a}_1$ ($= c_1$, the largest eigenvalue of $\mathbf{S}$). Hence, a simple indicator of the price paid, in terms of explained variance, for using the "sub-optimal" but more interpretable coefficients would be the percentage of excess variance, $100[(c_1 - \mathbf{a}_1^{*\prime}\mathbf{S}\mathbf{a}_1^*)/c_1]$. One can use this way of quantifying the price for simplicity of interpretation with any individual principal component. If one has modified a set of principal components for making them more interpretable and wishes to get a measure of how close the space spanned by the modified eigenvectors is to the space spanned by the original eigenvectors, then canonical correlation analysis (see Section 3.3) can be used.

To conclude the present discussion of linear principal components analysis, an example of application is considered next. The use of the technique will be discussed further in Section 6.4.

Example 1. This example is taken from Blackith & Roberts (1958) and has also been discussed as an application of linear principal components analysis by Blackith (1960) and by Seal (1964). It deals with measurements on 375 ($= n$) grasshoppers of 10 ($= p$) characters that were chosen to cover the major areas of the body. The 375 grasshoppers were, in fact, cross-classifiable into eight groups—two species, two sexes, and two color forms. One interest in the data was to study and to characterize basic patterns of growth of the grasshoppers. Suppose that, for $g = 1, 2, \ldots, 8, n_g$ denotes the number of grasshoppers measured in the gth group, and $\mathbf{X}_g, \bar{\mathbf{x}}_g$, and $\mathbf{S}_g$ are, respectively, the matrix of observations, the mean vector, and the covariance matrix for the gth group.

The mean vector and the covariance matrix for each group are obtained by using Eqs. 1 and 2 of this chapter. Blackith (1960) reports on a principal components analysis of the pooled 10 × 10 covariance matrix,

$$S = \frac{1}{n-8}\sum_{g=1}^{8}(n_g - 1)S_g,$$

where $n = \Sigma_{g=1}^{8} n_g = 375$. The pooled covariance matrix is based on 367 degrees of freedom. Exhibit 1*a*, taken from Blackith (1960), shows the first three eigenvalues and the corresponding three eigenvectors, which, therefore, define the first three principal components. The sum of the three eigenvalues is 16.924, and the first three principal components accounts for about 99% of the observed variation in 10-dimensional space.

Note that the normalization of these eigenvectors has been accomplished by making the largest element 1, instead of the more usual unit Euclidean norm scheme of making the squares of the elements add to 1. This, however, does not interfere with the interpretation of the results. Thus each of the three eigenvectors in Exhibit 1*a* is "close to" a corresponding unit eigenvector with a single nonzero element which is unity: the first to $(1, 0, 0, \ldots, 0)$, the second to $(0, 1, 0, \ldots, 0)$, and the third to $(0, 0, 1, 0, \ldots, 0)$. Utilizing the idea discussed above for assessing the price paid for using more interpretable principal components, in this example substituting the vector $(1, 0, 0, \ldots, 0)$, for the one containing the coefficients specifying the first principal component, results in paying a price of $100[(16.09 - 15.78)/16.09] \simeq 2\%$ in decreased variance. The use of the second and third variables, respectively, instead of the second and

Exhibit 1*a*. First three principal components for grasshoppers data (Blackith, 1960)

			Eigenvalues	
		16.087	0.516	0.321
Variate	Variance		Eigenvectors	
1. Reduced wt. (mg.)	15.7725	1.0000	−0.0678	−0.1056
2. # Antennal segments	0.5531	0.0523	1.000	−0.1027
3. Elytron length (mm.)	0.4155	0.0847	0.0694	1.000
4. Head width (mm.)	0.0138	0.0215	0.0141	0.0155
5. Pronotal width (mm.)	0.0150	0.0197	0.0146	0.0098
6. Hind femoral length (mm.)	0.2545	0.0929	0.0928	0.2688
7. Hind femoral width (mm.)	0.0198	0.0233	0.0024	0.0008
8. Prozonal length (mm.)	0.0097	0.0110	0.0055	−0.0095
9. Metazonal length (mm.)	0.0197	0.0150	0.0160	0.0555
10. Front femoral width (mm.)	0.0015	0.0046	−0.0025	0.0068

third principal components, on the other hand, results in approximately 7% and 29% increased variances.

This example also brings out another general feature of principal components analysis. The feature is that the first principal component weights almost exclusively in this example, the original variable that has the largest variance and, similarly, the second and third principal components, in turn, end up weighting the original variables with the second and third largest variances, respectively. The sensitivity of the principal component coordinates to the variances of the original variables implies a critical dependence of the derived coordinates on the choice of scales for observing the original variables. Moreover, in the present example, the two characteristics with largest variances—namely, reduced weight and number of antennal segments—also happen to be measured on different scales from the one (millimeters) used for the remaining eight characteristics. Thus an additional issue here is the effect of the commensurability of the observed responses on the derived principal component coordinates.

One way of handling this difficulty in this example is to omit the two responses measured on very different scales and then perform a principal components analysis on the remaining eight responses. Another approach, which was mentioned earlier, would be to perform the analysis on the correlation matrix instead of the covariance matrix. Exhibit 1*b* shows the eigenvalues (see Seal, 1964) obtained in principal components analyses performed on both covariance and correlation matrices for the full set of 10 responses as well as for the reduced set of 8 responses.

Lines indicating intuitively reasonable separations among the eigenvalues are also shown in Exhibit 1*b*, a dashed line denoting a weak separation and a

Exhibit 1*b*. Eigenvalues for four principal components analyses (Seat, 1964)

	S(10 × 10)	R(10 × 10)	S(8 × 8)	R(8 × 8)
	16.087	4.802	0.549	3.959
	0.516	0.970	0.145	0.923
	0.321	0.898	0.021	0.867
	0.103	0.852	0.015	0.634
	0.017	0.637	0.009	0.588
	0.012	0.587	0.006	0.501
	0.009	0.499	0.003	0.339
	0.006	0.351	0.001	0.189
	0.003	0.218		
	0.001	0.186		
Total	17.075	10.0	0.749	8.0

solid line suggesting stronger separation. The indicated number and location of the separations are seen to be different between analyses performed on covariance matrices and those done on correlation matrices. Thus, both with all 10 responses and with the subset of 8, the principal components analyses of correlation matrices suggest that only the largest eigenvalue is clearly separated from the remaining ones. The analyses based on the corresponding covariance matrices, however, seem to suggest two separations among the eigenvalues.

Seal (1964) provides a reasonable argument for the indicated single separation among the eigenvalues of the 8 × 8 correlation matrix. Many of the off-diagonal elements of the matrix appear to be essentially the same, thereby indicating a nearly equicorrelational structure among the variables. In the case of the eight responses, all of the correlation coefficients appear to be about 0.4. With exact equicorrelational structure, a $p \times p$ correlation matrix will have only two distinct eigenvalues, one being equal to $1 + (p - 1)r$ and the remaining $(p - 1)$ being equal to $(1 - r)$, where r is the common value of all the correlation coefficients. An interesting question in the present example is whether the equicorrelation is inherent and experimentally sensible or is induced by the pooling of the covariance matrices from the eight groups of grasshoppers. Pooling several widely different covariance structures may lead to an "average" equicorrelational structure, and the relatively low value of the "common" correlation coefficient (0.4) in the example raises the question of a possible artifactual nature of the observed equicorrelation. Had the covariance matrix within each group been available, a technique described in Section 6.3.2 could have been used to study the appropriateness of the preliminary pooling of the eight covariance matrices in this example.

A somewhat different issue here is the relevance of analyzing the data on the observed scales of measurement rather than transforming the observations before the analysis. An interesting and seemingly appropriate transformation in this case would be to use logarithms of the original observations as the starting point of the principal components analysis. Unfortunately, the raw observations in the example are unavailable and such an analysis is therefore not possible.

2.2.2. Factor Analysis

The so-called model in factor analysis is

$$\mathbf{y} = \mathbf{\Lambda f} + \mathbf{z}, \tag{7}$$

where $\mathbf{y}$ is a p-dimensional vector of observable responses, $\mathbf{\Lambda}$ is a $p \times q$ matrix of unknown parameters called *factor loadings*, $\mathbf{f}$ is a q-dimensional vector of hypothetical (unobserved) variables called *common factors*, and $\mathbf{z}$ is a p-dimensional vector of hypothetical (unobserved) variables called *unique factors*. [*Note*: To distinguish between $\mathbf{f}$ and $\mathbf{z}$ one needs to impose the condition that each column of $\mathbf{\Lambda}$ has at least two nonzero elements.] Generally, it is further assumed that the components of $\mathbf{z}$ are mutually uncorrelated as well as being uncorrelated with the elements of $\mathbf{f}$. In other words, the covariance matrix of $\mathbf{z}$ is a $p \times p$ diagonal matrix, $\mathbf{\Delta}$, with diagonal elements $\delta_1^2, \ldots, \delta_p^2$, and the cross-covariance matrix between $\mathbf{f}$ and $\mathbf{z}$ is null.

With n observations available on p responses which are being studied simultaneously, the above model may be written as

$$\mathbf{Y} = \mathbf{\Lambda F} + \mathbf{Z}, \tag{8}$$

where $\mathbf{Y}$ is $p \times n$, $\mathbf{F}$ is $q \times n$, and $\mathbf{Z}$ is $p \times n$. The factor-analytic model in Eq. 7 (or 8), taken together with the above assumptions, specifies the following relationship among the covariance matrices of the different sets of variables involved:

$$\mathbf{\Sigma}_{yy} = \mathbf{\Lambda \Sigma}_{ff}\mathbf{\Lambda}' + \mathbf{\Delta}, \tag{9}$$

where $\mathbf{\Sigma}_{yy}$ denotes the $p \times p$ covariance matrix of $\mathbf{y}$, and $\mathbf{\Sigma}_{ff}$ denotes the $q \times q$ covariance matrix of $\mathbf{f}$. If the q common factors are assumed to be standardized and mutually uncorrelated, then $\mathbf{\Sigma}_{ff} = \mathbf{I}$, and

$$\mathbf{\Sigma}_{yy} = \mathbf{\Lambda\Lambda}' + \mathbf{\Delta}. \tag{10}$$

Formally, the two cases represented by Eqs. 9 and 10 are indistinguishable. This is due to the fact that one can write $\mathbf{\Sigma}_{ff} = \mathbf{TT}'$, where $\mathbf{T}$ is a lower triangular matrix, and rewrite Eq. 9 as $\mathbf{\Sigma}_{yy} = \mathbf{\Lambda}^*\mathbf{\Lambda}^{*\prime} + \mathbf{\Delta}$, where $\mathbf{\Lambda}^* = \mathbf{\Lambda T}$. Despite this formal indistinguishability, however, the representations of the data in terms of correlated and of uncorrelated factors would be different. Thus, for purposes of interpretation, it may be important to distinguish between the two cases.

An alternative way of motivating the factor-analytic model, which may be more appealing statistically, is as follows: given p observable variables, $\mathbf{y}$, do there exist $q(<p)$ variables, $\mathbf{f}$, such that the partial correlations between every pair of the original variables upon elimination of the q f-variables are all zero? An affirmative answer to this question may be shown to be equivalent to the factor-analytic model as specified by Eq. 7 (or Eq. 9), for from Eq. 7 it follows that the conditional covariance matrix of $\mathbf{y}$ given $\mathbf{f}$ = covariance matrix of $\mathbf{y} - \mathbf{\Lambda f}$ = covariance matrix of $\mathbf{z} = \mathbf{\Delta}$. Hence, from the assumptions concerning $\mathbf{\Delta}$, the off-diagonal elements of the conditional covariance matrix, which are the partial covariances between pairs of the y-variables, given $\mathbf{f}$, are all 0, so that the partial correlations between pairs of the elements of $\mathbf{y}$, given $\mathbf{f}$, are also all 0. Conversely, suppose there exists $\mathbf{f}$ such that the partial correlation between every pair of y-variables, given $\mathbf{f}$, is 0. Then, from the definition of partial correlation, it follows that the covariance matrix of the "residuals" from the linear regression of $\mathbf{y}$ on $\mathbf{f}$ is diagonal. The linear regression of $\mathbf{y}$ on $\mathbf{f}$ is $\mathscr{E}(\mathbf{y} \mid \mathbf{f}) = \mathbf{\Sigma}_{yf}\mathbf{\Sigma}_{ff}^{-1}\mathbf{f}$, where $\mathscr{E}$ stands for expectation, and $\mathbf{\Sigma}_{yf}$ denotes the $p \times q$

cross-covariance matrix between $\mathbf{y}$ and $\mathbf{f}$. The conditional covariance matrix of $\mathbf{y}$ given $\mathbf{f}$ = covariance matrix of the "residuals," $\mathbf{y}-\boldsymbol{\Sigma}_{yf}\boldsymbol{\Sigma}_{ff}^{-1}\mathbf{f}=\boldsymbol{\Sigma}_{yy}-\boldsymbol{\Sigma}_{yf}\boldsymbol{\Sigma}_{ff}^{-1}\boldsymbol{\Sigma}'_{yf}$. If this is a diagonal matrix, $\boldsymbol{\Delta}$, then it follows that $\boldsymbol{\Sigma}_{yy} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}_{ff}\boldsymbol{\Lambda}' + \boldsymbol{\Delta}$, where

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}_{yf}\boldsymbol{\Sigma}_{ff}^{-1}. \tag{11}$$

Thus the equivalence claimed at the beginning of this paragraph follows.

One hope in using factor analyis is that the number of common factors, q, will be much smaller than the number of original variables, p, thus leading to a parsimony of description which may aid in interpretation and understanding. Formally, the problems to be solved in a factor-analytic approach include (*a*) finding $\boldsymbol{\Lambda}$ of minimal rank to satisfy the model as summarized by Eq. 9 (or by Eq. 10); (*b*) estimating $\boldsymbol{\Delta}$; and (*c*) making inferences about $\mathbf{F}$. For present purposes, most of the discussion in the rest of this section is devoted to (*a*). A more extensive and thorough discussion of factor analysis will be found in Harman (1967) and Lawley & Maxwell (1963).

Equation 11 suggests that one way of obtaining $\boldsymbol{\Lambda}$ is to regress $\mathbf{y}$ on $\mathbf{f}$. However, this is not a feasible direct approach since $\mathbf{f}$ is not observable. Two other methods for estimating the factor loadings—the principal factor and the maximum likelihood methods—are outlined below.

Before discussing these methods, a few additional concepts and terms need to be introduced. First, as a consequence of Eq. 10, one has

$$\text{variance }(y_i) = \sigma_{ii} = \sum_{j=1}^{q} \lambda_{ij}^2 + \delta_i^2 \qquad i = 1,\ldots,p,$$

where $\boldsymbol{\Sigma}_{yy} = ((\sigma_{il}))$ and $\boldsymbol{\Lambda} = ((\lambda_{ij}))$. The quantity

$$h_i^2 = \sum_{j=1}^{q} \lambda_{ij}^2 \tag{12}$$

is called the *communality* of the ith variable, while δ_i^2 is termed the *uniqueness* of the ith variable ($i = 1, 2, \ldots, p$). It follows that

$$\begin{aligned}\text{total variance} = \operatorname{tr}(\boldsymbol{\Sigma}_{yy}) &= \sum_{i=1}^{p} \sigma_{ii}\\ &= \sum_{i=1}^{p} (h_i^2 + \delta_i^2)\\ &= \sum_{i=1}^{p}\sum_{j=1}^{q} \lambda_{ij}^2 + \sum_{i=1}^{p} \delta_i^2\\ &= V + \delta^2, \end{aligned} \tag{13}$$

where $V = \Sigma_{i=1}^{p}\Sigma_{j=1}^{q}\lambda_{ij}^2$ is the *total communality*, and $\delta^2 = \Sigma_{i=1}^{p}\delta_i^2$.

Second, from Eq. 10, it is clear that any orthogonal transformation (rotation) of the factors $\mathbf{f}$ will still satisfy the constraint on the covariance structure as specified by Eq. 10. In fact, any transformation from $\mathbf{f}$ and Λ to $\mathbf{f}_1 = \mathbf{A}\mathbf{f}$ and $\Lambda_1 = \Lambda\mathbf{B}$, where $\mathbf{A}$ and $\mathbf{B}$ are $q \times q$ matrices such that their product $\mathbf{BA}$ is orthogonal, will satisfy the same constraint. In this general transformation, however, although $\mathbf{f}$ may be uncorrelated and standardized, the derived set $\mathbf{f}_1$ need not have either property. If one wishes to remain with standardized uncorrelated factors, the choice of $\mathbf{A}$ to be orthogonal and $\mathbf{B} = \mathbf{A}'$ will suffice. The indeterminacy implied by such transformations in any "solution" obtained for the factor loadings is used to advantage in the so-called practice of *rotating* a preliminary solution to obtain a more interpretable final solution. Further discussion of the issues and procedures involved in rotation is available in Harman (1967).

Without any loss of generality, the original variables may be assumed to be standardized ($\sigma_{ii} = 1$ for $i = 1, 2, \ldots, p$) so that, with standardized uncorrelated factors, Eq. 10 specifies the following structural representation of the $p \times p$ correlation matrix, $\Gamma = ((\rho_{il}))$:

$$\left.\begin{aligned} &\Gamma = \Lambda\Lambda' + \Delta, \\ \text{or}\quad & \\ &\rho_{ii} = 1 = h_i^2 + \delta_i^2 \qquad \text{for } i = 1, \ldots, p, \\ \text{and}\quad & \\ &\rho_{il} = \sum_{j=1}^{q} \lambda_{ij}\lambda_{lj} \qquad \text{for } i \neq l. \end{aligned}\right\} \qquad (14)$$

The $p \times p$ matrix Γ^*, whose diagonal elements are the communalities h_i^2 and off-diagonal elements are the correlation coefficients ρ_{il} between pairs of the observed variables, is called the *reduced correlation matrix*. This matrix plays an important role in the principal factor method of determining Λ and has the following properties: (i) as a consequence of Eq. 14, if every $\delta_i^2 > 0$, the diagonal elements of Γ^* are all less than 1; (ii) the rank of Γ^* = the minimum number of linearly independent factors required for reproducing the correlations among the observed variables = the dimensionality of the factor space; (iii) if $h_i^2 = 1$ for all i, the rank of $\Gamma^* = p$, and no reduction of dimensionality is accomplished by factor analysis.

The previously stated problem (*a*) of factor analysis may now be restated as follows: given the intercorrelations among a set of p observed responses, choose the set $\{h_i^2\}$ so as to minimize the rank of Γ^*.

The two methods of determining Λ under the model of Eq. 10 (or, equivalently, Eq. 14) are described next.

The Principal Factor Method. This method was proposed by Thurstone (1931)[†] and more fully described by Thomson (1934). It should not be confused with the principal components method described in Section 2.2.1, and the similarities and differences of the two techniques are discussed later.

From Eq. 12 it follows that for any factor, f_j, its contribution to the communality of the variable y_i is λ_{ij}^2. Hence the contribution of f_j to the communalities of all p observed responses is

$$V_j = \sum_{i=1}^{p} \lambda_{ij}^2 = \boldsymbol{\lambda}_j'\boldsymbol{\lambda}_j, \tag{15}$$

where $\boldsymbol{\lambda}_j$ denotes the jth column of Λ, and $j = 1, 2, \ldots, q$. The total communality defined by Eq. 13 is, of course, $V = \Sigma_{j=1}^{q} V_j$.

The principal factor method involves, as the first stage, choosing the coefficients, $\lambda_{11}, \ldots, \lambda_{p1}$, of the first factor f_1 so as to maximize the contribution of f_1 to the total communality subject to the constraints on the correlation structure as summarized by Eq. 14. In other words, we wish to choose $\boldsymbol{\lambda}_1$ so as to maximize $V_1 = \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1$ subject to the $p(p+1)/2$ constraints implied by $\boldsymbol{\Gamma}^* = \Lambda\Lambda'$.

The constrained maximization turns out to be equivalent to finding the eigenvalues and eigenvectors of $\boldsymbol{\Gamma}^*$ (see Harman 1967, for details); in fact, the maximum value of V_1 is the largest eigenvalue of $\boldsymbol{\Gamma}^*$, and the required maximizing value of $\boldsymbol{\lambda}_1$ is just proportional to the corresponding eigenvector. Thus, if γ_1 is the largest eigenvalue of $\boldsymbol{\Gamma}^*$ and $\boldsymbol{\alpha}_1$ is the corresponding eigenvector, which is normalized so that $\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = 1$, then

$$\boldsymbol{\lambda}_1 = \sqrt{\gamma_1}\cdot\boldsymbol{\alpha}_1, \tag{16}$$

and the maximum value of $V_1 = \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1 = \gamma_1$. This "solution" to the problem of determining $\boldsymbol{\lambda}_1$ is, however, artificial or circular in that the diagonal elements of $\boldsymbol{\Gamma}^*$ in turn involve $\lambda_{11}^2, \ldots, \lambda_{p1}^2$. The assumption is that the diagonal elements, namely, the communalities h_i^2's are independently known or specifiable. A method of specifying these is mentioned later.

At the second stage of the principal factor method, having determined $\boldsymbol{\lambda}_1$ as above, one seeks to determine $\boldsymbol{\lambda}_2$ so as to maximize $V_2 = \boldsymbol{\lambda}_2'\boldsymbol{\lambda}_2$ subject to a constraint on the residual reduced correlations after removal of the first factor. If

$$\boldsymbol{\Gamma}_1^* = \boldsymbol{\Gamma}^* - \boldsymbol{\lambda}_1\boldsymbol{\lambda}_1' \tag{17}$$

[†] Thurstone credits Walter Bartky with the mathematical solution associated with the technique. It may be of interest that this is the same Bartky, an astronomer, who is also said to have been the originator of the ideas of sequential sampling.

denotes the $p \times p$ matrix of residual reduced correlations after removing f_1, then the constraints are that

$$\mathbf{\Gamma}_1^* = [\boldsymbol{\lambda}_2 \cdots \boldsymbol{\lambda}_q]\begin{bmatrix} \boldsymbol{\lambda}_2' \\ \vdots \\ \boldsymbol{\lambda}_q' \end{bmatrix}. \tag{18}$$

The present constrained maximization problem, however, is of the same mathematical form as the one at the first stage of the method. Hence the required solution for $\boldsymbol{\lambda}_2$ is proportional to the eigenvector of $\mathbf{\Gamma}_1^*$ corresponding to its largest eigenvalue. An eigenanalysis of $\mathbf{\Gamma}_1^*$ is, however, not essential since the required solution for $\boldsymbol{\lambda}_2$ may be shown to be equivalent to choosing $\boldsymbol{\lambda}_2$ proportional to the eigenvector associated with the second largest eigenvalue of the original reduced correlation matrix, $\mathbf{\Gamma}^*$. Arguments for establishing this computationally convenient result follow.

If $\boldsymbol{\alpha}_k$ is the eigenvector of $\mathbf{\Gamma}^*$ corresponding to the eigenvalue γ_k $(k = 1, 2, \ldots, p)$, where $\gamma_1 \geqslant \gamma_2 \geqslant \cdots$, then

$$\begin{aligned} \mathbf{\Gamma}_1^*\boldsymbol{\alpha}_k = (\mathbf{\Gamma}^* - \boldsymbol{\lambda}_1\boldsymbol{\lambda}_1')\boldsymbol{\alpha}_k &= (\mathbf{\Gamma}^* - \gamma_1\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1')\boldsymbol{\alpha}_k \\ &= \gamma_k\boldsymbol{\alpha}_k - \gamma_1\boldsymbol{\alpha}_1\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_k. \end{aligned} \tag{19}$$

Using the orthonormal property of the set of eigenvectors $\{\boldsymbol{\alpha}_k\}$, it follows from Eq. 19 that $\boldsymbol{\alpha}_1$ is an eigenvector of $\mathbf{\Gamma}_1^*$ corresponding to the eigenvalue zero and that, for $k = 2, 3, \ldots, p, \gamma_k$ is an eigenvalue of $\mathbf{\Gamma}_1^*$ with associated eigenvector $\boldsymbol{\alpha}_k$. In particular, the largest eigenvalue of $\mathbf{\Gamma}_1^*$ is γ_2, the second largest eigenvalue of $\mathbf{\Gamma}^*$, and the corresponding eigenvector is $\boldsymbol{\alpha}_2$. The required $\boldsymbol{\lambda}_2$ is $\sqrt{\gamma_2}\cdot\boldsymbol{\alpha}_2$.

The remaining stages of the principal factor method now follow in exactly the same manner. Finding $\boldsymbol{\lambda}_3, \ldots, \boldsymbol{\lambda}_q$ so as to maximize the contributions of each corresponding factor to the total communality, subject to constraints on the residual reduced correlations at each stage, turns out to be equivalent to taking $\boldsymbol{\lambda}_j = \sqrt{\gamma_j}\cdot\boldsymbol{\alpha}_j$, for $j = 3, \ldots, q$.

The descriptions above have been presented in terms of population characteristics. With data one would use a sample reduced correlation matrix, $\mathbf{R}^*$, in place of $\mathbf{\Gamma}^*$ as the input to the eigenanalysis, where

$$\mathbf{R}^* = \begin{pmatrix} \hat{h}_1^2 & r_{12} & \cdots & r_{1p} \\ & \hat{h}_2^2 & \cdots & r_{2p} \\ & & & \vdots \\ & & \ddots & \\ & & & \hat{h}_p^2 \end{pmatrix}, \tag{20}$$

and $\{\hat{h}_i^2\}$ are communalities estimated from the data, while r_{il} is the correlation coefficient between the ith and lth responses as calculated from the data. Thus, with data, the principal factor solution for Λ is

$$\hat{\Lambda} = \mathbf{L} = ((l_{ij})) = [\mathbf{l}_1 \mathbf{l}_2 \cdots \mathbf{l}_q], \tag{21}$$

where $\mathbf{l}_j = \sqrt{c_j}\mathbf{a}_j$ for $j = 1, \ldots, q$, and $c_1 > c_2 > \cdots > c_q > 0$ are the q largest eigenvalues of $\mathbf{R}^*$ with corresponding eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_q$. Also, the estimate of δ_i^2 is

$$\hat{\delta}_i^2 = 1 - \sum_{j=1}^{q} l_{ij}^2, \qquad i = 1, \ldots, p. \tag{22}$$

The procedure described above is complete except for specification of $\hat{h}_i^2$'s. If $\hat{h}_i^2 = 1$ for $i = 1, \ldots, p$, then $\mathbf{R}^* = \mathbf{R}$, the ordinary correlation matrix of the responses, and the principal factor solution is exactly the same as the principal components solution for $\mathbf{R}$. Estimates of h_i^2 whose values are less than 1 include: (i) $\hat{h}_i^2$ = the highest observed positive correlation of variable y_i with the remaining $(p - 1)$ variables = the largest positive element in the ith row (column) of $\mathbf{R}$; (ii) $\hat{h}_i^2$ = the average (presumed positive) of the observed correlations of y_i with the other variables $= \Sigma_{\substack{l=1 \\ (l \neq i)}}^{p} r_{il}/(p - 1)$; (iii) $\hat{h}_i^2$ = square of the multiple correlation coefficient of y_i with the other variables $= 1 - (1/r^{ii})$, where $((r^{ij})) = \mathbf{R}^{-1}$; and (iv) iterative estimates obtained by starting with an arbitrary set of values for $\hat{h}_i^2$'s to get a principal factor solution, thence using the sums of squares of the factor loadings for each variable in such a solution as the new values of $\hat{h}_i^2$'s, and repeating the process until the sets of successive estimates do not differ greatly. An intuitive basis for the third of these choices is that the squared multiple correlation coefficient measures the proportion of the observed total variability in a specific response that is accounted for by its regression on the remaining $(p - 1)$ responses, and hence provides a measure of common or shared variance. A second reason for this choice is that, while 1 is an upper bound on $\hat{h}_i^2$, it can be shown that the squared multiple correlation coefficient involved is a lower bound. Many of the computer programs for performing a principal factor analysis use this choice for $\hat{h}_i^2$ as the standard one. In practice, except for small values of p (≤ 10), the different choices of $\hat{h}_i^2$ do not seem in general to lead to noticeably different outcomes.

If some or all of the diagonal elements of $\mathbf{R}^*$ are less than 1, then $\mathbf{R}^*$ need not be positive semidefinite. Hence some of the eigenvalues, $\{c_j\}$, may be negative with the consequence that the vectors of factor loadings, $\mathbf{l}_j$'s, associated with these will be imaginary. In practice, one discards these negative eigenvalues and the associated imaginary vectors of loadings. In fact, since the sum of the eigenvalues of $\mathbf{R}^*$ equals the total communality, the sum of just the positive eigenvalues will exceed the total communality if there are any negative eigenvalues at all. Hence, in extracting the factors, one would not proceed until their number q was as large as the number of positive eigenvalues but, rather, would stop when $\Sigma_{j=1}^{q} c_j$ was close to tr($\mathbf{R}^*$), the total communality.

Another useful procedure for guiding the choice of a value for q is to compute and study the residual correlation matrix after each factor has been fitted. Although all the numerical computations may be carried out on $\mathbf{R}^*$, from the standpoint of interpretation it is useful to compute the matrix of residual correlation coefficients,

$$\mathbf{R}_j^* = \mathbf{R}^* - \sum_{a=1}^{j} \mathbf{l}_a \mathbf{l}_a',$$

at the jth stage for $j = 1, 2 \ldots$.

Mention has been made that, when $\hat{h}_i^2$, for all i, are taken to be unity, the principal factor method is identical with a principal components analysis of the correlation matrix. In fact, if the communalities $\hat{h}_i^2$'s (or, equivalently, the $\hat{\delta}_i^2$'s) are all essentially equal and q is close to p, the principal factor method as described above and a principal components analysis of $\mathbf{R}$ would both lead to very similar results. The reason is that, if $\hat{\delta}_i^2 = d^2$ for all i, then $\mathbf{R}^* = \mathbf{R} - d^2\mathbf{I}$ and $c_j = b_j - d^2$ is an eigenvalue of $\mathbf{R}^*$ if b_j is an eigenvalue of $\mathbf{R}$. Hence the relationship $\mathbf{R}^*\mathbf{a}_j = c_j\mathbf{a}_j$ is equivalent to $(\mathbf{R} - d^2\mathbf{I})\mathbf{a}_j = (b_j - d^2)\mathbf{a}_j$, or to $\mathbf{R}\mathbf{a}_j = b_j\mathbf{a}_j$, so that the eigenvectors $\{\mathbf{a}_j\}$ of $\mathbf{R}^*$ are also those of $\mathbf{R}$. In practice, however, the $\hat{\delta}_i^2$'s (and $\hat{h}_i^2$'s) are often unequal and $q \ll p$, so that the principal factor method may lead to results that are different from those obtained by a principal components analysis of the correlation matrix. The use of values of $\hat{h}_i^2$ less than 1 has an interesting interpretation in terms of an idea utilized in ridge regression (Theil, 1963; Hoerl & Kennard, 1970). In multiple regression analysis (see Sections 3.3 and 5.2.1), the sum-of-products matrix of the independent variables may be nearly singular in some applications, perhaps because of round-off errors or high intercorrelations amongst the independent variables. The latter cause is referred to as "multicollinearity" in the econometric literature. The near singularity leads not only to numerical difficulties but also to estimates of regression coefficients that have undesirable statistical properties. The idea of ridge regression for handling this problem is to add a constant multiple of the identity matrix to the sum-of-products matrix (i.e., add a positive constant to each of the diagonal elements of the latter) and to utilize the resultant matrix in place of the nearly singular matrix. Thus, in ridge regression, a nearly singular covariance or correlation matrix is adjusted to become "more" positive definite (see also Section 5.2 and Devlin et al., 1975) by increasing the diagonal elements, whereas decreasing the diagonal elements involved in the principal factor method will have a "deridging" effect. An implication of this, and also of the discussion in the preceding paragraph, is that there is a strong implicit commitment in the principal factor method to a linear model for reducing dimensionality.

Indeed, some authors have tried to distinguish between the principal components and principal factor techniques on the grounds of their respective degrees of commitment to a linear model. This, however, does not seem to be a crucial distinction since both techniques have implicit, as well as explicit, linearity considerations underlying them, and both tend to be inadequate in the face of nonlinearity (see Examples 2–4). Perhaps a better distinction to be made is that the factor analysis model (Eqs. 7, 9, 10) is more explicit than the one underlying principal components in assuming a space (linear) of *reduced* dimensionality (i.e., $q \ll p$) for explaining the correlation structure of the original responses.

The Maximum Likelihood Method. This method, originally proposed by Lawley (1940), has received considerable attention from statisticians (see Anderson & Rubin, 1956; Howe, 1955), perhaps because of its usage of the criterion of maximizing a likelihood function, which is a familiar concept and method in statistics.

The assumption in this method is that the observations (viz., the columns of the $p \times n$ matrix Y of Eq. 8) constitute a random sample from a nonsingular p-variate normal distribution whose covariance matrix Σ_{yy} has a structure specified by Eq. 10. Furthermore, if $p \leqslant (n-1)$, the sample covariance matrix, S, will be nonsingular with probability 1 and will have a Wishart distribution. Using the Wishart density as a starting point, one obtains the log-likelihood function of Λ and Δ,

$$\mathscr{L}(\Lambda, \Delta \mid \mathrm{S}) = -\frac{n-1}{2}\left[\ln|\Lambda\Lambda' + \Delta| + \mathrm{tr}\{(\Lambda\Lambda' + \Delta)^{-1}\mathrm{S}\}\right]. \tag{23}$$

Hence maximizing $\mathscr{L}$ with respect to the elements of Λ and Δ is equivalent to minimizing $\ln|\Lambda\Lambda' + \Delta| + \mathrm{tr}\{(\Lambda\Lambda' + \Delta)^{-1}\mathrm{S}\}$, and the resulting values, $\hat{\Lambda}$ and $\hat{\Delta}$, are the required maximum likelihood estimates.

The indeterminacy of $\hat{\Lambda}$ up to rotation is handled in maximum likelihood estimation by imposing the constraint that the matrix

$$\hat{\mathrm{J}} = \hat{\Lambda}'\hat{\Delta}^{-1}\hat{\Lambda} \tag{24}$$

be diagonal. This constraint simplifies the solving of the likelihood equations. The actual equations that need to be solved by iterative methods are

$$\hat{\mathrm{J}}\hat{\Lambda}' = \hat{\Lambda}'\hat{\Delta}^{-1}(\mathrm{S} - \hat{\Delta}), \qquad \hat{\Delta} = \mathrm{diag}(\mathrm{S} - \hat{\Lambda}\hat{\Lambda}'), \tag{25}$$

where $\hat{\mathrm{J}}$ is defined by Eq. 24, and diag(M) denotes a diagonal matrix whose diagonal elements are those of the square matrix **M** (for details see Lawley & Maxwell, 1963; Howe, 1955).

Accumulated early experience with attempts at solving the likelihood equations (Eq. 25) indicated that convergence to a solution may be a serious problem. Subsequently, Jöreskog (1967) and Lawley (1967) developed numerical approaches for obtaining the maximum likelihood estimates that seem to circumvent this difficulty (see also the expository treatment by Jöreskog & Lawley, 1968). The modifications have two basic features. First, instead of solving the likelihood equations, a direct numerical maximization of the function $\mathscr{L}(\Lambda, \Delta \mid S)$ (or, equivalently, a minimization of $-[2/(n-1)]\ \mathscr{L}(\Lambda, \Delta \mid S)$) is attempted. Second, the maximization (or equivalent minimization) is carried out in two parts—for a given Δ find a Λ_Δ that maximizes $\mathscr{L}(\Lambda, \Delta \mid S)$, and then determine $\hat{\Delta}$ as that value of Δ which maximizes the function $\mathscr{L}_{\max}(\Delta) = \mathscr{L}(\Lambda_\Delta, \Delta \mid S)$. The required maximum likelihood estimates are $\hat{\Delta}$ and $\hat{\Lambda} = \Lambda_{\hat{\Delta}}$. The iterative scheme (see Jöreskog, 1967, for details) appears to work well, primarily because the determination of Λ_Δ for a given Δ is quite straightforward in that it simply involves the determination of the q largest eigenvalues and the associated eigenvectors of the matrix $\Delta^{-1/2} \cdot S \cdot \Delta^{-1/2}$. A general iterative numerical optimization technique (e.g., Fletcher & Powell, 1963) is then needed only at the second stage of maximization, namely, the maximization of $\mathscr{L}_{\max}(\Delta)$ with respect to the p diagonal elements of Δ.

An implication of Eq. 25 is that scaling any observed variable would induce proportional changes in the estimates of the factor loadings for that variable. Independence, in this sense, of the solution of Eq. 25 from the scales of the original variables has the numerical consequence that one may employ the covariance matrix or the correlation matrix of the original variables in seeking the solution. It should be recognized, however, that if one were to use the sample correlation matrix as the starting point, the Wishart density would not provide the basis of the initial likelihood function. The numerical implication of "invariance" of the solution $\hat{\Lambda}$ of Eq. 25 is, therefore, unrelated to the statistical considerations that underlie the formulation in terms of maximum likelihood estimation. The latter appears to be feasible only in terms of the sample covariance matrix.

A statistical advantage claimed for the maximum likelihood approach is that the asymptotic (viz., n large, $p < n$, and $q \ll p$) properties of such estimates are known and may be used for purposes of statistical inference (see Lawley, 1940; Anderson & Rubin, 1956). In particular, for example, one can obtain a likelihood ratio test for the adequacy of the hypothesized number, q, of common factors. The essential result derived by Lawley (1940) is that the observed value of the statistic

$$(n-1)\left\{\ln\left[\frac{|\hat{\Lambda}\hat{\Lambda}' + \hat{\Delta}|}{|S|}\right] + \mathrm{tr}[(\hat{\Lambda}\hat{\Lambda}' + \hat{\Delta})^{-1}S] - p\right\} \qquad (26)$$

may be referred to a chi-squared distribution with ν degrees of freedom, where $\hat{\Lambda}$ and $\hat{\Delta}$ are the estimates that satisty Eq. 25, S is the sample covariance matrix, and

$$\nu = \tfrac{1}{2}[(p-q)^2 - p - q]. \qquad (27)$$

Table 1. Comparison of Two Methods of Factor Analysis

Feature	Principal Factor Method	Maximum Likelihood Method
1. Estimates of communalities	Required	Not required
2. Dimensionality of common factors space	Inferable from manner of computing	Assumed for obtaining a solution but then may be statistically tested for adequacy
3. Distributional assumption	None specific	Multivariate normal
4. Formal statistical inference status	Not much is known	Large-sample theory is available
5. Iteration for obtaining the solution	Optional (i.e., not required unless one chooses to estimate communalities iteratively)	Necessary
6. Convergence of iterative procedure	Good	May be poor for solving Eq. 25, but modified method seems good
7. Scale "invariance"	No	Yes

[*Note*: For a given p, q has to be $\ll p$ for ν to be positive.] Statistically large observed values of the statistic imply that the number of common factors needed to adequately reproduce the correlations among the original variables is larger than q. Bartlett (1951) has suggested using the multiplicative factor $\{n - p/3 - 2q/3 - 11/6\}$ in place of $(n - 1)$ in Eq. 26 for improving the chi-squared approximation.

Table 1 provides a summary comparison of features of the principal factor and maximum likelihood methods.

A useful graphical technique, associated with both methods of factor analysis, is to represent the original variables in terms of their factor loadings in a space that corresponds to the common factors. Thus, using pairs (and/or triplets) of axes, one obtains p points whose coordinates are factor loadings with respect to pairs (and/or triplets) of the common factors. Such plots can often aid in interpreting the nature of the factors, as well as in suggesting "rotations" to more meaningful sets of coordinates for the factors.

A common practice in using factor analysis is to seek so-called "estimates" of the *factor scores*. In the notation of Eq. 8, an estimate, $\hat{\mathbf{F}}$, of $\mathbf{F}$ is desired. By analogy with multiresponse regression (see Sections 3.3 and 5.2.1) and considering the factors as the regression variables with the initial variables as the regressors, the desired estimate is defined by $\hat{\mathbf{F}} = \hat{\mathbf{\Lambda}}'\mathbf{R}^{-1}\mathbf{Z}$, where $\mathbf{Z} = \mathbf{D}_{1/\sqrt{s_{ii}}}(\mathbf{Y} - \bar{\mathbf{Y}})$ is the $p \times n$ matrix of standardized data.

Although factor analysis has been used most extensively as a tool in psychology and the social sciences, applications have been made to other fields as well. Seal (1964) summarizes various biological applications of factor analysis, and Imbrie (1963), Imbrie & Van Andel (1964), and Imbrie & Kipp (1971) have used it in analyzing certain geological problems.

Some applications of factor analysis, especially in the social sciences, raise questions concerning its usefulness for achieving parsimony of description or for incisively understanding a complex of observed variables in terms of a few underlying variables. Often in questionnaire survey data, for example, built-in or a priori groupings of the initial variables are the ones that are uncovered by using factor analysis. Even in such examples, however, the technique is perhaps useful in that it provides a more quantitative understanding of the qualitative prior groupings.

A different issue related to the usefulness of the method is its inadequacy in the face of nonlinearity of the underlying relationships. The work of McDonald (1962, 1967) and of Carroll (1969) is directed toward nonlinear factor analysis methodology. Other nonlinear techniques are considered next in this chapter.

2.3. NONMETRIC METHODS FOR NONLINEAR REDUCTION OF DIMENSIONALITY

A class of procedures, collectively designated as *multidimensional scaling techniques*, has been developed in connection with the following problem: given a set of observed measures of similarity or dissimilarity between every pair of *n* objects, find a representation of the objects as points in Euclidean space such that the interpoint distances in some sense "match" the observed similarities or dissimilarities. Some examples of measures of similarity are (i) confusion probabilities or the proportion of times one stimulus is identified as another among *n* stimuli, (ii) the absolute value of a correlation coefficient, and (iii) any index of profile similarity.

Several approaches have been proposed (see Coombs, 1964, for a general discussion) to the problem of multidimensional scaling, but for present purposes, only the technique developed by Shepard (1962a, b) and further refined by Kruskal (1964a b) is considered. A central feature of the Shepard-Kruskal approach is the specification of monotonicity as the sense in which interpoint distances are to match the observed dissimilarities among the objects; that is, the larger the specified dissimilarity between two objects, the larger should the interpoint distance be in the Euclidean representation of these objects. Kruskal (1964a, b) not only developed efficient algorithms for using the method but also proposed an explicit measure for judging the degree of conformity to monotonicity in any solution. Furthermore, as an integral part of their approach, Shepard and Kruskal obtain a graphical display of the data-determined monotone relationship between dissimilarity and distance (see details below).

Another important characteristic of the approach, demonstrated empirically by Shepard (1962a), is that one could start just from the nonmetric rank order information about the dissimilarities and still obtain quite "tightly determined" configurations. The technique exploits nonmetric information, when enough of it is available, to derive metric representations of the data (see also Abelson & Tukey, 1959). In this respect it is an interesting example of a data-analytic technique with a counterobjective to that underlying the practice of replacing metric observations by their ranks as a prelude to employing some nonparametric statistical procedures.

As motivation for the concepts and procedures involved in the Shepard-Kruskal approach, consider the case wherein one has four $(=n)$ objects and six observed values of dissimilarity for the six possible pairs of the objects. If δ_{ij} denotes the dissimilarity value for the pair of objects i and j, for $i, j = 1, 2, 3, 4$, then suppose, for example, that the following rank ordering among the six observed dissimilarity values holds:

$$\delta_{23} < \delta_{12} < \delta_{34} < \delta_{13} < \delta_{24} < \delta_{14}. \tag{28}$$

In other words, the second and third objects are judged to be least dissimilar (or most similar), the first and second objects next least dissimilar, and so on, with objects 1 and 4 ranked as most dissimilar (or least similar). Suppose that the objects are represented as points in a Euclidean space of a specified dimensionality, and let $\mathbf{y}_i$ denote the column vector of coordinates of the point corresponding to the ith object, $i = 1, \ldots, 4$. Then the familiar unweighted Euclidean distance between the points representing objects i and j is

$$d_{ij} = [(\mathbf{y}_i - \mathbf{y}_j)'(\mathbf{y}_i - \mathbf{y}_j)]^{1/2}, \tag{29}$$

$i < j = 2, 3, 4$. The monotonicity constraint, which is central in the approach, is said to be met perfectly in this simple example if, corresponding to the ordering of the observed dissimilarities shown in Eq. 28, the d_{ij}'s calculated by using Eq. 29 turn out to satisfy the following:

$$d_{23} \leqslant d_{12} \leqslant d_{34} \leqslant d_{13} \leqslant d_{24} \leqslant d_{14}. \tag{30}$$

In other words, the order relationship among the interpoint distances in the Euclidean representation of the objects is in exact concordance with the order relationship among the observed dissimilarities. Such a perfect match may, of course, not hold in a particular Euclidean representation, and one then needs both a measure to evaluate the closeness of match and a method of determining a configuration so as to achieve as close a match as possible.

A graphical representation that facilitates understanding the explicit measure of monotonicity and the mode of analysis proposed by Kruskal (1964a, b) is a scatter plot of points whose coordinates are (d_{ij}, δ_{ij}). Thus, in the above

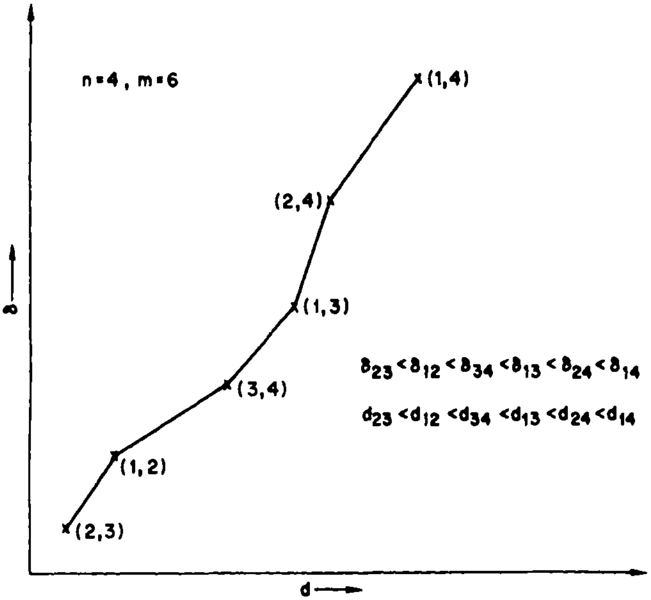


Fig. 2*a*. Illustrative scatter plot of dissimilarities versus distances, wherein monotonicity constraint is satisfied.

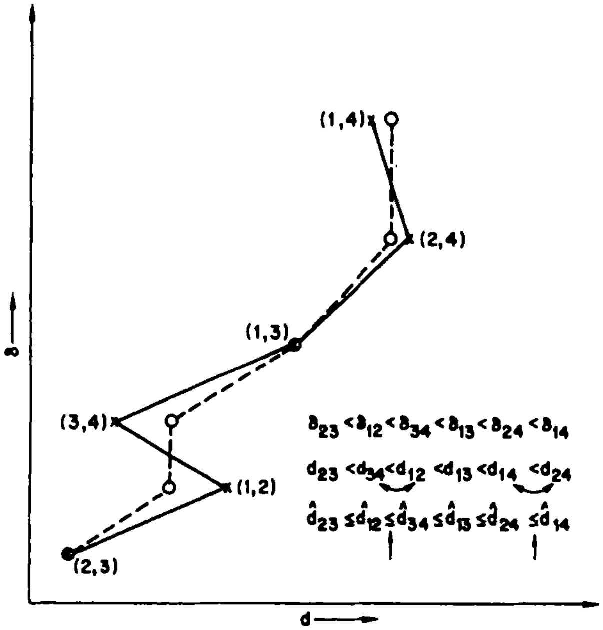


Fig. 2*b*. Illustrative scatter plot of dissimilarities versus distances, wherein monotonicity constraint is not satisfied.

simple example, one can obtain a plot of six points as shown, for instance, in Figure 2*a* by the crosses, which are labeled by the pair of object numbers to which each of them corresponds. Corresponding to the perfect monotonicity implied by Eqs. 28 and 30, the configuration of the crosses is such that the line segments joining the points form a chain in which, as one moves upward, one moves always to the right as well.

Complete conformity to monotonicity is always achievable by using a representation in a space of sufficiently high ($\geqslant n - 1$ with n objects) dimensionality. The primary interest, therefore, is to find a low-dimensional representation in which conformity to monotonicity is achieved to a reasonable degree if not perfectly.

Thus, suppose that, in the same example, one has a Euclidean representation of the four objects in which the ordered interpoint distances turn out to be

$$d_{23} < d_{34} < d_{12} < d_{13} < d_{14} < d_{24}. \tag{31}$$

The monotonicity constraint is now violated in that the order relationship between the interpoint distances between objects 3 and 4 and between objects 1 and 2 is not the same as the order relationship between the corresponding observed dissimilarities as specified by Eq. 28. There is also a violation of monotonicity by the interpoint distances between objects 1 and 4 and between objects 2 and 4. The scatter plot of (d_{ij}, δ_{ij}) corresponding to this situation is shown by the $\times$'s in Figure 2*b*, and the chain of lines joining the points is now observed to zigzag instead of always moving to the right as one moves upward. In this situation, one may wish to "fit" a set of values, $\hat{d}_{ij}$'s, such that the fitted values will indeed satisfy the monotonicity constraint so that

$$\hat{d}_{23} \leqslant \hat{d}_{12} \leqslant \hat{d}_{34} \leqslant \hat{d}_{13} \leqslant \hat{d}_{24} \leqslant \hat{d}_{14}, \tag{32}$$

corresponding to the order relationship in Eq. 28. A satisfactory set of fitted values in this example would be the abscissa values of the "fitted" points which are shown as o's and joined by dashed line segments in Figure 2*b*. Notice that only the $\hat{d}$ values for distances that did not conform to monotonicity are different from the corresponding d values. In fact, in the example the $\hat{d}$ value for both d_{12} and d_{34}, which, for instance, violate monotonicity, is just the average of d_{12} and d_{34}, that is, $\hat{d}_{12} = \hat{d}_{34} = (d_{12} + d_{34})/2$; and, similarly, $\hat{d}_{24} = \hat{d}_{14} = (d_{14} + d_{24})/2$. Apart from conforming to monotonicity, however, these fitted values may not, in fact, be distances in the sense that there is a configuration of points in Euclidean space whose interpoint distances are these values.

One measure for assessing the fit (viz., the conformity to monotonicity) of any proposed configuration is the sum of squares of deviations, $\Sigma_{i<j}(d_{ij} - \hat{d}_{ij})^2$. This measure of goodness of fit, although invariant under shifts (translations), reflections, and rotations (orthogonal transformations) of the coordinates in the Euclidean representation of the objects, is not invariant under uniform

shrinking or dilation. Hence Kruskal (1964a) proposed the following normalized measure of goodness of fit:

$$S = \left[\frac{\sum_{i<j} d_{ij} - \hat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2} \right]^{1/2}, \tag{33}$$

which he called the *stress*.

The stress may now be used as the basis for a systematic method of obtaining the fitted values. Given a set of d_{ij}'s, in fact, one may choose the $\hat{d}_{ij}$'s so as to minimize S subject to the constraint that they are to be monotone nondecreasing with the observed dissimilarity values, δ_{ij}'s. This minimization problem is equivalent to so-called *monotone least squares* or *monotone regression* and has been considered conceptually and algorithmically in other contexts of statistical applications (see Bartholomew, 1959; Miles, 1959; Barton & Mallows, 1961; van Eeden, 1957a, b). To avoid further notation, it is assumed that the fitted values, $\hat{d}_{ij}$'s, are in fact always obtained by this process of minimization, and the stress of a given configuration representing the initial objects in a Euclidean space is the value of S given by Eq. 33, using such fitted values. This value of S, of course, depends on the given configuration, and one may wish to make this relationship clearer by denoting it as $S(\mathbf{y}_1, \mathbf{y}_2, \ldots)$, where $\mathbf{y}_i$ is the vector of coordinates of the point corresponding to the ith object.

The next step is to determine the "best" configuration in a Euclidean space of specified dimensions. Such a configuration is one in the space of specified dimensionality whose stress is a minimum among all configurations in that space, that is, one wishes to determine $(\mathbf{y}_1^*, \mathbf{y}_2^*, \ldots, \mathbf{y}_n^*)$ so that

$$S(\mathbf{y}_1^*, \mathbf{y}_2^*, \ldots, \mathbf{y}_n^*) = \min_{\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}} S(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n).$$

Viewed as a trial and error process, what is involved is to start with a trial configuration, and then if $\hat{d}_{ij} < d_{ij}$ to move $\mathbf{y}_i$ and $\mathbf{y}_j$ closer, or if $\hat{d}_{ij} > d_{ij}$ to move $\mathbf{y}_i$ and $\mathbf{y}_j$ apart, so that in either case one is attempting to make d_{ij} resemble $\hat{d}_{ij}$ more closely. A systematic approach to this problem is provided by considering $S(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ as a function of the coordinates of all n points (i.e., a function of $n \times t$ variables if one is using a space of t dimensions for the representation), and then using a general numerical technique of function optimization, such as steepest descent, for determining the location of the minimum value of $S(\mathbf{y}_1, \ldots, \mathbf{y}_n)$.

Next, there is the issue of the choice of the dimensionality for the Euclidean representation. If $S_0(t) = S(\mathbf{y}_1^*, \mathbf{y}_2^*, \ldots, \mathbf{y}_n^*)$ denotes the minimum value of the stress associated with the minimum stress configuration in a t-dimensional space, Kruskal (1964a) suggests basing the choice of q, the minimum adequate value of t, on a study of a plot of $S_0(t)$ versus t. As t increases, $S_0(t)$ will decrease and, in fact, will be 0 for $t \geqslant (n-1)$. As general though not rigid benchmarks, Kruskal (1964a) proposes that a value of $S_0(t)$ of 20% be interpreted as suggesting a poor fit, 10% a fair fit, 5% a good one, and $2\frac{1}{2}$% an excellent one, with 0% being a perfect fit. In addition to these general guidelines, one may decide on a value for q by looking for an "elbow" in the plot of $S_0(t)$ versus t (see the discussion in Example 5).

The entire procedure can now be summarized in terms of the following steps:

1. For n objects, obtain the initial information, which is the rank ordering of the $m = n(n-1)/2$ dissimilarity values, δ_{ij}'s, among every pair of the n objects.

2. Given the m dissimilarity values, with the ordering

$$\delta_{i_1j_1} < \delta_{i_2j_2} < \cdots < \delta_{i_mj_m}, \tag{34}$$

and using some initial trial configuration of points $\mathbf{y}_{i0}$ $(i = 1, 2, \ldots, n)$, in $t(\geqslant 1)$ dimensions, determine the interpoint distances d_{ij}'s (see Eq. 29) and fit $\hat{d}_{ij}$'s so that

$$\hat{d}_{i_1j_1} \leqslant \hat{d}_{i_2j_2} \leqslant \cdots \leqslant \hat{d}_{i_mj_m}. \tag{35}$$

For a given configuration, the $\hat{d}_{ij}$'s are the chosen so as to minimize the stress S (see Eq. 33) subject to the monotonicity constraint, Eq. 35. The algorithm required here is that of so-called monotone regression.

3. Next, using these $\hat{d}_{ij}$'s and considering S as a function of the $n \times t$ coordinates of the n points in the representation, determine an improved configuration, $\{\mathbf{y}_{i1}\}$, and thence the new d_{ij}'s, $\hat{d}_{ij}$'s etc., until the best configuration in t dimensions is found as the one whose stress is $S_0(t) = \min\{S\}$, where the minimum is over all configurations in t dimensions. Steepest descent or some other general function minimization algorithm may be used for this step.

4. Finally, plot $S_0(t)$ versus t and choose q, the number of dimensions, as the minimum "adequate" value of t from the indications in such a plot.

For simplicity of exposition, the above discussion has assumed that the initially observed dissimilarity values are symmetric $(\delta_{ij} = \delta_{ji})$, that there are no ties among them, and that they are available for all possible pairs of the objects. Kruskal (1964a, b) suggests methods for handling asymmetries, ties, and missing observations, and also describes the details of the algorithms developed and implemented by him for using the technique.

Neither the final "best fitting" configuration nor the configuration at any stage is unique in that any similarity transformation (i.e., translation, rotation, reflection, or dilation) of the configuration will also have the same value of stress. In particular, if one so desires, one can rotate to the principal components axes of the configuration (see Section 2.2.1) and look at the projections of the points in the two- and three-dimensional spaces of pairs and triplets of these principal components axes.

The above ideas and methods of multidimensional scaling are directly relevant to reduction of dimensionality for a body of multiresponse data. Indeed, if n observed points in p-space are located close to a q-dimensional linear subspace, the use of the interpoint Euclidean distances of the points in the p-space as the initial measures of dissimilarity in multidimensional scaling could lead to a "solution" in q-space, in correspondence with the results of a principal components analysis of the original covariance matrix.

However, if the n points in p-space are located close to certain kinds of *curved* q-dimensional subspaces, multidimensional scaling may produce a solution in q-space which would not necessarily be indicated by the linear principal components analysis or usual factor analysis. The point is that multidimensional scaling attempts to preserve the monotone relation of distances, and, if the distances along the curved q-dimensional subspace are reasonably monotone with the Euclidean distances, the procedure will recognize the lower-dimensional curved space. For instance, in the oversimplified example of points lying on a semicircle, since the interpoint Euclidean distances (viz., chord lengths) are a strictly monotone function of distances measured along the curve (viz., arc lengths), multidimensional scaling will recover the spacing of the points along the unidimensional curve and a single dimension will be indicated as providing a perfect fit.

Example 2. This example derives from Ekman (1954), and the data, used by Shepard (1962b), consisted of similarity ratings by 31 subjects of every pair among 14 color stimuli, which varied primarily in hue. Thus $n = 14$ and $m = 91$ here. The subjective similarity rating of each pair by every subject was on a five-point scale, and the mean ratings from all 31 subjects were scaled to go from 0 ("no similarity at all") to 1 ("identical"). A 14 × 14 matrix of such mean similarity ratings was obtained and treated by Ekman (1954) as a correlation matrix for purposes of a factor analysis, which led to a five-factor description. The five factors were identified as violet, blue, green, yellow, and red. On the other hand, as mentioned by Shepard (1962b), intuition and the familiar concept of the "color circle" for representing colors differing in hue might suggest the reasonableness of a two-factor (or perhaps even a one-factor) solution. Even if experimentally unintended variations in "brightness" and "saturation" were involved in the subjective ratings, one would still expect three and not five factors.

Exhibit 2*a*, taken from Shepard (1962b), shows the two-dimensional solution obtained by a multidimensional scaling algorithm. [*Note*: The axes shown in the figure were obtained by rotation of the ones in the solution to principal axes; however, with the essentially circular configuration involved here this makes hardly any difference.] The multidimensional scaling solution, of course, consists merely of the coordinate representation of the 14 points, and the smooth line was drawn through the points by Shepard to emphasize the similarity of the configuration to the color circle.

Exhibit 2a. Multidimensional scaling solution for 14 colors (Ekman, 1954; Shepard, 1962b)

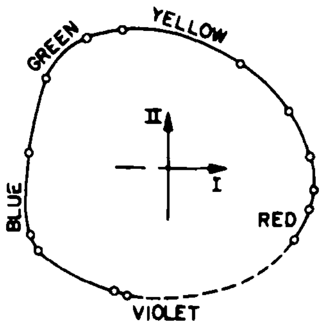
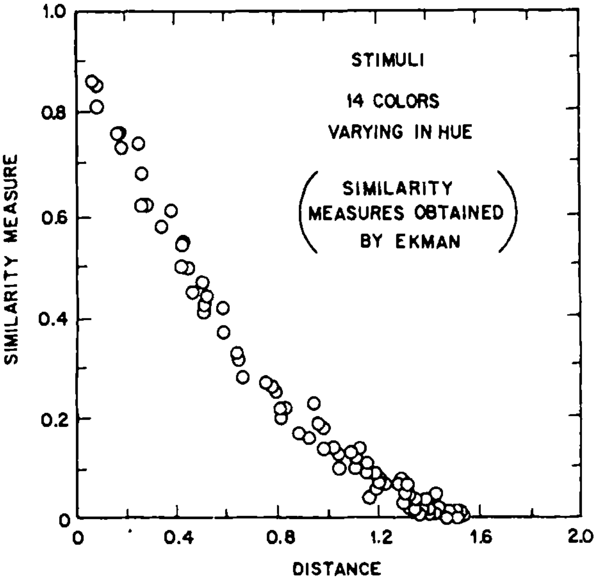


Exhibit 2*b* is a scatter plot of the original measures of similarity against the interpoint distances as computed from the two-dimensional solution shown in Exhibit 2*a*. The monotone relation between similarity and interpoint distance seen in this plot is, of course, a constraint of the multidimensional scaling procedure. The greater the observed similarity between two stimuli, the smaller is the distance between the two points representing the stimuli. The plot provides a graphical display of the data-determined monotone relationship involved, and in the present example it appears to be a relatively smooth, nonlinear (perhaps quadratic) relationship.

Exhibit 2*b*. Scatter plot of similarity versus distance for the example of Exhibit 2*a* (Shepard, 1962b)



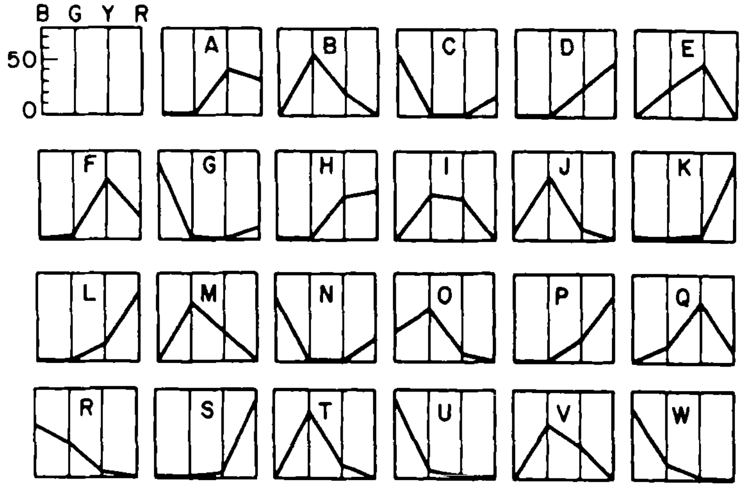
In this example, multidimensional scaling has produced both an intuitively appealing and a scientifically adequate parsimonious representation. It is perhaps reasonable to inquire about the details of the factor analysis solution and to try to understand the nature of and the reasons for the differences between the two solutions. In particular, after fitting the first two factors by the principal factor method, only 64% of the total variance (see Eq. 13) had been accounted for. Moreover, the residual correlations at that stage, that is, the off-diagonal elements of $\mathbf{R}_2^* = (\mathbf{R}^* - \mathbf{l}_1\mathbf{l}_1' - \mathbf{l}_2\mathbf{l}_2')$, where $\mathbf{R}^*$, $\mathbf{l}_1$, and $\mathbf{l}_2$ are defined by Eqs. 20 and 21, were still reasonably large. This implies that the original correlations (which in this example were, in fact, similarities) were not adequately reproducible from the two-factor solution. Shepard (1962b), however, fitted a quadratic to the plot shown in Exhibit 2*b*, obtained fitted values for similarities from such a quadratic, and demonstrated that such fitted values were adequate reproductions of (i.e., were sufficiently close in value to) the original measures of similarity. This is not surprising in view of the indication of "tightness" of the points about a quadratic which can be visualized in Exhibit 2*b*.

The main reasons for the difference in the dimensionalities suggested by the two methods is perhaps the inadequacy of the inherent linearity in the factor analysis approach to handle nonlinear reductions of dimensionality. In the present example, there may also be an effect on the factor analysis solution due to the use of similarity measures (with a range of 0 to 1) as inputs instead of the usual correlation coefficients (with a range of -1 to 1).

A modification of the scaling approach, due to Shepard & Carroll (1966), is directed toward improving the recognition of near singularities of a nonlinear nature among multidimensional observations. This modification focuses attention mainly on retaining the monotone relationship between interpoint distances and similarities only for nearby points rather than for all the points. The idea is illustrated by the next example, taken from Shepard & Carroll (1966).

Example 3. The data are from Boynton & Gordon (1965) and were used by Shepard & Carroll (1966) for illustrating the modified multidimensional scaling approach. The general concern and nature of the experiment that gave rise to the data are somewhat similar to those in the Ekman experiment described in Example 2, although the experimental detail and the nature of the data are different here. Specifically, 23 spectral colors differing only in their wavelengths were projected in random sequence several times to a group of observers. For each color the relative frequencies with which the observers denoted it as blue, green, yellow, or red were noted, thus giving a four-dimensional response associated with each color. In this example, $n = 23$ and $p = 4$. Exhibit 3*a* shows a pictorial representation of the data. The 23 colors (observations) are labeled *A* through *W*, and the four-dimensional vector of observations for each color is shown in a profile format. [*Note*: The four relative frequencies for any color are not required to add up to 100%, and they do not.]

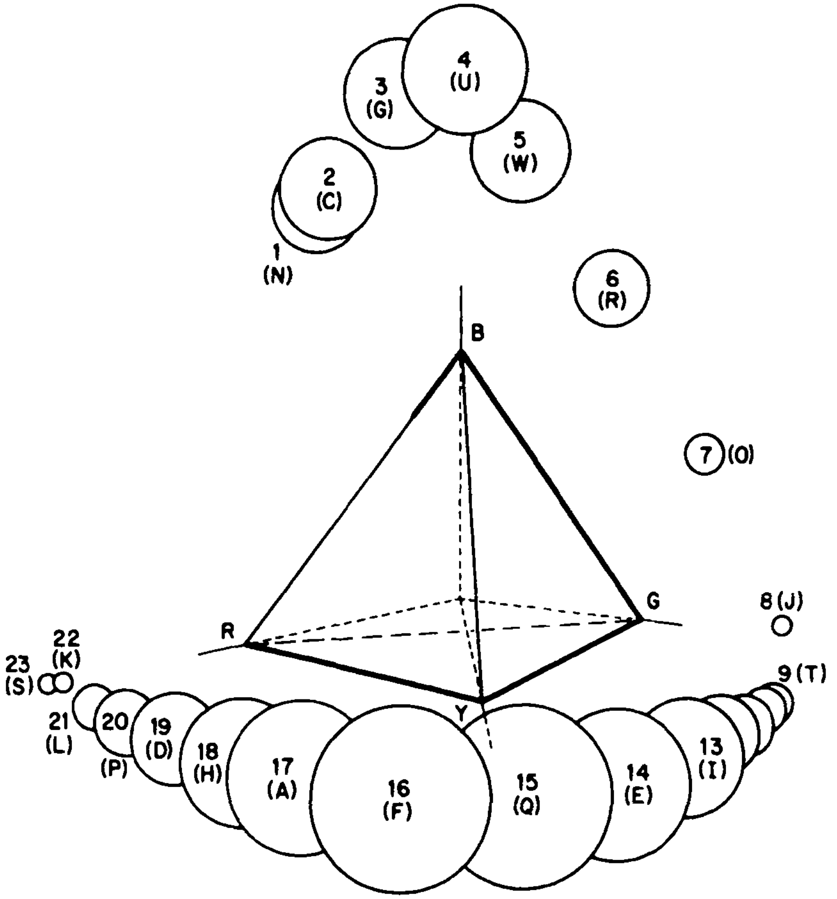
Exhibit 3a. Profiles of relative frequencies of identifications of each of 23 spectral colors as blue, green, yellow, or red (Boynton & Gordon, 1965; Shepard & Carroll, 1966)



A 4 × 4 correlation matrix may be calculated from the 23 observations; and, when a principal components analysis of the correlation matrix is performed, three of the four eigenvalues turn out to be relatively important, whereas the smallest eigenvalue is comparatively small and negligible. Thus, using a linear technique would lead one to conclude that a linear space of reduced dimensionality, $q = 3$, would be feasible and might be adequate in the present problem. Shepard & Carroll (1966) show a representation of the 23 stimuli in the space of the first three principal components, and Exhibit 3*b* is a reproduction of their ingenious two-dimensional display of the three-dimensional representation. Two of the axes are on the plane of the picture, while the third axis is to be visualized as emanating out toward the viewer. The 23 points are labeled *A* through *W* to correspond with the stimuli, and the size of the circle around a point corresponds to its distance away from the picture plane. Thus *F* and *Q* are about the closest to the viewer, while *S*, *K*, and *J* are among the farthest away.

From Exhibit 3*b* it is clear that, although $q = 3$, the 23 points are not scattered throughout three-dimensional space but, rather, appear to lie on a reasonably smooth one-dimensional curve that winds through the three-dimensional space. To emphasize this feature, the points in Exhibit 3*b* are also labeled 1 through 23 (shown alongside their original identifications by the letters *A* through *W*) to correspond with their positions on the one-dimensional curve. It is, of course, known that a single dimension (viz., wavelength) underlies the data; in fact, in the experimental setup the variation in wavelength of the 23 stimuli was accomplished by turning a single knob to different settings. (The numbering 1 through 23, in fact, corresponds with the known

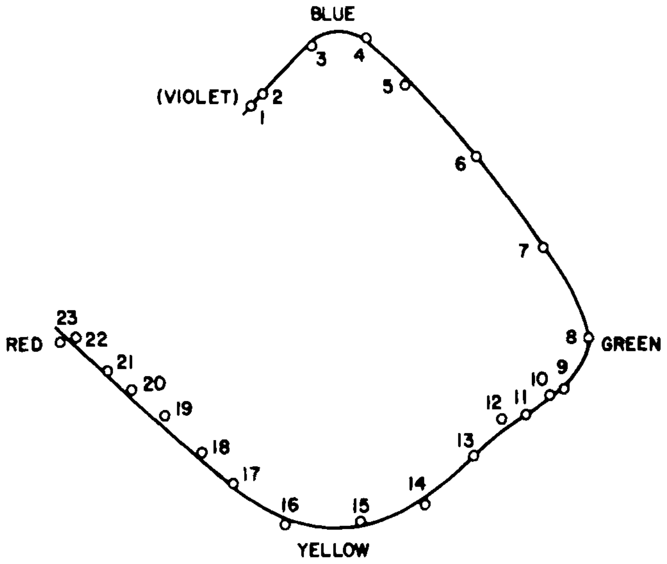
Exhibit 3*b*. Representation of three-dimensional principal factor solution for the data of Exhibit 3*a* (Shepard & Carroll, 1966)



ordering of the stimuli on the single dimension of wavelength!) Hence one might ask whether it is possible to obtain an adequate one-dimensional representation of the data.

If $\mathbf{y}_i' = (y_{i1}, \ldots, y_{i4})$ denotes the four-dimensional observation corresponding to the ith stimulus ($i = 1, \ldots, 23$), one could use the so-called city-block distance between the ith and jth observations, $\Sigma_{l=1}^{4} |y_{il} - y_{jl}|$, for $i < j = 1, \ldots, 23$, as a measure of dissimilarity (δ_{ij}) between stimuli i and j. Thus 253 ($= m$) dissimilarities may be obtained and used as input to multidimensional scaling. Shepard & Carroll (1966) performed such an analysis and found that the minimum stress in one-dimension, $S_0(1)$, was not adequately small but that $S_0(2)$ was sufficiently and markedly smaller. The two-dimensional solution obtained by Shepard & Carroll (1966) is shown in Exhibit 3*c*, with the points joined by a smooth line. Except for being confined to a two-dimensional space,

Exhibit 3c. Multidimensional scaling solution for the data of Exhibit 3*a* (Shepard & Carroll, 1966)

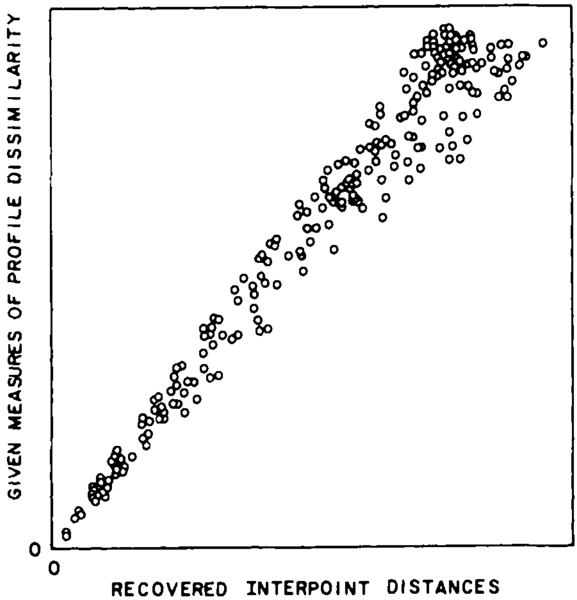


the curve in Exhibit 3*c* is qualitatively (including the location of bends) the same as the one which manifested itself in the principal components representation of Exhibit 3*b*. The tendency of both curves to close the loop is similar to the color circle concept and is explainable by the phenomenon that violet [stimulus *N* (or 1) with the lowest wavelength] is judged to contain some red [stimulus *S* (or 23) with the highest wavelength] along with a dominance by blue.

Exhibit 3*d* shows the scatter plot of interpoint Euclidean distances in the solution against the original dissimilarities (viz., the city-block distances). The near linearity of this configuration suggests that the interpoint Euclidean distances in the two-dimensional representation, determined as the output of the multidimensional scaling algorithm, are essentially linearly related to the city-block distances (calculated in the initial four-dimensional space of observations) that constituted the input measures of dissimilarity to the algorithm.

Next, since the use of the regular multidimensional scaling approach still does not lead to recovering the single dimension known to underlie the data in this example, Shepard & Carroll (1966) suggest that the monotonicity constraint not be imposed globally. The idea is not to try accommodating the dissimilarities between relatively remote profiles such as those for stimuli *N* (or 1) and *S* (or 23), since this might induce the "bending over" of a basically unidimensional phenomenon, and provision for such bending would necessitate the use of two dimensions. To focus on monotonicity only for pairs of stimuli that are likely to be "nearby" on the single dimension possibly

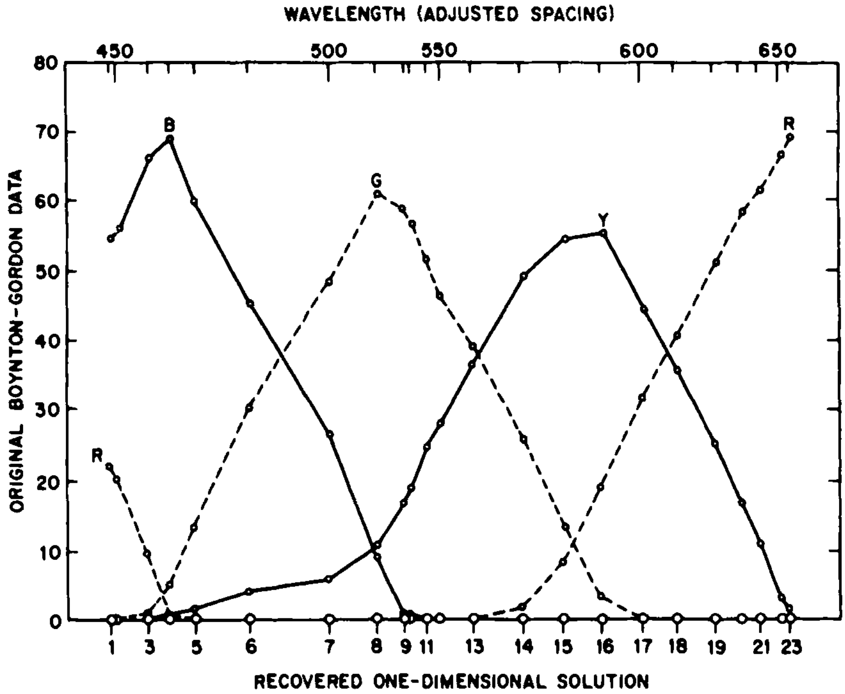
Exhibit 3d. Scatter plot of dissimilarity versus distance associated with Exhibit 3c (Shepard & Carroll, 1966)



underlying the data, one can ignore all pairs of objects whose observed dissimilarities exceed a specified cut-off value, and then require monotonicity between the distances and dissimilarities only for the remaining pairs of objects whose dissimilarities are smaller than the cut-off value. Using such a procedure and requiring monotonicity only for the pairs of stimuli with the smallest 100 dissimilarities (i.e., ignoring the 153 larger dissimilarity values) led Shepard & Carroll (1966) to the one-dimensional solution shown along the bottom of Exhibit 3*e*.

The configuration, including the spacing, is essentially the one that would be obtained by "unbending" the curve of Exhibit 3*c*. The ordering of the stimuli from 1 through 23 in Exhibit 3*e* does correspond to increasing wavelength. Their spacing, however, is not the same as it is on wavelength, as is evident from the scale at the top of Exhibit 3*e*, which shows the wavelengths corresponding to the 23 stimuli. Also shown in the figure are plots of values of each of the four original responses (B, G, Y, and R) for the 23 stimuli, against the spacings as determined in the one-dimensional solution. The observed values of the original responses are thus seen to be nonlinear functions of the single underlying dimension. Shepard & Carroll (1966) noticed the interesting fact that these curves are more regular and symmetrical than those obtained by Boynton & Gordon (1965), showing the responses plotted directly against wavelength. The experiment clearly involves the psychological perception of

Exhibit 3e. Multidimensional scaling solution with local monotonicity constraint for the data of Exhibit 3*a* (Shepard & Carroll, 1966)

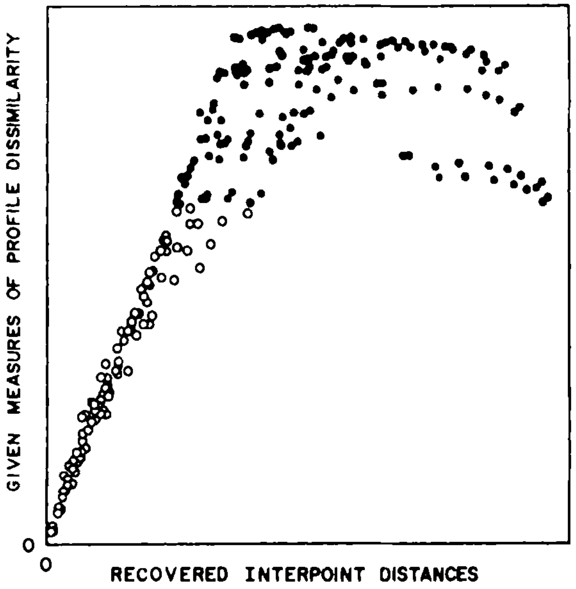


colors, and for this reason the data may not be reflecting only the experimentally controlled physical dimension of wavelength.

Exhibit 3*f* shows the scatter plot of interpoint Euclidean distances versus the initial dissimilarities for the one-dimensional solution shown in Exhibit 3*e*, and the departure from monotonicity at the top end may be seen clearly. The interpoint distances and dissimilarities, for stimuli such as *N* and *S* whose profiles (initial data) are very different, are indeed coordinates of the points at the top end of Exhibit 3*f*.

An important difficulty with the above simple modification of confining monotonicity to the smallest dissimilarities and the corresponding recovered distances is that, when the underlying dimensionality of the curved manifold is larger than 1, the procedure may not lead to detecting this. Hence the modified multidimensional scaling procedure may be adequate when $q = 1$ but not when $q > 1$. As an example, Shepard & Carroll (1966) mention the case in which the points lie on the surface of a fish bowl or a sphere with a hole. Here $p = 3$ and $q = 2$, and an appropriate solution would be a representation of the points on a two-dimensional disk obtained by pulling out the sphere at the hole and

Exhibit 3*f*. Scatter plot of dissimilarity versus distance associated with the multidimensional scaling solution in Exhibit 3*e* (Shepard & Carroll, 1966)



flattening out into a disk. However, in such solution, points at the rim of the hole which were close together in three-dimensional space end up being far apart on the two-dimensional disk; that is, points with small initial dissimilarities end up with large interpoint distances in the recovered configuration. An alternative modification of multidimensional scaling to handle this difficulty might be to impose regional monotonicity; in other words, monotonicity might be required separately within a region surrounding each point (see Bennett, 1965).

Rather than pursuing such a modification, however, Shepard & Carroll (1966) suggest a different procedure that involves maximizing an index of continuity, so as to find a representation of the original p-dimensional points in terms of $q(<p)$ new coordinates that are "smoothly" related to the old ones. Specifically, if $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ are points in an initial p-dimensional representation of n objects, they suggest finding a configuration $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ in q-space so as to minimize an index of the form

$$\kappa = \frac{\sum\sum_{i<j=1}^{n} (d_{ij}^2/D_{ij}^2)w_{ij}}{\text{Normalizing factor}}.$$

Here d_{ij} is the Euclidean distance between $\mathbf{y}_i$ and $\mathbf{y}_i$, while D_{ij} is the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, and w_{ij}'s are weights that decrease as the distance between the corresponding points in the x-space increases. The numerator of

κ is a multivariate generalization of the mean square successive difference whose ratio to the variance is a statistic which has been used as an inverse measure of trend in univariate time series. The smaller the value of κ, the "smoother" the relationship between $\mathbf{y}$ and $\mathbf{x}$ is considered to be.

Using the desiderata of invariance of the ratios of the weights under translations, and uniform shrinking or dilation of the x-space, Shepard & Carroll (1966) recommend the choice $w_{ij} = 1/D_{ij}^2$. Similarly, arguing that it would be desirable for the unconstrained minimum of the index to be attained when D_{ij}^2 is proportional to d_{ij}^2 for every i and j (i.e., the configurations in x-space and y-space match except for a similarity transformation), they suggest using the normalizing factor, $[\Sigma\Sigma_{i<j} 1/D_{ij}^2]^2$, in the denominator of κ.

Thus, given the initial configuration $\mathbf{y}_1, \ldots, \mathbf{y}_n$ in p-space, the approach suggested by Shepard & Carroll (1966) is to choose the nq coordinates involved in $\mathbf{x}_1, \ldots, \mathbf{x}_n$ so as to minimize

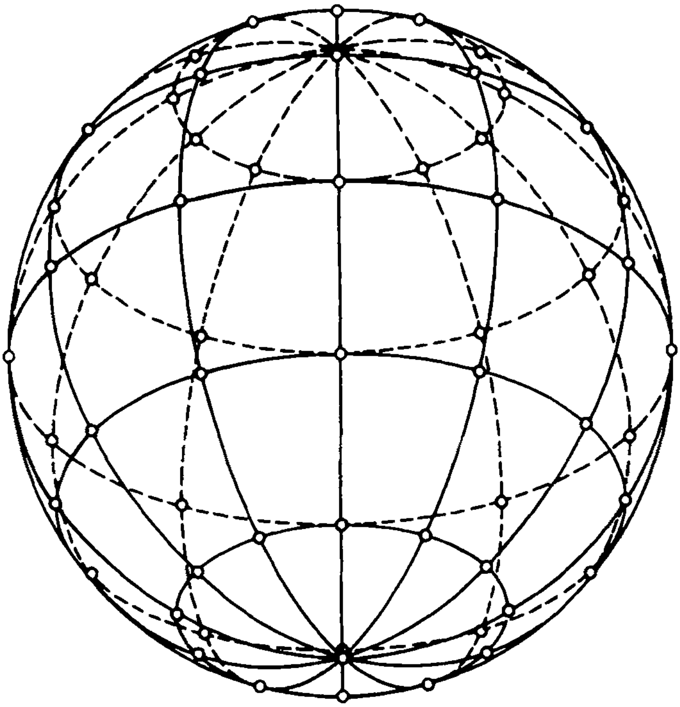
$$\kappa = \frac{\sum\sum_{i<j=1}^{n} (d_{ij}^2/D_{ij}^4)}{\left\{\sum\sum_{i<j=1}^{n} (1/D_{ij}^2)\right\}^2}. \tag{36}$$

Starting from a trial configuration in q-space, one could iterate to the desired configuration, with the minimum value of κ, by using a numerical optimization technique, such as the method of steepest descent. Also, as was the case in multidimensional scaling, one could repeat the process for a series of values of q and choose the smallest "adequate" value of q by studying the achieved values of κ for the different values of q.

In contradistinction to multidimensional scaling, the above procedure assumes the initial format of the data to be a Euclidean representation (i.e., n points in p-space) and not to consist only of rank order information about the pairwise dissimilarities. Also, little is known about the dependence of the final solution on the use of (i) other measures of distance besides the Euclidean measure in the x- or y-space, (ii) other weights, w_{ij}, and (iii) other normalizing factors. The experience with this approach, using Monte Carlo or real data, is too limited for specification of a yardstick for assessing the smallness of an observed value of κ. The unconstrained minimum of κ (corresponding to which there need not, of course, be a Euclidean configuration in x-space) is easy to compute, and in their published examples Shepard & Carroll (1966) seem to use this value for judging how small the κ is for the optimum configuration determined by the procedure. Despite these limitations, some interesting examples of the application of the procedure are discussed by Shepard & Carroll (1966), and the following example is taken from their work.

Example 4. The artificially generated data consisted of 62 points on the surface of a sphere — 12 points on each of five equally spaced parallels, and the two poles. Hence $n = 62$, $p = 3$, and it is known here that $q = 2$. Exhibit 4*a*

Exhibit 4a. Artificial data consisting of 62 points on the surface of a sphere (Shepard & Carroll, 1966)

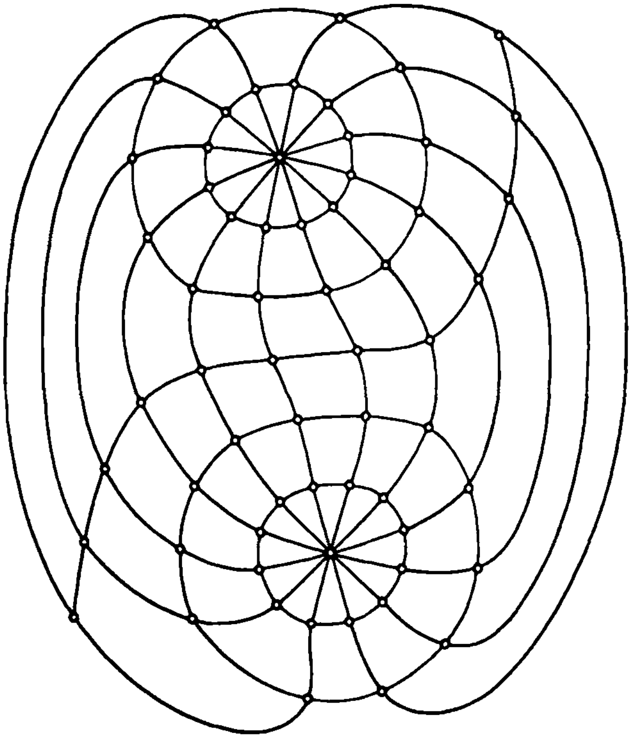


shows the data; Exhibit 4*b*, the solution obtained in two dimensions by minimizing κ. The solution consists of two hemispheres in three-dimensional space opened out on a hinge at the equator and then flattened out into a common plane. The equatorial circle has been distorted into an S-shaped curve. The reader is reminded, however, that the computer output in this solution (exactly as in the uses of multidimensional scaling) consists *only* of the coordinates of the points corresponding to the *n* objects, and the lines are drawn in from extraneous knowledge of some structure among the objects.

All of the procedures described in the present subsection of the book depend on the use of an index of achievement (viz., minimum stress or minimum κ) as an informal basis for assessing the comparative adequacy of the successive dimensions employed. This index is to be used with appropriate judgment relative to meaningfulness of interpretation in the subject matter area. A firm commitment to benchmarks for comparing achieved values of the index, without regard for issues such as interpretability, is neither necessary nor recommended for using these tools of data analysis.

***Example** 5.* The data, from an experiment of Rothkopf (1957) (see also Shepard, 1963; Kruskal, 1964a), were obtained from 598 subjects who were

Exhibit 4b. Two-dimensional parametric mapping of data of Exhibit 4a (Shepard & Carroll, 1966)



asked to judge whether or not pairs of successively presented Morse code signals were the same. Thirty-six signals were employed: 26 for the letters of the alphabet and 10 for the digits 0 through 9. Exhibit 5*a* is a matrix of the percentages of times that a signal corresponding to the row label was identified as being the same as the signal corresponding to the column label. These percentages may be considered as measures of similarity between the pairs of Morse code signals. [*Note*: Although large, the diagonal values in Exhibit 5*a* are not 100% (similarity of a signal to itself is not perfect), and also the matrix is not symmetric, so that $\delta_{ij} \neq \delta_{ji}$.] To use multidimensional scaling in a direct manner, the averages of each pair of symmetrically situated off-diagonal elements of this matrix may serve as input measures of similarity between the coresponding pair in the 36 signals.

Exhibit 5*b*, taken from Kruskal (1964a), shows the minimum stress achieved by the multidimensional scaling solution plotted against the number of dimensions employed for that solution. Using the benchmarks recommended by Kruskal (1964a) and mentioned earlier, as well as the rule of choosing the dimensionality by looking for an "elbow" in a plot such as Exhibit 5*b*, one might feel that a choice of $q = 2$ in this example would, from the point of view

Exhibit 5a. Data matrix of percentages of confusions between pairs of Morse code signals (Rothkopf, 1957; Shepard, 1963)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	1	2	3	4	5	6	7	8	9	0	
A	92	04	06	13	03	14	10	13	46	05	22	03	25	34	06	06	09	35	23	06	37	13	17	12	07	03	02	07	05	05	08	06	05	06	02	03	A
B	05	84	37	31	05	28	17	21	05	19	34	40	06	10	12	22	25	16	18	02	18	34	08	84	30	42	12	17	14	40	32	74	43	17	04	04	B
C	04	38	87	17	04	29	13	07	11	19	24	35	14	03	09	51	34	24	14	06	06	11	14	32	82	38	13	15	31	14	10	30	28	24	18	12	C
D	08	62	17	88	07	23	40	36	09	13	81	56	08	07	09	27	09	45	29	06	17	20	27	40	15	33	03	09	06	11	09	19	08	10	05	06	D
E	06	13	14	06	97	02	04	04	17	01	05	06	04	04	05	01	05	10	07	67	03	03	02	05	06	05	04	03	05	03	05	02	04	02	03	03	E
F	04	51	33	19	02	90	10	29	05	33	16	50	07	06	10	42	12	35	14	02	21	27	25	19	27	13	08	16	47	25	26	24	21	05	05	05	F
G	09	18	27	38	01	14	90	06	05	22	33	16	14	13	62	52	23	21	05	03	15	14	32	21	23	39	15	14	05	10	04	10	17	23	20	11	G
H	03	45	23	25	09	32	08	87	10	10	09	29	05	08	08	14	08	17	37	04	36	59	09	33	14	11	03	09	15	43	70	35	17	04	03	03	H
I	64	07	07	13	10	08	06	12	93	03	05	16	13	30	07	03	05	19	35	16	10	05	08	02	05	07	02	05	08	09	06	08	05	02	04	05	I
J	07	09	38	09	02	24	18	05	04	85	22	31	08	03	21	63	47	11	02	07	09	09	09	22	32	28	67	66	33	15	07	11	28	29	26	23	J
K	05	24	38	73	01	17	25	11	05	27	91	33	10	12	31	14	31	22	02	02	23	17	33	63	16	18	05	09	17	08	08	18	14	13	05	06	K
L	02	69	43	45	10	24	12	26	09	30	27	86	06	02	09	37	36	28	12	05	16	19	20	31	25	59	12	13	17	15	26	29	36	16	07	03	L
M	24	12	05	14	07	17	29	08	08	11	23	08	96	62	11	10	15	20	07	09	13	04	21	09	18	08	05	07	06	06	05	07	11	07	10	04	M
N	31	04	13	30	08	12	10	16	13	03	16	08	59	93	05	09	05	28	12	10	16	04	12	04	06	11	05	02	03	04	04	06	02	02	10	02	N
O	07	07	20	06	05	09	76	07	02	39	26	10	04	08	86	37	35	10	03	04	11	14	25	35	27	27	19	17	07	07	06	18	14	11	20	12	O
P	05	22	33	12	05	36	22	12	03	78	14	46	05	06	21	83	43	23	09	04	12	19	19	19	41	30	34	44	24	11	15	17	24	23	25	13	P
Q	08	20	38	11	04	15	10	05	02	27	23	26	07	06	22	51	91	11	02	03	06	14	12	37	50	63	34	32	17	12	09	27	40	58	37	24	Q
R	13	14	16	23	05	34	26	15	07	12	21	33	14	12	12	29	08	87	16	02	23	23	62	14	12	13	07	10	13	04	07	12	07	09	01	02	R
S	17	24	05	30	11	26	05	59	16	03	13	10	05	17	06	06	03	18	96	09	56	24	12	10	06	07	08	02	02	15	28	09	05	05	05	02	S
T	13	10	01	05	46	03	06	06	14	06	14	07	06	05	06	11	04	04	07	96	08	05	04	02	02	06	05	05	03	03	03	08	07	06	14	06	T
U	14	29	12	32	04	32	11	34	21	07	44	32	11	13	06	20	12	40	51	06	93	57	34	17	09	11	06	06	16	34	10	09	09	07	04	03	U
V	05	17	24	16	09	29	06	39	05	11	26	43	04	01	09	17	10	17	11	06	32	92	17	57	35	10	10	14	28	79	44	36	25	10	01	05	V
W	09	21	30	22	09	36	25	15	04	25	29	18	15	06	26	20	25	61	12	04	19	20	86	22	25	22	10	22	19	16	05	09	11	06	03	07	W
X	07	64	45	19	03	28	11	06	01	35	50	42	10	08	24	32	61	10	12	03	12	17	21	91	48	26	12	20	24	27	16	57	29	16	17	06	X
Y	09	23	62	15	04	26	22	09	01	30	12	14	05	06	14	30	52	05	07	04	06	13	21	44	86	23	26	44	40	15	11	26	22	33	23	16	Y
Z	03	46	45	18	02	22	17	10	07	23	21	51	11	02	15	59	72	14	04	03	09	11	12	36	42	87	16	21	27	09	10	25	66	47	15	15	Z
1	02	05	10	03	03	05	13	04	02	29	05	14	09	07	14	30	28	09	04	02	03	12	14	17	19	22	84	63	13	08	10	08	19	32	57	55	1
2	07	14	22	05	04	20	13	03	25	26	09	14	02	03	17	37	28	06	05	03	06	10	11	17	30	13	62	89	54	20	05	14	20	21	16	11	2
3	03	08	21	05	04	32	06	12	02	23	06	13	05	02	05	37	19	09	07	06	04	16	06	22	25	12	18	64	86	31	23	41	16	17	08	10	3
4	06	19	19	12	08	25	14	16	07	21	13	19	03	03	02	17	29	11	09	03	17	55	08	37	24	03	05	26	44	89	42	44	32	10	03	03	4
5	08	45	15	14	02	45	04	67	07	14	04	41	02	00	04	13	07	09	27	02	14	45	07	45	10	10	14	10	30	69	90	42	24	10	06	05	5
6	07	80	30	17	04	23	04	14	02	11	11	27	06	02	07	16	30	11	14	03	12	30	09	58	38	39	15	14	26	24	17	88	69	14	05	14	6
7	06	33	22	14	05	25	06	04	06	24	13	32	07	06	07	36	39	12	06	02	03	13	09	30	30	50	22	29	18	15	12	61	85	70	20	13	7
8	03	23	40	06	03	15	15	06	02	33	10	14	03	06	14	12	45	02	06	04	06	07	05	24	35	50	42	29	16	16	09	30	60	89	61	26	8
9	03	14	23	03	01	06	14	05	02	30	06	07	16	11	10	31	32	05	06	07	06	03	08	11	21	24	57	39	09	12	04	11	42	56	91	78	9
0	09	03	11	02	05	07	14	04	05	30	08	03	02	03	25	21	29	02	03	04	05	03	02	12	15	20	50	26	09	11	05	22	17	52	81	94	0
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	1	2	3	4	5	6	7	8	9	0	

Exhibit 5*b*. Plot of stress versus number of dimensions for the example in Exhibit 5*a* (Kruskal, 1964a; Shepard, 1963)

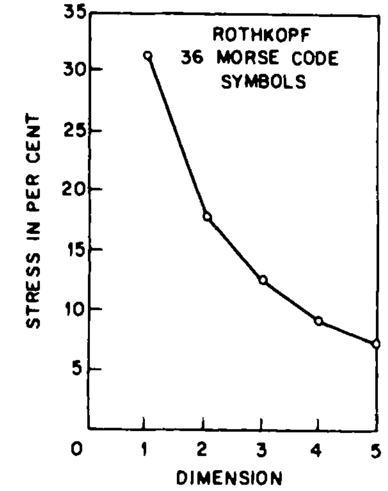
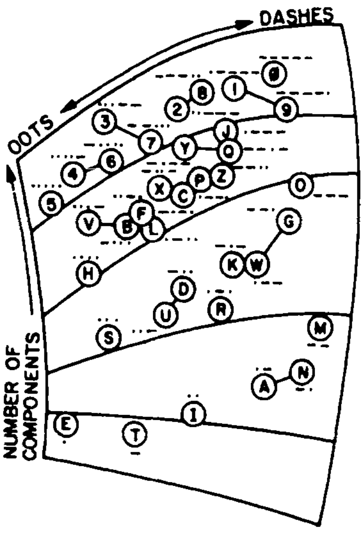


Exhibit 5c. Multidimensional scaling solution for the Morse code signals (Shepard, 1963)



of the index of achievement, be between fair and poor, and if not the value of 5 for *q* one should at least consider the value 3.

However, the two-dimensional solution obtained and interpreted by Shepard (1963) is shown in Exhibit 5*c*. The vertical axis in this solution is seen to correspond to the number of components in the Morse code symbol (i.e., the total number of dots and dashes), while the horizontal axis characterizes the composition of the symbol (i.e., the ratio of number of dots to number of dashes). Corresponding to the large value of stress ($\simeq 18\%$) for this solution, one observes that the distances between the representations of the signals do not closely match the observed similarities in Exhibit 5*a* [e.g., compare the distance between the pair (B, X) with those between (B, F) and (B, L)]. However, as Kruskal (1964a) points out, the fact that Shepard could not extract additional interpretable structure by going to three dimensions suggests that $q = 2$ would be a better choice than $q = 3$ for this problem regardless of any contraindication from Exhibit 5*b*. Interpretability and simplicity are important in data analysis, and any rigid inference of optimal dimensionality, in the light of the observed values of a numerical index of goodness of fit, may not be productive.

In this example, as well as the previous ones of the use of multidimensional scaling, the initial dissimilarity (or similarity) values were averages across several subjects, and no provision was made for differences among subjects. In later work, Carroll & Chang (1970) have proposed a scheme for multidimensional scaling that allows for individual differences. In this approach the coordinates recovered by multidimensional scaling are assumed to be the same for all individuals, but the weights assigned to the different coordinates are not assumed to be identical for all individuals.

The ideas and procedures involved in the above scaling types of approaches are imaginative and insightful. The procedures, however, have limitations in practical application in that they involve extensive iteration on $n(n-1)/2$ quantities—the interpoint similarities or distances. The currently available computer programs, for instance, can effectively and economically handle up to about 75 $(=n)$ objects only. Also, the solution space produced by these procedures does not have an analytic description that is simply interpretable in terms of the original response space. This becomes especially important when the solution space is of dimensionality greater than three and graphical representations are no longer available. Also, when one starts with a metric representation and uses some measure of distance in the initial space as the input measure of dissimilarity (see Example 3), questions arise concerning the nature of the dependence of the recovered configuration on the initial choice of distance function.

2.4. NONLINEAR SINGULARITIES AND GENERALIZED PRINCIPAL COMPONENTS ANALYSIS

2.4.1. Nonlinear Singularities

The problem of reduction in dimensionality concerns the recognition of lower-dimensional, possibly nonlinear, subspaces near which the multiresponse observations may, statistically speaking, lie. This is, of course, not a well-defined concept, in a sense very similar to the indefiniteness involved in the notion of "fitting a curve" to a scatter plot of y against x.

One major source of difficulty in the problem is the fact that in the analysis of high-dimensional data there are not available the informal, mainly graphical, internal comparisons procedures, such as scatter plots, that guide so much single-response, and some two-response, data analysis.

So far as near-linear singularities in a body of data are concerned, these may be statistically indicated by principal components analysis (but see Section 6.6 for a discussion of possible effects of outliers). However, nonlinear singularities will not necessarily be indicated by principal components, and one may not be able to infer their existence even from various obvious two-dimensional scatter plot representations of the data.

Example 6. The C-shaped configuration shown in Exhibit 6*a* is an oversimplified example of data configurations that may not be revealed by classical linear principal components analysis. Clearly, the three-dimensional analogue of this, namely, a cup-shaped configuration of data, may not be revealed even by a combination of principal components analysis in 3-space and two-dimensional marginal scatter plots. Example 7 discusses such an example.

Exhibit 6*b* is a plot of 50 computer-generated random bivariate normal samples with an underlying positive correlation coefficient. This is, therefore, an example of typical normal distribution scatter when $p = 2$.

The filled-in squares in Exhibits 6*a* and 6*b* are the centers of gravity (means) of the data.

One elementary technique for detecting the existence of curved configurations, such as the one considered in Example 6, involves the computation of

Exhibit 6*a*. Example of a curved configuration of data

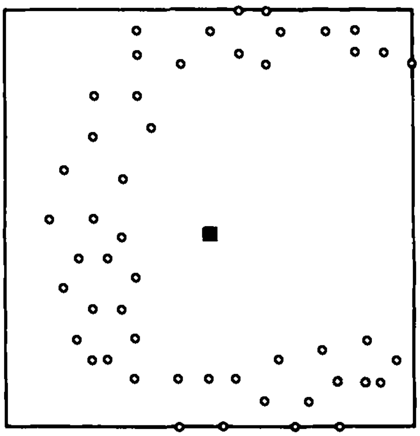
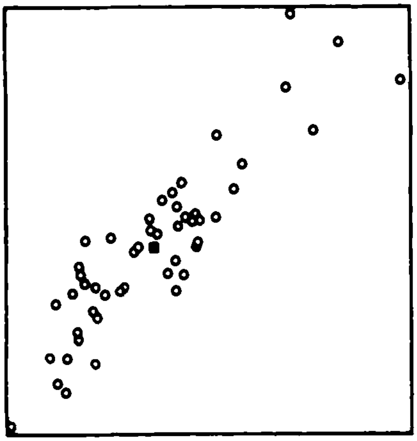


Exhibit 6b. Simulated sample of bivariate normal scatter



the squared generalized distances (using the inverse of the sample covariance matrix) of each point from the center of gravity, namely, $d_i^2=(\mathbf{y}_i-\bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_i-\bar{\mathbf{y}})$ for $i=1,\dots,n$. For a typical multivariate normal scatter (see Exhibit 6*b*), either throughout p dimensions or mainly on a linear subspace, these distances will have approximately a chi-squared or gamma distribution. Hence, on an appropriately selected gamma probability plot (Wilk, Gnanadesikan, & Huyett, 1962a; see also the discussion of probability plots in Section 6.2), using shape parameter values in the neighborhood of $p/2$, they will tend to show as a linear configuration oriented toward the origin. For curved singularity, such as the illustration, however, it is clear that there will be a deficiency of small distances. This will show on the gamma plot by orientation of the configuration toward a nonzero intercept at the "small" end. A histogram of the observed distances would also indicate the sparseness of small ones, but the probability plot may provide additional insights. The following simulated three-dimensional example illustrates some of these ideas and procedures.

Example 7. The 61 triads shown in Exhibit 7*a* were obtained by appending a standard normal deviate to each of the coordinates of points on the surface of a specified paraboloid. Exhibits 7*b*, 7*c*, and 7*d*, show the three possible two-dimensional scatter plots of these data with respect to pairs of the original three coordinates. None of these data displays is suggestive of the observations in three dimensions lying "near" a curved subspace.

The inverse of the sample covariance matrix, S, of these observations was employed to compute the generalized squared distance of each of the 61 points from their centroid. (See Section 4.2.1 for a discussion of distance measures.) A gamma probability plot of the ordered squared distances is shown in Exhibit 7*e*. The value of the shape parameter used for this plot was $\eta = 3/2$, and hence

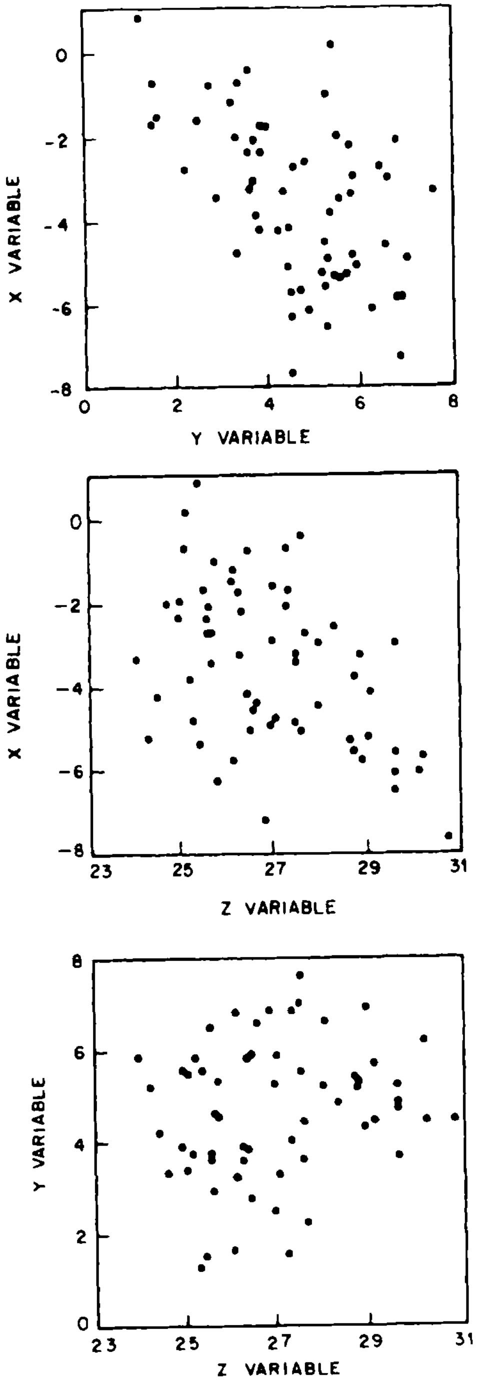
Exhibit 7a. Simulated data with observations scattered off the surface of a paraboloid

−2.732	6.557	25.507	−3.452	2.948	25.591
−5.264	5.253	24.200	−7.261	6.959	26.789
−5.103	5.986	26.446	−2.370	3.617	25.510
−3.335	5.888	23.947	−4.181	4.530	29.118
−5.420	5.607	25.321	−2.360	3.916	24.879
−3.261	7.697	27.479	−5.297	5.802	29.073
−4.607	6.651	26.518	−1.585	2.524	26.954
−4.236	4.220	24.416	−3.267	4.402	28.899
−4.947	5.363	26.918	−1.187	3.257	26.100
−2.189	5.881	26.282	−2.095	6.931	27.269
−2.913	5.953	26.962	−4.800	3.339	27.011
−4.838	5.909	25.196	−5.602	5.322	28.759
−3.448	5.610	27.489	−1.478	1.644	26.057
−0.990	5.391	25.667	−5.151	4.481	27.583
−6.116	6.326	30.189	−0.694	3.408	24.997
−2.715	4.645	25.613	−5.687	4.766	29.640
−5.849	6.876	26.070	−1.733	3.932	26.198
0.162	5.521	25.027	−6.154	4.932	29.631
−5.360	5.494	28.675	−3.823	3.784	25.123
−1.740	4.070	27.311	−2.588	4.923	28.343
−2.975	6.716	27.999	−3.237	3.648	26.249
−4.220	3.853	26.396	−5.740	4.537	30.277
−6.306	4.573	25.715	−0.709	1.542	27.240
−1.972	5.615	24.900	−6.568	5.335	29.631
−4.497	5.314	27.978	−1.669	1.501	25.413
−2.005	3.352	24.599	−7.690	4.578	30.863
−3.809	5.421	28.794	0.837	1.271	25.303
−2.081	3.795	25.542	−5.832	7.020	28.915
−4.907	7.120	27.449	−0.405	3.669	27.587
−0.742	2.800	26.394	−3.019	3.752	29.665
−2.750	2.233	27.669			

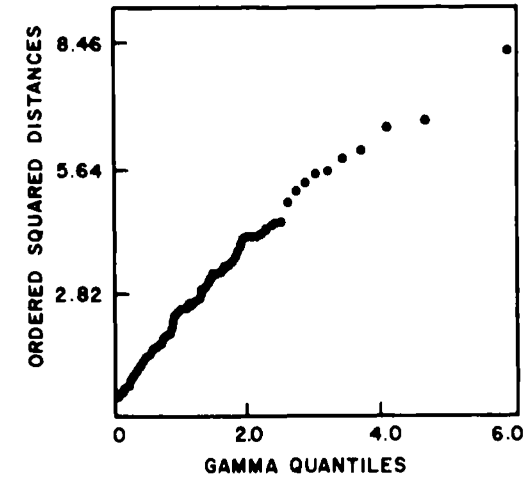
it is essentially a probability plot for a chi-squared distribution with 3 degrees of freedom. The configuration, which has a nonzero intercept, shows clearly the "deficiency" of small values, indicating a "hole" in the data. Furthermore, the slight tendency of the configuration to "bend over" at the top right-hand corner suggests that the data may be near a dish or a shallow paraboloid rather than a deep paraboloid. The nature of the indicated peculiarity may be investigated further by using the methods discussed below.

The simple-minded idea illustrated in Example 7 will, of course, not be indicative when the nonlinear surface near which the data lie has several bends, such as a sinusoidal shape. Furthermore, if a peculiarity is indicated, the probability plot does not tell very much about its nature.

Exhibits 7*b*–*d*. Scatter plots of bivariate subsets of the data of Exhibit 7*a*



Exhibits 7e. Gamma probability plot (with shape parameter = 3/2) of generalized squared distances for data of Exhibit 7a



A method proposed by Gnanadesikan & Wilk (1966, 1969) and considered independently by Van der Geer (1968) is useful for analyzing multidimensional linear or nonlinear singularities. The technique is a generalization of classical linear principal components analysis.

2.4.2. Generalized Principal Components Analysis

If one has nearly linear singularity of the data, what one wishes to do is to determine the *linear* coordinate system that is most nearly in concordance with the data configuration. Then the expression of the data in the new coordinate system is simpler, in that the effective description can be given by the use of fewer coordinates. This is accomplished by linear principal components analysis.

If one has nearly nonlinear singularity of the data, what one wishes to do is to determine the *nonlinear* coordinate system that is most nearly in agreement with the data configuration, just as in the linear case. Given a class of possible nonlinear coordinates, one needs to select that one along which the data variance is maximum, and then obtain another, uncorrelated with the first, along which the variance is next largest, and so on. For any class of coordinates that consist of an unspecified linear combination of arbitrary *functions* of the original coordinates, the solution to this problem is simply an enlarged eigenvalue-eigenvector problem. The essential concepts and computations may be illustrated by considering the bivariate ($p = 2$) case.

Suppose that $\mathbf{y}' = (y_1, y_2)$ denotes the original bivariate response, and for illustrative purposes suppose that one is seeking a quadratic coordinate system.

Thus, as a first step, one wishes to find

$$z = ay_1 + by_2 + cy_1y_2 + dy_1^2 + ey_2^2 \tag{37}$$

such that the variance of z is maximum among all such quadratic functions of y_1 and y_2.

Denoting y_1y_2 as y_3, y_1^2 as y_4, and y_2^2 as y_5, one can consider

$$\mathbf{y}^{*\prime} = (y_1\, y_2, y_3, y_4, y_5) \tag{38}$$

as a five-dimensional vector of responses, two of which are just the original variables and the remaining three are functions (squares and cross product) derived from them. If

$$\mathbf{a}^{*\prime} = (a, b, c, d, e) \tag{39}$$

denotes the vector of coefficients used in Eq. 37 in defining z, the first stage of a *quadratic principal components* analysis may be formulated as the problem of determining $\mathbf{a}^*$ so that the variance of $z(=\mathbf{a}^{*\prime}\mathbf{y}^*)$ is maximum subject to a normalizing constraint, such as $\mathbf{a}^{*\prime}\mathbf{a}^* = 1$, exactly as in linear principal components analysis.

On the basis of a sample, of n observations on the initial response variables y_1 and y_2, one can generate n observations on $\mathbf{y}^*$ and thence obtain the "sample" mean vector and covariance matrix:

$$\bar{\mathbf{y}}^{*\prime} = (\bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4, \bar{y}_5),$$

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

For a given $\mathbf{a}^*$ the observed variance of z will be $\mathbf{a}^{*\prime}\mathbf{S}^*\mathbf{a}^*$, and the first stage of the quadratic principal components analysis will then result in choosing $\mathbf{a}^* = \mathbf{a}_1^*$, where $\mathbf{a}_1^*$ is the eigenvector corresponding to the largest eigenvalue, c_1^*, of $\mathbf{S}^*$. Furthermore, if at the next stage one wishes a second quadratic function that is uncorrelated (in the sample) with the first and has maximum variance subject to being thus uncorrelated, then, exactly as in linear principal components, one will choose the set of coefficients for the second quadratic function as the eigenvector, $\mathbf{a}_2^*$, of $\mathbf{S}^*$ corresponding to its second largest eigenvalue, c_2^*. The process can be repeated to determine additional quadratic functions with decreasing variances at each stage and with zero correlations with each of the quadratic functions determined at earlier stages. Thus, if $c_1^* \geqslant c_2^* \geqslant \cdots \geqslant c_5^* \geqslant 0$ are the eigenvalues of $\mathbf{S}^*$ with corresponding eigenvectors $\mathbf{a}_1^*$, $\mathbf{a}_2^*$, $\mathbf{a}_3^*$, $\mathbf{a}_4^*$, and $\mathbf{a}_5^*$, one can derive five quadratic functions of the two original variables from the relationships

$$z_i = \mathbf{a}_i^{*\prime}\mathbf{y}^*, \qquad i = 1, 2, \ldots, 5. \tag{40}$$

As a method for the reduction of dimensionality, the interest will lie in the functions defined by "smallest" eigenvalues. For instance, if the bivariate observations lie on a quadratic curve, one will expect $c_5^* = 0$ and z_5 will be a constant for all observations, $\mathbf{y}_i^*$.

More generally, given n p-dimensional observations, if one wishes to perform a quadratic principal components analysis, one will augment the original p variables by $p + [p(p-1)/2]$ derived variables (viz., all squared and cross-product terms) and carry out a regular principal components analysis of the

$$\left[2p + \frac{p(p-1)}{2}\right] \times \left[2p + \frac{p(p-1)}{2}\right]$$

covariance matrix of the enlarged set of variables.

The generalization is immediate to cubic and higher-order polynomial principal components, as well as to any situation in which the system of derived coordinates sought consists of (unknown) linear combinations of (specified, possibly nonlinear) functions of the original variables. Thus a typical member of the class of derived coordinates that can be handled by the above generalization of principal components is

$$z = \sum_{j=1}^{k} a_j f_j(y_1, y_2, \ldots, y_p), \tag{41}$$

where the a_j's are to be determined, and the f_j's are completely specified, but otherwise arbitrary, functions of the original variables.

The above generalization, formulated in terms of properties of variances of and correlations among the derived coordinates, is an obvious extension to the nonlinear case of Hotelling's approach to linear principal components analysis. In particular, linear principal components may be viewed as a special case of polynomial principal components. In a linear principal components analysis, one starts with p correlated coordinates and derives a set of p uncorrelated (in the sample) coordinates that are linear functions of the original variables. In a quadratic principal components analysis, however, one starts with p correlated coordinates and obtains a set of $2p + [p(p-1)/2]$ uncorrelated quadratic coordinates that are quadratic functions of the original variables. As a method for nonlinear reduction of dimensionality, in using quadratic (or otherwise generalized) principal components analysis, the interest will lie in the functions defined by the "smallest" eigenvalues.

Example 8. The data consisted of 41 points lying on the parabola $Y_2 = 2 + 4Y_1 + 4Y_1^2$, with $Y_1 = -1.5(0.05)0.5$. A quadratic principal components anal-

Exhibit 8a. Eigenanalysis for example of quadratic principal components analysis

Eigenvalues				
2163.634	219.915	2.258	2.223	0.000009
Eigenvectors				
−.002513	.246077	.508757	−.442445	.696310
.169321	.011882	.758811	.604229	−.174076
−.094253	.932909	−.212548	.274991	.0000004
.044843	−.243106	.319056	.593499	.696311
.980015	.099425	−.135640	−.106239	.0000003

ysis was applied to this data, and the dimension of the enlarged eigenvalue problem was, therefore, $2p + [p(p-1)/2] = 5$, since $p = 2$. The resulting eigenvalues and eigenvectors are shown in Exhibit 8*a*. The largest eigenvalue is seen to be over 2000, while the smallest, which is "known" to be 0, is computed as 9×10^{-6}.

Each eigenvector provides a nonlinear (quadratic in this case) coordinate in the original space, and Exhibit 8*b* shows the coordinates determined by the eigenvectors corresponding to the smallest and largest of the five eigenvalues.

Exhibit 8b. Data and coordinates from the largest and smallest eigenvalues

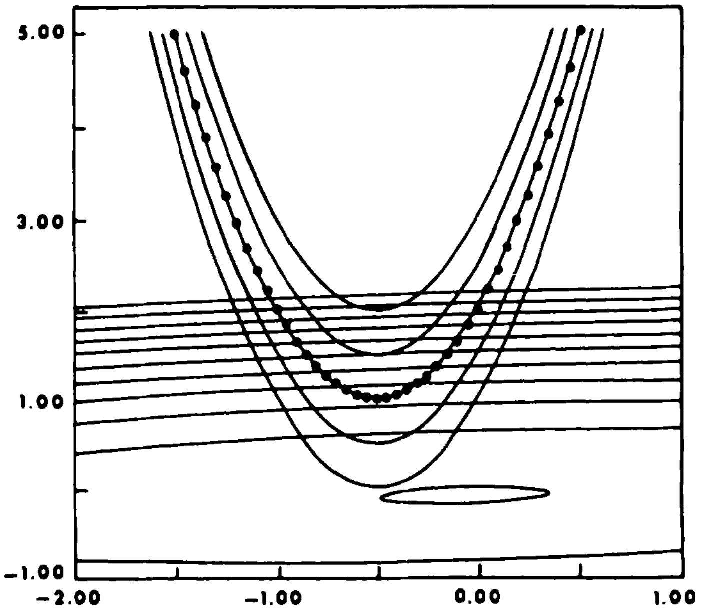


Exhibit 9a. Eigenanalysis for example of quadratic principal components analysis

Eigenvalues				
1969.563	285.938	5.717	2.034	1.081
Eigenvectors				
−.010915	.260105	−.269383	.610264	.698023
.172863	.012363	.189273	.761768	−.594853
−.106665	.930317	.301394	−.146763	−.103707
.062983	−.230187	.892766	.062053	.377047
.977064	.117118	−.061142	−.147978	.077413

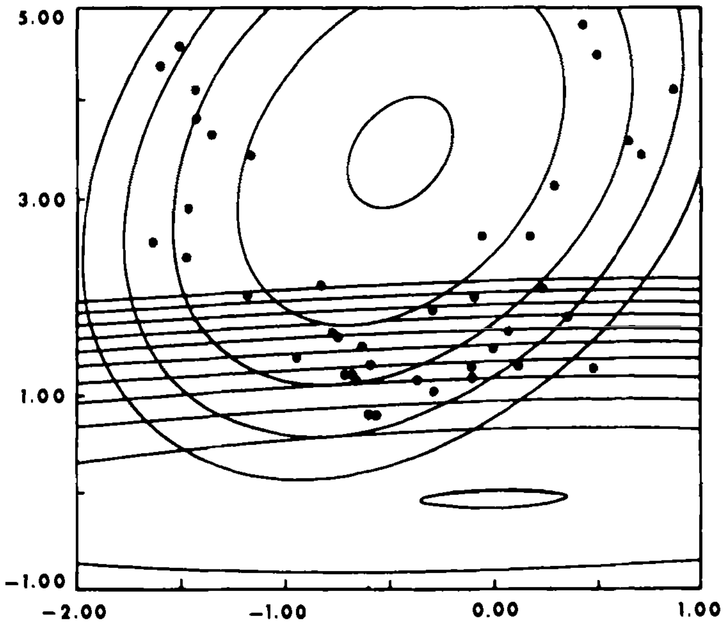
The parabolic coordinate is from the smallest eigenvalue, and the flat elliptic coordinate is from the largest. In the absence of statistical errors in the data, one of the parabolas (viz., the middle one) passes exactly through the points.

Example 9. To illustrate the effect of "noise" on the technique, the data of Example 8 were modified by adding random normal components (with mean 0 and variance $\frac{1}{16}$) to each of the two coordinates of every point. The results of an eigenanalysis, comparable to the one in Example 8, are shown in Exhibit 9*a*. It is seen that, although the largest eigenvalue is still about 2000, the smallest one is now 1.081. The first two eigenvalues, and especially the corresponding eigenvectors, are, in fact, quite similar to the ones in Exhibit 8*a*. The indications from the remaining three eigenvalues and eigenvectors, however, are different because of the added noise in the data of this example. In particular, in the eigenvector that corresponds to the smallest eigenvalue it is seen that the elements which provide the coefficients for Y_1Y_2 and Y_2^2 are no longer negligible and only the coefficient of Y_1 appears to be essentially the same as it was for the noiseless data.

Exhibit 9*b* shows the data and the quadratic coordinates defined by the eigenvectors corresponding to the smallest and largest eigenvalues. The smallest eigenvalue now leads to an elliptical coordinate system because of the influence of the statistical errors. This distortion of the parabolic coordinate into an elliptical one, however, does not seem to be unreasonable relative to the configuration of the data.

The illustration of the technique of generalized principal components in Examples 8 and 9 is trivial since $p = 2$ for both cases. With $p = 3$ more interesting possibilities begin to arise since the points may then lie on ($q =$) one- or two-dimensional curved subspaces. Example 10 is a case with $p = 3$ and $q = 2$. The purpose of Examples 8 and 9 is to illustrate the basic concept of nonlinear coordinate transformations and the analytical and algorithmic aspects of the method, which, of course, are valid for any value of p. The

Exhibit 9b. Data and coordinates from the largest and smallest eigenvalues



method is not crucial for two- or three-dimensional problems which lend themselves to graphical representation and study.

The above development of generalized principal components analysis in terms of minimizing (or maximizing) variances, and of requiring the different coordinates to be uncorrelated in the sample, is a statistical extension of Hotelling's (1933) approach for the linear case. However, to see the problem of nonlinear singularities in a broader context of nonlinear coordinate systems, one can formulate the question in function-fitting terms. Thus, in the linear case, the eigenvector corresponding to the smallest eigenvalue essentially determines a plane of closest fit, where closeness is measured by the sum of squares of perpendicular distances. Specifically, in the notation of Section 2.2.1, if $\mathbf{a}'_1, \ldots, \mathbf{a}'_p$ denote the eigenvectors of $\mathbf{S}$ corresponding, respectively, to the ordered eigenvalues $c_1 \geqslant c_2 \geqslant \cdots \geqslant c_p > 0$, the equation of the plane of closest fit to the data is $\mathbf{a}'_p\mathbf{y} = \mathbf{a}'_p\bar{\mathbf{y}}$, where $\bar{\mathbf{y}}$ is the sample mean vector (see Eq. 1). Also, among all planes orthogonal to the first, the equation of the plane of next closest fit to the data is $\mathbf{a}'_{p-1}\mathbf{y} = \mathbf{a}'_{p-1}\bar{\mathbf{y}}$, and so on. This indeed was Karl Pearson's (1901) formulation leading to the linear principal components.

In the linear case the statistical approach through variance minimization and mutual uncorrelatedness turns out to be identical with the approach of fitting mutually orthogonal planes by minimizing the sum of squares of perpendicular distances from the data to the planes. This equivalence of the two approaches does not, however, carry over to the general nonlinear case with "noisy" data. Developing algorithms for fitting even specific types (e.g., quadratic polynomials) of nonlinear functions to data by minimizing the sum of squares of "perpendicular" distances would be quite useful.

One value of such an algorithm would be the intuitive appeal of its criterion in function-fitting situations involving variables all of which may be subject to random errors, a circumstance which is not rare and in which the use of the usual least squares criterion may be questionable. Another feature of the function-fitting approach would make it particularly appealing as a statistical tool for the reduction of dimensionality. Unlike the eigenvector algorithm involved in the variance-minimization approach, the function-fitting algorithm is sensitive only to the scales of the original variables and not to the scales of additional nonlinear functions of them. For instance, in the bivariate quadratic case discussed earlier, the solution of the enlarged eigenvector problem is sensitive not only to the scales of Y_1 and Y_2 but also to those of Y_1^2, Y_2^2, and $Y_1 Y_2$, which are necessarily noncommensurable with the original coordinates. However, the criterion of minimizing the sum of squares of perpendicular deviations, although dependent on the scales of Y_1 and Y_2, does not depend in any way on the scales of the quadratic terms Y_1^2, Y_2^2, and $Y_1 Y_2$. This scale-resistant nature would be particularly desirable in many practical applications. One way of handling this issue, while still retaining the variance-minimization approach, is to carry out the eigenanalysis on the enlarged correlation matrix rather than on the corresponding enlarged covariance matrix, an idea discussed earlier (see pp. 11–12) in the context of linear principal components analysis.

A main advantage of the variance-minimization approach is the computational simplicity of the algorithm involved—merely an eigenanalysis. In the absence of a general algorithm for function fitting based on minimizing perpendicular deviations, one is limited to this approach anyway. In the absence of noise in the data, as in Examples 8 and 10, the issue of differences between the two approaches does not arise. When the variance-minimization approach is used for generalized principal components analysis, the equation of the nonlinear subspace to which the observations may possibly be confined will be defined by means of a procedure analogous to the one employed in the linear case. Specifically, for instance, in the bivariate case considered above, if the observations lie on a quadratic curve (as in Example 8), the variance-minimization approach will be expected to lead to $c_5^* = 0$ and the equation of the quadratic curve will be $(z_5 =)\ \mathbf{a}_5^{*\prime}\mathbf{y}^* = \mathbf{a}_5^{*\prime}\bar{\mathbf{y}}^*$, where $\mathbf{y}^*$, $\bar{\mathbf{y}}^*$, c_5^*, and $\mathbf{a}_5^*$ were defined earlier. In fact, the "middle" one of the five parabolas in Exhibit 8*b* and the "middle" one of the five ellipses in Exhibit 9*b* have equations that are specified in this manner. The same idea is also used in obtaining the equation of the quadratic surface involved in the next example.

Example 10. The data, shown in Exhibit 10*a*, consist of 19 points lying on the surface of the sphere, $X^2 + Y^2 + Z^2 = 25$. Thus, in this example, $p = 3$ and $q = 2$.

Exhibit 10a. Artificial data—19 points on the surface of a sphere

X	Y	Z
−5.0	0.0	0.0
−4.0	1.0	±2.828
−3.0	0.5	±3.969
−2.0	4.0	±2.236
−1.0	0.0	±4.899
0.0	3.0	±4.000
1.0	2.0	±4.472
2.0	4.0	±2.236
3.0	3.3	±2.261
4.0	2.4	±1.800

A quadratic principal components analysis, involving an eigenanalysis of dimension nine, yields the eigenvalues and eigenvectors shown in Exhibit 10*b*. The largest eigenvalue is seen to be about 96, while the smallest, which in the absence of "noise" in the data is 0, was computed to be less in value than 10^{-7} and hence is shown as being 0. Also shown at the bottom of Exhibit 10*b* is the equation of the sphere on which the data lie as determined by the eigenvector for the zero (smallest) eigenvalue. With the error-free data, the original sphere is recovered. Since $p = 3$, it is possible to represent the data of this example and the fitted sphere in a stereoscopic three-dimensional display, which can be obtained by using current capabilities in computer software and graphical aids.

In utilizing polynomial principal components, such as the quadratic one illustrated in Examples 8–10, the expectation is that these will respond to local nonlinearities. The use of quadratic analysis may produce a significant improvement in sensitivity to local nonlinearities, and the hope is that one will not need polynomials of very high degree for accommodating most nonlinearities met in practice.

The choice of the degree of polynomials to be used also has certain implications for the number, n, of observations that will be required. When $p = 1$, of course, one needs at least two observations to obtain a nonzero estimate of variance; and similarly, when $p = 2$, one needs $n \geqslant 3$ observations for obtaining a nonsingular estimate of the covariance matrix. In the same vein, for a problem in p-space, a quadratic principal components analysis will involve the eigenvector solution for a

$$\left[2p + \frac{p(p-1)}{2}\right] \times \left[2p + \frac{p(p-1)}{2}\right]$$

matrix, so that a nontrivial solution can be obtained only if $n > 2p + [p(p-1)/2]$; thus, for $p = 8$, n must exceed 44. Using a cubic coordinate system

Exhibit 10*b*. Eigenanalysis for quadratic principal components analysis of data in Exhibit 10*a*

				Eigenvalues				
96.3697	62.7167	61.6238	49.5563	34.4642	2.5909	0.9510	0.0779	0.0000
				Eigenvectors				
0.0496	0.2371	0.0000	0.0000	−0.3170	−0.0000	0.8900	−0.2209	−0.0001
0.0467	0.1698	−0.0000	0.0000	0.0494	−0.0000	0.2082	0.9608	−0.0006
−0.0000	−0.0000	0.2737	0.3037	−0.0000	0.9126	0.0000	0.0000	−0.0000
0.2259	0.2665	0.0000	0.0000	−0.8573	0.0000	−0.3723	0.0667	0.0001
0.0000	−0.0000	0.4116	−0.8946	−0.0000	0.1742	0.0000	0.0000	−0.0000
−0.0000	−0.0000	0.8693	0.3280	0.0000	−0.3699	−0.0000	−0.0000	0.0000
0.5745	−0.5639	0.0000	0.0000	−0.0665	−0.0000	0.1073	0.0515	−0.5774
0.1867	0.7103	0.0000	−0.0000	0.3121	−0.0000	−0.1196	−0.1251	−0.5773
−0.7612	−0.1464	0.0000	0.0000	−0.2457	−0.0000	0.0119	0.0726	−0.5774

Equation obtained by QPCA

$0.577x^2 + 0.577y^2 + 0.577z^2 = 14.4$

with $p = 5$ will require $n > 55$. In practice, one could handle the difficulty caused by such requirements on n by first performing a linear principal components analysis and then pursuing nonlinear analysis, using only the first few linear principal components (i.e., the ones with largest variances).

The magnitude of the eigenvector computation grows rapidly, both with the degree of the polynomial coordinate system considered and with the dimension of the response. With $p = 5$ a completely general cubic principal components analysis would involve a 55-dimensional eigenanalysis. For the numerical computations in these eigenanalyses, it would therefore be desirable to use the singular value decomposition method (see Businger & Golub, 1969) mentioned earlier in connection with linear principal components.

Certain advantages of developing a function-fitting approach to generalized principal components analysis have been mentioned already. A further, possibly intangible, advantage of viewing generalized principal components analysis in the framework of function fitting is that the latter area has received considerable attention in statistics, both conceptually and methodologically, under categories such as linear and nonlinear regression. The available methodology of these familiar areas may, after appropriate modifications, be relevant for further extending the usefulness of the generalized principal components approach. For instance, with a projected class of coordinates that involves arbitrary functions of the response variables, with some unspecified coefficients that may occur *nonlinearly*, the mathematical problem is still just that of finding members of the class which give closest fit, and nonlinear-fitting ideas and procedures may prove useful. Concepts and methods for linearizing the nonlinear problem and for iterative solutions may carry over. The eigenvector algorithms used earlier cannot be applied simply to yield the solution, although one may be able to use them iteratively to develop an approximate solution.

In the context of the new paradigm for multivariate data analysis techniques mentioned in the preface to this edition, Donnell et al. (1994) have suggested a nonlinear generalization of principal components called *additive principal components*.

REFERENCES

Section 2.2.1 Blackith (1960, Blackith & Roberts (1958), Businger (1965), Businger & Golub (1969), Golub (1968), Hotelling (1933), Pearson (1901), Rao (1965), Seal (1964).

Section 2.2.2 Anderson & Rubin (1956), Bartett (1951), Carroll (1969), Devlin et al. (1975), Fletcher & Powell (1963), Harman (1967), Hoerl & Kennard (1970), Howe (1955), Imbrie (1963), Imbrie & Kipp (1971), Imbrie & Van Andel (1964), Jöreskog (1967), Jöreskog & Lawley (1968), Lawley (1940, 1967), Lawley & Maxwell (1963), McDonald (1962, 1967), Seal (1964), Theil (1963), Thomson (1934), Thurstone (1931).

Section 2.3 Abelson & Tukey (1959), Bartholomew (1959), Barton & Mallows (1961), Bennett (1965), Boynton & Gordon (1965), Carroll & Chang (1970), Coombs (1964), Ekman (1954), Kruskal (1964a, b), Miles (1959), Rothkopf (1957), Shepard (1962a, b, 1963), Shepard & Carroll (1966), Van Eeden (1957a, b).

Section 2.4 Businger & Golub (1969), Donnell et al. (1994), Gnanadesikan & Wilk (1966, 1969), Van de Geer (1968), Wilk, Gnanadesikan & Huyett (1962a).

Methods for Statistical Data Analysis of Multivariate Observations, Second Edition
by R. Gnanadesikan

CHAPTER 3

Development and Study of Multivariate Dependencies

3.1. GENERAL

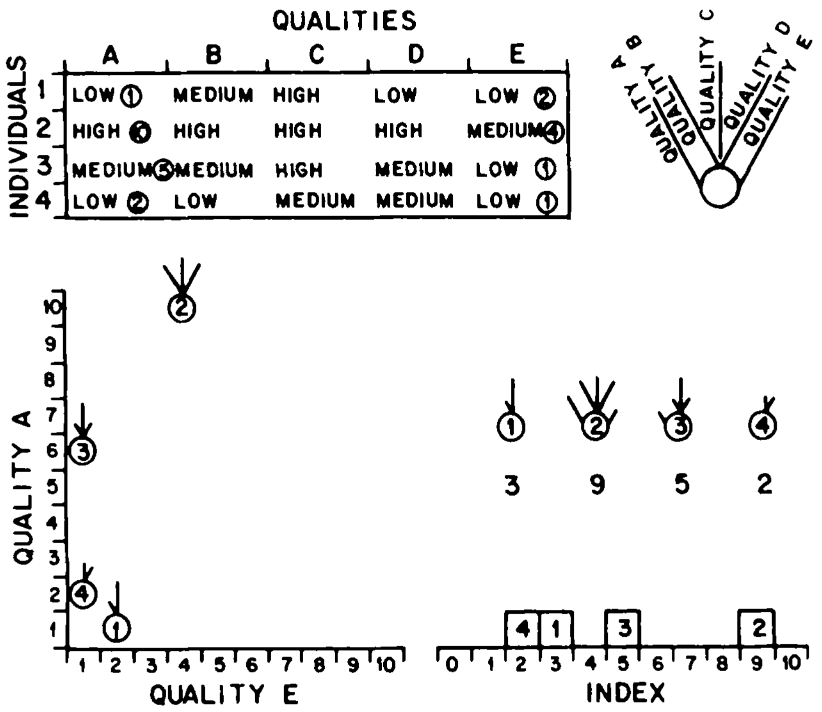
The general concern here is with the study of dependencies, both association and relationship, among several responses. It is possible to delineate two broad categories of multivariate dependencies: (i) those that involve only one set of multivariate responses, and (ii) dependencies of one set of responses on other sets of responses, or on extraneous design or regressor variables. The first category of dependencies may be called *internal*; the second, *external*.

3.2. INTERNAL DEPENDENCIES

For a case with n observations on a p-dimensional response vector, the familiar techniques of computing and studying simple and various partial correlation coefficients are examples of methods for studying the relative degrees of association among the p responses. A pictorial technique for displaying association, which can be useful at times, has been discussed by Anderson (1954, 1957, 1960) under the name *glyphs*, including generalized glyphs and metroglyphs. An illustrative example follows.

Example 11. Exhibit 11, taken from Anderson (1960), pertains to an example involving observed measures of five ($=p$) qualities as possessed by each of four ($=n$) individuals. The table of data is shown at top left. For each individual, a graphical representation called a glyph may be obtained in which each quality is pictured as a ray emanating from a circle corresponding to an individual (see top right of Exhibit 11). The position of a ray in such a glyph corresponds to one of the qualities, while the length of the ray reflects the level of the quality—a long ray indicating high level, a short one representing medium level, and no ray at all corresponding to low level. Thus, in the metroglyphs for all four individuals shown in the middle on the right side of

Exhibit 11. Metroglyph (Anderson, 1960)



INDIVIDUALS	A	B	C	D	E
1	LOW ①	MEDIUM	HIGH	LOW	LOW ②
2	HIGH ⑩	HIGH	HIGH	HIGH	MEDIUM ④
3	MEDIUM ⑥	MEDIUM	HIGH	MEDIUM	LOW ①
4	LOW ②	LOW	MEDIUM	MEDIUM	LOW ①

Exhibit 11, at a glance one can see that individual 2 is high on most qualities whereas individual 1 is low on many.

An alternative representation, which contains fewer rays and may be particularly useful for depicting associations among the qualities, is the scatter diagram shown at bottom left of Exhibit 11. Here the rays corresponding to two of the qualities (*A* and *E*) have been dropped from the glyphs, and coordinate axes with a finer degree of quantization of levels (10 instead of just 3) of *A* and *E* are introduced. The values of *A* and *E* actually observed on scales ranging from 1 to 10 are shown encircled for each of the four individuals in the data table at the top left of Exhibit 11. These values provide the coordinates for the placement of the glyphs in the bottom left picture. As stated by Anderson (1957, 1960), this representation of the four pentavariate observations in this example shows that qualities *B* and *C*, as also *B* and *D*, are associated. Moreover, one can see a fairly strong association between *B*, *C*, *D* (both individually and concurrently) and quality *A*, while only a weak association is manifest between the former set of qualities and quality *E*.

As an overall numerical summary, Anderson (1957, 1960) obtains an index for each glyph by scoring 2 for each long ray (i.e., high level of a quality), 1 for each short ray (i.e., medium level), and 0 for absence of a ray. Values of this index are given below the metroglyphs, and a histogram is shown at the bottom right of Exhibit 11.

Although glyphs and metroglyphs are seldom used these days, the essential idea of representing multivariate observations graphically so as to help get an overall impression of the data underlies many current schemes for graphical displays (see Chapter 5 of Chambers et al., 1983). For instance, a display known as a *snowflake plot* or a *star plot* is one in which the values of the different variables are coded into lengths of rays emanating from a center. Two other multiresponse data displays that use symbolic representation and/or more familiar scatter plotting are illustrated next. The first of these, called a *weathervane plot* by Bruntz et al. (1974), was developed for analyzing air pollution data, and the next example illustrates the techniquc.

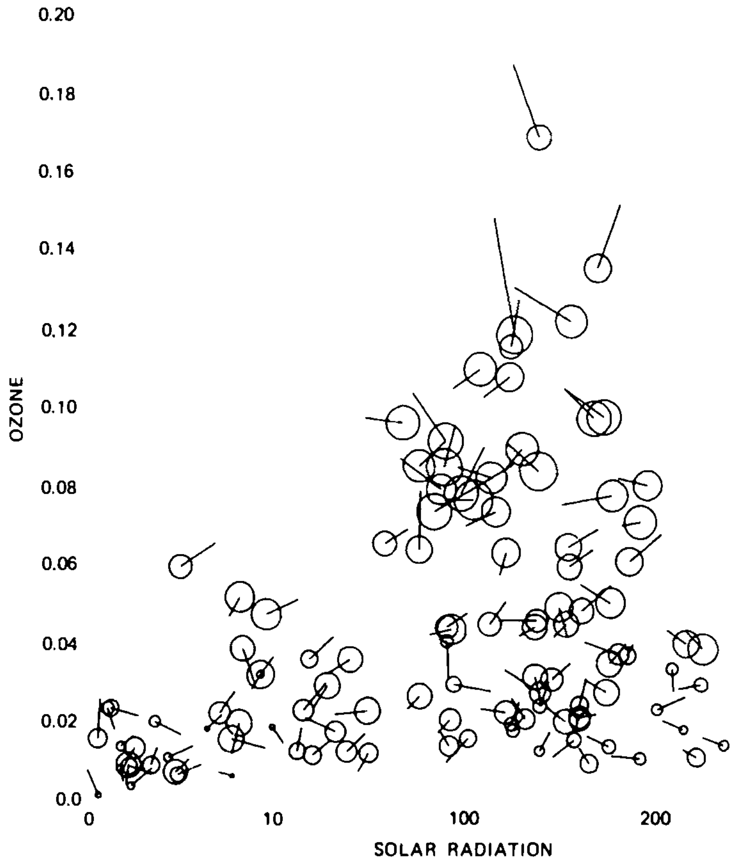
Example 12. In analyzing air pollution data it is often appropriate to consider not only the chemical reactions involved in producing the pollutants but also the prevailing meteorological conditions. Specifically, solar radiation, wind speed and direction, and temperature are some of the variables of interest.

Exhibit 12, taken from Bruntz et al. (1974), shows a weathervane plot in which the abscissa is the total solar radiation from 8 A.M. to noon, while the ordinate is the average ozone level observed from 1 P.M. to 3 P.M. The plot pertains to a specific site, and the points (centers of the circles) correspond to different days. As a two-dimensional scatter plot, the centers of the circles in Exhibit 12 provide information on the relationship between ozone levels and solar radiation. In addition, the diameter of the circle plotted has been coded to be proportional to the observed daily maximum temperature, while the line emanating from the circle is coded with information regarding the wind. The length of the line segment is inversely proportional to an average wind speed. If the lines are considered as arrows whose heads are at the centers of the circles, the orientations of the lines correspond to average wind directions. For instance, for the point at the top of Exhibit 12 (i.e., the day with highest ozone), the average wind direction is from the northwest.

One indication of the plot is that ozone levels do not become high when there is low solar radiation. However, the presence of points in the lower right-hand part of Exhibit 12 suggests that high solar radiation alone does not guarantee high ozone levels, and for the days in this part of the picture the temperature is generally low and the wind speed is generally high. Also, at a given level of solar radiation, ozone seems to increase as temperature increases and wind speed decreases. Wind direction does not seem to be a dominant factor in influencing patterns. Thus the five-dimensional display in Exhibit 12 enables one to get a "feeling" for some of the interrelationships among the five variables involved in this example.

The second pictorial representation that is analogous to glyphs in spirit is a novel scheme proposed by Chernoff (1973). The idea is to code the values of thc variables by associating them with different characteristics of a human face.

Exhibit 12. Weathervane plot of air pollution data (Bruntz et al., 1974)



An important issue in connection with Chernoff's scheme that needs further study is how to go about associating the variables with different aspects of a face in any specific application. Developing guidelines for doing this would be valuable. The next example illustrates the procedure.

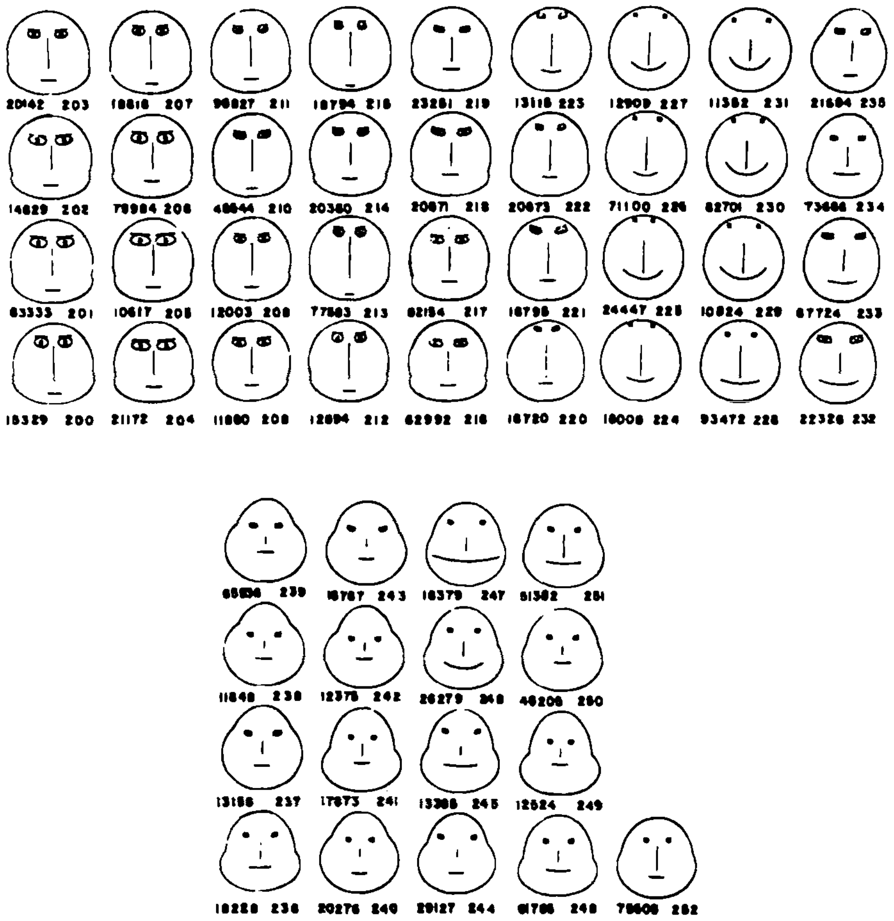
Example 13. This example, taken from Chernoff (1973), pertains to 12-dimensional data on mineral contents assayed on 53 equally spaced specimens taken from a 4500-foot core drilled into a Colorado mountainside. Exhibit 13*a* shows the numerical data, while Exhibit 13*b* shows the 53 faces as obtained by Chernoff. Some major changes in the values of the variables are noticed even on inspection of the numbers in Exhibit 13*a* (e.g., values of Z_5 for specimens 220–233). The breakpoints are clearly visible in the sequence of faces too. For instance, there is an abrupt change in the overall shape of the head and the location and shape of the eyes after the face for specimen 219. Also striking are the smile and the small, high eyes of the faces for specimens 224–231. In the

Exhibit 13*a*. Data for faces (Chernoff, 1973)

DATA ON 12 VARIABLES REPRESENTING MINERAL CONTENTS FROM A 4500-FOOT CORE DRILLED FROM A COLORADO MOUNTAINSIDE

ID	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	Z_{11}	Z_{12}
200	320	105	057	050	001	001	001	060	020	250	210	370
201	280	150	040	050	001	001	001	060	040	210	130	420
202	260	165	033	050	001	001	001	060	010	250	090	440
203	305	110	044	040	001	001	001	050	050	260	140	250
204	290	160	035	035	001	001	001	050	020	210	060	510
205	275	130	047	035	001	001	001	050	020	230	090	570
206	280	155	035	035	001	001	001	080	020	270	170	400
207	300	115	050	060	001	001	001	120	010	280	190	300
208	250	130	041	030	005	001	001	070	030	250	110	330
209	285	120	047	040	001	001	001	070	010	240	170	280
210	280	105	047	070	001	001	001	060	020	370	070	300
211	300	135	050	040	001	001	001	120	060	250	160	200
212	280	110	056	050	001	001	001	150	010	280	270	280
213	305	080	065	080	005	001	001	130	010	300	260	260
214	230	175	029	035	001	001	001	270	030	250	140	240
215	325	060	052	090	001	001	001	160	010	280	260	170
216	270	170	025	040	001	001	001	160	010	290	070	330
217	250	185	031	025	001	001	001	120	001	260	080	330
218	260	185	030	015	001	001	001	270	080	480	010	330
219	270	185	032	010	005	001	001	180	040	450	020	220
220	325	045	053	005	020	001	001	600	080	660	020	250
221	315	090	047	005	020	001	001	410	200	600	060	260
222	335	100	047	010	040	001	001	360	080	590	110	170
223	310	010	049	005	080	018	001	640	240	630	060	190
224	410	001	049	001	075	032	001	760	440	800	001	001
225	360	001	048	001	080	055	001	770	260	770	010	010
226	310	015	051	001	105	036	001	660	380	640	001	010
227	420	005	049	001	095	056	001	620	520	680	001	001
228	415	020	049	005	025	036	001	370	220	340	001	001
229	420	005	041	001	070	060	001	630	510	580	001	001
230	450	005	040	001	090	070	001	690	570	630	001	001
231	395	001	025	015	100	071	001	580	530	560	001	010
232	380	010	027	025	035	039	001	350	320	400	001	270
233	430	010	025	030	030	025	001	340	340	360	001	200
234	410	075	022	010	005	015	001	170	170	170	001	060
235	520	055	024	040	005	001	001	210	190	190	001	180
236	385	135	018	010	005	008	001	140	200	260	001	020
237	535	065	010	020	001	001	001	110	230	270	001	070
238	550	095	001	010	001	001	001	050	230	270	001	030
239	510	100	001	001	001	001	001	190	150	230	001	110
240	510	095	001	040	001	001	001	140	100	150	001	040
241	385	180	010	001	001	001	001	050	050	300	001	050
242	505	125	001	001	001	001	001	001	200	130	001	030
243	470	090	001	020	001	001	001	160	300	380	001	060
244	465	110	001	035	001	001	001	260	440	500	001	060
245	400	140	001	015	001	023	001	330	400	390	001	040
246	415	105	015	025	040	032	001	220	190	270	001	010
247	435	075	010	015	001	069	001	370	360	500	001	010
248	370	145	010	010	005	012	040	130	080	330	001	030
249	380	210	001	001	001	001	020	070	001	050	001	030
250	430	065	001	005	020	001	075	130	070	300	001	020
251	420	080	030	001	005	026	001	050	100	350	001	050
252	425	060	035	005	001	001	030	100	010	340	001	010
min	250	001	001	001	001	001	001	001	001	050	001	001
max	520	210	065	090	105	071	075	770	570	800	270	570

Exhibit 13*b*. Faces obtained by Chernoff (1973)



absence of more specific information regarding the coding employed in obtaining the faces in this example, however, it is difficult to provide any further detailed interpretations of the data here.

The detection and description of relationships (as against associations) among a set of response variables overlap, in obvious ways, the concerns and methods of reduction of dimensionality discussed in Chapter 2. The detection of linear and nonlinear singularities and the characterization of them are useful not only for uncovering redundancies among the response variables but also for studying the nature of the interrelationships among the variables. Thus most of the techniques discussed in Chapter 2 are also relevant to the objective of studying internal relationships.

3.3. EXTERNAL DEPENDENCIES

The multiple correlation coefficient, often discussed in the context of regressing a single variable on a set of extraneous variables, is one example of a measure of association between two sets of variables wherein one of the sets contains just a single variable. Canonical correlation analysis, developed by Hotelling (1936), is another classical technique for studying associations between two sets of variables. Given a set of variables, $\mathbf{x}$, and another set, $\mathbf{y}$, the basic idea is to find the two linear combinations, one of the x-variables and one of the y-variables, that have maximal correlation; then, from among the two sets of linear combinations orthogonal to those already determined, to select the two with maximal correlation, and so on. In general, if p and q are, respectively, the numbers of x- and y-variables, and if $p \leqslant q$, one can extract p pairs of linear combinations by this process. The derived linear functions will be called *canonical variates* (see also the remarks in Section 4.2). The correlations between pairs of the canonical variates were named *canonical correlations* by Hotelling.

An interesting use of canonical correlations was mentioned in passing in Section 2.2.1, in connection with principal components analysis. Frequently in using the principal components as summaries of multidimensional data, for simplicity of interpretation one may wish to round off the coefficients of some of the variates to "nice" values such as 0, or ± 1 (except for the usual normalization constraint on eigenvectors). The candidates for such rounding off are often based on a subjective assessment of the relative magnitudes of the coefficients in the principal components. The modified linear combinations of the variables obtained by such rounding will not necessarily be orthogonal. Yet, an interesting question would be how close the set of modified principal components is to the original set. One way of assessing the closeness is in terms of the angles between the two linear subspaces spanned by the original and the modified principal components. Since canonical correlations are measures of the cosines of angles between linear subspaces they can be used for this purpose—the higher the canonical correlation the closer the subspaces.

The concepts and techniques of canonical correlation analysis introduced by Hotelling have been extended to the case of more than two sets of variables by various authors (see Steel, 1951; Horst, 1965; Kettenring, 1969, 1971). Kettenring (1969, 1971) provides a unifying discussion of the various approaches, and the following material relies heavily on his treatment.

Given m sets of variables, $\mathbf{y}_j(p_j \times 1)$ for $j = 1, 2, \ldots, m$, suppose that $p_1 \leqslant p_2 \leqslant \cdots p_m$ and $p = \Sigma_{j=1}^{m} p_j$. It is assumed, without any loss of generality for present purposes, that $\mathscr{E}(\mathbf{y}_j) = \mathbf{0}$ for all j and that the $p_j \times p_j$ covariance matrix of $\mathbf{y}_j$ is nonsingular and denoted as $\boldsymbol{\Sigma}_{jj}$. The Cholesky decomposition, $\boldsymbol{\Sigma}_{jj} = \boldsymbol{\tau}_j\boldsymbol{\tau}_j'$, may then be utilized to obtain a linear transformation of $\mathbf{y}_j$ to $\mathbf{x}_j = \boldsymbol{\tau}_j^{-1}\mathbf{y}_j$ such that the covariance matrix of $\mathbf{x}_j$ is the identity matrix of order p_j. Now, if

$$\mathbf{y}' = (\mathbf{y}_1' \vdots \mathbf{y}_2' \vdots \cdots \vdots \mathbf{y}_m') \quad \text{and} \quad \mathbf{x}' = (\mathbf{x}_1' \vdots \mathbf{x}_2' \vdots \cdots \vdots \mathbf{x}_m'),$$

the covariance matrices of $\mathbf{y}$ and $\mathbf{x}$, denoted as $\mathbf{\Sigma}$ and $\mathbf{\Gamma}$, respectively, are

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} & \cdots & \mathbf{\Sigma}_{1m} \\ \mathbf{\Sigma}'_{12} & \mathbf{\Sigma}_{22} & \cdots & \mathbf{\Sigma}_{2m} \\ \multicolumn{4}{c}{\cdots\cdots\cdots\cdots\cdots\cdots} \\ \mathbf{\Sigma}'_{1m} & \mathbf{\Sigma}'_{2m} & \cdots & \mathbf{\Sigma}_{mm} \end{pmatrix} \tag{42}$$

and

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{I} & \mathbf{\Gamma}_{12} & \cdots & \mathbf{\Gamma}_{1m} \\ \mathbf{\Gamma}'_{12} & \mathbf{I} & \cdots & \mathbf{\Gamma}_{2m} \\ \multicolumn{4}{c}{\cdots\cdots\cdots\cdots\cdots\cdots} \\ \mathbf{\Gamma}'_{1m} & \mathbf{\Gamma}'_{2m} & \cdots & \mathbf{I} \end{pmatrix}, \tag{43}$$

where $\mathbf{\Gamma}_{ij} = \boldsymbol{\tau}_i^{-1}\mathbf{\Sigma}_{ij}(\boldsymbol{\tau}_j^{-1})'$.

Loosely speaking, one wishes to find linear functions of the variables (i.e., canonical variates) in each of the m sets so as to satisfy criteria that are specified in terms of the intercorrelations among the linear functions. Let ${}_jz_1 = {}_j\boldsymbol{\alpha}'_1\mathbf{x}_j = {}_j\boldsymbol{\beta}'_1\mathbf{y}_j$ (where ${}_j\boldsymbol{\beta}'_1 = {}_j\boldsymbol{\alpha}'_1\boldsymbol{\tau}_j^{-1}$), for $j = 1, 2, \ldots, m$, denote the m linear functions, one from each of the m sets, at the first stage. Let the coefficients of the linear combinations be required to satisfy the normalizing constraints ${}_j\boldsymbol{\beta}'_1\cdot\mathbf{\Sigma}_{jj}\cdot{}_j\boldsymbol{\beta}_1 = {}_j\boldsymbol{\alpha}'_1\cdot{}_j\boldsymbol{\alpha}_1 = 1$, so that the variance of ${}_jz_1$ is 1 for all j. Suppose that $\mathbf{z}'_1 = ({}_1z_1, {}_2z_1, \ldots, {}_mz_1)$ denotes the set of m first-stage canonical variates whose correlation (and covariance) matrix is

$$\mathbf{\Phi}(1) = \begin{pmatrix} 1 & \phi_{12}(1) & \cdots & \phi_{1m}(1) \\ & \ddots & & \vdots \\ \phi_{1m}(1) & \phi_{2m}(1) & \cdots & 1 \end{pmatrix} = \mathbf{D}_{\beta_1}\mathbf{\Sigma}\mathbf{D}'_{\beta_1} = \mathbf{D}_{\alpha_1}\mathbf{\Gamma}\mathbf{D}'_{\alpha_1}, \tag{44}$$

where the $m \times p$ matrices, $\mathbf{D}_{\alpha_1}$ and $\mathbf{D}_{\beta_1}$, are block-diagonal and are defined by

$$\mathbf{D}_{\alpha_1} = \begin{pmatrix} {}_1\boldsymbol{\alpha}'_1 & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & {}_2\boldsymbol{\alpha}'_1 & \cdots & \mathbf{0}' \\ \multicolumn{4}{c}{\cdots\cdots\cdots\cdots\cdots\cdots} \\ \mathbf{0}' & \cdots & & {}_m\boldsymbol{\alpha}'_1 \end{pmatrix} \tag{45}$$

and

$$\mathbf{D}_{\beta_1} = \begin{pmatrix} {}_1\boldsymbol{\beta}'_1 & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & {}_2\boldsymbol{\beta}'_1 & \cdots & \mathbf{0}' \\ \multicolumn{4}{c}{\cdots\cdots\cdots\cdots\cdots\cdots} \\ \mathbf{0}' & \cdots & & {}_m\boldsymbol{\beta}'_1 \end{pmatrix}. \tag{46}$$

The criteria for choosing the first-stage canonical variates $\mathbf{z}'_1$ (i.e., for choosing ${}_j\boldsymbol{\alpha}'_1$ or, equivalently, ${}_j\boldsymbol{\beta}'_1$) are all specifiable in terms of the matrix $\mathbf{\Phi}(1)$.

For instance, Horst (1965) proposed the criterion of maximizing the sum,

$$\sum_{i<j=1}^{m}\sum \phi_{ij}(1),$$

of the intercorrelations among the elements of $\mathbf{z}_1$, which is equivalent to maximizing the quadratic form,

$$\mathbf{1}'\mathbf{\Phi}(1)\mathbf{1}\left(=m+2\sum_{i<j=1}^{m}\sum \phi_{ij}(1)\right).$$

The method based on this criterion, which takes into account both the magnitudes and the signs of ϕ_{ij}'s, will be called the *SUMCOR* method. A second criterion, also due to Horst (1965), is to maximize the variance of the first principal component of $\mathbf{z}_1$. If $m > {}_1\lambda_1 \geqslant {}_2\lambda_1 \geqslant \cdots \geqslant {}_m\lambda_1 > 0$ denote the ordered eigenvalues of $\mathbf{\Phi}(1)$, the second criterion amounts to maximizing ${}_1\lambda_1$, and the method associated with this criterion will be called the *MAXVAR* procedure. Kettenring (1971) proposed two additional criteria leading to methods termed *SSQCOR* and *MINVAR*, respectively. The first of these attempts to maximize the sum of squares,

$$\sum_{i<j=1}^{m}\sum \phi_{ij}^2(1),$$

of the off-diagonal elements of $\mathbf{\Phi}(1)$, which is equivalent to maximizing the trace of

$$\mathbf{\Phi}^2(1)\left(=m+2\sum_{i<j=1}^{m}\sum \phi_{ij}^2(1)\right),$$

or the sum of squares of the eigenvalues, $\Sigma_{j=1}^{m}\,{}_j\lambda_1^2$. Unlike SUMCOR, SSQCOR takes account only of the magnitudes of the intercorrelations among the canonical variates. The SSQCOR criterion is also interpretable as maximizing the "distance" between $\mathbf{\Phi}(1)$ and the identity matrix of order m. MINVAR uses the criterion of minimizing the variance of the "smallest" principal component of $\mathbf{z}_1$, that is, minimizing ${}_m\lambda_1$. The first attempt at generalizing Hotelling's two-set canonical correlation analysis is due to Steel (1951), who used the criterion of minimizing the so-called generalized variance of $\mathbf{z}_1$, namely, $|\mathbf{\Phi}(1)|$ or, equivalently, the product, $\Pi_{j=1}^{m}\,{}_j\lambda_1$, of the eigenvalues of $\mathbf{\Phi}(1)$. The method associated with this criterion will be called the *GENVAR* technique.

Each of the five methods mentioned in the preceding paragraph employs a particular unidimensional summary of the matrix of intercorrelations among the canonical variates and determines the set of canonical variates optimally with respect to that summary. In each approach the higher-stage canonical variates, $\mathbf{z}_2, \mathbf{z}_3, \ldots, \mathbf{z}_{p_1}$, are chosen by using the same criterion at each stage and imposing additional constraints (e.g., mutual orthogonality of the different linear combinations within any given set of the original m sets of variables) to ensure that new canonical variates are being found at each stage. When $m = 2$, all five of the methods reduce to Hotelling's (1936) treatment of the problem.

Kettenring (1971) uses "factor-analytic" types of models (see Section 2.2.2) to motivate each of the five criteria and to discuss the similarities and differences that one might anticipate among the results of using the five methods for analyzing a given body of data. Thus, for instance, the SUMCOR and MAXVAR methods may be motivated by a single-common-factor model for the canonical variates:

$$\mathbf{z}_1 = \boldsymbol{\gamma}_1 f_1 + \mathbf{e}_1,$$

where f_1 is the single standardized common factor, and $\mathbf{e}_1$ is an m-dimensional vector of unique factors, which can also be considered in more familiar terms as a vector of residual errors, with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Psi}$. Then, assuming that $\boldsymbol{\gamma}_1$ is known and is proportional to the vector $\mathbf{1}$, it can be shown that choosing f_1 so as to minimize $\mathrm{tr}(\boldsymbol{\Psi})$ is equivalent to the SUMCOR procedure. In other words, the SUMCOR method generates a $\mathbf{z}_1$ having the best fitting (in the sense of minimizing the sum of the variances of the residual errors) single common factor, assuming that the factor contributes equally to each of the first-stage canonical variates. On the other hand, with the same single-factor model, choosing both $\boldsymbol{\gamma}_1$ and f_1 so as to minimize $\mathrm{tr}(\boldsymbol{\Psi})$ turns out to be equivalent to the MAXVAR method.

The SSQCOR and GENVAR methods may be motivated by using an m-factor model:

$$\mathbf{z}_1 = \sum_{j=1}^{m} \boldsymbol{\gamma}_{1j} f_j + \mathbf{e}_1.$$

Since the number of common factors is equal to the dimensionality of $\mathbf{z}_1$, the approach of fitting factors so as to minimize a criterion such as $\mathrm{tr}(\boldsymbol{\Psi})$ is no longer adequate for distinguishing between different sets of $\mathbf{z}_1$. However, suppose that one were to choose the f_j's to be the principal components transformations of $\mathbf{z}_1$ (see Section 2.2.1) so that they account for decreasing amounts of variance. In fact, if ${}_j\boldsymbol{\varepsilon}_1$ is the eigenvector corresponding to the (ordered) eigenvalue ${}_j\lambda_1$ of $\boldsymbol{\Phi}(1)$, then suppose that

$$\boldsymbol{\gamma}_{1j} = \sqrt{{}_j\lambda_1}\,{}_j\boldsymbol{\varepsilon}_1 \qquad \text{and} \qquad f_j = \frac{1}{\sqrt{{}_j\lambda_1}} \cdot {}_j\boldsymbol{\varepsilon}_1' \cdot \mathbf{z}_1 \qquad \text{for } j = 1, \ldots, m,$$

so that the f_j's are the standardized and mutually uncorrelated principal components derived from $\Phi(1)$. The factors derived from the largest and the smallest eigenvalues will be of particular interest, and seeking a $\mathbf{z}_1$ that corresponds to large separations among the eigenvalues will therefore be useful. One way of obtaining such a $\mathbf{z}_1$ is to choose them so as to maximize the measure of spread, $\Sigma_{j=1}^{m} {}_j\lambda_1^2$, subject to the constraint that the sum of the ${}_j\lambda_1$'s has to equal m. This, of course, is what the SSQCOR method attempts to do. Hence the SSQCOR method will tend to produce a $\mathbf{z}_1$ such that its first few principal components account for most of the variability. On the other hand, since the GENVAR method attempts to minimize $\Pi_{j=1}^{m} {}_j\lambda_1$, it will be expected to focus on the smallest eigenvalues and to minimize the contribution of the last few f_j's.

The inequalities, $m \leqslant \Sigma_{j=1}^{m} {}_j\lambda_1^2 \leqslant m^2$, can be established, and, furthermore, the lower bound can be shown to be attained when ${}_j\lambda_1 = 1$ for all j (i.e., $\Phi_{(1)} = \mathbf{I}$), while the upper bound is attained when ${}_1\lambda_1 = m$ and ${}_2\lambda_1 = \cdots = {}_m\lambda_1 = 0$. This result suggests that MAXVAR and SSQCOR will yield similar $\mathbf{z}_1$'s whenever most of the variability in $\mathbf{z}_1$ can be accounted for by a single factor.

The MINVAR method may be studied by using a $(m - 1)$-factor model:

$$\mathbf{z}_1 = \sum_{j=1}^{m-1} \gamma_{1j} f_j + \mathbf{e}_1.$$

For a given $\mathbf{z}_1$, choosing γ_{1j}'s and f_j's so as to minimize the trace of the covariance matrix of the residual error variables, $\mathbf{e}_1$, leads in this case to

$$\gamma_{1j} = \sqrt{{}_j\lambda_1}\, {}_j\boldsymbol{\varepsilon}_1 \qquad \text{and} \qquad f_j = \frac{1}{\sqrt{{}_j\lambda_1}} \cdot {}_j\boldsymbol{\varepsilon}_1' \cdot \mathbf{z}_1, \qquad \text{for } j = 1, \ldots, (m - 1),$$

so that the f_j's are the first $(m - 1)$ standardized principal components of $\Phi(1)$. The residual variance after fitting all $(m - 1)$ factors is $m - \Sigma_{j=1}^{m-1} {}_j\lambda_1$ $(= {}_m\lambda_1)$. Now choosing $\mathbf{z}_1$ to optimize the fit by such a set of $(m - 1)$ factors amounts to choosing $\mathbf{z}_1$ so as to minimize ${}_m\lambda_1$, i.e., the MINVAR method. This suggests that MINVAR and GENVAR may be expected to yield similar $\mathbf{z}_1$'s whenever the smallest eigenvalue, ${}_m\lambda_1$, is very small, that is, whenever almost all of the variability in $\mathbf{z}_1$ is confined to an $(m - 1)$-dimensional linear subspace (see the discussion in Example 14).

The above discussion has been presented in terms of "population" entities. With a sample of size n $(> p = \Sigma_{j=1}^{m} p_j)$ on the m sets of variables, one would have the following correspondences between the population entities and statistics computed from the observations:

$$\Sigma \leftrightarrow \mathbf{S} = ((\mathbf{S}_{ij})); \qquad \Gamma \leftrightarrow \mathbf{R} = ((\mathbf{R}_{ij})),$$

where $\mathbf{R}_{ij} = \mathbf{T}_i^{-1}\mathbf{S}_{ij}(\mathbf{T}_j^{-1})'$ and $\mathbf{S}_{jj} = \mathbf{T}_j\mathbf{T}_j'$;

$$\{{}_1\boldsymbol{\alpha}_1, \ldots, {}_m\boldsymbol{\alpha}_1\} \leftrightarrow \{{}_1\mathbf{a}_1, \ldots, {}_m\mathbf{a}_1\};$$
$$\{{}_1\boldsymbol{\beta}_1, \ldots, {}_m\boldsymbol{\beta}_1\} \leftrightarrow \{{}_1\mathbf{b}_1, \ldots, {}_m\mathbf{b}_1\};$$
$$\boldsymbol{\Phi}(1) \leftrightarrow \hat{\boldsymbol{\Phi}}(1); \qquad {}_j\lambda_1 \leftrightarrow {}_j\hat{\lambda}_1; \qquad {}_j\boldsymbol{\varepsilon}_1 \leftrightarrow {}_j\hat{\boldsymbol{\varepsilon}}_1.$$

Kettenring (1969) describes algorithms associated with each of the five methods that may be used with these sample statistics. SUMCOR, SSQCOR, and GENVAR involve iterative techniques, whereas MAXVAR and MINVAR depend only on a single eigenanalysis of the $p \times p$ matrix $\mathbf{R}$. Also, when $m = 2$, which is the case considered by Hotelling (1936), no iterative methods are involved, and all five methods reduce to utilization of the results from an eigenanalysis of $\mathbf{R}$. (See discussion below.)

In fact, for the MAXVAR method, the first-stage canonical variates, for instance, are obtained from the eigenvector corresponding to the largest eigenvalue of $\mathbf{R}$. If $c_1 \geqslant c_2 \geqslant \cdots \geqslant c_p > 0$ denote the ordered eigenvalues of $\mathbf{R}$ with corresponding p-dimensional eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$, where $\mathbf{v}_k'$ in partitioned form is $\{{}_1\mathbf{v}_k' \vdots {}_2\mathbf{v}_k' \vdots \cdots \vdots {}_m\mathbf{v}_k'\}$ for $k = 1, \ldots, p$, then the required solution for the coefficients of the first-stage MAXVAR canonical variates is

$$_j\mathbf{a}_1 = \pm \frac{{}_j\mathbf{v}_1}{\|{}_j\mathbf{v}_1\|}, \qquad j = 1, \ldots, m, \tag{47}$$

where $\|\mathbf{x}\|$ denotes the Euclidean norm (i.e., square root of the sum of squares of the elements) of $\mathbf{x}$. Similarly, the first-stage MINVAR canonical variates are derived from

$$_j\mathbf{a}_1 = \pm \frac{{}_j\mathbf{v}_p}{\|{}_j\mathbf{v}_p\|}, \qquad j = 1, \ldots, m. \tag{48}$$

Canonical variates may also be obtained at additional stages and will depend on the nature of the constraints imposed on them to ensure that they are different from ones determined at earlier stages. (See Kettenring, 1969, for a more detailed discussion.)

As to algorithms for computing the canonical correlations and variates, for the classical case involving just two sets of variables (i.e., $m = 2$), the preferred current method is to perform a singular value decomposition of the $p_1 \times p_2$ matrix, $\mathbf{R}_{12}$. Specifically, if this decomposition is denoted

$$\mathbf{R}_{12} = \mathbf{Q}_1[\mathbf{D}_r \,|\, \mathbf{0}]\mathbf{Q}_2',$$

the diagonal elements of the $p_1 \times p_1$ matrix, $\mathbf{D}_r$, are the required canonical correlations. The vectors of coefficients defining pairs of canonical variates are given by the columns of $\mathbf{T}_1^{-1'}\mathbf{Q}_1$ and $\mathbf{T}_2^{-1'}\mathbf{Q}_2$, respectively. For the case of more than two sets of variables (i.e., $m \geqslant 3$), Chen & Kettenring (1972) describe implementations of the methods discussed above.

Exhibit 14*a*. Correlation matrices of the original and the internally sphericized variables (Horst, 1965; Kettenring, 1971)

$$\mathbf{S} = \mathbf{R}_0 = \left[\begin{array}{ccc|ccc|ccc}
1 & 0.249 & 0.271 & 0.636 & 0.183 & 0.185 & 0.626 & 0.369 & 0.279 \\
 & 1 & 0.399 & 0.138 & 0.654 & 0.262 & 0.190 & 0.527 & 0.356 \\
 & & 1 & 0.180 & 0.407 & 0.613 & 0.225 & 0.471 & 0.610 \\
\hline
 & & & 1 & 0.091 & 0.147 & 0.709 & 0.254 & 0.191 \\
 & & & & 1 & 0.296 & 0.103 & 0.541 & 0.394 \\
 & & & & & 1 & 0.179 & 0.437 & 0.496 \\
\hline
 & & & & & & 1 & 0.291 & 0.245 \\
 & & & & & & & 1 & 0.429 \\
 & & & & & & & & 1
\end{array}\right]$$

$$\mathbf{R} = \left[\begin{array}{c|ccc|ccc}
 & 0.636 & 0.126 & 0.059 & 0.626 & 0.195 & 0.059 \\
\mathbf{I} & -0.021 & 0.633 & 0.049 & 0.035 & 0.459 & 0.129 \\
 & 0.016 & 0.157 & 0.521 & 0.048 & 0.238 & 0.426 \\
\hline
 & & & & 0.709 & 0.050 & -0.002 \\
 & & \mathbf{I} & & 0.039 & 0.532 & 0.190 \\
 & & & & 0.067 & 0.258 & 0.299 \\
\hline
 & & & & & \mathbf{I} &
\end{array}\right]$$

[*Note*: Values in the blocks below the diagonal blocks are obtained by symmetry.]

Example 14. This example, taken from Thurstone & Thurstone (1941) and also used by Horst (1965) and Kettenring (1971), deals with three $(= m)$ sets of scores by several people on three batteries of three tests each, that is, $p_1 = p_2 = p_3 = 3$. The three tests in each battery were intended to measure, respectively, the verbal, numerical, and spatial abilities of the persons tested. Exhibit 14*a* shows the 9×9 covariance matrix of the standardized scores or, equivalently, the correlation matrix $\mathbf{R}_0$ of the original scores. Also shown in Exhibit 14*a* is the 9×9 correlation matrix, $\mathbf{R}$, of the internally "sphericized" standardized variables (the $\mathbf{x}_j$-variables in terms of the earlier descriptions) derived from the standardized scores. The diagonal elements of the matrices in the off-diagonal blocks of $\mathbf{R}_0$ are all relatively large. Thus the scores on tests intended to measure the same ability tend to be highly correlated, whereas the correlations between scores on tests (even within the same battery) measuring different abilities, although positive, are not as high. After the internal transformations of the three sets, the off-diagonal terms of the matrices in the off-diagonal blocks tend to be even smaller (compare $\mathbf{R}$ with $\mathbf{R}_0$).

When the five methods of multiset canonical correlation analysis were applied in this case, with the exception of the MINVAR method the results were similar; that is, differences in the numerical answers occurred only in the third or higher decimal places. The results for the four methods other than

Exhibit 14*b*. Results of five methods of multiset canonical correlation analysis (Kettenring, 1971)

$$
\begin{aligned}
{}_1z_1 &= (0.73,\ 0.51,\ 0.45)\mathbf{x}_1\\
{}_2z_1 &= (0.66,\ 0.62,\ 0.42)\mathbf{x}_2\\
{}_3z_1 &= (0.68,\ 0.64,\ 0.36)\mathbf{x}_3
\end{aligned}
$$

$$
\hat{\Phi}(1) = \begin{pmatrix} 1 & 0.735 & 0.756 \\ & 1 & 0.743 \\ & & 1 \end{pmatrix}
$$

$$
\begin{aligned}
{}_1\hat{\lambda}_1 &= 2.49, & {}_1\hat{\mathbf{e}}'_1 &= (0.578,\ 0.574,\ 0.580)\\
{}_2\hat{\lambda}_1 &= 0.27, & {}_2\hat{\mathbf{e}}'_1 &= (-0.535,\ 0.803,\ -0.262)\\
{}_3\hat{\lambda}_1 &= 0.24, & {}_3\hat{\mathbf{e}}'_1 &= (-0.616,\ -0.159,\ 0.771)
\end{aligned}
$$

$$
\begin{aligned}
{}_1z_1 &= (0.68,\ 0.57,\ 0.45)\mathbf{x}_1\\
{}_2z_1 &= (0.96,\ -0.22,\ 0.16)\mathbf{x}_2\\
{}_3z_1 &= (-0.78,\ -0.53,\ -0.33)\mathbf{x}_3
\end{aligned}
$$

$$
\hat{\Phi}(1) = \begin{pmatrix} 1 & 0.345 & -0.736 \\ & 1 & -0.517 \\ & & 1 \end{pmatrix}
$$

$$
\begin{aligned}
{}_1\hat{\lambda}_1 &= 2.082, & {}_1\hat{\mathbf{e}}'_1 &= (0.591,\ \ 0.493,\ -0.638)\\
{}_2\hat{\lambda}_1 &= 0.683, & {}_2\hat{\mathbf{e}}'_1 &= (0.513,\ -0.839,\ -0.517)\\
{}_3\hat{\lambda}_1 &= 0.235, & {}_3\hat{\mathbf{e}}'_1 &= (0.621,\ \ 0.228,\ \ 0.751)
\end{aligned}
$$

MINVAR which lead to similar answers are shown in the upper portion of Exhibit 14*b*, while those for MINVAR are given in the lower portion. Each set of results in the exhibit pertains only to the first stage of analysis and contains the three first-stage canonical variates, their correlation matrix $\hat{\Phi}(1)$, and the eigenanalysis of $\hat{\Phi}(1)$.

The following features emerge from an inspection of the results in the upper portion of Exhibit 14*b*:

1. The largest eigenvalue ${}_1\hat{\lambda}_1$ is about 83% of $\mathrm{tr}\{\hat{\Phi}(1)\}$, and the corresponding eigenvector ${}_1\hat{\mathbf{e}}_1$ is approximately proportional to the vector $\mathbf{1}$; the latter may be considered an indication that the three sets of variables (viz., the batteries of tests) are so much alike that the three canonical variates which are derived, one from each of them, are contributing equally to the first principal component transformation of $\mathbf{z}_1$.

2. Eigenvalues ${}_2\hat{\lambda}_1$ and ${}_3\hat{\lambda}_1$ are approximately equal, each accounting for 8–9% of $\mathrm{tr}\{\hat{\Phi}(1)\}$; that is, $\hat{\Phi}(1)$ has one large eigenvalue and the other two

eigenvalues are essentially equal, a result that may also be surmised from the equicorrelational nature of $\Phi(1)$, as indicated by the near constancy of its off-diagonal elements (see the discussion of Example 1).

Combining the indications from features (1) and (2), and recalling the earlier discussion pertaining to the similarities and differences among the methods, one can understand the reasons why the SUMCOR, MAXVAR, and SSQCOR methods yield similar results. The reason why GENVAR is not as similar to MINVAR but is more similar to SSQCOR in this example lies, perhaps, in the fact that the smallest eigenvalue is not "small enough." The product function, $\Pi_{j=1}^{m} {}_j\lambda_1$, is especially sensitive to the smallest eigenvalue only for extremely small values of it, and in the present example this is not the case. The results for the MINVAR method in Exhibit 14*b* show that ${}_3\hat{\lambda}_1$ accounts for about 8% (not negligible) of $\text{tr}\{\hat{\Phi}(1)\}$, while ${}_1\hat{\lambda}_1$ and ${}_2\hat{\lambda}_1$ contribute approximately 70% and 22%, respectively.

One use of eigenvectors, such as ${}_1\hat{\mathbf{e}}_1$ for the MAXVAR method and ${}_3\hat{\mathbf{e}}_1$ for the MINVAR method, is to study them for selecting "important" subsets of the sets of variables for further analysis. This is generally done by looking at the relative magnitudes of the elements of the eigenvector involved. Thus, in this example, an examination of ${}_3\hat{\mathbf{e}}_1$ associated with the MINVAR method (see Exhibit 14*b*, lower portion) reveals that the first and third elements are much larger than the second. If one decides, on the basis of this indication, to choose the first and third sets of variables for doing a pairwise canonical correlation analysis, then in this example it does indeed turn out that one would have selected the two sets with the highest two-set canonical correlation. [*Note*: One could also have utilized $\hat{\Phi}(1)$ for this, since the element in its top right corner indicates that the first and third canonical variates at the first stage have a large (in magnitude) correlation.]

An interesting alternative analysis in this example (left as an exercise to the reader) would be to regroup the nine variables into three sets corresponding to the three abilities measured rather than the three batteries of tests. A quite different approach with somewhat different objectives would be to use an analysis-of-variance approach (see Chapters 5 and 6) for studying the relative importance of various "effects" (e.g., differences of batteries, or a time or trend effect if the tests were administered across time). This, however, would require the original scores on the tests.

From the viewpoint of data analysis, analyzing subsets of responses is very important and should not be replaced by a single overall multiresponse analysis. In the context of canonical correlation analysis for m sets of multiple responses, analyses of subsets of the m sets, as well as the study of subsets (pairs, triplets, etc.) of the canonical variates from the m-set analysis, are important. Specifically, plots of the original observations transformed according to the canonical variate transformations taken two, and three, at a time may be valuable. In the case of two-set canonical correlation analysis, such plots are actual "displays" of the computed canonical correlations and may lead to uncovering possibly aberrant observations or peculiar relationships.

Example 15. The use of pairwise canonical variate plots is illustrated with data from a questionnaire study, which was concerned with assessing employees' readership of, and attitudes toward, a company magazine published by their employer for communicating general information.

Exhibit 15*a* shows a plot for the two canonical variates corresponding to the largest canonical correlation, derived from the answsers of 645 employees to two subsets consisting of four questions each. The questions in one subset pertained to the expectations of the respondent regarding the publication, while the other subset was concerned with the respondent's evaluation of its actual performance. Each of the eight questions was answered on a six-point scale, and, as indicated in Exhibit 15*a*, the observed largest canonical correlation was 0.4023. The striking features about the configuration are the "bunching" of points on the right-hand boundary of the plot and the vertical striations evident in it. A subsequent inspection of the data, stimulated by these indications of peculiarities, revealed that a large proportion of the respondents tended to use only the higher values of the six-point scale when dealing with their expectations and only the middle values of the scale in evaluating the performance of the publication. Such tendencies would lead to the peculiarities indicated in the plot of canonical variates, although one could detect their existence by other methods (e.g., histograms of original observations) of displaying the data as well.

Exhibit 15*a*. Pairwise canonical variate plot (canonical correlation = 0.4023; $n = 645$, $p = 4$, $q = 4$)

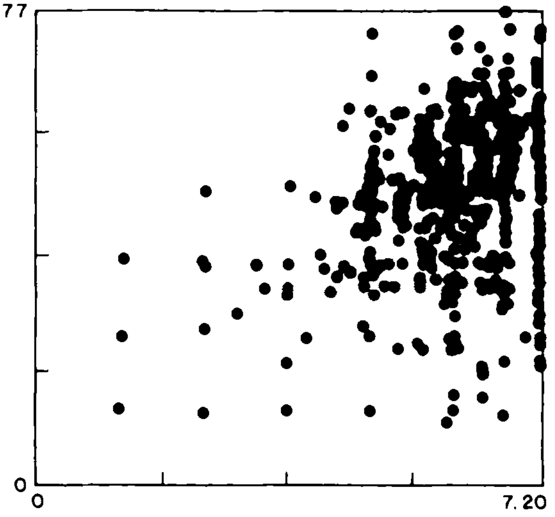


Exhibit 15*b*. Pairwise canonical variate plot (canonical correlation = 0.4833; $n = 580$, $p = 14$, $q = 14$)

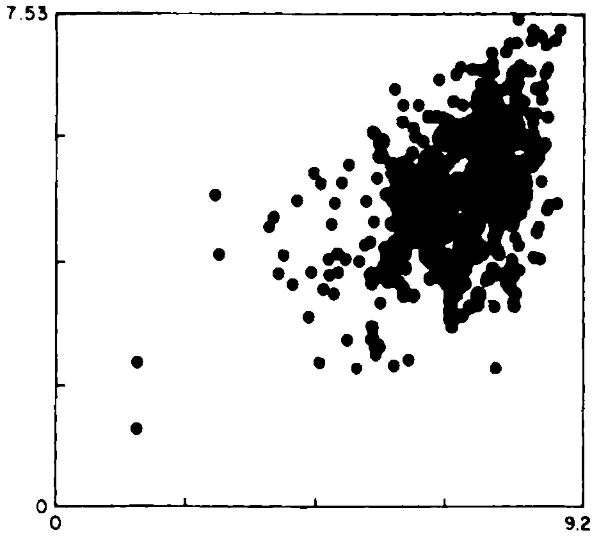


Exhibit 15*b* shows a plot derived from a canonical correlation analysis of two other subsets of questions in the same study. Each subset consisted of 14 questions, and answers from 580 respondents were used in the analysis. The observed value of the largest canonical correlation was 0.4833. The scatter of the points appears to be bounded above by a straight line parallel to a "diagonal line" drawn through the configuration, thus suggesting possible asymmetry in, and departure from normality of, the joint distribution of the canonical variates. There is also a mild suggestion of two outliers in the lower left-hand corner of the plot.

With multiple sets of multiresponse observations, one can also define and determine canonical correlational analogues of partial correlations between scalar variables. Thus, for instance, if $\mathbf{y}_j$ denotes a set of p_j responses, for $j = 1, 2, 3$, one can use, as measures of the first-order partial canonical correlations between any pair of sets $\mathbf{y}_j$ and $\mathbf{y}_k$, given the third set $\mathbf{y}_l$, just the two-set canonical correlations between the "residuals," $\mathbf{r}_j$ and $\mathbf{r}_k$, from the multivariate multiple regressions of $\mathbf{y}_j$ and $\mathbf{y}_k$, respectively, on $\mathbf{y}_l$ $(j \neq k \neq l = 1, 2, 3)$. Similarly, with more than three sets, one can define higher-order partial canonical correlations as well. At each stage only a two-set canonical correlational analysis is involved between sets of "residuals" derived from multivariate multiple regressions of pairs of the original sets of responses on the remaining sets.

The concepts and methods involved in multivariate multiple regression mentioned in the preceding paragraph are utilized widely for studying relationships between a set of response variables, $\mathbf{y}$, and a set of so-called independent variables or regressor variables, $\mathbf{x}$. The multivariate multiple regression model, or the so-called multivariate general linear model (see Roy et al., 1971), is usually specified as follows:

$$\mathbf{Y}' = \mathbf{X} \cdot \mathbf{\Theta} + \boldsymbol{\varepsilon}, \tag{49}$$

where the n rows of $\mathbf{Y}'$ are the n observations on the p-dimensional response variable; the rows of the $n \times k$ matrix, $\mathbf{X}$, are the corresponding observations on k regressor variables; the elements of the $k \times p$ matrix, $\mathbf{\Theta}$, are the unknown regression coefficients; and the n rows of $\boldsymbol{\varepsilon}$ are p-dimensional error variables which are generally assumed to have a mean vector, $\mathbf{0}$, and a common $p \times p$ unknown covariance matrix, $\mathbf{\Sigma}$. The rows of $\boldsymbol{\varepsilon}$ are also generally assumed to be mutually uncorrelated and, for some purposes of formal statistical inference, p-variate normally distributed as well. Thus the n p-dimensional observations are considered to be mutually uncorrelated with means specified by the regression relationships, $\mathscr{E}(\mathbf{Y}' \mid \mathbf{X}) = \mathbf{X\Theta}$, and a common unknown covariance matrix, $\mathbf{\Sigma}$.

The multivariate multiple regression model of Eq. 49 may be rewritten in its equivalent form,

$$\mathbf{Y}' = [\mathbf{Y}_1\mathbf{Y}_2 \cdots \mathbf{Y}_p] = \mathbf{X}[\boldsymbol{\theta}_1\boldsymbol{\theta}_2 \cdots \boldsymbol{\theta}_p] + [\boldsymbol{\varepsilon}_1\boldsymbol{\varepsilon}_2 \cdots \boldsymbol{\varepsilon}_p], \tag{50}$$

where $\mathbf{Y}_j$, the jth column of $\mathbf{Y}'$, consists of the n observations on the jth response, $\boldsymbol{\theta}_j$ consists of the regression coefficients in the univariate multiple linear regression of the jth response variable on the k regressor variables, and $\boldsymbol{\varepsilon}_j$ is an n-dimensional vector of mutually uncorrelated errors pertaining to the jth response variable ($j = 1, 2, \ldots, p$). In this form it is clear that the multivariate model is merely a simultaneous statement of p univariate multiple regression models. In particular, in this treatment the matrix $\mathbf{X}$ is assumed to be the same for all p response variables. When the regressor variables are dummy variables corresponding to factors or treatments in a designed experiment, this means that all p responses are observed under the same design.

In the usual treatment of multivariate multiple regression, the estimate of $\mathbf{\Theta}$ is taken to be $\hat{\mathbf{\Theta}} = [\hat{\boldsymbol{\theta}}_1\hat{\boldsymbol{\theta}}_2 \cdots \hat{\boldsymbol{\theta}}_p]$, where $\hat{\boldsymbol{\theta}}_j = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_j$, for $j = 1, \ldots, p$, are the least squares estimates of the regression coefficients for the jth response analyzed individually. A more detailed discussion of the formal issues, such as the statistical estimation, involved in this approach is presented later in Chapter 5 and may also be found in Roy et al. (1971). For present purposes, however, it is probably worth reiterating that adapting a multivariate view in multiresponse multiple regression situations may be important because the estimated regression coefficients may be statistically dependent because of the intercorrelations of the responses. In other words, although the $\hat{\boldsymbol{\theta}}_j$'s are obtained from separate analyses of the responses, the corresponding elements of the $\mathbf{Y}_j$'s (viz., all their first elements, all second elements, etc.) are assumed to be simultaneously observed on an experimental unit and may therefore be expected to be statistically correlated in many situations. Recognition of this may play an important role in the subsequent analysis and interpretation of the results. (See Example 44 in Chapter 6.)

A slightly more general form of the above multivariate general linear model is provided by $\mathscr{E}(\mathbf{Y}' \mid \mathbf{X}, \mathbf{G}) = \mathbf{X}\boldsymbol{\Xi}\mathbf{G}$, whee $\boldsymbol{\Xi}$ is a $k \times q$ matrix of unknown parameters and $\mathbf{G}$ is a $q \times p$ matrix, with known elements, of rank $q \leqslant p$. This generalization enables one to include polynomial growth-curve models in the class of general linear models (see, for example, Section 6 of Chapter IV in Roy et al., 1971). The application of nonlinear (in the parameters) models for studying multivariate relationships has been considered recently, but, perhaps because of the inherent difficulties of nonlinear modeling even in uniresponse problems, the use of these models in practice is not widespread.

Remarks. There is a close relationship between two-group canonical correlation analysis and a number of methods of analyzing data that are known by other names. For example, if one of the two sets of variables consists of indicator variables that designate if an observation belongs to a group (scored 1) or does not (scored 0), the two-group canonical correlation analysis is the same as the multi-group discriminant analysis procedure described in Section 4.2. Also, if both sets of variables are indicator variables, with one set associated with the row categories in a two-way contingency table and the other associated with the column categories, then the sum of squares of the canonical correlations from the analysis of such data is χ^2/n, where χ^2 is the well-known chi-squared statistic for testing the independence of the rows and columns and n is the total number of observations.

Despite such an interesting generality of canonical correlation analysis, evidence for its use as a tool for analyzing multivariate observations is fairly limited. The reasons for these are many, including the lack of aids for inference and interpretation as well as the lack of efficiency/flexibility of computations involved in deleting variables or observations.

REFERENCES

Section 3.2 Anderson (1954, 1957, 1960), Bruntz et al. (1974), Chambers et al. (1983), Chernoff (1973).

Section 3.3 Chen & Kettenring (1972), Horst (1965), Hotelling (1936), Kettenring (1969, 1971), Roy et al. (1971), Steel (1951), Thurstone & Thurstone (1941).

Methods for Statistical Data Analysis of Multivariate Observations, Second Edition
by R. Gnanadesikan

CHAPTER 4

Multidimensional Classification and Clustering

4.1. GENERAL

A wide variety of objectives, concepts, and techniques is encompassed under the heading "multidimensional classification and clustering." Loosely speaking, the concern is with respect to categorization of objects or experimental units and problems of classification and clustering lie at the core of the concerns, not only of the well-known multivariate topic of discriminant analysis, but also of more modern areas such as pattern recognition, neural networks, and so-called supervised and unsupervised learning in artificial intelligence. A dichotomy into two broad types of approaches to problems is possible. First, there are situations in which the categorization is based on prespecified groups. The term used here for this case will be *classification*; other terms used in the literature for describing it include "discriminant analysis," "classificatory analysis," "supervised learning," and "allocation." Second, there are situations in which the categorization is done in terms of groups that are themselves determined from the data. The term used here for describing the concern in such situations, wherein one is seeking meaningful data-determined groupings of objects, is *clustering*; other terms for this situation include "unsupervised learning." There are, of course, many problems that tend to fall somewhere between classification and clustering, rather than entirely into one of these two cases (see Example 17 in this chapter). Methods for systematic analysis of such in-between problems need to be developed.

Typically, problems both of classification and of clustering tend, in their primitive form, to be multidimensional in nature. Categorizations on the basis of measurements of a single feature or variable are often suspect and of limited use. The discussion in this chapter will first be concerned with classification problems and procedures, and will then consider cluster analysis.

Both classification and clustering have been the foci of considerable development of new methods. For example, in addition to the more classical nonparametric approaches to classification (see, for example, Chapter 5 of Hand, 1981), the recent computer-intensive and more data-driven developments include the work of Breiman et al. (1984) on *CART.* Also, Hastie et al. (1994) have recently proposed a method called *flexible discriminant analysis,* which relies on an analogy between regression and classification, and utilizes a nonparametric, local smoothing algorithm as its core. The discussion in this chapter is confined to the more classical approaches and algorithms.

4.2. CLASSIFICATION

Even when the concern is with classifying an object in terms of prespecified groups, distinctions will usually exist in regard to the kind and amount of background information in individual problems. For example, given a set of fingerprints of some unknown person, it is one problem to check on whether they do or do not correspond to a specific individual. It is quite another problem to attempt to determine to which one, if any, of a large population of alternative possibilities the unknown might correspond. Clearly, the strategy of the procedures, including the characteristics used, may differ between the verification and the identification problems.

From the viewpoint of data analysis, apart from correct formulation of the problem and initial choice of variables, there appear to be two other basic aspects of multidimensional classification: (i) the choice of an effective space, or representation, for discrimination, and (ii) the choice of a distance measure or metric for use in such a space.

Perhaps the simplest guise of the classification problem, although not usually considered as such, is the one-group problem wherein one wishes to decide whether or not an item belongs to a particular group. A test of significance and methods for assessing whether an observation is an outlier are simple examples of this case. A multivariate quality control procedure suggested by Hotelling (1947) is essentially a test of significance viewed as a one-group classification problem. Jackson (1956) has suggested a bivariate graphical implementation of Hotelling's procedure involving the plotting of points in an elliptical frame defined by Hotelling's T^2 statistic.

The classical form of the classification problem is the two-group case considered by Fisher (1936, 1938), leading to the so-called discriminant function. Suppose that, given two groups, G_1 and G_2, one has a reference set of observations (also referred to as *training samples*), $\mathbf{Y}_1$ and $\mathbf{Y}_2$, respectively, from them, that is, the n_1 columns of $\mathbf{Y}_1$ are p-dimensional observations on n_1 units known to come from G_1, and, similarly, the n_2 columns of $\mathbf{Y}_2$ are observations on n_2 units from G_2. Utilizing the observations in the reference set, one can obtain the sample mean vectors, $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$, as well as the sample covariance matrices, $\mathbf{S}_1$ and $\mathbf{S}_2$. Fisher's discriminant function is that linear combination of the p original responses which exhibits the largest ratio of variance between the two groups relative to that within the groups. More explicitly, if the linear combination of the original variables is denoted as $z = a_1y_1 + a_2y_2 + \cdots + a_py_p = \mathbf{a}'\mathbf{y}$, a two-sample t statistic for the variable z may be written as

$$t_{\mathbf{a}} = \frac{\mathbf{a}'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)}{\{\mathbf{a}'\mathbf{S}\mathbf{a}(1/n_1 + 1/n_2)\}^{1/2}},$$

where $(n_1 + n_2 - 2)\mathbf{S} = (n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2$. Fisher's discriminant function is obtained by choosing $\mathbf{a}$ so as to maximize $|t_{\mathbf{a}}|$ or, equivalently,

$$t_{\mathbf{a}}^2 = \left(\frac{n_1 n_2}{n_1 + n_2}\right)\left\{\frac{\mathbf{a}'(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{a}}{\mathbf{a}'\mathbf{S}\mathbf{a}}\right\}.$$

The required solution for $\mathbf{a}$ can be shown to be proportional (i.e., equal except for a multiplicative constant) to $\mathbf{S}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$. In the p-dimensional space of the responses $y_1, y_2, \ldots, y_p$, such a vector $\mathbf{a}$ defines the direction of maximal group separation in the sense that the means of the projections of the observations from the two groups are maximally apart relative to the variance of the projections around their respective means. Choosing $\mathbf{a} \propto \mathbf{S}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$ leads to the maximum value, $(n_1 n_2/n_1 + n_2) \times (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$, for $t_{\mathbf{a}}^2$, and this maximum is thus seen to be the value of the two-sample Hotelling's T^2 statistic. Also, the quadratic form, $(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$, is just the so-called Mahalanobis' D^2 statistic.

For the two-group problem, one can consider the unidimensional space of the derived variable $z = \mathbf{a}'\mathbf{y}$ (with $\mathbf{a}$ chosen as above) as an effective space for discriminating between the two groups. Since the dimensionality of the space is 1, the issue of selecting a distance measure is relatively simple in this case. Specifically, given an "unknown" object which is known only to belong to either G_1 or G_2 and for which the values of the p variables are observed to be $\mathbf{u}' = (u_1, \ldots, u_p)$, one can project the points $\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2$, and $\mathbf{u}$ onto the unidimensional space corresponding to z (viz., the vector $\mathbf{a}$) and assign the unknown to G_1 or G_2 according as the projection of $\mathbf{u}$ is closer to the projection of $\bar{\mathbf{y}}_1$ or of $\bar{\mathbf{y}}_2$. Algebraically, this amounts to calculating the value of the discriminant function for the unknown, namely, $\mathbf{a}'\mathbf{u} = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\mathbf{S}^{-1}\mathbf{u}$, and classifying the unknown in G_1 or G_2 according as $\mathbf{a}'\mathbf{u} \gtrless \mathbf{a}'(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2)/2$.

A generalization of the two-group procedure to several groups is described, for example, by Rao (1952, Section 9c). Suppose that one has g groups, $G_1, \ldots, G_g$, with the reference set of observations consisting of n_i p-dimensional observations (constituting the columns of a $p \times n_i$ matrix $\mathbf{Y}_i$) from G_i $(i = 1, \ldots, g)$. Using the observations from the ith group, one can compute the sample mean vector, $\bar{\mathbf{y}}_i$, and the sample covariance matrix, $\mathbf{S}_i$, for $i = 1, \ldots, g$. For the total set of $n = \Sigma_{i=1}^{g} n_i$ observations, one can calculate an overall mean vector, $\bar{\mathbf{y}} = \Sigma_{i=1}^{g} n_i \bar{\mathbf{y}}_i/n$, and a $p \times p$ pooled *within-groups* covariance matrix,

$$\mathbf{W} = \frac{1}{n - g} \sum_{i=1}^{g} (n_i - 1)\mathbf{S}_i. \qquad (51)$$

Furthermore, one can define a $p \times p$ *between-groups* covariance matrix,

$$\mathbf{B} = \frac{1}{g-1} \sum_{i=1}^{g} n_i(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})', \tag{52}$$

which provides a summary of the dispersion among the group means, $\bar{\mathbf{y}}_i$'s, in p-space. In some situations, when the n_i's are extremely disparate, one may wish not to weight the deviations of the group centroids from the overall centroid by the n_i's as in $\mathbf{B}$ but rather to use the $p \times p$ matrix,

$$\mathbf{B}^* = \frac{1}{g-1} \sum_{i=1}^{g} (\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})', \tag{53}$$

in place of $\mathbf{B}$ for the subsequent analysis.

Next, exactly as in the two-group problem, if $z = \mathbf{a}'\mathbf{y}$ denotes a linear combination of the original variables, a one-way analysis of variance for the derived variable z will lead to the following F-ratio of the between-groups mean square to the within-groups mean square:

$$F_{\mathbf{a}} = \frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}}. \tag{54}$$

If now one were to choose $\mathbf{a}$ so as to maximize this F-ratio, the required $\mathbf{a}$ would be the eigenvector, $\mathbf{a}_1$, corresponding to the largest eigenvalue, c_1, of $\mathbf{W}^{-1}\mathbf{B}$. The maximum value of the F-ratio would be $F_{\mathbf{a}_1} = \mathbf{a}_1'\mathbf{B}\mathbf{a}_1/\mathbf{a}_1'\mathbf{W}\mathbf{a}_1 = c_1$. Having determined $\mathbf{a}_1$, one can seek a second linear combination of the original variables which has the next largest F-ratio. The required solution for the coefficients in the second linear combination turns out to be the eigenvector, $\mathbf{a}_2$, corresponding to the second largest eigenvalue, c_2, of $\mathbf{W}^{-1}\mathbf{B}$. The process may be repeated for determining additional linear combinations. To ensure that new linear combinations are being found at each stage, some constraints (e.g., linear independence) have to be imposed on the sets of coefficients. All that is involved computationally is an eigenanalysis of $\mathbf{W}^{-1}\mathbf{B}$, leading to the ordered eigenvalues $c_1 \geqslant c_2 \geqslant \cdots \geqslant c_r > 0$ and the corresponding eigenvectors, $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r$, which will satisfy the constraints $\mathbf{a}_j'\mathbf{W}\mathbf{a}_k = \delta_{jk}$, the Kronecker delta, for $j, k = 1, \ldots, r$. The eigenanalysis may be performed by using a singular-value decomposition algorithm which is appropriate for this case involving the two matrices $\mathbf{B}$ and $\mathbf{W}$ (see Chambers, 1977, Section 5.k). The computations involved may also be viewed in terms of an initial transformation to sphericize the within-groups dispersion, followed by an eigenanalysis of the between-groups dispersion in this transformed space. More explicitly, one can first linearly transform the initial variables, $\mathbf{y}$, to p new variables, $\mathbf{x} = \mathbf{T}^{-1}\mathbf{y}$, where $\mathbf{W} = \mathbf{T}\mathbf{T}'$ is the so-called Cholesky decomposition of $\mathbf{W}$. Then the within-groups covariance matrix for the x-variables will be the identity matrix, and the between-groups covariance matrix will be $\mathbf{B}_x = \mathbf{T}^{-1}\mathbf{B}(\mathbf{T}^{-1})'$, where $\mathbf{B}$ is defined in Eq. 52. Next an eigenanalysis on the $p \times p$ symmetric mtrix, $\mathbf{B}_x$, may be performed. The eigenvalues of $\mathbf{B}_x$ are, in fact, also the eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$, and the eigenvectors, $\{\mathbf{a}_j\}$, of $\mathbf{W}^{-1}\mathbf{B}$ are related to the eigenvectors, $\{\mathbf{l}_j\}$, of $\mathbf{B}_x$ by the equations, $\mathbf{a}_j = (\mathbf{T}')^{-1}\mathbf{l}_j$, for $j = 1, 2, \ldots, r$.

In general, if there are g groups and the problem is p-dimensional, the number, r, of positive eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ will be equal to the smaller of $(g - 1)$ and p. This is a consequence of the fact that, if g is less than p, the g group means are contained in a $(g - 1)$-dimensional hyperplane. In particular, when $g = 2$, the analysis is exactly equivalent to the two-group discriminant analysis, considered earlier, leading to a single discriminant function. More generally, with $g > 2$, one can determine up to r linear combinations, $z_i = \mathbf{a}_i'\mathbf{y}$, for $i = 1, 2, \ldots, r$, and the z_i's will be called *discriminant coordinates* or *CRIMCOORDS*. [*Note*: Other authors have referred to these as "canonical variates" (e.g., Rao, 1952; Seal, 1964), but the present author's preference is to use the term "canonical variates" only in the context of canonical correlational analysis, discussed in Chapter 3.] The space defined by the *CRIMCOORDS*, or by a subset of the first t ($\leqslant r$) of them, will be called the *discriminant space*. Since the CRIMCOORDS are determined so that they account for group separation in decreasing order, there is an issue of how many of them (viz., choice of a value for t) one ought to use. The nature of the diminishing returns from using the later CRIMCOORDS has to be studied in any given problem, and often t has to be chosen by trying several alternative values.

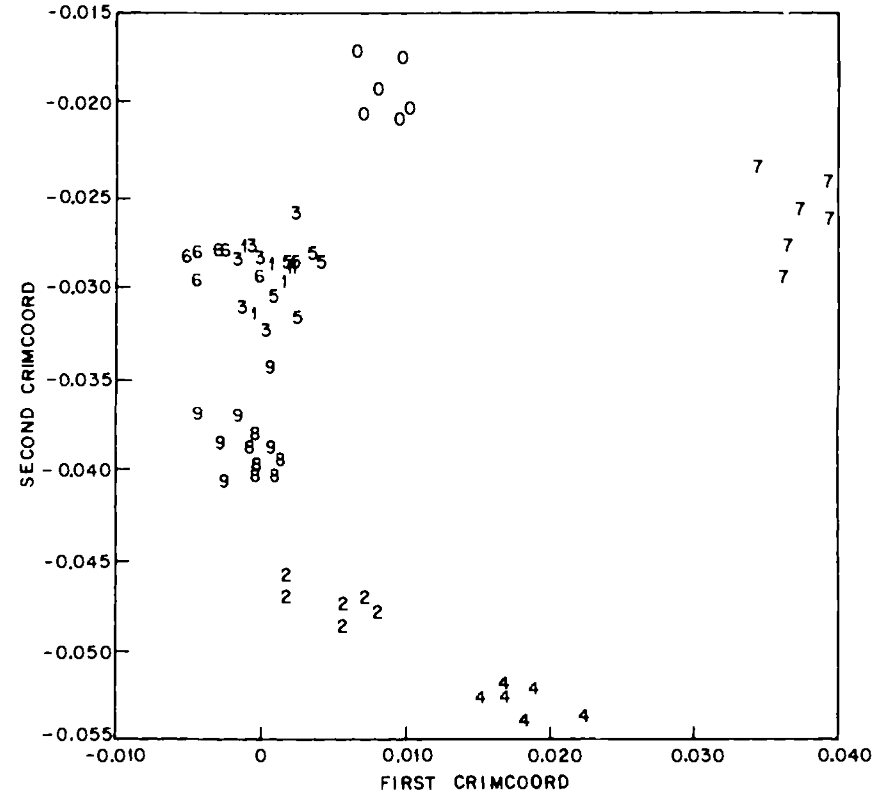
In the multigroup case, the discriminant space, which is a specifically chosen linear transformation of the original space, may be considered as an effective space for use in classifying "unknown" objects. The original reference set of observations, as well as the observations corresponding to the (unknown) objects which are to be classified into one of the g groups, may be represented in the discriminant space of dimension $t(\leqslant r)$. The representation of the original data consists in making the transformation

$$\mathbf{Z} = \mathbf{A}_t'\mathbf{Y}, \tag{55}$$

where $\mathbf{A}_t'$ is a $t \times p$ matrix whose rows are the eigenvectors $\mathbf{a}_1', \ldots, \mathbf{a}_t'$ $(t \leqslant r)$, and $\mathbf{Y} = [\mathbf{Y}_1 \mid \mathbf{Y}_2 \mid \cdots \mid \mathbf{Y}_g]$ is the $p \times n$ set of all the reference observations. The columns of $\mathbf{Z}$ may, of course, be partitioned according to the partitioning of $\mathbf{Y}$ so as to provide the original group identities for the representations in the discriminant space. If $\mathbf{u}' = (u_1, \ldots, u_p)$ denotes the p-dimensional observation on an object which is to be classified as belonging to one of the g groups, a representation of $\mathbf{u}$ in the t-dimensional discriminant space is given by $\mathbf{A}_t' \cdot \mathbf{u}$.

For data-analytic purposes, two- and three-dimensional graphical representations of the columns, respectively, of $\mathbf{Z}(2 \times n)$ and $\mathbf{Z}(3 \times n)$ may be obtained. Such plots are often useful for studying the degree and nature of group separations, for suggesting possible metrics for use in the discriminant space,

Exhibit 16. Representation of utterances in the space of first two CRIMCOORDS



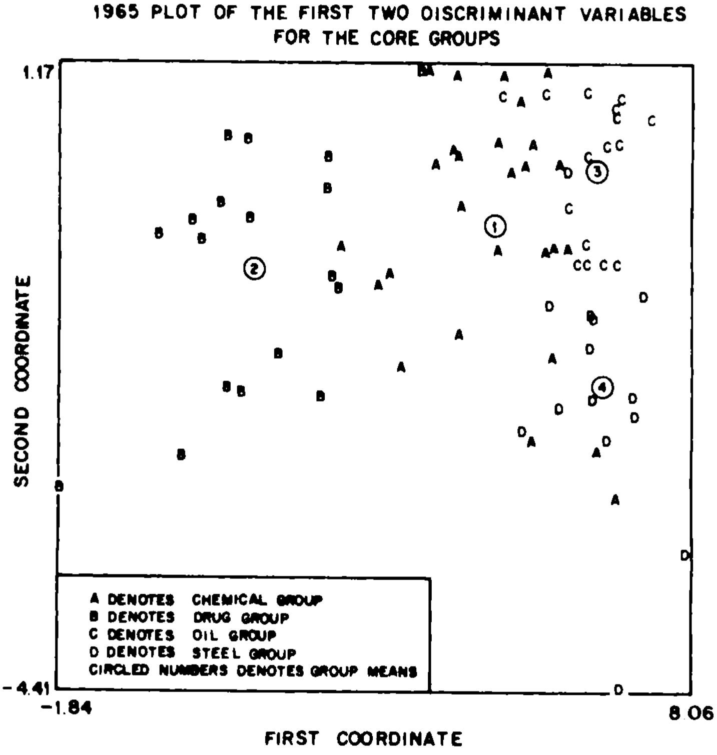
and for indicating stray or outlying observations. The approach is illustrated by the next two examples.

Example 16. This example, which pertains to the talker-identification problem (for details see Becker et al., 1965; Bricker et al., 1971), involves data from 10 talkers, each of whom repeated a given word six times. The initial representation of each utterance in this particular example was a 16-dimensional summary derived from raw data whose dimensionality was much higher. Exhibit 16 shows a representation of the 60 resultant utterances in the two-dimensional discriminant space of the first two CRIMCOORDS, that is, coordinates that are obtained from the eigenvectors corresponding to the two largest eigenvalues of a $\mathbf{W}^{-1}\mathbf{B}$ matrix calculated from the initial 16-dimensional data. The utterances are labeled by the 10 digits 0 through 9 to correspond to the talkers with whom they are known to be associated.

The clustering of the points in Exhibit 16 corresponds generally to the known categorization of the utterances and indicates the clear separations of and among talkers 0, 2, 4, and 7, as well as the considerable overlapping of talkers, 1, 3, 5, and 6 and of talkers 8 and 9. There are no indications of outlying observations. Also, despite some indications of possible differences in the dispersions of the utterances when they are represented in the space of the first two CRIMCOORDS, one may feel that it is not unreasonable to use a simple Euclidean metric in the two-dimensional discriminant space (see the discussion in Section 4.2.1 on distance measures).

Example 17. A second example of the value of graphical representations in discriminant space is taken from a study of Chen et al. (1970, 1974) concerned with developing empirical bases for grouping industrial corporations into categories such as chemicals, drugs, oils, etc. One part of the study, using observations on 14 economic and financial variables, was concerned with the validity and appropriateness of such prespecified categories. [*Note*: Since one is interested both in utilizing useful prior groups where these are appropriate and in evolving data-determined groups when these are meaningful, this problem really does not fall totally under either classification or cluster

Exhibit 17. Representation of core group companies in the space of first two CRIMCOORDS



analysis.] A four-group analysis of the chemical, drug, oil, and steel groups of companies in this investigation led to three CRIMCOORDS, and Exhibit 17 shows a representation of the companies in the discriminant space defined by the first two CRIMCOORDS. [*Note*: In this problem an initial analysis of each of the groups internally led to the identification of a few outliers (see also Example 50 in Section 6.4.1), and the determination of the CRIMCOORDS was then based only on the companies retained in the core groups.]

Apart from its usefulness in studying group separations, the configuration in Exhibit 17 indicates a relatively tight grouping of the oil companies and a very widely dispersed chemical group, thus hinting at possibly large disparities among the covariance matrices of the different groups in the space of the original 14-dimensional observations. The pooling involved in obtaining **W** might then be questionable, and other approaches (see Section 4.2.1) might prove more appropriate.

As a further aid in using plots such as Exhibits 16 and 17, one can draw circular "confidence regions," defined by

$$n_i(\bar{\mathbf{z}}_i - \boldsymbol{\mu}_i)'(\bar{\mathbf{z}}_i - \boldsymbol{\mu}_i) \leqslant \chi_2^2(\alpha) \qquad \text{for } i = 1, \ldots, g, \tag{56}$$

where $\bar{\mathbf{z}}_i = \mathbf{A}_2'\bar{\mathbf{y}}_i$ is the centroid (i.e., mean) of the representations of the n_i observations in group G_i in terms of the first two CRIMCOORDS, $\boldsymbol{\mu}_i$ is the unknown expected value of $\bar{\mathbf{z}}_i$, and $\chi_2^2(\alpha)$ denotes the upper $100\alpha\%$ point of the chi-squared distribution with 2 degrees of freedom. For three-dimensional representations in the space of the first three CRIMCOORDS, one can define spheres centered again at centroids by analogy with the two-dimensional case. The required percentage point would be from a chi-squared distribution with 3 degrees of freedom in this case. These circular and spherical regions may help in assessing the degree of group separation.

The pooling of the individual group dispersions to obtain the within-groups covariance matrix **W** merits a few comments. First, there is the question of inappropriateness of averaging across dissimilar covariance structures and the statistical effects of such averaging on the details of the classification or discriminant procedures. Suppose, for instance, that **W** includes one covariance matrix from a very widely dispersed group (see Example 17). Then the determination of the CRIMCOORDS, and hence the associated discriminant space, may be distorted considerably by the inclusion in **W** of the "large" covariance matrix. A second question raised by the presence of widely varying covariance structures among the groups is the general issue of the meaningfulness of looking for location types of differences in the presence of such dispersion disparities—the so-called Behrens-Fisher problem of statistical inference. Third, particularly important from the point of view of data analysis is the following question: if, in fact, there are large discrepancies in the dispersion characteristics of the groups, and one is interested in discriminating among the groups, should not one attempt to use the dispersion information for the discrimination? As a partial answer to this question, one way of incorporating dispersion differences in the analysis is described in Section 4.2.1.

4.2.1. Distance Measures

Given a space (either the one for the original variables or a derived discriminant space) for representing the objects, the fundamental problem in classification is reduced to choosing a metric or a distance measure. For, if such a metric is available, an object which needs to be assigned to one of the groups may be identified with the group to which it is closest as judged by the metric.

Theoretical formulations have, by and large, been confined to the derivation of specific distance measures to satisfy narrowly defined optimality criteria under a body of assumptions, which themselves are often beyond empirical check by the data on hand. For instance, the optimal Bayes discriminant function minimizes expected loss, using prior probabilities, as well as other distributional assumptions.

From the point of view of data analysis, the prescription of a distance function will generally be a trial and error task in which the use of some general techniques needs to be aided by insight, intuition and, perhaps, good luck!

One useful general class of squared distance functions is provided by a class of positive semidefinite quadratic forms. Specifically, if $\mathbf{u}' = (u_1, u_2, \ldots, u_p)$ denotes the p-dimensional observation on an object that is to be assigned to one of the g prespecified groups, then, for measuring the squared distance between $\mathbf{u}$ and the centroid of the ith group, one may consider the function

$$D^2(i) = (\mathbf{u} - \bar{\mathbf{y}}_i)'\mathbf{M}(\mathbf{u} - \bar{\mathbf{y}}_i), \tag{57}$$

where $\mathbf{M}$ is a positive semidefinite matrix to ensure that $D^2(i) \geqslant 0$. The object will be assigned to the group for which $D^2(i)$ is smallest as i takes on the values from 1 through g. Different choices of the matrix $\mathbf{M}$ lead to different metrics,

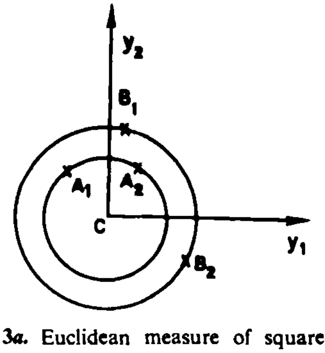


Fig. 3*a*. Euclidean measure of squared distance.

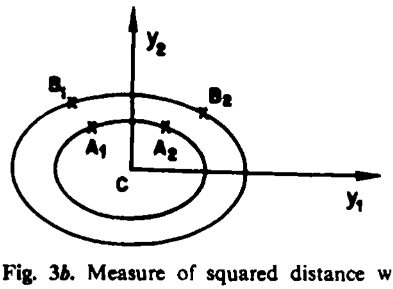


Fig. 3*b*. Measure of squared distance with different weights for the variables.

and the class of squared distance functions represented by Eq. 57 is not unduly narrow.

Thus, when $\mathbf{M} = \mathbf{I}$, one obtains the familiar Euclidean squared distance between the "unknown" and the centroid of the ith group in the p-dimensional space of the responses. Geometrically, as shown in Figure 3*a* for the case when $p = 2$, the use of such a measure of squared distance amounts to measuring distances by circles (or spheres when $p > 2$)—points A_1 and A_2 lying on the same circle are considered to be the same distance away from the center C, while points B_1 and B_2 lying on the outer circle are considered to be farther away from C than are A_1 and A_2. For statistical uses, when the different responses are noncommensurable and likely to have very different variances, the use of this unweighted Euclidean metric may be inappropriate. For instance, if $p = 2$ and y_1 has a larger variance than y_2, one may wish to weight a deviation in the y_1-direction less than an equal deviation in the y_2-direction. A way of accomplishing this would be to use "elliptical" (or ellipsoidal) distance measures as shown in Figure 3*b*—again A_1 and A_2 are considered to be equidistant from C, while B_1 and B_2 are considered to be farther from C than A_1 and A_2. Algebraically, this measure of squared distance corresponds to specifying $\mathbf{M}$ in Eq. 57 to be a diagonal matrix with diagonal elements equal to the reciprocals of the variances of the different variables. Still another extension of the distance measure may be made to accommodate intercorrelations among the responses as well as possible differences among their variances. When $p = 2$ and the statistical correlation between y_1 and y_2 is positive, Figure 3*c* shows how one may use "elliptical" distance measures by tilting the ellipses so that their major axis is oriented in a direction reflecting the positive correlation—once again, points on the same ellipse are considered equidistant from C, while points, such as A_1 and B_1, on the different ellipses are considered to be at increasing distances away from C. A way of reflecting this choice formally in Eq. 57 is to use for $\mathbf{M}$ the inverse of the covariance matrix of the variables.

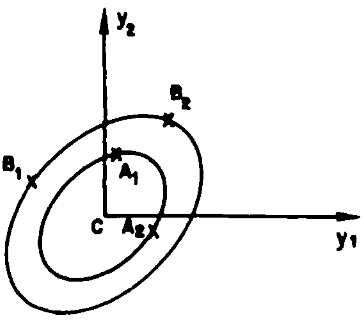


Fig. 3c. Generalized squared distance measure.

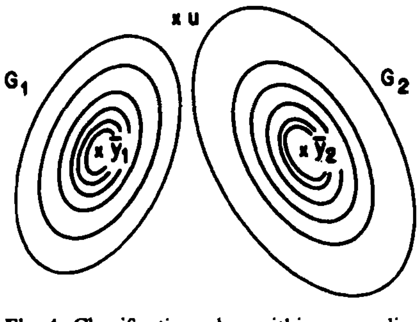


Fig. 4. Classification when within-group dispersions are different.

Three specific choices for $\mathbf{M}$ in Eq. 57 are worth considering in more detail. The first is $\mathbf{M} = \mathbf{S}_i^{-1}$, yielding

$$D_1^2(i) = (\mathbf{u} - \bar{\mathbf{y}}_i)'\mathbf{S}_i^{-1}(\mathbf{u} - \bar{\mathbf{y}}_i), \tag{58}$$

where $\mathbf{S}_i$ is the covariance matrix derived from the reference set of observations, $\mathbf{Y}_i$, in the ith group, $i = 1, \ldots, g$. An important practical constraint which needs to be met to ensure the nonsingularity of $\mathbf{S}_i$ is that $n_i > p$, so that to be able to use the metric $D_1(i)$ for classifying the "unknown" object one would, in general, require the number of reference observations in *every* group to exceed the dimensionality p. Also, since $\mathbf{M}$ changes from group to group, the use of $D_1(i)$ implies a considerable increase in the computational effort involved in classifying several "unknowns." Despite these limitations, however, one appealing feature of $D_1(i)$ is that it uses a dispersion standard that is internal to the group being considered as a possibility for assignment of an "unknown," and hence it may be able to exploit differences in the dispersion characteristics of the different groups. If $p = 2$ and one has two groups, G_1 and G_2, Figure 4 illustrates the geometry involved in using the metric D_1 in the presence of dispersion differences. In this example, although the "unknown" (shown as an ×) is closer, in Euclidean distance, to the centroid of G_1 than to that of G_2, in terms of D_1 it is likely to be assigned to G_2 rather than G_1. An important feature in using D_1 is that, if one looks for boundaries dividing the p-dimensional space of the responses into regions, one for each of the g groups, such boundaries are nonlinear. The use of a likelihood-ratio approach (see Anderson, 1984; Rao, 1952) to classification in the presence of heterogeneity of covariance matrices of the groups would lead to a similar but not identically the same procedure. For instance, with two groups, the likelihood-ratio approach based on assuming multivariate normality for the distributions of the observations would lead to classifying $\mathbf{u}$ in G_1 or G_2 according as $D_1^2(1) - D_1^2(2) \lesseqgtr \ln[|\mathbf{S}_2|/|\mathbf{S}_1|]$. On the other hand, the procedure described above would assign $\mathbf{u}$ to G_1 or G_2 according as $D_1^2(1) - D_1^2(2) \lesseqgtr 0$. More generally, with g groups, the likelihood-ratio approach would assign $\mathbf{u}$ to the ath group if

$$D_1^2(a) + \ln|\mathbf{S}_a| = \min_{i=1,\ldots,g} \{D_1^2(i) + \ln|\mathbf{S}_i|\},$$

whereas the procedure based on the metric D_1 would do so merely if

$$D_1^2(a) = \min_{i=1,\ldots,g} D_1^2(i).$$

A second choice for $\mathbf{M}$ in Eq. 57 is associated with the derivation of the discriminant space. If $\mathbf{M} = \mathbf{A}_t\mathbf{A}_t'$, where $\mathbf{A}_t'$ is defined following Eq. 55, then

$$D_2^2(i) = (\mathbf{u} - \bar{\mathbf{y}}_i)'\mathbf{A}_t\mathbf{A}_t'(\mathbf{u} - \bar{\mathbf{y}}_i) \tag{59}$$

is the measure of the squared distance of the "unknown" from the ith group. Using the metric D_2 in the p-dimensional space of original responses can be seen to be exactly equivalent to using the simple unweighted Euclidean metric in the t-dimensional discriminant space. The constraint on the eignvectors of $\mathbf{W}^{-1}\mathbf{B}$ used in obtaining the CRIMCOORDS is that $\mathbf{A}_t'\mathbf{W}\mathbf{A}_t = \mathbf{I}$. Hence, under the assumptions used in deriving the discriminant space (including homogeneity of the group covariance structures), the CRIMCOORDS would be mutually uncorrelated and have unit variance each. This is a reason for using the simple Euclidean metric in the discriminant space, though not in the original space. The choice of $\mathbf{M}$ that leads to the metric D_2 does not vary from group to group. However, it does depend on the number, t, of eigenvectors to be employed from the eigenanalysis of $\mathbf{W}^{-1}\mathbf{B}$.

A third choice of $\mathbf{M}$ leads to the so-called generalized distance of the "unknown" from the centroid of the ith group in the p-dimensional space of the original variables. Specifically, choosing $\mathbf{M} = \mathbf{W}^{-1}$ leads to

$$D_3^2(i) = (\mathbf{u} - \bar{\mathbf{y}}_i)'\mathbf{W}^{-1}(\mathbf{u} - \bar{\mathbf{y}}_i). \tag{60}$$

This choice of $\mathbf{M}$ also does not change from group to group. To ensure nonsingularity of the within-groups covariance matrix, $\mathbf{W}$, the constraint $p \leqslant (n - g)$ must be met, where $n = \Sigma_{i=1}^{g} n_i$ is the total number of observations from all g groups in the reference set. This constraint on the relationship between the dimensionality of response and the number of observations is less restrictive than the one underlying the choice of $\mathbf{M}$ that led to D_1. If pooling the dispersions of the different groups is reasonable and justified, one can thus have significant gains in the dimensionality to be used for the initial representation. A method of assessing the homogeneity of the dispersions of the groups is described in Section 6.3.2.

In the sense that both D_1^2 and D_3^2 use inverses of covariance matrices of the responses, one can think of D_3^2 as a generalization of D_1^2 when all the groups have similar dispersion characteristics. However, in the sense that D_1^2 is applicable when the groups have dissimilar dispersion characteristics, it is a generalization of D_3^2. In a somewhat less obvious sense, D_3^2 is interpretable in terms of a discriminant analysis approach which leads to D_2^2. In fact, performing a two-group discriminant analysis (with $\mathbf{W}$ in place of $\mathbf{S}$ in the earlier description of Fisher's two-group procedure) for every possible pair of groups [i.e., $g(g - 1)/2$ analyses in all] is equivalent to using D_3^2. Also, using the maximum number, r, of eigenvectors corresponding to the nonzero eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ in D_2^2 would be entirely equivalent to using D_3^2.

These equivalences and relationships between D_2^2 and D_3^2 are easier to see in terms of the sphericized coordinates $\mathbf{x} = \mathbf{T}^{-1}\mathbf{y}$, where $\mathbf{W} = \mathbf{T}\mathbf{T}'$ (see the discussion on pp. 84–85). If $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{x}}_j$ denote the centroids of the ith and jth groups, respectively, in this space, Figure 5 provides a geometrical demonstration of the equivalence between D_3^2 and the $g(g - 1)/2$ pairs of two-group discriminant analyses. For the two-group analysis involving the ith and jth

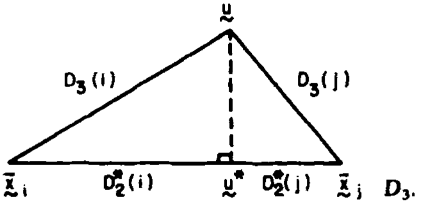


Fig. 5. Relationship between the uses of D_2 and D_3.

groups, the unknown $\mathbf{u}$ is projected onto the line joining $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{x}}_j$ to obtain $\mathbf{u}^*$ and is assigned to the ith or jth group according as $[D_2^*(i)] \lessgtr [D_2^*(j)]$. From Figure 5, however, it is clear that for the metric D_3 the relationship $D_3(i) \lessgtr D_3(j)$ holds according as $D_2^*(i) \lessgtr D_2^*(j)$, so that $g(g-1)/2$ comparisons in terms of D_2^* are equivalent to a comparison of g values of D_3.

Also, in the x-space, $D_3^2(i)$ will be just the Euclidean squared distance of $\mathbf{u}_0$ $(=\mathbf{T}^{-1}\mathbf{u})$ from $\bar{\mathbf{x}}_i$ $(=\mathbf{T}^{-1}\bar{\mathbf{y}}_i)$. Hence, if $r = p$, D_3^2 is not only equivalent to D_2^2 but also identical with it, since $D_2^2(i) = (\mathbf{u}_0 - \bar{\mathbf{x}}_i)'\mathbf{LL}'(\mathbf{u}_0 - \bar{\mathbf{x}}_i)$, where $\mathbf{L}$ is now a $p \times p$ orthogonal matrix (i.e., $\mathbf{LL}' = \mathbf{I}$) whose columns are the eigenvectors of $\mathbf{B}_x$, the between-groups covariance matrix in the x-space. If, however, $r = (g-1) < p$, then, to establish the equivalence between D_2^2 and D_3^2, it has to be shown that $D_3^2(i) \leqslant D_3^2(j)$ if and only if $D_2^2(i) \leqslant D_2^2(j)$, where D_2^2 is based on all r eigenvectors corresponding to the nonzero eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$. This follows from Pythagoras' theorem in p-space since

$$D_3^2(i) = D_2^2(i) + \left\{\begin{array}{l}\text{squared length of the perpendicular} \\ \text{from } \mathbf{u}_0 \text{ to the } (g-1)\text{-dimensional} \\ \text{hyperplane containing } \bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_g\end{array}\right\},$$

and the second term on the right-hand side of this equation is seen not to depend on i.

Using D_3^2 has the merit of conceptual simplicity, avoidance of the eigenvector computations involved in D_2^2, and a performance in accurately classifying "unknowns" that may be as good as the result obtained by the use of any subset of the eigenvectors in D_2^2. On the other hand, the computation of the eigenvectors for use in D_2^2 may lead to reduction in dimensionality of the problem and perhaps some insight. Also, sometimes when the last few CRIMCOORDS are merely reflecting "noise," using a subset consisting of the first few eigenvectors for calculating D_2^2 may improve its performance over that of D_3^2.

One can also consider the use of squared distance measures that are approximations, in varying degrees of appropriateness, to D_1^2, D_2^2, and D_3^2 respectively. For instance, when the number of observations in the reference set is not sufficiently large for obtaining nonsingular estimates of the covariance matrices involved, one may decide merely to incorporate in the distance measures the differences in the variances of the coordinates and to neglect intercorrelations. Thus, in such a case, one may obtain an "approximation" to D_2^2, for example, by using for **M** in Eq. 57 a diagonal matrix whose diagonal elements are the ratios of between-groups to within-groups sums of squares for each of the p variables. (See Becker et al., 1965, for further discussion of these "approximations" in the context of a specific application.)

Example 18. The relative performances of the three metrics D_1, D_2, and D_3 may be illustrated in the context of the corporation-grouping study (see Chen et al., 1970, 1974) used also in Example 17. Exhibit 18 shows, for a particular year, the proportion of companies from each of the four core groups that are classified into their "proper" (i.e., according to the prespecified identification of a company as chemical, drug, oil, or steel) groups when D_1^2, D_2^2, and D_3^2 are used for the assignment. There is an element of bias in the classifiction procedure in this example since each of the companies being classified has influenced the estimates of the group centroids and covariance matrices, as well as the matrices **B** and **W** used in the eigenanalysis for deriving the discriminant space. Thus there is no clear separation of "unknowns" from the reference set of observations in this example. Nevertheless, since the core groups were determined after an initial elimination of extreme outliers, the proportions in Exhibit 18 may be viewed as indicators of "percent correctly identified" by the three measures of distance.

The metric D_1 has a better overall performance and is seen to be particularly good in handling the chemical group, which happens also to be the most dispersed. Using the first two CRIMCOORDS for the metric D_2 is, of course, equivalent to assigning companies to groups on the basis of Euclidean distance in Exhibit 17. The use of an additional CRIMCOORD, which would amount to employing D_3, is seen not to make any difference for three of the four groups, although for the chemical group the use of D_3 results in a noticeable improvement over the performance of D_2. In fact, it turns out that the third CRIMCOORD mainly pulls the chemical and oil groups apart so that this improvement is explainable.

Exhibit 18. Proportion of core-group companies classified into their initial groups for 1965

	Initial Group				
Metric	Chemical	Drug	Oil	Steel	Overall Proportion
D_1	26/27	18/18	16/16	8/14	68/75
D_2 with $t = 2$	16/27	17/18	15/16	12/14	60/75
D_3	21/27	17/18	15/16	12/14	65/75

4.2.2. Classification Strategies for Large Numbers of Groups

When the number of groups, g, is large, classifying an "unknown" by comparing its distances from all of the group centroids can become prohibitively expensive even with present-day high-speed computers. Some means of initially limiting the number of contenders to which an "unknown" may be assigned have to be developed. The rest of this subsection describes an ad hoc procedure based on using the first few CRIMCOORDS for this purpose. The essential ideas are developed in the context of the talker-identification problem, but their general applicability whenever g is large will also, it is hoped, emerge from their description.

Since the first few CRIMCOORDS provide a linear transformation of the original variables so as to maximally separate the groups, one natural approach would be to use a representation of the observations, together with an "unknown," in the space of the first few CRIMCOORDS as the basis for delineating "most likely" contenders for the "unknown." As seen in Exhibit 16 pertaining to the talker-identification example, with only 10 talkers one can see both separations and clusterings among the talkers even in the two-dimensional representation with respect to the first two CRIMCOORDS. When the number of talkers increases however, such indications may not be as clear. Thus in Figure 6*a*, which shows a representation of only the centroids of the utterances of a given word by 172 talkers in the space of the first two CRIMCOORDS, there are no obvious clusters.

One approach here is to divide the two-dimensional discriminant space arbitrarily into boxes as a first step. The boundaries of the boxes may be determined by using specified quantiles of the distributions of the group centroids along the two CRIMCOORDS, and it would be appropriate to employ a larger number of quantiles for the distribution along the first CRIMCOORD than for the one along the second. For the talker-identification example, Figure 6*b* shows a division of the space in Figure 6*a* into 40 boxes, using nine deciles (i.e., values that divide the distribution into 10 equal parts) of the distribution of the 172 centroids along the first CRIMCOORD and three quartiles (i.e., values that divide the distribution into four quarters) of the distribution with respect to the second CRIMCOORD. Using such an arbitrarily partitioned two-dimensional discriminant space, one can determine the box into which an unknown under consideration for assignment falls (see Figure 6*c*), and then can initially limit the comparison of the "unknown" to only the groups whose centroids fall in the same box or a few nearby ones. Figure 6*d* shows a case in which the initial comparison is limited to nine boxes, with the one containing the "unknown" in the center. In the particular example used for Figures 6*a*–*d*, while the 0 denotes the "unknown," the × corresponds to the centroid of the talker from whom the "unknown" arose. Although the × is not in the same box as the 0 in this example, it is seen to be in a neighboring box, which is included for comparison. This may not always

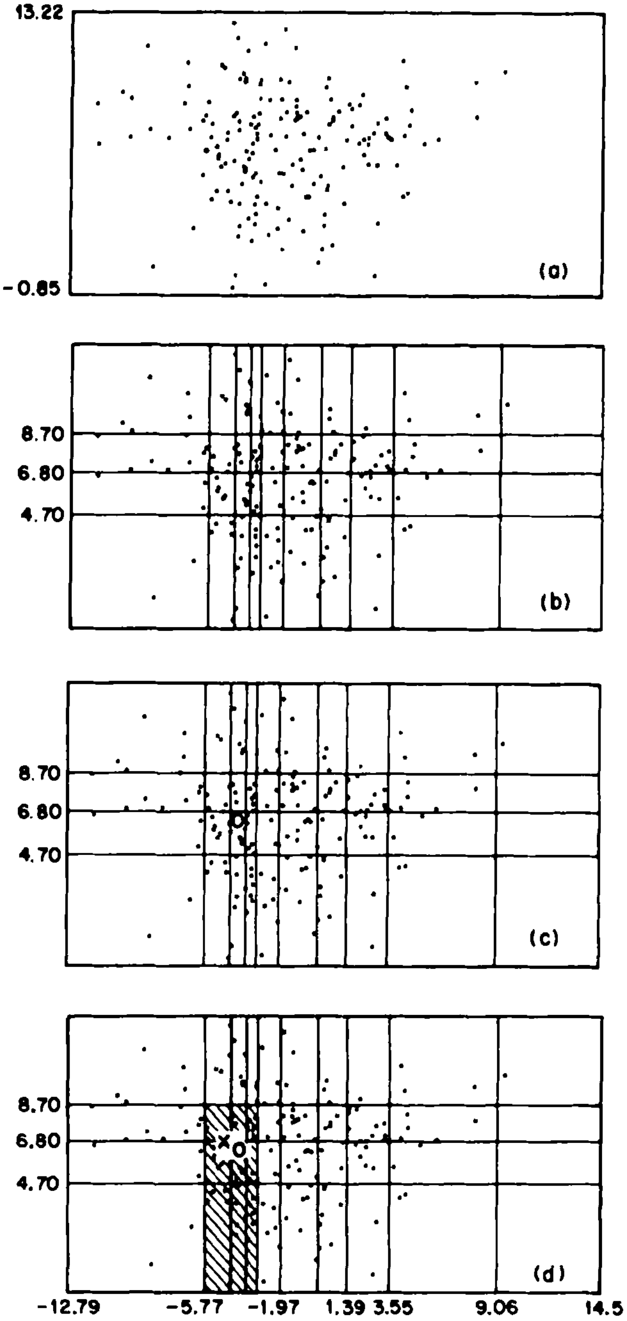


Fig. 6. Illustration of method for subselecting groups for classifying an unknown.

happen, however, and sometimes additional boxes may have to be included for picking up the "true" contender. A statistical strategy for expanding the base of comparisons by considering additional boxes is described below.

The actual comparison of the "unknown" with the groups whose centroids are in nearby boxes is made by calculating distances *not* just in the space of the first two CRIMCOORDS but in terms of all the t CRIMCOORDS that one has decided to include. In other words, the metric D_2 defined by Eq. 59 is used with the chosen value of t, but the centroids, $\bar{y}_i$, are initially limited to those of groups that are nearby in the space of the first two CRIMCOORDS.

The decision to include additional boxes may be based on two considerations: (i) the number of groups considered for the assignment of an "unknown" is inadequately small, a situation that may, for instance, occur when the "unknown" falls in a box toward the outer edges in Figure 6*b*; and (ii) the evidence for assigning the "unknown" to a group included in the initial set of boxes is not sufficiently strong.

To evaluate the strength of the evidence for associating an "unknown" with a group, two statistics that depend on the observed values of D_2^2 may be used. For a given set of contending groups, the ratio of the second smallest value of D_2^2 to the smallest value, as well as the latter value by itself, is a useful indicator. Thus, while the smallest value of D_2^2 determines the group to which the "unknown" is assigned, its numerical magnitude is an indicator of actual closeness between the "unknown" and the group. The ratio of the second smallest value to the smallest value is a measure of the closeness of the "unknown" to the group it is assigned to, as compared to its closeness to the next nearest group. Thus a large value of the ratio and/or a small value of the minimum observed D_2^2 lend strength to an assignment. Statistical benchmarks are needed for comparing the observed values of statistics such as the ratio and the smallest distance. If one is dealing with a situation in which there are sufficient data under "null" conditions (i.e., correct classification), one can obtain adequate estimates of the "null" statistical distributions of the statistics; that is, using only the reference set of observations, one can "simulate" the classifiction procedures, obtain the values of the statistics when the procedures lead to a correct assignment, and study the empirical distribution of these values. Such empirical distributions and their percentage points may then be used for comparing observed values of the statistics in assigning an "unknown" to decide whether they are large (or small) enough to confirm a "safe" assignment.

The essential features in the above type of stategy are, first, initial limitation of contenders by including for the primary comparisons only groups that are near the "unknown" in the space of the first two CRIMCOORDS; and second, enlargement of the popılation of contenders only when the assignment based on the primary comparisons is suspected of not being statistically sufficiently unequivocal. The hope is that for most of the "unknowns" one will not need to include groups from very many boxes to arrive at a satisfactorily clear assignment and that for only a few of the "unknowns" will one need to consider a large number of groups (possibly even all of them). Obviously, the properties of the strategy depend on various facets, including the number of CRIMCOORDS used initially (two is simplest and is generally recommended), the number and size of the boxes, the cut-off values for comparing statistics, such as the smallest distance and the ratio of the second smallest to the smallest distance, etc. In any example where g is very large, the specific values for these quantities may have to be chosen on a trial and error basis.

In the talker-identification example, which was used to motivate the strategy for large g, for the case of 172 talkers with one utterance from each serving as an "unknown," the use of the above type of strategy led to 81% (140/172) correct identification. An exhaustive comparison of each "unknown" against every talker, at a computing cost almost twice that for this strategy, led only to an improvement of 3%, namely, 84% (144/172) correct identification. A more detailed discussion of the talker-identification problem, including additional means that were employed to increase the percentage of correct identifications, is provided by Bricker et al. (1971).

4.2.3. Classification in the Presence of Possible Systematic Changes among Replications

The classification process described in the preceding subsections of this chapter may be summarily described as follows: given g group centroids and an "unknown," $\mathbf{u}$, assign $\mathbf{u}$ to the group to whose centroid it is closest in terms of some metric. However, in some situations there may be an arbitrary or systematic change, for artifactual or other reasons, from observation to observation even within a specified group. For example, in repeated utterances of a word by a given talker, the general level of the jointly observed energies may shift because of varying proximity to the microphone. A second example would be a situation in which the groups are different species and the observations within a group are made on members at different stages of growth. In such circumstances a modified view of the classification problem is in order.

Thus suppose that with repeated utterances of a specific word by the same person one observation leads to the vector $\mathbf{x}$ and the next to $\mathbf{x} + \mathbf{c}$, where all the components of the vector $\mathbf{c}$ are equal but unknown. Of course, this is perhaps an oversimplified model for the true effect of proximity to the microphone. However, the essential point is that, when such possibilities exist, it is no longer wise or proper to classify the unknown with the group to whose center it is closest. Now each group is represented, in concept, not by a point, but by the line joining the group center, $\bar{\mathbf{y}}$, and the point, $\bar{\mathbf{y}} + \mathbf{c}$, for any $\mathbf{c}$. The proper classification is then based on the shortest generalized distance to such lines.

Similarly, one may need to allow for possible joint scale change, or even higher-order change, affecting all the coordinates of the multiresponse vector identically. For instance, if both scale and origin are artifactual, so that one observation in a group is $\mathbf{x}$ and another is $b\mathbf{x} + \mathbf{c}$, each group is defined as a plane and classification is based on shortest generalized distances to these group planes.

A simple version of these problems and one approach to them have been considered by Burnaby (1966) and by Rao (1966). A different approach is suggested here by casting the problem in a more familiar and suggestive form.

Instead of considering the ith group centroid, $\bar{\mathbf{y}}_i = (\bar{y}_{i1}, \ldots, \bar{y}_{ip})'$, and the unknown, $\mathbf{u} = (u_1, \ldots, u_p)'$ as two points in p-space, consider them as p points in two-dimensional space with coordinates $(\bar{y}_{ij}, u_j)$ for $j = 1, 2, \ldots, p$. One can then make a scatter plot of these p points.

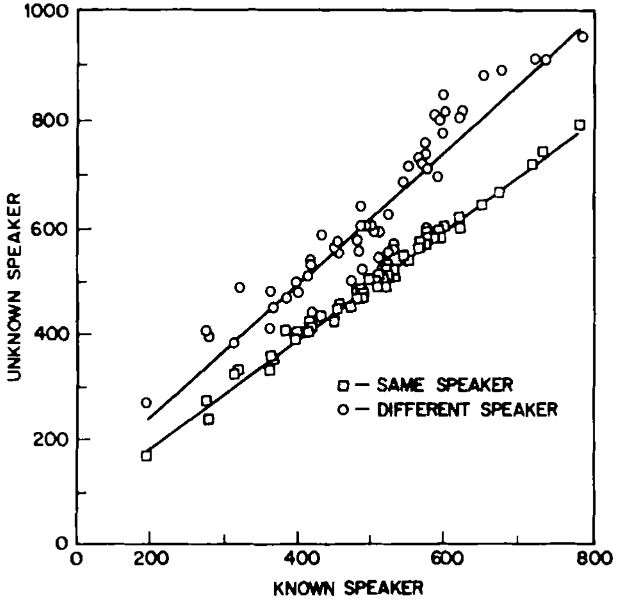
Clearly, in the absence of any systematic changes between replications within a group, perfect correspondence between the unknown and the ith group will lead to a linear configuration having unit slope and passing through the origin. If the unknown is a member of the group, one expects a good linear configuration, and, indeed, the generalized distance in p-dimensional space between the unknown, $\mathbf{u}$, and the ith group centroid [i.e., $D_3^2(i)$ of Eq. 60] is just an appropriately defined quadratic form in the residuals of the observations from the line of unit slope through the origin in this two-dimensional representation. A joint additive shift and common scale change, if present, will show as a nonzero intercept and a slope not equal to unity.

The above type of scatter plot can be made for each unknown against the centroid of every group, and for classification purposes a linear regression line may be determined corresponding to each plot and the unknown may be assigned to a group by comparing the magnitudes of the g residual sums of squares in the g regressions. In general, since the p points are associated with p responses that may have widely differing variances in addition to being intercorrelated, the fitting may have to be performed by generalized (i.e., weighted) least squares rather than by simple least squares. The classification will then be based on a comparison of g values of a quadratic form in the residuals of the observations from the generalized linear least squares fits in each of the scatter plots. For an initial exploratory analysis in many problems, the simpler approach through ordinary least squares may be adequate.

Example 19. The approach is illustrated by application to data from the talker-identification problem. One summary employed in this problem consisted of a 57-dimensional vector of energies for characterizing each utterance of a given word by each talker.

Exhibit 19 shows a scatter plot of the values of the 57 components for two "unknown" utterances of a word against the corresponding values in the average of four "known" utterances of the same word by a specific talker. One of the unknowns was chosen from the same talker, and the points for this are shown as □'s; the other was from another talker, and the corresponding points are shown in Exhibit 19 as ○'s. Also shown in Exhibit 19 are the simple least squares linear fits to the two sets of points.

Exhibit 19. Linear regression of unknown versus centroid: residual sum of squares for □ is 8258 and for ○ is 180,747



The existence in these data of artifacts of the type discussed above is evident in this plot. The configuration of the □'s, although quite linear, has a nonzero (small positive) intercept. Also the slope of the fitted line is very slightly smaller than unity. Thus there is some evidence of a shift (and, perhaps, no scale) artifact.

The configuration of the ○'s exhibits poor linearity, with considerably more scatter about the linear fit. A comparison of the two configurations suggests the possible utility of a classification procedure based on a quadratic form in the residuals from a least squares linear fit. In the present example the ordinary sums of squares of the residuals, for instance, turn out to be about 8260 for the configuration of the □'s and over 180,000 for that of the ○'s. Also, in this example the use of simple least squares fits and a comparison of the associated residual sums of squares led to almost 70% correct identifications, and the utilization of weighted least squares employing estimates of variances (and neglecting the correlational aspects) improved the percentage to about 75%. Of course, in other examples, wherein the variances may be more disparate and the intercorrelations perhaps high, the performance of the approach based on simple least squares may not be as good.

The approach just illustrated has a number of attractions. First, it involves familiar regression ideas. Second, it permits a graphical representation. Third, largely as a consequence of the second point, the procedure enables the use of a flexible internal comparisons process, in that the data themselves may help to suggest the nature of the possible corrections which may be desirable, such as the detection of coordinate outliers or the form of the regression (e.g., quadratic or other nonlinear regressions) which may be more appropriate to use.

In this approach there can be additional methodological problems when *both* the covariance matrix (needed for the generalized least squares fitting) and the regression function have to be estimated from the data. In that case some iterative process is possible, if necessary. There are also problems of strategy and implementation of any iterative technique.

4.3. CLUSTERING

The area of cluster analysis, which had its origins outside the mainstream of statistics largely in fields such as numerical taxonomy and psychology, has in recent decades received considerable attention in the statistical literature. Entire books, such as those by Everitt (1974), Hartigan (1975) and Kaufman & Rousseeuw (1990), in addition to surveys [e.g., Cormack (1971); Gnanadesikan & Kettenring (1989)], bear testimony to the extensiveness of the field. The essential concern of cluster analysis is to find groupings of things (e.g., objects, experimental units, variables) such that the things within groups are more "similar" (in some sense to be indicated by the measurements on the things) than the things across groups. Despite the intuitive appeal of such a goal, however, performing a cluster analysis sensibly and obtaining meaningful results is far from simple. Questions of what to measure, how to quantify similarity, what methods to use for performing the clustering, and most importantly, how to assess the results of using clustering algorithms are all critical. For convenience of exposition, one can distinguish three stages of cluster analysis: (1) the input stage, (2) the algorithm stage, and (3) the output stage.

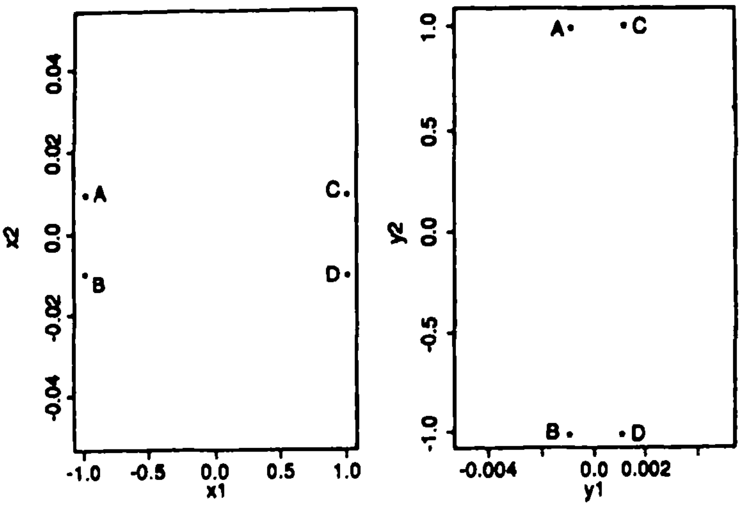
Of these three stages, the second one concerned with different types of algorithms for clustering is the one that has received the lion's share of attention in the literature. Partly under the stimulus of modern computing technology, there has been an explosion in the variety and the number of algorithms and very little is known about the relative statistical behaviors of the myriad of the currently available methods. The input stage, where one needs for example to consider what an appropriate measure of similarity to use would be in the light of the data at hand, has received a reasonable amount of attention. Unfortunately, the output stage concerned with assessing and interpreting the results of cluster analyses, is the one that has received scant attention thus far.

The three subsections that follow discuss issues and methods pertaining to the three stages, respectively.

4.3.1. Inputs

Issues that need consideration prior to carrying out any cluster analysis include the following: appropriate scaling or weighting of the variables, or transformations of them; measures of proximity or metrics to use as indicators of closeness among the items to be clustered. Choices made at this stage can have a determining influence on the outputs of the subsequent analysis. A simple example illustrates what can happen. Figure 7*a* shows four items, A, B, C, and D, in a scatter plot. In this picture, one would consider it natural to group A and B into one cluster and C and D into a second cluster. However, if one were to scale the two variables differently (e.g., measure things in different units), for example by dividing the abscissa variable by 1000 and multiplying the ordinate variable by 100, the configuration changes to the one in Figure 7*b*. It would now be more natural to group A and C together in one cluster and B and D in a second one. The outcome is a result entirely of scaling the variables or choice of the units of measurement, and one may not find this desirable in many applications wherein the clusters sought should be scale invariant.

The nature of the data, as well as the type of clustering algorithm one wishes to use in a specific situation, will influence the choice of the inputs. For instance, with metric data represented as *n* points in *p*-space, if the interest is in grouping the *n* *p*-dimensional observations using a non-hierarchical clustering method (see Section 4.3.2 for a description of such methods), one could use the $p \times n$ matrix, Y, of raw data as the input. If one wants the results not to depend on the scales of the variables, on the other hand, one would use the $p \times n$ matrix, $\mathbf{Z} = \mathbf{D} \cdot \mathbf{Y}$, where **D** is a diagonal matrix of reciprocals of



(a) (b)

Fig. 7. Effect of scaling of variables on clustering.

estimates of scale of the p variables. Widely used choices of the scale estimate are the range or the standard deviation of each variable, where neither of these choices takes any account of the possible cluster structure in the data. More generally, with metric data to be clustered by a non-hierarchical algorithm, if one wants the results to be "invariant" under affine transformations of the initial variables, then the input has to be the transformed data, $\mathbf{Z} = \mathbf{A} \cdot \mathbf{Y}$, where $\mathbf{A}$ is either the inverse of the triangular matrix from the Cholesky decomposition of an estimate of the covariance matrix, or the inverse of the symmetric square-root of such an estimate. Using the covariance matrix, $\mathbf{S} = (1/n - 1)(\mathbf{Y} - \bar{\mathbf{Y}})(\mathbf{Y} - \bar{\mathbf{Y}})'$, as the estimate, while simple and obvious, also ends up ignoring possible clusters in the data. Ideally, one would like to use an estimate of the "within-cluster" covariance matrix analogous to the within-group covariance matrix, $\mathbf{W}$, in discriminant analysis defined by Eq. 51. The difficulty in the cluster analysis situation is that the clusters are not known a priori and have to be determined. Various schemes have been proposed for handling this difficulty (see Art et al., 1982; Gnanadesikan et al., 1993, 1995, and references therein). The basic idea of the method proposed by Art et al. (1982) is that, although one does not know the clusters ahead of time it is likely that nearest neighbors among the observations belong to the same cluster. This idea is then used to develop an estimate of the within-cluster covariance matrix (except for a multiplicative constant) based on nearest neighbors.

More explicitly, the motivation of the scheme proposed by Art et al. (1982) is the following decomposition of the total sum-of-cross-products matrix in terms of pairwise differences among the observations:

$$\begin{aligned}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{y}_{ij}-\bar{\mathbf{y}})(\mathbf{y}_{ij}-\bar{\mathbf{y}})' &= (1/n)\sum_{i<i'=1}^{n}(\mathbf{y}_i-\mathbf{y}_{i'})(\mathbf{y}_i-\mathbf{y}_{i'})' \\ &= (1/n)\sum_{\substack{i<i'\\ \text{within}}}(\mathbf{y}_i-\mathbf{y}_{i'})(\mathbf{y}_i-\mathbf{y}_{i'})' \\ &\quad + (1/n)\sum_{\substack{i<i'\\ \text{between}}}(\mathbf{y}_i-\mathbf{y}_{i'})(\mathbf{y}_i-\mathbf{y}_{i'})',\end{aligned}$$

or $\mathbf{T} = \mathbf{W}^* + \mathbf{B}^*$. [*Note*: In the notation used above, $\mathbf{y}_{ij}$, denotes the jth observation in the ith cluster. However, in the expressions on the right-hand side, $\mathbf{y}_i$ and $\mathbf{y}_{i'}$, are the ith and i'th columns of the $p \times n$ data matrix, $\mathbf{Y}$, without any association with any particular clusters, and is thus more in concordance with the cluster analysis case where one has no prior knowledge of the clusters or their compositions.] In the above equation, $\mathbf{W}^*$ is based solely on within-cluster pairs while $\mathbf{B}^*$ is based on between-cluster pairs of observations. In the cluster analysis situation, the difficulty is that one lacks advance information on g, n_i and the cluster labels. Using the intuitive reasoning that, despite this difficulty, if there are any clusters present in the data then the nearest neighbors are likely to belong to the same cluster, an estimator similar in spirit to $\mathbf{W}^*$ is proposed. The steps in developing the estimator are:

(i) set $\mathbf{W}^{*(o)}_{(m)} = \mathbf{I}$, the identity matrix, and set the index of iteration $t = 1$;

(ii) find the m closest pairs of observations according to the squared generalized distance

$$(\mathbf{y}_i - \mathbf{y}_{i'})' \mathbf{W}^{*(t-1)^{-1}}_{(m)} (\mathbf{y}_i - \mathbf{y}_{i'});$$

(iii) define

$$\mathbf{W}^{*(t)}_{(m)} = \frac{1}{n} \sum_{A} (\mathbf{y}_i - \mathbf{y}_{i'})(\mathbf{y}_i - \mathbf{y}_{i'})',$$

where A is the set of pairs (i, i'), $i < i'$, corresponding to the closest pairs found in step (ii);

(iv) compute

$$E^{(t)} = \operatorname{tr}(\mathbf{W}^{*(t-1)^{-1}}_{(m)} \mathbf{W}^{*(t)}_{(m)} - \mathbf{I})^2;$$

if $E^{(t)} \leqslant E$, a user-specified number, or if $t = t_{\max}$, the maximum number of iterations allowed, stop and let $\mathbf{W}^*_{(m)} = \mathbf{W}^{*(t)}_{(m)}$; otherwise replace t by $t + 1$ and return to step (ii). ($E = 0.001$ and $t_{\max} = 20$ were used by Gnanadesikan et al., 1993.)

The above $\mathbf{W}^*$-algorithm is fully defined except for a value of m to be used in it. Care needs to be taken to choose an "appropriate" value of m so that only within-cluster pairs are used to form $\mathbf{W}^*_{(m)}$. If m is too small the estimate may be highly variable because it is based on too few within-cluster pairs, but if m is too large then bias would creep in due to the inclusion of between-cluster pairs of observations. This is another example of the common phenomenon in statistical practice of having to trade off bias and efficiency. Gnanadesikan et al. (1993) describe a graphical aid for choosing m. Also, they suggest a conservative (i.e., more concerned with avoiding bias) "2/3 rule" of using a value of m about $(n/3)(n/g - 1)$ with a guessed value for the number of clusters, g.

The method as described is clearly iterative. In practice, the number of iterations needed for convergence seems to be quite small, most often less than 10.

The fact that the estimator, $\mathbf{W}^*_{(m)}$, needs a multiplicative constant to make it an estimator of the underlying common within-cluster covariance matrix does not limit its usefulness in the context of cluster analysis since the effect of omitting the constant is an inability to distinguish among scatters of the points when subjected to uniform dilation or shrinking. The missing constant, therefore, has no *relative* effect on the scatter of the points or their interpoint distances. The $\mathbf{W}^*$-algorithm, while distinctly different in its motivation and setting, is similar in spirit to the ellipsoidal trimmed robust estimator of dispersion defined in Eq. 75 in Section 5.2.3.

Gnanadesikan et al. (1995) provide comparisons, and demonstrate the advantages, of using the W*-algorithm as against others that ignore possible cluster structure in the data. The appendix on computer programs and software mentions currently available implementations of the W*-algorithm.

For hierarchical clustering algorithms (see Section 4.3.2), the input needed is a set of proximity or similarity values. With metric data, for instance, if the interest is in clustering the n "observations" then the input should be a set of inter-observation distances. Choices for the distance function or metric would include the so-called city-block or Manhattan metric defined by

$$d_{ii'} = \sum_{j=1}^{p} |y_{ij} - y_{i'j}|,$$

and squared distance functions such as those described in Section 4.2.1. For the squared distance functions, in the clustering context there are once again the issues of the desirability of basing estimates of variances and of covariance matrices on within-cluster information, and the W*-algorithm described earlier is useful. If one were interested in clustering the p variables (instead of the n observations) using a hierarchical algorithm, then one could use a measure of association such as the values of the correlation coefficient between every pair of variables.

A major practical advantage of hierarchical clustering algorithms is that they can handle non-metric data. All that they need as input is a set of values of proximities. As such, if the data consist of subjective similarity (or dissimilarity) judgements, as they tend to be in market research or psychological data, they can be used directly as input. In fact, for two of the most widely used versions of hierarchical clustering methods called the maximum and minimum methods in Section 4.3.2b, the rank orders of similarities (dissimilarities) is all that is needed as input. Also, if the data pertains to binary variables (e.g., presence or absence of traits), one can use a variety of measures of association for such data as inputs to hierarchical clustering. More specifically, if there are p binary variables each assuming the values 0 or 1, suppose the following 2 × 2 table summarizes the frequencies of 0's and 1's in the ith and jth observations, $\mathbf{y}_i$ and $\mathbf{y}_j$:

$\mathbf{y}_j$ \ $\mathbf{y}_i$	1	0	Total
1	a	b	$a + b$
0	c	d	$c + d$
Total	$a + c$	$b + d$	p

Then, one measure of similarity that one could use is the familiar chi-squared statistic computed from this table:

$$\chi^2 = \frac{p(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}.$$

One could also use the square root of this statistic. Other measures of dissimilarity for such data include

$$d = 1 - [(a + d)/p],$$

as well as alternatives such as the one based on the Jaccard coefficient,

$$d = 1 - [a/(a + b + c)].$$

Kaufman & Rousseeuw (1990) have an extensive discussion of these and other measures of association for binary data.

There are, of course, situations in which one might have a mix of types of variables. For example, some of the variables might be metric while others are qualitative, perhaps even binary. For hierarchical clustering of such data, one suggestion for measuring interobservation dissimilarity is to use the sum of two pieces, each of which is a measure of dissimilarity: one piece based on a meaningful distance function for the metric variables and the second piece based on a measure of dissimilarity such as the above examples for the binary variables. This way of combining the information from variables of different types is simple, but clearly does not take into account any information on the "association" between the metric and the binary variables.

4.3.2. Clustering Algorithms

Graphical displays of the data can be useful in revealing clusters. With two- and three-dimensional data, scatter plots can reveal clusters. With higher dimensional data, looking at all pairwise scatter plots can be useful but may not always be revealing and other displays may be needed. Friedman & Tukey (1974) initiated *projection pursuit* as a means of finding interesting projections of high-dimensional data, including those that reveal clusters. Glyphs and Chernoff Faces (see Section 3.2) have been used for visually grouping similar observations (see Example 13). Andrews's Curves, a method of mapping multivariate observations into curves plotted in two dimensions (see Section 6.2), have the property that observations that are close in the p-dimensional space of the variables are mapped close together as curves (see Example 41) and hence have particular value for detecting clusters of observations when n is relatively small ($\leqslant 50$, say).

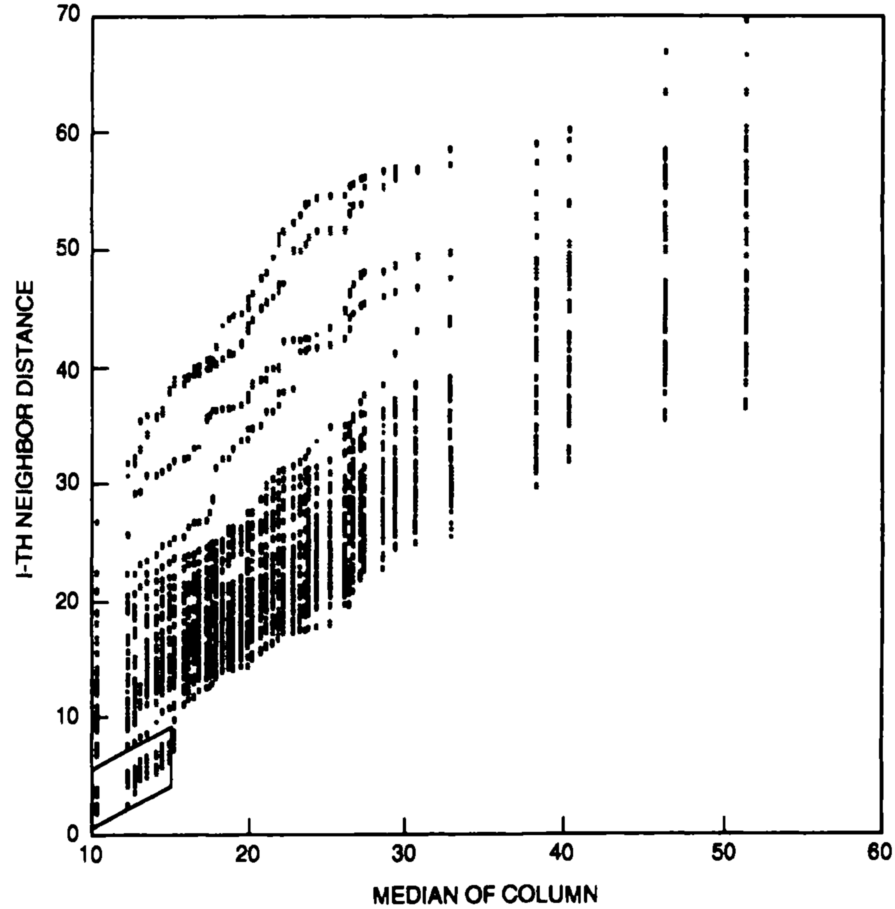
Cohen et al. (1977) describe a method of displaying nearest-neighbor distances and illustrate its use in finding clusters. Given an $n \times n$ matrix of inter-observation distances, d_{ij}, the method starts by ignoring the uninteresting zero distance of each observation from itself and sorting the remaining $(n - 1)$ distances within each row from smallest to largest. The $n \times (n - 1)$ matrix of sorted distances that results from this process will thus contain values of the nearest-neighbor distances in the first column, the second nearest-neighbor distances in the second column, and so on with the farthest-neighbor (i.e., the $(n - 1)$th nearest-neighbor) distances appearing in the last column. The so-called *nearest-neighbors distances plot* is a scatter plot of the values in each column of the sorted distances matrix along the ordinate axis against the median of the values in that column along the abscissa axis. Thus, if $d_{i(1)} \leqslant d_{i(2)} \leqslant \cdots \leqslant d_{i(n-1)}$ denote the sorted values (after omission of d_{ii}) in the ith row of the $n \times n$ matrix of distances, the nearest-neighbors distances plot is just a scatter plot of

$$\left(\underset{i}{\text{median}}\, d_{i(j)}, d_{i(j)}\right), \ i = 1, \ldots, n; j = 1, \ldots, (n - 1).$$

Information on clustering present is revealed by the configuration at the lower left end of the nearest-neighbors distances plot while the presence of "outliers" (which can be visualized as singleton clusters) is detected by the top portion of the plot.

Example 20. The technique is illustrated by Cohen et al. (1977) using data on the quarterly rates of return on $n = 52$ investment portfolios over a period of 11 successive quarters. If r_{ij} denotes the rate of return on the ith portfolio in the jth quarter, then the value of the Manhattan metric, $d_{ii'} = \Sigma_{j=1}^{11} |r_{ij} - r_{i'j}|$, was used as a measure of distance between the ith and i'th portfolios. Before proceeding with a clustering of the 52 portfolios using these distances, a nearest-neighbors distances plot was made following the steps mentioned above. The resulting plot is shown in Exhibit 20. The largest two distances in all but the last four columns in this display involved either portfolio #47 or #52, that is, $\max_i d_{i(j)} = d_{47(j)}$ or $d_{52(j)}$, $j = 1, \ldots, 47$. These two portfolios along with two others are nearly always associated with the four largest distances in each column. These may therefore be "outliers" which can then be isolated for further study (e.g., are they portfolios with consistently high rates of return?). Returning to Exhibit 20, there is a noticeable blob in the lower left-hand section of the plot, highlighted by a box around the points. In particular, the smallest eight distances in the first eight columns turned out to involve a set of nine portfolios which thus seem to belong to a cluster in that their profiles of rates of return are very close. Subsequent investigation revealed that these nine portfolios were in fact managed by the same person! For privacy reasons, no information was provided initially about the identities of

Exhibit 20. Nearest-neighbor distances plot for portfolio returns



the managers of the different portfolios. The "back door" discovery of the fact that the performances of nine portfolios were so similar as to suggest the possibility that they were managed by the same person, or a group of people with very similar investment strategies, impressed the suppliers of the original data!

Aside from static graphical displays mentioned above, in recent decades dynamic displays have been developed for looking at high-dimensional data. Many of these systems (see, for example, Fisherkeller et al., 1974; Azimov et al., 1988; Buja & Hurley, 1990; Cook et al., 1993; Swayne et al., 1991) have as a major motivation the finding and displaying of interesting projections of the data, including those that exhibit clustering.

As to numerical algorithms for clustering, there is a bewildering choice of types and instances of methods available. Despite the considerable computing power at one's disposal today, looking at all possible partitions of the data for determining the clustering that is optimal with respect to some criterion continues to be prohibitively expensive and practically impossible. The state of the art has not changed dramatically from the one pointed out by Gower (1967) that the computations involved in looking at the $(2^{n-1} - 1)$ possible partitions of n units into two sets for choosing the partition with minimum within-sets sum of squares would take approximately $(n - 1)^2 2^{n-11}$ seconds on a 5-microsecond-access-time machine, so that with $n = 21$ units the time involved would be approximately 114 hours and with $n = 41$ it would be approximately 54,000 years! (See also Scott & Symons, 1971.) One reason for the large number of algorithms available today is perhaps the fact that they are all attempts at "approximating," in some sense, the optimal partitions, and it is not surprising that one source of differences among currently available clustering schemes is their relative computational efficiency. Apart from computational issues, different methods of clustering a given data set may lead to different results and insights. From a data analysis viewpoint, this is not necessarily bad, and what is needed is help in assessing and interpreting the results of a cluster analysis. The approaches and aids discussed in Section 4.3.3 are addressed to this need.

At any rate, at the highest level, one can distinguish between algorithms that lead to mutually exclusive clusters and those that yield overlapping clusters. ADCLUS and MAPCLUS are examples of the latter category (see Shepard & Arabie, 1979; Arabie & Carroll, 1980). The more commonly used algorithms lead to mutually exclusive clusters. Such algorithms may be categorized broadly as being *hierarchical* (e.g., Hartigan, 1967; Johnson, 1967; Sokal & Sneath, 1963) or *nonhierarchical* (e.g., Ball & Hall, 1965; Friedman & Rubin, 1967). The former class is one in which every cluster obtained at any stage is a merger of clusters at previous stages. In this case, therefore, it is possible to visualize not only the two extremes of clustering, namely, n clusters with one unit per cluster (*weak clustering*) and a single cluster with all n units (*strong clustering*), but also a monotonically increasing strength of clustering as one goes from one level to another. In the nonhierarchical procedures, on the other hand, new clusters are obtained by both lumping and splitting of old clusters and, although the two extremes of clustering are still the same, the intermediary stages of clustering do not have this natural monotone character of strength of clustering.

The format of the input data for clustering procedures may be metric or nonmetric, that is, as a representation of n points in p-space or only as rank order information regarding the similarities of pairs of the n units. The descriptions of most nonhierarchical schemes seem to assume a metric input with an implied choice of p as well. This, however, is not a necessary limitation, since the observed ordering of the similarities may be utilized as input to multidimensional scaling (see Section 2.3) for obtaining a representation of the n units in a Euclidean space whose dimensionality is data determined. Also, even with metric data inputs, if redundancy among the p coordinates is suspected, the original data may first be transformed to a reduced dimensional space by using linear or generalized principal components analyses (see Sections 2.2.1 and 2.4.2), and then the clustering may be performed in the lower-dimensional linear or nonlinear subspace of the original p-dimensional space. Caution must, however, be exercised in any reduction of dimensionality which ignores the presence and nature of clusters in the data. The key issue here is the nature of the spread among the clusters relative to the within-cluster dispersions, and the possible misleading indications of any reduction of dimensionality that ignores this.

In any specific application, whether one uses hierarchical or nonhierarchical methods is largely dependent on the meaningfulness, in the particular situation, of the tree structure imposed by hierarchical clustering procedures. For instance, in biological applications concerned with groupings of species, clusters of species, subclusters of subspecies, and so on may be of interest, and hierarchical clustering may then be a sensible approach to adopt. Even the area of numerical taxonomy, however, is not without controversy as to the biological meaningfulness of clusters (hierarchical or otherwise) determined by the use of statistical data-analytic techniques.

Section 4.3.2a will discuss hierarchical clustering methods, and Section 4.3.2b will be concerned with nonhierarchical clustering.

4.3.2a. Hierarchical Clustering Procedures

Hierarchical clustering algorithms come in two flavors: *agglomerative* (where one starts with each of the n units in a separate cluster and ends up with a single cluster that contains all n units) and *divisive* (where the process is to start with a single cluster of all n units and then form new clusters by dividing those that had been determined at previous stages until one ends up with n clusters containing individual units). Only agglomerative techniques are considered here.

The discussion of methods for hierarchical clustering in this subsection follows closely the development due to Johnson (1967). The essential idea of a hierarchical clustering scheme is that n units are grouped into clusters in a nested sequence of, say, $(m + 1)$ clusterings, $C_0, C_1, \ldots, C_m$, where C_0 is the weak clustering, C_m is the strong clustering, and every cluster in C_i is the union or merger of some clusters in C_{i-1} for $i = 1, \ldots, m$. Also, corresponding to C_i we have its "strength," α_i, where $\alpha_0 = 0$ and $\alpha_i < \alpha_{i+1}$ for $i = 0, 1, \ldots, (m - 1)$. The α's, therefore, are an increasing sequence of nonnegative numbers.

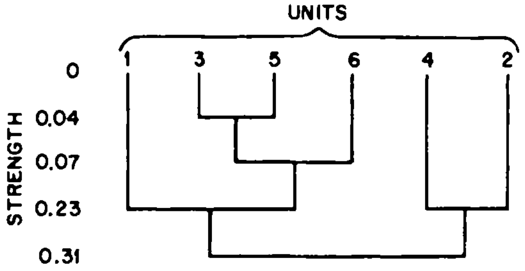
Johnson (1967) demonstrates that, for any such hierarchical clustering scheme, a metric for measuring the distance between every pair of the n units is implied, and, conversely, that, given such a metric, one can recover the hierarchical clustering scheme from it. Given two units, x and y, and the above hierarchical clustering scheme, let j be the smallest integer in the set $[0, 1, \ldots, m]$ such that in clustering C_j the units x and y belong to the same cluster; then define the distance between x and y, $d(x, y)$, to be the strength, α_j, of the clustering C_j. In other words, the distance between any pair of units is defined as the strength of the clustering at which the units first appear together in the same cluster. This definition leads to a distance measure with properties generally associated with metrics. Thus $d(x, y) = 0$ if and only if x and y first appear together in C_0, which means that x and y are not distinct units or that $x = y$. Also, from the definition it follows that $d(x, y) = d(y, x)$. Finally, if x, y, and z are three units, the triangle inequality $d(x, z) \leqslant d(x, y) + d(y, z)$ may be shown to hold. In fact, a stronger inequality, namely, $d(x, z) \leqslant \max\{d(x, y), d(y, z)\}$, which implies the triangle inequality, may be shown to be satisfied by d. This stronger inequality, which states that the distance between x and z cannot exceed the larger of the two distances, $d(x, y)$ and $d(y, z)$, has been called the *ultrametric inequality* by Johnson (1967). That the above definition of distance between pairs of units in a hierarchical clustering scheme satisfies the ultrametric inequality may be established as follows. Let $d(x, y) = \alpha_i$ and $d(y, z) = \alpha_j$, so that the units x and y appear together in the same cluster for the first time in clustering C_i, and the units y and z do so in clustering C_j. Then, because of the hierarchical nature, one of these clusters includes the other, namely, the one which appears in the clustering whose index corresponds to the larger of i and j includes the other. Hence, if $k = \max(i, j)$, in clustering C_k the units x, y, and z are all in the same cluster and, clearly, $d(x, z) \leqslant \alpha_k = \max(\alpha_i, \alpha_j)$. Thus, given a hierarchical clustering scheme such as the one in the preceding paragraph, one may derive a metric satisfying the ultrametric inequality (and hence the triangle inequality).

The converse—namely, given a set of $n(n - 1)/2$ interunit values of a metric that satisfies the ultrametric inequality, one may recover a hierarchical clustering of the n units—is also demonstrated by Johnson (1967). The equivalence between hierarchical clustering and a metric that satisfies the ultrametric inequality is perhaps most easily shown by the following simple example, taken from Johnson (1967).

Example 21. Exhibit 21*a* shows a hierarchical clustering of six $(= n)$ units involving five $(= m + 1)$ stages of clustering, $C_0, \ldots, C_4$, with respective associated strengths $\alpha_0, \ldots, \alpha_4$ ranging from 0 to 0.31. [*Note*: The value of the strength of each clustering is, for the moment, assumed to be specified, and the later discussion in this subsection will deal with how these strengths are actually obtained in various hierarchical clustering algorithms.]

Using the earlier-mentioned definition of a distance between a pair of units, one may derive the matrix of interunit distances shown in Exhibit 21*b*. Thus, since every unit "appears with itself in the same cluster" for the first time in C_0 and $\alpha_0 = 0$, the diagonal elements are all 0. Also, for instance, since units 3 and 5 are clustered for the first time in C_1 with $\alpha_1 = 0.04$, the distance between these units is 0.04, while the distance between units 1 and 2 is 0.31, the strength of C_4, the strong clustering, which is the first stage in which units 1 and 2 are clustered together. All of the metric properties claimed for the definition of distance used in obtaining Exhibit 21*b* from Exhibit 21*a* can be verified in this example in terms of the elements of the distance matrix shown in Exhibit 21*b*.

Exhibit 21*a*. Example of hierarchical clustering tree



Next, the inverse process of going from Exhibit 21*b* to Exhibit 21*a* may be demonstrated in terms of this simple example. To start the process, at the first level we form the weak clustering C_0 with six clusters containing one unit each. Next, by scanning the elements of the distance matrix in Exhibit 21*b*, we identify the smallest interunit distance as being 0.04, the distance between units 3 and 5. In a natural manner, suppose that we decide to create a cluster (3, 5) and to leave the remaining four units by themselves, thus leading to five clusters at level C_1 with associated strength α_1 equal to the smallest distance, 0.04, in the distance matrix of Exhibit 21*b*. To repeat the process for constructing the higher-level clusterings, we now need a way of defining distances between cluster (3, 5) and the four units 1, 2, 4, and 6. An interesting property (shown below to be a consequence of the distance function here satisfying the ultrametric inequality) of the values in Exhibit 21*b* is that $d(3, x) = d(5, x)$ for $x = 1, 2, 4$, and 6. Hence a natural measure of the distance between cluster (3, 5) and unit x would be $d([3, 5], x) = d(3, x) = d(5, x)$ for $x = 1, 2, 4$, and 6. Exhibit 21*c* shows a 5 × 5 distance matrix obtained by using this definition.

Scanning Exhibit 21*c* for the smallest element, we find that it is $0.07 = d([3, 5], 6)$, and we can then form a cluster (3, 5, 6) while leaving units 1, 2, and 4 by themselves, thus obtaining four clusters in stage C_2 with associated

Exhibit 21*b*. Initial distance matrix for the example in Exhibit 21*a*

6 × 6 *DISTANCE MATRIX*

	1	2	3	4	5	6
1	0	0.31	0.23	0.31	0.23	0.23
2	0.31	0	0.31	0.23	0.31	0.31
3	0.23	0.31	0	0.31	0.04	0.07
4	0.31	0.23	0.31	0	0.31	0.31
5	0.23	0.31	0.04	0.31	0	0.07
6	0.23	0.31	0.07	0.31	0.07	0

Exhibit 21*c*. Distance matrix after forming first cluster in Exhibit 21*a*

5 × 5 *DISTANCE MATRIX*

	1	2	(3,5)	4	6
1	0	0.31	0.23	0.31	0.23
2	0.31	0	0.31	0.23	0.31
(3,5)	0.23	0.31	0	0.31	0.07
4	0.31	0.23	0.31	0	0.31
6	0.23	0.31	0.07	0.31	0

strength $\alpha_2 = 0.07$. It is now clear that by repeating this process of constructing distance matrices and scanning them for a minimum value, for deciding on which clusters to merge and for determining the value for the strength of clustering, we can recover the entire hierarchical clustering shown in Exhibit 21*a*. At any stage in this process, there may be a tie, that is, more than one interentity distance [e.g., the distance between cluster (3, 5, 6) and unit 1 and the distance between units 2 and 4 are both 0.23] may correspond to the smallest value in the distance matrix. This merely implies that parallel clusters are being formed at such a stage.

The foregoing simple example provides a basis for the following summarization of the steps involved in going from a matrix of values of a metric satisfying the ultrametric inequality to a hierarchical clustering of n units in a nested sequence of $(m + 1)$ clusterings, $C_0, C_1, \ldots, C_m$:

1. Form C_0 of strength 0 by considering each unit as a cluster.
2. Given a clustering C_j with a corresponding matrix of interentity (where an entity may be a single unit or a cluster of units) distances that satisfy the ultrametric inequality, merge the pair of entities with the smallest nonzero distance, α_{j+1}, to create C_{j+1} with strength α_{j+1}.
3. Create a new distance matrix corresponding to C_{j+1}.
4. Starting with $j = 0$, by repeated use of steps 2 and 3, generate $C_1, C_2, \ldots$ and C_m (the strong clustering).

The assumption in step 3 is that the distance matrix corresponding to C_{j+1} can be constructed in an *unambiguous* manner. That this is a consequence of the distances satisfying the ultrametric inequality can be demonstrated. Suppose that x and y are two entities in C_j that become clustered together in C_{j+1}, so that $d(x, y) = \alpha_{j+1}$, and suppose that z is any other entity in C_j. Then the unambiguous construction of the distance matrix corresponding to C_{j+1} is possible because $d(x, z)$ necessarily has to be equal to $d(y, z)$ if

d satisfies the ultrametric inequality. If $d(x, z) \neq d(y, z)$, then suppose that $d(x, z) > d(y, z)$. But, as a consequence of the ultrametric inequality, $d(x, z) \leqslant \max\{d(x, y), d(y, z)\}$, so that the inequality in the preceding sentence would imply that $d(y, z) < d(x, z) \leqslant d(x, y) = \alpha_{j+1}$. But α_{j+1} is the smallest nonzero distance in clustering C_j, and hence $d(y, z)$ cannot be smaller than α_{j+1}. Thus, assuming that $d(x, z) \neq d(y, z)$ leads to a contradiction, and therefore $d(x, z)$ has to be equal to $d(y, z)$.

In practice, the observed measures of distance between pairs of units may be either subjective measures of dissimilarity (or similarity) or measures of distance computed from metric data representing the n units, perhaps as n points in a p-dimensional space. Such measures of distance may not, and need not, satisfy the ultrametric inequality, with the consequence that the distance between cluster (x, y) and entity z may not be definable unambiguously since $d(x, z)$ need not necessarily equal $d(y, z)$ unless d satisfies the ultrametric inequality. Hence ways of defining $d([x, y], z)$, given $d(x, z)$ and $d(y, z)$, need to be devised for most situations. Considering $d([x, y], z)$ as a function, $f\{d(x, z), d(y, z)\}$, which is required to equal the common value of $d(x, z)$ and $d(y, z)$ whenever these are equal, leads to a fairly wide class of functions, f, including any weighted average, the geometric mean, etc. Motivated partially by the considerations underlying multidimensional scaling (viz., a monotone relationship between distance and dissimilarity), Johnson (1967) proposes two specific choices of f which, while satisfying the above constraint, are also invariant under monotone transformations of the distances. The two choices are $f =$ the min (or the smaller of) function and $f =$ the max (or the larger of) function. The methods based on these two choices have been called, respectively, the ***minimum method*** and the ***maximum method*** by Johnson (1967). In the numerical taxonomy literature, Sneath (1957) has proposed a hierarchical method called the *single linkage method*, and Sørensen (1948) a method called the *complete linkage method*. These two methods are, respectively, is the minimum and maximum methods suggested by Johnson (1967). At any rate the two methods of hierarchical clustering described by Johnson (1967) may now be summarized.

The Minimum Method (See Johnson, 1967; also Sneath, 1957)

1. Form C_0, consisting of n clusters with one unit each, with corresponding strength $\alpha_0 = 0$.
2. Given C_j with associated distance (or dissimilarity) matrix (where the observed values at stage C_0, for example, may not satisfy the ultrametric inequality), merge the entities whose distance, α_{j+1} (>0), is smallest to obtain C_{j+1} of strength α_{j+1}.
3. Create a matrix of distances for C_{j+1}, using the following rules: (a) if x and y are entities clustered in C_{j+1} but not in C_j [i.e., $d(x, y) = \alpha_{j+1}$],

then $d([x, y], z) = \min\{d(x, z), d(y, z)\}$; (*b*) if x and y are separate entities in C_j that remain unclustered in C_{j+1}, then do not change $d(x, y)$.

4. Repeat the process until the strong clustering is obtained.

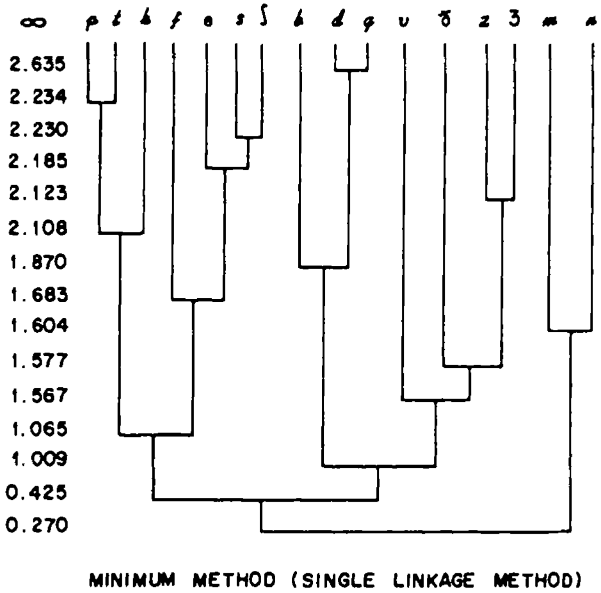
The Maximum Method. (See Johnson, 1967; also Sørensen, 1948). For this method, steps 1, 2, 3*b*, and 4 are the same as those in the minimum method. For step 3*a*, however, the following is substituted: if x and y are entities clustered in C_{j+1} but not in C_j, define $d([x, y], z) = \max\{d(x, z), d(y, z)\}$.

The two methods are directed toward different objectives and may not necessarily yield similar results when applied to the same body of data. The maximum method is concerned essentially with minimizing the maximum intracluster distance at each stage and hence tends to find compact clusters, sometimes forming several small clusters in parallel. The minimum method, on the other hand, tends to maximize the "connectedness" of a pair of units through the "intermediary" units in the same cluster (see Johnson, 1967, for a discussion of these interpretations), with a tendency to create fewer distinct clusters than the maximum method. When the initial data consist of p-dimensional representations of the n units and the interpoint distances in the representation are used as the elements of a distance matrix, both methods may tend to be unduly sensitive to outliers. For this reason, defining $d([x, y], z)$ in step 3*a* of the preceding descriptions as the average of $d(x, z)$ and $d(y, z)$ may be preferable. This method, which may be called the *averaging method*, does not, however, possess the property of invariance under monotone transformations of the distances. A different but critical issue, in the case when the initial representation of the n units is metric, is the dependence of the clusters obtained on the type of distance function used for measuring interunit distance. The results of the clustering techniques described heretofore seem to be highly dependent on the specification of a metric for measuring interunit distance.

Example 22. This example, taken from Johnson (1967), deals with data from Miller & Nicely (1955) pertaining to the confusability of 16 consonant sounds. The observed data were the values of the frequency, $f(x, y)$, with which the consonant phoneme x was heard as the consonant phoneme y by a group of human listeners. Different levels of both filtering and noise were involved in the experiment, and the frequency of confusions for each pair of consonant sounds was observed separately for each experimental condition.

For purposes of hierarchical clustering of the 16 consonants under each experimental condition, Johnson (1967) defines the symmetric measure of similarity, $s(x, y) = f(x, y)/f(x, x) + f(y, x)/f(y, y)$, and considers it as an inverse measure of distance (i.e., the similarity increases as distance decreases, and, in particular, $d = 0$ is taken to correspond to $s = \infty$, which would be the strength of C_0 in terms of s).

Exhibit 22a. Hierarchical clustering obtained by minimum (or single linkage) method for data on confusability of 16 consonant sounds (Johnson, 1967)

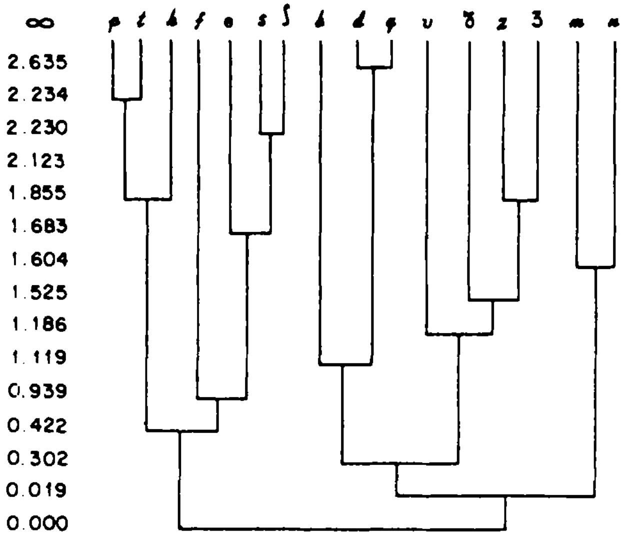


Exhibits 22*a* and *b* show the solutions obtained by Johnson (1967) by using the minimum and maximum methods, respectively. The two clustering solutions in this example are generally similar, although there are differences both in the numerical values of the strengths of the clusterings and in the stage of clustering when specific clusters are formed. For instance, two of the so-called unvoiced stop consonants, p and t, join the third one, k, earlier in the maximum than in the minimum method. Also, in respect to the unvoiced fricatives, f, θ, s, ʃ, f joins (θ, s, ʃ) earlier in the minimum than in the maximum method.

The similarity between the two solutions also extends to the order in which the consonant phonemes or groups of them come together, with the exception of the manner in which the last three groups merge. In the maximum method, the voiced consonant phonemes (both the voiced stops, b, d, g, and the voiced fricatives, v, ð, z, ʒ) merge with the nasals, m, n, and their combination then merges with the unvoiced consonant phonemes (the unvoiced stops, p, t, k, and the unvoiced fricatives, f, θ, s, ʃ). With the minimum method, the voiced and the unvoiced consonants merge before their combined group joins the nasals.

In this example the manner of the hierarchical grouping of the consonants makes sense in terms of what is known about the discrimination of consonant phonemes. First, the stops and the fricatives in both the unvoiced and the voiced categories come together in four separate groups, whereas the nasals combine by themselves to form a fifth group. Then the unvoiced and voiced

Exhibit 22*b*. Hierarchical clustering obtained by maximum (or complete linkage) method for data on confusability of 16 consonant sounds (Johnson, 1967)



MAXIMUM METHOD (COMPLETE LINKAGE METHOD)

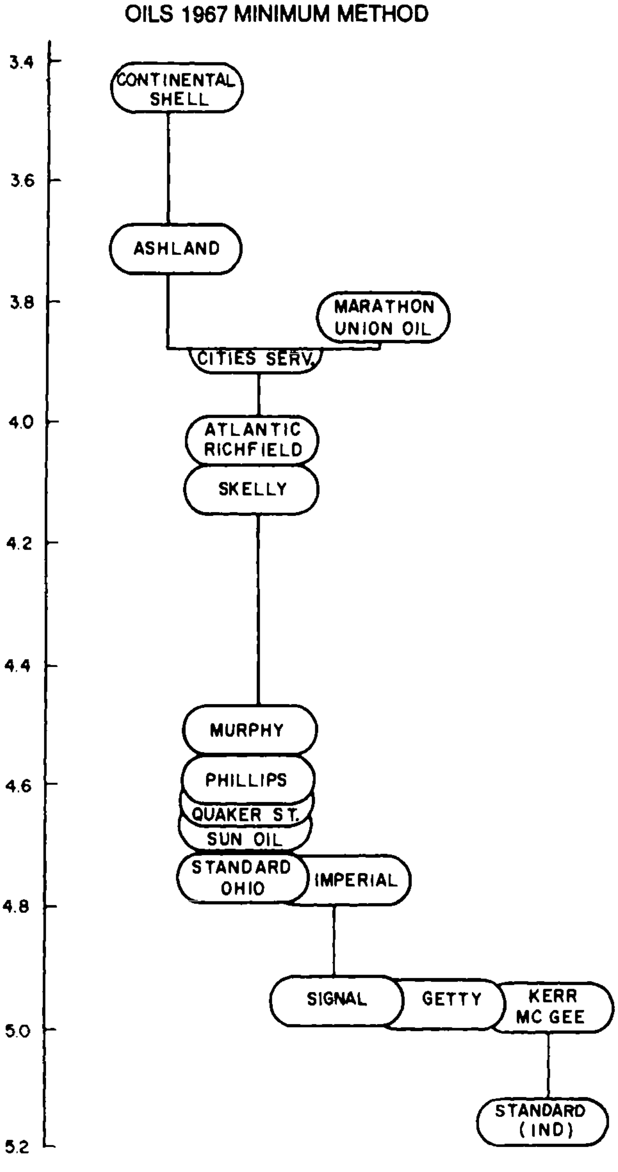
consonants coalesce into two separate goups, while the nasals constitute the third group.

The close similarity between the solutions obtained by the two methods in this example is, although comforting, not necessarily to be expected in all applications, as illustrated by the next example.

Example 23. As a part of the corporation-grouping study (see Examples 17 and 18), clustering procedures were employed to investigate the structure among specific groups of companies as indicated by the observations on the 14 variables involved. Thus, for the year 1967, hierarchical cluster analyses were performed for 18 of the domestic oil companies, and Exhibits 23*a–c* show, respectively, the results obtained by the minimum, average, and maximum methods. [*Note*: The type of representation used in these figures is different from the one in Exhibits 22*a* and *b*. Instead of a listing of all the objects at the top, with lines emanating downward from them being joined together by horizontal lines at the different clustering levels, the representation in Exhibits 23*a–c* shows the companies by their names as the clusters are formed. Thus, in Exhibit 23*a*, the first cluster to form at a clustering strength of about 3.4 consists of two oil companies, Continental and Shell, which are next joined by Ashland, then merged with a cluster consisting of Marathon and Union Oil, and so on. The scale on the left depicting the values of the strength of clustering is the same for all three figures.]

A striking feature of Exhibits 23*a–c* is that, in this example, the minimum method exhibits its characteristic tendency of stringing out the clusters,

Exhibit 23*a*. Hierarchical cluster of 18 domestic oil companies obtained by minimum method



bringing the companies in one at a time, whereas the maximum method appears to form several compact clusters (see the four branches in Exhibit 23*c*), which come together quite late in the hierarchical clustering, and the average method is intermediary between the minimum and maximum methods. Except for this feature, however, the general indications of the clustering from the three methods are not markedly different in terms of which companies appear to be together and of the strength of clustering at which a particular company is brought into an existent cluster.

Exhibit 23*b*. Hierarchical cluster of 18 domestic oil companies obtained by averaging method

OILS 1967 AVERAGING METHOD

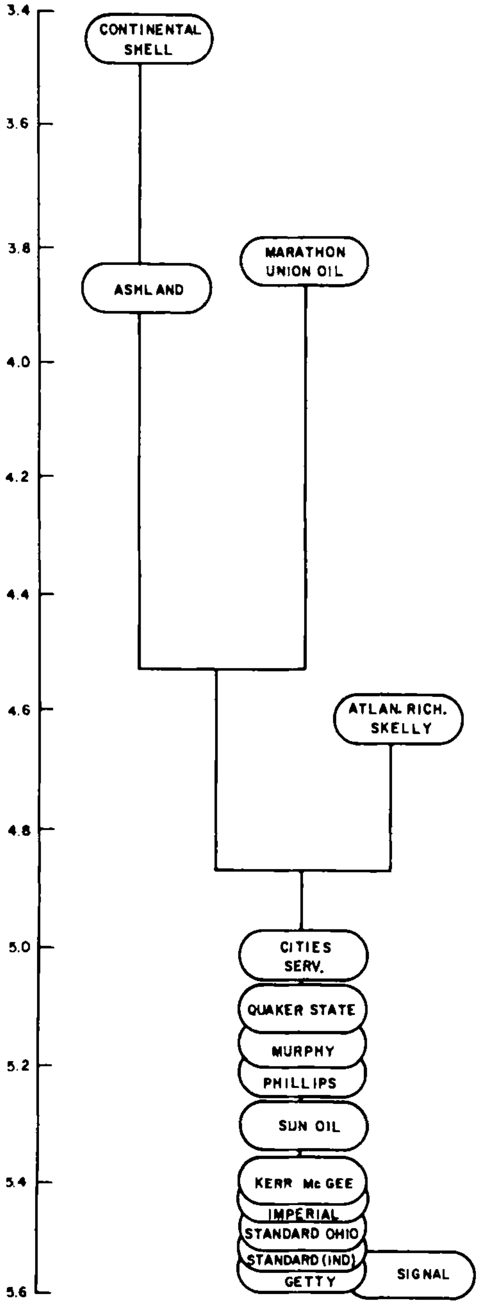
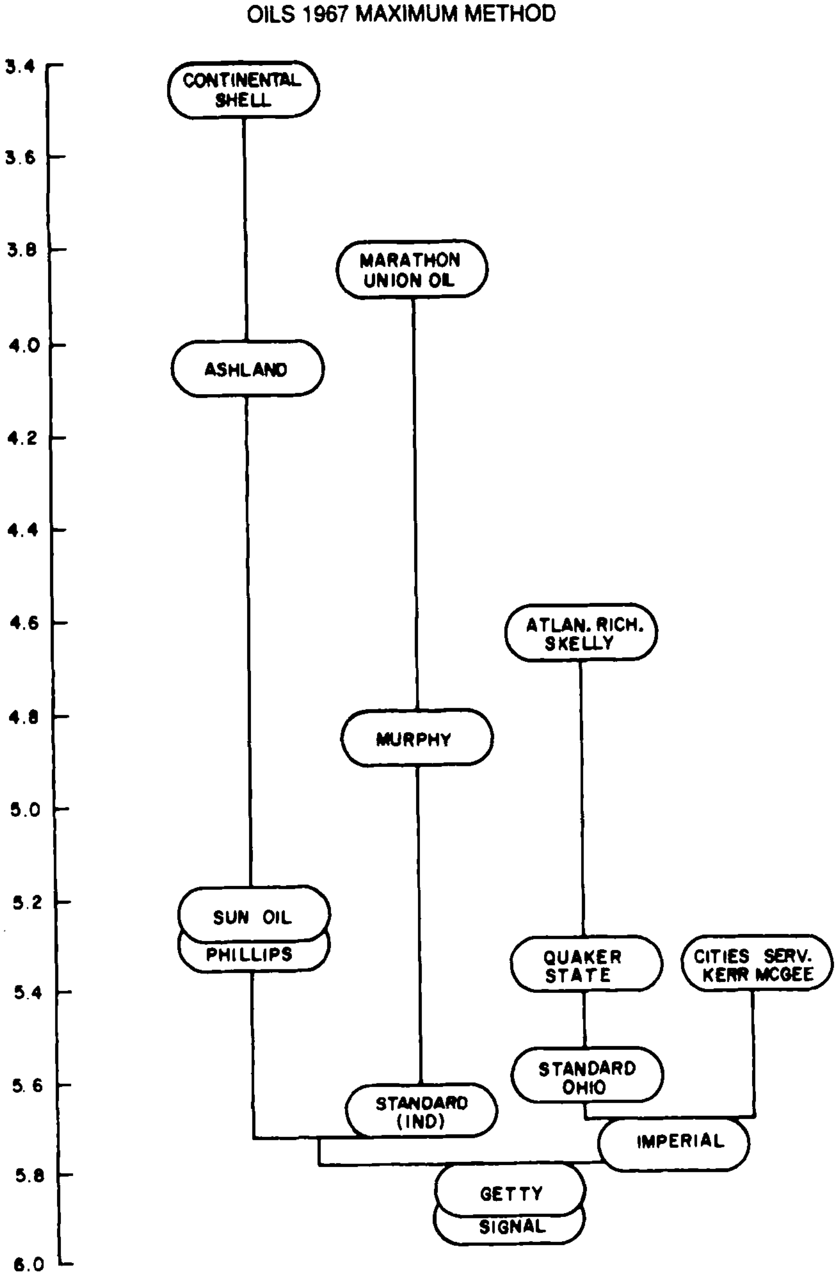


Exhibit 23c. Hierarchical cluster of 18 domestic oil companies obtained by maximum method



4.3.2b. Nonhierarchical Clustering Procedures

For this class of procedures, the starting point is the $p \times n$ data matrix $\mathbf{Y}$. The best known of the nonhierarchical methods is the so-called k-means method. MacQueen (1965) contains an early description of the method but different versions have since appeared as implementations. The steps involved in a generic description of the k-means algorithm are:

(i) Determine an initial set of k clusters.

(ii) Move each observation to the cluster whose centroid/mean is closest in distance.

(iii) Recalculate the cluster centroids/means and repeat step (ii) until no observation is reassigned to a new cluster.

Choices of more than one type of metric for use in step (ii), and of a range of values for k, constitute two sources of variation in implementing the technique. Standard implementations of k-means are based on the Euclidean metric. In this case, it may be prudent to sphericize the data initially, using the Cholesky decomposition of a $\mathbf{W}^*_{(m)}$ matrix computed by the method described in Section 4.3.1, before carrying out k-means clustering.

Different ways of choosing the initial clusters also result in different versions. Suggestions for the initial choice of clusters range from a random partitioning into k nonempty clusters to a variety of things based on statistics computed from the data. The hope is that the final results will not depend critically on the choice of initial clusters.

A different nonhierarchical clustering method called ISODATA (Iterative Self Organizing Data Analysis Technique A) was proposed by Ball & Hall (1965). The input to the procedure consists of the n p-dimensional observations, $\mathbf{y}_i$, $i = 1, 2, \ldots, n$, on the n units to be clustered. The method also requires an initial specification of the number, k $(<n)$, of clusters desired and a set of so-called cluster points in p-dimensions, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$.

The first stage of the process is referred to as *sorting* and consists of assigning each of the n units to one of the k clusters, $C_1, C_2, \ldots, C_k$, by using the criterion of closeness (as measured by Euclidean distance) of the observation on the unit to the cluster point. Thus the ith unit is assigned to cluster C_l if

$$(\mathbf{y}_i - \mathbf{x}_l)'(\mathbf{y}_i - \mathbf{x}_l) = \min_{a=1,\ldots,k} (\mathbf{y}_i - \mathbf{x}_a)'(\mathbf{y}_i - \mathbf{x}_a).$$

If cluster C_r consists of n_r, units with corresponding observations denoted, $\mathbf{y}_{rs}$ $(r = 1, \ldots, k;\ s = 1, \ldots, n_r;\ \Sigma n_r = n)$, then in the sorting stage all clusters with no units assigned to them are discarded, and for the l $(\leqslant k)$ remaining clusters the initially specified cluster points are replaced by the mean vectors, $\bar{\mathbf{y}}_r$ (centroids), of the observations within each of the clusters. Thus the accomplishment of the sorting phase of ISODATA is to form preliminary clusters and to utilize them to define more appropriate cluster points or typical values of the clusters.

At the next stage additional cluster statistics are computed. Specifically, the following statistics are calculated: (i) for each cluster C_r, the average distance of points from the cluster centroid, that is,

$$\bar{d}_r = \frac{1}{n_r} \sum_{s=1}^{n_r} d(\mathbf{y}_{rs}, \bar{\mathbf{y}}_r),$$

where $d(\mathbf{x}, \mathbf{z})$ denotes a defined measure of distance between the points $\mathbf{x}$ and $\mathbf{z}$; (ii) for each cluster C_r, the $p \times p$ covariance matrix,

$$\mathbf{S}_r = \frac{1}{n_r - 1} \sum_{s=1}^{n_r} (\mathbf{y}_{rs} - \bar{\mathbf{y}}_r)(\mathbf{y}_{rs} - \bar{\mathbf{y}}_r)';$$

and (iii) the average intracluster distance across all clusters, that is,

$$\bar{d} = \frac{1}{n} \sum_{r=1}^{l} n_r \bar{d}_r.$$

The statistics computed under (i) and (iii) are of descriptive value in that they provide measures of "tightness" of the clusters. The covariance matrix, $\mathbf{S}_r$, is used for basing decisions pertaining to the formation of new clusters at the next stage.

The third stage of the ISODATA scheme is the formation of new clusters by *splitting* apart, or *lumping* together, existent clusters. Roughly speaking, if one has too few clusters, splitting will be desirable, if there are too many clusters, lumping will be more appropriate. The splitting or lumping is actually accomplished by comparing certain cluster properties against user-specified values (benchmarks) of two parameters, θ_S and θ_L, called, respectively, the splitting and lumping parameters. Specifically, (*a*) if the maximum coordinate variance in a cluster C_l exceeds the specified value of θ_S, that is, if the largest diagonal element of $\mathbf{S}_l > \theta_S$, then C_l is split along that coordinate into two new clusters; or (*b*) if the variance of the first principal component within a cluster C_l exceeds the specified θ_S, that is, if the largest eigenvalue of $S_l > \theta_S$, then C_l is split along the direction of the first principal component into two new clusters. The user has a choice between the two methods of splitting. The decision to lump two clusters, C_r and C_s, is based on comparing the distance between the two cluster points, $\bar{\mathbf{y}}_r$ and $\bar{\mathbf{y}}_s$, with the specified value of the lumping parameter, θ_L. If the distance, $d(\bar{\mathbf{y}}_r, \bar{\mathbf{y}}_s)$, is smaller than the value of θ_L, C_r and C_s are combined into a single cluster whose centroid is then computed and used as the cluster point of the merged cluster.

The remaining step of the ISODATA process is to iterate the three above-mentioned stages. In the early iterations the method tends to alternate

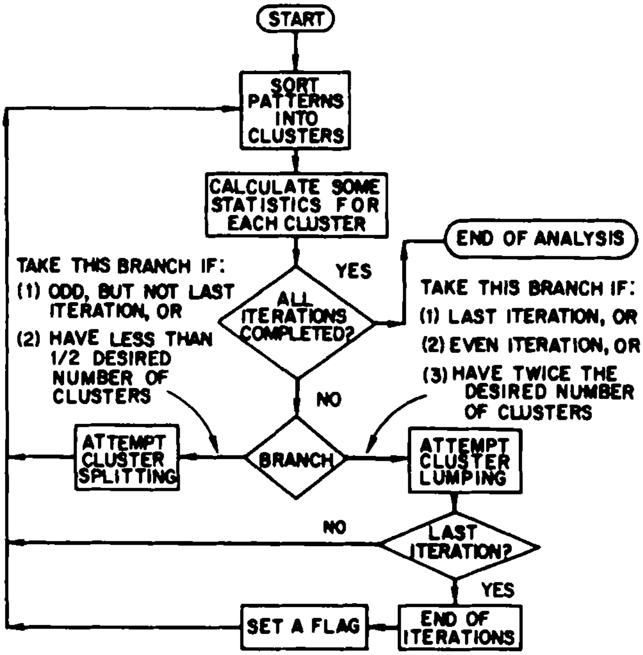

START
SORT PATTERNS INTO CLUSTERS
CALCULATE SOME STATISTICS FOR EACH CLUSTER
ALL ITERATIONS COMPLETED?
YES
END OF ANALYSIS
TAKE THIS BRANCH IF:
(1) ODD, BUT NOT LAST ITERATION, OR
(2) HAVE LESS THAN 1/2 DESIRED NUMBER OF CLUSTERS
TAKE THIS BRANCH IF:
(1) LAST ITERATION, OR
(2) EVEN ITERATION, OR
(3) HAVE TWICE THE DESIRED NUMBER OF CLUSTERS
NO
BRANCH
ATTEMPT CLUSTER SPLITTING
ATTEMPT CLUSTER LUMPING
LAST ITERATION?
NO
YES
END OF ITERATIONS
SET A FLAG

Fig. 8. Flowchart of steps in ISODATA algorithm.

between splitting and lumping, but in the later iterations a comparison of the number of clusters found at the end of a given iteration with the initially desired number, k, of clusters also influences the decision to split or lump clusters in the next iteration. The number of iterations is also a specification under the user's control. In fact, the process will terminate at the end of the specified number of iterations, and the number of clusters found may not be exactly k, the initially desired number. Figure 8 shows a summary flowchart of the steps involved in one computer implementation of the ISODATA procedure (see Warner, 1968).

Although ISODATA is nonhierarchical in that it allows splitting (in addition to lumping) of existent clusters to form new ones, because it starts off with all units in a single cluster and then alternates between splitting and lumping there is a tendency in the process to impose a loose tree (hierarchical) structure on the clusters.

The algorithm requires the user to specify several things (the number of clusters desired, the number of iterations, the initial values of cluster points, and the values of θ_S and θ_L), and, in the present state of the art, such specifications tend to be arrived at by a trial and error process. Little is known regarding the statistical-inferential aspects of cluster analysis techniques, and no general guidance or simple data-dependent method is available for choosing particular values of the quantities that need to be specified. All that one can say is that the number of clusters desired may not be a critical specification in that it is not guaranteed anyway and that ISODATA appears to be reasonably robust to the initial choice of cluster points provided that the number of iterations specified is not inadequate. This robustness implies that, if the user does not wish to specify the initial cluster points, the use of default values will probably not excessively distort the final results.

Since the splitting methods in ISODATA are based on variances (either of original variables or of the first principal component), and since the Euclidean metric is employed as the basis for measuring closeness, the results are scale dependent in that different scalings of the initial variables can in general lead to different clusterings of the objects. The nonhierarchical clustering method proposed by Friedman & Rubin (1967) uses eigenvalues and eigenvectors of a $\mathbf{W}^{-1}\mathbf{B}$ type of matrix ($\mathbf{W}$ would be the pooled within-clusters covariance matrix, and $\mathbf{B}$ the between-clusters covariance matrix for a given set of clusters) for basing the decisions regarding splitting and lumping. Although computationally more involved and expensive, this method does, of course, have the property that a solution obtained by its use will be invariant under all affine transformations of the original set of variables, including simple scaling of each of them separately.

ISODATA is similar in spirit to the k-means method, one main difference being in the cluster splitting and the other being that the number of clusters initially specified may not be the number of final chapters. Whereas ISODATA bases splitting on intracluster dispersion characteristics, in the k-means method the creation of new clusters depends on the distance of the proposed new cluster centroid from the nearest existent cluster centroid exceeding some prespecified value of a splitting parameter. MacQueen (1965) has established some asymptotic properties of the k-means method, and, in view of the similarities between the methods, it would be interesting to ask whether these properties carry over to ISODATA.

4.3.3. Outputs

The typical output of a hierarchical clustering algorithm is a tree which can be cut at different levels to produce clusters. With a nonhierarchical method, the output often is a list of clusters and their members. From a data analysis viewpoint, what is needed is help in understanding and interpreting the clusters that have emerged. When there is strong clustering present in the data, perhaps they will stand out and be recognizable virtually independent of the method of clustering used or the details. However, this is often not the case in practice and statistical aids are needed for assessing various aspects of the results of a cluster analysis. Despite the lack of attention to the development of such tools, there are a few available and these will be discussed in the present section.

The facets of the results of a cluster analysis which one would like some help in understanding include: (i) separations among the clusters; (ii) the relative tightnesses of the different clusters; (iii) the orientations or shapes of the clusters; and (iv) the relative stabilities of the clusters that have been found, that is, which clusters if any are "strong" or "real" enough to be believable. By formulating the last of these as a test of significance problem, based on certain assumed distributional or probabilistic models, some authors have proposed formal tests. (See, for example, Baker & Hubert, 1975; Hubert, 1974; Ling, 1973.) These procedures have limitations, aside from the models assumed, in terms of the numbers of clusters and observations that can be handled, and of the distribution of the observations across clusters that can be accommodated. In the remainder of this section, a number of informal data-analytic aids, many reliant on graphical displays, are described and illustrated for purposes of understanding and interpreting the outputs of clustering algorithms.

A simple starting point for understanding the clusters determined by any algorithm is to see what characteristics, as measured by the different variables, are shared by members of the same cluster and how the clusters differ from each other. A *profile plot* of the deviations of the cluster means (or medians) from the overall mean (or median) of each variable is useful for this. Also, scatter plots of the observations, identified by their cluster membership, for all pairs of variables can be helpful.

Example 24. Fowlkes et al. (1988) illustrate the use of such plots. The data involved microwave attenuation measurements at $n = 51$ locations in the U.S., as well as measurements on seven environmental variables that were thought to be important influencers of the attenuation. Fitting a global model to all of the data rather than to subsets consisting of meaningfully comparable entities can be misleading. To avoid this in the present example, it was decided to do a preliminary clustering of the 51 locations in terms of the environmental variables to be followed up with modeling (e.g., via regression) of the relationship between attenuation and the explanatory variables within each cluster. For the specific data, Fowlkes et al. (1988) identified four of the environmental variables as being particularly important for clustering the 51 locations which were clustered into three groups.

Exhibit 24*a* shows a profile plot of the deviations of the means of the three clusters from the overall mean for each of the four variables, humidity (mg/cubic meter), terrain (measure of roughness in meters), temperature (average annual temperature in degrees Fahrenheit), and average annual number of days with thunderstorms. Since the variables are on very different units of measurement, the deviations for the profile plot are all standardized. From Exhibit 24*a* one would infer that Cluster #3 has the roughest terrain, the lowest annual temperature and the lowest humidity. Geographically, this made sense since the sites belonging to Cluster #3 belonged to Pennsylvania, New York, Wyoming, and New Mexico. Cluster #2, which consisted of sites in Florida and other southeastern states, shows up in the profile plot as one with the highest humidity, highest temperature and flat terrain.

Exhibit 24*b* is a scatter plot of the 51 locations, labeled by their cluster identity, in the space of two of the four variables, namely, terrain and

Exhibit 24a. Profile plot showing separation of clusters on individual standardized variables

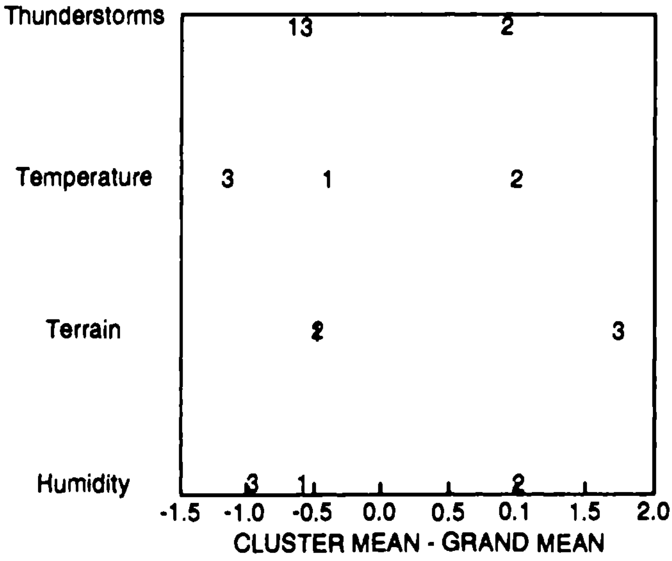
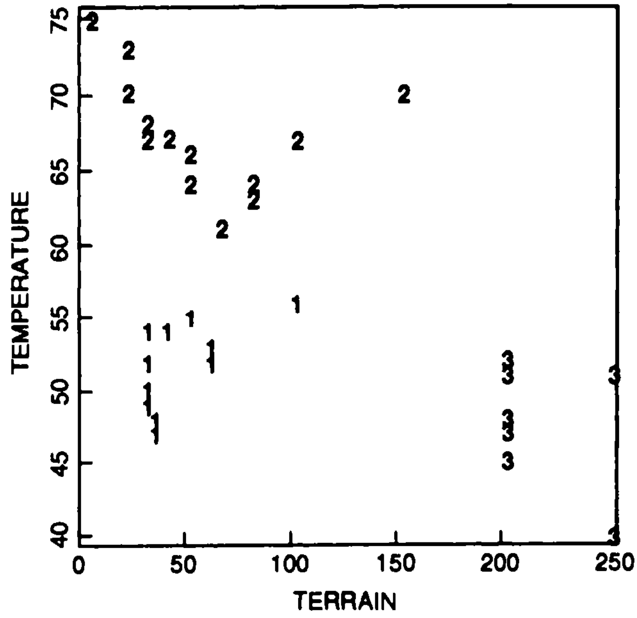


Exhibit 24b. Scatter plot of clusters in space of two initial variables



temperature. The quantized nature of the terrain measurement is evident, but also the separations among the clusters as well as the relative tightness of the three clusters can be seen in this scatter plot, at least in terms of two of the variables.

Getting a feel for the separations among clusters and the relative tightnesses of the different clusters in the space of all the variables used for the cluster analysis, if possible, is a common need in most applications of cluster analysis. With metric data, looking at distances among objects belonging to the same cluster and those among objects in different clusters can be useful for this purpose. The schematic displays in the four panels of Figure 9, taken from Gnanadesikan et al. (1977), show examples of plots of distances that can be helpful. For purposes of these displays it is assumed that a cluster analysis of metric data with $n = 5$ and $p = 2$ has yielded three clusters, labeled A, B, and C, with two objects in each of A and B, and the remaining single item belonging to C. Panel (*a*) of Figure 9 shows a plot of the squared distances of every object from each cluster centroid. In this artificial case, the three clusters appear to be well separated from each other with cluster A being particularly isolated and relatively tight. Panel (*b*) of Figure 9 is a different way of plotting

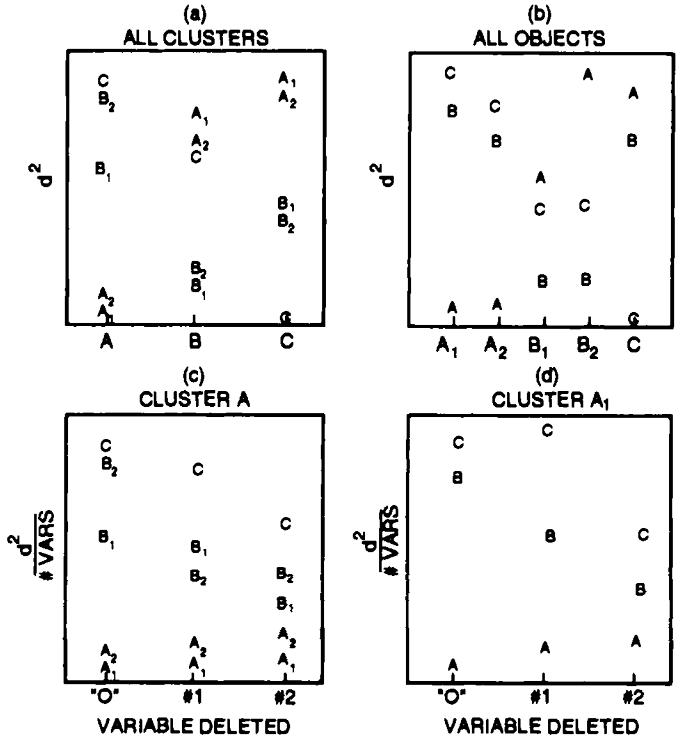


Fig. 9. Schematic plots of object-to-cluster distances. 3 clusters: $A = (A_1, A_2)$; $B = (B_1, B_2)$; C.

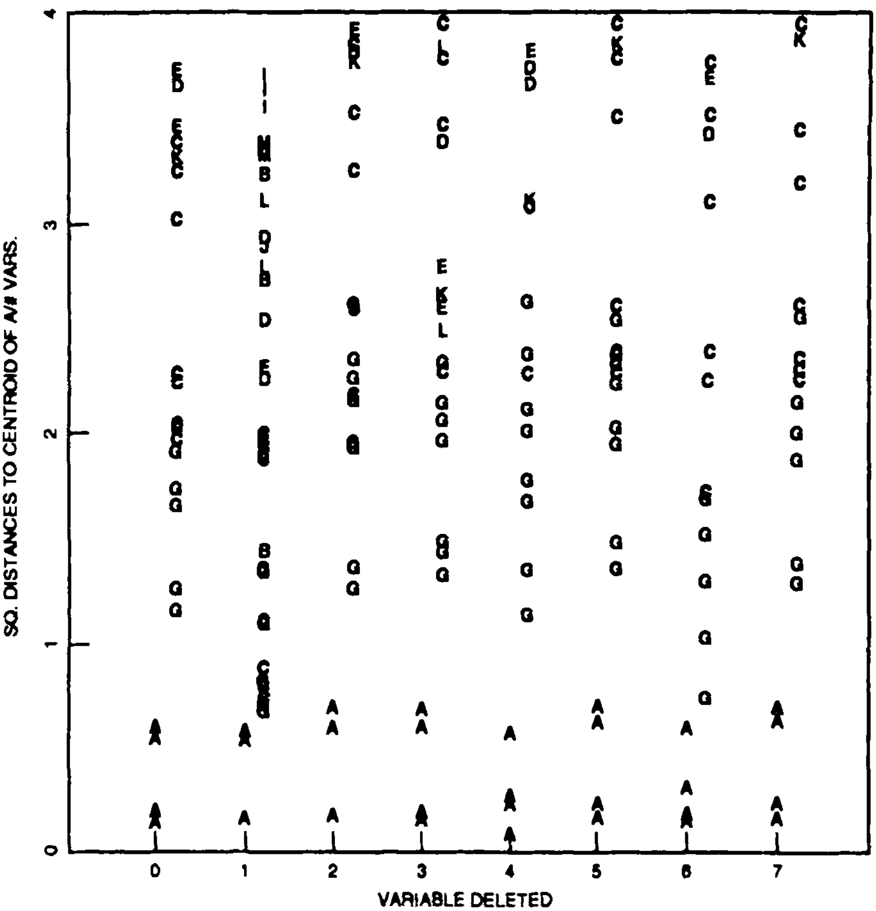
the same squared distances with the focus now being on objects and the strength of their classification into clusters. This plot shows how far away each cluster centroid is from each object.

Panels (*c*) and (*d*) of Figure 9 are plots of distances for help in judging the relative importance of each of the initial variables in determining the clusters. In these pictures, the ordinate is the ratio of the squared distance between object and cluster centroid to the number of variables used in computing the distance. Dividing by the number of variables used for the distance computation is a crude normalization that enabes comparisons across columns. The first column of panel (*c*) is exactly the same as the first column of panel (*a*) since no variables have been deleted. The remaining columns of panel (*c*) show the normalized squared distances of the five objects from the same cluster A after deleting each of the two variables in turn. The configuration indicates that variable #2 is important since its deletion decreases the separation of cluster A from the others. Panel (*d*), focused on the distances of the cluster centroids to object A1, is a detailed look at the relative importance of the variables in studying a specific column of panel (*b*) in the same way that panel (*c*) elucidates the same issue with respect to a specific column of panel (*a*).

Example 25. The example is from Gnanadesikan et al. (1977) and concerns the comparison of 48 subsidiaries of a single parent company. The subsidiaries operate in different environments and seven variables were defined to capture these environmental differences. Thus, in this example, $n = 48$ and $p = 7$. The 48 subsidiaries were clustered using the seven variables so that comparisons of subsidiaries with respect to their business performance could be made within clusters, that is, within similar operating environments. Exhibit 25 is a plot for this data similar to the schematic display in panel (*c*) of Figure 9 and it focuses on the effects of each of the seven variables on the formation of cluster A. The first column, labeled zero, shows the normalized squared distances of members of cluster A to the centroid of cluster A and also the distances of the other entities to the same centroid. The latter are shifted slightly to the right in the display to make them distinct from the A's. (*Note*: The numbers which identify the objects within clusters have not been shown as subscripts to avoid complicating the visual appearance.) The columns labeled 1–7 display the normalized squared distances obtained by deleting each of the seven variables one at a time. Since the gap between the A's and the other letters largely disappears when variable #1 or variable #6 is deleted, these two variables would seem to be providing cluster A with much of its distinctive character. In the context of this example, this made sense since cluster A was composed of subsidiaries in urban-industrial areas and variable #1 turned out to be a measure of urbanization while variable #6 was an indicator of industrialization.

While the simple method of plotting certain distances illustrated above can be useful for identifying individual variables that account for the clustering of

Exhibit 25. Plot of type in Fig. 9c to assess the effects of variables on cluster separations



the objects, its use for such things as looking at a large number of subsets of variables for picking out groups of variables that account for the clusters is clearly more complicated. The problem of variable selection for cluster analysis is an important one in its own right. A closely related problem is the one of weighting variables, for example in measuring inter-object distance, so as to reflect their differential importance for the clustering that might be present in the data. Modest beginnings for developing aids for variable selection and weighting have been made but a lot more needs to be done (see, for example, Fowlkes et al., 1987, 1988; Gnanadesikan et al., 1993; Gnanadesikan et al., 1995; and references therein).

Despite the fact that the groups are not prespecifiable in a cluster analysis situation, once they have been determined by an algorithm, one can use the clusters as *ipso facto* groups and use techniques from the known groups situation for informally studying such things as degree of separation among the clusters, their relative tightnesses and shapes. For instance, one can obtain the CRIMCOORDS (see Section 4.2) using the clusters as known groups and then display the objects in the space of the first two CRIMCOORDS. This and other projection plots for help in interpreting separations, tightness and shapes are discussed and illustrated by Gnanadesikan et al. (1982).

Another idea borrowed from the situation where the groups are prespecified is the one of using a "t-statistic" for informal guidance in assessing the separation between pairs of clusters, as suggested by Gnanadesikan et al. (1977). The following example illustrates what is involved.

Example 26. Kettenring et al. (1976) described a cluster analysis of $n = 452$ workers in terms of $p = 24$ variables (which themselves were composites of several directly measured variables) that pertain to the perceived needs for training to perform a certain complex job. One objective of the cluster analysis here was to group the workers into people with similar training needs. In addition to the 24 variables which formed the basis for a hierarchical clustering of the 452 workers, there were some exogenous variables, such as age, experience and educational level, of interest in the data.

In examining the hierarchical tree from the analysis, at some step of the tree formation wherein two branches (clusters) are being merged, one can compute a t-statistic for each of several variables to measure the separation between the branches with respect to the variables. The variables can be either ones on which the clustering is based or ones that are extraneous to the clustering algorithm but nevertheless useful for interpreting the clusters. To enable comparisons of branches across different levels of the tree, which might involve looking at t-statistics with possibly different "degrees of freedom," the p-value (i.e., the probability of exceeding the observed absolute value of the statistic) associated with the t-value can be computed and, as a further transformation, converted to logarithms. Large value of $|\log(p)|$ would suggest that the two branches (clusters) in question be considered as separate clusters rather than be merged into a single cluster. Variants of the t-statistic to accommodate differing variances in the clusters and extensions, including non-parametric and multivariate versions of statistics for measuring separations between locations of two groups, can also be tried. The use of the procedure based on the t-statistic (or any modification or extension) as a formal test of significance is, of course, questionable. Among other reasons for this, one that is important to keep in mind is that the two groups being compared are data-determined and not "independent random samples." An internal comparison of the relative magnitudes of the $|\log(p)|$-values can, however, be helpful in identifying worthwhile separations and in interpreting the differences among clusters. Bearing this in mind, the idea was applied to the training needs example and Exhibit 26 shows a table of values of $|\log(p)|$ for a few of the variables to enable a comparison of two specific clusters that merged at the last step of the tree formation. A comparison of the values indicates that the two clusters are far less different on the first variable than on the others used as the basis for

Exhibit 26. Statistics measuring cluster separation

VARIABLE	$\|LOG_{10}(p)\|$
1	0.8
2	6.9
3	7.0
⋮	⋮
22	6.8
23	7.0
24	6.8
AGE	0.5
EXPERIENCE	0.6
EDUCATION	2.1
⋮	⋮

clustering. Also, among the exogenous variables, the two clusters are far more different with respect to the educational levels of the workers belonging to them than with respect to their ages or experience.

If a hierarchical clustering algorithm has been used to generate clusters, a natural thing to do in identifying strong clusters is to look for well-defined branches in the tree. In terms of the output, one is thus tempted to look at the strengths associated with the steps at which clusters are formed and, more specifically, to look at the spacings between strengths, with large spacings in the early stages of the tree corresponding to strong clusters. Despite the obvious appeal of studying strengths and their spacings, the task is fairly difficult. One reason for this is that the configuration of strengths, as pointed out in the discussion at the end of Section 4.3.2a, depends on the particular method of clustering. Another reason is that there is no satisfactory statistical model of the "null" (i.e., case of no clusters) behavior of strengths against which one can study departures for assessing the strengths observed in clustering real data. Tests of significance and other approaches that are based on tightly specified distributional models for the null case, for instance, may not be appropriate for many real-life situations.

An informal graphical procedure, with a "nonparametric" flavor in that it avoids specifying a particular distribution as the null model for data, was proposed by Cohen et al. (1977). It too, however, involves a model that is less realistic in one of its aspects than one would like. The method is reasonably simple and seems informative enough in practice to warrant its use especially as an informal aid. The steps involved in it are described next.

First, the n observed values of each of the p variables are replaced by their ranks within that variable, thus changing the original data matrix to a new

$p \times n$ matrix each of whose rows contain integers from 1 to n. The transformation to ranks enables the use of a nonparametric approach. With no cluster structure in the data, and if the p variables are independent (this being the unrealistic assumption involved in the method), every permutation of the integers $1, 2, \ldots, n$ in a row is equally likely, and the rows are independent. This is the null model. Under non-null conditions, such as the existence of clusters, all permutations would not be equally likely. Therefore, for assessing the presence of real clusters in the data, the clusters from ranked data can be compared to clusters obtained from simulations of the null model. Specifically, Cohen et al. (1977) suggest an informal use of box plots in conjunction with such simulations, and the next example illustrates their proposal.

***Example** 27.* The portfolios data used in Example 20 constituted the starting point. The quarterly rates of return of each of the 52 investment portfolios were replaced by their ranks within each of the 11 quarters and the Manhattan metric was used to measure the distance between every pair of portfolios in terms of these ranks. Employing the interportfolio distances, a hierarchical clustering of the portfolios was carried out resulting in 51 strengths. Then 250 randomly permuted data ranks matrices were generated and each such matrix was used with the Manhattan metric again to obtain the input to the same clustering algorithm, leading to 51 new strengths each time.

Exhibit 27a. Comparison of strengths from clustering 52 portfolios against null distributions

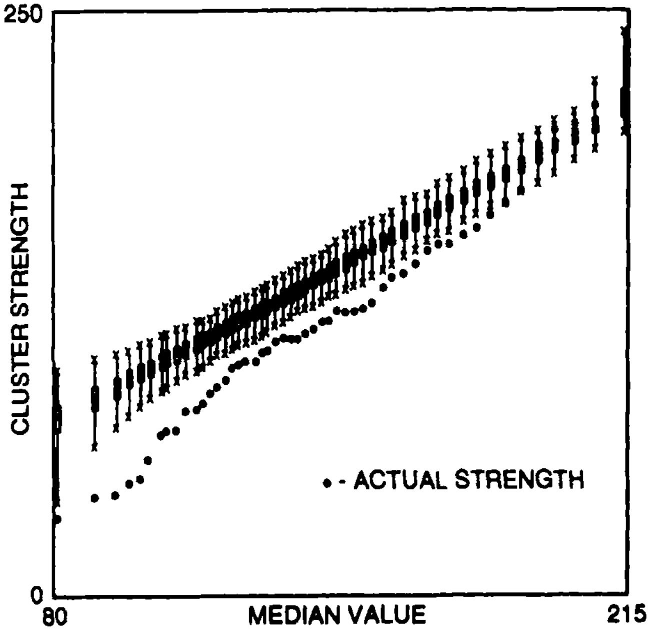


Exhibit 27*b*. Comparison of strengths of clustering 37 portfolios against null distributions

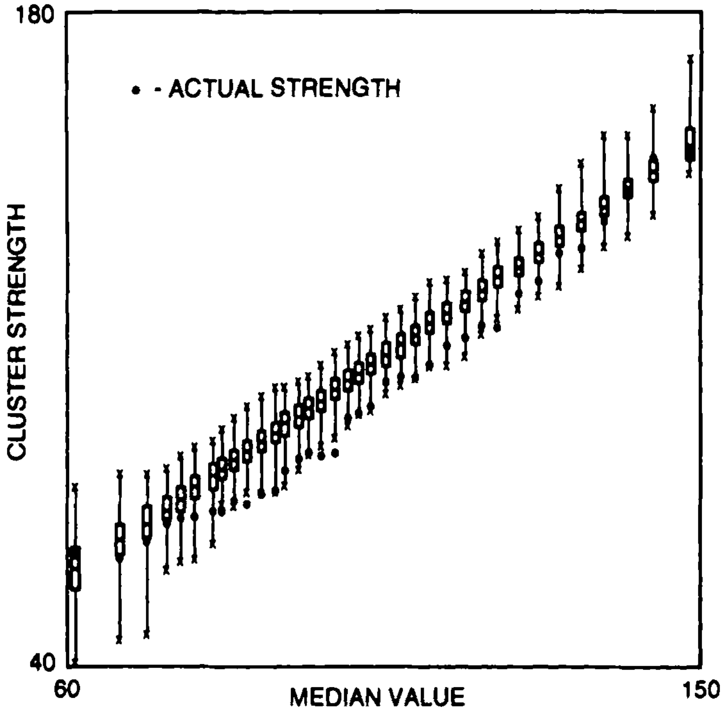


Exhibit 27*a* shows the box plots for displaying the distributions of the 250 values of each of the 51 strengths, there being 51 box plots in all. The boxes are placed along the *x*-axis so that their medians fall along the 45° line. Superimposed on the box plots are the strengths derived from the clustering of the actual data ranks matrix for the portfolios. These strengths, except for a few of the top end ones, are seen to be consistently lower than the "null regions" spanned by the box plots, thus suggesting smaller distances between certain portfolios than would be expected by chance.

Recalling the discussion in Example 20 for using the nearest-neighbors distances plot with the same data, there was evidence for a cluster of nine portfolios and four "outliers." Removing these portfolios (plus two additional possible "outliers" revealed by other analyses), and then repeating the above steps on the remaining 37 portfolios led to Exhibit 27*b*. The fact that the 36 strengths from the hierarchical clustering of the actual data now lie within the ranges of the box plots summarizing the distributions of the strengths from simulations of the null model suggests that these portfolios look more like a random sample from the null, "no cluster" situation.

Ideas based on permuting the similarities that constitute the input to a hierarchical clustering algorithm, rather than the above idea of permuting the ranks of the initial data values for each variable, for assessing the "significance" of clusters in a real data set have also been considered by several authors. (See, for example, Ling, 1973; Hubert, 1974; Baker & Hubert, 1975.)

A slightly different conceptualization of the problem of assessing the believability of clusters in a real data set is to cast it in terms of evaluating the relative stability of the clusters. Gnanadesikan et al. (1977), for instance, suggest "shaking the tree" by adding "noise" to the data, then clustering the resultant data, and finally comparing the trees for the original data and the perturbed data to determine which clusters, if any, remain unaltered. If the initial data are $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, then the proposal is to perturb the data by adding randomly generated values, $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$, to the initial data to obtain $\mathbf{z}_i = \mathbf{y}_i + \mathbf{e}_i$, $i = 1, \ldots, n$. Two possible choices for the distribution of the e's are: (i) a multivariate uniform (perhaps defined over the hypercube $[-c, +c]$ for specified c that can be varied for controlling the scale of the "noise"); and (ii) a multivariate normal (say with mean $\mathbf{0}$ and covariance matrix, $c\mathbf{V}$, where $\mathbf{V}$ could be chosen to be the covariance matrix of the initial data, and c is a specified constant which can be varied to reflect differing degrees of "noise"). [*Note*: When c is "small enough" presumably the data will only be perturbed slightly but as c increases the "noise" will become more dominant. In practice, what one would be interested in is to identify which clusters remain intact when c is neither too small to change any of the clusters nor so large that all the clusters fall apart.] For any specific value of c in either of the choices above for the distribution of the e's, what one would have are two trees that result from a hierarchical cluster analysis, one for the initial data and the other for the perturbed data. Comparing the two trees, to see how similar the clusters derived from them are, is a tedious if not impossible task. What is needed is a systematic scheme for comparing the two trees in terms of the contents of the clusters obtained by cutting the two trees at different levels to produce the same number of clusters from each tree.

A graphical method proposed by Fowlkes & Mallows (1983) is one tool for this purpose. The steps involved are the following:

1. Given two hierarchical trees, cut each of them to produce k clusters.
2. Form the $k \times k$ matrix, $\mathbf{M}$, whose (i, j)th element, m_{ij}, is the number of objects common to the ith cluster from the first tree and the jth cluster from the second tree; if the two sets of k clusters from the two trees are similar one would expect the diagonal elements (after suitably permuting columns in each row if necessary) of $\mathbf{M}$ to be large while the off-diagonal elements are close to zero.
3. Calculate a statistic, B_k, from the elements of $\mathbf{M}$ for measuring the similarity of the contents of the two sets of k clusters.
4. Repeat steps 1–3 for $k = 2, \ldots, (n - 1)$, where n is the total number of objects clustered.

5. Calculate the expected value, $\mathscr{E}(B_k)$, and the standard deviation, $S(B_k)$, of B_k under assumptions that capture the null case that the two sets of clusters are totally unrelated to each other.
6. Plot the observed values of B_k vs. k, with superimposed connected curves showing values of $\mathscr{E}(B_k)$ and $\mathscr{E}(B_k) \pm 2S(B_k)$; values of B_k that are large and clearly outside the bands indicated by the superimposed curves suggest great similarity of the two sets of clusters while values lying within the bands indicate lack of such similarity.

The statistic needed in step 3, as proposed by Fowlkes & Mallows (1983) is,

$$B_k = T_k/\sqrt{P_k Q_k},$$

where

$$T_k = \sum_{i=1}^{k} \sum_{j=1}^{k} m_{ij}^2 - n,\ P_k = \sum_{i=1}^{k} m_{i\cdot}^2 - n,\ Q_k = \sum_{j=1}^{k} m_{\cdot j}^2 - n,$$

$$m_{i\cdot} = \sum_{j=1}^{k} m_{ij} \qquad \text{and} \qquad m_{\cdot j} = \sum_{i=1}^{k} m_{ij}.$$

B_k lies between 0 and 1, taking on the value 0 if the two sets of clusters are completely different and the value 1 if they are the same. The statistic has the desirable property that it is invariant to the numbering of the clusters drawn from each tree.

The null model involved in step 5 is described by the two assumptions that the marginal totals, $m_{i\cdot}$ and $m_{\cdot j}$, are fixed for all i and j, and that subject to this the m_{ij}-values are random allocations across **M**. Under such assumptions,

$$\mathscr{E}(B_k) = \sqrt{P_k Q_k}/n(n-1),$$

and

$$\begin{aligned} S^2(B_k) = {} & 2/n(n-1) + 4P'_k Q'_k/n(n-1)(n-2)P_k Q_k \\ & + [(P_k - 2 - 4P'_k/P_k)(Q_k - 2 - 4Q'_k/Q_k)]/n(n-1)(n-2)(n-3) \\ & - P_k Q_k/n^2(n-1)^2, \end{aligned}$$

where

$$P'_k = \sum_{i=1}^{k} m_{i\cdot}(m_{i\cdot} - 1)(m_{i\cdot} - 2),\ Q'_k = \sum_{j=1}^{k} m_{\cdot j}(m_{\cdot j} - 1)(m_{\cdot j} - 2).$$

The technique of plotting B_k vs. k can be used in many contexts besides the one of studying the stability of clusters. Whenever one wishes to compare two hierarchical trees in their entirety with an eye to determining the degree of similarity of the clusters that results from cutting the two trees, this display might be helpful. Thus, comparing the effects of using two different metrics or measures of similarity, or of using two different hierarchical clustering algorithms, for the same data set are situations in which such a display can be used. Fowlkes & Mallows (1983) discuss the application of the method to a number of simulated and real data sets.

Exhibit 28a. Hierarchical tree for data simulating 4 clusters

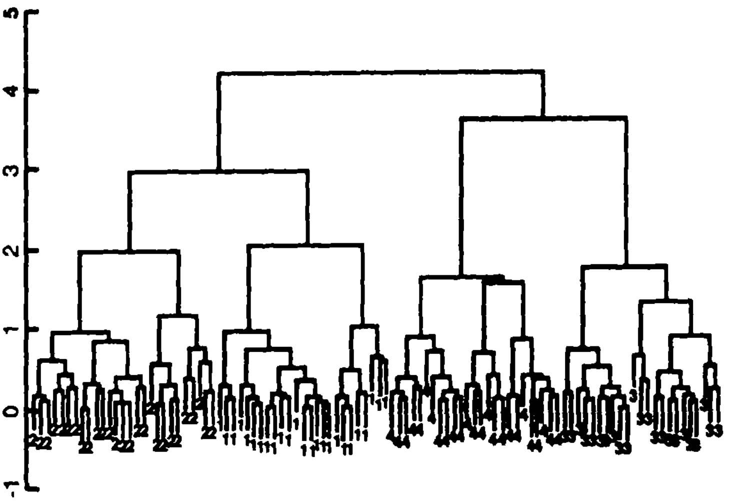
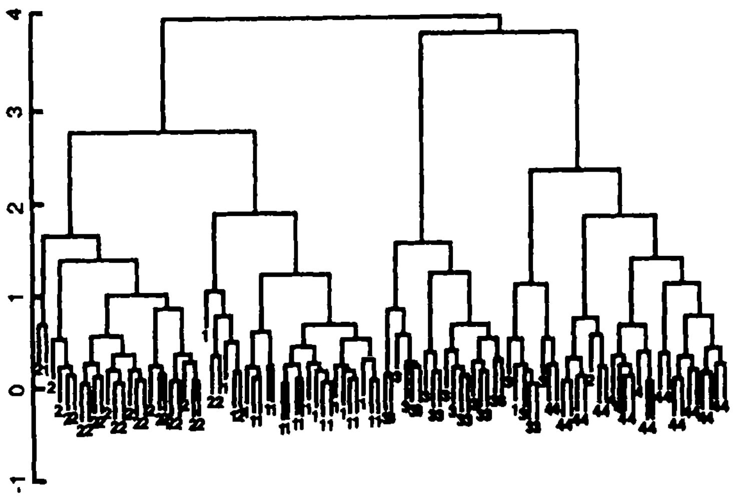


Exhibit 28b. Hierarchical tree for data of Exhibit 28a perturbed by adding random noise



Example 28. In this example, the use of the (k, B_k) plot for comparing a tree with a "shaken" tree obtained by perturbing the data is illustrated. The computer-generated data for the example simulated a random sample of 100 observations from a mixture of four spherical (i.e., with the identity matrix as the covariance matrix) multinormal distributions with different means. Thus, one would expect that there are four underlying clusters.

Exhibit 28*a* shows the tree resulting from the maximum method with Euclidean distances between pairs of points as input. The individual items are labeled at the bottom of the tree by their known cluster identification, and if the tree were cut to produce four clusters, the picture shows that a perfect recovery of all four clusters would result. The initial data were then perturbed by adding "noise" which consisted of a random sample from the multinormal distribution $N[0, 0.04 \times I]$, and the shaken tree that results from using these data is shown in Exhibit 28*b*. The items are again labeled by their known cluster identifications and it can be seen that if one were to cut this tree to produce four clusters, the original clusters are not recovered perfectly. Items originally in cluster #s 1 and 2, for example, are now mixed. However, it is difficult to get a feel for the similarity of the two trees and appreciate which set of clusters (produced by cutting the tree at different heights) are alike and which are not.

Exhibit 28c. Plot of B_k vs k for assessing the similarity of the trees in Exhibits 28*a* & *b*

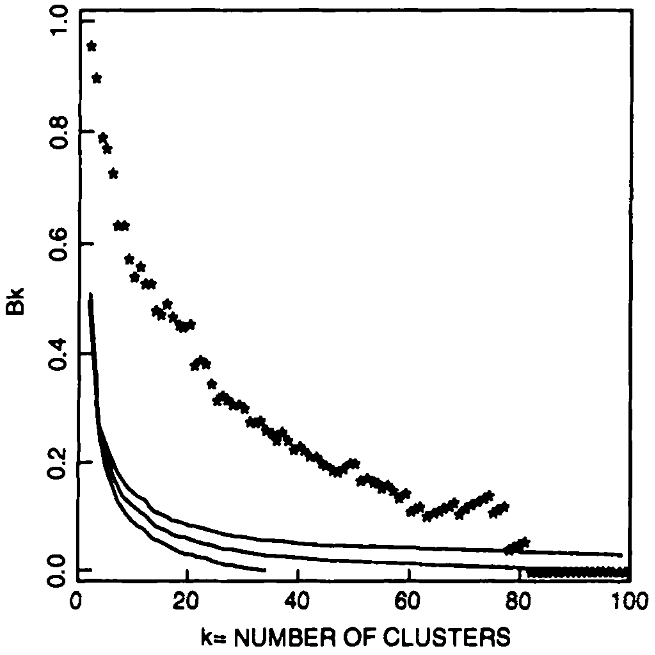


Exhibit 28*c* shows the (k, B_k) plot for the two trees. The configuration of the values of B_k relative to the "null" bands shows that the clusterings determined by the two trees are quite similar for small values of k and, in particular, for $k = 4$. Only at the lowest levels of the tree that correspond to large values of k (say, $k > 80$) are the two sets of clusterings very different. In this artificial data example, the small degree of perturbation has resulted in leaving most of the tree unchanged. Clearly, with larger amounts of noise added to the data, the clusters corresponding to smaller values of k will start to look less and less similar.

REFERENCES

Section 4.1 Breiman et al. (1984), Buja & Hastie (1994), Hand (1981).

Section 4.2 Becker et al. (1965), Bricker et al. (1971), Chambers (1977), Chen et al. (1970, 1974), Fisher (1936, 1938), Hotelling (1947), Jackson (1956), Rao (1952), Seal (1964).

Section 4.2.1 Anderson (1984), Becker et al. (1965), Chen et al. (1970, 1974), Rao (1952).

Section 4.2.2 Bricker et al. (1971).

Section 4.2.3 Burnaby (1966), Rao (1966).

Section 4.3 Cormack (1971), Everitt (1974), Gnanadesikan & Kettenring (1989), Hartigan (1975), Kaufman & Rousseeuw (1990).

Section 4.3.1 Art et al. (1982), Gnanadesikan et al. (1993), Gnanadesikan et al. (1995), Kaufman & Rousseeuw (1990).

Section 4.3.2 Arabie & Carroll (1980), Azimov et al. (1988), Ball & Hall (1965), Buja & Hurley (1990), Cohen et al. (1977), Cook et al. (1993), Fisherkeller et al. (1974), Friedman & Rubin (1967), Friedman & Tukey (1974), Gower (1967), Hartigan (1967, 1975), Johnson (1967), MacQueen (1965), Miller & Nicely (1955), Scott & Symons (1971), Shepard & Arabie (1979), Sneath (1957), Sokal & Sneath (1963), Sorenson (1948), Swayne et al. (1991), Warner (1968).

Section 4.3.3 Baker & Hubert (1975), Cohen et al. (1977), Fowlkes et al. (1987, 1988), Fowlkes & Mallows (1983), Gnanadesikan et al. (1977, 1982), Gnanadesikan et al. (1993), Gnanadesikan et al. (1995), Hubert (1974), Kettenring et al. (1976), Ling (1973).

CHAPTER 5

Assessment of Specific Aspects of Multivariate Statistical Models

5.1. GENERAL

The major portion of formal multivariate statistical theory has been directed toward the assessment of specific aspects of an assumed (and often unverified) mathematicostatistical model; that is, one assumes a model and then is concerned with formal statistical inferences about particular aspects of it. Examples include theories and methods of estimating multivariate parameters, and tests of hypotheses concerning location and/or dispersion parameters under either a multivariate normal or a more general model.

The assessment of statistical models is a legitimate and important concern of data analysis. Typically, however, the multivariate procedures for assessment have been developed by mathematical analogy with corresponding univariate methods, and often it is not clear that the complex aspects of the multivariate problem are incorporated in such an analogy or are better appreciated or otherwise benefit from it. For instance, more varied departures from a null hypothesis are possible in a multivariate situation than in the "analogous" univariate problem, and having a test per se of the null hypothesis against a completely general alternative is not of much value for multiresponse data analysis.

The standard results, pertaining to assessments of tightly posed questions in the framework of statistical models, are well organized and easily accessible in the multivariate literature (see, for example, Anderson, 1984; Puri & Sen, 1971; Rao, 1965; Roy, 1957). Therefore the present chapter provides no more than a cursory review of some of the standard "classical" results, whereas it concentrates on some later developments and covers them in greater detail (see Sections 5.2.3, 5.3, and 5.4).

5.2. ESTIMATION AND TESTS OF HYPOTHESES

The classical multivariate theory has been based largely on a multivariate normal distributional assumption. One consequence of this has been the concentration of almost all of the work on just location and dispersion parameters, with relatively little attention paid to questions of shape and other high-order characteristics.

5.2.1. Location and Regression Parameters

The so-called multivariate general linear model, mentioned toward the end of Chapter 3, has been the focus of much of the developments. Specifically, the model, which was defined earlier in Eqs. 49 and 50, is a simultaneous statement of p univariate general linear models:

$$\mathbf{Y}_j = \mathbf{X}\boldsymbol{\theta}_j + \boldsymbol{\varepsilon}_j \qquad \text{for } j = 1, \ldots, p, \tag{61}$$

where $\mathbf{Y}_j$ is the vector of n observations on the jth response, $\mathbf{X}$ is the $n \times k$ matrix of known values of design and/or regressor variables, $\boldsymbol{\theta}_j$ is a $k \times 1$ vector of unknown parameters (treatment effects or regression coefficients) associated with the jth response, and $\boldsymbol{\varepsilon}_j$ is a vector of n random errors associated with the observations on the jth response. The multivariate nature of the formulation is introduced by assuming that the p corresponding elements of the vectors $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_p$ are not necessarily statistically independent. The usual assumptions, in fact, are that each of the n sets of p elements has a mean (expected) value $\mathbf{0}$ and a common $p \times p$ covariance matrix $\boldsymbol{\Sigma}$, which is generally unknown and has to be estimated.

This general statement of the model subsumes both the multivariate multiple regression case and the more standard multivariate designed experiment situation, although it does not include any of the nonstandard multivariate experimental designs. The term "standard" is used to denote the case in which the design does not vary for the different responses; that is, one of the familiar univariate designs is used, but on each experimental unit p responses are observed simultaneously. In the multivariate multiple regression case it is generally assumed that the matrix $\mathbf{X}$ is of "full" rank k $(<n)$, whereas in the usual experimental design situations $\mathbf{X}$ will be a design matrix of rank r $(<k<n)$. In the latter case, however, one can reparametrize the problem in terms of a set of r parameters that are linear functions of the θ's and rewrite the general linear model in terms of the derived parameters and a new $n \times r$ design matrix of "full" rank r (see, for example, Scheffé, 1959, pp. 15–16). Hence, for present purposes, it is assumed that the matrix $\mathbf{X}$ is of full rank k so that $\mathbf{X}'\mathbf{X}$, in particular, will be a nonsingular matrix.

A useful device in the formal treatment of the estimation problem is the so-called rolled-out version of the model stated in Eq. 61 (see Section 4.a of Chapter III in Roy et al., 1971). By stringing out all the elements of the vector $\mathbf{Y}_1$, followed by those of $\mathbf{Y}_2$, and so on, one can obtain an np-dimensional column vector, $\mathbf{y}^*$, for which the following linear model is a consequence of

Eq. 61:

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}^*, \tag{62}$$

where (i) $\boldsymbol{\theta}^*$ and $\boldsymbol{\varepsilon}^*$ are, respectively, *kp*- and *np*-dimensional column vectors obtained by rolling out the $\boldsymbol{\theta}_j$'s and $\boldsymbol{\varepsilon}_j$'s of Eq. 61 in exactly the same manner as the $\mathbf{Y}_j$'s, (ii) the $np \times kp$ matrix, $\mathbf{X}^*$, can be written compactly as the Kronecker product $\mathbf{I}(p) \otimes \mathbf{X}$, and (iii) the covariance structure for the elements of $\boldsymbol{\varepsilon}^*$ is the $np \times np$ matrix, $\boldsymbol{\Sigma}^*$, which is the Kronecker product $\boldsymbol{\Sigma} \otimes \mathbf{I}(n)$. Then, if ξ is a linear function of the elements of $\boldsymbol{\theta}^*$, that is, $\xi = \mathbf{c}'\boldsymbol{\theta}^* = \mathbf{c}_1'\boldsymbol{\theta}_1 + \mathbf{c}_2'\boldsymbol{\theta}_2 + \cdots + \mathbf{c}_p'\boldsymbol{\theta}_p$, it can be established (see Section 4.a of Chapter III of Roy et al., 1971) that the minimum variance unbiased linear estimate of ξ is $\hat{\xi} = \mathbf{c}_1'\hat{\boldsymbol{\theta}}_1 + \mathbf{c}_2'\hat{\boldsymbol{\theta}}_2 + \cdots + \mathbf{c}_p'\hat{\boldsymbol{\theta}}_p$, where $\mathbf{c}_j'\hat{\boldsymbol{\theta}}_j$ is the least squares estimate of $\mathbf{c}_j'\boldsymbol{\theta}_j$, obtained by considering only the *j*th response and ignoring the rest of the variables. Specifically, the unknown covariance matrix $\boldsymbol{\Sigma}$ is not involved in computing $\hat{\xi}$, although the variance of $\hat{\xi}$ would involve $\boldsymbol{\Sigma}$.

This formal result, which is derived within the framework of linear models and the method of least squares, is perhaps one reason for the well-worn and widespread practice of constructing multivariate location estimates simply by assembling together univariate location estimates which have themselves been obtained by separate univariate analyses that ignore the multivariate nature of the data. In particular, as mentioned in the discussion following Eqs. 49 and 50 in Chapter 3, the classical result on linear estimation under the multivariate general linear model is that the estimate of $\boldsymbol{\Theta}$ is

$$\hat{\boldsymbol{\Theta}} = [\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2, \ldots, \hat{\boldsymbol{\theta}}_p]; \qquad \hat{\boldsymbol{\theta}}_j = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_j \qquad \text{for } j = 1, \ldots, p. \tag{63}$$

For the simple case of an unstructured (or single) sample, the matrix $\mathbf{X}$ would just be a column vector of 1's; $\boldsymbol{\Theta}$ would be the *p*-dimensional row vector, $\boldsymbol{\mu}'$, the unknown mean vector of the *p*-variate population from which the sample is presumed to be drawn; and Eq. 63 would yield as an estimate of $\boldsymbol{\Theta}$ the sample mean vector, $\bar{\mathbf{y}}'$, defined in Eq. 1 in Section 2.2.1. An interesting theoretical sidelight, the critical importance of which for analyzing multivariate data is unclear, is the result due to Stein (1956) on the inadmissibility of the sample mean vector as an estimator of $\boldsymbol{\mu}'$ when $p \geqslant 3$. From the viewpoint of data analysis, a far more significant objection to estimators, such as the sample mean vector, associated with the least squares criterion is their nonrobustness or susceptibility to the influence of a few outliers (see Section 5.2.3 for a discussion of robust estimators).

In regard to the problem of testing linear hypotheses concerning the parameters $\boldsymbol{\Theta}$, for the classical normal theory treatment of the multiresponse general linear model of Eq. 49 it is assumed that the rows of $\boldsymbol{\varepsilon}$ are *p*-variate normally distributed (or, equivalently, that in the rolled-out version given in Eq. 62 $\boldsymbol{\varepsilon}^*$ is *np*-variate normal with mean $\mathbf{0}$ and covariance structure $\boldsymbol{\Sigma}^*$).

Under this additional distributional assumption, the usual null hypothesis considered is

$$H_0\colon \mathbf{C\Theta U} = \mathbf{O}, \tag{64}$$

where the $s \times k$ matrix, $\mathbf{C}$ $(s \leqslant k)$, and the $p \times u$ matrix, $\mathbf{U}$ $(u \leqslant p \leqslant n - k)$, are matrices of specified constants with ranks s and u, respectively. For testing H_0 of Eq. 64 against the completely general alternative, $H_1\colon \mathbf{C\Theta U} \neq \mathbf{O}$, several statistics have been proposed in the literature (see, for example, Anderson, 1984; Roy et al., 1971). The test statistics are different functions of the eigenvalues of the matrix $\mathbf{S}_h\mathbf{S}_e^{-1}$, where

$$\left.\begin{aligned} \mathbf{S}_h &= \mathbf{U'YX(X'X)^{-1}C'[C(X'X)^{-1}C']^{-1}C(X'X)^{-1}X'Y'U}, \\ \text{and}\qquad & \\ \mathbf{S}_e &= \mathbf{U'Y[I}(n) - \mathbf{X(X'X)^{-1}X'}]\mathbf{Y'U}. \end{aligned}\right\} \tag{65}$$

For instance, if $\infty > c_1 \geqslant c_2 \geqslant \cdots \geqslant c_t > 0$ denote the t $[=\min(u, s)]$ ordered positive eigenvalues of $\mathbf{S}_h\mathbf{S}_e^{-1}$, the likelihood ratio test of H_0 is based on the statistic $\Lambda = \Pi_{j=1}^{t} 1/(1 + c_j)$, whereas the so-called largest-root test proposed by Roy (1953) is based on c_1 and the sum-of-the roots test on $\Sigma_{j=1}^{t} c_j$. Only when at least one of the quantities u and s equals 1 are the different tests entirely equivalent, and in situations where this condition does not apply the application of the different tests may indeed lead to different conclusions regarding the tenability of the null hypothesis. A more detailed discussion of tests of hypotheses (including not only the general null hypothesis of Eq. 64 but also more specialized hypotheses) and of formal power properties of such tests will be found in Chapters IV and V of Roy et al. (1971). (See also the work of Pillai and his students, for example, Pillai & Jayachandran, 1967, on power comparisons of the tests.) For present purposes it suffices to say that these tests of hypotheses tend to be of very limited value in multivariate data analysis, especially of an exploratory nature, when tightly specified models are either unavailable or unreasonable to commit oneself to.

A geometrical description may help to elucidate the concepts and methods associated with the general linear model stated in Eq. 61. One can think of the n observations on each of the p responses, $\mathbf{Y}_1, \ldots, \mathbf{Y}_p$, as the coordinates of p points, $P_1, P_2, \ldots, P_p$, in n-dimensional space (although, as stated in Chapter 1, the usual view is to consider the multiresponse observations as n points in p-space). Let O denote the origin in the n-space. Then the least squares estimate, $\hat{\boldsymbol{\theta}}_j$, from the univariate analysis of the jth response is known (see, for example, Scheffé, 1959) to be associated with the projection of P_j onto the k-dimensional subspace of n-space spanned by the k columns of $\mathbf{X}$, and the matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, which relates $\mathbf{Y}_j$ to its projection, is called the projection matrix involved in the operation of obtaining the least squares estimate. In the multivariate linear model which assumes a common design or regression structure, $\mathbf{X}$, for all the responses, $\hat{\boldsymbol{\Theta}}$ of Eq. 63 is then associated with the set of *p projections of* $P_1, \ldots, P_p$ onto the same space, and the same projection operator is applied to each of the p points.

In more specialized situations the matrix $\mathbf{X}$ may derive from an orthogonal design (e.g., a balanced multifactorial experiment), and then its columns will specify a decomposition of the n-space into mutually orthogonal linear subspaces, each associated with a meaningful source of variation (e.g., blocks, main effects, error) incorporated in the experimental design. In such situations a univariate analysis of variance of the observations on a single response, $\mathbf{Y}_j$, for instance, will yield, for the particular variable, a decomposition of the total sum of squares into sums of squares associated with each of the meaningful orthogonal sources of variation, and this process can be viewed geometrically as decomposing (á la Pythagoras' theorem) the squared length of the vector OP_j in terms of the squared lengths of the vectors joining O to the projections of P_j onto the different mutually orthogonal linear subspaces. Similarly, in orthogonal multivariate analyses of variance, the decomposition of the total sum-of-products matrices, associated with orthogonal sources of variation underlying the experiment, may be visualized as a decomposition of the squared lengths of and angles between the p vectors, $OP_1, OP_2, \ldots, OP_p$, in terms of the squared lengths of and the angles between the vectors joining O to the projections of $P_1, \ldots, P_p$ onto the different mutually orthogonal linear subspaces. Sum-of-products matrices associated with tests of hypotheses, such as $\mathbf{S}_h$ in Eq. 65, may be given similar interpretations in these geometrical terms.

In addition to point estimation and tests of hypotheses, under the normal theory various confidence regions and interval estimation schemes have also been proposed (see, for example, Chapter VI of Roy et al., 1971). For instance, a set of simultaneous confidence intervals can be obtained for bilinear functions of $\boldsymbol{\Theta}$, and these are multiresponse analogues of the univariate result due to Scheffé (1953) and to Roy & Bose (1953).

With regard to carrying out the computations involved in multiresponse multiple regression or analysis of variance, although the algebraic representations in some of the formulae have involved expressions such as inverses of certain matrices (e.g., Eq. 63), they are not intended to suggest that it would be appropriate to develop computational algorithms directly from them. In fact, the recent literature in statistical computing is rich in its emphasis on avoiding not only pitfalls in inverting matrices but also round-off errors in forming sum-of-products matrices. Approaches involving different types of matrix decompositions (e.g., singular value decompositions, Givens rotations) are the ones recommended currently, and Chambers (1977), Fowlkes & Lee (Appendix C in Roy et al., 1971), Golub & Reinsch (1970), and Wilkinson (1970) are a few of the relevant references.

5.2.2. Dispersion Parameters

A familiar example of estimation of dispersion is the estimation of the unknown covariance matrix Σ that specifies the error structure in the multiresponse general linear model stated in Eq. 49 or 62. An unbiased estimate of Σ is

$$\mathbf{S}_{\text{error}} = \frac{1}{n-k}\mathbf{Y}[\mathbf{I}(n) - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}', \tag{66}$$

while the (biased) maximum likelihood estimate is

$$\hat{\Sigma} = \frac{n-k}{n}\mathbf{S}_{\text{error}}.$$

With $p \leqslant (n-k)$, $\mathbf{S}_{\text{error}}$ will be nonsingular with probability 1.

In the special case of an unstructured sample, the expression in Eq. 66 simplifies to yield the sample covariance matrix S, defined earlier in Eq. 2 (see Section 2.2.1). Some (see Lindley, 1972) have worried about the inadmissibility issue à la Stein (1956, 1965) regarding such estimates of dispersion, but once again the nonrobustness of these estimates is perhaps more worrisome for data-analytic purposes than the theoretically fascinating issue of inadmissibility. Section 5.2.3 discusses some ideas regarding the robust estimation of dispersion parameters.

The literature pertaining to the standard normal theory treatment of multivariate analysis contains much material on formal statistical inference procedures (including estimation, tests of hypotheses, and distribution theory) associated with covariance matrices. The results on tests of hypotheses range from methods for testing the equality of two or more covariance matrices to procedures for testing hypotheses concerning specific structures for covariance matrices (e.g., sphericity, mutual independence of subsets of variables) and concerning eigenvalues and eigenvectors of covariance matrices. Chapters 9–11 and 13 (especially Chapter 10) of Anderson (1984) and various chapters of Rao (1965) contain many of the available results. Some results on confidence bounds for dispersion parameters are given by Roy & Gnanadesikan (1957, 1962). Methods associated with the study of structured covariance matrices have been discussed by several authors, including Srivastava (1966), Anderson (1969), and Jöreskog (1973).

5.2.3. Robust Estimation of Location and Dispersion

For the uniresponse situation, problems and methods of robust estimation have received considerable attention, especially over the recent two decades. The start of the thrust was the work of Tukey (1960). Huber (1964) provided the first theoretical framework for considering the uniresponse situation.

The initial focus was on obtaining estimates of uniresponse location that are resistant or less sensitive to outliers, and the distributions considered as alternatives to the normal (for purposes of modeling outliers and comparison of the relative behaviors of the proposed estimates) have almost always been symmetrical with heavier or longer tails than the normal. A useful analogy for robustness has been the one with insurance: if one has an unforeseen accident (which, with data, would be unanticipated outliers) then the insurance is intended to pay off but, on the other hand, if one is fortunate to have no accident (which would correspond to data being "well behaved", i.e., data conform to the usual assumptions underlying the classical statistical methods including behaving like a random sample from a normal distribution) then the insurance premium should not be prohibitively high. The "premium" involved has been measured in terms of the efficiency of the estimator, that is, one would like the estimator to be highly efficient in the presence of outliers and yet be only moderately less efficient than the optimal estimator for the normal case. The uniresponse location problem was studied extensively and reported on by Andrews et al. (1972). (See also Huber, 1972.)

The problem of robust estimation of scale, on the other hand, has received less attention (see, for example, Lax, 1975). Considerable attention has also been given to the problem of robust estimation of uniresponse multiple regression parameters (see, for example, Huber, 1973; Mallows, 1983; Chapter 6 of Hampel et al., 1986, and the references therein). Indeed the literature on robust estimation for the uniresponse case continues to grow steadily, especially focused on many subtle and esoteric theoretical aspects. Hampel et al. (1986) provide a comprehensive state-of-the-art account of the field.

Two fundamental and useful formulations that have arisen from the theoretical work in robust estimation are the concepts of the so-called *influence function* and the *breakdown point*, both proposed by Hampel (1971, 1974). The influence function is useful in that it provides a measure of the influence of an observation on an estimator. The observation can be either an actual one (see Section 6.4.2) or a conceptualized one, and, in the latter case, one can define an influence function with respect to a parameter in addition to the one with respect to an estimator of that parameter. The influence function has proven to be useful not only in understanding the sensitivity of existent estimators (e.g., the sample mean and variance, median) but, more importantly, it has led to designing new robust estimators. For example, it has played a central role in formalizing estimators on which outlying observations have *bounded influence* (thus providing resistance to such outliers), or even "zero" influence if the outliers are truly extreme, leading to the class of estimators with so-called *redescending influence* functions. The concept of the breakdown point of an estimator is essentially the largest fraction of observations in a sample that can be arbitrarily bad (i.e., be extreme outliers) without distorting the value of the estimator. Thus, a highly resistant estimator would have a high breakdown point. As an estimator of uniresponse location, the sample mean, for instance, has a breakdown point of zero, whereas the sample median has a breakdown value of 0.5. While the influence function and the breakdown point are useful aids for designing and assessing estimators, in practice one should be wary of emphasizing these to the exclusion of other important considerations such as the computational difficulty associated with calculating an estimator from data. There are also questions, especially with moderate sized data sets, of the meaningfulness of emphasizing high breakdown values (e.g., 0.5): What does it mean to estimate some parameter when half the sample consists of outliers?

Gnanadesikan & Kettenring (1972) have discussed some issues of and techniques for the robust estimation of multiresponse location and dispersion, and this section is based largely on their discussion and proposals. Although these methods are useful in protecting data summaries against certain kinds of outliers, it should be emphasized that the variety, both in kind and in effect, of outliers in multiresponse data can indeed be large, and the routine use of any robust estimate without exploring the data for the existence of specific peculiarities in them is neither wise nor necessary. (See Chapter 6 for a discussion of additional techniques for facilitating the exposure of peculiarities in data.)

The usual (nonrobust) estimate of multivariate location is the sample mean vector, $\bar{\mathbf{y}}$, whose elements are just the uniresponse means. The initial approaches (see Mood, 1941; Bickel, 1964) to the problem of robust estimation of multivariate location consisted of looking at just vectors of univariate robust estimators by analyzing the observations on each of the response variables separately. Gentleman (1985) was the first to propose a robust estimate of multivariate location which involved the simultaneous manipulation of all response variables. Gnanadesikan & Kettenring (1972) proposed a class of ad hoc procedures for robust estimation of multivariate location and dispersion, that are much in the spirit of the initial attack on the univariate problems of location and scale by Tukey (1960). The work of Maronna (1976) cast the multivariate problems in the general theoretical framework of *m*-estimates (see also Huber, 1981). While drawing heavily from Gnanadesikan & Kettenring (1972), the results on *m*-estimates are also included in this section for completeness.

Some possibilities for robust estimators of multivariate location that are simply vectors of univariate robust estimators are the following:

1. $\mathbf{y}_M^*$, the vector of medians of the observations on each response, as suggested by Mood (1941).

2. $\mathbf{y}_{HL}^*$, the vector of Hodges-Lehmann estimators (i.e., the median of averages of pairs of observations) for each response, as proposed and investigated by Bickel (1964).

3. $\mathbf{y}_{T(\alpha)}^*$, the vector of α-trimmed means (i.e., the mean of the data remaining after omitting a proportion, α, of the smallest and of the largest observations) for each response, as considered by Gnanadesikan & Kettenring (1972), or in a similar vein $\mathbf{y}_{W(\alpha)}^*$, the vector of α-Winsorized means.

4. $\mathbf{y}_m^*$, a vector of any of the so-called m-estimates of univariate location for each response. An m-estimate of location for the jth response, when the scale is unknown, is generally defined as the solution, T_j, of the equation

$$\sum_{i=1}^{n} \psi\left(\frac{y_{ij} - T_j}{s_j}\right) = 0,$$

where y_{ij} is the ith observation on the jth response ($i = 1, \ldots, n, j = 1, \ldots, p$), and s_j is a simple robust estimate of scale of the jth respose such as the median absolute deviation (MAD) of the observations from the median. Two widely used choices for the function, ψ, are

$$\text{(a)} \qquad \psi(u) = \begin{cases} -k, & \text{if } u < -k, \\ u, & \text{if } |u| \leqslant k, \\ +k, & \text{if } u > +k, \end{cases} \qquad \text{with } k = 1.5,$$

leading to the so-called Huber m-estimate; and

$$\text{(b)} \qquad \psi(u) = \begin{cases} 0, & \text{if } u < -c, \\ (u/c)[1 - (u/c)], & \text{if } |u| \leqslant c, \\ 0, & \text{if } u > c, \end{cases}$$

with values of c in the range $[6, 9]$, leading to the so-called *bisquare* or *biweight* estimates proposed by Tukey. Calculation of these m-estimates involves iterative computations and, in fact, the process entails iteratively weighted least squares with weights that are data dependent and change from iteration to iteration. As such, m-estimates of location can be written in the form, $\Sigma_{i=1}^{n} w_{(i)} y_{(i)j} / \Sigma_{i=1}^{n} w_{(i)}$, where $y_{(1)j} \leqslant y_{(2)j} \leqslant \cdots \leqslant y_{(n)j}$ denote the ordered observations and the iteratively determined weights, $w_{(i)}$, decreases as $|y_{(i)j} - T_j|$ increases. Qualitatively, this behavior of the weights explains why the resultant estimate of location will be robust since extreme observations will be given far less weight than ones in the "middle" of the data. For starting the iterative computations involved in m-estimates of location such as those above, the common practice is to use the median in the first step.

5. More generally, a vector each of whose elements is any univariate robust estimator (see Andrews et al., 1972, for a variety of possibilities) of the location for a single response variable.

Unlike the above estimators, each of which is just a collation into vector form of univariate estimators, the estimator proposed by Gentleman (1965) is multivariate in character in that the analysis involves a combined manipulation of the observations on the different responses. The essential idea is to choose the estimator, $\mathbf{y}^* = (y_1^*, y_2^*, \ldots, y_p^*)'$, of $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)'$ so as to minimize the criterion, $\Sigma_{j=1}^{p} |y_j^* - \mu_j|^k$, for a specified value of k in the range $1 \leqslant k \leqslant 2$. For $k = 2$ the estimator is the usual sample mean vector, while for $k < 2$ one obtains an estimator less sensitive to possible outliers. For a general value of k between 1 and 2, a closed-form expression of the estimator is not available, but Gentleman (1965) describes a numerical logorithm which can be used to compute the estimator for any set of multiresponse data. Gentleman also discusses some issues of modifying the criterion of kth power deviations to reflect the existence, if any, of both differences in the scales of the responses and intercorrelations among the responses.

Another approach to specifying a location estimate which also involves considering the responses simultaneously is discussed briefly in Section 6.2 in conection with the uses of a technique for plotting high-dimensional data.

A theoretical issue, which has been raised by Bickel (1964) in connection with $\mathbf{y}_M^*$ and $\mathbf{y}_{HL}^*$ but, in fact, applies to all of the estimators mentioned above, is the lack of affine commutativity of the robust estimators of multivariate location, in contrast to the usual mean vector $\bar{\mathbf{y}}$, which does possess this property. (A location estimator is affine commutative if the operations of affine transformation and formation of the estimate can be interchanged without affecting the outcome.) From a practical viewpoint the issue may be viewed in terms of commitment to the coordinate system for the observations. At one extreme the interest may be confined entirely to the observed variables, and, if so, any issue of commutativity will perhaps be remote. At the other extreme one may feel that the location problem is intrinsically affine commutative (e.g., one may wish to require that when one works with metric and nonmetric scales the effect of transforming from one scale to the other and then computing the robust location estimate be the same as directly transforming the robust estimate on the former scale) and insist that all location estimators have this property. As an intermediate position one may seek more limited commutativity (e.g., just linear transformations of each variable separately as in the above metric-nonmetric example, or just orthogonal transformations) than the very general affine commutativity.

In regard to the robust estimation of multivariate dispersion, there are at least two aspects of the problem, namely, the facet that depends on the scales (i.e., variances) of the responses and the one that is concerned with orientation (i.e., intercorrelations among the responses). For some purposes it may be desirable to consider the robust estimation of each of these aspects separately, whereas for other purposes a combined view may be in order.

An approach that separates the two aspects has the advantage of using all the available and relevant information for each estimation task, whereas an approach that combines the two will involve retaining only observations (perhaps fewer in number) which pertain to both aspects simultaneously. On the other hand, in many cases the combined approach may be computationally simpler and more economical.

The problem of robust estimation of the variance of a univariate population has been considered (see Tukey, 1960; Johnson & Leone, 1964, Section 6.9;

Hampel, 1968; Lax, 1975), although not as intensively or extensively as the location case. When one leaves the location case, certain conflicting aims seem to emerge in estimating higher-order characteristics of a distribution. Thus for the variance (and maybe even more so for the shape) there is a possible conflict between the desire to protect the estimate from outliers and the fact that the information for estimating the variance relies more heavily on the tails.

This conflict raises certain questions about the routine use of robust estimation procedures for these higher-order characteristics, especially in relatively small samples. Thus, with a sample of size 10, for instance, the use of a 10% (the minimum possible in this case) trimmed sample to provide a robust estimate may lead to an estimator whose efficiency is unacceptably and unnecessarily low when the data are reasonably well behaved. The main point is that, with relatively small, and yet reasonable, samples sizes, it may be both expedient and wise to study the observations more closely, omitting only clearly indicated outliers (see Chapter 6 for outlier-detection methods) or possibly transforming the observations to make them more nearly normally distributed (see Section 5.3).

The usual unbiased estimator of the variance for the jth response ($j = 1, \ldots, p$) based on n observations may be denoted as s_{jj}, and a corresponding robust estimator, s_{jj}^*, may be developed by any of the following three methods:

1. The square of the median absolute deviation, $(\text{MAD})^2$, of the observations from the median for the jth response.
2. Trimmed variance from an α-trimmed sample, as suggested by Tukey (1960) and further studied by Hampel (1968).
3. Winsorized variance from an α-Winsorized sample, as sugested by Tukey & McLaughlin (1963).
4. The slope of the lower end of a $\chi^2_{(1)}$ probability plot (see Section 6.2 for a brief discussion of probability plots) of the $n(n-1)/2$ squared differences between pairs of observations.

The second and third methods need an estimate of location, and a direct suggestion would be to use a trimmed mean for the trimmed variance and a Winsorized mean for the Winsorized variance. Huber (1970), however, suggests using a trimmed mean for getting the Winsorized variance, and for t-statistic types of considerations associated with the trimmed mean this may be appropriate. But even for applying a trimmed mean for the trimmed variance, or a Winsorized mean for the Winsorized variance, because of the considerations mentioned above it would seem advisable to use a smaller proportion of trimming (or Winsorizing) for the variance estimation than for the location estimation in samples even as large as 20.

To obtain unbiased, or even consistent, estimates from $(\text{MAD})^2$, or a trimmed or Winsorized variance, multiplicative constants are needed. These constants are based on an underlying assumption that the "middle" of the sample is sufficiently normal, and their values derive from the moments of the order statistics of the normal distribution. The required adjusted estimate of variance based on the median absolute deviation is $(\mathrm{MAD}/.675)^2$. Johnson and Leone (1964, p. 173) give a table of the required constants for small ($n \leqslant 15$) samples, and tables provided by McLaughlin and Tukey (1961), together with the tabulation by Teichroew (1956) of the expected values of cross products and squares of normal order statistics, may be used for calculating the required constant for samples of sizes up to 20. Unfortunately, asymptotic results do not appear to be adequate at $n = 20$, and further work is needed on developing the required multiplicative constant for larger values of n.

One advantage of the third method mentioned above is that it does not involve an estimate of location. A second is that the type of adjustment provided by the multiplicative constant in the trimmed and Winsorized variances is contained in the probability plot itself—namely, the abscissa (or quantile axis) is used to scale the ordinate for determining the slope (which will be an estimate of twice the variance). A third advantage is that, by looking at $n(n-1)/2$ pieces of information (some of which may be redundant because of statistical correlations), the error configuration on the $\chi^2_{(1)}$ probability plot may often be indicated more stably than on a normal probability plot of the n observations. A fourth, and perhaps the most significant, advantage of the approach is its exposure value in facilitating the detection of unanticipated peculiarities in the data. On the negative side a disadvantage of the technique is that it may not be useful, and may even be misleading, for estimating the variance in circumstances where a large proportion of the observations may be outliers.

The multivariate nature of dispersion is introduced inevitably by considering the estimation of covariance and correlation. A simple idea for estimating the covariance between two variables, Y_1 and Y_2, is based on the identity

$$\operatorname{cov}(Y_1, Y_2) = \tfrac{1}{4}\{\operatorname{var}(Y_1 + Y_2) - \operatorname{var}(Y_1 - Y_2)\}. \tag{67}$$

One robust estimator, s^*_{12}, of the covariance between Y_1 and Y_2 may, therefore, be obtained from

$$s^*_{12} = \tfrac{1}{4}\{\hat{\sigma}^{*2}_1 - \hat{\sigma}^{*2}_2\}, \tag{68}$$

where $\hat{\sigma}^{*2}_1$ and $\hat{\sigma}^{*2}_2$ are robust estimators of the variances of $(Y_1 + Y_2)$ and $(Y_1 - Y_2)$, respectively, and may be obtained by any of the methods mentioned above.

When such a robust estimator of the covariance is available, a natural way of defining a corresponding robust estimator of the correlation coefficient between Y_1 and Y_2 is

$$r^*_{12} = \frac{s^*_{12}}{\{s^*_{11} s^*_{22}\}^{1/2}}, \tag{69}$$

where s^*_{jj} is a robust estimator of the variance of the jth response.

Since the robust estimators involved in Eqs. 68 and 69 are determined with no considerations of satisfying the well-known Cauchy–Schwarz inequality relationship between the covariance and the variances, therefore, r_{12}^* as obtained from Eq. 69 may not necessarily lie in the admissible range, $[-1, +1]$, for a correlation coefficient. To ensure an estimate of the correlation coefficient in the valid range, while still retaining the above approach of obtaining the covariance estimate as the difference between two variance estimates, a modification may be suggested. Let $Z_j = Y_j/\sqrt{s_{jj}^*}$ denote the "standardized" form of Y_j, where s_{jj}^* is a robust estimate of the variance of Y_j. Then define

$$\hat{\rho}_{12}^* = \frac{\hat{\sigma}_3^{*2} - \hat{\sigma}_4^{*2}}{\hat{\sigma}_3^{*2} + \hat{\sigma}_4^{*2}}, \tag{70}$$

where now $\hat{\sigma}_3^{*2}$ and $\hat{\sigma}_4^{*2}$ are robust estimators of the variances of $(Z_1 + Z_2)$ and $(Z_1 - Z_2)$, respectively. One can use any robust estimate of the variances of the standardized sum and difference, but Devlin et al. (1975) have studied the use of trimmed variances in particular and have denoted by r^*(SSD) the associated $\hat{\rho}_{12}^*$. Corresponding to $\hat{\rho}_{12}^*$, which necessarily lies in the range $[-1, +1]$, a covariance estimator may be defined by

$$\hat{\sigma}_{12}^* = \hat{\rho}_{12}^*\{s_{11}^* s_{22}^*\}^{1/2}. \tag{71}$$

An interesting consequence of estimating the correlation coefficient by Eq. 69 or 70 is that the multiplicative constant, which is required for removing the biases involved in trimmed or Winsorized variances, cancels out by appearing in both the numerator and the denominator of the defining equations 69 and 70. Hence, for any sample size, the trimmed (or Winsorized) variances that provide the bases for obtaining r_{12}^* and $\hat{\rho}_{12}^*$ can be used directly without any multiplicative constant. This does not, however, imply that r_{12}^* and $\hat{\rho}_{12}^*$ are unbiased estimators of the population correlation. In fact, just as the usual product moment correlation coefficient is biased, these robust estimates are biased in small (but not large) samples. Devlin et al. (1975) have studied the biases and efficiencies of the above-mentioned as well as other robust estimators of correlation, including some well-known nonparametric estimators such as Kendall's τ.

A full-fledged consideration of multiresponse dispersion would be necessary if one were interested in the estimation of not just a single covariance or correlation but a collection of these, say a covariance or a correlation matrix. The usual estimates of the covariance and correlation matrices are, respectively, the sample covariance matrix, **S**, and the associated sample correlation matrix, **R** (see defining Eqs. 2 and 3 in Section 2.2.1). When robust estimates of the variances and covariances have been obtained by the methods discussed above, a direct method of obtaining a robust estimate of the covariance matrix is just to "put these together" in a matrix. Thus, corresponding to each of the two methods described above (see Eqs. 69 and 70) for obtaining an estimate of the correlation coefficient, a robust estimate of the covariance matrix would be

$$\mathbf{S}_a^* = \mathbf{D}\mathbf{R}_a^*\mathbf{D} \qquad \text{for } a = 1, 2. \tag{72}$$

where $\mathbf{D}$ is a diagonal matrix with diagonal elements $\sqrt{s_{jj}^*}$ $(j = 1, \ldots, p)$, $\mathbf{R}_1^* = ((r_{jj'}^*))$, and $\mathbf{R}_2^* = ((\hat{\rho}_{jj'}^*))$.

For some purposes of analyzing the multiresponse data, when the underlying distribution is not singular, it may be desirable to have a positive definite estimate of the covariance matrix. For instance, in analyzing the configuration of the sample in terms of the generalized squared distances of the observations from the sample centroid (see Example 7 in Section 2.4), the inverse of the estimate of the covariance matrix is used.

If the dimensionality, p, does not exceed the number of independent observations [$(n - 1)$ in the case of an unstructured sample], the usual estimator, S, is positive definite with probability 1. However, neither of the estimators, $\mathbf{S}_1^*$ and $\mathbf{S}_2^*$, defined above is necessarily positive definite. The positive definiteness of these estimators is equivalent to the positive definiteness of the corresponding estimators, $\mathbf{R}_1^*$ and $\mathbf{R}_2^*$, of the correlation matrix, and even though each off-diagonal element of $\mathbf{R}_2^*$ necessarily lies in the range $[-1, +1]$, this does not necessarily imply positive definiteness of $\mathbf{R}_2^*$, except for the bivariate case. [*Note*: Positive definiteness of a correlation matrix may be conceptualized as a high-dimensional analogue of the property that a single correlation coefficient lies between -1 and $+1$, and the constraint of positive definiteness seems to introduce the need to consider all the responses simultaneously with respect to their dispersion or orientational summary, although superficially such a summary might appear to be based only on a pairwise consideration of the responses.] Devlin et al. (1975) suggest a way of modifying $\mathbf{R}_2^*$ when $p > 2$ so as to obtain a positive definite estimate of the correlation matrix, which can then be employed in Eq. 72 to obtain a positive definite estimate of the covariance matrix. The essential idea is to "shrink" each of the estimates, $r_{jj'}^*$(SSD), of bivariate correlation sufficiently to ensure positive definiteness. It would be desirable for the shrinking scheme to decrease the high (in absolute value) correlations only slightly and the low correlations more drastically. Thus, the objective is to shrink $\mathbf{R}_2^*$ nonlinearly toward the identity matrix, I. Dropping the identification, SSD, and denoting the jj'th element of $\mathbf{R}_2^*$ as $r_{jj'}^*$, Devlin et al. (1975) propose the following specific nonlinear shrinking scheme: replace $r_{jj'}^*$ by

$$g(r_{jj'}^*) = \begin{cases} z^{-1}[z(r_{jj'}^*) + \Delta], & \text{if } r_{jj'}^* < -z(\Delta), \\ 0, & \text{if } |r_{jj'}^*| \leqslant z(\Delta), \\ z^{-1}[z(r_{jj'}^*) - \Delta], & \text{if } r_{jj'}^* > z(\Delta), \end{cases} \tag{72a}$$

where $z = \tanh^{-1}(r^*_{jj'})$ is Fisher's z-transform and Δ is a prespecified small positive constant (e.g., $\Delta = 0.05$). The matrix of resulting correlations is checked for positive definiteness and the process is repeated until positive definiteness is achieved. The variance stabilizing z-transform puts the correlations on roughly the same footing before the Δ-shift is applied, and hence has the desirable effect of nonlinear shrinkage of the correlations themselves. The resulting positive definite robust estimate of the correlation matrix can be suggestively denoted as $\mathbf{R}^*_2(+)$ or $\mathbf{R}^*_+(\text{SSD})$.

An entirely different context, of some practical interest for using this scheme of shrinkage, is one with incomplete multivariate observations, that is, where not all p variables are measured on each of the n units. Such missing data are particularly likely in very large data sets. Many archaeological and paleontological data sets seem prone to incomplete observations as a consequence of uncontrollable factors. One idea for developing a positive definite estimate of the overall correlation matrix from such data would be to use all of the observations available for every pair of variables to calculate bivariate correlations first, and then to adjust these by the shrinking scheme in Eq. 72a.

Gnanadesikan & Kettenring (1972) tentatively proposed some other methods for obtaining positive definite robust estimators of covariance (and thence correlation) matrices, and these are described next. The essential idea underlying all of them is to base the estimate on a "sufficiently large" number, v, of the observations (i.e., $v > p$), which are, nevertheless, subselected from the total sample so as to make the estimate robust to outliers. A second feature of these estimators is that they are based on a combined consideration of both scale and orientational aspects, unlike $\mathbf{S}^*_1$ and $\mathbf{S}^*_2$, which were built up from separate considerations of these aspects.

The first method for ensuring a nonsingular robust estimator of the covariance matrix is based on an approach suggeted by Wilk et al. (1962), who were concerned with developing appropriate compounding matrices for a squared distance function employed in an internal comparisons technique suggested by Wilk & Gnanadesikan (1964) for analyzing a collection of single-degree-of-freedom contrast vectors (see Section 6.3.1). The first step in the procedure is to rank the multiresponse observations, $\mathbf{y}_i$ $(i = 1, \ldots, n)$, in terms of their Euclidean distance from some robust estimate of location, $\mathbf{y}^*$, that is, $\|\mathbf{y}_i - \mathbf{y}^*\|$ [or, equivalently, the squared Euclidean distance, $(\mathbf{y}_i - \mathbf{y}^*)'(\mathbf{y}_i - \mathbf{y}^*)$]. Next, a subset of the observations whose ranks are the smallest $100(1 - \alpha)\%$ is chosen and used for computing a sum-of-products matrix,

$$\mathbf{A}_0 = \sum_{\substack{l \in \text{chosen subset} \\ \text{of observations}}} (\mathbf{y}_l - \mathbf{y}^*)(\mathbf{y}_l - \mathbf{y}^*)'. \tag{73}$$

(The fraction α of the observations not included in $\mathbf{A}_0$ is assumed to be small enough to ensure that $\mathbf{A}_0$ is positive definite.) After all n observations have been ranked in terms of the values of the quadratic form,

$(\mathbf{y}_i - \mathbf{y}^*)'\mathbf{A}_0^{-1}(\mathbf{y}_i - \mathbf{y}^*)$, a subset of the observations whose ranks are the smallest $100(1-\beta)\%$ may be chosen and employed for defining a robust estimator of the covariance matrix,

$$\mathbf{S}_3^* = \frac{k}{n(1-\beta)} \sum_{\substack{r \in \text{chosen subset} \\ \text{of observations}}} (\mathbf{y}_r - \mathbf{y}^*)(\mathbf{y}_r - \mathbf{y}^*)', \tag{74}$$

where k is a constant that will hopefully make the estimator "sufficiently unbiased," and again β has to be small enough so that $[n(1-\beta)] > p$ and $\mathbf{S}_3^*$ is positive definite with probability 1. It may be convenient, but it is not imperative, to have $\alpha = \beta$. The above steps can be repeated using the sum of products on the right-hand side of Eq. 74 in place of $\mathbf{A}_0$, repeating the ranking of the observations, subselecting a major fraction of them for obtaining a further estimate, and iterating the process until a stable estimate is obtained. The limited experience of the authors with this method seems to suggest that, unless α, β, and n are moderately large (viz., α and $\beta \geqslant 0.2$ and $n \geqslant 50$) and unless the underlying correlation structure for the observations is nearly singular, many iterations will not be necessary to improve the estimate defined by Eq. 74. On the other hand, the work of Devlin et al. (1975) indicates that some care in the starting point (viz., not starting with ranking on simple Euclidean distances) may yield significant improvements in the estimator obtained.

The scheme involved in obtaining $\mathbf{S}_3^*$ depends on having an estimate, $\mathbf{y}^*$, of location. A natural way of obtaining the needed location estimator would be to calculate it as the mean of the subset of the observations (i.e., the untrimmed ones) at each stage of the iteration. In this case, at convergence, the above iterative scheme would lead to estimates of both location and dispersion, $\mathbf{y}^*$ and $\mathbf{S}_3^*$, respectively. While this is the preferred scheme, one can also use any of the earlier mentioned location estimators derived from univariate analyses of the responses (e.g., vector of medians) without any change from iteration to iteration. In fact, even for the iterative determination of $\mathbf{y}^*$ and $\mathbf{S}_3^*$, it is sensible to use a vector of simple univariate robust estimators of location such as the vector of medians, as the starting value of $\mathbf{y}^*$ for the iterations.

If one is interested in obtaining an estimator of dispersion not involving a location estimator, however, exactly as in the estimation of univariate variance one can, in the multiresponse situation, work with pairwise differences, $(\mathbf{y}_i - \mathbf{y}_{i'})$, the p-dimensional observations. Specifically, an estimator, $\mathbf{S}_4^*$, can be obtained by repeating each of the steps involved in getting $\mathbf{S}_3^*$ with $(\mathbf{y}_i - \mathbf{y}^*)$ there replaced by $(\mathbf{y}_i - \mathbf{y}_{i'})$, working with rankings of these $n(n-1)/2$ differences, and obtaining as an estimator analogous to $\mathbf{S}_3^*$ the matrix

$$\mathbf{S}_4^* = \frac{k'}{n(n-1)(1-\beta)} \sum_{\substack{r,s \in \text{chosen subset} \\ \text{of observations}}} (\mathbf{y}_r - \mathbf{y}_s)(\mathbf{y}_r - \mathbf{y}_s)'. \tag{75}$$

Just as in the univariate variance situation mentioned earlier, this estimator may be poor when a large fraction of the observations are outliers.

The multiplicative constants k and k' in Eqs. 74 and 75 are not as simply conceptualized or computed as the constants involved in the trimmed or Winsorized variances and covariances. The hope is that, although these constants may depend on n p, α, and β, they will not depend on the underlying variances and/or correlations and also, for practical convenience, that a single multiplicative constant will be adequate for "blowing up" the estimator to make it sufficiently unbiased or consistent. This aspect of the problem needs further research.

The iterative ellipsoidal trimming of a fraction of the most distant multiresponse observations can be thought of as a multivariate analogue of univariate trimming. One advantage of this conceptualization is to ask if it would be better, in some sense, to downweight distant observations more smoothly than abruptly trimming them. Indeed, such downweighting is what is involved in formulating the problem as one of m-estimation. In particular, the m-estimates proposed by Maronna (1976) and Huber (1977) are examples of such iteratively weighted estimates with weights decreasing more smoothly than the ellipsoidal trimming method implies. These m-estimates are affine commutative.

The basic equations defining the m-estimators of multivariate location, $\mathbf{y}^*$, and of dispersion, $\mathbf{S}^*$, are:

$$\mathbf{y}^* = \sum_{i=1}^{n} \{w_1(d_i)\mathbf{y}_i\} \Big/ \sum_{i=1}^{n} w_1(d_i)$$

$$\mathbf{S}^* = (1/n) \sum_{i=1}^{n} w_2(d_i^2)(\mathbf{y}_i - \mathbf{y}^*)(\mathbf{y}_i - \mathbf{y}^*)', \tag{75a}$$

and

$$d^2(\mathbf{y}) = (\mathbf{y} - \mathbf{y}^*)'\mathbf{S}^{*-1}(\mathbf{y} - \mathbf{y}^*).$$

[*Note*: Since the weights, $w_1(d_i)$ and $w_2(d_i^2)$, are themselves functions of $\mathbf{y}^*$ and $\mathbf{S}^*$, the m-estimators have to be computed iteratively.] The specific m-estimators proposed by Maronna and by Huber involve explicit suggestions for the functions, $w_1(d_i)$ and $w_2(d_i^2)$.

Maronna's suggestion is to use

$$w_1(d) = (p + f)/(f + d) = w_2(d^2), \tag{75b}$$

where f is an integer. This suggestion is related to the maximum likelihood estimates of location and dispersion for the so-called multivariate t-distribution based on f degrees of freedom (see Section 5.4.3).

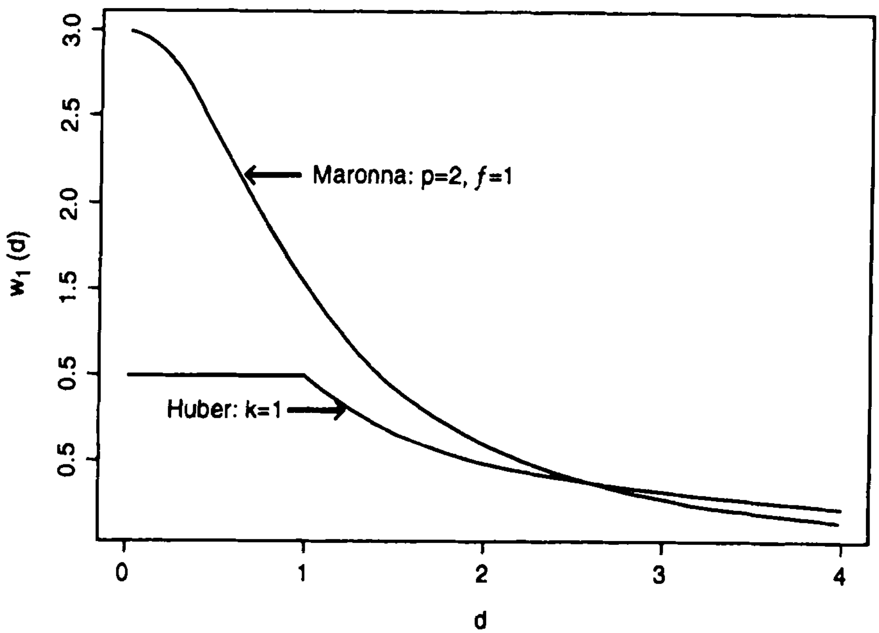


Fig. 10. Weight functions for two *m*-estimates

Huber's proposal, arguing by analogy with the univariate case, is to use

$$w_1(d) = \begin{cases} 1, & \text{if } d \leqslant k, \\ k/d, & \text{if } d > k, \end{cases} \tag{75c}$$

and

$$w_2(d^2) = \{w_1(d)\}^2/\beta,$$

where β is a constant chosen so as to make the estimator, S*, "unbiased".

Looked at as iteratively weighted estimates of location, the weights $w_1(d)$ involved in the Huber scheme do not approach zero for distant observations as quickly as those for the Maronna scheme. Figure 10 sketches the shapes of these two schemes of weighting and indicates this difference at least qualitatively. From the formulae defining the two weighting schemes and this picture, one can see that, by appropriately choosing the weights, the Maronna scheme could be made to weight distant observations even less than the Huber scheme.

Associated with each of the robust estimators of the covariance matrix, such as S_3^* and S_4^*, is a robust estimator of the correlation matrix, which may be obtained by pre- and postmultiplying the covariance matrix estimate by a diagonal matrix whose elements are reciprocals of the square roots of the diagonal elements of the covariance matrix estimate. For instance, one such estimate would be $\mathbf{R}_3^* = \mathbf{D}\mathbf{S}_3^*\mathbf{D}$, where the jth diagonal element of $\mathbf{D}$ would be the reciprocal of the square root of the jth diagonal element of $\mathbf{S}_3^*$. One implication of this is that the robust estimators of the correlation matrix derived in this way can be obtained without knowing the multiplicative constants, such as k and k', as long as these do not depend on the underlying (and unknown) variances and/or correlations. An estimate of bivariate correlation obtained in this manner by trimming whole observations is denoted as $r^*(\text{BVT})$ and included in the comparative study of robust estimators by Devlin et al. (1975). An important feature of this robust estimator is its ability to provide protection against asymmetric outliers.

As to the influence functions of the estimators of location and dispersion discussed above, the important feature of the robust estimators is that their influence functions are bounded unlike those of the classical estimators. As to breakdown points, estimates such as the median for univariate location and MAD for univariate scale have the high value of 0.5, whereas the α-trimmed estimators have the value α. For the multiresponse situation, Tyler (1983) obtains asymptotic efficiencies of robust estimates of dispersion, including m-estimates, in the context of likelihood ratio tests. Devlin et al. (1981) demonstrate empirically that the α-trimmed estimators, such as $\mathbf{y}^*$ and $\mathbf{S}_3^*$ and $\mathbf{R}_3^*$ associated with Eq. 74, have breakdown values equal to α. A curious theoretical result due to Maronna (1976) and Huber (1977) is that the m-estimates of multiresponse location and dispersion described above have a breakdown value $\leqslant (1/p)$ regardless of the fraction of outliers! This implies that these m-estimates would break down when p is large, that is, in high enough dimensions. The simulation study of Devlin et al. (1981) provides empirical support for this aspect of the m-estimates of multiresponse location and dispersion. Motivated in part by seeking an estimator with a breakdown property comparable to the univariate median and MAD, Rousseeuw (1983) has recently been advocating an estimator known as the minimal volume ellipsoid estimator (MVE) for multivariate location and dispersion.

Robust estimators, such as $\mathbf{R}_3^*$, of correlation matrices can serve as starting points for more complex analyses such as principal components and factor analysis discussed in Chapter 2 and canonical correlation analysis discussed in Chapter 3. In fact, a situation in which robust estimators might be extremely useful is one that involves a very large amount of data which are subjected to a series of reasonably complex statistical analyses, with the output of one analysis constituting the input of another. In such a situation one may not want a few observations to influence excessively the final outcome or conclusions.

Example 29. Data on the incidence rates of five types of cancer for white males in 41 states is used to illustrate the use of a robust estimate of a correlation matrix as the input to a principal components analysis instead of the usual correlation matrix. The rates were, in fact, calculated as averages of

Exhibit 29*a*. Correlation matrix, **R**, for cancer rates

	2. Stomach	3. Small Intest.	4. Colon	5. Rectum
1. Esophagus	.49	.22	.89	.87
2. Stomach		.30	.51	.66
3. Small Intest.			.25	.12
4. Colon				.93

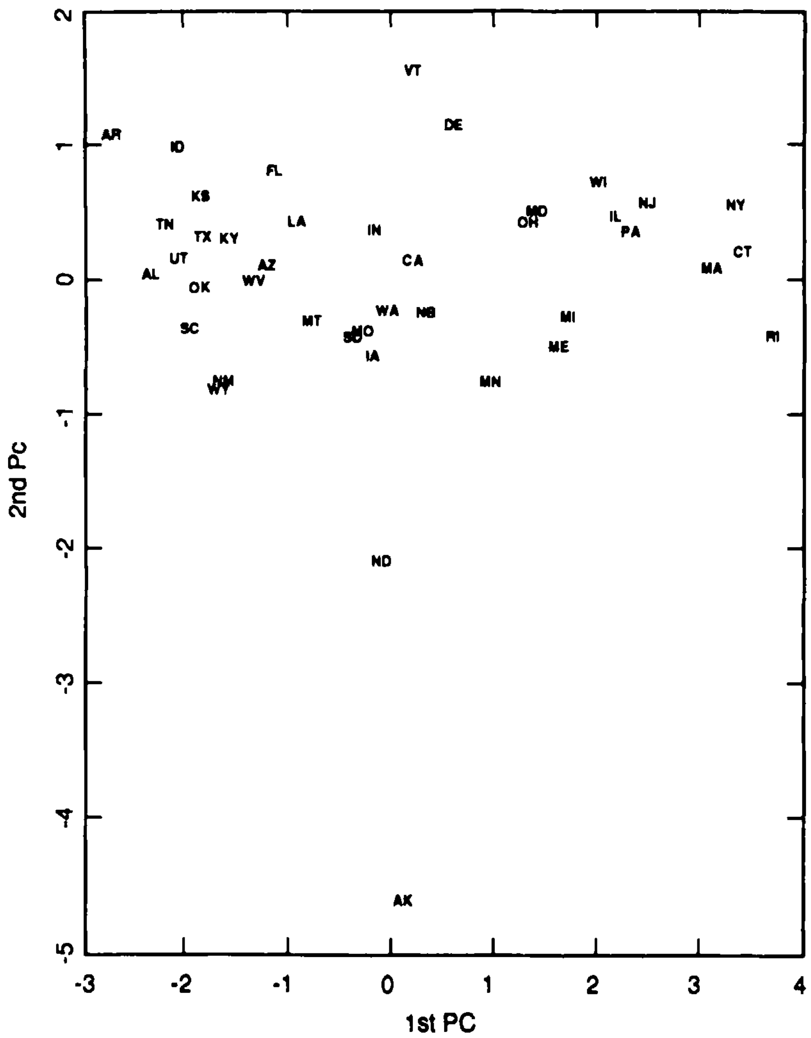
Eigenvalues:	c_1	c_2	c_3	c_4	c_5
	3.28	.97	.58	.14	.03
Eigenvectors:	.50	.14			−.06
	.40	−.20			−.20
	.19	−.93			.14
	.52	.12			−.66
	.53	.23			.71

the annual rates for the period 1950–1967, lending some support to considering the data as being distributed approximately normally. Another feature of the variables here is that, being on very similar scales, they are commensurable, so that one might hope that the findings of a principal components analysis of the covariance matrix and the correlation matrix would be comparable.

Using the 41 ($=n$) 5-dimensional observations, the standard correlation matrix, **R**, was computed and is shown in the top half of Exhibit 29*a*. The correlation between the rates for esophagus and colon cancers, as well as that between the rates of esophagus and rectum cancers, are seen to be relatively high (0.89 and 0.85, respectively), while the correlation between the rates of small intestinal and rectal cancers is low (0.12). The bottom portion of the exhibit shows the eigenvalues and eigenvectors of **R** resulting from a principal components analysis of **R**. [All five eigenvalues are shown but only the eigenvectors associated with the largest two eigenvalues and the smallest one are displayed.]

The first two principal components are seen to account for approximately 85% of the total variance of the five standardized variables. The first principal component seems to be a weighted average of the five standardized variables with roughly equal weights for all but the standardized small intestinal cancer rate, which receives a smaller weight. The second principal component, on the other hand, places most weight on the small intestinal cancer rate. Switching to the last principal component, the smallness of the eigenvalue (0.03) suggests that the linear combination defined by this component is essentially a constant, that is, an approximate linear relationship among the five standardized

Exhibit 29*b*. Plot of 41 states in the space of the first two principal components of the usual correlation matrix



variables is identified. The linear relationship seems not to be simple but a somewhat complex contrast between rectal cancer and three of the other cancer rates, colon, small intestine and stomach.

Exhibit 29*b* shows a plot of the 41 states, using their zip code abbreviations as labels, in the space of the first two principal components. An interesting indication is that, while there appear to be no outliers with respect to the first principal component, Alaska stands out as an outlier in the second principal component. North Dakota is also a moderate outlier in the second principal component.

Exhibit 29c. Stem-and-leaf display of squared distances

```
n = 41      Median = 4.9
Quartiles = 3.1, 6.9


Decimal point is at the colon
1:57889
2:26788
3:145789
4:005899
5:023557
6:089
7:3
8:
9:066
10:8
11:1

High: 25.1 (New Mexico), 26.9 (Vermont), 35.5 (North Dakota), 156.8 (Alaska)
```

To study the possibility of outliers in the original 5-dimensional data, a robustified version of the Mahalanobis squared distances of the 41 observations was calculated. The robustification consisted of using the ellipsoidal trimming scheme described by Eq. 72 and iteratively trimming the observations from the four states with the largest distances, that is, using a trimming fraction $\alpha \simeq 10\%$ for computing the robust estimates of location and dispersion which were then utilized in calculating the Mahalanobis squared distances. Exhibit 29*c* shows the stem-and-leaf display of the values of the squared distances at the last step of the iteration, along with the names of the states whose distances were among the largest. Most of the values lie between 1 and 12, then there are two between 25 and 27, one at about 35, and one way out at over 156. Thus, although not revealed in the two-dimensional space of the principal components, in addition to Alaska and North Dakota, both Vermont and New Mexico are also well removed from the middle of the original 5-dimensional data.

The robust correlation matrix, $\mathbf{R}_3^*$, derived from the ellipsoidal trimming was computed next. The resulting correlations, and the results of the principal components analysis of $\mathbf{R}_3^*$ are shown in Exhibit 29*d*. A striking comparison of $\mathbf{R}$ and $\mathbf{R}_3^*$ is that all of the pairwise correlations, except one, have increased. In the one exception, the already high correlation between esophagus and colon cancer rates remained at the same high value of 0.89. Many of the smaller and moderate correlations have increased noticeably. Thus, in this example, the effects of the outliers have been to deflate the pairwise correlations among the five variables.

Exhibit 29*d*. Ellipsoidally trimmed correlation matrix, $\mathbf{R}_3^*$, for cancer rates

	2. Stomach	3. Small Intest.	4. Colon	5. Rectum
1. Esophagus	.63	.46	.89	.87
	2. Stomach	.57	.74	.83
		3. Small Intest.	.54	.57
			4. Colon	.96

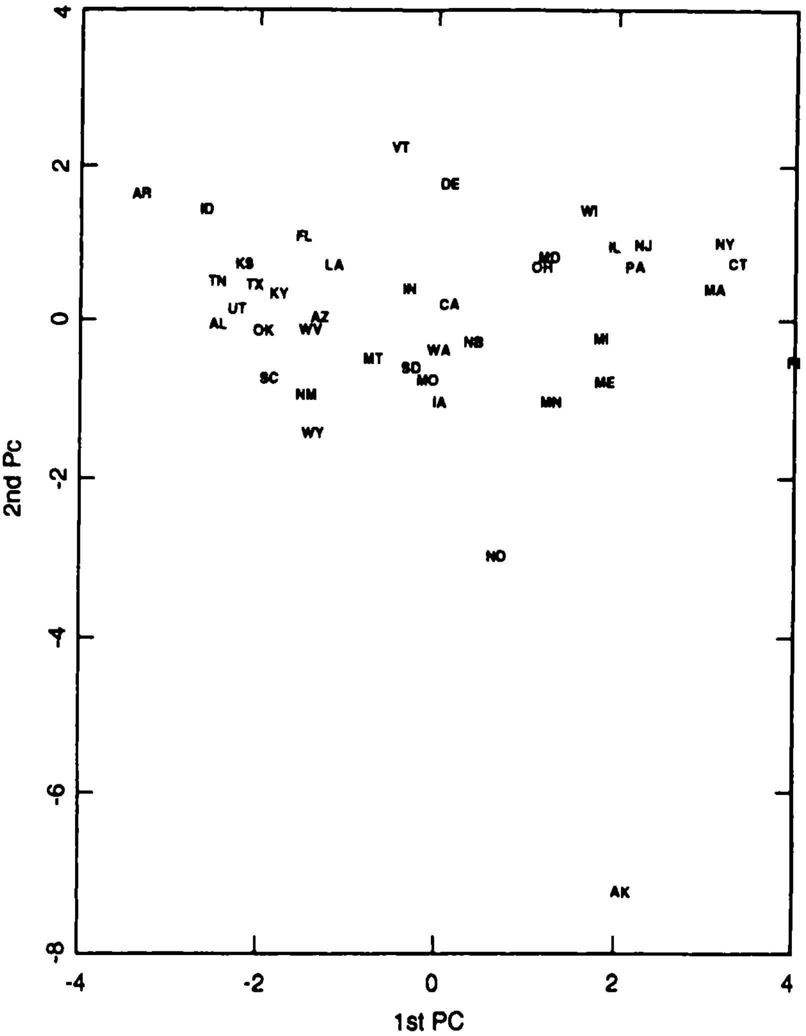
Eigenvalues:	c_1	c_2	c_3	c_4	c_5
	3.87	.64	.36	.11	.02
Eigenvectors:	.45	.37			−.01
	.44	−.13			−.19
	.35	−.87			.00
	.49	.23			−.61
	.50	.16			.77

The principal components based on $\mathbf{R}_3^*$ are also interesting. The first two principal components now account for about 90% of the total variance as against the 85% in the case of the first two principal components of **R**. Also, the first principal component now is a more nearly equally weighted average of the five standardized cancer rates. The second principal component still tends to weight the standardized small intestine rate heavily. Turning to the last principal component, the eigenvalue is a bit smaller than the corresponding eigenvalue of **R**. More interestingly, with two very small weights for esophagus and small intestine cancer rates, the last principal component now emerges as a simpler contrast between the standardized rectum cancer rate and the standardized rates of colon (primarily) and stomach (secondarily).

Exhibit 29*e* shows all 41 states in the space of the first two principal components of $\mathbf{R}_3^*$. Despite the similarity in appearance, the scales in Exhibits 29*b* and 29*e* are very different. Alaska is even more of an outlier as is North Dakota. Both states continue to be outliers in the second principal component and there are no indications of outliers with respect to the first principal component. Other interesting patterns noticed in the display are that, with respect to the first principal component which is essentially a weighted average or index derived from the five individual standardized cancer rates, the northeastern states (Rhode Island, Connecticut, New York, Massachusetts, New Jersey, and Pennsylvania) are at the high end, and states such as Arkansas, Alabama, Tennessee, Idaho, and Utah sit at the low end.

An interesting question about the findings is why the outiers affect, and are revealed by, the second principal component and not the first. The reason is that the first principal component of a correlation matrix is sensitive to the

Exhibit 29e. Plot of 41 states in the space of the first two principal components of a robust estimate of the correlation matrix



larger pairwise correlations. Alaska has a noticeably high small intestine cancer rate and the correlations of small intestine cancer rate with the other cancer rates are relatively small (see Exhibits 29*a* and *d*). These small correlations may be affecting the second principal component but clearly not the first. The second principal component gives highest weight to the small intestine rate, on the other hand, and hence Alaska stands out.

The preceding discussion has dealt with robust estimation of location and dispersion for unstructured multiresponse data. More important, however, is the case of structured mutiresponse data. Analogously to the treatment of the simple location problem, one could of course approach the multiresponse multiple regression problem by simply considering as a robust estimator of the multiresponse regression coefficient vectors the vectors whose elements are just the univariate robust regression coefficients. In other words, with the multiresponse multiple regression structure

$$\mathbf{Y}' = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_p) = \mathbf{X}\mathbf{B} + \boldsymbol{\varepsilon} = \mathbf{X}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_p) + (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_p),$$

given a robust estimator $\hat{\boldsymbol{\beta}}_j^*$ of $\boldsymbol{\beta}_j$, obtained by analyzing the observations on the jth response alone considered according to a uniresponse multiple regression model,

$$\mathbf{Y}_j = \mathbf{X}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j \qquad (j = 1, \ldots, p),$$

a straightforward way of developing a robust estimator $\hat{\mathbf{B}}^*$ of $\mathbf{B}$ is to take

$$\hat{\mathbf{B}}^* = \{\hat{\boldsymbol{\beta}}_1^*, \boldsymbol{\beta}_2^*, \ldots, \hat{\boldsymbol{\beta}}_p^*\}.$$

The estimators $\hat{\boldsymbol{\beta}}_j^*$ can be obtained by using any of the currently available univariate methods (e.g., Huber, 1973; Mallows, 1973; Krasker & Welsch, 1982). This approach to multivariate robust estimation mimics the usual practice of doing separate univariate analyses of the individual responses, "putting together" the univariate results, and considering the amalgamated result as a solution for the multiresponse problem. Once again, although this approach is simple and appealing in certain ways, it seems to be not fully satisfying in the sense that it does not explicitly exploit the multivariate nature of the data.

One approach to simultaneous manipulation of the responses for obtaining $\hat{\mathbf{B}}^*$ is, initially, to get a robust estimator of the $(p + q) \times (p + q)$ covariance matrix of $\begin{pmatrix}\mathbf{Y}\\ \mathbf{X}'\end{pmatrix}$, that is, a robust estimator of

$$\boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22}\end{pmatrix}.$$

[*Note*: in the notation employed here, $\mathbf{X}$ is an $n \times q$ matrix of values of the q regression variables.] If $\hat{\boldsymbol{\Sigma}}^*$ denotes such a robust estimator, then a robust estimator, $\hat{\mathbf{B}}^*$, which is based on the linearity of all regressions for elliptical distributions (see Section 5.4.3), would be defined by

$$\hat{\mathbf{B}}^* = \hat{\boldsymbol{\Sigma}}_{22}^{*-1}\hat{\boldsymbol{\Sigma}}_{12}^{*\prime}.$$

Such an estimator would be sensible for a wide class of distributions, including the normal and heavier-tailed alternatives that might serve as models of certain types of outliers. However, it may not be a reasonable estimator in the presence of other types of outliers such as asymmetric ones perhaps. These questions, as well as issues of, and methods for, robust estimation of parameters occurring in multiresponse designed experiments (the multivariate analysis of variance or MANOVA setup), need to be addressed by future research.

Apart from getting robust estimators of the multiresponse regression coefficient matrix, there is the very important question of the statistical behavior of the "robustified residuals," $\mathbf{Y}' - \mathbf{X}\hat{\mathbf{B}}^*$, as opposed to the behavior of the usual least squares residuals (see Section 6.4 and Examples 51 and 52 for further discussion).

A final comment about robust methods for multivariate data analysis may be in order. Formulating most problems as ones of "fitting" a model, one approach to robustness represented by the *m*-estimation framework is to replace the more classical scheme of minimizing a squared-error criterion (e.g., least squares in location estimation; squares of orthogonal deviations in principal components analysis) by criteria that are less sensitive to outliers and develop new specialized algorithms for each problem. Such specialization may provide methods that have particular advantages for the problem at hand, but would typically involve building the algorithms from the ground up. A simpler approach that is computationally appealing would be to robustify the input to a classical technique and not alter the basic algorithm. This is the approach illustrated in Example 29, where a robust correlation matrix was used as input to a standard eigenanalysis algorithm for determining the principal components. From a practical viewpoint, what is important to keep in mind is the need to protect oneself against the possible presence of outliers and keep the computational efforts from becoming too cumbersome. Also, the most important thing to do is to carry out both a standard (perhaps nonrobust), often well-understood, analysis of a data set and a robust analysis of it, and then compare the two sets of results. One often gains valuable insights into the data from both the similarities of, and the differences between, the results of the two analyses.

5.3. DATA-BASED TRANSFORMATIONS

As stated in Section 5.2, the classical multivariate theory has been based largely on the multivariate normal distribution and the paucity of alternative models for the useful guidance of multiresponse data analysis is a well-recognized limitation. One way of handling this limitation has been to develop nonparametric or distribution-free methods for specifically posed problems such as the formal inferential ones mentioned in Section 5.2. (See, for example, Puri & Sen, 1971, for an extensive treatment of multivariate nonparametric inference.) Although such methods may serve the specific purpose for which they are designed, the statistics employed by them are not always useful for revealingly summarizing the structure in a body of data. On the other hand, the serendipitous value of many classical methods lies in their utility for summarizing the structure underlying data. Hence it is natural and appropriate to inquire about ways of transforming the data so as to permit the use of more familiar statistical techniques based implicitly or explicitly on normal distributional theory. The choice of a transformation, of course, should depend on the nature of the objectives of the data analysis, and transforming to obtain more nearly normally distributed data is only one of several possible reasonable motivations. Moreover, even if a transformation of variables does not accomplish normality, it may often go a long way toward symmetrizing the data, and this can be a significant improvement of the data as a preliminary to computing standard statistical summaries such as correlation coefficients and covariance matrices.

A transformation may be based on theoretical (or a priori) considerations or be bootstrapped (or estimated) from the data that are being analyzed. Examples of the former type are the logistic transformation of binary data proposed by Cox (1970, 1972) and the well-known variance-stabilizing transformations of the binomial, the Poisson, the correlation coefficient, etc. Techniques for developing data-based transformations of univariate observations have also been proposed by several authors (see, for example, Moore & Tukey, 1954; Tukey, 1957; Box & Cox, 1964). Andrews et al. (1971) have extended the approach of Box & Cox (1964) to the problem of estimating a power transformation of multiresponse data so as to enhance normality, and the present section is a summary of their proposals and results.

If $\mathbf{y}' = (y_1, y_2, \ldots, y_p)$ denotes the set of p response variables, the general problem may be formulated as follows: determine the vector of transformation parameters, $\boldsymbol{\lambda}$, such that the transformed variables $\{g_1(\mathbf{y}'; \boldsymbol{\lambda}), g_2(\mathbf{y}'; \boldsymbol{\lambda}), \ldots, g_p(\mathbf{y}'; \boldsymbol{\lambda})\}$ are "more nearly" p-variate normal, $N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, than the original p variables. The elements of $\boldsymbol{\lambda}$ are unknown, as are those of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Provided that one can obtain an appropriate estimate, $\hat{\boldsymbol{\lambda}}$, of $\boldsymbol{\lambda}$ (as well as of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) from the data, the original observations, $\mathbf{y}_i'$ ($i = 1, \ldots, n$), can be transformed one at a time to yield new observations, $\{g_1(\mathbf{y}_i'; \hat{\boldsymbol{\lambda}}), \ldots, g_p(\mathbf{y}_i'; \hat{\boldsymbol{\lambda}})\}$, which may then be considered as more nearly conforming to a p-variate normal model than the original observations.

The work of Andrews et al. (1971) is concerned with transformation functions, g_j, which are direct extensions of the power transformation of a single nonnegative response, X, to $X^{(\lambda)}$, considered by Moore & Tukey (1954) and by Box & Cox (1964), where

$$X^{(\lambda)} = \begin{cases} (X^\lambda - 1)/\lambda & \text{for } \lambda \neq 0, \\ \ln X & \text{for } \lambda = 0. \end{cases}$$

Furthermore, for simplicity of both the exposition and the computations, some of the details are developed only for the bivariate case, that is, $p = 2$.

For ease of interpretation it may be desirable to look for transformations that operate on each of the original variables separately. A simple family of such transformations is defined by

$$g_j(\mathbf{y}'; \boldsymbol{\lambda}) = y_j^{(\lambda_j)} = \begin{cases} (y_j^{\lambda_j} - 1)/\lambda_j & \text{for } \lambda_j \neq 0, \\ \ln y_j & \text{for } \lambda_j = 0, \end{cases} \tag{76}$$

where $j = 1, 2$ for the bivariate case and $j = 1, 2, \ldots, p$ for the general p-variate case.

A natural starting point is to choose λ_j so as to improve the marginal normality of $y_j^{(\lambda_j)}$. Although it is recognized that marginal normality does not imply joint normality, the choice of transformations to improve marginal normality may in many cases yield data more amenable to standard analyses. The procedure is merely to apply the method proposed by Box & Cox (1964) to each response separately so that only univariate computations are involved and the theory and techniques for each are identical with those of Box & Cox. Specifically, one of the approaches suggested by Box & Cox leads to estimating λ_j by maximum likelihood, using only the observations on the jth response variable. The logarithm of a likelihood function (which has been initially maximized with respect to the unknown mean and variance for given λ_j), $\mathscr{L}_{\max}(\lambda_j)$, is maximized to provide the estimate $\hat{\lambda}_j$. If $\mathbf{Y}' = [\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_p]$ denotes the $n \times p$ matrix of original observations, and if the transformed observations obtained by using Eq. 76 are denoted as

$$(\mathbf{Y}^{(\lambda)})' = [\mathbf{Y}_1^{(\lambda_1)}, \ldots, \mathbf{Y}_j^{(\lambda_j)}, \ldots, \mathbf{Y}_p^{(\lambda_p)}],$$

where $\mathbf{Y}_j^{(\lambda_j)}$ denotes the vector of n observations on the jth variable, each of which has been obtained by transforming according to Eq. 76, then

$$\mathscr{L}_{\max}(\lambda_j) = -\frac{n}{2} \ln \hat{\sigma}_{jj} + (\lambda_j - 1) \sum_{i=1}^{n} \ln y_{ij}, \tag{77}$$

where y_{ij} denotes the ith observation on the untransformed jth response, and $\hat{\sigma}_{jj}$ is the maximum likelihood estimate of the variance of the presumed normal distribution of $\mathbf{Y}_j^{(\lambda_j)}$ [i.e.,

$$\hat{\sigma}_{jj} = \frac{1}{n} (\mathbf{Y}_j^{(\lambda_j)} - \hat{\boldsymbol{\xi}}_j)'(\mathbf{Y}_j^{(\lambda_j)} - \hat{\boldsymbol{\xi}}_j),$$

where $\hat{\boldsymbol{\xi}}_j$ is the maximum likelihood estimate of $\boldsymbol{\xi}_j = \mathscr{E}(\mathbf{Y}_j^{(\lambda_j)})$; specifically, for an unstructured sample, $\hat{\boldsymbol{\xi}}_j$ would be an $n \times 1$ vector all of whose elements are equal to the mean of the transformed observations on the jth variable, while for the more general case of a linear model specification, $\boldsymbol{\xi}_j = \mathbf{X}\boldsymbol{\theta}_j$, the appropriate estimate would be $\hat{\xi}_j = \mathbf{X}\hat{\boldsymbol{\theta}}_j$]. In addition to the second term on the right-hand side of Eq. 77, $\hat{\sigma}_{jj}$ is also a function of λ_j, and the required maximum likelihood estimate, $\hat{\lambda}_j$, is the value of λ_j which maximizes $\mathscr{L}_{\max}(\lambda_j)$ as defined by Eq. 77. Despite the complication caused by $\hat{\sigma}_{jj}$ being a function of λ_j, since the maximization is with respect to a single unknown parameter λ_j the computations involved are quite simple. In fact, one can compute the value of $\mathscr{L}_{\max}(\lambda_j)$ for a sequence of values of λ_j and empirically determine the value, $\hat{\lambda}_j$, for which it is a maximum. Also, for this case of a single parameter, one can graph $\mathscr{L}_{\max}(\lambda_j)$ and study its behavior near $\hat{\lambda}_j$.

Following Box & Cox (1964), by using approximate asymptotic theory one can also obtain an approximate confidence interval for λ_j. The essential result is that a 100(1 − α)% confidence interval for λ_j is defined by

$$2\{\mathscr{L}_{\max}(\hat{\lambda}_j) - \mathscr{L}_{\max}(\lambda_j)\} \leqslant \chi_1^2(\alpha),$$

where $\chi_\nu^2(\alpha)$ denotes the upper 100α% point of a chi-squared distribution with ν degrees of freedom.

The preceding discussion has been concerned with estimating power transformations of multiresponse data so as to improve marginal normality. Next, a method is described for choosing the transformations of Eq. 76 so as to enhance joint normality. To keep the computations simple, this description will be presented just in terms of a bivariate response situation. Thus the $n \times 2$ matrix $\mathbf{Y}' = ((y_{ij}))$, $i = 1, \ldots, n$; $j = 1, 2$ is the data matrix whose rows, $\mathbf{y}_i'$, are the bivariate observations, and it is assumed that after a transformation of the form in Eq. 76 the transformed data $(\mathbf{Y}^{(\boldsymbol{\lambda})})'$ may be statistically described by a *bivariate* normal density function with mean $\boldsymbol{\mu}'$ and covariance matrix $\boldsymbol{\Sigma}$.

Let $\boldsymbol{\Xi}' = \mathscr{E}[(\mathbf{Y}^{(\boldsymbol{\lambda})})'] = \mathbf{1}\cdot\boldsymbol{\mu}'$. [*Note*: For simplicity the sample is considered to be unstructured; however, the treatment for a structured sample with a general linear model specification is quite straightforward, requiring only that $\mathbf{X}\boldsymbol{\Theta}$ be substituted for $\mathbf{1}\cdot\boldsymbol{\mu}'$.] If

$$\boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}$$

is the set of transformation parameters yielding bivariate normality, the density function of the original data, $\mathbf{Y}$, is

$$f(\mathbf{Y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-n} \exp[-\tfrac{1}{2} \operatorname{tr} \boldsymbol{\Sigma}^{-1}(\mathbf{Y}^{(\boldsymbol{\lambda})} - \boldsymbol{\Xi})(\mathbf{Y}^{(\boldsymbol{\lambda})} - \boldsymbol{\Xi})']J,$$

where J, the Jacobian of the transformation from $\mathbf{Y}^{(\boldsymbol{\lambda})}$ to $\mathbf{Y}$, is

$$\prod_{j=1}^{2} \prod_{i=1}^{n} y_{ij}^{\lambda_j - 1}.$$

Thus the log likelihood of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\lambda}$ is given (aside from an additive constant) by

$$\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda} \mid \mathbf{Y}) = -\frac{n}{2} \ln |\boldsymbol{\Sigma}| - \tfrac{1}{2} \operatorname{tr} \boldsymbol{\Sigma}^{-1}(\mathbf{Y}^{(\lambda)} - \boldsymbol{\Xi})(\mathbf{Y}^{(\lambda)} - \boldsymbol{\Xi})' + \sum_{j=1}^{2} \left[(\lambda_j - 1) \sum_{i=1}^{n} \ln y_{ij} \right].$$

For specified λ_1 and λ_2, the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given, respectively, by

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \mathbf{Y}^{(\lambda)} \cdot \mathbf{1},$$

and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} (\mathbf{Y}^{(\lambda)} - \hat{\boldsymbol{\Xi}})(\mathbf{Y}^{(\lambda)} - \hat{\boldsymbol{\Xi}})',$$

where $\hat{\boldsymbol{\Xi}}' = \mathbf{1} \cdot \hat{\boldsymbol{\mu}}'$. If these estimates are substituted in the above log-likelihood function, the resulting maximized function (up to an additive constant) is

$$\mathscr{L}_{\max}(\lambda_1, \lambda_2) = -\frac{n}{2} \ln |\hat{\boldsymbol{\Sigma}}| + \sum_{j=1}^{2} \left[(\lambda_j - 1) \sum_{i=1}^{n} \ln y_{ij} \right], \tag{78}$$

a function of two variables that may be computed and studied. The maximum likelihood estimates $\hat{\lambda}_1$ and $\hat{\lambda}_2$ may be obtained by numerically maximizing Eq. 78. Also an approximate $100(1 - \alpha)\%$ confidence region for λ_1 and λ_2, obtained on the basis of asymptotic considerations, is

$$2\{\mathscr{L}_{\max}(\hat{\lambda}_1, \hat{\lambda}_2) - \mathscr{L}_{\max}(\lambda_1, \lambda_2)\} \leqslant \chi_2^2(\alpha),$$

where $\chi_2^2(\alpha)$ is the upper $100\alpha\%$ of the chi-squared distribution with 2 degrees of freedom.

It is easy to see that Eq. 77 is the univariate version of the bivariate version in Eq. 78, and that both result from using the power transformations in Eq. 76 and a likelihood approach, except that Eq. 77 is the result of specifying marginal normality whereas Eq. 78 is a consequence of specifying bivariate normality. In fact, for the general case of p responses, if one were to start with the transformations in Eq. 76 and specify a p-variate normal distribution, $N[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$, for the transformed observations, then, following the same arguments used in arriving at Eq. 78 for the bivariate case, one would obtain for the general case the following log-likelihood function of $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)'$ after initial maximization with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

$$\mathscr{L}_{\max}(\lambda_1, \lambda_2, \ldots, \lambda_p) = -\frac{n}{2}\ln|\hat{\boldsymbol{\Sigma}}| + \sum_{j=1}^{p}\left[(\lambda_j - 1)\sum_{i=1}^{n}\ln y_{ij}\right], \tag{79}$$

where y_{ij} is the ith observation on the (untransformed) jth response ($i = 1, \ldots, n; j = 1, 2, \ldots, p$), and the $p \times p$ matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}(\mathbf{Y}^{(\lambda)} - \hat{\boldsymbol{\Xi}})(\mathbf{Y}^{(\lambda)} - \hat{\boldsymbol{\Xi}})',$$

$$\hat{\boldsymbol{\Xi}}' = \begin{cases} \dfrac{1}{n}\;\mathbf{1}\cdot\mathbf{1}'(\mathbf{Y}^{(\lambda)})' & \text{for an unstructured sample,} \\ \mathbf{X}\hat{\boldsymbol{\Theta}} & \text{for a general linear model specification.} \end{cases}$$

For this general p-response case, however, $\mathscr{L}_{\max}(\boldsymbol{\lambda})$ is a function of p variables, $\lambda_1, \lambda_2, \ldots, \lambda_p$, and thus the problem of studying and numerically maximizing it is more complex than in the bivariate case (see Chambers, 1973, for a discussion of optimization techniques). Formally, however, if $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ are the values that maximize $\mathscr{L}_{\max}(\lambda_1, \ldots, \lambda_p)$, an approximate confidence region for $(\lambda_1, \ldots, \lambda_p)$, analogous to the bivariate one mentioned above, is defined by

$$2\{\mathscr{L}_{\max}(\hat{\lambda}_1, \ldots, \hat{\lambda}_p) - \mathscr{L}_{\max}(\lambda_1, \ldots, \lambda_p)\} \leqslant \chi_p^2(\alpha),$$

where $\chi_p^2(\alpha)$ is the upper $100\alpha\%$ point of the chi-squared distribution with p degrees of freedom.

In certain situations, data may exhibit nonnormality in some but not all directions in the space of the original responses. One way of thinking about the two approaches discussed thus far is that the one directed toward improving marginal normality is concerned with p directions, one for each of the original coordinates, whereas the approach directed toward enhancing joint normality is concerned with all possible directions. The method to be described next is concerned with identifying directions (not necessarily confined to prespecified directions such as those of the coordinate axes) of possible nonnormality and then estimating a power transformation of the projections of the original observations onto these directions so as to improve normality along them. The specification of a direction will in general depend on several and possibly all, coordinates, and hence the method no longer involves just transformations of each coordinate separately.

As before, let $\mathbf{Y}'$ denote the data matrix whose rows, $\mathbf{y}_i'$, $i = 1, \ldots, n$, are the multiresponse observations. With a general multivariate linear model specification, $\mathscr{E}(\mathbf{Y}') = \mathbf{X}\boldsymbol{\Theta}$ (which includes the case of an unstructured sample by specifying $\mathbf{X}$ as an n-vector of unities, $\mathbf{1}$, and $\boldsymbol{\Theta}$ as the unknown mean vector, $\boldsymbol{\mu}'$), one can obtain the residual error covariance matrix, $\mathbf{S}_{\text{error}}$, defined in

Eq. 66. For brevity of notation, S_{error} will be denoted as S in the following discussion.

If $S^{1/2}$ denotes the symmetric square root of S, one can obtain the set of sphericized residual vectors,

$$\mathbf{z}_i' = (\mathbf{y}_i' - \mathbf{x}_i' \cdot \hat{\mathbf{\Theta}})\mathbf{S}^{-1/2}, \qquad i = 1, \ldots, n,$$

where $\mathbf{x}_i'$ denotes the ith row of the design matrix X. [*Note*: Once again, for an unstructured sample, $\mathbf{x}_i' \cdot \hat{\mathbf{\Theta}}$ will just be the sample mean vector, $\bar{\mathbf{y}}'$.] Any nonnormal characteristics of the observations $\mathbf{y}_i'$ will be reflected in corresponding (nonnormal) characteristics of the $\mathbf{z}_i'$, and the direction of any nonnormal clustering of points, if present, may perhaps be identified by studying a normalized weighted sum of the $\mathbf{z}_i'$:

$$\mathbf{d}_\alpha' = \frac{\sum_{i=1}^{n} w_i \mathbf{z}_i'}{\left\| \sum_{i=1}^{n} w_i \mathbf{z}_i \right\|}, \qquad w_i = \|\mathbf{z}_i\|^\alpha,$$

where $\|\mathbf{x}\|$ denotes the Euclidean norm, or length, of the vector $\mathbf{x}$, and α is a constant to be chosen.

The vector $\mathbf{d}_\alpha'$ provides a parametrization of directions in the z-space (and hence in the y-space of the original observations) in terms of the single parameter α. If $\alpha = -1$, $\mathbf{d}_\alpha'$ is a function only of the orientation of the $\mathbf{z}_i$'s, while if $\alpha = +1$, $\mathbf{d}_\alpha'$ becomes sensitive primarily to the observations, $\mathbf{y}_i'$, that are far from the mean $\bar{\mathbf{y}}'$. More generally, for $\alpha > 0$ the vector $\mathbf{d}_\alpha'$ will tend to point toward any clustering of observations far from the mean, while for $\alpha < 0$ the vector $\mathbf{d}_\alpha'$ will point in the direction of any abnormal clustering near the center of gravity of the data. If the scaled residuals are skewed in one direction, $\mathbf{d}_\alpha'$ will tend to point in that direction.

For a specified α, the vector $\mathbf{d}_\alpha'$ (chosen to be sensitive to particular types of nonnormal clusterings if any are present) corresponds to the vector $\mathbf{d}_\alpha^{*\prime} = \mathbf{d}_\alpha'\mathbf{S}^{1/2}$ in the space of the original observations. The projections of the original observations onto the unidimensional space specified by the "direction" $\mathbf{d}_\alpha^{*\prime}$ constitute a univariate sample, and one can estimate a power transformation to improve the normality of these projections by using the univariate technique of Box & Cox (1964) on the unidimensional "sample" of the projections. The effect of the transformation is to alter the data only in the direction $\mathbf{d}_\alpha^{*\prime}$.

The advantage of this method of enhancing directional normality is that the relatively small class of power transformations may be applied to very complex data. The procedure may be applied iteratively, using a different value of α at each stage so as to transform along a different direction. The computations for estimating transformations along each direction are univariate (in the sense that one is working only with the projections onto the unidimensional space specified by each direction), and this is an important pragmatic advantage of this approach.

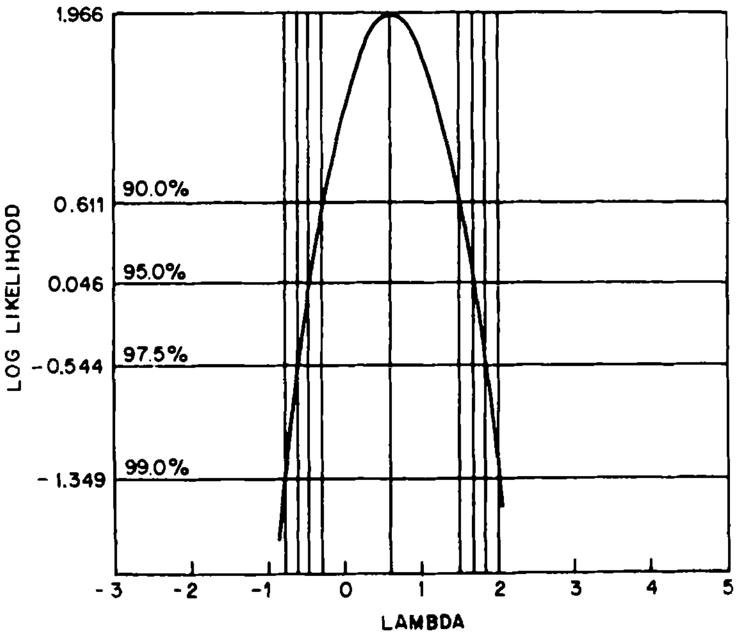
As mentioned in the initial definition of the power transformation, $X \rightarrow X^{(\lambda)}$, a requirement for using this transformation is that the data be nonnegative since otherwise the transformed values may become imaginary for fractional values of λ. A simple way of conforming to this requirement is to shift all of the observations by an arbitrary amount to make them all nonnegative. A different way of handling the problem is to use the more general shifted-power class of transformations, $X \rightarrow (X + \xi)^{(\lambda)}$, where $(X + \xi)$ replaces X, and to treat ξ as an unknown parameter as well. For using the shifted-power transformations the requirement is that X not be smaller than $-\xi$, rather than that it be nonnegative. The main difficulties in using the shifted-power instead of the power transformation are that the computations become more complex and that the interpretation of the resulting estimates, $\hat{\xi}$ and $\hat{\lambda}$, may be complicated because of high correlations between them. Nevertheless, for the transformation approaches that involve only univariate computations (i.e., the schemes aimed at improving marginal normality and directional normality), it is not out of the question to use the shifted-power class of transformations. On the other hand, for the method aimed at improving joint normality, if one were to use the shifted-power transformation the log-likelihood function corresponding to Eq. 79 would be a function of $2p$ parameters, $\{\lambda_1, \xi_1, \lambda_2\, \xi_2, \ldots, \lambda_p, \xi_p\}$, so that even for bivariate response data one would in general have to consider maximizing a function of four variables to obtain the required maximum likelihood estimates of the transformation parameters.

Example 30. The data consist of 50 ($=n$) sets of bivariate normal deviates generated on a computer. Pairs of random standard normal deviates, (x_{1i}, x_{2i}),

Exhibit 30a. Monte Carlo normal data ($p = 2, n = 50$)

	I		II		III	
ρ	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}$	$\mathbf{d}^{*\prime}_{1.0}$
0	0.896	0.957	0.878	0.984	0.688	−0.7, 0.7
0.1	0.896	1.021	0.892	1.009	0.811	−0.6, 0.8
0.3	0.896	1.085	0.890	1.035	0.714	−0.9, 0.4
0.5	0.896	1.041	0.882	1.011	0.728	−1, 0.2
0.75	0.896	0.810	0.887	0.887	0.725	−0.6, 0.8
0.8	0.896	0.745	0.886	0.852	0.729	−0.5, 0.8
0.9	0.896	0.614	0.884	0.782	0.735	−0.3, 0.9
0.95	0.896	0.596	0.884	0.769	0.719	−0.3, 0.9
0.975	0.896	0.642	0.883	0.784	0.737	−0.3, 0.9
0.999	0.896	0.833	0.884	0.859	0.715	−0.6, 0.8

Exhibit 30*b*. Plot of log-likelihood function with associated confidence intervals (Method I); mle of $\lambda = 0.596$



were transformed using the relationships

$$\left.\begin{aligned} y_{1i} &= x_{1i} \\ y_{2i} &= \rho x_{1i} + \sqrt{1-\rho^2 x_{2i}} \end{aligned}\right\}, \qquad i = 1, 2, \ldots, 50,$$

to obtain the 50 samples, (y_{1i}, y_{2i}), from

$$N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right].$$

To avoid negative values (so that power transformations could be employed), the mean vector was shifted sufficiently away from the origin by adding a constant vector (c, c) to each of the observations. A range of values for ρ was used to provide a basis for comparing the different approaches discussed above for transforming observations. For convenience in referring to the approaches, the method of Box & Cox (1964) applied to each variable separately so as to improve marginal normality is called Method I, the method for enhancing joint normality is termed Method II, and the one aimed at improving directional normality is designated as Method III.

Exhibit 30c. Contour plot of log-likelihood surface with associated confidence regions (Method II) for data with $\rho = 0.5$; mle of $\lambda' = (0.882, 1.011)$

```
(*) ∈ 90% conf. set
(*&=) ∈ 95% conf. set
(*,=&X) ∈ 97.5% conf. set
(*,=,X&O) ∈ 99% conf. set

       |
-0.118 |
       |
       |
       |
       |         0000000
       |        0X=====XX0
       |       0X=******=X0
       |      0X=********=X0
       |      0=**********=X0
       |      X=**********=X0
       |      X=**********=X0
 0.882 |      X=**********=X0
       |      0=**********=X0
       |      0X*********=X0
       |       X=*******==00
       |       0X=*****=XX0
       |        0X=====X00
       |         00XXX00
       |           000
       |
       |
       |
 1.882 |
       |
       +------------------------------------
       0.011      1.011       2.011
```

Exhibit 30*a* shows the estimates of the transformation parameters obtained by the three approaches described earlier. The actual outputs of the analyses consist not only of the maximum likelihood estimates involved in each case but also, for Methods I and III, plots of the log-likelihood functions involved, together with the associated approximate confidence intervals, and for Method II a contour plot of the log-likelihood surface displayed with the approximate confidence sets for this case. To minimize the number of displays, only a few sample plots are included here.

Over the range of 10 values of ρ shown in Exhibit 30*a*, it can be seen that the estimates of λ_1 and λ_2 obtained by Method I vary between 0.596 and 1.085. [*Note*: Because of the scheme used in generating the data, the sample of values of the first variable does not change as ρ changes and hence the estimate of λ_1 obtained by Method I remains the same for all ρ.] Moreover, every 95% confidence interval includes not only the "true" value of $\lambda = 1$ (since the

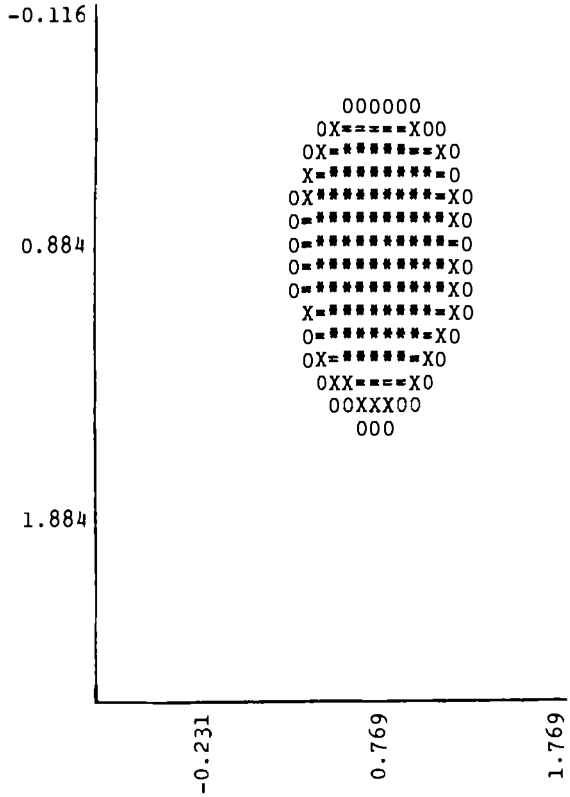
Exhibit 30*d*. Contour plot of log-likelihood surface with associated confidence regions (Method II) for data with $\rho = 0.95$; mle of $\lambda' = (0.884, 0.769)$

```
(*) ∈ 90% conf. set

(*&=) ∈ 95% conf. set

(*,=&X) ∈ 97.5% conf. set

(*,=,X&O) ∈ 99% conf. set
```

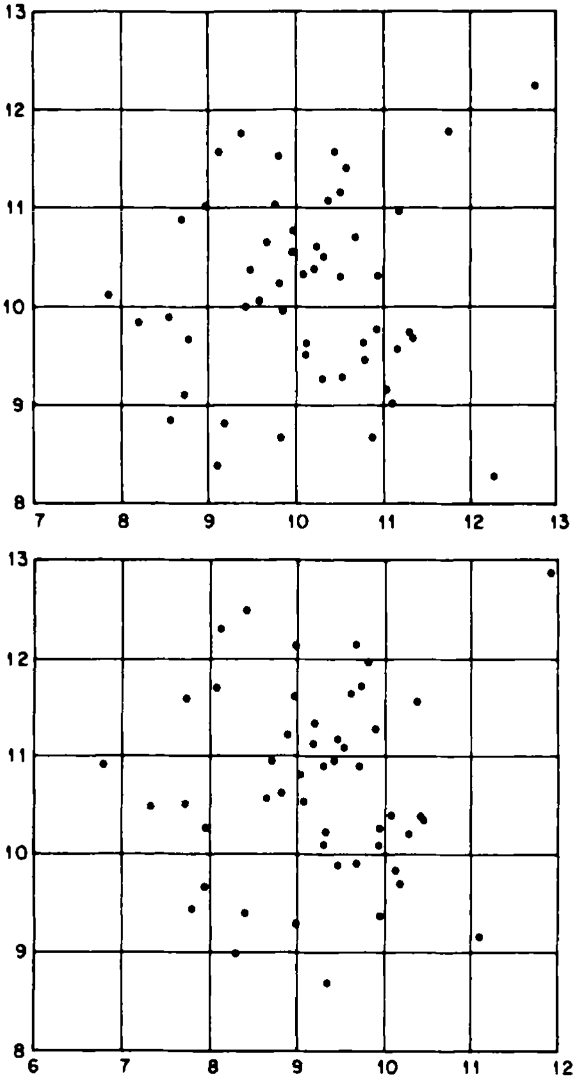


```
    000000
  0X=====X00
 0X=*****==X0
 X=********=0
0X*********=X0
0=**********X0
0=**********=0
0=**********X0
0=**********X0
 X=********=X0
 0=*******=X0
 0X=*****=X0
  0XX====X0
   00XXX00
     000
```

original distributions are all normal) but also every other estimate of λ. Exhibit 30*b* shows a plot of the log-likelihood function of λ_2 when $\rho = 0.95$, the case in which Method I yielded the smallest (and farthest from 1) estimate of the transformation parameter.

Method II yielded estimates of λ_1 and λ_2 that range between 0.878 and 1.035, and Exhibits 30*c* and *d* show the contour plots of the log-likelihood surfaces for the cases when $\rho = 0.5$ and $\rho = 0.95$. Even on the coarse grid used for generating these plots, there is some indication that the contours are tighter for the higher value of ρ.

Exhibits 30e, f. Scatter plots of data before and after transformation by Method III



A very interesting feature of the results in Exhibit 30*a* is the greater stability shown by the estimates obtained by Method II as compared to the ones yielded by Method I. The stability is particularly noticeable as ρ increases, although it is evident even for small values of ρ. This is encouraging in that the bivariate approach, that is, seeking joint normality while still using coordinatewise transformations, is yielding "more" than the repeated application of the univariate approach with each variable separately. It is always legitimate to ask whether one gains anything significant by using a multivariate approach. In the present case it seems that a multivariate approach may be able to exploit the intercorrelations among the variables to advantage and lead to more stable estimates.

The results of applying Method III are also included in Exhibit 30*a*. The value of α used for obtaining the direction of possible nonnormality $\mathbf{d}_\alpha^{*\prime}$, was 1. The estimate of λ as well as the direction, $\mathbf{d}_{1.0}^{*\prime}$, is shown. In this Monte Carlo example, the method appears to identify an arbitrary direction; and, as seen by the $\hat{\lambda}$ values and from the fact that all the 95% confidence intervals included the value 1, the transformation has not altered the data very much. This is also evident in Exhibits 30*e* and *f*, which show, respectively, the data before and after the transformation.

Example 31. To illustrate the use of the method for improving directional normality in a "nonnull" case, bivariate observations were generated for which the first coordinate was distributed lognormally whereas the second was distributed normally independent of the first. For these data, using $\alpha = 1$ in the directional method leads to identifying the direction of nonnormality as $\mathbf{d}_{1.0}^{*\prime} = (1, 0)$, as it should, and $\hat{\lambda} = -0.003$, which again is sufficiently close to 0, the value one would expect. Exhibits 31*a* and *b* show scatter plots of the data before and after transformation, and the achievements of the transformation are clear.

Exhibit 31*a*. Scatter plot of untransformed data

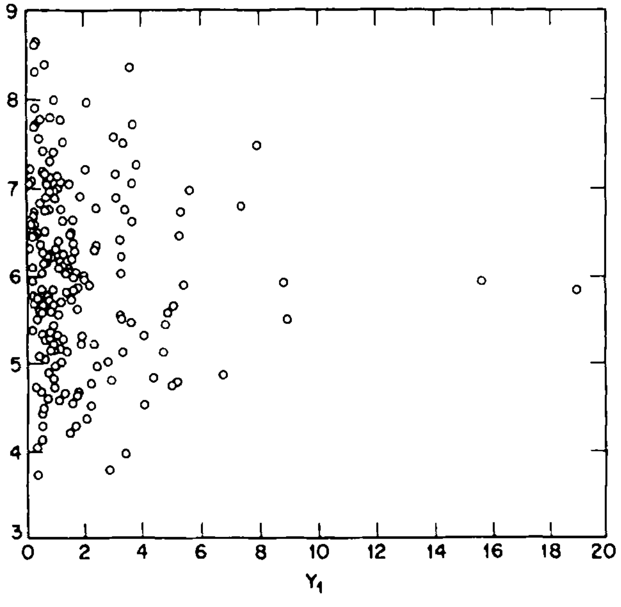
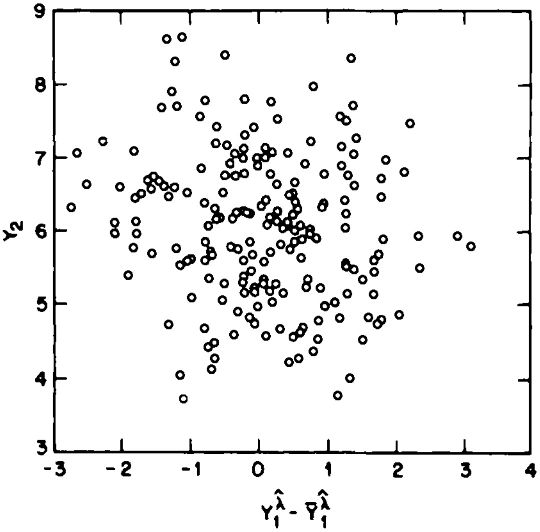


Exhibit 31b. Scatter plot of data transformed by Method III



Other examples of the use of the data-based transformation methods discussed in this section will be given in Chapter 6 (see Example 45 in Section 6.3.1). A key point regarding the transformation methods described heretofore is that, although the objective in estimating a transformation is perhaps to improve normality, there is no guarantee that the resultant transformation will actually achieve adequate normality in any particular application. In other words, some kinds of nonnormality may not be ameliorated by relatively simple types of nonlinear transformations.

5.4. ASSESSMENT OF DISTRIBUTIONAL PROPERTIES

Statistical distributions play a useful role in modeling data. Both fitting and assessing the fit of various distributions to univariate data are common exercises. One reason for the interest in using appropriate distributions for modeling data is the feasibility of obtaining parsimonious representations of data in terms of the parameters (hopefully much fewer in number than the observations) of such distributions.

The variety of univariate distributional models is, of course, very rich, whereas this is not so in the multivariate situation. In fact, the multivariate normal distribution has been almost exclusively at the center of much of the development of multivariate methodology, and, although other multivariate distributions have been proposed as alternative models, far less use has been made of these in practice (see Kendall, 1968). For example, stimulated in part by the need for alternative models in the study of properties of robust estimators such as those discussed in Section 5.2.3, a class of distributions called the elliptical distributions has been considered. This class includes the multivariate normal as a member as well as others which are all elliptically symmetric like the multinormal but having either longer (heavier) or shorter (lighter) tails. Section 5.4.3 contains a brief discussion of elliptical distributions..

The lack of availability of a variety of alternatives is perhaps one explanation for the relative lack of emphasis, in the multiresponse situation as opposed to the univariate one, on assessing distributional properties and assumptions in the light of the data. Nevertheless, certain questions can be posed and solutions to them proposed, and the two subsections that follow are concerned with methods addressed to two sets of questions. Section 5.4.1 will discuss methods for evaluating the similarity of the marginal distributions of the responses, and Section 5.4.2 will describe techniques for assessing the normality of multiresponse data.

5.4.1. Methods for Evaluating Similarity of Marginal Distributions

Many of the theoretical multivariate distributions that have been proposed as bases or models for statistical analyses of data have the feature that the marginal distributions are either identical or common up to origin (or location) and/or scale parameters. For instance, in addition to the multivariate normal, the usual definitions of the multivariate t, F, and beta (or Dirichlet) distributions all incorporate this feature.

Many multivariate summaries (e.g., correlation coefficients or the covariance matrix) seem to depend, for sense and interpretability, on the degree of similarity of the marginal distributions of the components of a multiresponse observation. Also, to the extent that multivariate normality motivates many of the usual multivariate methods, a preliminary step for matching a body of data to such methods might be the assessment of the degree of *commonality* of the marginal distributions.

The problem to be considered here is the following: given a set of multiresponse observations which, for purposes of analysis, is viewed as a random sample from a single multivariate distribution, provide ways of assessing the degree of similarity or commonality of the marginal distributions of the components. Two types of approaches to this problem have been proposed by Gnanadesikan (1972) and will be described here. The first consists of informal graphical methods in the spirit of probability plotting techniques, while the second is based on the methods of Section 5.3 for developing data-based transformations to improve the normality of the observations.

One simple approach to the problem of assessing commonality is to ignore the multivariate nature of the observations and to study quantile-versus quantile (Q-Q) probability plots (see Wilk & Gnanadesikan, 1968, and the brief description in Section 6.2) of the observations on the individual components separately, using a common distribution (e.g., univariate normal) as the standard for comparison. Although this method can often be useful, it is not parsimonious in some ways and in certain applications may lead to findings in the separate analyses that are difficult to integrate into a cohesive overall conclusion.

A second approach, which is more parsimonious than the above one, is to calculate the individual averages of the corresponding ordered observations for all of the components (i.e., average of smallest observation on each component, average of second smallest, etc.) and to make *Q-Q* probability plots of these averages. This is merely an adaptation, to the multiresponse case, of a method proposed by Laue and Morse (1968) for studying the assumed common distribution underlying several mutually independent, but comparable, sets of univariate data. Possible noncommensurability of the components in the multiresponse case may be handled by averaging the ordered observations after standardizing the individual components. This method, too, can be quite useful in some circumstances. However, like the first approach, it also uses an external standard distribution for comparison purposes; moreover, the averaging involved is likely to mask the differences among the marginal distributions, and thus, for the purpose of assessing the degree of similarity of marginal distributions, the method may not be sufficiently sensitive.

A third approach, which avoids the drawbacks of the first two while having its own limitations, is the joint plotting of component order statistics as now described. Let the rows of the $n \times p$ matrix, $\mathbf{Y}' = ((y_{ij}))$, $i = 1, \ldots, n$; $j = 1, \ldots, p$, denote the n p-dimensional observations. The jth column of $\mathbf{Y}'$ then consists of the n observations on the jth response, and one can order these observations to obtain

$$y_{[1j]} \leqslant y_{[2j]} \leqslant \cdots \leqslant y_{[nj]}$$

for each value of j separately. The first approach mentioned above consists in obtaining, for each standard distribution chosen, p *Q-Q* plots of these p sets of ordered observations. The second approach involves obtaining n averages, $\Sigma_{j=1}^{p} y_{[ij]}/p$ for $i = 1, \ldots, n$, and studying a single *Q-Q* plot of these for every chosen standard distribution. [*Note*: One version of the first approach which would avoid the need for choosing an external standard distribution would be to plot the n ordered observations on the jth response against the ordered values of the n averages defined in the second approach. For obtaining usefully stable averages, however, this would require a reasonably large value of p and hence a very much larger value of n, which might prove to be a severe requirement in some applications.]

For the third approach, a set of n sample *multivariate order statistics* is obtained by collecting together the corresponding ordered observations,

$$\mathbf{y}'_{[i]} = (y_{[i1]}, y_{[i2]}, \ldots, y_{[ip]}) \qquad \text{for } i = 1, 2, \ldots, n.$$

A plot of these n derived points in p-space is called a *component probability plot*

(CPP for short). Actual graphical displays may be obtained for two- and three-dimensional projections of the n points, that is, for subsets of sizes two and three from among the original p variates.

The motivation for this method of intercomparing the distributions of the components of the multivariate observation is that if, in the original coordinate system, the marginal distributions are the same up to origin and scale parameters, one may expect that the combined (i.e., multivariate) order statistics will conform to a linear configuration. Departures from linearity would indicate noncommonality of the marginal distributions. The procedure does not depend on any specific distributional assumptions and is addressed to the assessment of the composite hypothesis that the marginal distributions are the same up to origin and scale. A negative indication from this analysis may suggest the need for a nonlinear transformation on one or more of the coordinates.

In practice, the procedure may be particularly relevant and revelant when all the correlations among the variates are nonnegative. One implication of this is that two-dimensional CPP's are likely to be particularly useful (since a change of sign of one of the variables will accomplish this) and may be employed for assessing the similarity of marginal distributions of bivariate observations.

For the case of two variates, when there is no dependence between the variates the CPP is just a Q-Q probability plot since one is essentially plotting one set of empirical quantiles (sample order statistics) against another. Also, it is apparent that, when there is perfect positive correlation between the two variates, the two-dimensional CPP will be an exact linear configuration. In general, as the correlation decreases toward 0, the scatter about a linear configuration may be expected to increase. A useful supplement to the CPP is to fit a straight line to the scatter of the n points in p-space by minimizing the sum of squares of perpendicular deviations of the points from the line and to compute the value of the achieved minimum orthogonal sum of squares (MOSS). The algorithms for fitting the MOSS line and computing the MOSS value are simply linear principal components analysis ones (i.e., eigenanalysis of covariance matrices of the points plotted in the component probability plot).

Noncommensurability of the p components will introduce the usual difficulties of principal components analysis. Hence, both as a more reasonable graphical scaling technique and as a way of standardizing the above fitting procedure, one can define, display, and work with a *standardized component probability plot* (SCPP), which is a component probability plot whose coordinates have been standardized to have unit variance.

The crux of the graphical nature of the CPP or SCPP is the linearity of the configuration under null conditions and the departures from linearity otherwise. In two- or three-dimensional representations the picture is an adequate conveyor of information on conformity to linearity, but in higher-dimensional (and perhaps even in three-dimensional) space one needs some summary statistics to facilitate the assessment. The covariance and correlation matrices, $\mathbf{S}_0$ and $\mathbf{R}_0$, respectively, of the points in the CPP are natural starting points. Eigenvalues, and functions derived from them, of $\mathbf{S}_0$ and/or $\mathbf{R}_0$ may also be studied. Specifically, for instance, the MOSS associated with a p-dimensional CPP (SCPP) is the sum of the $(p - 1)$ smallest eigenvalues of $\mathbf{S}_0$ ($\mathbf{R}_0$). For this and other summary statistics, it is useful to obtain some idea of their null distributions (i.e., distributions when the marginal distributions are the same up to origin and/or scale) so that some benchmarks will be available against which to compare observed values.

A different type of approach to the question of evaluating the similarity of marginal distributions can be based on the transformation techniques discussed in Section 5.3. The basic idea in this approach is, first, to transform the observations on each coordinate of a multivariate random variable so as to make the distributions of the transformed quantities more nearly the same; and, second, to intercompare the transformations, deciding that if they are in some sense identical or similar the original marginal distributions must have been equally similar. The choice of the class of transformations to be employed is an important issue. For present purposes only the power class of transformations (see Section 5.3) is considered. Specifically, one looks for a set of p transformation parameters, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$, to transform the set of observations, $\mathbf{Y}'$, to $(\mathbf{Y}^{(\lambda)})' = ((y_{ij}^{(\lambda_j)}))$, where

$$
\begin{aligned}
y_{ij}^{(\lambda_j)} &= \frac{y_{ij}^{\lambda_j} - 1}{\lambda_j} \qquad \text{for} \qquad \lambda_j \neq 0, \\
&= \ln y_{ij} \qquad\; \text{for} \qquad \lambda_j = 0,
\end{aligned}
$$

$y_{ij} > 0$; $i = 1, \ldots, n$; $j = 1, \ldots, p$. The objective of transforming the initial observations is to make the transformed observations have more nearly the same marginal distributions, and a natural choice for the common base distribution is the normal distribution. Hence the problem is to estimate the parameters, $\lambda_1, \ldots, \lambda_p$, from the data so as to enhance normality on the transformed scales (this is exactly the problem discussed earlier in Section 5.3) and then to develop methods for comparing the estimates, $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p$, to assess the reasonableness of assuming that they are all essentially estimates of a common parameter, λ. In this formulation, if indeed $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ turn out to be a cohesive set of estimates of a common parameter, it will not be unreasonable to conclude that the original marginal distributions are quite similar except possibly for differences in location and/or scale.

Corresponding to the methods of Section 5.3, there are two possibilities for specifying normality of the transformed scales, namely, improving marginal normality and enhancing joint normality. As discussed in Section 5.3, the method concerned with improving marginal normality would involve a consideration of the p log-likelihood functions, $\mathscr{L}_{\max}(\lambda_j)$, for $j = 1, \ldots, p$, defined in Eq. 77, and the associated maximum likelihood estimates, $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, as well as the p approximate confidence intervals for $\lambda_1, \lambda_2, \ldots$ and λ_p involved here (see the discussion in Section 5.3). As a procedure for assessing the similarity of marginal distributions, one can study plots of $\mathscr{L}_{\max}(\lambda_j)$ for $j = 1, \ldots, p$ on a single plot, or the p confidence intervals for $\lambda_1, \ldots, \lambda_p$, respectively, and infer the cohesiveness of the estimates. Thus, if the plots of $\mathscr{L}_{\max}(\lambda_j)$ overlap considerably, or, equivalently, if the confidence interval for λ_j includes not only $\hat{\lambda}_j$ but also $\hat{\lambda}_k$ for every $k \neq j$ and this happens for every j, one can conclude that $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ behave as if they are estimates of a common parameter, λ. [*Note*: For convenience in scaling the superimposed plots, it is desirable to plot the likelihood ratios, $L(\lambda_j)/L(\hat{\lambda}_j)$, where $L(\lambda_j) = \exp(\mathscr{L}_{\max})$, instead of the log-likelihood functions, $\mathscr{L}_{\max}$, since all the ratios have a maximum value of 1.]

Adopting the more explicitly multivariate approach, one would obtain the maximum likelihood estimates that enhance joint normality as the values $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$ that maximize the log-likelihood function, $\mathscr{L}_{\max}(\lambda_1, \ldots, \lambda_p)$, defined in Eq. 79. A simple test of the significance of the hypothesis that $\lambda_1 = \lambda_2 = \cdots = \lambda_p = \lambda$, say, may be obtained by using the approximate asymptotic theory associated with the likelihood approach. In particular, the statistic

$$2\{\mathscr{L}_{\max}(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_p) - \mathscr{L}_{\max}(\hat{\lambda}, \hat{\lambda}, \ldots, \hat{\lambda})\} \tag{80}$$

may be referred to the chi-squared distribution with $(p - 1)$ degrees of freedom. The first term within the curly brackets in Eq. 80 is just the maximum value of $\mathscr{L}_{\max}(\lambda_1, \ldots, \lambda_p)$ of Eq. 79. The second term is the maximum value of $\mathscr{L}_{\max}(\lambda_1, \ldots, \lambda_p \mid \lambda_1 = \cdots = \lambda_p = \lambda)$, which may be defined by analogy with Eq. 79 just by replacing the $y_{ij}^{(\lambda_j)}$ by $y_{ij}^{(\lambda)}$, wherein a common value λ is used in place of the separate λ_j. The second term in Eq. 80, therefore, involves just a one-dimensional maximization, whereas the first term entails a p-dimensional maximization that may require considerable computational effort. (See Chambers, 1973, for a discussion of available numerical optimization algorithms.)

The computational effort involved in the transformation approach would increase considerably if the class of transformations were to be enlarged to include the shifted-power transformation [viz., with $(y_{ij} + \xi_j)$ in place of y_{ij}], which, among other advantages, would enable one to handle negative observations as well as positive ones. Apart from this important consideration, however, once again in principle the above approach can handle the shifted-power class of transformations.

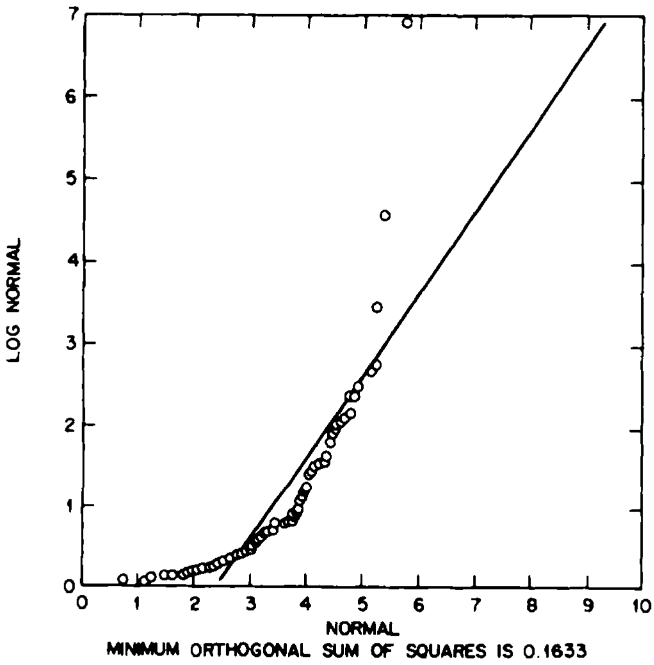
Other classes of transformations, which remain simple and yet provide additional flexibility, need to be considered. There are, of course, limitations to the transformation approach, including general conceptual ones such as the possible nontransformability of some distributions by simple classes of transformations. Also, in some circumstances, it may be misleading to conclude that the marginal distributions are similar in shape just because the power transformations of the variables are essentially identical. This is illustrated in Example 34.

The two approaches to evaluating the similarity of marginal distributions have been applied to a variety of computer-generated two- and three-dimensional data. For instance, with bivariate normal data it was found repeatedly (and comfortingly) that both the graphical technique and the transformation test led to no striking or significant departures from null expectations (see Gnanadesikan, 1972, for a typical example of this sort). The performances of the techniques under nonnull conditions would, of course, be more interesting to study, and the next three examples illustrate specific aspects of the two approaches as they are revealed in the context of particular types of departures from the case of similar marginal distributions. For simplicity of discussion and display, the data in each of these examples are two-dimensional.

Example 32. The data for this example are a computer-generated sample of 100 bivariate observations in which one component has a standard normal distribution and the other an independent lognormal distribution, $\Lambda(0, 1)$, in the notation of Aitchison & Brown (1957, p. 7). All of the observations on the first coordinate were shifted to make them positive so as to allow the use of the simple power transformation.

Exhibit 32 shows the SCPP for this example. The MOSS value here is 0.16, which is about an order of magnitude larger than the typical values observed

Exhibit 32. SCPP of lognormal vs. normal



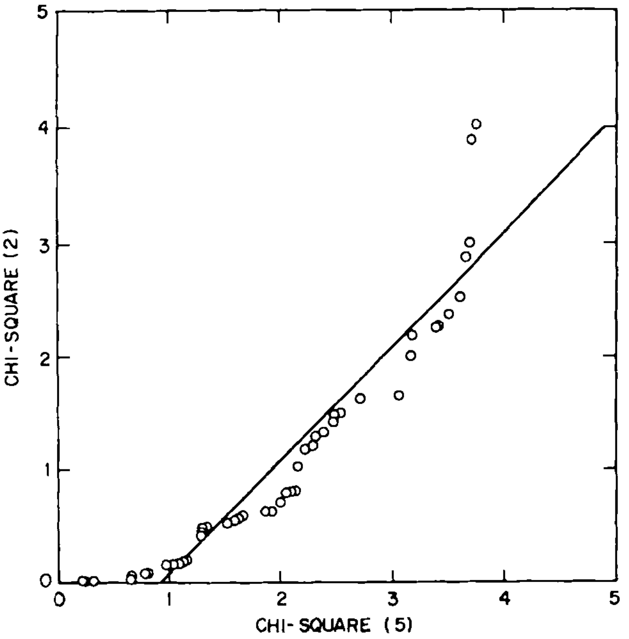
with bivariate normal data (see Example 1 of Gnanadesikan, 1972). The departure from linearity is striking and clearly suggests the extreme dissimilarity of the two marginal distributions.

The value of the log-likelihood ratio test statistic defined in Eq. 80 turns out to be 16.06 in this example. The associated probability of exceedance in the $\chi^2_{(1)}$ distribution is 6×10^{-5}, indicating a highly significant departure from commonality.

Example 33. This example is based on a bivariate subset of trivariate data in which one component has a $\chi^2_{(2)}$ distribution, another has an independent $\chi^2_{(3)}$ distribution, and the third component is derived as the sum of the first two components, so that it has a $\chi^2_{(5)}$ distribution that is not independent of the first two distributions. The value of n is 50, and the subset chosen is the $[\chi^2_{(2)}, \chi^2_{(5)}]$ combination. Exhibit 33 shows the SCPP for this example, together with the fitted straight line and the associated MOSS value of 0.06. The systematically curved nature of the configuration on this plot would suggest dissimilarity of the marginal distributions.

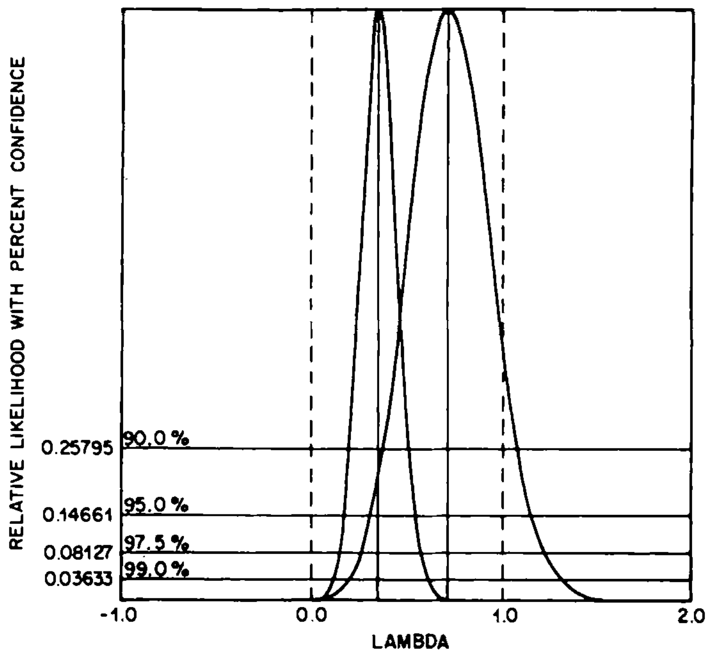
On the other hand, the statistic defined in Eq. 80 turns out, in this example, to have the value 2.14, which is exceeded in $\chi^2_{(1)}$ distribution with a probability of 0.14, thus suggesting no highly significant departure from commonality. The estimated values of λ_1 and λ_2, in this bivariate transformation approach, are 0.41 and 0.69, respectively, while the estimate of λ (the hypothesized common

Exhibit 33a. SCPP of $\chi^2(2)$ vs. $\chi^2(5)$



MINIMUM ORTHOGONAL SUM OF SQUARES IS 0.0608

Exhibit 33*b*. Superimposed likelihood-ratio plots with associated confidence intervals



value of λ_1 and λ_2) is 0.49. The corresponding univariate transformation approach leads to the estimates $\hat{\lambda}_1 = 0.33$ and $\hat{\lambda}_2 = 0.73$, and Exhibit 33*b* shows the superimposed plots of $L(\lambda_j)/L(\hat{\lambda}_j)$, $j = 1, 2$, with approximate confidence intervals for λ_1 and λ_2 also displayed on the figure. Although the univariate approach tends to pull the transformations of the two variables apart to a greater degree than does the bivariate approach, the overall indication from both approaches is of a moderate but not very strong difference in the two marginal distributions in this example.

In this example, therefore, the graphical display via the SCPP tends to be more revealing than the more formal test of significance based on the transformation approach. The next example brings out the same result even more forcefully.

Example 34. The bivariate data for this example consist of 100 observations simulated to be a random sample from the bivariate lognormal distribution, $\Lambda\,[\mu, \Sigma]$, where

$$\mu' = (0, 5) \quad \text{and} \quad \Sigma = \begin{pmatrix} 1 & 2.7 \\ & 9 \end{pmatrix}.$$

Exhibit 34*a* shows the SCPP for this example, and the departure from linearity is very striking. The MOSS value for this SCPP is 0.25.

Exhibit 34*a*. SCPP for bivariate lognormal data

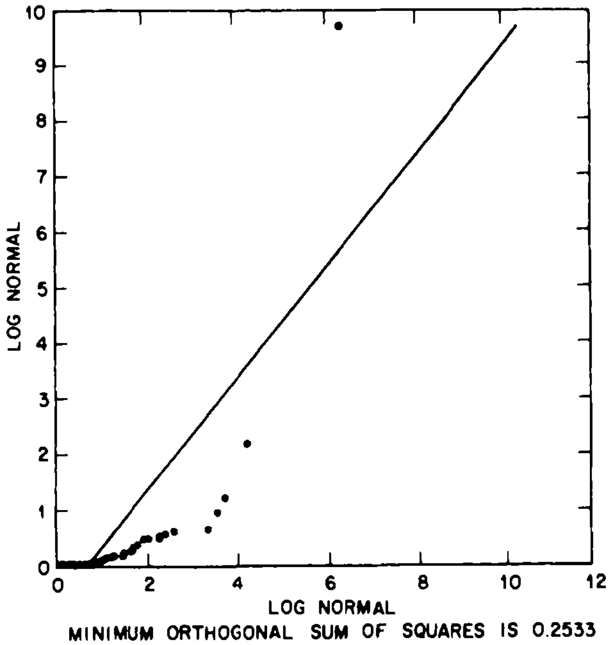
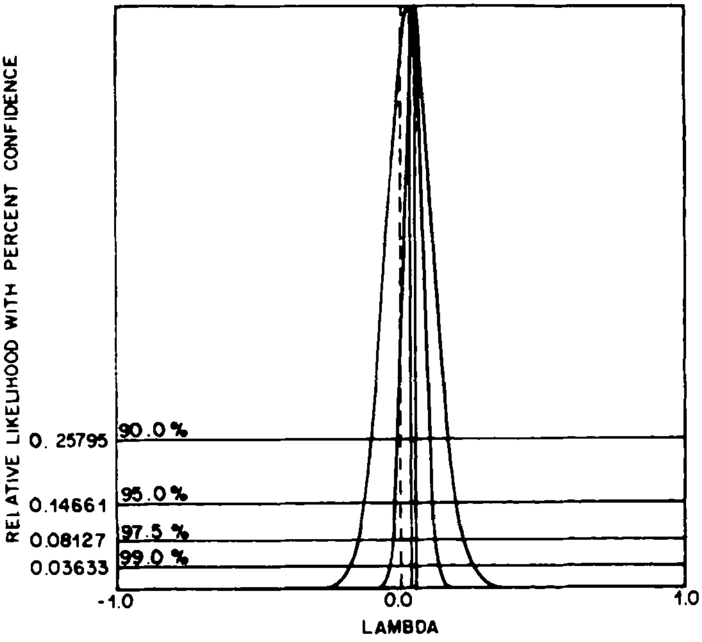


Exhibit 34*b*. Superimposed likelihood ratio plots with associated confidence intervals



The results of using the transformation approach, however, are totally unrevealing in this example. The bivariate approach leads to the estimates $\hat{\lambda}_1 = 0.02$, $\hat{\lambda}_2 = 0.04$, and $\hat{\lambda} = 0.03$, and the value of the log-likelihood ratio statistic is 0.07, with an associated exceedance probability of 0.79. The univariate transformation approach yields $\hat{\lambda}_1 = 0.03$ and $\hat{\lambda}_2 = 0.05$, and Exhibit 34*b* shows the superimposed plots of $L(\lambda_j)/L(\hat{\lambda}_j)$, $j = 1, 2$. The closeness of the estimates of λ_1 and λ_2 and the considerable overlapping of the curves in Exhibit 31*b* sdhould be anticipated in this example, since, although the two lognormal distributions are distinctly different in shape (as judged by the difference between the diagonal elements of Σ above), the power transformation needed to transform both lognormal distributions to normal distributions is just the logarithmic one, that is, the one corresponding to $\lambda = 0$. In fact, all the above estimates of λ_1 and λ_2 are statistically close to this zero value.

This example thus illustrates a limitation of the transformation approach in that the closeness of the transformations (within a class such as the power one) required to enhance normality of the marginal distributions is not a sufficient condition for similarity of the distributions of the untransformed variables.

5.4.2. Methods for Assessing Normality

The assumption of multivariate normality underlies much of the standard "classical" multivariate statistical methodology. The effects of departures from normality on the methods are not easily or clearly understood. Moreover, for analyzing multiresponse data, while techniques that are resistant to outliers are currently available (see Section 5.2.3), others that are more generally robust against a variety of departures from idealized models are still at a relatively nascent stage, especially in terms of experience in using them. Thus it would be useful to have procedures for verifying the reasonableness of assuming normality for a given body of multiresponse observations. If available, such a check would be helpful in guiding the subsequent analysis of the data, perhaps by suggesting the need for and nature of a transformation of the data to make them more nearly normally distributed, or perhaps by indicating appropriate modifications of the models and methods for analyzing the data.

Not only is there a paucity of multivariate nonnormal distributional models, but also most of the proposed alternative distribution (e.g., multivariate lognormal, exponential) are defined so as to have properties that are similar to those of the multivariate normal (e.g., that all marginal distributions belong to the same class). Real data will, of course, not necessarily conform to such specialized forms of multivariate nonnormality.

With multiresponse data it is clear that the possibilities for departure from joint normality are indeed many and varied. One implication of this is the need for a variety of techniques with differing sensitivities to the different types of departures: seeking a single best method would seem to be neither pragmatically sensible nor necessary. Developing several techniques and enabling an accumulation of experience with, and insight into, their properties is a crucial first step. Aitkin (1972), Andrews et al. (1973), and Malkovich & Afifi (1973) have proposed different methods for assessing normality, and the discussion in this subsection draws heavily from the work of Andrews and his colleagues. Mardia (1980) provides a survey of various tests for normality.

One way of seeing the need for a variety of techniques in the multivariate case is in terms of the degree of commitment one wishes to make to the coordinate system for the multiresponse observations. (See also the discussion of this issue in Section 5.2.3 regarding robust estimates of multivariate location.) At one extreme is the situation in which interest is completely confined to the observed coordinates. In this case the marginal distributions of each of the observed variables and conditional distributions of certain of these, given certain others, will be the objects of interest. On the other hand, the interest may lie in the original coordinates as well as all possible orthogonal transformations of them, and here summaries (such as Euclidean distance) that remain invariant under orthogonal transformations will be the ones of interest. More generally, the class of all nonsingular linear transformations of the observed variables may be the one of interest, and then affine invariance will guide the analysis. Aside from linear transformations, one may sometimes be willing to make simple nonlinear transformations (perhaps of each coordinate separately) so as to be able to use simple models and techniques. In this case the methods used should reflect an awareness of this degree of flexibility and should attempt to incorporate it statistically. Much of the formal theory of multivariate analysis has been concerned solely with affine invariance, thus limiting the class of available procedures. The present subsection will consider techniques that are applicable to situations with different degrees of commitment to the observed coordinate system, including the classical one requiring affine invariance.

Another important issue with multivariate techniques is that, although some complexity of the methods is to be expected, they should, if possible, be kept computationally economical. The feasibility of extensive computing, made easily accessible by modern computers, does not imply that every technique is economically tenable. One objective used in developing the methods to be described below was that, computationally, they be reasonably economic and efficient.

The methods for assessing normality to be discussed here may be grouped under the following headings: (i) univariate techniques for evaluating marginal normality; (ii) multivariate techniques for evaluating joint normality; and (iii) other procedures based on unidimensional views of the multiresponse data. As mentioned earlier (see also Section 5.3), performing an initial transformation on the data and then using "standard" methods of analysis constitute a prevalent and often useful approach in analyzing data. Hence, as a general approach under each of the three categories of methods mentioned above, the assessment of normality may be made by inquiring about the need for a transformation. However, an approach that is not explicitly dependent on data-based transformations is also possible. Techniques of both types are discussed below.

Evaluating Marginal Normality. In practice, a single overall multivariate analysis of data is seldom sufficient or adequate by itself, and almost always it needs to be augmented by analyses of subsets of the responses, including univariate analyses of each of the original variables. Although marginal normality does not imply joint normality, the presence of many types of nonnormality is often reflected in the marginal distributions as well. Hence a natural, simple, and preliminary step in evaluating the normality of multiresponse data is to study the reasonableness of marginal normality for the observations on each of the variables. For this purpose one can use a variety of well-known tests for univariate normality, some of which are described next.

Perhaps the most classical method of evaluating the normality of the univariate observations $X_1, \ldots, X_n$ is by means of the well-known skewness and kurtosis coefficients:

$$\sqrt{b_1} = \frac{\sqrt{n} \sum_{i=1}^{n} (X_i - \bar{X})^3}{\left\{ \sum_{i=1}^{n} (X_i - \bar{X})^2 \right\}^{3/2}},$$

$$b_2 = \frac{n \sum_{i=1}^{n} (X_i - \bar{X})^4}{\left\{ \sum_{i=1}^{n} (X_i - \bar{X})^2 \right\}^{2}}.$$

Tables of approximate 5% and 1% points of these two statistics may be found in Pearson & Hartley (1966, pp. 207–208).

D'Agostino & Pearson (1973) have proposed improved schemes for using $\sqrt{b_1}$ and b_2 to test normality rather than employing these coefficients directly. Specifically, they provide (i) graphs (based on extensive computer simulations) for calculating the empirical probability integral of b_2 (under the null hypothesis of sampling from a normal) for a specified sample size n ($20 \leqslant n \leqslant 200$), and (ii) a table for calculating a standardized normal equivalent deviate $X(\sqrt{b_1})$ corresponding to $\sqrt{b_1}$—the table gives values of δ and $1/\lambda$ for use in the definition, $X(\sqrt{b_1}) = \delta \sinh^{-1}(\sqrt{b_1}/\lambda)$, for values of $n = 8(1)50(2)100(5)250(10)500(20)1000$.

Given an observed couplet of values $\sqrt{b_{1,0}}$ and $b_{2,0}$ derived from a sample of n_0 observations, one may use the graphs of the empirical probability integral of b_2 to obtain a value of the cumulative probability, $P(b_{2,0} \mid n_0) = P(b_2 \leqslant b_{2,0} \mid n = n_0)$, and then the equivalent standard normal deviate, $X(b_{2,0})$, corresponding to this probability. Also, the table of values of δ and $1/\lambda$ can be used to calculate $X(\sqrt{b_{1,0}})$. These standard normal deviates, $X(\sqrt{b_{1,0}})$ and $X(b_{2,0})$, may be utilized individually for testing skewness and kurtosis departures, and, in addition, they can be combined into a single omnibus test statistic,

$$\chi^2_{(2)} = X^2(\sqrt{b_{1,0}}) + X^2(b_{2,0}),$$

which can be referred to a chi-squared distribution with 2 degrees of freedom. D'Agostino & Pearson (1973) also suggest a second omnibus test based on tail probabilities rather than the equivalent normal deviates, but they state that the two tests are likely to produce very similar results.

Shapiro & Wilk (1965) have suggested a different omnibus test for normality that has appealing power properties including generally good sensitivity to a wide variety of alternatives to the normal (see Shapiro et al., 1968). The statistic proposed for assessing the univariate normality of $X_1, \ldots, X_n$ is

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2},$$

where $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$ denote the ordered observations, and the unit-length vector $\mathbf{a}' = (a_1, \ldots, a_n)$ is defined in terms of the vector of expected values, $\mathbf{m}'$, of standard normal order statistics and their covariance matrix, $\mathbf{V}$, as

$$\mathbf{a}' = \frac{\mathbf{m}'\mathbf{V}^{-1}}{\|\mathbf{m}'\mathbf{V}^{-1}\|}.$$

The numerator of W is, except for a multiplicative constant, the square of the best linear unbiased estimate of the standard deviation from the order statistics of a sample assumed to be from a normal population (see Sarhan & Greenberg, 1956, Section 10C), and the denominator is, of course, $(n - 1)$ times the usual unbiased estimate of the variance. Shapiro & Wilk (1965) provide tables of values of the coefficients $\{a_i\}$ for $n = 2(1)50$.

Small values of W correspond to departure from normality, and percentage points are given by Shapiro & Wilk for $n = 3(1)50$.

For handling $n > 50$ without extensive tabulation of coefficients (or percentage points), D'Agostino (1971) has proposed an alternate test statistic,

$$D = \frac{T}{n^{3/2}\left\{\sum_{i=1}^{n} (X_i - \bar{X})^2\right\}^{1/2}},$$

where

$$T = \sum_{i=1}^{n} \left[i - \frac{n+1}{2}\right] X_{(i)}$$

is essentially Gini's mean difference and also, except for the multiplicative constant $2\sqrt{\pi}/n(n - 1)$, the estimator of the standard deviation of a normal distribution proposed by Downton (1966). Thus D is a constant times the ratio of two estimates of the standard deviation, and both large and small deviations from its expected value correspond to departures from normality. D'Agostino (1971) gives a brief table of percentage points of a standardized version of D for sample sizes up to 1000.

For moderately large samples another simple test for normality has been proposed by Andrews et al. (1972). The test is based on the normalized gaps,

$$g_i = \frac{X_{(i+1)} - X_{(i)}}{m_{i+1} - m_i}, \qquad i = 1, \ldots, (n-1),$$

where $\mathbf{m}' = (m_1, \ldots, m_n)$, as before, is the vector of expected values of standard normal order statistics.

If the distribution of X is normal with mean μ and variance σ^2, the g_i will be approximately independently and exponentially distributed with scale parameter σ. Under an alternative, the configuration of the ordered observations may be expected to depart from the m_i with a corresponding effect on the configuration of the g_i. One approach for studying relatively smooth departures from the null configuration of the g_i proceeds via an examination of means of adjacent g_i. Specifically, one can compute sums of the first quarter, the middle half, and the last quarter of the g_i:

$$S_L = \sum_{i=1}^{[(n-1)/4]} g_i, \qquad S_M = \sum_{i=[(n-1)/4]}^{[3(n-1)/4]} g_i, \qquad S_U = \sum_{i=[3(n-1)/4]}^{n-1} g_i.$$

Let n_1 be the number of normalized gaps involved in S_L and S_U, and n_2 the number involved in S_M, so that $2n_1 + n_2 = (n-1)$. Then, under null conditions,

$$g_L = \frac{1}{n_1} S_L \qquad \text{and} \qquad g_U = \frac{1}{n_1} S_U$$

will have mean σ and variance σ^2/n_1, while

$$g_M = \frac{1}{n_2} S_M$$

has mean σ and variance σ^2/n_2.

On the basis of the approximate exponential distribution of the normalized gaps, the ratios $r_L = g_L/g_M$ and $r_U = g_U/g_M$ will each have an F distribution with degrees of freedom $2n_1$ and $2n_2$. Thus, for large n, r_L and r_U may each be treated as approximately normal with mean 1 and variance $(1/n_1 + 1/n_2)$. Also, a test statistic (distributed approximately as chi-squared with 2 degrees of freedom) which will tend to be more omnibus by combining the information in r_L and r_U is the quadratic form in $(r_L - 1)$ and $(r_U - 1)$ with compounding matrix $n_1\mathbf{I} - [n_1^2/(2n_1 + n_2)]\mathbf{J}$, where $\mathbf{I}$ is the identity matrix and $\mathbf{J}$ is a matrix of unities. With $n_2 = 2n_1$, this $\chi^2_{(2)}$ statistic is

$$q = \frac{n_1}{4}\{3(r_L - 1)^2 - 2(r_L - 1)(r_U - 1) + 3(r_U - 1)^2\}.$$

Thus, three statistics, r_L, r_U, q, together with approximate significance levels, may be calculated. The statistics r_L and r_U may be useful in interpreting and acting on significant nonnormality detected by the more omnibus q.

In addition to the direct tests for univariate normality discussed thus far, on can inquire into tests based on transforming the data. One such test can be developed in conjunction with the method proposed by Box & Cox (1964) for estimating shifted-power transformations,

$$X \rightarrow X^{(\xi,\lambda)} = \begin{cases} [(X + \xi)^\lambda - 1]/\lambda & \text{for } \lambda \neq 0, \\ \ln(X + \xi) & \text{for } \lambda = 0. \end{cases}$$

The estimation problem and an approach of Box & Cox to it was discussed in Section 5.3, where the detailed development was presented for the case in which ξ, the shift parameter, is taken to be 0. For the purpose of deriving an associated test for univariate normality, using both the shift and power parameters would seem to be more advantageous than using just the power parameter, λ. Basically, λ appears to be sensitive to skewness, whereas ξ seems to respond to kurtosis and heavy-tailedness. Also, in the univariate situation, including ξ implies a two-parameter effort in computational aspects, and this is not too difficult to handle.

Thus, if $X_1, X_2, \ldots, X_n$ are univariate observations, which are to be transformed by a shifted-power transformation of the above form so as to improve normality on the transformed scale, then, following Box & Cox (1964) and essentially the same steps as outlined in Section 5.3, one can obtain a log-likelihood function (initially maximized with respect to the mean and the variance for given ξ and λ) quite analogous to the one in Eq. 77 of Section 5.3:

$$\mathscr{L}_{\max}(\xi, \lambda) = -\frac{n}{2}\ln\hat{\sigma}^2 + (\lambda - 1)\sum_{i=1}^{n}\ln(X_i + \xi),$$

where $\hat{\sigma}^2$, a function of both ξ and λ, is the maximum likelihood estimate of the variance of the presumed normal distribution of the transformed observations, for example,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}[X_i^{(\xi,\lambda)} - \bar{X}^{(\xi,\lambda)}]^2, \qquad \bar{X}^{(\xi,\lambda)} = \frac{1}{n}\sum_{i=1}^{n}X_i^{(\xi,\lambda)}$$

for an unstructured sample. The above log-likelihood function may be maximized to obtain the maximum likelihood estimates, $\hat{\xi}$ and $\hat{\lambda}$, and approximate asymptotic theory yields a $100(1-\alpha)\%$ confidence region for ξ and λ, defined by

$$2\{\mathscr{L}_{\max}(\hat{\xi},\hat{\lambda}) - \mathscr{L}_{\max}(\xi,\lambda)\} \leqslant \chi_2^2(\alpha),$$

where $\chi_2^2(\alpha)$ denotes the upper $100\alpha\%$ point of a chi-squared distribution with 2 degrees of freedom. A simple transformation-related procedure for assessing the normality of the distribution of X consists in not rejecting (at a $100\alpha\%$ level of significance) the hypothesis of normality if the above confidence region overlaps with the line $\lambda = 1$. A related, more stringent "likelihood-ratio test" would consist of comparing the value of $2\{\mathscr{L}_{\max}(\hat{\xi},\hat{\lambda}) - \mathscr{L}_{\max}(\hat{\xi},1)\}$ to a chi-squared distribution with 1 degree of freedom. [Note that $\mathscr{L}_{\max}(\xi,1)$ is independent of ξ so that any value, including $\hat{\xi}$, maximizes $\mathscr{L}_{\max}(\xi,1)$ as a function of ξ.]

When the observations on the variable X are structured (i.e., some design or regression structure underlies the observations), Andrews (1971) has proposed exact procedures (confidence regions as well as tests) for formal inferences regarding ξ and λ, and one can use these in place of the approximate procedures described above. In many applications the conclusions from using the exact procedures are not likely to be markedly different from those arrived at by the approximate methods.

The preceding discussion has been concerned with formal tests of significance for detecting departures from univariate normality. For data-analytic purposes, plotting on normal probability paper or making a normal Q-Q (quantile-versus-quantile) probability plot (see Section 6.2) is often a very useful method of assessing the univariate normality of observations. The technique consists in plotting the ordered observation, $X_{(i)}$, against the quantile, q_i, of the standard normal distribution corresponding to the cumulative probability $(i-\frac{1}{2})/n$ (or $i/n + 1$ or similar fraction) for $i = 1, \ldots, n$. [*Note*: $q_i = \Phi^{-1}(p_i)$, that is, q_i is defined by the equation,

$$\int_{-\infty}^{q_i} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}t^2)\, dt = p_i,$$

where $p_i = (i-\frac{1}{2})/n$ or similar fraction.] A linear configuration on such a plot would correspond to adequate normality of the observations, while systematic and subtle departures from normality would be indicated by deviations from linearity.

Although a normal probability plot does not provide a single-statistic-based formal test, as a graphical tool it conveys a great deal more information about the configuration of the observations than any single summary statistic is likely to do. In fact, one motivation for the W statistic of Shapiro & Wilk (1965) mentioned earlier is that it provides a comparison of the square of the slope of a normal probability plot of the observations against the usual estimate of variance and hence is directed towards assessing the linearity of such a plot. Devising tests directed toward detecting specific departures from linearity (e.g., quadratic or cubic) would be natural extensions of the W test. Also, Filliben (1975) has proposed a test based on the correlation coefficient from the normal probability plot, that is, for the points $(q_i, X_{(i)})$, $i = 1, \ldots, n$. The normal probability plot is, however, likely to be a valuable supplement to any single test procedure.

A graphical display of the normalized gaps is also possible. A plot of g_i versus i, for $i = 1, \ldots, (n - 1)$, should appear as a random horizontal scatter revealing no systematic patterns or extremely deviant observations, provided that the original data are reasonably normally distributed. Under several nonnormal alternatives, the g_i have expected values that deviate smoothly but noticeably in the tails, and this will show up as deviations from horizontality at the left and right ends of the plot of g_i versus i. To reduce the "noisy" appearance, some smoothing of such a plot may be helpful. Exponential probability plots of the normalized gaps [i.e., a plot of the ith ordered value, $g_{(i)}$, versus the "corresponding quantile," viz., the quantile for a fraction such as $(i - \frac{1}{2})/n$, for the exponential distribution] can also be made and studied. Another variant is to make a normal probability plot of the cube roots of the normalized gaps.

Evaluating Joint Normality. In practice, except for rare or pathological examples, the presence of joint nonnormality is likely to be detected quite often by methods directed at studying the marginal normality of the observations on each variable. However, there is a need for tests that explicitly exploit the multivariate nature of the data in order, it is hoped, to yield greater sensitivity. Some methods addressed to this need are discussed next.

Classical goodness-of-fit tests, such as the chi-squared and Kolmogorov-Smirnov tests, would be possibilities for use in testing for multivariate normality. However, the drawbacks of these tests in univariate circumstances (e.g., choice of the number and boundaries of cells for the chi-squared test) are likely to be magnified for the multivariate case, and this may be part of the reason for the noticeable lack of use of these procedures with multivariate data. Also, Weiss (1958) and Anderson (1966) have suggested tests based on local densities of the observations, but perhaps because of the difficulty of the computations involved, neither of these has seen wide application.

A relatively simple test, called the *nearest distance test*, has been proposed by Andrews et al. (1972) for testing joint normality. In this test nearest neighbor distances for each point are transformed through a series of steps to standard normal deviates. Under the null hypothesis these transformed distances are independent of the coordinates of points from which they are measured. This independence may be tested by multiple regression techniques.

The first step in the procedure is to transform the data to the unit hypercube, using the sample version of a transformation discussed by Rosenblatt (1952). One way of implementing the transformation is to initially transform the observations, using the sample mean vector and covariance matrix so as to make the transformed data have zero mean and identity covariance matrix. Then one applies the probability integral transformation to each "observation" on each coordinate separately, using the standard normal distribution as the null basis for the probability integral transformation. [*Note*: The degree of nonuniformity in small samples, resulting from using the univariate probability integral transformation with estimated values of the parameters substituted for the parameters, has been studied by David & Johnson (1948).] For adequately large sample sizes, it is perhaps not unreasonable to expect the data, if they conform to the null hypothesis of joint normality, to be transformed to the unit hypercube by this means. Also, for large sample sizes (>50 when p is small), the occurrence of points in disjoint parts of this space may be usefully approximated by independent Poisson events.

For each point $\mathbf{x}_i$ in this hypercube, a nearest neighbor distance may be calculated by using the metric

$$d(\mathbf{x}_i, \mathbf{x}_j) = \max\{\min[|x_{ik} - x_{jk}|, ||x_{ik} - x_{jk}| - 1|]\}.$$

[*Note*: To avoid boundary effects, the metric "wraps around" opposite faces of the unit hypercube. Other ways of handling this problem may also be worth considering.] Other metrics, such as the Euclidean one, may also be used. However, with moderate-sized samples, many distances have to be calculated, and the above metric is relatively inexpensive to compute. It seems well suited to algorithms that make use of sorted arrays of each coordinate.

The volume of the set enclosed by a distance d from the point $\mathbf{x}_i$,

$$\{\mathbf{x}: d(\mathbf{x}_i, \mathbf{x}) \leqslant d\},$$

is given by

$$V(d) = (2d)^p.$$

Since the points are assumed to be uniformly distributed in the space, the variable $V(d)$, where d is the distance to the nearest neighbor, has an exponential distribution, and

$$P[V(d) \leqslant V(d_i)] = 1 - \exp\{-\lambda V(d_i)\}.$$

Conditionally, given that $d \leqslant d_0$, the probability

$$p(d_i) = P[V(d) \leqslant V(d_i) \,|\, d_i \leqslant d_0]$$
$$= \frac{1 - \exp\{-\lambda V(d_i)\}}{1 - \exp\{-\lambda V(d_0)\}}.$$

To this probability there corresponds a standard normal deviate,

$$w_i = \Phi^{-1}\{p(d_i)\}.$$

If the w_i are calculated from disjoint parts of the unit p-cube, they should not show any dependence on $\mathbf{x}_i$, the coordinates of the center from which nearest neighbors are measured. Such dependence may be tested by examining the regression sum of squares associated with fitting to w a quadratic surface in the elements of $\mathbf{x}$. Under the null hypothesis this regression sum of squares has a chi-squared distribution with $(p + 1)(p + 2)/2$ degrees of freedom. Using this distribution, one may readily assess the significance level associated with the observed regression sum of squares. If only a first-order (i.e., linear in elements of $\mathbf{x}$) model is used, the degrees of freedom are $(p + 1)$.

For the $n \times p$ multiresponse data matrix $\mathbf{Y}'$, whose rows $\mathbf{y}_i'$ $(i = 1, \ldots, n)$ are taken for simplicity of discussion to constitute an unstructured sample, the computations involved in the nearest distance test are outlined by the following steps:

1. Compute the sample mean vector, $\bar{\mathbf{y}}$, and covariance matrix, $\mathbf{S}$; obtain the sphericized residuals, $\mathbf{z}_i = \mathbf{S}^{-1/2}(\mathbf{y}_i - \bar{\mathbf{y}})$; and, if z_{ij} denotes the jth element of $\mathbf{z}_i$, compute the standard normal probability integral value, $x_{ij} = \Phi(z_{ij})$, $i = 1, \ldots, n$; $j = 1, \ldots, p$. Let $\mathbf{x}_i$ denote the p-dimensional vector whose jth element is x_{ij}.

2. Calculate the distances

$$d(i, i') = \max_k \left[\min\{|x_{ik} - x_{i'k}|, ||x_{ik} - x_{i'k}| - 1|\}\right]$$

and

$$d_{\min}(i) = \min_{i' \neq i} d(i, i').$$

3. For each point $\mathbf{x}_i$, if $d_{\min}(i) < 1/2n^{1/p}$ and if $d(i, i') > 1/2n^{1/p}$, $i' < i$, calculate

$$w_i = \Phi^{-1}\left[\frac{1 - \exp\{-n[2d_{\min}(i)]^p\}}{1 - \exp\{-1\}}\right].$$

4. For the $\mathbf{x}_i$ used in step 3, regress w_i on 1, $x_{i1}, \ldots, x_{ip}, x_{i1}^2, \ldots, x_{ip}^2, x_{i1}x_{i2}, \ldots, x_{i(p-1)}x_{ip}$; that is, fit the quadratic relationship $\mathscr{E}(w) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \beta_{11}x_1^2 + \cdots + \beta_{pp}x_p^2 + \beta_{12}x_1x_2 + \cdots + \beta_{(p-1)p}x_{p-1}x_p$, using the n' [$\leq n$ and, it is hoped, $> (p + 1)(p + 2)/2$] points that survive step 3, and thus obtain a regression sum of squares with $(p + 1)(p + 2)/2$ degrees of freedom.

5. Compare the obtained value of the regression sum of squares to a chi-squared distribution with $(p + 1)(p + 2)/2$ degrees of freedom, rejecting joint normality for large values of the regression sum of squares.

Just as the univariate transformation approach of Box & Cox (1964) was utilized to obtain a test of marginal normality, the transformation approach of Andrews et al. (1971) directed toward enhancing the joint normality of multiresponse data (see Section 5.3) may be used for providing a transformation-related test of multivariate normality. The essential idea in a transformation-related approach is that evidence suggesting that a nonlinear transformation is required to significantly improve joint normality is considered as evidence that the untransformed data are nonnormal. (See, however, the discussion near the end of this subsection regarding a limitation of this formulation.)

For present purposes, even when p is not larger than 2, in order to keep the computational effort down and also be able to display some of the analyses graphically, the transformations actually employed are just power transformations of each variable separately, namely, $Y_j^{\lambda_j}$, with no shift parameters involved. (See, however, some of the earlier comments and the discussion of Example 36 for possible limitations imposed by not including shift parameters.)

For the power family, the linear transformation $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)' = \mathbf{1}$ is the only transformation consistent with the hypothesis that the data are normally distributed. A likelihood-ratio test of the hypothesis $\boldsymbol{\lambda} = \mathbf{1}$ may be based on the asymptotically approximate $\chi^2_{(p)}$ distribution of

$$2\{\mathscr{L}_{\max}(\hat{\boldsymbol{\lambda}}) - \mathscr{L}_{\max}(\mathbf{1})\},$$

where $\mathscr{L}_{\max}(\boldsymbol{\lambda})$ is the log-likelihood function defined in Eq. 79 of Section 5.3, and $\hat{\boldsymbol{\lambda}}$ is the value of $\boldsymbol{\lambda}$ that maximizes $\mathscr{L}_{\max}(\boldsymbol{\lambda})$. This $\chi^2_{(p)}$ distribution may be used to obtain both a significance level, α, associated with the observed $\hat{\boldsymbol{\lambda}}$ and a confidence set for $\boldsymbol{\lambda}$. In that the estimation method discussed earlier in Section 5.3 is built into this procedure, it not only indicates when data are nonnormal—which we may be willing to grant for many large samples—but also suggests data transformations that may be used to enhance normality.

The discussion heretofore of methods for assessing joint normality was oriented toward numerical rather than graphical techniques. For evaluating univariate normality, normal probability plots were mentioned as having particular appeal as a graphical aid in analyzing data. For evaluating joint normality, Andrews et al. (1973) have suggested an informal graphical procedure that utilizes a *radius-and-angles* representation of multiresponse data.

The first step in conceptualizing the method, in the simple context of an unstructured sample, is to obtain the sphericized residuals

$$\mathbf{z}_i = \mathbf{S}^{-1/2}(\mathbf{y}_i - \bar{\mathbf{y}}), \qquad i = 1, \ldots, n,$$

which were defined and used in the nearest distance test discussed earlier. Under the null hypothesis the sphericized residuals are approximately spherically symmetrically distributed. The squared radii, or squared lengths of the $\mathbf{z}_i$,

$$r_i^2 = \mathbf{z}_i'\mathbf{z}_i = (\mathbf{y}_i - \mathbf{y})'\mathbf{S}^{-1}(\mathbf{y}_i - \mathbf{y}),$$

will have approximately a chi-squared distribution with 2 degrees of freedom in the bivariate case (and p degrees of freedom in the p-variate case). Also, in the bivariate case the angle θ_i that $\mathbf{z}_i$ makes with, say, the abscissa direction will be approximately uniformly distributed over $(0, 2\pi)$. All quantities, namely, the r_i^2's and the θ_i's, will be approximately independent for large n. The dependence enters, among other routes, via the estimates of the mean and the covariance matrix, and it is hoped that for adequately large samples this dependence will have no serious effects. A further comment which may be in order is that the exact *marginal* distribution of r_i^2 is known to be a constant multiple of a beta rather than a chi-squared distribution; but again, even for moderate samples (i.e., $n = 20$ or 25 in the bivariate case), the difference between using the beta and the chi-squared approximation appears to be insignificant (see Gnanadesikan & Kettenring, 1972).

The properties mentioned above suggest that summaries in terms of radii and angles may be useful for assessing joint normality. Indeed some authors (e.g., Healy, 1968; Kessel & Fukunaga, 1972) have suggested procedures based purely on the squared radii. The simple graphical procedures to be described next are based on both radii and angles.

In the bivariate case the procedure is to make a $\chi^2_{(2)}$ probability plot of the r_i^2 and a uniform probability plot of the normalized form of θ_i, namely, $\theta_i^* = \theta_i/2\pi$. (See Section 6.2 for a brief discussion of probability plots.) Specifically, the n squared radii, r_i^2 $(i = 1, \ldots, n)$, are ordered in magnitude, and the ith-ordered value is plotted against the quantile of a $\chi^2_{(2)}$ distribution corresponding to a cumulative probability of $(i - \frac{1}{2})/n$, for $i = 1, \ldots, n$. Also, the n values of the normalized angles, θ_i^* $(i = 1, \ldots, n)$, are ordered, and the ith-ordered value is plotted against $(i - \frac{1}{2})/n$, for $i = 1, \ldots, n$. If the data conform statistically to the null hypothesis of bivariate normality, the configurations on these two probability plots should be reasonably linear. Departures from linearity on either or both of the plots would indicate specific types of departure from null conditions. [*Note*: The origin on the plot of the θ_i^* is arbitrary. Also, the θ_i^* that correspond to large r_i^2 may be more statistically stable than those with very small r_i^2, and therefore one may wish to "trim" the observations with the smallest values of r_i^2 and to study an appropriate uniform probability plot of the θ_i^* only for the remaining observations.]

For bivariate data one can also combine the information in the radii and angles in a single two-dimensional display. Let u_i denote the probability integral transformation of r_i^2 based on a $\chi^2_{(2)}$ distribution of the latter, that is $u_i = P\{\chi^2_{(2)} \leqslant r_i^2\}$ for $i = 1, \ldots, n$. Then a plot of the n points whose coordinates are (u_i, θ_i^*), $i = 1, \ldots, n$, may be made. Under the null hypothesis one would expect to get a uniform scatter of points on the unit square. Nonuniformity of scatter, or indication of any relationship between the two coordinates in the plot, would suggest departures from the null hypothesis. [*Note*: Formal tests for uniformity can also be made; however, the main value and appeal of the procedure is its graphical character.]

For higher-dimensional data (say, p-dimensional with $p > 2$), the radius-and-angles representation in terms of the elements of the sphericized residual $\mathbf{z}_i$ ($i = 1, \ldots, n$) is:

$$\begin{aligned}
z_{i1} &= r_i \cos\theta_{i1},\\
z_{i2} &= r_i \sin\theta_{i1}\cos\theta_{i2},\\
&\vdots\\
z_{ij} &= r_i \sin\theta_{i1}\cdots\sin\theta_{i,j-1}\cos\theta_{ij}, \text{ for } j \text{ up to } (p-1),\\
z_{ip} &= r_i \sin\theta_{i1}\cdots\sin\theta_{i,p-1},
\end{aligned}$$

so that $r_i^2 = \mathbf{z}_i'\mathbf{z}_i = \Sigma_{j=1}^p z_{ij}^2$. The initial p-dimensional observations are thus representable in terms of a radius and $(p-1)$ angles. The relevant approximate distributional results, if the initial observation have a p-dimensional normal distribution, are that r_i^2 will have approximately a chi-squared distribution with p degrees of freedom, θ_{p-1} will be approximately uniformly distributed over $(0, 2\pi), \theta_j$ for $j = 1, \ldots, (p-2)$ will have approximately a distribution whose density is

$$f(\theta_j) = \frac{1}{B\left(\frac{p-j}{2}, \frac{1}{2}\right)} \sin^{p-1-j}\theta_j;\ 0 \leqslant \theta_j \leqslant \pi,$$

and the distribution of the radius and all angles are approximately mutually independent.

These distributional results are useful in suggesting appropriate probability plots for checking the p-dimensional normality of the observations. Specifically, there are p separate probability plots that one could make: (a) a plot of the n-ordered squared radii values against the corresponding quantiles of the chi-squared distribution with p degrees of freedom; (b) a uniform probability plot of the n-ordered values of the normalized angle $\theta_{p-1}^* = \theta_{p-1}/2\pi$; and (c) for each of the $(p-2)$ angles, θ_j ($j = 1, \ldots, p-2$), a probability plot of the ordered values of θ_j against the corresponding quantiles of the distribution with density $f(\theta_j)$ given above. Regarding the probability plot of θ_j ($j = 1, \ldots, p-2$), the transformation $V_j = \sin^2\theta_j$ leads to a beta distribution for V_j with parameters $(p-j)/2$ and $1/2$. Using this fact, one can either obtain the quantiles of the distribution of θ_j from those of the associated beta distribution, or transform to V_j and make a beta probability plot (see Gnanadesikan et al., 1967) of the n-ordered values of V_j. In terms of the sphericized random variables, $Z_1, Z_2, \ldots, Z_p$, the random variable, $V_j = (\Sigma_{k=j+1}^{p} Z_k^2)/(\Sigma_{k=j}^{p} Z_k^2)$, $j = 1, 2, \ldots, (p-2)$, and $\theta_p = arctan(Z_p/Z_{p-1})$. Also, $r^2 = \Sigma_{k=1}^{p} Z_k^2$. Thus, in practice, all the quantities involved in the probability plots can be computed directly from the values of the sphericized random variables.

Analogous to the bivariate case discussed earlier, in addition to the separate probability plots of the radius and $(p-1)$ angles one can also make plots on unit squares for pairs of appropriately transformed (viz., the probability integral transforms of the radius and the θ_j for $j = 1, \ldots, p-2$) radius and angle values, and study these for uniformity of scatter. With modern graphical display facilities, one can also study triplets of such transforms plotted on the unit cube.

Mardia (1970, 1975) has proposed a large-sample test for multivariate normality based on measures of multivariate skewness and kurtosis. The measure of multivariate skewness suggested by him is

$$b_{1,p} = \frac{1}{n^2} \sum_{i,k=1}^{n} \{(\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_k - \bar{\mathbf{y}})\}^3$$
$$= \frac{1}{n^2} \sum_{i,k=1}^{n} \{r_i r_k \cos\theta_{ik}\}^3,$$

where θ_{ik} is the angle between the scaled residual vectors $\mathbf{z}_i$ and $\mathbf{z}_k$. [*Note*: θ_{ik} is referred to as the Mahalanobis angle by some authors.] The dependence of $b_{1,p}$ on θ_{ik} implies that it reflects the orientation of the data. The large-sample test for joint normality based on $b_{1,p}$ would be to refer the observed value of the statistic $A = nb_{1,p}/6$ to a chi-squared distribution with $p(p+1)(p+2)/6$ degrees of freedom.

The multivariate kurtosis measure proposed by Mardia (1970) is the arithmetic mean of the squares of the Mahalanobis generalized distances of the observations from the sample mean, that is,

$$b_{2,p} = \frac{1}{n} \sum_{i=1}^{n} \{(\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_i - \bar{\mathbf{y}})\}^2 = \frac{1}{n} \sum_{i=1}^{n} r_1^4,$$

where the r_i^2's are the squared radii discussed earlier. This measure depends on how far observations are from the mean and thus reflects only the tail behavior of the data, and not their orientation. The proposed large-sample test for kurtosis departures from joint normality is to compare $b_{2,p}$ to a normal distribution with mean $p(p+2)$ and variance $8p(p+2)/n$. In other words, the statistic

$$B = \frac{b_{2,p} - p(p+2)}{[8p(p+2)/n]^{1/2}}$$

is to be compared against the percentage points of the standard normal distribution. Limited investigation of the normality of B, using simulated bivariate normal samples, suggests that one would need extremely large samples for the normal approximation to be adequate and that, even for moderately large n, the distribution of B can be positively skewed. Hence, in small and moderately large samples, the exact significance level associated with the above test may be quite different from the assumed nominal level.

Tests Based on Unidimensional Views. One attractive property of tests for marginal normality is that the computational effort involved increases only linearly with p, the dimensionality of the data. It is therefore not inappropriate to examine the possibility of using various unidimensional views of the data in addition to just the marginal variables. A study of the squared radii by themselves, as proposed by Healy (1968) and by Kessell & Fukunaga (1972), is one example. Investigating the degree to which the regression of each variable on all the others is linear (a property of the multivariate normal) is another example of using a collection of unidimensional views of the data.

Another obvious class of techniques to seek is based on the characterization of the multivariate normal distribution in terms of univariate normality of all linear combinations of the variables. Tests of multivariate normality that look at "all possible" unidimensional projections and utilize the union-intersection principle of Roy (1953) have received some attention (see Aitkin, 1972; Malkovich & Afifi, 1973). The computational efforts involved in some of these tests, however, tend to be prohibitive. A different scheme, based on looking at unidimensional projections of the multivariate data along specified, rather than "all possible," directions is described next.

Marginal analysis of each of the original variables considers the projections of the data onto each of the coordinate axes separately. Other one-dimensional projections may also be considered. It is of some interest to use the projections that are likely to exhibit certain types of marked nonnormality.

One approach to this problem is to look at projections of the data along directions that are in part determined by the data, but also in part chosen to be sensitive to particular types of nonnormality. The work of Andrews et al. (1971) described in Section 5.3 in the context of estimating transformations to enhance directional normality provides a contact point for the testing problem of present concern.

From the discussion in Section 5.3, it will be recalled that the method consists in first obtaining the projections of the observations onto the unidimensional space specified by the direction vector $\mathbf{d}_\alpha^{*\prime}$, which has been chosen to be sensitive to particular types of nonnormality by appropriately specifying a value for α. Then, since these projections constitute a univariate sample, they may be studied by any of the univariate procedures (described earlier in the context of evaluating marginal normality) for detecting departures from univariate normality. For instance, the D'Agostino & Pearson (1973) test, the Shapiro-Wilk test, the shifted-power transformation test, and a normal probability plot are all candidates for use.

Because of the data-dependent, as well as certain other, aspects of the approach the significance levels associated with the formal tests are probably not formally applicable when used with the univariate "sample" of the projections. However, they do provide useful benchmarks for measuring the nonnormality along particular directions (employing different values of α would enable one to look in many directions) in the space of the original variables. If this measure is not significant, there is some hope that subsequent methods of analysis will behave as expected. If, on the other hand, this measure is highly significant, a further transformation may make the subsequent analysis more meaningful. Since the transformation test derives from the estimation technique described in Section 5.3 for enhancing directional normality, it provides an indication of what transformation will ameliorate the abnormalities when the data are viewed in specified directions.

The various methods for assessing normality described in this subsection are applied to specific sets of data in the three examples discussed next. The examples, which for simplicity are limited to bivariate observations, are based both on computer-simulated (Examples 35 and 36) and "real" (Example 37) data. The examples involving computer-simulated data are useful because the departure from normality is known since it is part of the data-generation process. The two such examples included here are extracted from a larger set studied by Andrews et al. (1972), who discuss a greater variety of nonnull (i.e., nonnormal) data.

The scheme involved in generating the computer data was to start with observations on two independent standard normal variables, X_1 and X_2, then to transform the observations on each of these variables separately to yield observations on two independently distributed variables, Z_1 and Z_2, with a specified (but same for Z_1 and Z_2) nonnormal distribution, and, finally, to combine the variables Z_1 and Z_2 to form correlated variables, Y_1 and Y_2, by using the linear transformation

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \to \begin{cases} Y_1 = Z_1, \\ Y_2 = \rho Z_1 + \sqrt{1-\rho^2}\, Z_2. \end{cases} \tag{81}$$

The correlation coefficient between Y_1 and Y_2 would thus be ρ. A different scheme for generating correlated bivariate nonnormal distributions is also discussed and used by Andrews et al. (1972, 1973), but Examples 35 and 36 apply only the scheme just described.

Example 35. This mildly nonnormal example involves 100 observations from a bivariate correlated $\chi^2_{(10)}$ distribution. Two independent $\chi^2_{(10)}$ (i.e., Z_1 and Z_2 of Eq. 81 were $\chi^2_{(10)}$ variables) samples were taken and then correlated as in Eq. 81 with $\rho = 0.9$. The two variables have marginal distributions that are relatively close to normal, the second being "expected" to be more nearly normal than the first.

Exhibit 35a. Results of Box-Cox transformation test

	Variable	
	Y_1	Y_2
Parameter estimates		
$\hat{\xi}$	-0.032	-5.429
$\hat{\lambda}$	0.382	0.604
Log likelihood-ratio value	5.724	7.639
Approximate significance level	0.0007	0.0001

The first analyses to be performed on the data were addressed to assessing the univariate normality of each variable separately. For instance, the two-parameter family of transformations

$$y \rightarrow (y + \xi)^{\lambda}$$

yielded parameter estimates and a likelihood-ratio test for each marginal variable. These results are summarized in Exhibit 35*a*. This test gives strong evidence of nonnormality of both marginal distributions.

The skewness and kurtosis measures, $\sqrt{b_1}, b_2$, were calculated for both marginal variables, and the results are recorded in Exhibit 35*b*. There is some statistical evidence of skewness in the distributions of both variables but not of kurtosis.

The Shapiro-Wilk test, applied to the marginal distributions of these data, yielded values of W of 0.954 and 0.965 for the two variables. Without precise tables of percentage points for the present sample size, it is difficult to conclude anything regarding statistical significance other than that the first value (viz., 0.954) is "possibly" mildly statistically significant. The D'Agostino test, when applied to the two variables, led to values of the D statistic both of which had significance levels greater than 0.2.

Exhibit 35b. Marginal skewness and kurtosis

Measure		Y_1	Y_2
Skewness	$\sqrt{b_1}$	0.699	0.395
Approximate significance level		<0.01	<0.05
Kurtosis	b_2	3.20	2.784
Approximate significance level		Not sig.	Not sig.
Omnibus D'Agostino-Pearson test	$\chi^2_{(2)}$	8.382	2.850
Significance level		$\simeq 0.015$	Not sig.

Exhibit 35c. Standardized gaps test

		Variable 1	Variable 2
Left-hand gaps	$r_L - 1$	-0.267	-0.325
Approximate significance level		$2\Phi(-1.08)$ $\simeq 0.28$	$2\Phi(-1.31)$ $\simeq 0.19$
Right-hand gaps	$r_U - 1$	0.351	0.027
Approximate significance level		$2\Phi(-1.42)$ $\simeq 0.16$	$2\Phi(-0.11)$ Not sig.
Combined statistic	q	4.626	1.951
Approximate significance level		0.1	0.4

For applying the gaps test, the standardized spacings or gaps were calculated for both marginal distributions. In both cases the difference between left and right tail lengths was manifested by gaps shorter on the left and longer on the right. The values of $(r_L - 1)$ and $(r_U - 1)$ for each variable, together with approximate significance levels, are recorded in Exhibit 35*c*. The combined statistic, q, was also computed, and its value, together with the approximate significance level, is also shown in the exhibit. Both variables show a specific skewness departure from normality in that the tails of the distributions appear to be short on the left and long on the right. However, these departures are not extremely statistically significant as measured by the formal gaps test.

Exhibits 35*d* and *e* are normal probability plots of the data for the first and second variables, respectively. The departure from normality is quite striking in Exhibit 35*a*. In Exhibit 35*e* departure from normality in the second variable, although not as striking, can be detected in the gentle curvature away from the hypothesized linear configuration.

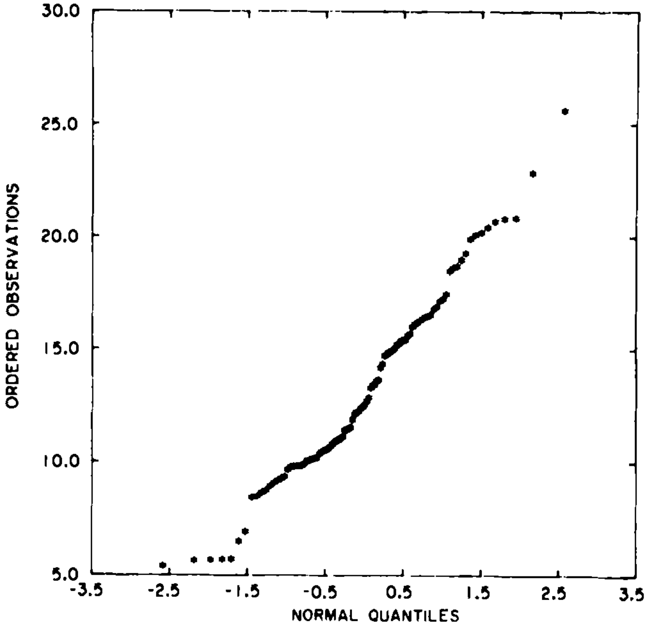
The marginal techniques have indicated with differing degrees of strength the apparent nonnormality in the marginal distributions for this example. The results of applying the techniques for assessing joint normality are described next.

Exhibit 35*f* (see page 206) is a scatter plot of the bivariate data. The bivariate transformation technique described earlier for assessing bivariate normality was applied, and two transformation parameters, λ_1, λ_2, were estimated by maximum likelihood. The asymptotic properties of the likelihood ratios yielded an approximate test of the null hypothesis $\lambda_1 = \lambda_2 = 1$. The results of this procedure are summarized in Exhibit 35*g* (see page 206). From this test there is some, though not very strong, evidence of nonnormality.

The nearest distance test did not yield significant results. The significance level was about 0.4. This test seems to have relatively low power against smooth departures from normality, as is the case in the present example.

Next, the techniques based on radii and angles were applied. Exhibit 35*h* (see page 207) is a scatter plot of the radius-and-angle reparametrization of

Exhibits 35d,e. Normal probability plots for the two variables



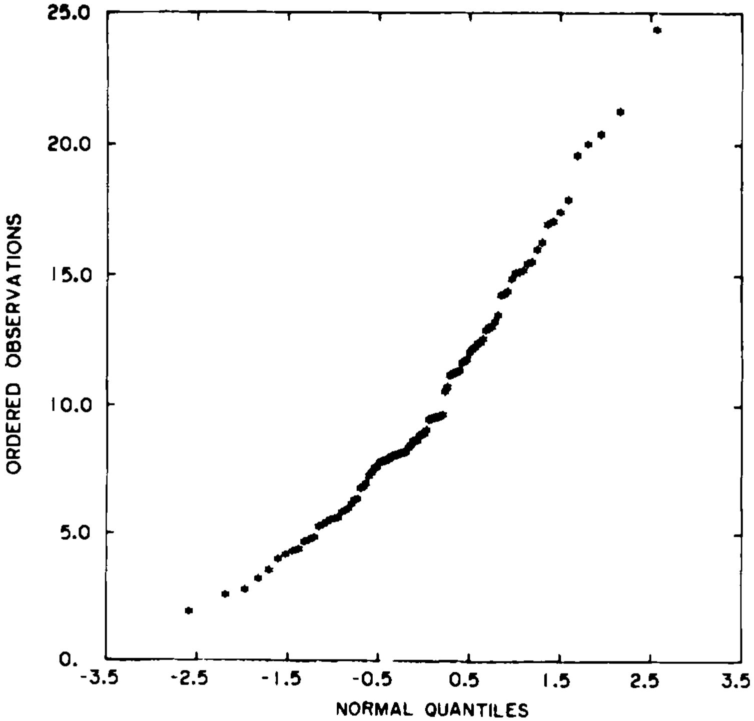
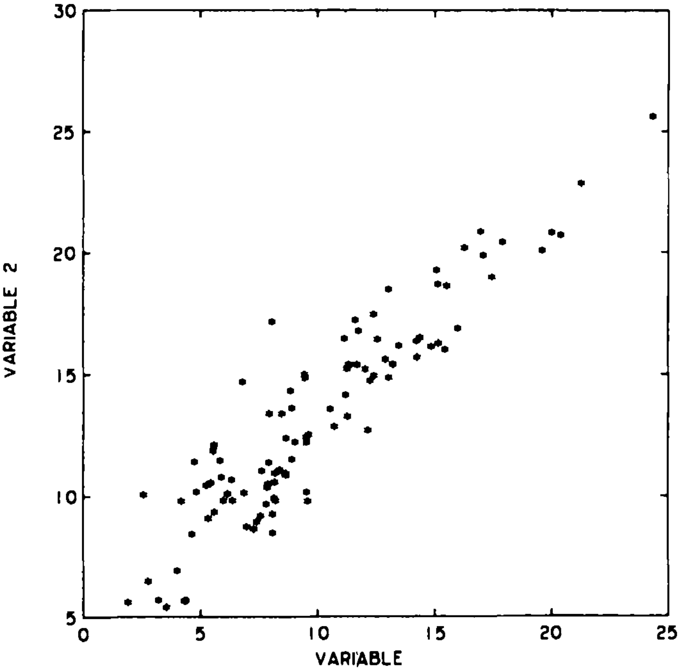


Exhibit 35*f*. Scatter plot of data



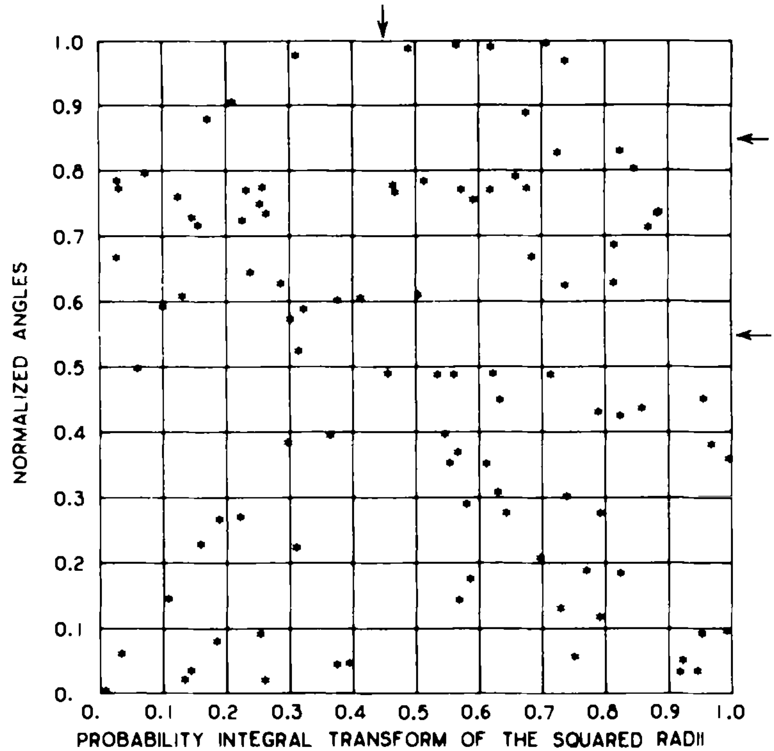
these data. Under the null hypothesis the plotted points have a uniform distribution on the unit square. Departures from this null hypothesis are quite apparent in this plot in that several cells are empty and also several horizontal and some vertical strips (e.g., the ones marked with arrows) are sparse in points relative to other strips. Exhibit 35*i* (see page 208) is a $\chi^2_{(2)}$ probability plot of the squared radii for this example. This plot appears to be reasonably linear, exhibiting no marked departures of the squared radii from null expectations.

Exhibit 35*j* (see page 209) is a uniform probability plot of the normalized angles. Under the null hypothesis these normalized angles should have a uniform distribution. This plot, however, appears quite irregular, especially at

Exhibit 35*g*. Bivariate power transformation test

$\hat{\lambda}_1$	0.937
$\hat{\lambda}_2$	0.706
$\mathscr{L}_{\max}(\hat{\lambda}_1, \hat{\lambda}_2) - \mathscr{L}_{\max}(1, 1)$	2.8
Approximate significance level	0.061

Exhibit 35k. Scatter plot of normalized angles vs. probability integral transform of squared radii



the upper end. A chi-squared goodness-of-fit test based on 10 equal cells yields a statistic of 25.4 with a corresponding significance level of ~0.002.

The multivariate skewness and kurtosis tests proposed by Mardia (1970) were also used with these data. Whereas the skewness statistic revealed a striking departure from bivariate normality, the kurtosis statistic was not statistically significant.

Finally, the technique of testing for directional normality was applied to the data in this example. This procedure (with $\alpha = 1.0$) selected a direction $\mathbf{d}_{1.0}^{*\prime} = (0.788, 0.616)$ and investigated the projections of the data on this one-dimensional subspace. The univariate shifted-power transformation procedure was then applied, with the results summarized in Exhibit 35*k* (see page 209). In the direction chosen, the data exhibit extreme nonnormality, much more marked than either of the marginal variables (see Exhibit 35*a*). Exhibit 35*l* (see page 210) shows a normal probability plot of the projections onto the unidimensional space specified by $\mathbf{d}_{1.0}^{*}$, and the departure from linearity here is just as striking as the one in Exhibit 35*d*.

Exhibit 35*l*. Chi-squared (df = 2) probability plot of squared radii

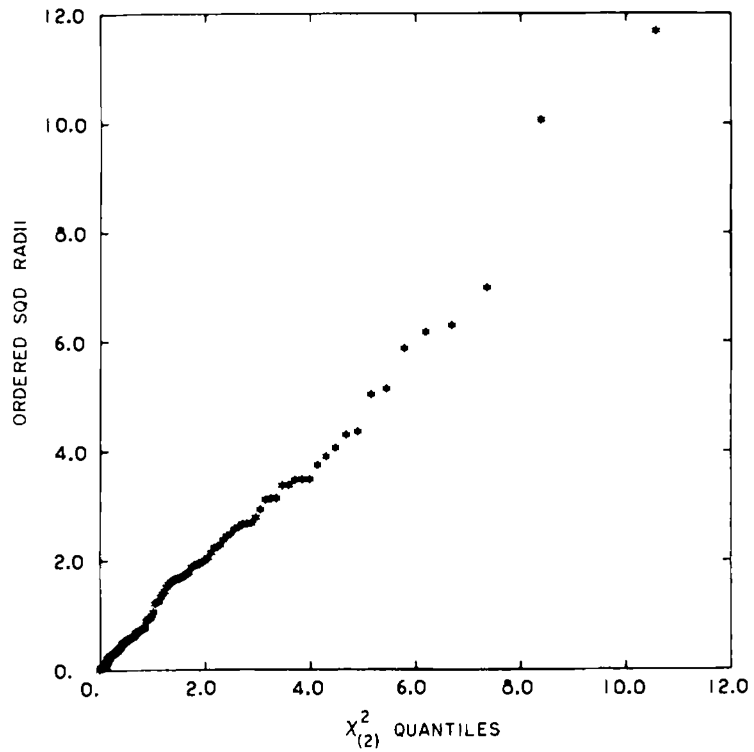
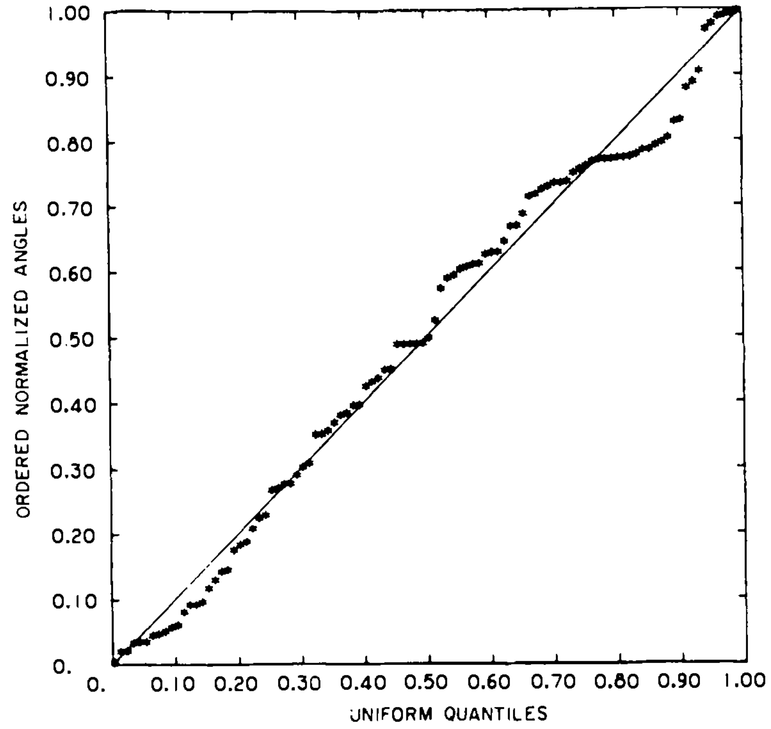


Exhibit 35*m* (see page 210) summarizes the results of applying the various techniques to this example. Many of the techniques indicate significant departures from multivariate normality. The exceptions include the D'Agostino test and the univariate gaps test for marginal normality, and the nearest distance test for bivariate normality. An important aspect of this example is the discovery of which methods did not detect the sort of departure incorporated in the data.

Example 36. The data for this example are 100 points from a correlated Laplace distribution, correlated by using Eq. 81 with $\rho = 0.9$. The distribution is long tailed but quite symmetric. Only the results of using the techniques for assessing joint and directional normality are described here.

The bivariate transformation procedure was applied to these data after initially shifting the observations to make them all lie in the first quadrant, and the results are summarized in Exhibit 36*a* (see page 211). The transformation utilized here involved only power parameters and no shift parameters. For this

Exhibit 35*j*. Uniform probability plot of normalized angles



reason, one would expect sensitivity to skewness but not to long-tailedness in the presence of symmetry. Including shift parameters in the transformations of the variables would most probably remedy the situation, but the computational effort required would be substantially higher. At any rate, the nonsignificant result in Exhibit 36*a* is at least interpretable.

Exhibit 35*k*. Directional normality test

Shift parameter estimate	$\hat{\xi}$	−5.66
Power parameter estimate	$\hat{\lambda}$	0.53
$\mathcal{L}_{\max}(\hat{\xi}, \hat{\lambda})$		−168.2
$\mathcal{L}_{\max}(\hat{\xi}, 1)$		−178.3
Approximate significance level		0.00001

Exhibit 35*l*. Normal probability plot of projections onto direction of nonnormality

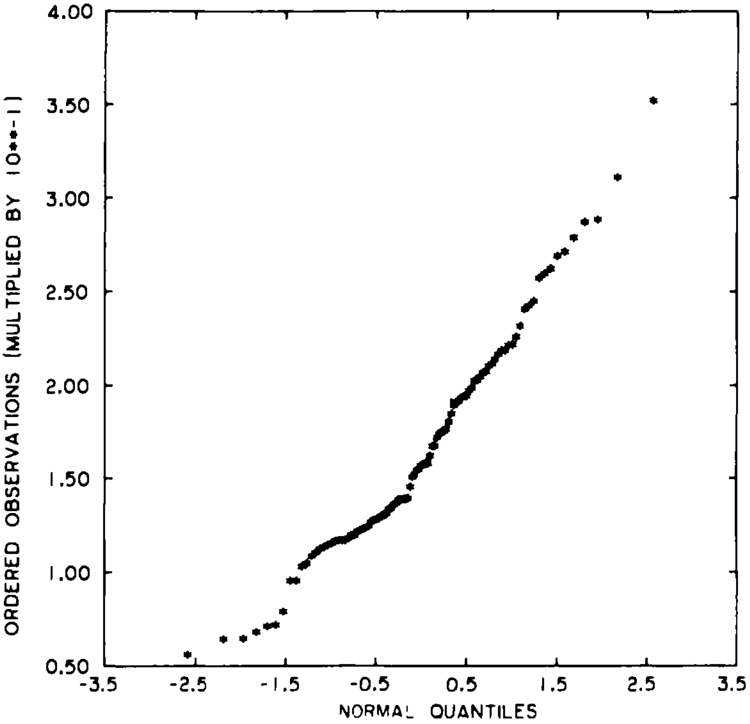


Exhibit 35*m*. Example 35 summary

Technique	Significance Level	
Marginal		
Marginal Box-Cox	0.0007	0.0001
Skewness	<0.01	<0.05
Kurtosis	Not sig.	Not sig.
D'Agostino-Pearson	$\simeq 0.015$	Not sig.
Shapiro-Wilk	?	?
D'Agostino	>0.02	>0.2
Univariate gaps	~ 0.2	
Marginal probability plots	Some evidence of nonnormality	
Joint		
Scatter plot	Some evidence of nonnormality	
Bivariate transformation	0.06	
Nearest distance	0.4	
Radius and angles	Good evidence of nonnormality	
Mardia's tests	0.0014($b_{1,2}$)	$b_{2,2}$ not sig.
Directional transformation	0.00001	

Exhibit 36a. Bivariate power transformation test

Estimates of transformation parameters	
$\hat{\lambda}_1$	1.22
$\hat{\lambda}_2$	1.20
$\mathscr{L}_{\max}(\hat{\lambda}_1, \hat{\lambda}_2) - \mathscr{L}_{\max}(1, 1)$	0.91
Approximate significance level	0.40

The nearest distance test also failed to detect any significant departure from normality in this case—the observed significance level was 0.7. On the other hand, the multivariate skewness and kurtosis tests revealed significant skewness and kurtosis departures; observed levels were 0.0055 and $< 10^{-4}$, respectively. Examination of a scatter plot of the data suggested that it is quite reasonable to reject bivariate normality on grounds of both skewness and kurtosis.

The plotting procedures for radii and angles also proved useful once again. Exhibits 36*b*–*d* show the combined scatter and marginal probability plots of

Exhibit 36*b*. Scatter plot of normalized angles vs. probability integral transform of squared radii

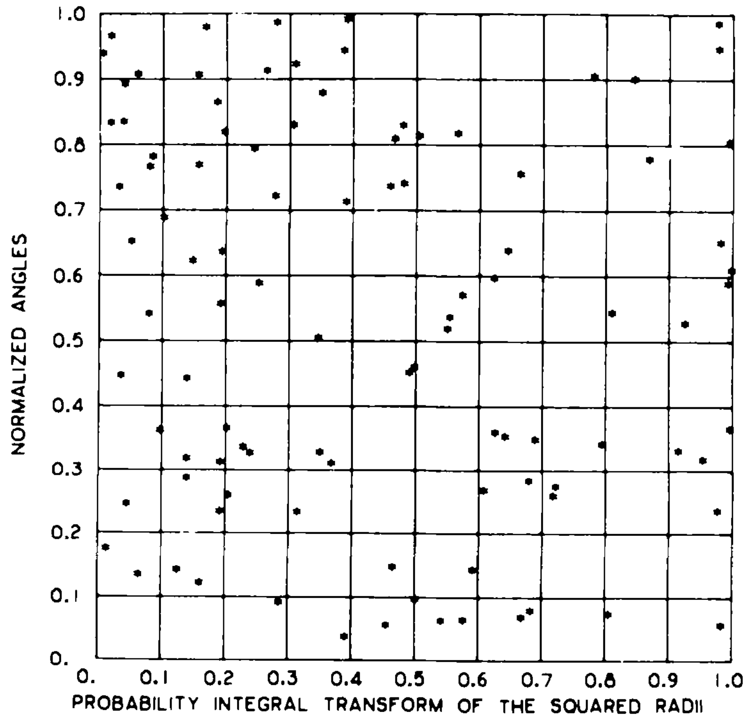


Exhibit 36*c*. Chi-squared (*df* = 2) probability plot of squared radii

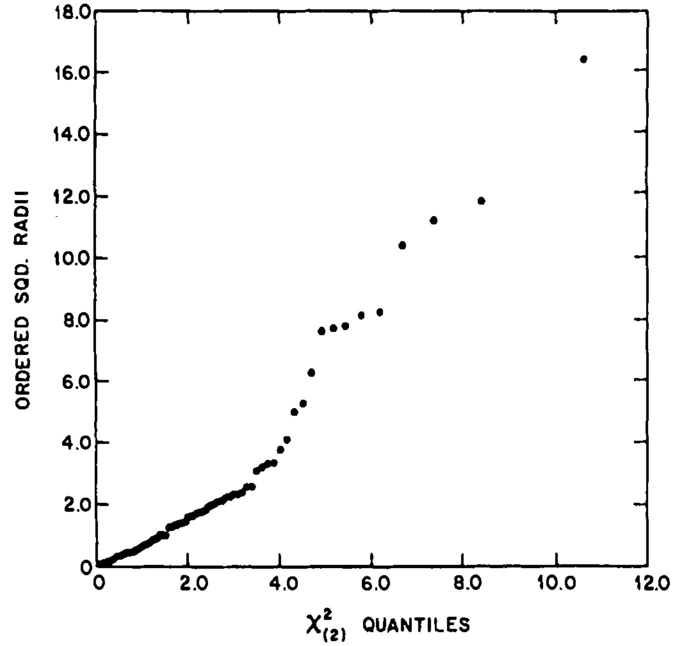


Exhibit 36*d*. Uniform probability plot of normalized angles

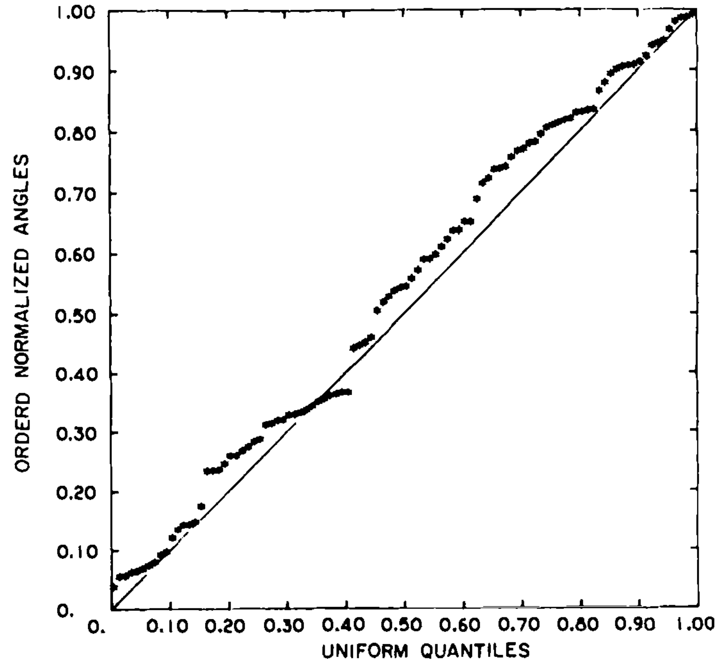


Exhibit 36e. Directional normality tests

α	$\mathbf{d}_\alpha^{*\prime}$	$\hat{\xi}$	$\hat{\lambda}$	$\mathscr{L}_{\max}(\hat{\xi},\hat{\lambda}) - \mathscr{L}_{\max}(\hat{\xi},1)$	Approximate Significance Level
1.0	−0.65, −0.76	23.2	−1.0	3.2	0.011
0.5	−0.63, −0.78	24.5	−1.0	3.2	0.011
0.1	−0.59, −0.81	24.5	−1.0	3.1	0.013
−0.1	0.56, 0.83	10.0	1.8	4.1	0.004
−0.5	0.48, 0.88	9.7	1.8	4.0	0.005
−1.0	0.31, 0.95	8.9	1.7	3.9	0.005

radii and angles for these data. The long-tailedness of the data is clearly evident in the $\chi^2_{(2)}$ probability plot of the squared radii (Exhibit 36*c*). Some evidence of lack of spherical symmetry of the sphericized residuals is provided by the uniform probability plot of the angles (Exhibit 36*d*).

Lastly, as a means of studying directional normality in this example, projections of the data were explored by employing directions $\mathbf{d}_\alpha^{*\prime}$ for a range of values of α, namely, $\alpha = -1, -0.5, -0.1, 0.1, 0.5, 1$. Exhibit 36*e* gives for each value of α the resulting direction, $\mathbf{d}_\alpha^{*\prime}$ together with the results of the univariate shifted-power transformation test procedure applied to the projections on the direction involved. The directions determined by using $\alpha < 0$, being sensitive to the center of the data, do indicate more significant departures, and this is not very surprising in view of the difference between the densities of the Laplace and the normal in the center.

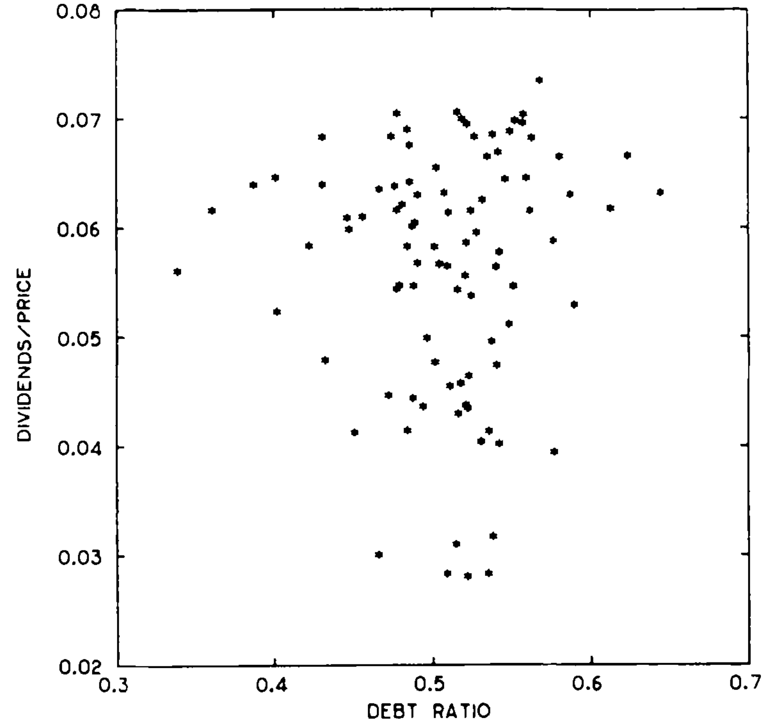
In summary, as expected, this symmetric nonnormality, which is an important though not sufficiently extreme departure, was not clearly detected by some procedures. The results for this example are summarized in Exhibit 36*f*.

Example 37. Since real data may not conform to any prespecified type of nonnormality of the kinds reflected in Examples 35 and 36, it is instructive to

Exhibit 36*f*. Example 36 summary

Technique	Approximate Significance Level
Bivariate power transformation	0.4
Nearest distance	0.7
Mardia's tests	$0.0055(b_{1,2})$, $<10^{-4}(b_{2,2})$
Radius-and-angles decomposition	Indication of departures from normality
Directional normality	<0.015 for $\alpha > 0$, $\simeq 0.005$ for $\alpha < 0$

Exhibit 37*a*. Scatter plot of dividends/price vs. debt ratio for 94 utilities in 1969



apply the techniques for assessing normality to observations that are not simulated. Thus the data for this example, which is taken from Standard and Poor's COMPUSTAT tape, consist of observed values of debt ratio and the dividends/price ratio for each of 94 utilities for the year 1969. Exhibit 37*a* is a scatter plot of the observations, and departures from normality are evident even in this simple plot.

The test proposed by D'Agostino (1971) was applied to both marginal distributions and did not indicate any strikingly significant departures from normality. The results of estimating a shifted-power transformation of each variable by the methods of Box & Cox (1964), and of applying the associated likelihood-ratio test of univariate normality to each variable separately, are summarized in Exhibit 37*b*. The transformation-based test indicates a highly significant departure from normality for the distribution of values of the dividends/price ratio.

Exhibits 37*c* and 37*d* (see page 216) are normal probability plots of debt ratio and dividends/price ratio, respectively. Both plots exhibit noticeable deviations from the null straight line configuration to be expected for normal-

Exhibit 37*b*. Results of Box-Cox transformation test

	Debt Ratio	Dividends/Price
Parameter estimates		
$\hat{\xi}$	-0.127	0.195
$\hat{\lambda}$	1.980	10.873
Log likelihood-ratio value	1.826	7.854
Approximate significance level	$0.05 \leqslant p < 0.058$	$0.00006 \leqslant p \leqslant 0.00008$

ity. Exhibit 37*d* with marked curvature indicates an abnormally short upper tail of the distribution of the dividends/price ratio. Also, six observations in the lower tail are distinctly separated from the rest of the data.

The bivariate transformation procedure also detected significant nonnormality. Exhibit 37*e* presents the results of this procedure. Some evidence of nonnormality appears in both variables, as indicated by the values of $\hat{\lambda}_1$ and $\hat{\lambda}_2$.

Exhibit 37*c*. Normal probability plot for debt ratio

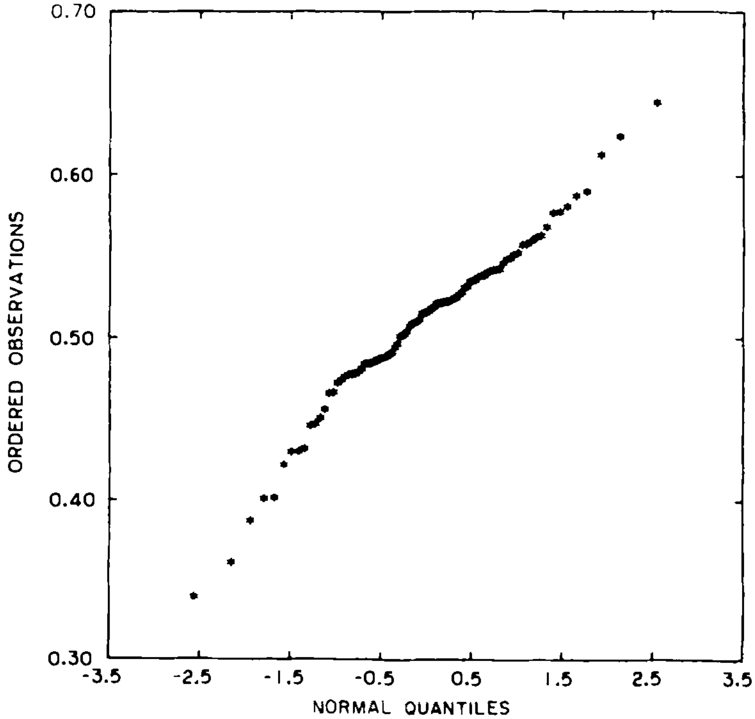
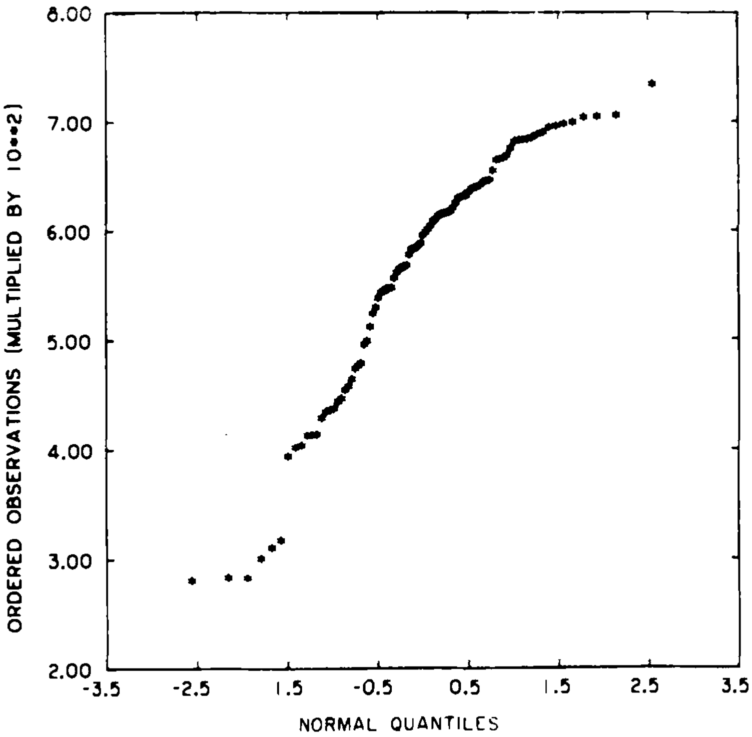


Exhibit 37*d*. Normal probability plot for dividends/price



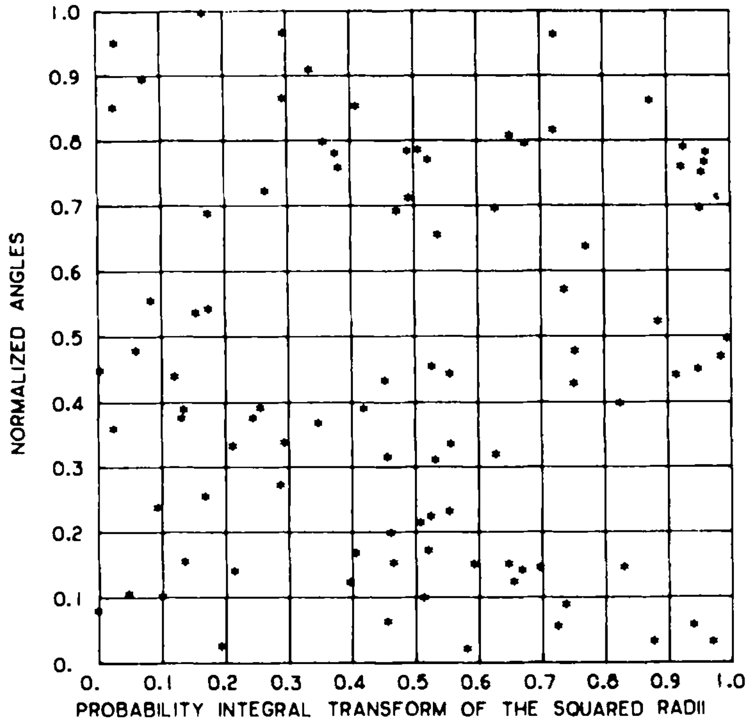
The nearest distance test was used with both linear and quadratic regressions as the basis for studying the dependence of transformed nearest neighbor distances on the location of the point from which the distances were measured. The significance levels of these two regression tests were 0.031 and 0.093, respectively. In this example, therefore, this test provides some indication, although not very strong evidence of nonnormality.

Exhibit 37*f* is a scatter plot of the radii-and-angles decomposition. Exhibit 37*g* (see page 218) is a $\chi^2_{(2)}$ probability plot of the squared radii, and Exhibit 37*h* (see page 219) is the uniform probability plot of the normalized angles. The

Exhibit 37*e*. Bivariate power transformation test

$\hat{\lambda}_1$	2.719
$\hat{\lambda}_2$	2.375
$\mathscr{L}_{\max}(\hat{\lambda}_1, \hat{\lambda}_2) - \mathscr{L}_{\max}(1, 1)$	8.227
Approximate significance level	0.00027

Exhibit 37*f*. Scatter plot of normalized angles vs. probability integral transform of squared radii



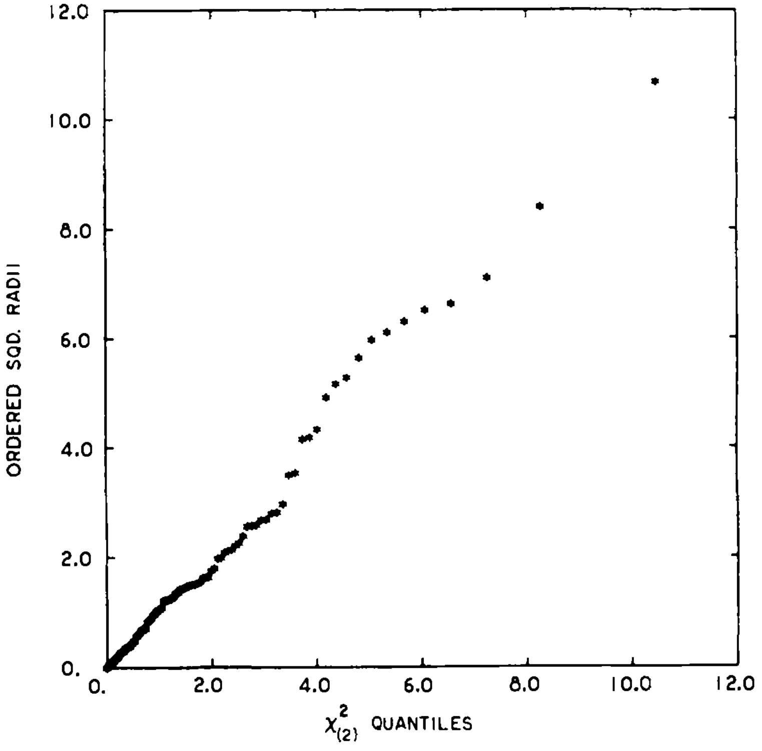
nonuniform scatter in Exhibit 37*f*, and especially the sparseness in several contiguous blocks, indicate departures from bivariate normality. Exhibit 37*g* shows peculiarities in the upper tail of the distribution of the squared radii, Exhibit 37*h* also manifests some departures from spherical symmetry in the distribution of the sphericized residuals.

The directional normality procedure also suggested a somewhat, but not strikingly, significant departure from normality. Exhibit 37*i* (see page 219) presents the results of the directional normality test for the case of $\alpha = 1.0$. Here $\mathbf{d}^*_{1.0}$ is clearly influenced heavily by the second variable (dividends/price ratio). This is not too surprising in the light of the more striking nonnormality of the second variable, as revealed by the tests for marginal normality discussed earlier.

The results for this example are summarized in Exhibit 37*j* (see page 220).

In this subsection various techniques have been described for assessing the normality of the distribution of multiresponse data. Many of the new techniques (e.g., the nearest distance test, the radius-and-angles plots) need further theoretical investigation, as well as practical use and exposure. On the theoretical front some refinement of the distributional approximations in-

Exhibit 37g. Chi-squared ($df = 2$) probability plot of squared radii

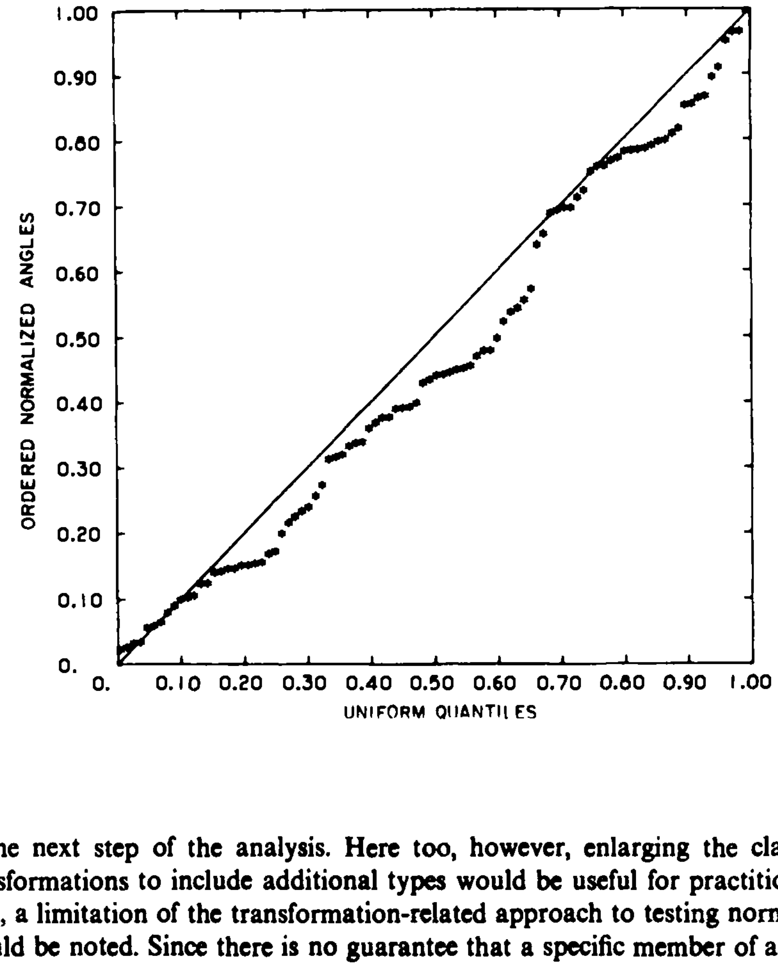


volved in some of the procedures (e.g., the univariate gaps test and the multivariate nearest distance test) may be in order. Also, a better understanding is needed of issues such as the nonuniqueness of some of the preliminary transformations of the data (e.g., the Rosenblatt, 1952, transformation mentioned in connection with the nearest distance test) and the effects of using the sample mean vector and covariance matrix in place of the corresponding population quantities in some of the methods.

More work, of course, is needed to promote understanding of the relative sensitivities of the different procedures. This is necessary, not for picking an optimal test for normality, but for general guidance in interpreting the results in specific applications of these techniques.

General indications concerning the newer techniques are that the transformation-related methods and the plotting procedures based on the radius-and-angles representation of multivariate data appear to be promising tools for data analysis. The transformation-related methods have appeal above and beyond serving as tests of significance because of the fact that estimates of the transformation (admittedly within some class such as the shifted-power one) are included as an integral part of the method and are likely to be very useful

Exhibit 37k. Uniform probability plot of normalized angles



in the next step of the analysis. Here too, however, enlarging the class of transformations to include additional types would be useful for practitioners. Also, a limitation of the transformation-related approach to testing normality should be noted. Since there is no guarantee that a specific member of a class of transformations, such as the shifted-power class, will necessarily achieve normality, the evidence, as provided by the transformation test, that no transformation is required cannot be taken entirely at face value as adequate support for normality. Specifically, if one were to transform a set of data by

Exhibit 37l. Directional normality
$\mathbf{d}_{1.0}^{*\prime} = (-0.462, -0.887)$

Shift parameter estimate	$\hat{\xi}$	4.415
Power parameter estimate	$\hat{\lambda}$	−10.703
$\mathscr{L}_{\max}(\hat{\xi}, \hat{\lambda}) - \mathscr{L}_{\max}(\hat{\xi}, 1)$		2.487
Approximate significance level		0.026

Exhibit 37j. Example 37 summary

Technique	Approximate Significance Level	
Marginal Box-Cox	0.06	0.00008
Marginal probability plots	Good evidence of marginal nonnormality	
Bivariate power transformation	0.0003	
Radii and angles	Good evidence of joint nonnormality	
Directional normality	0.03	

the techniques described in Section 5.3 and treat the resulting transformed data as input to the same transformation techniques, one would necessarily get indications that no further transformation (within the class considered) was required, but this would be just an artifact of iterating the transformation technique. The "direct" techniques (i.e., those not related to transformations) for assessing normality do not suffer from such a limitation.

The plotting procedures associated with the radius-and-angles decompositions have particular appeal as informal but informative graphical aids for data analysis. Additional graphical methods, especially directed toward assessing the normality of high-dimensional data, would indeed be worth developing (see the discussion of one such tool in Section 6.2). Such graphical techniques often exemplify the significant value of a statistical tool which may have been designed for one purpose but turns out to have a variety of additional applications. Thus, for instance, the usefulness of the techniques considered in this subsection for assessing distributional normality is greatly enhanced by their possible utility in detecting additional data anomalies such as outliers.

5.4.3. Elliptical Distributions

Most of the distributions (see, for example, Johnson & Katz, 1972) that have been proposed as alternatives to the multivariate normal, on which much of the classical multivariate theory and methodology are based, have been defined by mathematical analogy with univariate distributions. Indeed, conceptualization of the sense in which distributions are alternatives to the multivariate normal would be a more natural starting point than such formal analogies. For some purposes (e.g., empirical study of robust estimators), relatively simple alternatives may be desirable. One may wish, for instance, to consider alternatives whose density functions have the same ellipsoidally shaped contours as the multivariate normal but still are flexible enough to provide longer- and shorter-tailed alternatives to the multivariate normal. Elliptical distributions, defined below, are such a class that has received wide attention (e.g., Kelker, 1970; Chu, 1973; see also Devlin et al., 1976).

While these elliptical distributions provide a flexible class of alternatives to the multinormal, their very simplicity implies of course that they cannot capture all types of departures from multinormality that can occur in real data. They serve as useful starting points for certain kinds of investigations, including theoretical analyses as well as computer simulations studies, of properties of statistical procedures under such alternative distributions. Due to the interest in the class of elliptical distributions for providing simple alternatives to the multinormal, in this section a number of results pertaining to the properties of this class taken from Devlin et al. (1976) are collected together. Details, including proofs, can be found in the references indicated throughout the discussion.

A variety of definitions of an elliptically distributed random vector, $\mathbf{Y}$, exists. [*Note*: In this section, the convention of using capital letters for denoting random variables is used. Hence $\mathbf{Y}$ is a vector and not the usual data matrix.] They include definitions in terms of:

(a) linear combinations
   (i) all $\mathbf{a}'\mathbf{Y}$ with the same variance should have the same distribution (see Vershik, 1964);
(b) probability density functions, $f(\mathbf{y})$, of $\mathbf{Y}$
   (ii) $f(\mathbf{y})$ should be a function only of a positive definite quadratic form $\mathbf{y}'\mathbf{C}^{-1}\mathbf{y}$ (see Chu, 1973); and
(c) characteristic functions, $c(\mathbf{t})$, of $\mathbf{Y}$
   (iii) $c(\mathbf{t})$ should depend only on a quadratic form $\mathbf{t}'\mathbf{C}\mathbf{t}$ (see Kelker, 1970).

   (For convenience, it is assumed here and in what follows that $\mathbf{Y}$ has been centered at the origin.)

The three definitions are not completely equivalent. Vershik's definition, for instance, implicitly assumes the existence of the first two moments and would thus exclude elliptical t distributions with 1 or 2 degrees of freedom. Most of the usual elliptical distributions will, however, satisfy all three definitions.

The matrix $\mathbf{C}$ appearing in (ii) and (iii) is called the *characteristic matrix* by Chu. It is determined only up to a multiplicative constant. Its role is like that of the covariance matrix and, indeed, when the latter exists, $\mathbf{C}$ must be proportional to it.

Items J1–J7 list a few of the important joint distributional properties of elliptical distributions, while C1–C3 summarize their useful conditional distributional characteristics.

J1. The components of $\mathbf{Y}$ are mutually independent *iff* $\mathbf{C}$ is diagonal and $\mathbf{Y}$ is multinormal, $N[\mathbf{0}, \mathbf{C}]$ (see Kelker, 1970).

J2. If $\mathbf{C} = \mathbf{I}$, then a polar coordinate transformation applied to $\mathbf{Y}$ produces new variables $(D, \Theta_1, \ldots, \Theta_{p-1})$, that is, a "radius" and $(p-1)$ "angles," which are mutually independent. The distributions of the Θ's are the same for all $\mathbf{Y}$, with density $f(\theta_k) \propto \sin^{p-1-k}\theta_k$. In particular, Θ_{p-1} has a uniform density on $[0, 2\pi]$. (See Goldman, 1974, and also

Section 5.4.2.) The density of $D^2 = \mathbf{Y}'\mathbf{Y}$, however, does depend on the distribution of $\mathbf{Y}$:

$$f(d^2) = \{\pi^{p/2}/\Gamma(p/2)\}(d^2)^{p/2-1}g(d^2),$$

where $g(d^2)$ is the density of $\mathbf{Y}$ evaluated at $\mathbf{y}'\mathbf{y} = d^2$ (see Kelker, 1970).

J3. If the covariance matrix $\mathbf{V}(\mathbf{Y}) = \mathbf{C}$ and $D^2 = \mathbf{Y}'\mathbf{C}^{-1}\mathbf{Y}$, then $\mathscr{E}(D^2) = p$.

J4. Let $\mathbf{Z} = \mathbf{T}\mathbf{Y}$ where $\mathbf{T}$ is an $(r \times p)$ matrix of rank r with $r \leqslant p$. Then $\mathbf{Z}$ is also elliptically distributed with characteristic matrix $\mathbf{TCT}'$ (see Kelker, 1970, Chu, 1973).

J5. If $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent with the same characteristic matrix $\mathbf{C}$, then $\mathbf{Y}_1 + \mathbf{Y}_2$ is also elliptically distributed with the same characteristic matrix (this follows from definition (iii); see also Yao, 1973). If, moreover, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are identically distributed with $\mathbf{V}(\mathbf{Y}_1) = \mathbf{V}(\mathbf{Y}_2) = \mathbf{C}$, then $1/\sqrt{2}(\mathbf{Y}_1 + \mathbf{Y}_2)$ will have the same distribution *iff* $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are multinormal, $N[\mathbf{0}, \mathbf{C}]$ (see Das Gupta et al., 1972).

J6. The correlation matrix of $\mathbf{Y}$, assuming it is defined, is given by $\mathbf{\Gamma} = ((c_{ij}/\sqrt{c_{ii}c_{jj}}))$ and is, therefore, the same for all elliptical distributions with the same (or only rescaled) characteristic matrices (see Kelker, 1970).

J7. The density function of $\mathbf{Y}$ can be represented as

$$f(\mathbf{y}) = \int_0^\infty n(\mathbf{y}; v)\, dW(v),$$

where dW is a weighting function ($\int_0^\infty dW(v) = 1$) which may assume negative values and n is the density function corresponding to $N[\mathbf{0}, v^{-2}\mathbf{C}]$ (see Chu, 1973).

C1. Let $\mathbf{Y}' = (\mathbf{Y}_1', \mathbf{Y}_2')$ with $\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}$ partitioned accordingly. Then $\mathscr{E}(\mathbf{Y}_1 \mid \mathbf{Y}_2 = \mathbf{y}_2) = \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{y}_2$, that is, the "regression" of $\mathbf{Y}_1$ on $\mathbf{Y}_2$ is linear (see Kelker, 1970). In particular, if $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are uncorrelated, then $\mathscr{E}(\mathbf{Y}_1 \mid \mathbf{Y}_2 = \mathbf{y}_2) = \mathbf{0} = \mathscr{E}(\mathbf{Y}_1)$, that is, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are semi-independent (see Vershik, 1964; Blake & Thomas, 1968). This property indicates that the well-known linearity of all regressions for the multinormal distribution is shared by other members of the class of elliptical distributions.

C2. The conditional variance, $\mathbf{V}(\mathbf{Y}_1 \mid \mathbf{Y}_2 = \mathbf{y}_2)$, is independent of $\mathbf{y}_2$ *iff* $\mathbf{Y}$ is multinormal, $N[\mathbf{0}, \mathbf{C}]$ (see Kelker, 1970). More generally, $\mathbf{V}(\mathbf{Y}_1 \mid \mathbf{Y}_2 = \mathbf{y}_2) = h(\mathbf{y}_2)(\mathbf{C}_{11} - \mathbf{C}_{12}\mathbf{C}_{22}^{-1}\mathbf{C}_{21})$ for some function h (see Chu, 1973). This property implies that the well-known homoscedasticity property of the multinormal distribution is not shared by other distributions in the elliptical class.

C3. Suppose $Y_1, Y_2 \ldots$ is an infinite sequence with $\mathbf{Y} = (Y_1, \ldots, Y_p)'$. If for any p, $\mathbf{Y}$ has an elliptical distribution with $\mathbf{C}_p = \mathbf{I}$, then there is a positive random variable V such that the conditional distribution of $\mathbf{Y}$ given $V = v$ is $N[\mathbf{0}, v^{-2}\mathbf{I}]$ (see Kelker, 1970; Kingman, 1972). For arbitrary covariance structures, $\mathbf{C}_p$, the conditional distribution is $N[\mathbf{0}, v^{-2}\mathbf{C}_p]$ (this is essentially the result of Yao, 1973).

Under the conditions of C3, it follows that W in J7 is the cumulative distribution function of a positive random variable, namely, V. The resulting special class of elliptical distributions to be called *compound multinormal* distributions, are of particular interest. (See also Picinbono, 1970 and Yao, 1973.) Rogers & Tukey (1972) give several examples of univariate distributions which are in this class.

The multinormal distribution is a member of the compound multinormal class obtained by taking V to be a constant. The other members, because they are *longer tailed* than the multinormal, offer a variety of simple, flexible and symmetric alternatives to this standard reference distribution.

Even this special class of elliptical distributions contains an infinity of members since every choice of W leads to a new case. On the other hand, an arbitrary symmetric probability density may or may not be representable as a random mixture of normal components. This is really a univariate issue, and Andrews & Mallows (1974) have developed necessary and sufficient conditions in terms of the derivatives of a univariate density for a representation of this type to be possible. These conditions express in a precise way the *long-tailed* requirements on the distribution of $\mathbf{Y}$ which were mentioned earlier. Andrews & Mallows also discuss how W can be determined.

The following are special properties of compound multinormal distributions:

S1. $\mathbf{Y}$ can be represented as $V^{-1}\mathbf{X}$ where V is a positive random variable, $\mathbf{X}$ is $N[\mathbf{0}, \boldsymbol{\Sigma}]$, and V and $\mathbf{X}$ are independent (this follows from J7).

S2. If $\mathscr{E}(V^{-4}) < \infty$, then the kurtosis of $\mathbf{a}'\mathbf{Y}$ is $3\phi^2$, where $\phi^2 = \mathscr{E}(V^{-4})/\{\mathscr{E}(V^{-2})\}^2 \geqslant 1$. Equality occurs *iff* $\mathbf{Y}$ is $N[\mathbf{0}, \mathscr{E}(V^{-2})\boldsymbol{\Sigma}]$.

S3. If $\mathbf{V}(\mathbf{Y}) = \mathbf{C}$, then

$$D^2 = \mathbf{Y}'\mathbf{C}^{-1}\mathbf{Y} = \{\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}\}/(V^2\mathscr{E}(V^{-2})\}.$$

Also, $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$ has a chi-squared distribution with p degrees of freedom, and is independent of V^2.

S4. If $\mathscr{E}(V^{-4}) < \infty$, then

$$\mathscr{E}(D^4) = p(p+2)\phi^2 \geqslant p(p+2)$$

with equality *iff* $\mathbf{Y}$ is $N[\mathbf{0}, \mathscr{E}(V^{-2})\boldsymbol{\Sigma}]$.

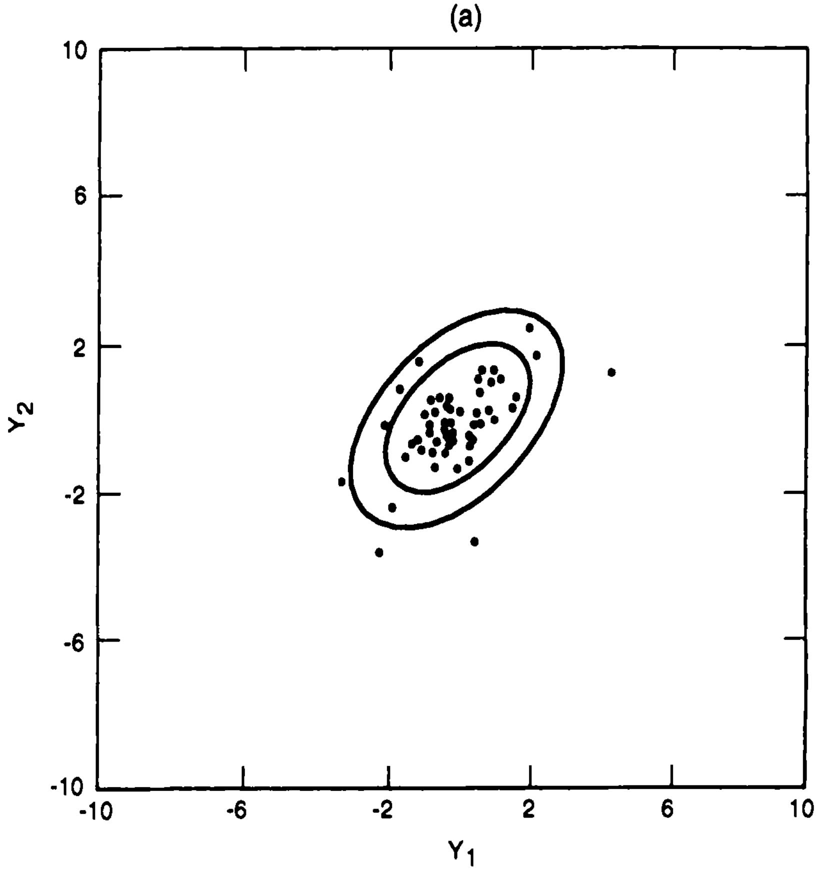


Fig. 11a. Bivariate t_4 sample with 75% and 90% probability contours

Property S4 reflects the fact that the distribution of D^2, like that of Y, becomes longer tailed when $\phi^2 > 1$. Even though $\mathscr{E}(D^2) = p$ (see J3), whatever the value of $\phi^2, \mathscr{E}(D^4)$ increases with ϕ^2.

A convenient way of generating samples of observations from compound multinormal distributions is suggested by S1: combine a multinormal sample of X with an independent sample of an appropriate V. This strategy was used to form the examples in Figures 11*a* and *b*. The first shows a sample of $n = 60$ from a bivariate elliptical t_4 (i.e., t with four degrees of freedom) distribution with correlation $\rho = .5$. The second is for a like sample from an elliptical Cauchy distribution. Technically, ρ is not defined in this case, but it is convenient to think of it as the correlation in the associated bivariate normal distribution of X, which again is $\rho = .5$. To facilitate comparison, ellipses which theoretically contain 75 and 90 percent of the probability have been drawn and the same scale has been kept in both figures. The difference in the sizes of the ellipses corresponding to the same percent in the two figures shows how much

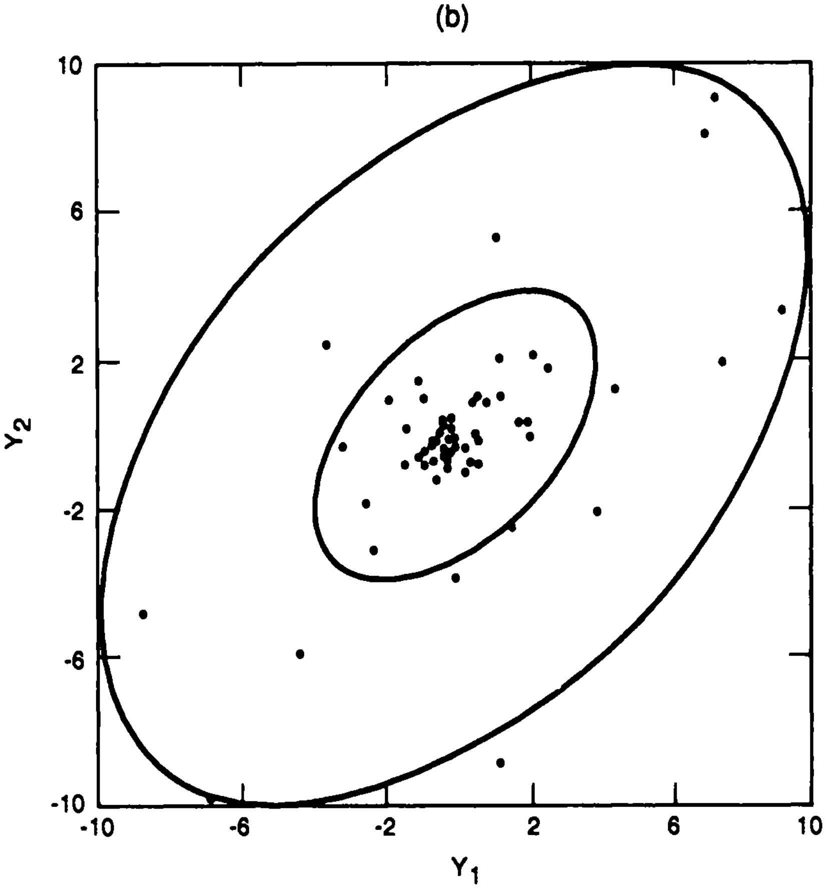


Fig. 11b. Bivariate Cauchy sample with 75% and 90% probability contours

longer the "tail regions" of the Cauchy are relative to the t with four degrees of freedom.

Devlin et al. (1976) discuss a number of additional properties and uses of elliptical distributions and compound normal distributions, including the behavior of Fisher's z-transformation of the correlation coefficient, influence functions, and the distribution of Mahalanobis' D^2.

For instance, the asymptotic distribution of $\sqrt{n}(z(r) - z(\rho))$, where $z(\ )$ is the z-transformation, for samples from a compound multinormal distribution is $N(0, \phi^2)$. This result demonstrates that Fisher's z-transform is variance stabilizing for sampling from all compound normal distributions with finite fourth moments. Devlin et al. (1976) also discuss radii-and-angles method for assessing goodness-of-fit of compound multinormal distributions analogous to those described in Section 5.2.2 for assessing multivariate normality. Indeed, as a consequence of J2, there is no difference (and, hence, no discrimination capabilities) at all in the distributional properties of the angles while the distribution of the radii will be longer tailed for compound multinormal distributions, such as the multivariate t-distribution, than for the multivariate normal case. Specifically, for sampling from multivariate t-distributions with degrees of freedom $\geqslant 3$, the distribution of the squared radius (i.e., Mahalanobis' D^2) is approximately $(f - 2)\chi_p^2/\chi_f^2$. This property can be used to make a probability plot of the ordered observed squared radii values against the corresponding quantiles of an F-distribution with p and f degrees of freedom. The null configuration on such a plot, confirming that the data have a multivariate t-distribution, would be linear with a slope, $p(f - 2)/f$.

REFERENCES

Section 5.1 Anderson (1985), Puri & Sen (1971), Rao (1965), Roy (1957).

Section 5.2.1 Anderson (1984), Chambers (1977), Golub & Reinsch (1970), Pillai & Jayachandran (1967), Roy & Bose (1953), Roy et al. (1971), Scheffé (1953, 1959), Stein (1956), Wilkinson (1970).

Section 5.2.2 Anderson (1984, 1969), Jöreskog (1977), Lindley (1972), Rao (1965), Roy & Gnanadesikan (1957, 1962), Srivastava (1966), Stein (1956, 1965).

Section 5.2.3 Andrews et al. (1972), Bickel (1964), Devlin et al. (1975, 1981), Gentleman (1965), Gnanadesikan & Kettenring (1972), Hampel (1968, 1971), Hampel et al. (1986), Huber (1964, 1970, 1972, 1973, 1977, 1981), Johnson & Leone (1964), Krasker & Welsch (1982), Lax (1975), Mallows (1973, 1983), Maronna (1976), McLaughlin & Tukey (1961), Mood (1941), Rousseeuw (1983), Teichroew (1956), Tukey (1960), Tukey & McLaughlin (1963), Tyler (1983), Wilk & Gnanadesikan (1964), Wilk et al. (1962).

Section 5.3 Andrews et al. (1971), Box & Cox (1964), Chambers (1973), Cox (1970, 1972), Moore & Tukey (1954), Puri & Sen (1971), Tukey (1957).

Section 5.4 Kendall (1968).

Section 5.4.1 Aitchison & Brown (1957), Chambers (1973), Gnanadesikan (1972), Laue & Morse (1968), Wilk & Gnanadesikan (1968).

Section 5.4.2 Aitkin (1972), Anderson (1966), Andrews (1971), Andrews, Gnanadesikan & Warner (1971, 1972, 1973), Box & Cox (1964), D'Agostino (1971), D'Agostino & Pearson (1973), David & Johnson (1948), Devlin et al. (1975), Downton (1966), Filliben (1975), Gnanadesikan & Kettenring (1972), Gnanadesikan et al. (1967), Healy (1968), Kessell & Fukunaga (1972), Malkovich & Afifi (1973), Mardia (1970, 1975, 1980), Pearson & Hartley (1966), Rosenblatt (1952), Roy (1953), Sarhan & Greenberg (1956), Shapiro & Wilk (1965), Shapiro et al. (1968), Weiss (1958).

Section 5.4.3 Andrews & Mallows (1974), Blake & Thomas (1968), Chu (1973), Das Gupta et al. (1972), Devlin et al. (1976), Goldman (1974), Johnson & Katz (1972), Kelker (1970), Kingman (1972), Picinbono (1970), Rogers & Tukey (1972), Vershik (1964), Yao (1973).

Methods for Statistical Data Analysis of Multivariate Observations, Second Edition
by R. Gnanadesikan

CHAPTER 6

Summarization and Exposure

6.1. GENERAL

The main function of statistical data analysis is to extract and explicate the informational content of a body of data. The processes of description and communication of the information involve *summarization*, perhaps in terms of a statistic (e.g., a correlation coefficient) which may be undergirded by some reasonably tightly specified model or, perhaps, in terms of a simple plot (e.g., a scatter plot). In addition to the well-recognized traditional role of summarization, however, the meaningful exercise of processes of data analysis requires *exposure*, that is, the presentation of analyses so as to facilitate the detection not only of anticipated but also unexpected characteristics of the data. For instance, an x–y scatter plot of data is a pictorial representation that is useful not only for interpreting the computed value of the correlation coefficient for that body of data (see also the scatter plots described in Section 6.4.2), but also for indicating the adequacy of assuming linearity of the relationship between x and y. A more substantial example of the twin-pronged process of summarization and exposure is fitting a straight line to y versus x data and then studying the residuals in a variety of ways, especially through different plots of them, such as plots of residuals against observed values of x and of y, perhaps against values of relevant extraneous variables such as time, and also probability plots of the residuals. Although the fitting provides summarization, the study of the residuals is often crucial in exposing the inadequacies of various assumptions that underlie the fitting procedure (e.g., constancy of variance).

Pedagogy, publications and, more generally, the codification of statistical theory and methods have been concerned almost exclusively with formal procedures such as tests of hypotheses, confidence region estimation, and various optimality criteria and associated methodology. Even when the concern has been with developing methods for summarization, with a clear awareness of the possible inadequacies or inappropriateness of certain "standard" assumptions, the goal of summarization has often been somewhat artificially separated from that of exposure. For instance, much of the work on robust estimation, while usefully concerned with summary statistics that are not unduly influenced by a small fraction of possibly deviant observations, has adopted the formal and familiar framework of point estimation with its criteria of bias, efficiency, and so on, and relatively little attention has been given to the exposure value of such things as the residuals obtained from using the robust estimates in place of the standard estimates.

The manifold theories of statistical inference that have been advanced as the focal points for "unifying" statistics have only relatively recently (Cox, 1973; Cox & Hinkley, 1974) been considered and carefully scrutinized in terms of their relevance for and relationship to the needs of applications of statistics. No single formal theory of statistical inference seems able to encompass and completely subsume the flexible and interactive processes involved in summarization and exposure. There are, however, less formal techniques that are, perhaps, not in the mainstream of any formal statistical theory but that are nevertheless useful tools of informative inference directed toward the dual objectives of summarization and exposure.

The treatment in this book has attempted, even when a problem has been formulated fairly narrowly (e.g., tests for commonality of marginal distributions), to combine formal methods, where available, with informal procedures for revealing the relevant information in multiresponse data. A feature common to most of the informal procedures is their graphical nature. In the following sections of this chapter, some general problem areas of summarization and exposure are distinguished (and inevitably these overlap to some degree the concerns of the earlier chapters), and some techniques of relevance to these problems are discussed. The emphasis throughout this chapter is on graphical methods.

6.2. STUDY OF AN UNSTRUCTURED MULTIRESPONSE SAMPLE

One is often interested in examining a body of data *as if* it were an unstructured collection or sample, and many of the techniques discussed in the earlier chapters of this book have been described in the context of analyzing an unstructured sample (e.g., linear and generalized principal components analysis in Chapter 2, robust estimates of location and dispersion in Chapter 5, and the assessment of distributional properties in Chapter 5). With uniresponse data several graphical and semigraphical techniques are available for analyzing an unstructured sample. Some examples of such techniques are stem-and-leaf displays and box plots (see Chapters 1 and 5 of Tukey, 1970) and the more familiar histogram, empirical cumulative distribution function (or *ecdf*, which may be defined as a plot of the ith ordered observation against $(i - \frac{1}{2})/n$), and the class of techniques loosely called *probability plotting methods* (see Wilk & Gnanadesikan, 1968).

Wilk & Gnanadesikan (1968) describe two basic types of probability plots, called *P-P* and *Q-Q* plots, respectively. Figure 12 may be used for defining the two types. In comparing two distribution functions, a plot of points whose coordinates are the cumulative probabilities $\{p_x(q), p_y(q)\}$ for different values

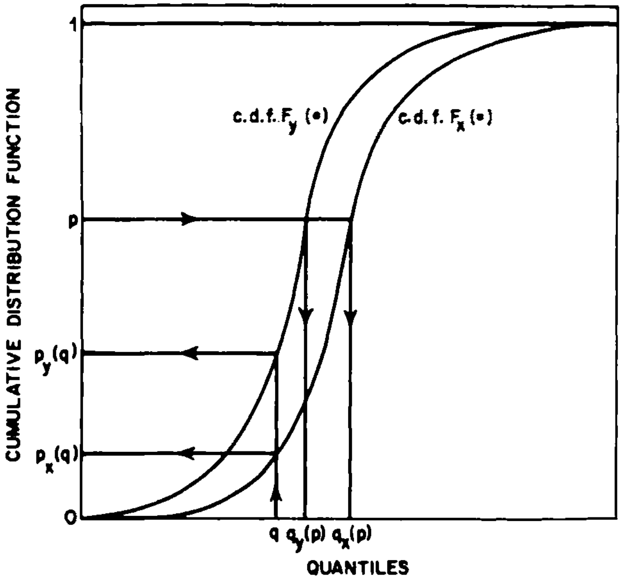


Fig. 12. Illustration for *P*-*P* and *Q*-*Q* plots.

of q is a *P*-*P* plot, while a plot of the points whose coordinates are the quantiles $\{q_x(p), q_y(p)\}$ for different values of p is a *Q*-*Q* plot. For conceptual convenience both of the distribution functions displayed in Figure 12 are shown as smooth curves, but this is not an essential requirement, in that one or both of the distribution functions involved can be a step function or an ecdf. In fact, the usual form of the comparison is one in which an ecdf for a body of univariate data is compared with a specified (or theoretical) distribution function, that is, a step function is compared to a continuous one. Also, *Q*-*Q* probability plots tend to be more widely used than *P*-*P* probability plots. Perhaps one reason for this is a property of linear invariance possessed by *Q*-*Q* but not *P*-*P* plots, namely, when the two distributions involved in the comparison are possibly different only in location and/or scale, the configuration on the *Q*-*Q* plot will still be linear (with a nonzero intercept if there is a difference in location, and/or a slope different from unity if there is a difference in scale), whereas the configuration on a *P*-*P* plot will in general be *necessarily* linear (with zero intercept and unit slope) only if the two distributions are identical in all respects, including location and scale.

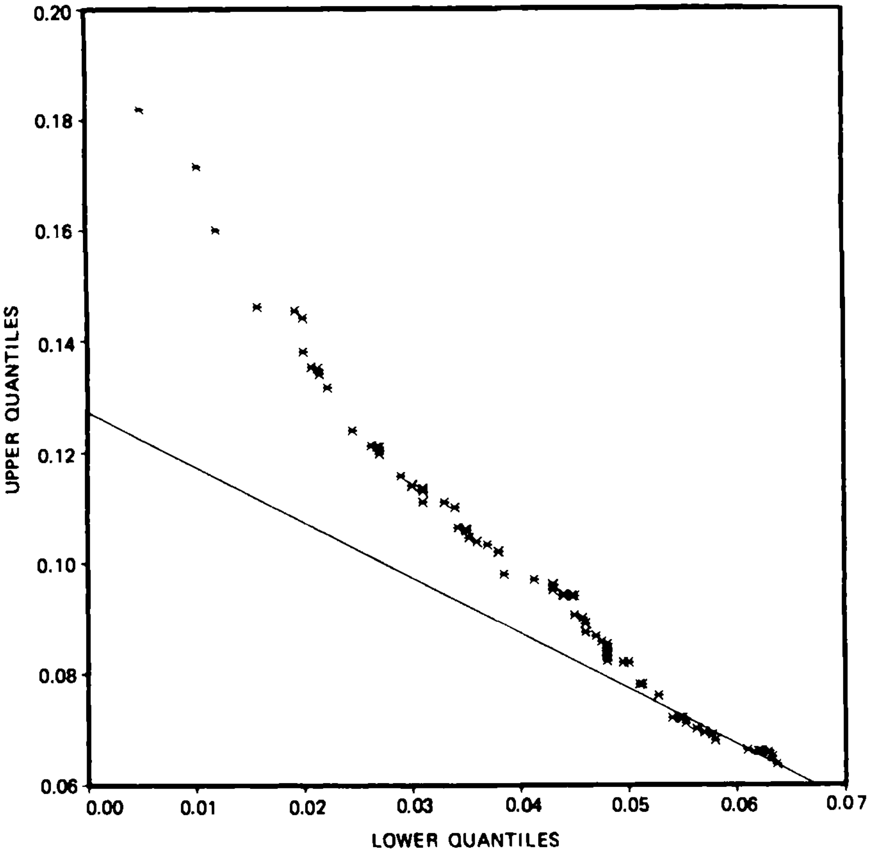
At any rate the following is a canonical description of a *Q*-*Q* probability plot in its most widely used form, wherein an ecdf of an unstructured sample, $x_1, \ldots, x_n$, of size n is to be compared with a hypothesized standardized (i.e., origin or location parameter is zero and scale parameter is unity) distribution function $F(x; \boldsymbol{\theta})$ (where the parameters $\boldsymbol{\theta}$, which do not include origin or location and/or scale parameters, have specified values): if $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$

are the ordered observations, plot the n points $\{\tilde{x}_i, x_{(i)}\}$, $i = 1, \ldots, n$, where $\tilde{x}_i$ is the quantile of the distribution function F corresponding to a cumulative probability p_i $[=(i - \alpha)/(n - 2\alpha + 1)$ with $\alpha = \frac{1}{2}, \frac{1}{3}$, or 0 as some of the possible choices$]$, that is, $\tilde{x}_i$ is defined by $F(\tilde{x}_i; \boldsymbol{\theta}) = p_i$. The well-known use of normal probability paper for plotting data (which was mentioned, for instance, in Section 5.4.2 in connection with evaluating marginal normality) is an example of the making of a Q-Q plot with F taken as the distribution function Φ of the standard normal distribution. Other examples of specification of F are a chi-squared distribution with a specified degree of freedom, a gamma distribution with a specified shape parameter, and a beta distribution with values for both of its shape parameters. (See Wilk et al., 1962a; Gnanadesikan et al., 1967.)

For uniresponse observations, in addition to ecdf's, probability plots, and the other displays mentioned above, there are some simple graphical displays for aiding in the assessment of symmetry of the data distribution. From the viewpoint of multiresponse data analysis, symmetry of the marginal distributions of each response is not an unreasonable requisite for the meaningful use of several summary statistics such as correlation coefficients and covariance matrices. If the raw data are quite asymmetric, a preliminary transformation of the observations (perhaps by the methods of Section 5.3) to enhance symmetry will often be a sensible first step before the subsequent univariate or multivariate analyses that may be performed on the transformed data.

The ecdf itself is often a good means of studying symmetry. However, other plots specifically useful for investigating possible asymmetry in data can also be made. For instance, if $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n-1)} \leqslant x_{(n)}$ denote the ordered observations, Wilk & Gnanadesikan (1968) have suggested plotting the points whose coordinates are $\{x_{(1)}, x_{(n)}\}$, $\{x_{(2)}, x_{(n-1)}\}$, etc. If the observations are symmetric around a center of symmetry $x = b$, such a plot should look reasonably linear with intercept approximately equal to $2b$ and slope approximately equal to -1. Departures from such a linear configuration will indicate the type of asymmetry present in the data. For instance, an upward bow to the plot will indicate a longer upper tail; a downward bow, a longer lower tail. In a variant of this plot for assessing symmetry the points plotted have coordinate values that are deviations from the median of the observations, that is, the points plotted are $\{x_M - x_{(1)}, x_{(n)} - x_M\}$, $\{x_M - x_{(2)}, x_{(n-1)} - x_M\}$, etc., where x_M denotes the median. For symmetric data the configuration on such a plot will, therefore, be linear with zero intercept and unit slope, and departures from such a "null" configuration can be easily appreciated and interpreted. Another plotting procedure for studying symmetry, proposed by Tukey (see Wilk & Gnanadesikan, 1968), consists in plotting the points whose coordinates are, respectively, differences and sums of the symmetrically situated pairs of ordered observations $x_{(i)}$ and $x_{(n-i+1)}$, that is, the plotted points are $\{x_{(n)} - x_{(1)}, x_{(1)} + x_{(n)}\}$, $\{x_{(n-1)} - x_{(2)}, x_{(2)} + x_{(n-1)}\}$, etc. This is a scheme for "tilting" the plots so that the "null" configuration becomes a horizontal linear one, and departures of the data from symmetry will be indicated by deviations from horizontality. The next example illustrates the use of these three graphical methods for assessing symmetry.

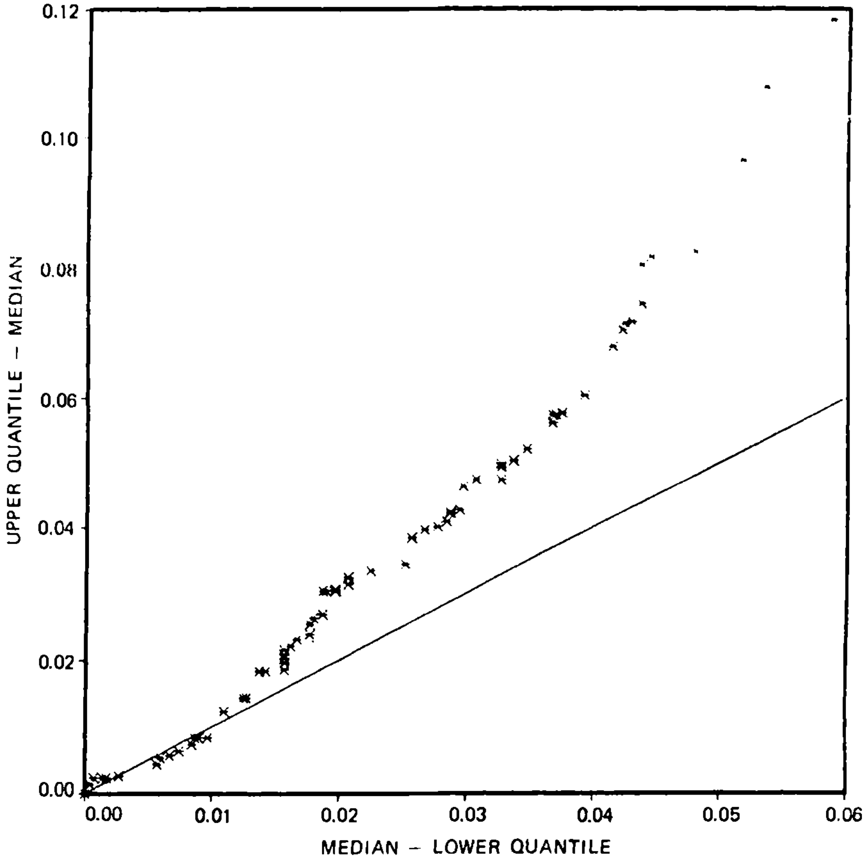
Exhibit 38a. Plot of upper vs. lower half of the sample for ozone data



Example 38. The observations are maximum daily ozone measurements (in ppm) as observed at a particular air monitoring site in New Jersey during certain months in 1973. Exhibit 38*a* shows a plot of the symmetrically situated ordered observations $\{x_{(i)}, x_{(n-i+1)}\}$, $i = 1, \ldots, [n/2]$, and the upward bow of the plot suggests a positively skewed distribution for the observations. Exhibit 38*b* (see page 232) shows a plot of the deviations from the median, viz. $\{x_M - x_{(i)}, x_{(n-i+1)} - x_M\}$ for $i = 1, \ldots, [n/2]$, and the departure from the line of zero intercept and unit slope is quite strikingly indicative of a positively skewed distribution. Exhibit 38*c* (see page 233) shows the plot suggested by Tukey; the systematic deviation from a horizontal linear configuration is evident here too.

In an attempt to improve the symmetry of the data in this example, square roots of the observations were taken, and Exhibit 38*d* (see page 234) shows a plot of the deviations from the median for the transformed data. A comparison with Exhibit 38*b* reveals the clear accomplishment of the square-root trans-

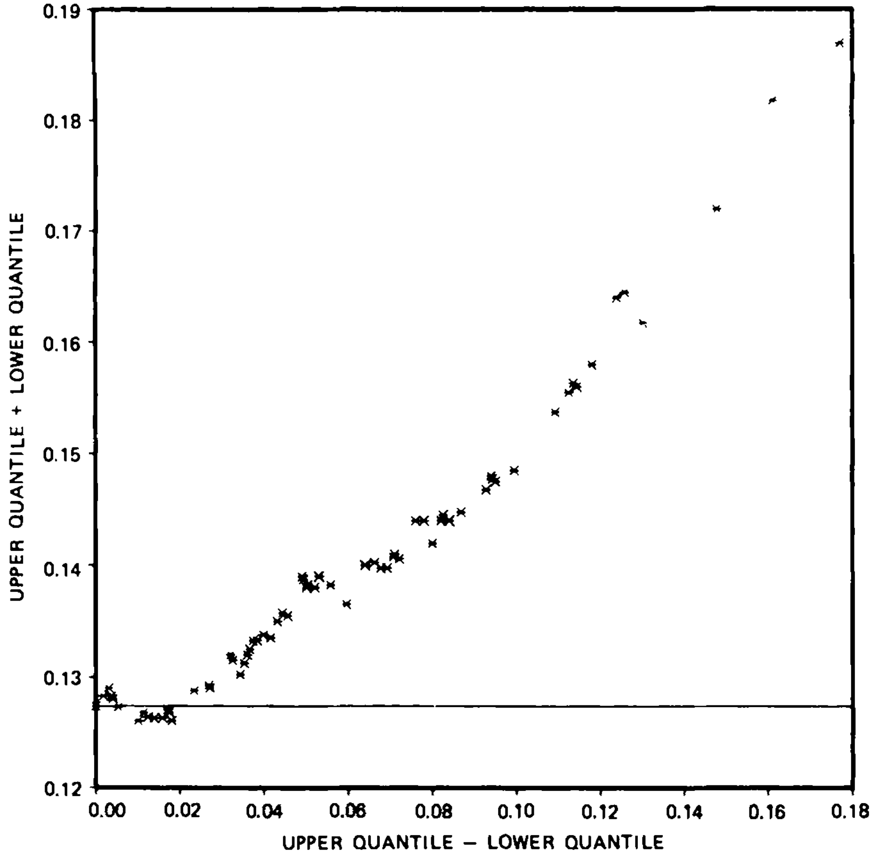
Exhibit 38*b*. Plot of deviations of upper quantiles from median vs. deviations of lower quantiles from median for ozone data



formation in symmetrizing the data. The lognormal distribution has often been used as the model for ambient air quality measurement data (see for example, Zimmer & Larsen, 1965). For the present data Exhibit 38*e* (see page 235) shows a plot of deviations from the median when logarithms of the observations are used, and a comparison of Exhibits 38*b*, *d*, and *e* reveals that the square-root transformation is better for symmetrizing the data than the logarithmic transformation, which results in a negatively skewed distribution for the data in this example. More extensive evidence in favor of the square-root transformation in connection with ambient air quality data is contained in the work of Cleveland et al. (1975).

For multiresponse data there does not seem to be any extension of uniresponse *Q*-*Q* probability plotting, perhaps because no unique (or even generally useful) way of defining quantiles is available. Even more basically,

Exhibit 38c. Plot of sum of and difference between upper and lower quantiles for ozone data

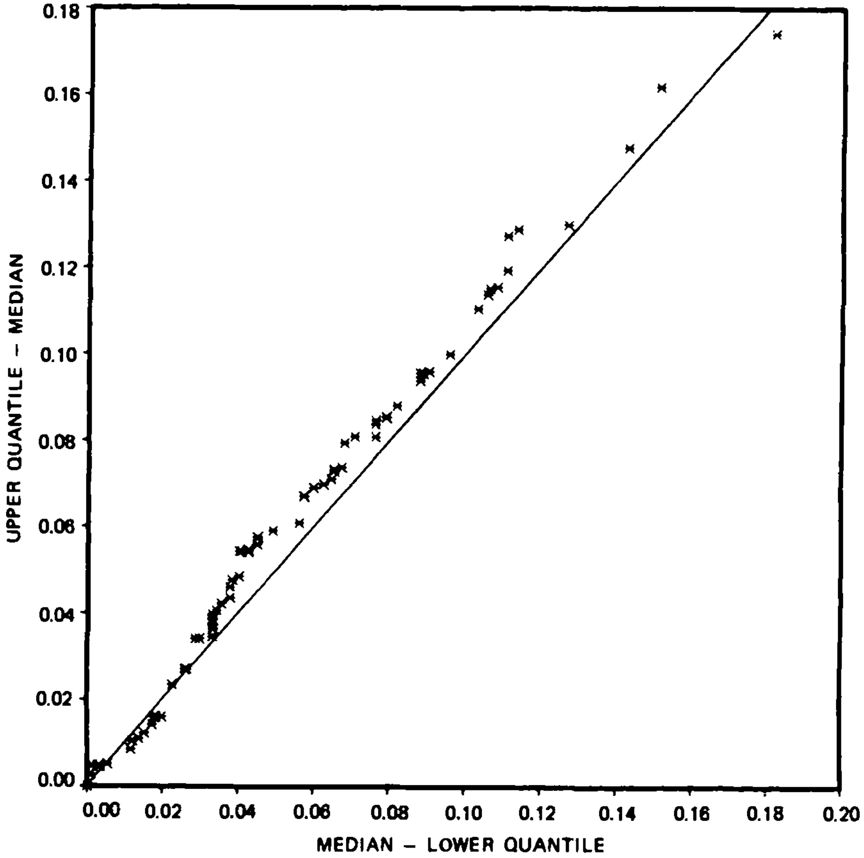


convenient graphical representations of multivariate (especially with $p \geqslant 3$) histograms of the kind proposed by Hartigan (1973) have not been widely implemented.

Nevertheless some things *can* be done to provide insights into multiresponse data configurations, and a few graphical techniques (some of which have been described and used in earlier chapters) are worth explicit mention here.

1. Two- and three-dimensional scatter plots of bivariate and trivariate subsets of the original data can be useful for studying cohesiveness, separations within the sample, possible outliers, and general shape. Devlin et al. (1975) have suggested a way of augmenting the pictorial value of two-dimensional scatter plots for judging the effects of individual observations on a correlation coefficient computed from the points exhibited in a scatter plot. The suggestion is to display on the scatter plot the contours of a so-called influence function

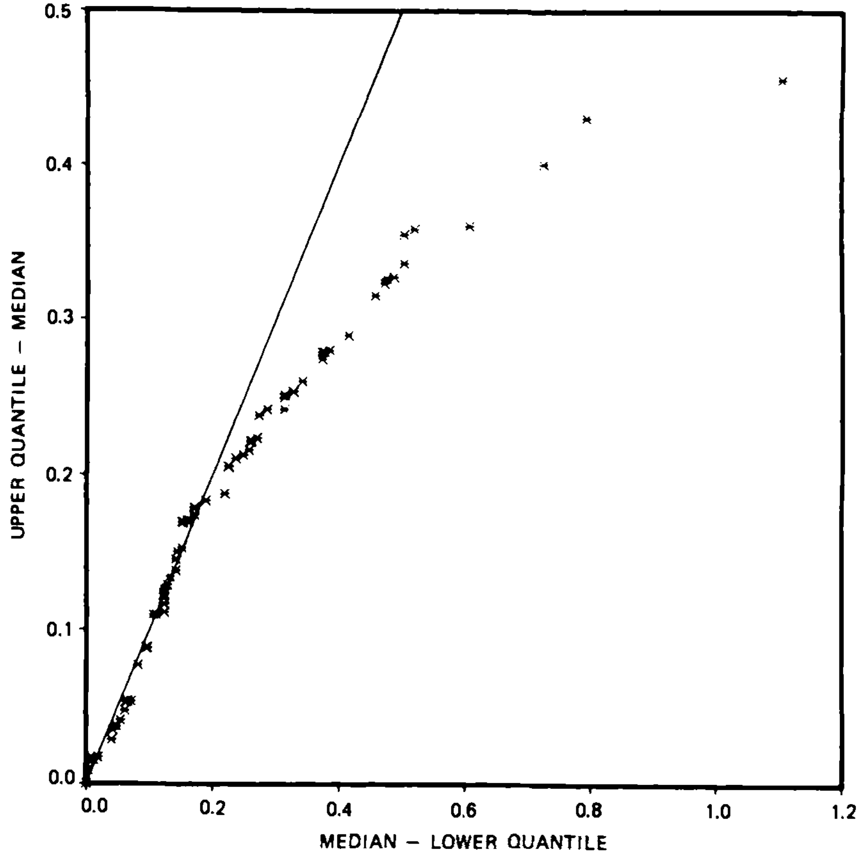
Exhibit 38d. Plot of deviations of upper quantiles from median vs. deviations of lower quantiles from median for square roots of ozone data



(see Hampel, 1968, 1974) of the correlation coefficient so as to enable one not only to gain an overall appreciation of the strength of the correlation but also to gauge how much the computed value of the correlation coefficient can be altered (inflated or deflated) by the omission of individual observations. An example of a scatter plot with superimposed influence function contours is given in Section 6.4.2.

With the availability of interactive graphical display facilities, one could sweep through a series of two-dimensional projections in addition to those onto the original coordinate planes and gain a good appreciation of the structure of high-dimensional data. The PRIM-9 system developed by Fisherkeller et al. (1974) was an early implementation of such an interactive, dynamic graphical display. A number of later schemes for dynamic displays, incorporating both flexible user-interaction and aids for selecting interesting projections of high-dimensional observations, have emerged subsequently. Many of these schemes have been motivated by finding clusters and were mentioned in

Exhibit 38e. Plot of deviations of upper quantiles from median vs. deviations of lower quantiles from median for logarithms of ozone data



Section 4.3.2 (e.g., Azimov et al., 1988; Buja & Hurley, 1990; Cook et al., 1993; Swayne et al., 1991).

2. Probability plots of the observations on each response separately will generally be useful, in conjunction with other multivariate analyses. A natural base or starting point for such plotting will often be the normal distribution. Such plots may indicate the desirability of marginal transformations or of more appropriate and insightful kinds of probability plots.

3. Scatter plots and/or probability plots of projections onto eigenvectors from either linear or generalized principal components analysis can also be made and studied with benefit in many cases.

4. Joint evaluation of the eigenvalues of a sample covariance or correlation matrix is a problem often associated with the analysis of an unstructured multiresponse sample. Such evaluations may, for instance, provide the key to possible reduction of dimensionality. The adequate assessment of specific

sample eigenvalue results is not, however, an elementary task. The fact is that, even for large samples from spherical multivariate distributions, the eigenvalues may exhibit substantial variability (see Example 13 of Gnanadesikan & Wilk, 1969). A *scree* plot, which is a plot of the ordered eigenvalues against their ranks, is often used for studying the separations amongst eigenvalues (see Cattell, 1966). One looks for "elbows" in the plot to decide where the separations might be. However, this can turn out to be very difficult in many problems where the fall off in the eigenvalues tends to look like a smooth exponential curve. Buja & Eyuboglu (1993) have suggested a useful augmentation of scree plots for assessing the eigenvalues of a correlation matrix. Their proposal is discussed below. Example 40 below illustrates the use of the standard scree plot and the enhancement proposed by Buja & Eyuboglu.

Rather than plotting the eigenvalues against their rank order, a *Q-Q* type of probability plot has been proposed (Gnanadesikan, 1968, 1973). The idea is to plot the ordered eigenvalues against their expected (or some other typical) values, determined by using the null assumption of sampling from a standard spherical normal distribution. The plotting positions (i.e., the "expected" values) may be determined by either a simple or a more sophisticated and efficient (Hastings, 1970) Monte Carlo approach. The work of Stein (1969) and of Mallows & Wachter (1970) on asymptotic configurations of eigenvalues of Wishart matrices also provides a basis for plotting the ordered eigenvalues of a covariance (albeit not a correlation) matrix against a set of corresponding quantiles of a particular distribution.

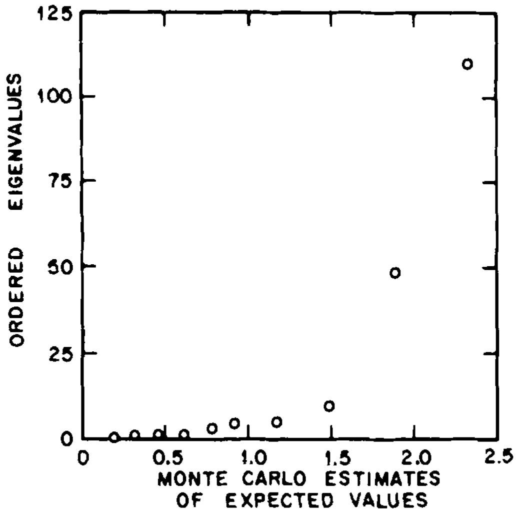
Example 39. For this example 25 random deviates from a 10-dimensional normal distribution were generated. The underlying dispersion of the computer-generated data was much larger along two of the coordinates than along the other eight, so that the variability observed in the 10-dimensional sample would be expected to be confined largely to a two-dimensional subspace.

Exhibit 39 shows a plot of the 10 ordered eigenvalues of the sample covariance matrix against simple Monte Carlo estimates of their respective expected values under sampling from a 10-dimensional standard spherical normal distribution. The two largest eigenvalues clearly deviate from the configuration indicated by the smaller eight eigenvalues. Replotting (i.e., redetermination of the plotting positions on the basis of sampling from an eight-dimensional standard spherical normal) the smaller eight eigenvalues would be useful for studying the cohesiveness of and/or groupings among them.

This *Q-Q* type of graphical analysis of eigenvalues is useful not only for isolating large eigenvalues but also for identifying cases in which the overall space of the responses is decomposable into subspaces within each of which the dispersion of the points is essentially spherical; this will be indicated by a plot that has several linear pieces with differing slopes.

Buja & Eyuboglu (1993) suggest a different method for assessing the separations among the eigenvalues of a correlation matrix. The central idea in

Exhibit 39. "*Q-Q*" plot of eigenvalues



the approach is to randomly permute the initial data and thence generate permutation distributions of the eigenvalues which can then be used to provide the quantiles of the distribution of each ordered eigenvalue. The quantiles are added to a scree plot and serve as aids to judging the deviations of the observed eigenvalues from a "null" model that specifies them all to be equal. The null model is thus equivalent to specifying the true correlation matrix to be the identity matrix.

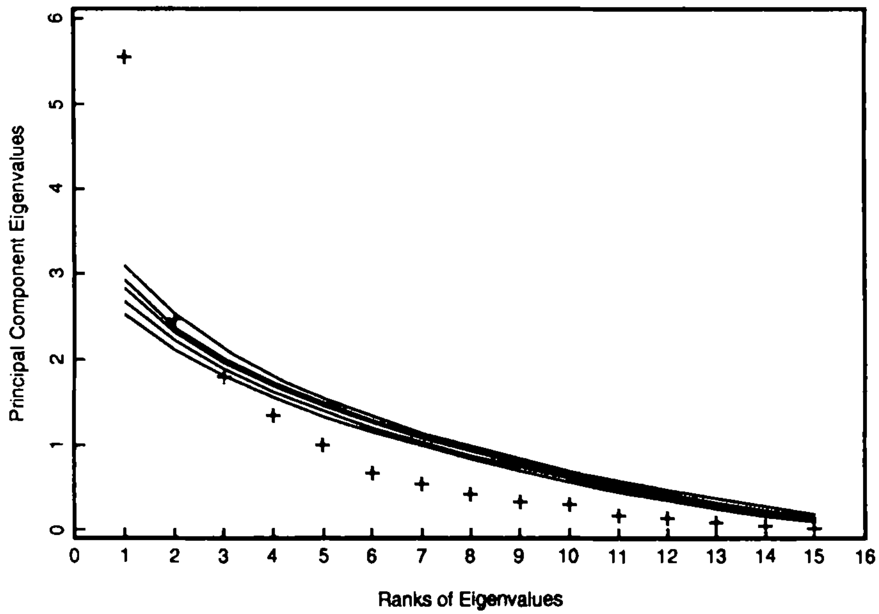
More specifically, starting with the $p \times n$ data matrix, **Y**, Buja & Eyuboglu suggest the following as "null" assumptions: (a) the ith row of **Y**, consisting of the n observations on the ith response, is a random sample from a univariate distribution with distribution function, F_i, for $i = 1, \ldots, p$; and (b) the p response variables are mutually independent (which would imply that the correlation matrix is **I**). Under these assumptions, the joint null distribution of the multiresponse observations (and hence the null distribution of statistics such as the eigenvalues of the sample correlation matrix) is invariant under permutations within rows, there being $(n!)^p$ possible permutations in all. In fact, given that any single row of **Y** is itself a random realization, one can limit the number to $(n!)^{p-1}$ permutation, which is still a very large number even for moderate values of n and p. To generate the required permutation distributions of the eigenvalues of the correlation matrix, the suggestion is to take a finite, perhaps large, number of permutations and hope that the required permutation distributions under null assumptions are reasonably well determined. In practice, for moderate values of p and n, using 500–1000 random permutations may suffice but experimentation in the given context may be wise. At any rate, the method consists of generating an adequate number of row-permuted "samples" from the data, computing the correlation matrix and its eigenvalues from each permutation, and, finally, determining a set of empirical quantiles (e.g., median, lower and upper 25%, 10%, 5%, 1% values) from the distribution of each eigenvalue separately and superimposing the loci of these quantiles onto a scree plot of the observed eigenvalues of the original data. Those eigenvalues that are beyond the outer, upper quantile curves may be judged as being really different from other eigenvalues which are within the bands of the inner quantiles (e.g., quartiles).

The simulations involved in the *Q-Q* type of plot suggested earlier can, of course, be used for calculating such things as standard deviations and even the quantiles of the distributions of the eigenvalues, which can then be used as aids in assessing departures, if any, from the null linear configuration of the plot to be expected when the sampling is from $N[\mathbf{0}, \mathbf{I}]$. A major difference between the method proposed by Buja & Eyuboglu and the *Q-Q* type of plot is that the latter is based on simulated samples from a normal distribution, whereas the former is nonparametric in nature, since it is based on random permutations of the observed data. On the other hand, the *Q-Q* type plot can be generated for studying separations among the eigenvalues of either the covariance matrix, S, or the correlation matrix, **R**, or even robust versions of these, whereas the method of Buja and Eyuboglu is confined to correlation matrices including robust versions (e.g., $\mathbf{R}_2^*(+)$ or $\mathbf{R}_3^*$ defined in Section 5.2.3) and also the reduced correlation matrix involved in the principal factor analysis method (see Section 2.2.2).

Example 40. This example, taken from Buja & Eyuboglu (1993), illustrates the use of the augmented scree plot. The data concern 15 questionnaire items that are intended to measure the bargaining behavior of opponents as rated by 28 subjects involved in a stylized psychological experiment. Using 499 random permutations of the data matrix, Buja & Eyuboglu obtained the median, the upper quartile, and the 90th, 95th, and 99th percentiles of the permutation distributions of the 15 eigenvalues of the correlation matrix. Exhibit 40 shows the augmented scree plot obtained by them. The fact that the largest eigenvalue is well separated from the second and later eigenvalues would perhaps have been evident even from the simple scree plot for this example. The augmented scree plot, however, provides a clearer calibration of how far out in the tail of the "null" distribution the largest eigenvalue is and, moreover, is helpful in aiding the conclusion that the second eigenvalue is beyond the 95th percentile. The conclusion in this example would be that the top two eigenvalues are worth considering as distinct and well separated from the remaining ones which appear to be estimates of a common value.

5. The use of one, or preferably several, distance functions to convert the multiresponse data to single numbers, followed by the probability plotting of these numbers, can be very effective. A common useful class of distance functions is that of positive semidefinite quadratic forms, $\mathbf{x}'\mathbf{A}\mathbf{x}$, where both the vector, $\mathbf{x}$, and the matrix, $\mathbf{A}$, may be some functions of the multiresponse observations themselves. Example 7 discussed in Section 2.4, as well as Examples 44–46 described in Section 6.3.1, illustrate the idea involved here.

Exhibit 40. Scree plot with superimposed percentiles from permutation distributions of the eigenvalues (Buja & Eyuboglu, 1993). [Percentiles from the bottom: median, 75th, 90th, 95th, and 99th].



6. The CPP and SCPP techniques of component probability plotting described in Section 5.4.1, and the plotting procedures associated with the radius-and-angles decomposition discussed in Section 5.4.2, are additional examples of graphical methods that are useful in studying the distributional characteristics of multiresponse data.

7. Last, a technique proposed by Andrews (1972) and certain ramifications of it constitute promising developments in the graphical display of high-dimensional data. The rest of this subsection is devoted to a discussion and illustration of this class of displays.

The essential idea in Andrews's proposal is to map each multiresponse observation into a function, $f(t)$, of a single variable, t, by defining $f(t)$ as a linear combination of orthonormal functions in t with the coefficients in the linear combination being the observed values of the responses. For instance, given the p-dimensional observations, $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ip})'$, $i = 1, \ldots, n$, one can map each observation, $\mathbf{y}_i$, into

$$\begin{aligned} f_{\mathbf{y}_i}(t) &= f_i(t) \\ &= y_{i1}a_1(t) + y_{i2}a_2(t) + \cdots + y_{ip}a_p(t) \\ &= \mathbf{y}_i'\mathbf{a}_t, \qquad i = 1, \ldots, n, \end{aligned} \tag{82}$$

where the functions $\{a_1(t), a_2(t), \ldots, a_p(t)\}$ are orthonormal in an interval, say

$0 \leqslant t \leqslant 1$. Specifically, Andrews (1972) suggests the set of functions

$$\begin{aligned}\mathbf{a}_t' &= \{a_1(t), a_2(t), \ldots\} \\ &= \left\{\frac{1}{\sqrt{2}}, \sin t, \cos t, \sin 2t, \cos 2t, \ldots\right\},\end{aligned} \tag{83}$$

which are orthonormal on $(-\pi, +\pi)$. [*Note*: Simply by taking $2\pi t$ in place of t in Andrews's definition one would obtain a set of functions orthonormal on $(0, 1)$ instead of $(-\pi, +\pi)$.]

The n functions, $f_1(t), f_2(t), \ldots, f_n(t)$, may then be plotted simultaneously against values of t in the permissible range, for example, $(0, 1)$ or $(-\pi, +\pi)$. Thus the initial multiresponse observations, which are n points in p-space, will now appear as n curves in a two-dimensional display whose ordinate corresponds to the function value and whose abscissa is the range of values of t. At a specific value of t, say $t = t_0$, $f_i(t_0)$ is the length of the projection of the ith observation, $\mathbf{y}_i$, onto the vector (or one-dimensional subspace) $\mathbf{a}_{t_0}' = \{a_1(t_0), a_2(t_0), \ldots, a_p(t_0)\}$. Thus, on the Andrews *function plot*, at a specific value of t one is looking at the lengths of the projections of each of the n observations onto a specific one-dimensional subspace of the original p-space, and, as one scans the plot across several values of t, one is looking at a collection of several such one-dimensional views. An equivalent algebraic way of thinking about the function plot is that at each value of t one is looking at a specific linear combination of the p responses, and thus across different values of t one is looking at several different linear combinations.

Andrews (1972) has established various statistical properties of these function plots. For instance, since the definition of the functions in Eq. 82 is linear in the p variables, the technique preserves the mean in the sense that, if $\bar{\mathbf{y}}$ denotes the mean vector of the observations, then

$$f_{\bar{\mathbf{y}}}(t) = \frac{1}{n} \sum_{i=1}^{n} f_i(t),$$

so that the centroid of the observations will correspond to an "average curve" on the function plot. Another property of the function plot is that it preserves distances in a certain sense. Specifically, as a consequence of the orthonormality of the functions $\{a_j(t)\}$, $j = 1, \ldots, p$, the squared distance between the pair of functions $f_i(t)$ and $f_l(t)$, defined as

$$\int\limits_{\substack{\text{over the}\\ \text{total range}\\ \text{of } t}} [f_i(t) - f_l(t)]^2 \, dt,$$

is just proportional to the squared Euclidean distance between $\mathbf{y}_i$ and $\mathbf{y}_l$ in the p-space of the original observations. This property enables one to think of close curves on the function plot as corresponding to close data points (at least as judged by Euclidean distance, which may not itself be a statistically appropriate measure of distance for certain kinds of multiresponse data, as discussed in Section 4.2.1) in the p-space of the responses.

Yet another property of the plot is that, provided the p responses have equal variances and are mutually uncorrelated, the variance across values of t in the function plot is essentially constant. This is so because $\mathbf{a}_t'$ as defined by Eq. 83, for instance, is of constant length $(=\sqrt{p/2})$ when p is odd, and when p is even and large it is of approximately constant length since its length is between $\sqrt{(p-1)/2}$ and $\sqrt{(p+1)/2}$. Consequently, if the p responses have a common variance σ^2 and, furthermore, are mutually uncorrelated, it follows from Eq. 82 that the variance σ_f^2 of $f(t)$ (defined with $\mathbf{a}_t'$ as in Eq. 83) is $\sigma^2 p/2$ when p is odd and lies between $\sigma^2(p-1)/2$ and $\sigma^2(p+1)/2$ when p is even. The requirement of a common variance for, and no intercorrelations among, the responses is, however, not only unrealistic but also self-defeating, in that if this were so the case for a multivariate approach to analyzing the data would not be very cogent. In practice, two different ways of moving the data toward meeting the requirement for constancy of variance of the function plot are (i) rotating the data to standardized principal component coordinates, and (ii) standardizing the variables initially (e.g., by scaling the observations on a response by either the standard deviation or the interquartile range) without any attempts to uncorrelate the data.

From the definition of $f_i(t)$ in Eq. 82, it is clear that the choice of the specific elements of $\mathbf{a}_t'$ to associate with each of the variables can be important. For instance, the suggestion in Eq. 83 would associate $1/\sqrt{2}$ with the first variable, $\sin t$ with the second, and so on. A different permutation of the coefficients would, of course, lead to weighting the variables differently. One suggestion for using a specific ordering of coefficients such as the one in Eq. 83 is to take the variables in the order of their importance; however, such an ordering by importance may not always be feasible, and as a general rule it may be advisable to try a few different permutations of the coefficients with a given set of variables. Since the appearance of the function plot is not invariant under permutations of the coefficients, the use of different permutations may lead to different insights into the data and thus prove to be valuable.

Ideally, as t varies across its total range of values, the values assumed by the vector $\mathbf{a}_t'$ will "cover" the sphere in p dimensions systematically and thoroughly so that no interesting unidimensional views (or linear combinations) are neglected. This seems to be too much to expect or require, however, even for moderately large p, especially if the set of coefficients is prespecified and not based on indications from the data. For providing a more complete coverage of the sphere and also for including the case of assigning equal weights to the p variables, a suggestion due to Tukey is the choice

$$\mathbf{a}_t' = (\cos t, \cos\sqrt{2}\,t, \cos\sqrt{3}\,t, \cos\sqrt{5}\,t, \ldots), \qquad 0 \leqslant t \leqslant k\pi, \tag{84}$$

for an appropriate value of k. Normalization of $\mathbf{a}_t'$ to constant length would seem advisable for comparisons across different values of t. At $t = 0$, the weights for the variables are all equal. [*Note*: The lack of orthogonality among the elements of $\mathbf{a}_t'$ in Eq. 84 would imply that interpreting closeness among the curves directly in terms of closeness of the original p-dimensional observations would not be as easy as it would be in the case of Andrews's original suggestion for $\mathbf{a}_t'$, namely, Eq. 83.]

A different issue in using the plotting scheme as proposed initially by Andrews is its use in the situation where one has a very large number of multivariate observations. Since each observation is mapped into a curve, a routine Andrews plot with a very large number of curves would tend to look quite messy and not particularly revealing of anything but global aspects (e.g., clearly separated clusters or outliers) of the configuration of the data. For this case when n is large, an adaptation of the Andrews plot is, however, feasible and appears to be quite useful for studying the configurational and distributional aspects of multivariate data. The essential idea in the adaptation is to plot, for each point in a specified grid of values of t, only selected quantiles or percentage points (e.g., median, upper, and lower quartiles) of the distribution of the n values of f and, in addition, perhaps plot selected individual observations such as extreme values. The appearance and appreciation of such a *quantile contour plot*, or indeed of any of the versions mentioned above, can sometimes be improved by plotting an internally standardized set of values (such as deviations of f from its median divided by the interquartile range) rather than the values of f itself.

In addition to issues of choice of $\mathbf{a}_t$, that are shared by both function and quantile contour plots, the latter also involve the choice of quantiles for display purposes. As a general rule, plotting the median and quartiles (or, in the standardized version, centering the quantile contour plot at the median and scaling by the interquartile range) is useful. With regard to choosing specific quantiles beyond the quartiles, however, flexibility in the light of the specific application is in order. Thus, choosing the upper and lower 10%, 5%, and 1% quantiles is one possibility, while the upper and lower $12\frac{1}{2}$%, $6\frac{1}{4}$%, and $3\frac{1}{8}$% quantiles (i.e., equally spaced in probability) constitute another useful choice.

Since, in general, the function representing a specific multiresponse observation need not correspond to a particular quantile for all values of t [i.e., for instance, if $f_i(t_1)$ is the median value of the function at $t = t_1$ and $f_j(t_2)$ is the median value at $t = t_2$, it is not necessarily true that $i = j$], the quantile contours will not enable one to study the behavior of specific observations but will only aid in assimilating the general distributional aspects of the high-dimensional data. The functions corresponding to specific observations that are of particular interest can, of course, be displayed on a quantile plot, provided that the number of such observations is not so large as to interfere with appreciation of the plot as a whole.

Function plots and quantile contour plots are useful devices for detecting clusters and/or outliers. In view of the properties mentioned earlier, on an

Andrews function plot the curves corresponding to the multiresponse observations in a cluster would cohere together, and distinct clusters (or outliers) would be indicated by clear separations among the curves (or sets of them). On a quantile contour plot the existence of strong clusters may be revealed by a disproportionate squeezing together of particular quantiles at some values of t. For instance, clustering that is revealed by multimodality of the distribution of the projections along the vector corresponding to a specific value of t will tend to show up as such a squeezing together of certain of the quantiles at that value of t since the multimodality implies that for a small change in some quantile values (perhaps usually the "outer" quantiles) there will be a large change in the corresponding cumulative probability values.

Also, the quantile contour plot may be useful for studying the shape and more subtle configurational aspects of high-dimensional data distributions. Symmetry (as revealed by the appearances of the contours of pairs of upper and lower quantiles, especially the outer ones) is most easily appreciated. More specifically, if the data distribution is essentially spherical, one would expect to see approximately equal spacings between any specified pair of quantiles across the entire plot. The existence of high intercorrelations among the responses, which will tend to induce "ellipsoidal" types of configurations for the data, is likely to be revealed by an approximately proportionate squeezing together of almost all the quantiles for some values of t. [*Note*: An interesting use of function and/or quantile contour plots, as a consequence of this property, occurs in the context of multiple regression analysis. When one suspects difficulties caused by possible multicollinearity in the data, a plot of this type for the observations (either just on the independent variables or on both the dependent and independent variables) may be useful for identifying the singularities, as well as the essentially linearly independent combinations of the variables—the singularities would correspond to directions of (or linear combinations with) zero variance, which would be revealed by all the curves (or all quantiles) going through a single point at one or more values of t, while directions in (or linear combinations for) which curves (or quantiles) spread out considerably would be useful for picking the essentially linearly independent combinations.]

With quantile contour plots, as a check on the normality of the distribution of the data, one can compare the ratios of the observed spacings between specified pairs of quantiles with the values of such ratios for the normal distribution. Thus, for instance, the spacing between the 10% quantile and the median is approximately twice (1.9, more precisely) the spacing between the quartile and the median for a normal distribution; and if this relationship is not adequately satisfied for one or more values of t by the three quantiles involved, one will have reason to question the normality of the distributions of the linear combinations of the variables corresponding to these values of t, and hence also to question the joint normality of the distribution of the initial observations. Since the values of t spanned in a quantile contour plot do not generally yield *all possible* linear combinations of the original variables, indications of reasonable conformity to normality for every value of t in the grid chosen for a quantile contour plot, although not equivalent to a confirmation of joint normality, will nevertheless be useful evidence for deciding to use methods based on normality assumptions for further analyses of the data. Also, if normality is singularly inapplicable for only a few values of t, one may be able to transform the data initially so as to improve directional normality (see Section 5.3) along just these directions without altering the data in other directions, and then use standard methods with the transformed data. At any rate the quantile contour plot at least provides an informal basis for verifying normality. [*Note*: Since several comparisons of ratios of spacings between quantiles may be involved, one may wish to automate this process and have the computer not only do the plotting but also provide printout flagging situations in which the departures from normality are sufficiently striking.]

A considerably different problem, which can be motivated in terms of the function and quantile contour plots, is that of choosing a "typical" multiresponse observation. The sample mean vector, $\bar{\mathbf{y}}$, and the more robust estimators of location discussed in Section 5.2.3 are examples of statistics that summarize one typical aspect of multiresponse data, namely, overall location. Even in the context of location estimation, for some applications one may be interested in choosing an actual observation as a typical value instead of a summary statistic. One approach to this problem is to choose as a typical observation the one whose representation as a curve on a function plot is closest (in some specified metric) to the set of median values on the corresponding quantile contour plot. In fact, this idea, in addition to its use in developing a location estimate, may be worth investigating as a means of defining multivariate order statistics and quantiles.

Example 41. This example, taken from Andrews (1972), pertains to a discriminant analysis described by Ashton et al. (1957) of eight measurements ($p = 8$) on teeth of different "races" of men and apes, so as to aid in the classification of some fossils on the basis of their measurements with respect to the same eight characteristics used for the men and the apes. Nine groups— three "races" of men and six groups (three types × two sexes) of apes were involved, so that there were eight CRIMCOORDS (see Section 4.2) in the discriminant analysis.

Exhibit 41*a* (see page 245) shows the coordinates of the nine group centroids in the eight-dimensional space of the CRIMCOORDS and also the representations for the six fossils in this space. Exhibit 41*b* (see page 246) shows a graphical representation, obtained by Ashton et al. (1957), of the group centroids as well as the fossils in the space of the first two CRIMCOORDS; also included in this picture are approximate 90% confidence regions for the locations of the nine groups in the two-dimensional CRIMCOORDS space. (See the discussion in Section 4.2 pertaining to the methodological details of such graphical displays.) In Exhibit 41*b* the coordinates of the centers of the circles and of the points that correspond to the fossils are the values shown in

Exhibit 41*a*. Results of discriminant analysis of fossil data (Ashton et al., 1957; Andrews, 1972)

PERMANENT FIRST LOWER PREMOLAR

	means of groups in the space of CRIMCOORDS							
A. West African	−8.09	+0.49	+0.18	+0.75	−0.06	−0.04	+0.04	+0.03
B. British	−9.37	−0.68	−0.44	−0.37	+0.37	+0.02	−0.01	+0.05
C. Australian aboriginal	−8.87	+1.44	+0.36	−0.34	−0.29	−0.02	−0.01	−0.05
D. gorilla: male	+6.28	+2.89	+0.43	−0.03	+0.10	−0.14	+0.07	+0.08
E. female	+4.82	+1.52	+0.71	−0.06	+0.25	+0.15	−0.07	−0.10
F. orang-outang: male	+5.11	+1.61	−0.72	+0.04	−0.17	+0.13	+0.03	+0.05
G. female	+3.60	+0.28	−1.05	+0.01	−0.03	−0.11	−0.11	−0.08
H. chimpanzee: male	+3.46	−3.37	+0.33	−0.32	−0.19	−0.04	+0.09	+0.09
I. female	+3.05	−4.21	+0.17	+0.28	+0.04	+0.02	−0.06	−0.06
			fossils					
J. Pithecanthropus	−6.73	+3.63	+1.14	+2.11	−1.90	+0.24	+1.23	−0.55
K. pekinensis	−5.90	+3.95	+0.89	+1.58	−1.56	+1.10	+1.53	+0.58
L. Paranthropus robustus	−7.56	+6.34	+1.66	+0.10	−2.23	−1.01	+0.68	−0.23
M. Paranthropus crassidens	−7.79	+4.33	+1.42	+0.01	−1.80	−0.25	+0.04	−0.87
N. Meganthropus palaeojavanicus	−8.23	+5.03	+1.13	−0.02	−1.41	−0.13	−0.28	−0.13
O. Proconsul africanus	+1.86	−4.28	−2.14	−1.73	+2.06	+1.80	+2.61	+2.48

Exhibit 41*b*. Representations of the fossil groups and the unknowns in the space of the first two CRIMCOORDS (Ashton et al., 1957; Andrews, 1972)

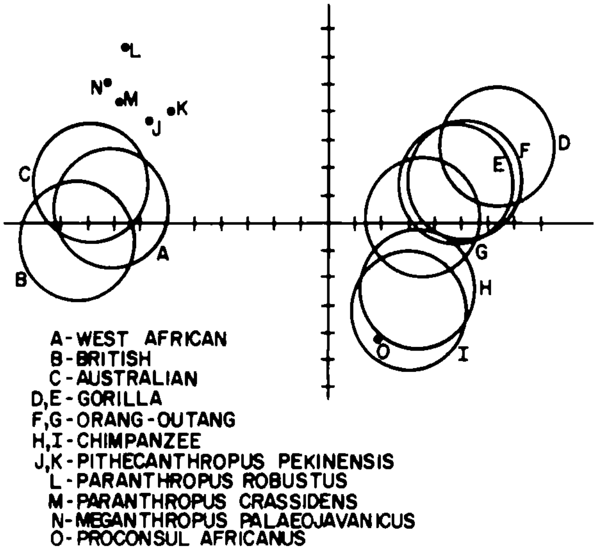


Exhibit 41*c*. High-dimensional plot of the centroids of fossil groups, using Eq. 83 for coefficients (Andrews, 1972)

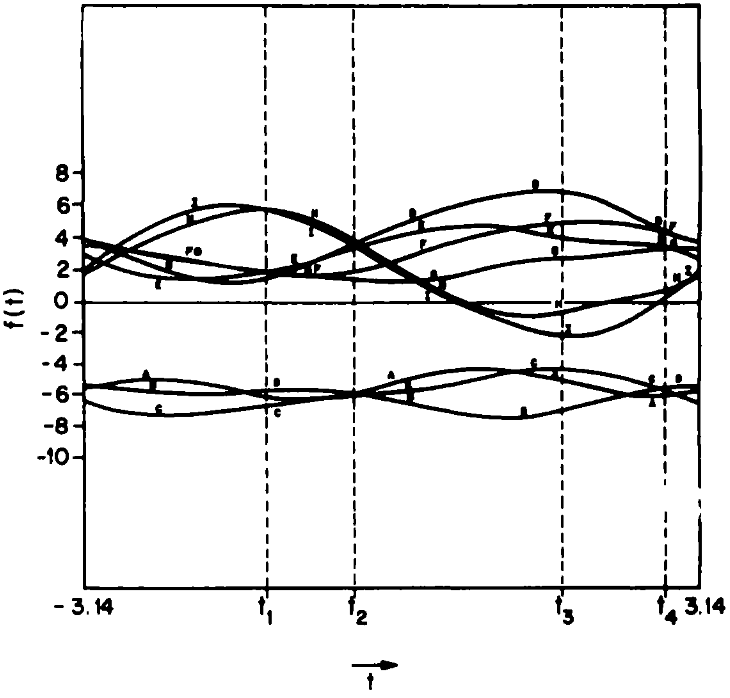


Exhibit 41d. High-dimensional plot of the centroids of the fossil groups and the unknowns, using Eq. 83 for coefficients (Andrews, 1972)

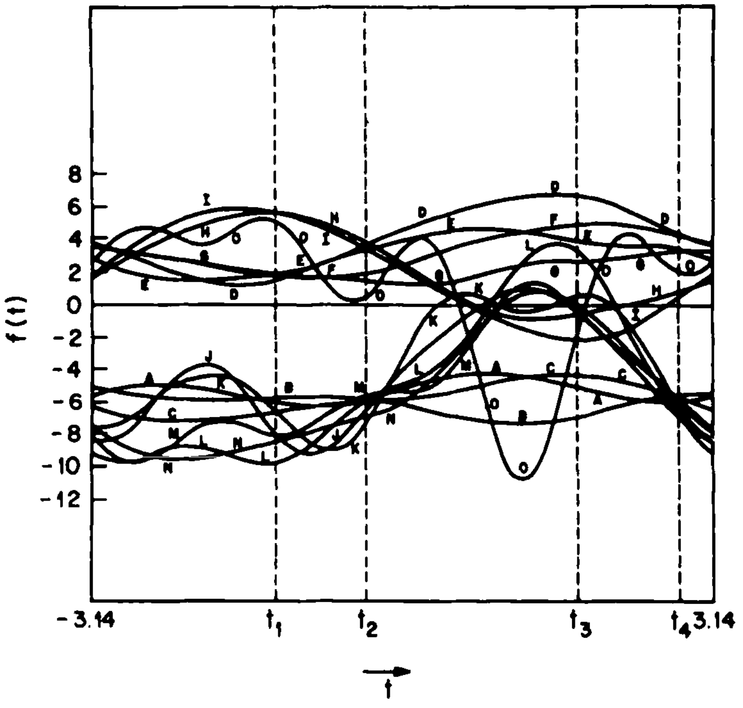
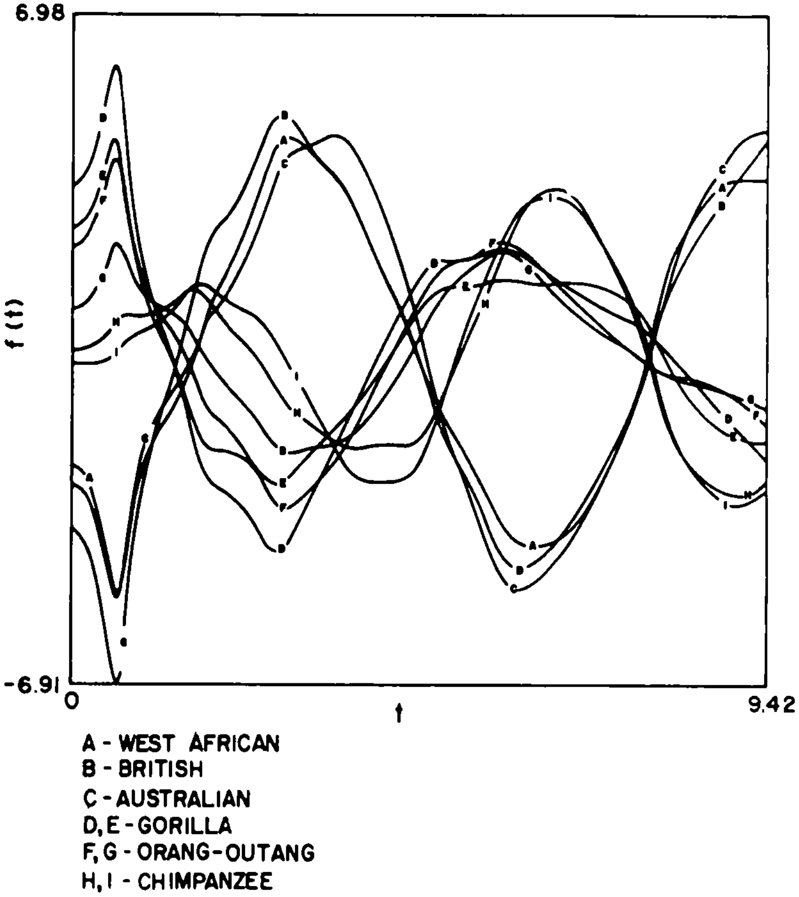


Exhibit 41*a* for just the first two CRIMCOORDS. Thus an inspection of the values in Exhibit 41*a* or of the visual portrayal in Exhibit 41*b* shows the clear separation (especially on the first CRIMCOORD) of the three "races" of men from the six groups of apes. Ashton et al. (1957) went further, on the basis of Exhibit 41*b*, to conclude that the fossil *Proconsul africanus* is very much like a chimpanzee, whereas the other five fossils are more akin to the "races" of men.

Andrews (1972), on the other hand, studied the eight-dimensional data in Exhibit 41*a* by means of his function plots and came to interesting but different conclusions regarding the classification of the fossils. Specifically, Exhibit 41*c* (see page 246) shows the function plot obtained by Andrews (1972) for the nine group centroids given in Exhibit 33*a*. The choice of $\mathbf{a}_t'$ in this plot is the one in Eq. 83 with $1/\sqrt{2}$ being associated with the first CRIMCOORD, sin *t* with the second, and so on. [*Note*: The CRIMCOORDS would have equal variances and be uncorrelated, ***under the usual assumptions*** of discriminant analysis, even if the original variables did not have these properties.] In Exhibit 41*c* the curves for the three human groups are quite well separated from the curves of the six groups of apes, and, among the apes, the chimpanzees stand out from the other two types. Also, the two sexes within each ape group tend to have relatively closely spaced curves, especially in the left part of the plot.

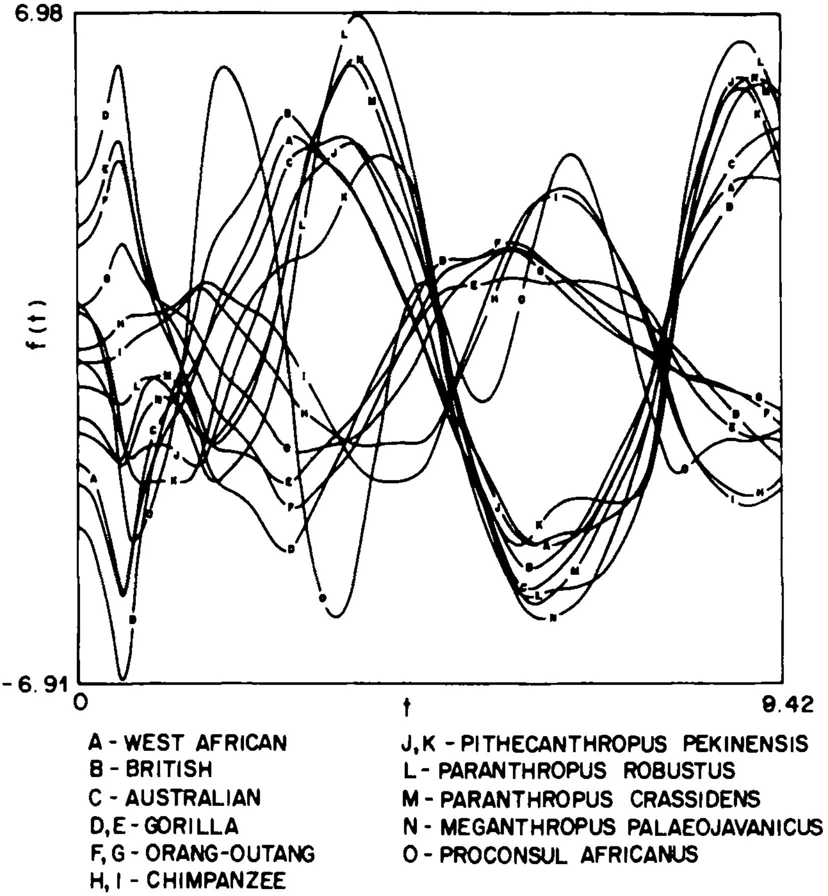
Exhibit 41*e*. High-dimensional plot of the centroids of fossil groups using Eq. 84 for coefficients



In addition, at specific values of t the separations among the groups are very pronounced in relation to the separations within the groups. For instance, at t_2 and t_4 the human groups are very cohesive and more clearly distinguished from the values for the curves for any of the apes. At t_1 and t_3 the chimpanzees seem to stand out more clearly from the remaining two groups of apes. Thus Exhibit 41*c* has been useful for detecting directions of clusterings among the groups.

In Exhibit 41*d* (see page 247) the curves for the six fossils in the study have been superimposed on the curves of Exhibit 41*c*. Immediately, one fossil, *Proconsul africanus*, stands out as being different, although for specific (but different) values of t it comes close to each of the groups. The remaining fossils are quite similar to man, especially at t_2 and t_4. Andrews (1972) discusses more

Exhibit 41*f*. High-dimensional plot of the centroids of the fossil groups and the unknowns, using Eq. 84 for coefficients



formal tests (based on certain confidence bands obtained by using the σ_f discussed earlier) to support the conclusions drawn from the function plot shown in Exhibit 41*d*. The feature that *Proconsul africanus* does not seem to belong to any of the groups illustrates the fact that in some applications of discriminant analysis it is wise to have the option of not classifying an unknown as necessarily belonging to any one of the prespecified groups. (See also Rao, 1960, 1962, regarding this issue.)

Exhibits 41*e* and *f* show function plots that correspond to Exhibits 41*c* and *d* when $\mathbf{a}_t$ is specified according to Eq. 84 rather than Eq. 83. Although the appearance of the plots in Exhibits 41*e* and *f* is more noisy and the separation of the human and ape groups is no longer as striking, the general indications and conclusions in this example are quite similar for the two choices of $\mathbf{a}_t$. In fact, in this example, a star plot and a Chernoff's faces display of the "data" in Exhibit 41*a* would also have led to the same conclusions.

Example 42. The data collected by Anderson (1935), and also used by Fisher (1936), shown in Exhibit 42*a* consist of 50 quadrivariate observations (viz., logarithms of sepal length and width and of petal length and width) for *Iris setosa*. The original data on *Iris setosa*, as well as on two other species of iris (*Iris versicolor* and *Iris virginica*), are well known in the multivariate literature and have been utilized by many authors as the basis for testing different classification and clustering algorithms (e.g., Friedman & Rubin, 1967). The data set is considered to be generally well behaved with no particular peculiarities, and it has been found that *Iris setosa* is easily distinguishable from the other two species (see, for example, Fisher, 1936; Friedman & Rubin, 1967; and Exhibit 42*c*).

Exhibit 42*a*. *Iris setosa* data (Anderson, 1935; Fisher, 1936)

Sepal Length (ln cm)	Sepal Width (ln cm)	Petal Length (ln cm)	Petal Width (ln cm)
1.629	1.253	0.336	−1.609
1.589	1.099	0.336	−1.609
1.548	1.163	0.262	−1.609
1.526	1.131	0.405	−1.609
1.609	1.281	0.336	−1.609
1.686	1.361	0.531	−0.916
1.526	1.224	0.336	−1.204
1.609	1.224	0.405	−1.609
1.482	1.065	0.336	−1.609
1.589	1.131	0.405	−2.303
1.686	1.308	0.405	−1.609
1.569	1.224	0.470	−1.609
1.569	1.099	0.336	−2.303
1.459	1.099	0.095	−2.303
1.758	1.386	0.182	−1.609
1.740	1.482	0.405	−0.916
1.686	1.361	0.262	−0.916
1.629	1.253	0.336	−1.204
1.740	1.335	0.531	−1.204
1.629	1.335	0.405	−1.204
1.686	1.224	0.531	−1.609
1.629	1.308	0.405	−0.916
1.526	1.281	0	−1.609
1.629	1.194	0.531	−0.693

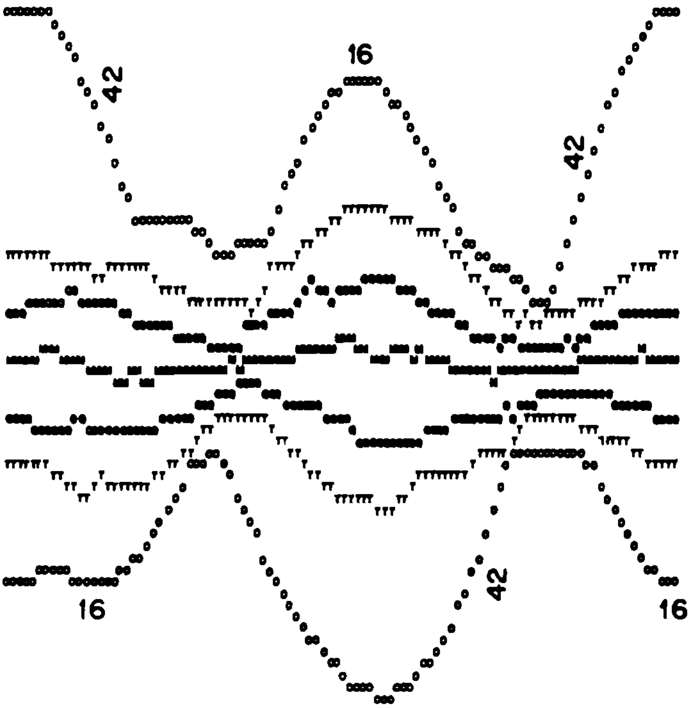
Exhibit 42*a*. (*Continued*)

Sepal Length (ln cm)	Sepal Width (ln cm)	Petal Length (ln cm)	Petal Width (ln cm)
1.569	1.224	0.642	−1.609
1.609	1.099	0.470	−1.609
1.609	1.224	0.470	−0.916
1.649	1.253	0.405	−1.609
1.649	1.224	0.336	−1.609
1.548	1.163	0.470	−1.609
1.569	1.131	0.470	−1.609
1.686	1.224	0.405	−0.916
1.649	1.411	0.405	−2.303
1.705	1.435	0.336	−1.609
1.589	1.131	0.405	−1.609
1.609	1.163	0.182	−1.609
1.705	1.253	0.262	−1.609
1.589	1.281	0.336	−2.303
1.482	1.099	0.262	−1.609
1.629	1.224	0.405	−1.609
1.609	1.253	0.262	−1.204
1.504	0.833	0.262	−1.204
1.482	1.163	0.262	−1.609
1.609	1.253	0.470	−0.511
1.629	1.335	0.642	−0.916
1.569	1.099	0.336	−1.204
1.629	1.335	0.470	−1.609
1.526	1.163	0.336	−1.609
1.668	1.308	0.405	−1.609
1.609	1.194	0.336	−1.609

For present purposes the data of Exhibit 42*a* were initially "standardized" on each variable by subtracting the median from each observation and then dividing by the interquartile range. With as many as the 50 observations in this example, a quantile contour plot rather than a function plot is the appropriate choice, and Exhibit 42*b* (see page 252) shows such a plot. The choice for $\mathbf{a}'_t$ in this case was $\{\sin t, \cos t, \sin 2t, \cos 2t\}$, and the quantiles chosen for display were the median, the lower and upper quartiles, and the lower and upper tenths. In Exhibit 42*b* the median is labeled M, the two quartiles are denoted as Q, and the tenths are shown as T's. Exhibit 42*b* is actually a printer plot and is an example of graphical output that does not require any exotic, expensive, or specialized hardware.

Also shown on Exhibit 42*b* are the two "outermost" points, plotted as 0's. An investigation of these points showed that one of them corresponded to the

Exhibit 42*b*. Quantile contour plot of *Iris setosa* data

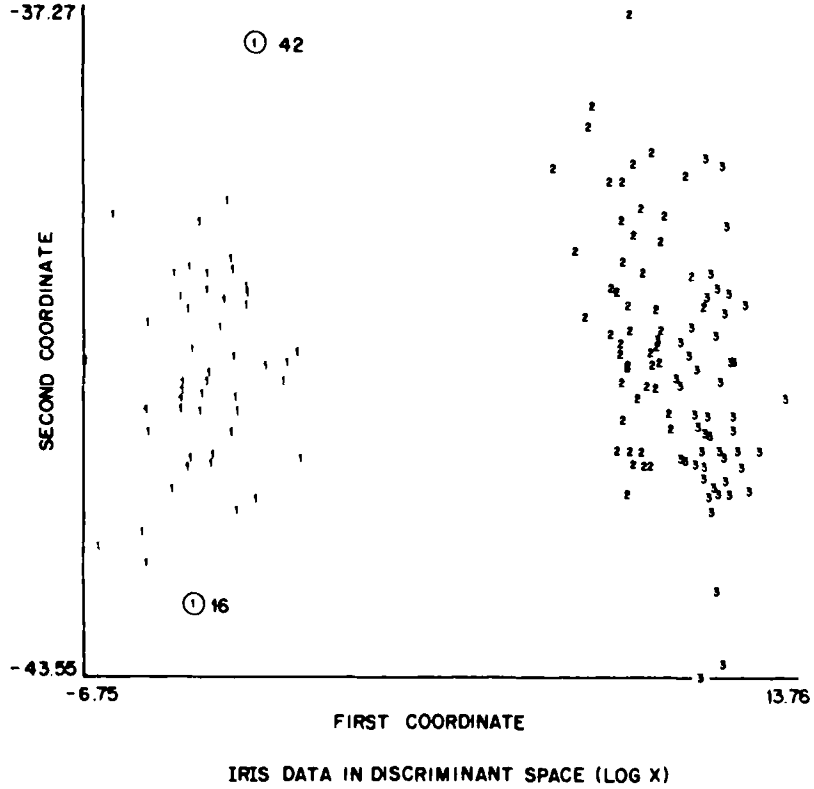


16th observation in the original set, and the other to the 42nd observation. The process of identifying such consistently outlying observations, if they exist, can be automated by requiring the computer program to superimpose, with appropriate labels, the curves for all observations that are consistently (by some quantitive definition such as "for more than half the values of *t*") well separated from the majority. At any rate Exhibit 42*b* shows directly and simply that the 16th and 42nd observations are symmetrically and oppositely situated observations which seem to be quite clearly separated from the remaining observations as one views the data in several unidimensional directions.

Exhibit 42*c* shows a representation of the three groups of irises in the two-dimensional discriminant space for this problem. The fifty *Iris setosa* points are labeled 1 in this figure; those for *Iris versicolor*, 2; and those for *Iris virginica*, 3. In this picture the 16th and 42nd observations in the *Iris setosa* group are seen to lie at opposite ends of the data configuration with respect to the second discriminant coordinate, although the separation of the two points is by no means as striking as it is in the quantile contour plot.

Aside from the indications regarding the 16th and 42nd observations, by scanning the spacings among the quantiles across the whole picture in Exhibit 42*b* one can get a "feeling" for the shape of the quadrivariate distribution of

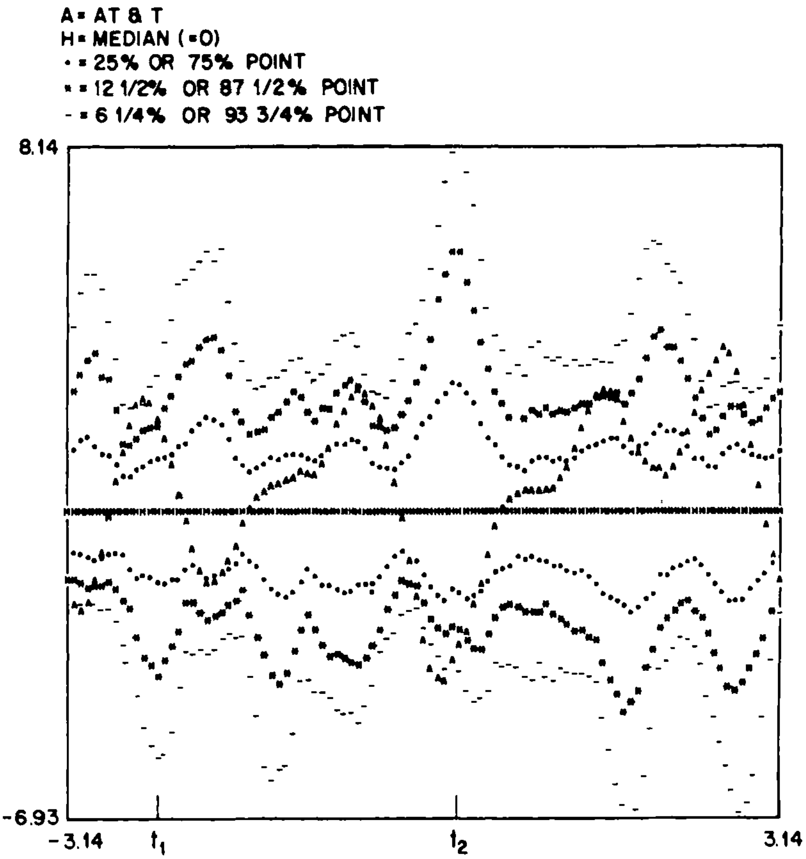
Exhibit 42c. Representation of the three groups of irises in the space of the two CRIMCOORDS



the data in Exhibit 42*d*. The configurations of the *M*'s, *Q*'s, and *T*'s in Exhibit 42*b* indicate that the data in this example are quite symmetrically, although not spherically (perhaps because of the known intercorrelations among the variables here), distributed in 4-space.

Example 43. This example derives from a study of empirical groupings of corporations (see Chen et al., 1970, 1974, and also Examples 17, 18, and 23 in Chapter 4) on the basis of yearly observations on several variables chosen to represent the financial and economic characteristics of the firms. One question of interest in this study was what an appropriate classification would be for AT&T (American Telephone & Telegraph Company) vis-à-vis the dichotomy of corporations as either industrials or utilities. Standard and Poor's COMPUSTAT tape was used for deriving annual values for 13 variables (see Chen et al. 1970, 1974, for a list and definition of the variables), and the investigation was carried out by performing separate analyses of each of the years 1960–1969 in particular.

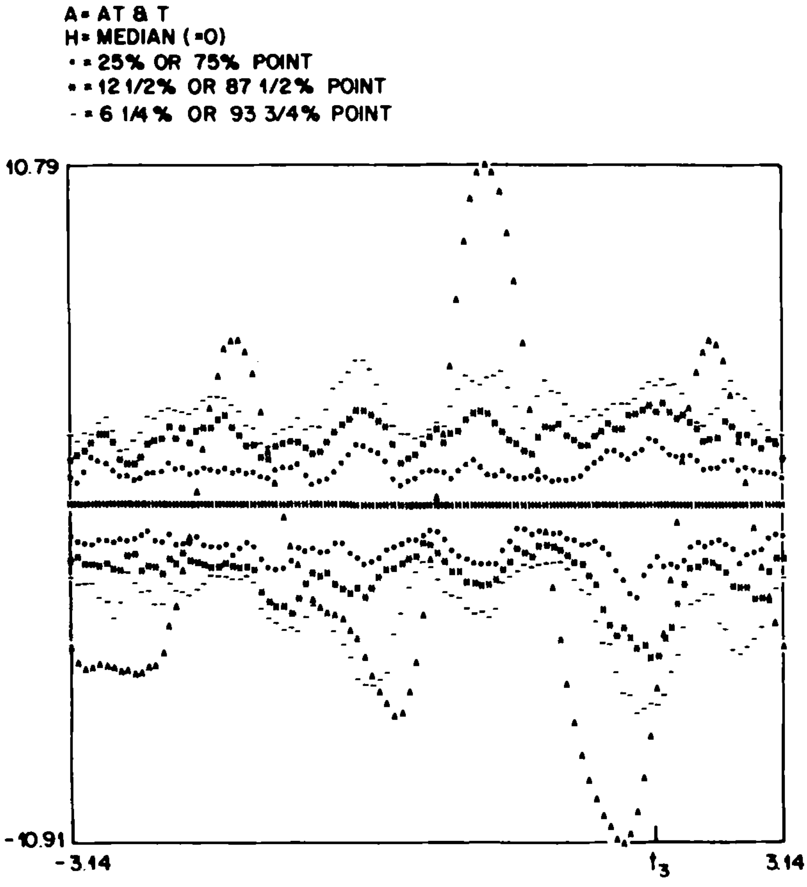
Exhibit 43*a*. Quantile contour plot of 495 industrials and AT&T for 1969



Quantile contour plots can be employed to summarize the findings regarding AT&T's classification. Exhibit 43*a* shows a quantile contour plot of the 13-dimensional data for 495 industrial firms for 1969; also included on the plot is the curve for AT&T, labeled *A*. [*Note*: As a preliminary standardization of each of the 13 variables, the median was substracted and the resulting deviations were divided by the inter-quartile range; the 13-dimensional observation for AT&T was subjected to the same standardization as was performed for the industrial firms displayed in Exhibit 43*a*.] The choice of $\mathbf{a}_t'$ for Exhibit 43*a* was the one in Eq. 83, and the displayed quantiles (whose associated probabilities are defined in the legend for the figure) are in fact deviations from the median, which therefore appears as a steady level line (labeled *H* for "half") across the middle of the picture.

Exhibit 43*b* shows a similar quantile contour plot for the 94 utilities involved in the study for the same year (1969), and again AT&T's curve is

Exhibit 43*b*. Quantile contour plot of 94 utilities and AT&T for 1969



shown as a series of *A*'s across the plot. [*Note*: The preliminary standardization for Exhibit 43*b* was, of course, based on the 94 utilities in this case.] A comparison of Exhibits 43*a* and *b* gives a clear visual impression that AT&T fits in better with the industrials than with the utilities. A more quantitative summary of this point is that for the industrials AT&T falls within the lower and upper quartiles for about 50% of the *t* values, within the lower and upper $12\frac{1}{2}\%$ points about 75% of the time, and within the band, defined by the lower and upper $6\frac{1}{4}\%$ points at a frequency of about 88%. [*Note*: These frequencies are exactly those to be expected for a typical industrial firm.] The corresponding frequencies for the utilities shown in Exhibit 43*b* are much smaller, being, respectively, 10%, 20%, and 40%.

Also, the configurations of the quantiles exhibit strong asymmetries of the data distributions along several directions (e.g., those that correspond to values of $t = t_1$ and t_2 in Exhibit 43*a* and $t = t_3$ in Exhibit 43*b*), and thus as an adjunct indication one has clear evidence of departures from joint normality of the distribution of the initial 13-dimensional data. The evidence for nonnormality of the distribution of these data is indeed plentiful, and Exhibits 43*a* and *b* are by no means the only indicators of this facet of the data. (See also Example 37 in Section 5.4.2.)

The above examples have demonstrated the utility of function plots and quantile contour plots. Despite their limitations the techniques have significant appeal because, at least in part, of the simplicity involved in their two-dimensional character, although the data being represented in them may be, and often are, quite high-dimensional. More work leading to other choices for $\mathbf{a}_t'$ (to provide, for instance, better coverage of "interesting" directions in p-space) and to methods for choosing $\mathbf{a}_t'$ in the light of the data would indeed be worthwhile.

6.3. COMPARISON OF SEVERAL MULTIRESPONSE SAMPLES

Many situations require the presentation of data from several identified groups for comparative purposes. The analysis of variance is a widely used technique for comparing two or more samples. In the uniresponse case Student's t and F statistics are familiar examples of summary statistics that are used for formal comparisons among two or more samples. It will often be useful in analyzing uniresponse data to supplement the computation of such summary statistics by probability plotting techniques, such as a Q-Q plot of one sample versus another, or perhaps superimposed normal probability plots of several samples on a single picture.

In the multiresponse situation Mahalanobis' D^2 or Hotelling's T^2 can be computed and utilized for formally comparing the locations of two groups. Also, for the two-group location problem, when the dimensionality (i.e., number of responses) exceeds the degrees of freedom available for estimation of the error covariance matrix, Dempster (1958) has proposed a test. Multivariate analysis of variance (MANOVA) is concerned with certain generalizations of the two-group procedures proposed by Mahalanobis and by Hotelling (see Sections 5.2.1 and 5.2.2). However, the formal analyses involved in MANOVA are often not sufficiently revealing. They need to be augmented by various graphical analyses, and the discussion of such graphical tools is the concern of this section.

It is perhaps typical of analysis of variance situations that one wishes to ask not one or two questions of the same body of data but several. One would also like to have a climate for the statistical analysis in such a situation that would allow unanticipated characteristics to be spotted. Examples of nonobvious but interesting indications are the presence of possibly real treatment effects, the existence of outliers in the data, and heteroscedasticity.

For these reasons it is reasonable to provide statistical procedures that use some sort of statistical model to aid in comparisons of various collections of comparable quantities, and yet enable one to make such comparisons without the need to commit oneself to any narrow specification of objectives. Examples of collections of comparable quantities are a collection of single-degree-of-freedom contrasts, a collection of mean squares in ANOVA, and a collection of sum-of-products matrices in MANOVA. Procedures for such comparisons have been called *internal comparisons methods* by Wilk & Gnanadesikan (1961, 1964). Specifically, some kinds of probability plotting techniques have been developed for internal comparisons of the relative magnitudes that are involved in ANOVA and MANOVA. These procedures provide a statistical measure for facilitating the assessment of relative magnitudes, which probably becomes rather nonintuitive when one is dealing with a large collection of comparable quantities. Moreover, the procedures can provide some insight into various possible inadequacies of the statistical model used to generate the analysis. The procedures are not excessively influenced by some data-independent aspects, such as the need to prechoose an error term.

In particular, Table 2 shows a categorization of orthogonal analysis of variance situations according to the multiplicity of response and the degrees-of-freedom decomposition of the experimental design or model for the data. Also given beneath the two-way categorization is a list of specific references that describe techniques of relevance to each of the cells in the table.

Table 2. Categorization of ANOVA and MANOVA Cases

	Response Structure	
DF Decomposition	Uniresponse	Multiresponse
All 1 df	I	IV
All ν df	II	V
Mixed df	III	VI

I. A half-normal plot of absolute values of contrasts—C. Daniel, *Technometrics* 1 (1959), 311-41.

II. A ν-df chi-squared plot of sums of squares (or gamma plot of sums of squares with shape parameter $\eta = \nu/2$)—M. B. Wilk, R. Gnanadesikan, and M. J. Huyett, *Technometrics* 4 (1962), 1-20.

III. A generalized probability plot of mean squares—R. Gnanadesikan and M. B. Wilk, *J. R. Stat. Soc.* B32 (1970), 88-101.

IV. A gamma plot of squared distances with an estimated shape parameter—M. B. Wilk and R. Gnanadesikan, *Ann. Math. Stat.* 35 (1964), 613-31; also, Chapter VII of Roy et al. (1971).

V. Gamma plots of certain functions of eigenvalues with an estimated shape parameter—R. Gnanadesikan and E. T. Lee, *Biometrika* 57 (1970), 229-37.

Gnanadesikan (1980) provides complete step-by-step descriptions of the techniques for cells I–V. The two subsections that follow will be concerned with describing the methods for cells IV and V, respectively. Methods for cell VI are not yet available.

6.3.1. Graphical Internal Comparisons among Single-Degree-of-Freedom Contrast Vectors

The techniques to be described in this subsection may be employed in any situation in which there is a meaningful decomposition of effects (in the sense of the analysis of variance) into orthogonal single-degree-of-freedom components. For instance, in multifactor experiments in which the factors are quantitative and are used at several levels for obtaining the treatment combinations, one has the familiar decomposition into linear, quadratic, etc., components for the treatment effects, and these form a natural set of orthogonal single-degree-of-freedom components that one may wish to intercompare. Two-level factorial (full and/or fractional) experiments, of course, yield a meaningful decomposition into main effects and interactions which together constitute an orthogonal single-degree-of-freedom set of effects whose interpretations are of prime interest in such experiments. For simplicity the prototype experimental situation for developing the methodology here will be taken to be that of a 2^N factorial experiment, that is, N factors each of which has two levels, but it should be kept in mind that the methods have wider applicability, as indicated by the preceding discussion. More specifically, the setup will be one in which there are $n = 2^N$ treatment combinations in all, and corresponding to the ith treatment combination one has a p-dimensional observation, $\mathbf{y}_i'$ ($i = 1, \ldots, n$), whose coordinates are the observed values of the p responses for the particular treatment combination.

In fact, for motivating some of the basic concepts underlying the methodology, consider a 2^3 experiment ($N = 3$, $n = 8$) with two responses ($p = 2$) measured on each experimental unit after "application" of one of the eight treatment combinations involved. Hence one has eight bivariate observations which can be represented as points in a two-dimensional space, as shown, for example, in Figure 13. The eight points in the plot are labeled by the respective treatment combinations associated with them, and the notation for the treatment combinations is the standard one for two-level factorial experiments. To illustrate how one may think about a treatment effect in this bivariate case, consider the problem of defining the main effect of factor A. The set of eight points can be divided into two groups of four observations each; in one group (shown in Figure 13 as unshaded circles) all the treatment combinations are ones in which the factor A is at its lower level, and in the other group (shown in Figure 13 as unshaded squares) the factor A is at its higher level. One can define a centroid of each of the sets of four observations, and these are shown in Figure 13 as a filled-in circle and a filled-in square. The univariate estimate of the main effect of A with respect to the first response, for instance, is just the

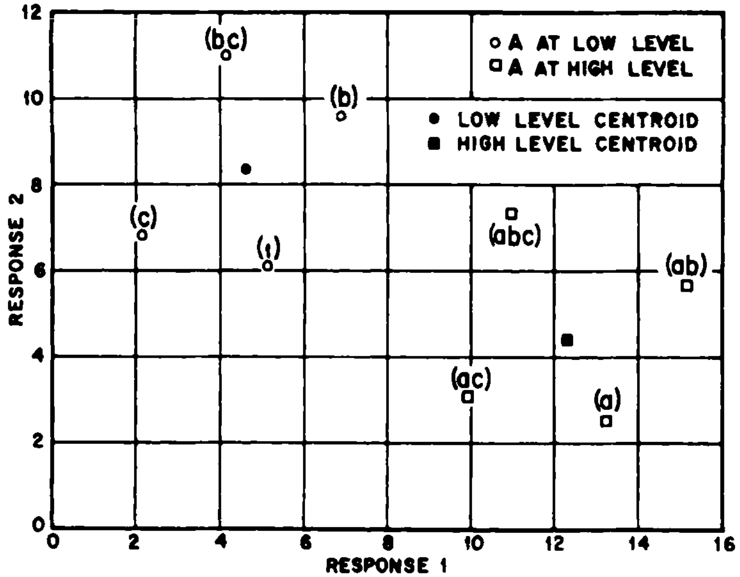


Fig. 13. Pictorial representation for two-dimensional main effect of A in a 2^3 experiment involving bivariate observations.

distance between the projections of the two centroids on the horizontal axis. Similarly, the main effect of A with respect to the second response is the distance between the projections of these two centroids on the vertical axis. Next, a natural and not unreasonable conceptualization of the main effect of A in the two-dimensional situation would be as a distance between the two centroids in the two-dimensional space. Hence, in particular, one can think of a vector going from the filled-in circle to the filled-in square and consider that the "larger" (in some sense) this vector is, the greater is the two-dimensional main effect of factor A.

Clearly one can partition the eight observations in the example of the 2^3 experiment in other specific ways to get various pairs of groups of four observations each, and by analogous reasoning to that used above define the main effects of, as well as the various interactions among, all the factors. One will then have vectors going between the centroids for these different partitions corresponding to the seven effects involved, and these vectors, labeled according to the effects to which they correspond, can be represented in a two-dimensional space. There will be seven such vectors emanating from the origin, corresponding to the three main effects, the three two-factor interactions, and the one three-factor interaction in this case of the 2^3 experiment (see Figure 14). The transformation involved in obtaining these vector effects from the initial observations is essentially an orthogonal transformation—in fact, the same one that is well known in the analysis of univariate two-level factorial experiments. Explicitly, if $y'_1, y'_2, \ldots, y'_8$ denote the bivariate observations in a

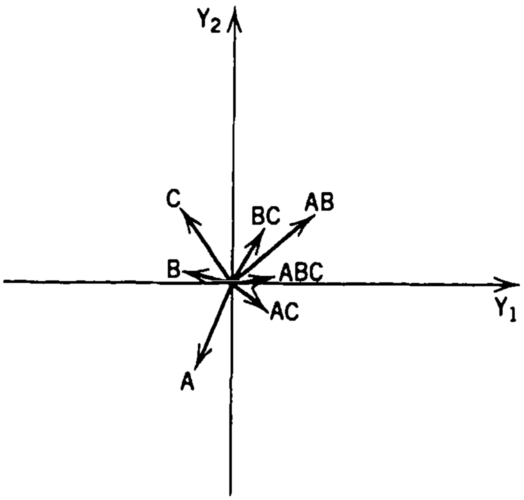


Fig. 14. Representation of contrast vectors in a 2^3 experiment involving bivariate observations.

2^3 experiment, where the treatment combinations are taken in so-called standard order [i.e., (1), (*a*), (*b*), (*ab*), (*c*), (*ac*), (*bc*), (*abc*)], the seven bivariate treatment effect vectors, $\mathbf{x}'_1, \ldots, \mathbf{x}'_7$ (such as the ones in Figure 14), are defined by the transformation

$$\begin{pmatrix} \mathbf{m}' \\ \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_7 \end{pmatrix} = \mathbf{R} \begin{pmatrix} \mathbf{y}'_1 \\ \vdots \\ \mathbf{y}'_8 \end{pmatrix},$$

where

$$\mathbf{R} = \begin{bmatrix} + & + & + & + & + & + & + & + \\ - & + & - & + & - & + & - & + \\ - & - & + & + & - & - & + & + \\ + & - & - & + & + & - & - & + \\ - & - & - & - & + & + & + & + \\ + & - & + & - & - & + & - & + \\ + & + & - & - & - & - & + & + \\ - & + & + & - & + & - & - & + \end{bmatrix} \tag{85}$$

with + and − standing, respectively, for +1 and −1. Thus $\mathbf{m}'$, associated with an overall effect, is proportional to (i.e., 8 times) the mean vector for the data, while $\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_7$ are the vectors associated with the effects A, B, AB, C, AC, BC, and ABC, respectively. Since each of the rows of $\mathbf{R}$ that leads to one of the treatment effect vectors defines a *contrast* (i.e., the number of +1's is equal to the number of −1's, so that their sum is 0), the treatment effect vectors, $\mathbf{x}'_1, \ldots, \mathbf{x}'_7$, may also be called *single-degree-of-freedom contrast vectors*. [*Note*: The usual definitions of the effects generally multiply the first row of $\mathbf{R}$ by $\frac{1}{8}$ so as to yield the mean vector itself, and also the remaining rows by $\frac{1}{4}$ so as to yield differences in the means of four observations as described in the discussion of Figure 13. Also, to make $\mathbf{R}$ an orthogonal matrix all that is required is to multiply it by the scalar $1/\sqrt{8}$.] In practice, the transformation involved in obtaining the $\mathbf{x}'_i$ from the initially observed $\mathbf{y}'_i$ is generally carried out by Yates's algorithm, which is not only simple but also computationally sound in the sense of numerical accuracy and stability.

Comparisons of the relative magnitudes of treatment effects are one important goal of the analysis of variance; and, returning to Figure 14, for this purpose one needs some way of measuring the "sizes" of the treatment effect vectors (or single-degree-of-freedom contrast vectors), $\mathbf{x}'_1, \ldots, \mathbf{x}'_7$, displayed there. The problem here can be treated as being just the same as the one of choosing a distance measure for classification procedures (see Section 4.2.1), since the issues in choosing a measure of "size" for the vectors displayed in Figure 14, so that a "large" vector will correspond to a large effect, also arise in choosing a metric for measuring distances between centroids of various partitions of the observations in the two-dimensional space of the observations shown in Figure 13. At any rate, the methodology developed here depends on using a squared distance measure, $\mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is some positive semidefinite matrix, for measuring the size of the treatment effect, $\mathbf{x}$.

More generally, with $n = 2^N$ treatment combinations and p responses observed on each experimental unit, if $\mathbf{Y}'$ denotes the $n \times p$ matrix whose rows are the p-dimensional observations, let

$$\mathbf{Y}' = \begin{pmatrix} \mathbf{y}'_1 \\ \vdots \\ \mathbf{y}'_n \end{pmatrix} = [\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_p],$$

so that $\mathbf{Y}_j$ consists of the n observations on the jth response ($j = 1, \ldots, p$). Then the univariate analysis of variance for the jth response will yield

$$\begin{pmatrix} m_j \\ \mathbf{X}_j \end{pmatrix} = \mathbf{R}\mathbf{Y}_j, \qquad j = 1, \ldots, p, \tag{86}$$

where $\mathbf{R}$ is an $n \times n$ matrix that can be built up by analogy with the one in Eq. 85, m_j corresponds to an overall (or mean) effect, and the $(n - 1)$ elements of

$\mathbf{X}_j$ correspond to the measures of the main effects and interactions for the jth response. Thus one obtains

$$\mathbf{RY}' = \begin{pmatrix} m_1 & m_2 & \cdots & m_p \\ \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_p \end{pmatrix} = \begin{pmatrix} \mathbf{m}' \\ \mathbf{X}' \end{pmatrix}, \tag{87}$$

where the $(n - 1) \times p$ matrix, $\mathbf{X}'$, has as its columns $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$, and as its rows the $(n - 1)$ *single-degree-of-freedom contrast vectors* $\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_{n-1}$, that is,

$$\mathbf{X}' = \begin{pmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_{n-1} \end{pmatrix} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p]. \tag{88}$$

Equations 86, 87, and 88 suggest that one way of obtaining the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$ is to perform univariate analyses of variance (perhaps via Yates's algorithm) of each response separately and then collect the p individual measures for each effect (main or interaction) together as a p-dimensional vector.

For assessing the relative magnitudes of the contrast vectors, the values

$$d_i = \mathbf{x}'_i \mathbf{A} \mathbf{x}_i, \qquad i = 1, \ldots, n - 1, \tag{89}$$

for some choice of the positive semidefinite compounding matrix $\mathbf{A}$ (more will be said later regarding the choice of $\mathbf{A}$) are obtained. Exactly as in the simple 2^3 example discussed earlier, the d_i's, which are measures of the sizes of the $\mathbf{x}_i$'s, can be interpreted as squared distances between the centroids of certain partitions of the original observations for defining the different treatment effects.

To assess the contrast vectors by means of these squared distances, one needs an "evaluating distribution" or a "null distribution" of such squared distances. In other words, one needs a distribution that is reasonable under the usual kinds of null assumptions, such as multivariate normality (which may be a more reasonable assumption for the contrast vectors than for the original observations because of the "averaging" involved in obtaining the contrast vectors), homoscedasticity, and the absence of any real treatment effects.

More explicitly, under the usual linear model assumptions (see Section 5.2.1), the observations $\mathbf{y}'_i$ are p-variate normally distributed, with location parameters (or expected values) that reflect their factorial experimental structure and a common unknown covariance matrix, $\mathbf{\Sigma}$. Under these assumptions, taken in conjunction with the further null assumption that there are no real treatment effects, the contrast vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$ will be mutually independently distributed as $N[\mathbf{0}, \Sigma]$. [*Note*: In order for the contrast vectors to have exactly the same covariance structure as the initial observations, the transformation matrix **R** in Eq. 86 must be specified to be orthogonal, that is, with the multiplicative constant $1/\sqrt{n}$.] These null assumptions are to be used only as a basis for generating internal comparisons techniques, and an appealing characteristic of these techniques is that, in any specific application, they provide some indications of the appropriateness and adequacy of the assumptions themselves.

The question of an evaluating distribution thus can be formulated as follows: given that $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$ are a random sample from $N[\mathbf{0}, \Sigma]$, what is the distribution of $d_1, \ldots, d_{n-1}$, where $d_i = \mathbf{x}_i' \mathbf{A} \mathbf{x}_i$? In developing an answer to this question, one needs to consider the role of the compounding matrix **A**, which itself may be, as mentioned earlier, and in practice often is, computed from the observations, $\mathbf{y}_i'$. For instance, one may decide to use $\mathbf{A} = \mathbf{I}$ (which amounts to measuring squared Euclidean distances between centroids of partitions of the data), or in order to reflect differences in the variances of the responses one may wish to use a diagonal matrix of reciprocals of either prespecified variances or estimates of these from the current data. More generally, one may wish to scale the contrast vectors to allow both for different variances of the responses and for intercorrelations among them, and then the choice for **A** will be the inverse of a prespecified or estimated covariance matrix of the responses. In general, the use of several choices of **A** in analyzing a single set of data may be productive since the different choices may lead to different findings about the data. Whatever the choice is for **A**, however, since it is common to all the squared distances that are to be internally compared, it is treated in the approach taken here as being a fixed (i.e., nonrandom) quantity.

With this in mind, under the null assumptions outlined above, each of the squared distances d_i (for a selected compounding matrix **A**) is distributed as the linear combination $c_1\chi_1^2 + c_2\chi_2^2 + \cdots + c_r\chi_r^2$, where $c_1, \ldots, c_r$ are the positive eigenvalues of $\mathbf{A}\Sigma$, r is the rank of **A**, and the χ^2's are mutually independent central chi-squared variates, each with 1 degree of freedom. This well-known distributional result is not very useful as it stands; rather, its value lies in suggesting an equally well-known approximate result (see Satterthwaite, 1941; Patnaik, 1949; Box, 1954). The approximate result in question is that the distribution is represented reasonably adequately by a gamma distribution. Thus, specifically, under the null assumptions one can consider $d_1, d_2, \ldots, d_{n-1}$ approximately as a random sample from the gamma distribution with scale parameter λ and shape parameter η, that is, with density

$$f(d; \lambda, \eta) \begin{cases} = \dfrac{\lambda^\eta}{\Gamma(\eta)} d^{\eta-1} \exp(-\lambda d) & \text{for } d \geqslant 0, \\ = 0 & \text{for } d < 0, \end{cases} \tag{90}$$

where both λ and $\eta > 0$.

To be able to use this evaluating distribution, one needs estimates of λ and η since these are in general unknown. In particular, if a "proper" estimate, $\hat{\eta}$, of η is available, one can obtain a gamma probability plot (i.e., a *Q-Q* plot whose abscissa corresponds to a gamma distribution; see the discussion of *Q-Q* plotting in Section 6.2) of the ordered squared distances. The ordinate on such a plot will correspond to values of the ordered squared distances, and the abscissa will represent the corresponding quantiles of a standard gamma distribution with a shape parameter equal to this "properly estimated" value, $\hat{\eta}$. Under null conditions the resulting configuration will be linear, oriented toward the origin with a slope that is an estimate of $1/\lambda$. [*Note*: For gamma probability plotting one does not need a knowledge of the scale parameter since it affects, not the linearity, but only the slope of the configuration; however, the estimation of λ and of η will be carried out simultaneously, although only the estimate of η is needed.]

If the null conditions are not in accord with the data—say, for instance, that there are some real treatment effects—the largest squared distances will be too "large" and will exhibit themselves as departures from a linear "error" configuration defined by the smaller squared distances. Departures from other null assumptions (e.g., homoscedasticity, normality) may also be expected to show up as systematic departures from the "null" linear configuration of the "null" d_i's (i.e., those that conform adequately to the null assumptions).

The next question is what a "proper" estimate of η (and λ) might be. It is desirable that the estimate be based on a null subset of the squared distances (i.e., only those d_i's that satisfy the null assumptions), so that the d_i's which do not conform to such assumptions will stand out against a background defined by d_i's that do. In particular, when a d_i reflects a real treatment effect, its distribution will also be a linear combination of independent χ^2's, but now involving a noncentral χ^2. It is, however, known (Patnaik, 1949) that such a combination involving a noncentral χ^2 can also be approximated by a suitably chosen gamma distribution. Hence, to minimize the influence of possibly real treatment effects on the estimation of the parameters required for the evaluting gamma distribution, it is wise to base the estimation on an order statistics formulation. In other words, one orders the squared distances to obtain $0 \leqslant d_{(1)} \leqslant d_{(2)} \leqslant \cdots \leqslant d_{(M)} \leqslant \cdots \leqslant d_{(K)} \leqslant \cdots \leqslant d_{(n-1)}$. Then, on the basis of judgment, one chooses a number, $K[\leqslant(n-1)]$, as the number of squared distances that are likely to conform to the null assumptions. As additional insurance one bases the actual estimation on the M smallest squared distances considered as the M smallest observations in a random sample of size K. The actual method of estimation to be used with this formulation will be maximum likelihood. The maximum likelihood estimates, $\hat{\lambda}$ and $\hat{\eta}$, obtained from $d_{(1)}, \ldots, d_{(M)}$ considered as the M smallest order statistics in a random sample of size K [where $M \leqslant K \leqslant (n-1)$] from the gamma distribution with density as specified in Eq. 90 are functions only of $d_{(M)}$ and the ratios of the geometric and arithmetic means of $d_{(1)}, \ldots, d_{(M)}$ to $d_{(M)}$, that is,

$$P = \frac{\prod_{i=1}^{M} [d_{(i)}]^{1/M}}{d_{(M)}} \quad \text{and} \quad S = \frac{\sum_{i=1}^{M} d_{(i)}}{M d_{(M)}}.$$

It is necessarily true that $0 \leqslant P \leqslant S \leqslant 1$. Wilk et al. (1962b) provide tables that enable one to obtain the maximum likelihood estimates, $\hat{\lambda}$ and $\hat{\eta}$. They also describe numerical methods for computing these estimates (see also Gnanadesikan, 1980, and the appendix on computer software at the end of this book).

The above discussion has been cast in terms of an interest in internal comparisons among all $(n - 1)$ single-degree-of-freedom contrast vectors. This is, however, not a requisite in any application, and in some situations either one may wish to consider all n single degrees of freedom (although in most analysis of variance situations the overall mean effect would be real a priori and hence set aside in making the other assessments), or, more realistically, one may wish to compare internally only a subset of the $(n - 1)$ single-degree-of-freedom contrast vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$. The steps involved in the graphical internal comparisons procedure for assessing the relative magnitudes of $L[\leq(n - 1)]$ contrast vectors can therefore be summarized now:

1. Calculate the $(n - 1)$ single-degree-of-freedom contrasts for each response separately.
2. Form the p-dimensional contrast vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$.
3. Choose the subset of $L[\leq(n - 1)]$ contrast vectors to be internally compared—$\mathbf{x}_i$, $i = 1, \ldots, L$.
4. Select the positive semidefinite compounding matrix, $\mathbf{A}$.
5. Compute the measures of size (or the squared distances), $d_i = \mathbf{x}_i'\mathbf{A}\mathbf{x}_i$, $i = 1, 2, \ldots, L$, and order them to obtain $d_{(1)} \leqslant d_{(2)} \leqslant \cdots \leqslant d_{(L)}$.
6. Select the number K $(\leqslant L)$ on the basis of judgment.
7. Select the number M $(\leqslant K)$, and using $d_{(1)} \leqslant \cdots \leqslant d_{(M)}$, calculate

$$P = \frac{\prod_{i=1}^{M} [d_{(i)}]^{1/M}}{d_{(M)}} \quad \text{and} \quad S = \frac{\sum_{i=1}^{M} d_{(i)}}{M d_{(M)}}.$$

8. Using K/M, $d_{(M)}$, P, and S, determine the maximum likelihood estimates, $\hat{\lambda}$ and $\hat{\eta}$.
9. Plot $d_{(1)}, d_{(2)}, \ldots, d_{(L)}$ against the corresponding quantiles of the gamma distribution with parameters $\lambda = 1, \eta = \hat{\eta}$; that is, plot the points $\{\bar{x}_i, d_{(i)}\}$

for $i = 1, \ldots, L$, where $\tilde{x}_i$ is defined by

$$\int_0^{\tilde{x}_i} \frac{1}{\Gamma(\hat{\eta})} u^{\hat{\eta}-1} \exp(-u)\, du = p_i,$$

for a specified cumulative probability p_i [e.g., $(i - \frac{1}{2})/L$, $(i - \frac{1}{3})/L + \frac{1}{3}$ or $i/(L + 1)$].

Before presenting examples of application of this graphical internal comparisons method, a few comments on certain features of the method may be appropriate. First, with regard to the choice of the compounding matrix A, it has already been stated that data-analytic wisdom suggests the use of several A's in analyzing any given set of data. Once again, the point is that any truly multivariate situation cannot usually be fully described by any single unidimensional representation, and the implication of this here is that different choices for A may lead to quite different and possibly interesting insights into the multivariate nature of the data. A flexible collection to use has, however, been developed (see Section II of Appendix C in Roy et al., 1971, and also Wilk et al., 1962), and the following is a list of the set:

1. $\mathbf{A}_1 = \mathbf{I}$, the identity matrix.
2. $\mathbf{A}_2 = \mathbf{S}_L^{-1}$, the inverse of a covariance (or sum-of-products) matrix obtained from all L contrast vectors to be internally compared.
3. $\mathbf{A}_3 = \mathbf{S}_R^{-1}(\mathbf{A}_1)$, the inverse of a sum-of-products matrix obtained from the R ($<L$) contrast vectors whose associated distances based on the compounding matrix $\mathbf{A}_1$ are the R smallest distances (see also the discussion of robust estimators of dispersion in Section 5.2.3).
4. $\mathbf{A}_4 = \mathbf{S}_R^{-1}(\mathbf{A}_2)$, the inverse of a sum-of-products matrix obtained from the R ($<L$) contrast vectors whose associated distances based on the compounding matrix $\mathbf{A}_2$ are the R smallest distances.
5. $\mathbf{A}_5 = \mathbf{S}^{-1}$, the inverse of a sum-of-products matrix based on a user-specified subset of contrast vectors.
6. $\mathbf{A}_6 = \mathbf{D}(1/s_{ii}(\mathbf{A}_2))$, a diagonal matrix of the reciprocals of the diagonal elements of $\mathbf{A}_2^{-1}$.

7–9. For $j = 7, 8, 9$, $\mathbf{A}_j = \mathbf{D}(1/s_{ii}(\mathbf{A}_l))$, a diagonal matrix of the reciprocals of the diagonal elements of $\mathbf{A}_l^{-1}$, $l = 3, 4, 5$.

10. $\mathbf{A}_{10} = \mathbf{a}_l\mathbf{a}_l'$, where $\mathbf{a}_l$ is the eigenvector (or principal component) corresponding to the lth largest eivenvalue of either the correlation matrix or the covariance matrix computed from all L contrast vectors (see Section 2.2.1 for a discussion of principal components).

A second issue in using the gamma probability plotting procedure outlined earlier involves the choice of values for K and M, which is deliberately left to

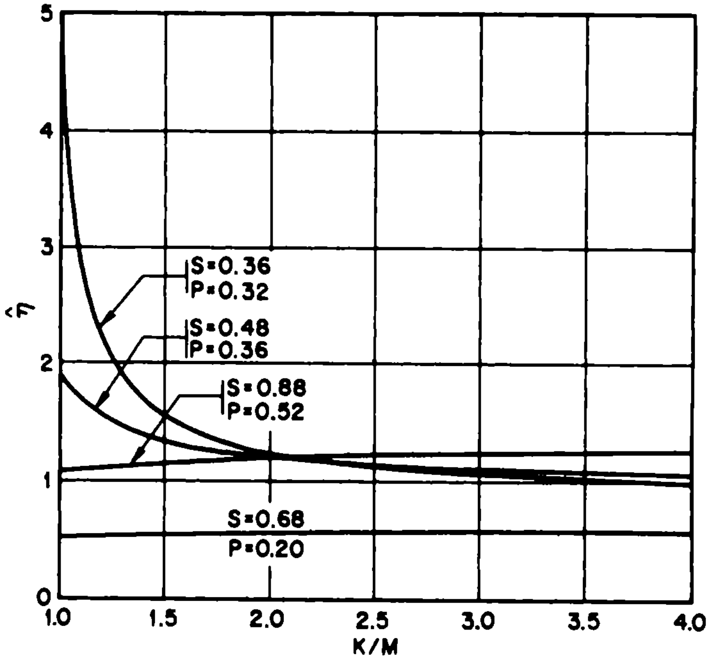


Fig. 15. Dependence of $\hat{\eta}$ on K/M for various P and S values.

the user's discretion. Since this is an informal statistical tool, which is not concerned with such things as the precise significance levels to be associated with formal tests of hypotheses, it would be expected that both prior and posterior (i.e., after seeing the data) considerations would influence the choice of values for K and M. The choice of these values would, of course, affect the estimates of λ and η, and one concern may be the sensitivity of these estimates to the choices of K and M. Figure 15 shows a plot of the maximum likelihood estimate of η as a function of K/M for various values of P and S (defined earlier), and it is seen that the estimate of η is quite insensitive to the value of K provided that M is not too close to K. For many two-level factorial experiments a choice of K and M such that $K/M > 3/2$ seems to be recommendable as a relatively safe rule. In most situations the loss of efficiency in estimating η due to choosing a small value of M relative to K appears to have little or no effect on the interpretations of the configurations observed on the gamma probability plots (see Wilk & Gnanadesikan, 1964; and Chapter VII of Roy et al., 1971).

A third facet of the method is that as a multiresponse technique it is intended to augment rather than to replace analyses of various subsets of the responses, including the responses considered individually. The analyses of subsets by comparable gamma probability plots are accomplished quite easily by suitably modifying the $p \times p$ compounding matrix $\mathbf{A}$. For example, if one is interested in studying a subset of q ($<p$) of the initial variables, a choice of zero, for all the elements in rows and columns of $\mathbf{A}$ that correspond to the complementary subset of $(p - q)$ of the variables, will yield squared distances based only on the chosen set of q variables. [*Note*: If $\mathbf{A}$ is to be an inverse of a covariance or sum-of-products matrix, an appropriate procedure may be to invert the matrix for the q responses of interest rather than extracting a $q \times q$ matrix from the inverse of the full $p \times p$ matrix.] When $q = 1$ (i.e., a single response is to be analyzed by itself), $\mathbf{A}$ may be specified as a matrix all of whose elements are 0 except for a single element (which can be taken to be equal to 1) in the diagonal position corresponding to the particular response. For this choice of $\mathbf{A}$, to treat the case when $q = 1$, the squared distances are just the squared contrasts for the particular response, and it is customary to treat these as having a chi-squared distribution with 1 degree of freedom, which is equivalent to specifying $\eta = \frac{1}{2}$ instead of estimating a value for it. Example 44 contains some discussion of the issues pertaining to the productive interplay between an overall multiresponse analysis and separate univariate analyses of the individual responses.

Example 44. The data are from a study of nine factors thought to affect Picturephone® quality, and the experiment was organized as a one-half replicate of a 2^9 factorial in a split-plot design (see Wilk & Gnanadesikan, 1964, and also Chapter VII of Roy et al., 1971). There were $p = 8$ responses

Exhibit 44a. Gamma probability plot for PICTUREPHONE® data; $L = 129$, $\mathbf{A} = \mathbf{I}$, $M = 64$, $K/M = 2$, $\hat{\eta} = 2.33$

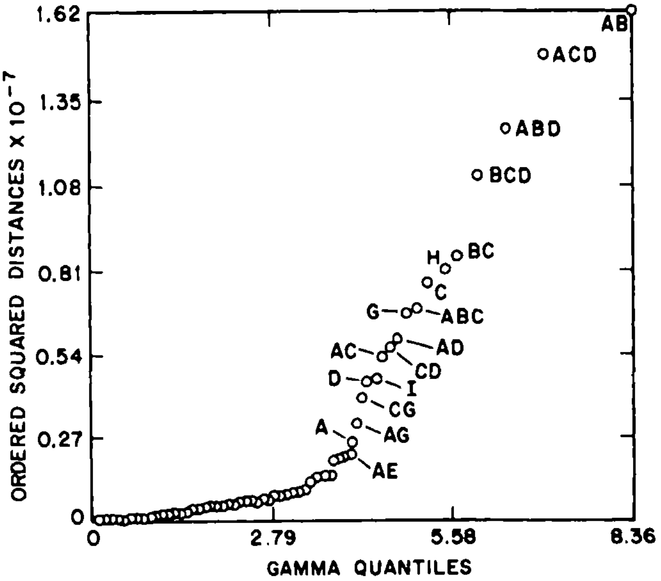
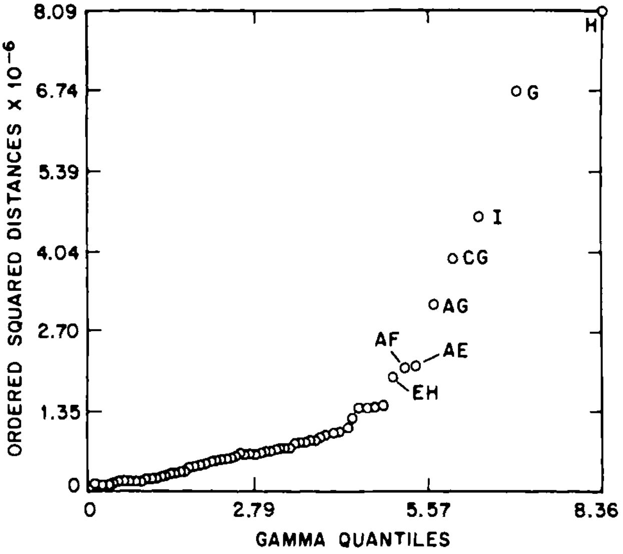


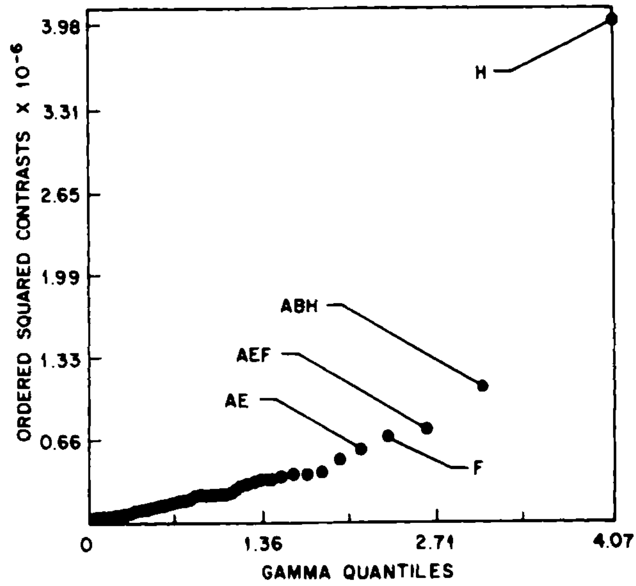
Exhibit 44*b*. Gamma probability plot for PICTUREPHONE® data; $L = 115$, $\mathbf{A} = \mathbf{I}$, $M = 57$, $K/M = 2$, $\hat{\eta} = 2.40$



per experimental unit, and each was a subjective assessment of picture quality on a 10-point scale. Exhibit 44*a* shows a gamma probability plot of squared distances obtained by the above method for the 129 main effects and two- and three-factor interactions involved. The squared distances are labeled by the treatment effects to which they correspond. The choice for **A** in this example was the identity matrix, and the values of K/M, M, and the estimated shape parameter are all shown in the figure caption.

One interpretation of this figure is that the "large" points (viz., those that correspond to the larger squared distances and appear in the plot toward the right-hand top) all correspond to real treatment effects. On the other hand, one might also think that the configuration is suggestive of two intersecting straight lines. Since the experiment was of a split-plot type, there are whole-plot factors and subplot factors. Typically, in a split-plot experiment, the whole-plot factors will have a variance or, in this multiresponse case, a covariance structure that is more dispersed than the covariance structure for the subplot factors. It turns out here that 14 of the 17 points on the suggested line of steeper slope correspond to effects involving the whole-plot factors, *A*, *B*, *C*, and *D*. Thus the two intersecting lines correspond to groups of treatment effects with different covariance structures, and one ought to split up the collection of treatment effects into those that have a whole-plot covariance structure and another set that has a subplot covariance structure. A gamma probability plot of squared distances for the 115 main effects and two- and three-factor interactions not solely confined to the whole-plot factors is shown in Exhibit 44*b*. On this plot

Exhibit 44c. Gamma probability plot ($\eta = \frac{1}{2}$) of 115 squared contrasts for second variable



one can identify the "top" 7 or 8 points as being indicative of real treatment effects.

As with all multiresponse methods, it would be legitimate in this example to inquire what might have happened if one had carried out the analysis of this experiment by doing separate analyses of the eight responses involved, perhaps using the uniresponse probability plotting technique of cell I in Table 2. Exhibits 44c, *d*, and *e* (see pages 270–271) show typical $\chi^2_{(1)}$ (actually, gamma with shape parameter $\frac{1}{2}$) probability plots of the 115 squared contrasts for three of the responses. [*Note*: A chi-squared-with-one-degree-of-freedom probability plot of the squared contrasts should yield a configuration "equivalent" to one obtained by making a half-normal plot of the absolute contrasts.] These three plots are typical of the eight plots that one can get by analyzing the responses separately. Comparing Exhibit 44*b* with these three figures, one sees that one is able to identify many more possibly real effects in the multivariate analysis than in the separate univariate analyses. In this example the combined evidence from the eight separate uniresponse analyses is that three or four of the treatment effects are possibly real, whereas the multiresponse analysis provides evidence that seven or eight of the effects may be real.

One possible explanation for the greater sensitivity of the multiresponse analysis in this example is provided by the estimated value of the shape parameter, namely, $\hat{\eta} = 2.4$. If the responses were indeed statistically independent, one might expect the squared distances to be distributed as a chi-squared variate with 8 $(=p)$ degrees of freedom or, equivalently, as a gamma variate

Exhibit 44d. Gamma probability plot ($\eta = \frac{1}{2}$) of 115 squared contrasts for third variable

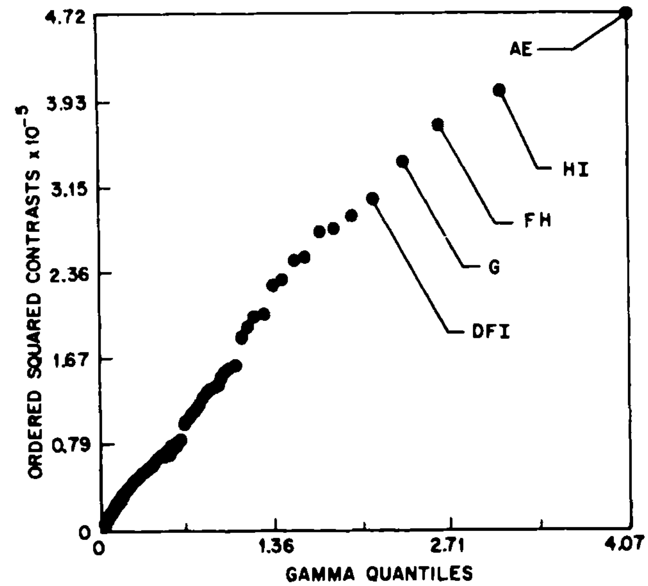
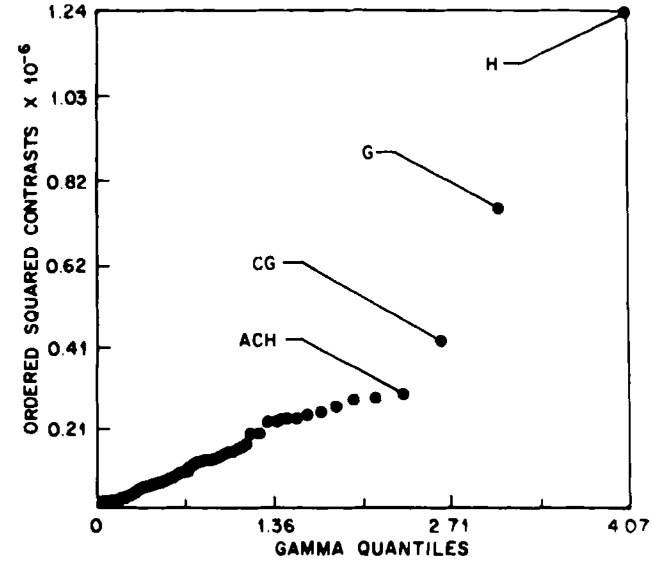


Exhibit 44e. Gamma probability plot ($\eta = \frac{1}{2}$) of 115 squared contrasts for fifth variable



with shape parameter 4 $(=p/2)$. Thus the lower value for $\hat{\eta}$ suggests that there is probably an accumulation of several fairly small real effects on the separate response scales into a smaller-dimensional space, which is then revealed better by the multiresponse analysis. [*Note*: An interesting modification of the separate uniresponse analyses suggested by this is to estimate a shape parameter for the probability plot of the squared contrasts, rather than using the prespecified $\chi^2_{(1)}$ distribution for them.] Also, there is perhaps a stabilizing effect on the error configuration (i.e., the linear part of the plot) due to the intercorrelations among the responses.

Example 45. This example is based on data (cf. Chapters IV and VII of Roy et al., 1971) from a one-quarter replicate of a 2^7 experiment concerned with seven factors that might affect the operation of a detergent manufacturing process. The original study involved measurements on seven responses, but for present purposes only a bivariate subset of the original seven is considered. The two responses are called *rate* (bins/hour) and *stickiness*, and Exhibit 45*a* shows the 32 bivariate observations involved, together with the treatment combinations that label them. The seven experimental factors involved were as follows: A—air injection, B—nozzle temperature, C—crutcher amperes, D—inlet temperature, E—tower air flow, F—number of baffles, and G—nozzle pressure.

Andrews et al. (1971) used these data as an example for applying their methods (see Section 5.3) for developing data-based transformations, and the discussion here is drawn largely from their paper. An indirect way of assessing the data-based transformation methods described earlier in Section 5.3 is to study the effects of the transformations on the outputs of statistical analyses, such as analyses of variance, performed before and after the transformations. In the present example, for instance, one can obtain 31 estimated treatment effects (or single-degree-of-freedom contrasts) of interest for each response. One can use the graphical internal comparisons technique of cell I in Table 2 for simultaneously assessing the 31 effects, namely, via either a half-normal probability plot of the absolute values of the estimated effects, or equivalently, a $\chi^2_{(1)}$ probability plot of the squared values. One can do this for effects estimated on both the untransformed and the transformed scales of the responses and compare the resulting configurations. It is perhaps reasonable to expect that, because of the averaging involved in obtaining the estimated treatment effects, except for bad nonnormality of the original observations the estimated effects would be adequately normal in distribution. For instance in the last example, despite the initial 10-point scale for the eight responses, all the probability plots associated with the contrasts (viz., Exhibits 44*c*, *d*, and *e*) are quite linear for the major part, and the lack of systematic curvilinearity in these plots lends credence to the "expected" normality of the contrasts. However, in the present example, Exhibits 45*b* and *c* (see page 274), which are $\chi^2_{(1)}$ (or corresponding gamma) probability plots of the squared contrasts on each of the two untransformed response scales, appear to indicate the presence

Exhibit 45a. Data matrix for 2^{7-2} experiment on detergent manufacturing process (cf. Roy et al., 1971, p. 54)

Run No.	Treatment Combination	Rate (bins/hour)	Stickiness
1	(1)	38.0	5.40
2	*afg*	38.0	5.90
3	*bfg*	35.0	2.95
4	*ab*	36.0	5.38
5	*cg*	38.0	5.22
6	*acf*	37.0	5.33
7	*bcf*	37.0	4.90
8	*abcg*	36.0	4.50
9	*df*	34.5	3.15
10	*adg*	38.0	3.06
11	*bdg*	36.0	5.70
12	*abdf*	37.0	4.20
13	*cdfg*	38.5	4.70
14	*acd*	38.0	4.20
15	*bcd*	38.0	5.17
16	*abcdfg*	39.0	5.66
17	*eg*	37.0	4.60
18	*aef*	38.0	5.20
19	*bef*	32.0	2.49
20	*abeg*	39.0	6.10
21	*ce*	39.0	3.84
22	*acefg*	37.0	4.90
23	*bcefg*	35.0	4.30
24	*abce*	34.0	3.50
25	*defg*	37.0	3.24
26	*ade*	37.0	3.79
27	*bde*	39.0	5.80
28	*abdefg*	39.0	5.30
29	*cdef*	39.0	5.60
30	*acdeg*	40.0	6.20
31	*bcdeg*	40.0	5.47
32	*abcdef*	40.0	4.77

of considerable distributional peculiarities. The extremely "choppy" appearance of the lower left-hand end of Exhibit 45*b* (see page 274) can perhaps be attributed to the essentially discrete nature of the response termed *rate*, as evident in Exhibit 45*a*, whereas the distributional departure indicated in Exhibit 45*c* (see page 274) and associated with the other response seems to be more subtle.

The transformation method of Box & Cox (1964) to improve marginal normality and the one proposed by Andrews et al. (1971) for enhancing joint

Exhibit 45*b*. Gamma probability plot ($\eta = \frac{1}{2}$) of 31 squared contrasts for untransformed rate data

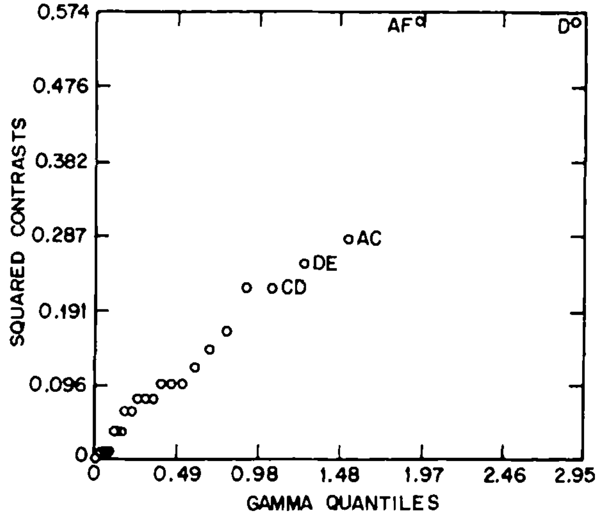
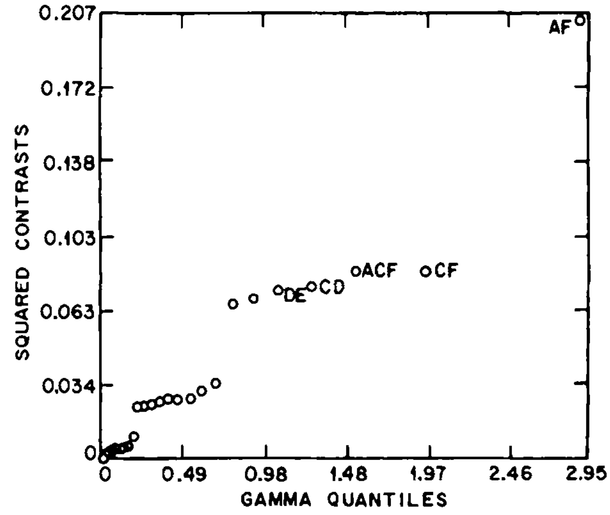


Exhibit 45*c*. Gamma probability plot ($\eta = \frac{1}{2}$) of 31 squared contrasts for untransformed stickiness data

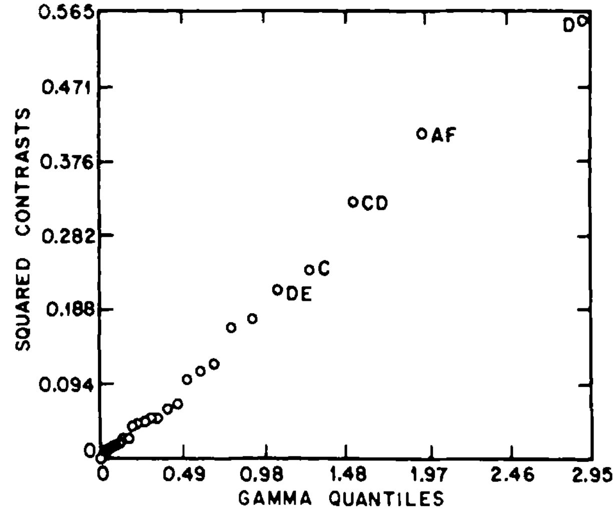


normality (see Section 5.3) were employed, and the estimated values of the power transformation parameters are shown in Exhibit 45*d* (see page 275). In estimating these transformations, in addition to enhancing normality an attempt was made to reduce nonadditivities at the same time by specifying a fit (or linear model) solely in terms of the seven main effects on the transformed scales of the responses.

Exhibit 45d. Estimates of transformation parameters for detergent manufacture data (cf. Andrews et al., 1971)

Box-Cox Method Estimates		Andrews et al. Method Estimates	
$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$
8.88	2.06	7.22	1.88

Exhibit 45e. Gamma probability plot ($\eta = \frac{1}{2}$) of 31 squared contrasts for rate data transformed by Method I



The improvements achieved by using the transformations determined by the Box & Cox method are evident in Exhibits 45*e* and *f*, which show the $\chi^2_{(1)}$ probability plots for squared effects on the transformed scales of the two variables involved. The smoother configurations of these two plots, especially at the lower end, suggest not only possible improvement of underlying normality but also the delineation of a more homogeneous grouping of fairly small effects, from which one can hopefully derive a "cleaner" estimate of error variance.

A similar evaluation of the method proposed by Andrews et al. (1971) for enhancing joint normality can be made by comparing the two gamma probability plots of the squared distances obtained from both the untransformed bivariate data and the transformed bivariate observations obtained by

Exhibit 45*f*. Gamma probability plot ($\eta = \frac{1}{2}$) of 31 squared contrasts for stickiness data transformed by Method I

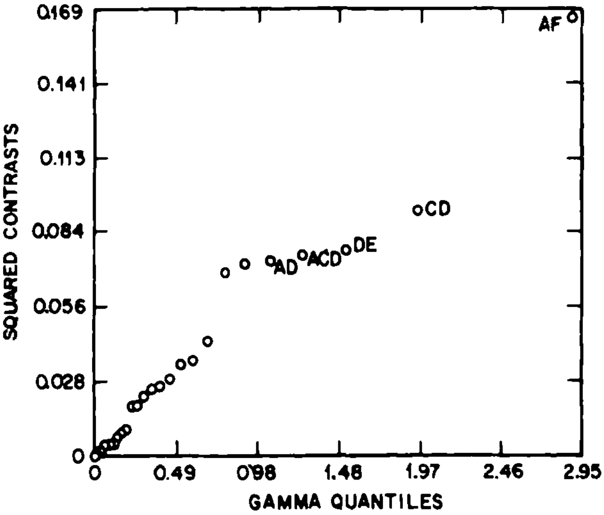
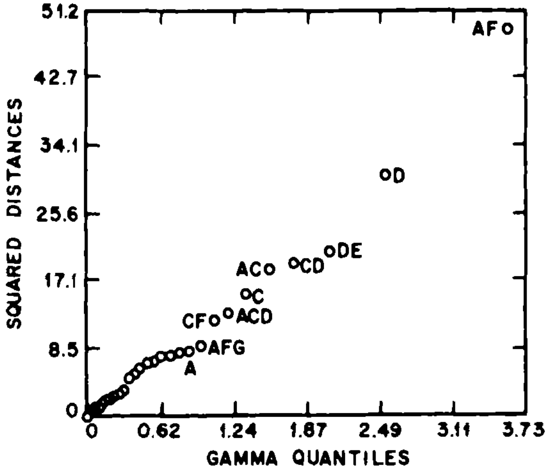
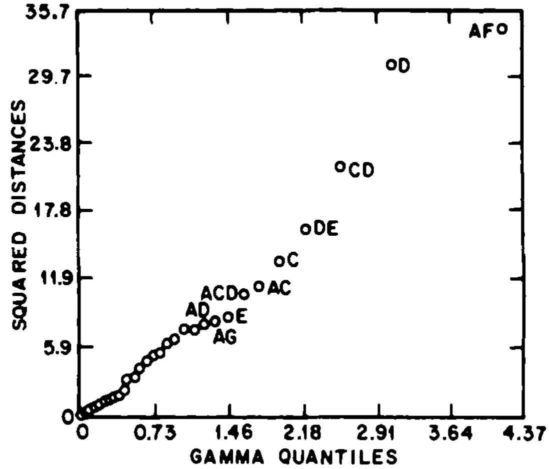


Exhibit 45*g*. Bivariate gamma probability plot ($\hat{\eta} = 0.75$) for detergent data untransformed



using the powers in the second set of columns of Exhibit 45*d* for the two variables. Exhibits 45*g* and *h* show the gamma probability plots, the former derived from the untransformed observations and the latter from observations transformed by using $\hat{\lambda}_1 = 7.22$ and $\hat{\lambda}_2 = 1.88$. The choice for the compounding matrix in both these plots was of the A_3 type, involving a subselection of R contrast vectors with smallest Euclidean lengths (see the list of choices for A given earlier); in particular, R was taken to be 15. Also, for both plots, K was

Exhibit 45*h*. Bivariate gamma probability plot ($\hat{\eta} = 1.01$) for detergent data transformed by Method II



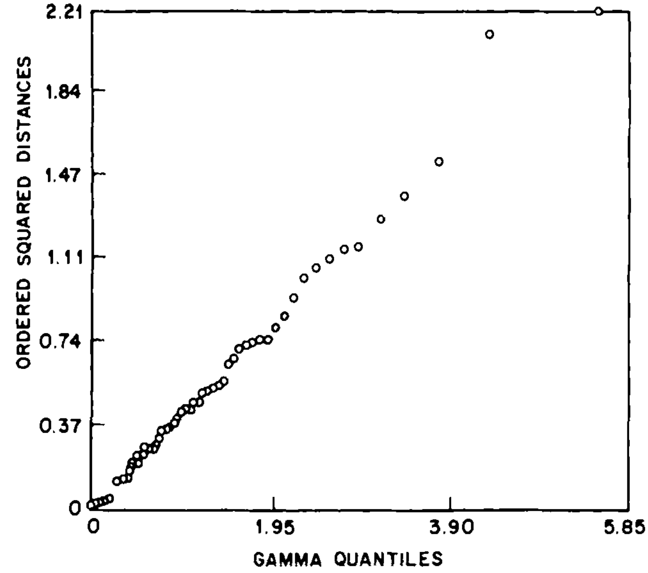
considered to be 31 and the value of M was taken as 15, so that $K/M = 2.07$. The estimated values of the shape parameter are indicated in the captions for the figures.

Not only is the null configuration of the "smaller" points (i.e., the ones in the lower left-hand corner) in Exhibit 45*h* smoother, but also the delineation of the departures at the "large" end (viz., the upper right-hand corner) is clearer than in Exhibit 45*g*. An interesting feature, however, is that the improvement in the null configuration (i.e., the lower end) in going from Exhibit 45*g* to *h* is nowhere near as striking as the improvements from Exhibit 45*b* to *e* or from Exhibit 45*c* to *f*. This suggests that the approach used in the internal comparisons method for estimating the shape parameter may be introducing a very valuable robustness into the process, inasmuch as the suspected marked nonnormality of the responses does not seem to unduly distort the configurations on the gamma probability plot.

Example 46. The main purpose of this example, taken from Wilk & Gnanadesikan (1964), is to provide a rather dramatic illustration of the importance of using more than one choice for the compounding matrix **A**. Sixty random deviates were generated from a five-dimensional normal distribution with zero mean vector and a distinctly nonspherical covariance matrix,

$$\Sigma = \begin{pmatrix} 1 & & & & \\ 2 & 5 & & & \\ -3 & -7 & 11 & & \\ 4 & 10 & -16 & 25 & \\ 2 & 5 & -8 & 9 & 16 \end{pmatrix},$$

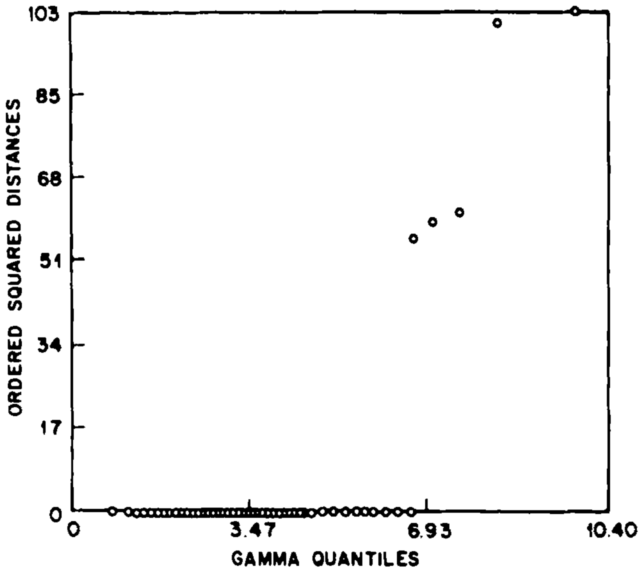
Exhibit 46a. Gamma probability plot for artificial five-dimensional data; $L = 55$, $\mathbf{A} = \mathbf{I}$, $M = 30$, $K/M = 1.4$, $\hat{\eta} = 1.4$



where the elements above the diagonal are, of course, obtainable by symmetry. To a random selection of 10 of these 60 observations were added certain constants to shift their means and thus simulate "real" effects. The shifts chosen were as follows: three vectors equal to (3, 7, 10, 12, 11), three others equal to (5, 5, 5, 5, 5), three more equal to (7, 2, 0, 5, 4), and a last one equal to (5, 8, 15, 20, 18).

Exhibits 46*a* and *b* are gamma probability plots of the squared distances derived from these observations for two choices of $\mathbf{A}$, namely, $\mathbf{A} = \mathbf{I}$ and $\mathbf{A} = \mathbf{S}^{-1}$, the inverse of a sum-of-products matrix based on a random selection of 30 out of the 50 "central" (i.e., zero mean) observations. Each figure is actually a plot of only the 55 smallest squared distances instead of all 60 of them, and the values of M and K/M, as well as the resulting estimate of η, are all indicated in the captions. The ranges of the values of the squared distances are quite different, as are the two values of $\hat{\eta}$, but the most striking thing about the two figures is that Exhibit 46*a* contains no indication of the five known nonnull observations, whereas Exhibit 46*b* clearly delineates them. In light of the prespecified nonsphericity of the covarince matrix used in generating the observations, it is perhaps not surprising that the identity matrix is not a very appropriate choice. The main implication in practice, however, is that without a considerable amount of knowledge about the data a safe rule would be to use different choices of $\mathbf{A}$ and then to compare the results to gain further insights into the structure of the data.

Exhibit 46*b*. Gamma probability plot for artificial five-dimensional data; $L = 55$, $A = S_{30}^{-1}$, $M = 30$, $K/M = 1.4$, $\hat{\eta} = 3.83$



6.3.2. Graphical Internal Comparisons among Equal-Degree-of-Freedom Groupings

This subsection deals with probability plotting techniques for Cell V of Table 2. In the geometrical terms used in Section 5.2.1 for describing the concepts and processes of orthogonal multivariate analysis of variance, the situation represented by this cell arises when the decomposition of n-space into orthogonal subspaces contains a set of r mutually orthogonal linear subspaces, each of dimensionlity ν (>1), and the observations are p-dimensional. Thus the prototype here is a situation in which there are r $p \times p$ sum-of-products matrices, $S_1, \ldots, S_r$, each based on ν degrees of freedom, and one wishes to compare simultaneously the "sizes" of the dispersions summarized by these matrices, or equivalently, by the mean sum-of-products matrices, S_i/ν's, using probability plotting techniques. One example of this occurs when internal comparisons are desired among all the main effects, or interactions of the same order, in an m-level factorial experiment. In this case all the main effects will have $\nu = (m - 1)$ degrees of freedom, and each qth order interaction will have $\nu = (m - 1)^{q+1}$. Another example occurs when one has $(\nu + 1)$ replications within cells and wishes to assess the validity of the assumption of a common within-cell covariance structure.

An intrinsic difficulty of the present problem is to define measures of "size" of a dispersion matrix. One should not expect that any single measure will provide an adequate summary of the dispersion information contained in the matrix. Certain functions of the eigenvalues of a dispersion matrix may be used as unidimensional summaries of the size of the dispersion; see, for example, Roy et al. (1971, Chapter II, Section 3). Two such functions are the arithmetic mean, or sum, and the geometric mean of the eigenvalues. Since the arithmetic mean is sensitive to very large and very small eigenvalues, whereas the geometric mean tends to be particularly sensitive only to very small eigenvalues, the two functions may lead to different insights concerning the dispersion structure. The use of both functions is recommended for data-analytic purposes, and the two methods described below are based, respectively, on the two functions.

In analysis of variance applications, such as the factorial experiment mentioned earlier, the dimensionality of response p may often exceed the value v. In this case the matrices $\mathbf{S}_1, \ldots, \mathbf{S}_r$ will have v positive eigenvalues and $(p - v)$ zero eigenvalues. A natural modification of the second function, therefore, is to consider the geometric mean of the nonzero eigenvalues. Specifically, then, the two functions to be considered as measures of size of a sum-of-products matrix, $\mathbf{S} = ((s_{ij}))$, with eigenvalues $c_1 \geqslant \cdots \geqslant c_t > 0$, are

$$\mathscr{A} = \sum_{i=1}^{t} c_i = \operatorname{tr}(\mathbf{S}) = \sum_{j=1}^{p} s_{jj},$$

$$\mathscr{G} = \left(\prod_{i=1}^{t} c_i\right)^{1/t}.$$

When the different responses in a multiresponse analysis of variance are measured on very different scales, it may be desirable to weight the responses accordingly, so that deviations from null conditions on the different response scales are not given the same weight. For incorporating this feature in the present mode of analysis, one can use as starting points in the analysis not just the sum of products matrices, $\mathbf{S}_i$'s, but also scaled versions of them, namely, $\mathbf{S}_1\mathbf{A}, \ldots, \mathbf{S}_r\mathbf{A}$, where $\mathbf{A}$ is a positive semidefinite matrix. [*Note*: Computationally, the eigenvalues required for $\mathscr{G}$ may be obtained either from a singular-value decomposition appropriate to problems, such as discriminant analysis, that involves two covariance matrices, or from eigenanalyses of the symmetric matrices, $\mathbf{Z}_i'\mathbf{A}\mathbf{Z}_i$, where $\mathbf{S}_i = \mathbf{Z}_i\mathbf{Z}_i'$, rather than eigenanalyses of the asymmetric forms, $\mathbf{S}_i\mathbf{A}$, using the mathematical property that the nonzero eigenvalues of $\mathbf{Z}_i\mathbf{Z}_i'\mathbf{A}$ are also the nonzero eigenvalues of $\mathbf{Z}_i'\mathbf{A}\mathbf{Z}_i$.] However, when the $\mathbf{S}_i$'s and $\mathbf{A}$ are all positive definite, then, since the product of the eigenvalues of $\mathbf{S}_i\mathbf{A}$ is

$$\prod_{j=1}^{p} c_j = |\mathbf{S}_i\mathbf{A}| = |\mathbf{S}_i|\,|\mathbf{A}|,$$

the scaling by $\mathbf{A}$ is immaterial for purposes of internal comparisons among the $\mathbf{S}_i$'s in terms of the statistic $\mathscr{G}$.

The matrix $\mathbf{A}$ plays the same role here as the compounding matrix $\mathbf{A}$ did in the method discussed in Section 6.3.1. Hence, as before, possible choices for the

$p \times p$ matrix $\mathbf{A}$ include (i) the identity matrix, which may be appropriate when $\mathbf{S}_1, \ldots, \mathbf{S}_r$ pertain to a decomposition of the error covariance structure; (ii) a diagonal matrix of reciprocals of variances of the responses; and (iii) the inverse of a covariance matrix of the responses. Once again, under choices (ii) and (iii), the matrix $\mathbf{A}$ may be either prespecified or estimated from the data on hand, and in either case, since it is used as a common factor to scale all the $\mathbf{S}_i$'s, it is considered a fixed matrix for the subsequent internal comparisons analyses, just as in Section 6.3.1. For the rest of this subsection, it is to be understood that the internal comparisons of the "magnitudes" of $\mathbf{S}_1, \ldots, \mathbf{S}_r$ are to be made via the r associated values of either $\mathscr{A}$ or $\mathscr{G}$,

$$a_i = \mathrm{tr}(\mathbf{S}_i\mathbf{A}), \qquad g_i = \left\{\prod_{j=1}^{t} c_j(\mathbf{S}_i\mathbf{A})\right\}^{1/t}.$$

Next an evaluating distribution is needed for each of these collections. If such a distribution were available for the statistic $\mathscr{A}$, for instance, one could obtain a probability plot of the ordered values, $0 < a_{(1)} \leqslant \cdots \leqslant a_{(r)}$, against the corresponding quantiles of the distribution. A similar use may be made of the distribution of $\mathscr{G}$. Under null conditions the distributions of both $\mathscr{A}$ and $\mathscr{G}$ turn out to be well approximated by gamma distributions.

That this is so for $\mathscr{A}$ can be seen by recognizing that $\mathscr{A}$ is the sum of v mutually independent positive semidefinite quadratic forms, each of whose distributions may itself be adequately approximated by a gamma distribution, as stated in Section 6.3.1. Specifically, each matrix $\mathbf{S}_i$ can be represented as $\mathbf{Z}_i\mathbf{Z}_i'$, where $\mathbf{Z}_i' = \mathbf{R}_i\mathbf{Y}'$. Furthermore, $\mathbf{Y}'$ is the $n \times p$ matrix of original observations, and the $v \times n$ matrix, $\mathbf{R}_i$, is such that $\mathbf{R}_i\mathbf{R}_i' = \mathbf{I}(v)$ and $\mathbf{R}_i\mathbf{R}_j' = \mathbf{O}$ $(i \neq j)$. Then

$$a_i = \mathrm{tr}(\mathbf{Z}_i\mathbf{Z}_i'\mathbf{A}) = \sum_{j=1}^{v} \mathbf{z}_{ij}'\mathbf{A}\mathbf{z}_{ij},$$

where $\mathbf{z}_{ij}$ is the jth column of $\mathbf{Z}_i$. [*Note*: This way of looking at the measure of size $\mathscr{A}$ is discussed also in Section 5 of Chapter VII of Roy et al., 1971.] Null assumptions, which may be employed to develop methodology for studying specific departures from them, are that the original observations are mutually independent with identical, but unknown, covariance matrices $\boldsymbol{\Sigma}$ and that there are no real effects associated with the r groups. Under such null conditions, $\mathbf{z}_{ij}$ $(i = 1, \ldots, r; j = 1, \ldots, v)$ may be considered as a random sample from $N(\mathbf{0}, \boldsymbol{\Sigma})$, so that a_i is the sum of v mutually independent positive semidefinite quadratic forms, and, furthermore, $a_1, \ldots, a_r$ are mutually independent. The normality assumption concerning the $\mathbf{z}_{ij}$'s is not unreasonable since they are linear combinations of the original observations. As stated in Section 6.3.1, under null conditions the distribution of each quadratic form can be adequately approximated by a gamma distribution with scale parameter λ_a and

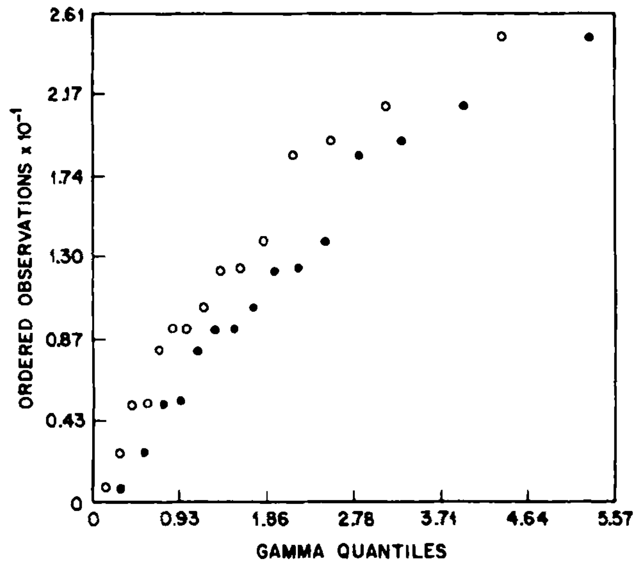


Fig. 16. Gamma probability plots for two estimates of shape parameter; ○—method of moments estimate; ●—maximum likelihood estimate.

shape parameter η_a/ν, and hence $a_1, \ldots, a_r$ may be considered as approximately a random sample from a gamma distribution with scale parameter λ_a and shape parameter η_a.

The use of $\mathscr{A}$ is thus seen to be a direct extension of the method discussed in Section 6.3.1 for the single-degree-of-freedom case, and one may wonder why an analysis of the νr quadratic forms, $\mathbf{z}'_{ij}\mathbf{A}\mathbf{z}_{ij}$, by that method is not adequate for the present problem. The issue, of course, is that the orthogonal decomposition yielding the individual p-dimensional vectors, $\mathbf{z}_{ij}$, is arbitrary and may not have any meaningful interpretation, whereas the a_i's are defined uniquely and meaningfully. The problem is the same here as in uniresponse analysis of variance, where a sum of squares with ν degrees of freedom does not necessarily have a unique meaningful decomposition into ν orthogonal single degrees of freedom.

In connection with the null distribution of $\mathscr{G}$, Hoel (1937) suggests approximating the distribution of the geometric mean of the eigenvalues of a $p \times p$ sample covariance matrix based on a sample of size n with $p \leqslant (n-1)$ by a gamma distribution whose shape parameter is a function only of p and n but whose scale parameter is $|\Sigma|^{1/p}$ times a quantity involving p and n, where Σ is the unknown underlying covariance matrix. The unknown scale and shape parameters are then obtained by equating the first two moments. For present purposes the null distribution is approximated by a gamma distribution with unknown scale and shape parameters, λ_g and η_g, respectively, and the parameters are then estimated by maximum likelihood instead of the method of moments. A Monte Carlo investigation was carried out to check on the adequacy of the approach. The dots in Figure 16 constitute a typical gamma probability plot of geometric means from the Monte Carlo study, and the reasonably good linear configuration indicates that the present approach is adequate.

The O's in Figure 16 provide the probability plot for the same set of geometric means, employing Hoel's estimate of the shape parameter instead of the maximum likelihood estimate. A comparison of the two configurations suggests that the maximum likelihood method of fitting the approximating distribution is to be preferred to the method of moments.

In the analysis of variance application, in order to minimize the effects of possibly real sources of variation on the estimation of the scale and shape parameters, an order statistics formulation along the lines of Section 6.3.1 may be employed once again. Specifically, if

$$a_{(1)} \leqslant \cdots \leqslant a_{(r)} \qquad \text{or} \qquad g_{(1)} \leqslant \cdots \leqslant g_{(r)}$$

denote the ordered values of the trace or geometric mean, then, considering the M ($\leqslant K$) smallest of these as the M smallest order statistics in a random sample of size K ($\leqslant r$) from a gamma distribution, one can find the maximum likelihood estimate of the scale and shape parameters using only these M values (see Wilk et al., 1962b).

Next, using the estimate of the shape parameter, one can obtain a gamma probability plot of the r ordered values, $a_{(1)} \leqslant \cdots \leqslant a_{(r)}$ or $g_{(1)} \leqslant \cdots \leqslant g_{(r)}$. Under null conditions the resulting configuration would be expected to be linear with zero intercept and slope $1/\lambda_a$ on the plot of the a_i's (and $1/\lambda_g$ on the plot of the g_i's). Departures from linearity may then be studied for pinpointing violations of the null assumptions, such as the presence of possibly real sources of variation, the existence of more than one underlying error covariance structure, and other distributional peculiarities. The interpretation of these probability plots is similar to that of other probability plotting techniques that have been proposed for augmenting analyses of variance; in particular, it is quite analogous to the technique discussed in Section 6.3.1.

Two examples, taken from Gnanadesikan & Lee (1970), are given next to demonstrate the use of the techniques described above.

Example 47. This example (see also Example 4 in Chapter VII of Roy et al., 1971) consists of computer-generated trivariate normal data that simulate the results of a 30-cell experiment with four replications per cell. The data for 15 of the 30 cells had an underlying covariance matrix **I**, while in the remaining cells the covariance matrix was 9**I**. Exhibit 47*a* shows a gamma probability plot of the 30 ordered values of the trace of the within-cell covariance matrices for these data. The shape parameter required for this plot was based on the 15 ($=M$) smallest observed trace values, and K was taken as 30. Each point on the plot is labeled 1 or 2 according as it derives from a cell with one or the

Exhibit 47a. Gamma probability plot for $\mathscr{A}$ in the artificial data example; $A = I$, $r = K = 30$, $M = 15$, $\hat{\eta} = 1.996$

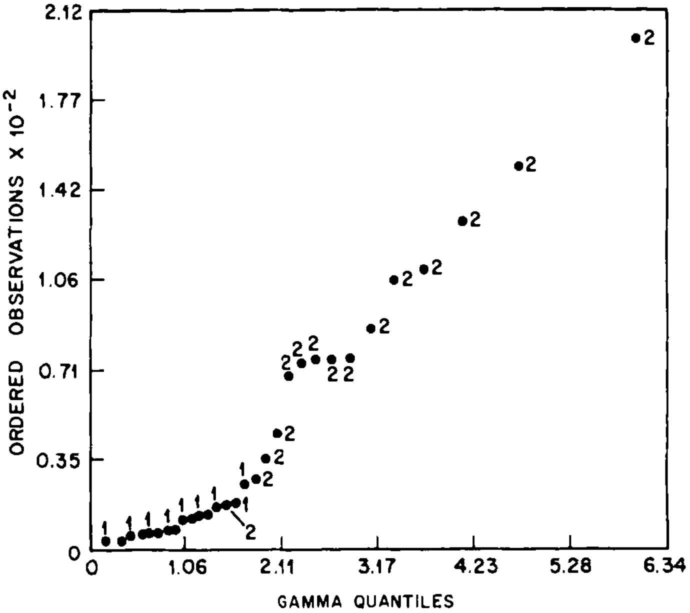
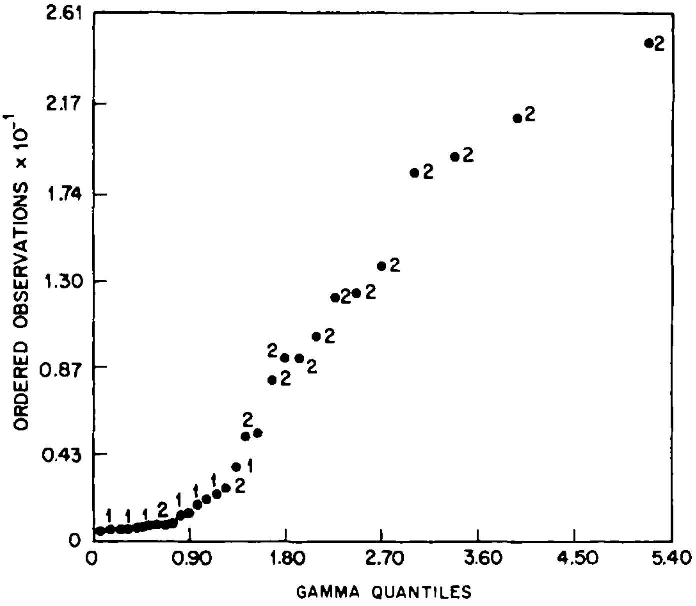


Exhibit 47b. Gamma probability for $\mathscr{G}$ in the artificial data example; $A = I$, $r = K = 30$, $M = 15$, $\hat{\eta} = 1.514$



other of the two covariance structures employed in generating the data. The configuration is suggestive of two intersecting straight lines, each of which consists of points that correspond to the cells with a common covariance structure. Exhibit 47*b* shows the analogous gamma probability plot for the 30 ordered values of the geometric mean statistic, and the same phenomenon of two intersecting straight lines is seen again, although the general configuration in Exhibit 47*b* is smoother than the one in Exhibit 47*a*.

Example 48. The set of data derives from the talker-identification problem used also as the basis for earlier examples (e.g., Examples 16 and 19 in Chapter 4). As part of the analysis for obtaining a discriminant space for representing the utterances (see Example 16) and for assigning an unknown to one of the contending speakers, it is usual to pool the within-speaker covariance matrices of the utterances to obtain an overall within-speakers covariance matrix. It is legitimate in such multivariate classification problems to inquire about the validity of such a pooling procedure, and it would be useful to have an informal statistical procedure for a preliminary, simultaneous intercomparison of the covariance matrices from the different speakers. Specifically, for a set of 10 speakers, one input to a classification analysis consisted of a six-dimensional

Exhibit 48*a*. Gamma probability plot for $\mathscr{A}$ in talker-identification example; $\mathbf{A} = \mathbf{I}$, $r = K = 10$, $M = 5$, $\hat{\eta} = 4.236$

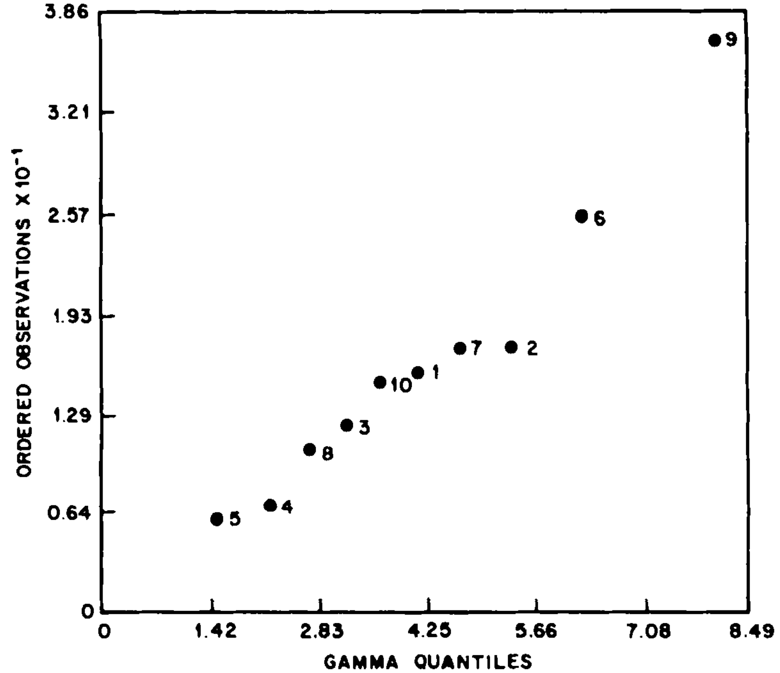
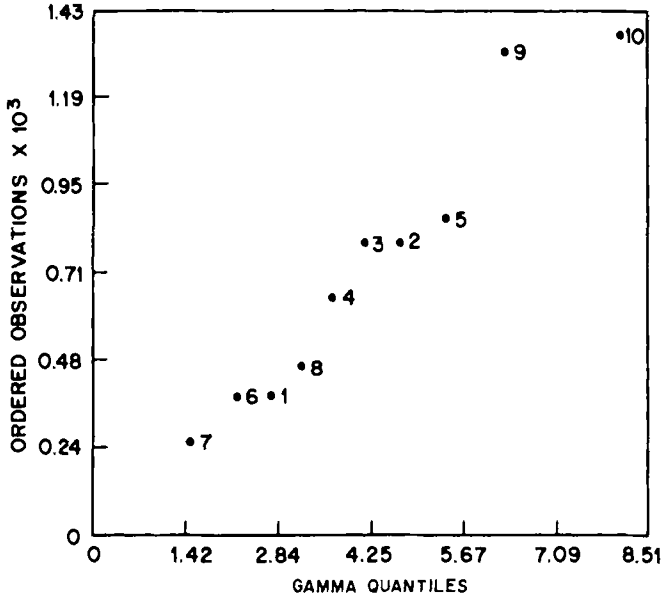


Exhibit 48*b*. Gamma probability plot for $\mathscr{G}$ in talker-identification example; $\hat{\eta} = 4.245$



representation of each utterance of a given word, and there were seven utterances available per speaker. As a preliminary to pooling the 10 within-speaker covariance matrices, one can assess their similarity in "size" by using the methods described above in this subsection. Exhibits 48*a* and *b* show the gamma probability plots for the 10 ($=r$) values of each of the functions $\mathscr{A}$ and $\mathscr{G}$, respectively, in this example. No scaling was performed on the covariance matrices, so that $\mathbf{A} = \mathbf{I}$, and the estimation of the parameters of the evaluating gamma distribution was based, in each case, on the $M = 5$ smallest observed values with $K = r = 10$. In this example the configurations obtained by using the estimates from the complete sample, that is, $M = K = 10$, were quite similar to the ones in Exhibits 48*a* and *b*.

The points in Exhibits 48*a* and *b* are labeled 1 through 10 to correspond with the speaker from whose covariance matrix a particular point derives. Although the point corresponding to speaker 9 appears to depart from the linear configuration suggested by the other points in Exhibit 48*a*, the departure is not marked. The general indication of both plots is that the 10 covariance structures form a reasonably homogeneous group in terms of their "sizes." The two plots exhibit different internal orderings of the covariance matrices for the 10 speakers, in accordance with the sensitivities of $\mathscr{A}$ and $\mathscr{G}$ to different aspects of the covariance structures. Thus, for example, speaker 6 who is second from the top in Exhibit 48*a*, turns out to be second from the bottom in Exhibit 48*b*, suggesting that the covariance matrix for that speaker may have a noticeably small eigenvalue, as was indeed the case. Also, the covariance matrix of speaker 9, the top point in Exhibit 48*a* and the next-to-top point in Exhibit 48*b*, is indicated as possibly having a markedly large eigenvalue and no significantly small eigenvalue, and this again was found to be true.

The essential concepts in the probability plotting approaches described in the last two subsections (and indeed also the others mentioned in Table 2) are first to obtain meaningful summary statistics to serve as the medium for making the simultaneous assessments and, second, to display the internal comparisons against a null background by means of a probability plot of the ordered observed values of a statistic versus the corresponding quantiles of an appropriate null statistical distribution. Thus, with the method of Section 6.3.1, the d_i's constitute the summary statistics, whereas the a_i's and g_i's are the corresponding entities for the method of Section 6.3.2, and the appropriate evaluating distribution in each case turns out to be a gamma distribution whose parameters are fitted by maximum likelihood, using an order statistics approach.

The probability plotting methods discussed heretofore have been concerned with internal comparisons of relative magnitudes and not with orientational aspects of multiresponse data. Orientational information is contained in eigenvectors associated with covariance matrices, whereas summary statistics such as $\mathscr{A}$ and $\mathscr{G}$ are based on eigenvalues. A simple starting point, for example, for comparing overall similarities of orientations of k sets of multiresponse data, would be in terms of the corresponding principal components of the covariance (or correlation) matrices of the data sets. For instance, one could ask if the k first principal components are similar, if the k second principal components are similar, and so on. Krzanowski (1979) has defined a measure of similarity between two sets of corresponding eigenvectors, and Flury (1984) has proposed a likelihood ratio test, based on normal distributional assumptions, for the null hypothesis that *all* the eigenvectors of several covariance matrices are the same. Keramidas et al. (1987) have suggested an informal, graphical, data analytic approach to the problem of assessing the similarities of sets of corresponding eigenvectors of covariance matrices, and their proposal is described next.

Given k sample covariance matrices, $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_k$, with $\mathbf{S}_i$ based on n_i (typically $>p$) observations, one has the spectral decomposition,

$$\mathbf{S}_i = \mathbf{X}_i \mathbf{D}_i \mathbf{X}_i', = 1, \ldots, k,$$

where $\mathbf{X}_i$ is an orthogonal matrix with p column vectors, ${}_j\mathbf{x}_i$ $(j = 1, \ldots, p)$, representing the eigenvectors of $\mathbf{S}_i$. If the n_i's are either equal, or all large enough, one can treat the eigenvectors from the k sets of data to be determined with roughly the same degree of precision. The diagonal matrix, $\mathbf{D}_i$, has as its diagonal elements the nonnegative eigenvalues of $\mathbf{S}_i$ in decreasing order of magnitude. The problem of comparing the orientations of the k data sets can then be formulated as one of assessing the statistical similarity of a set of k

eigenvectors, $\{{}_j\mathbf{x}_i\}$, $i = 1, \ldots, k$, where j indicates the eigenvector corresponding to the jth eigenvalue of each covariance matrix.

The problem of comparing several p-dimensional vectors is clearly complex. To simplify the task, Keramidas et al. (1987) propose a one-dimensional measure of the similarity among the vectors, and then use the observed values of such a measure as the basis of the assessment of similarity. It should, however, be reemphasized that any one-dimensional representation of inherently high-dimensional information may not be complete despite its attractive simplicity.

In considering a comparison of k eigenvectors, there are two key issues that have to be addressed. First, for all k covariance matrices, does there exist a sufficient separation between the jth eigenvalue and the remaining eigenvalues so that the corresponding eigenvector, ${}_j\mathbf{x}_i$, is well determined? The difficulty is that, without such clear separation, the eigenvectors are merely a random set of orthogonal directions in a subspace in which the scatter of the data is essentially spherical. If k is large, Keramidas et al. (1987) suggest using p side-by-side box plots of the k realizations of each ordered eigenvalue to evaluate the separations amongst eigenvalues. Also, the Q-Q type of probability plot, or the augmented scree plot, described in Section 6.2 for assessing separations amongst eigenvalues can be used for this purpose. There will be k such plots involved.

The second issue in comparing $\{{}_j\mathbf{x}_i\}$, $i = 1, \ldots, k$, is what p-dimensional vector one should use as the standard for comparison. The choice can be either an a priori one, depending on what the data analyst supposes the intrinsic nature of the data to be, or a data-based estimate of a common underlying eigenvector. Let ${}_j\boldsymbol{\xi}$ denote the choice based on a priori specification and ${}_j\tilde{\mathbf{x}}$ denote a data-determined "typical" vector.

Given ${}_j\boldsymbol{\xi}$, the measure of dissimilarity between ${}_j\mathbf{x}_i$ and ${}_j\boldsymbol{\xi}$ proposed by Keramidas et al. (1987) is the Euclidean distance between the two points defined by the pair of vectors:

$${}_j\delta_i^2 = \min[({}_j\boldsymbol{\xi} - {}_j\mathbf{x}_i)'({}_j\boldsymbol{\xi} - {}_j\mathbf{x}_i), ({}_j\boldsymbol{\xi} + {}_j\mathbf{x}_i)'({}_j\boldsymbol{\xi} + {}_j\mathbf{x}_i)].$$

Analogously, the measure of dissimilarity between ${}_j\mathbf{x}_i$ and the data-based typical vector, ${}_j\tilde{\mathbf{x}}$ is denoted ${}_j\hat{\delta}_i^2$ and is obtained by using ${}_j\tilde{\mathbf{x}}$ in place of ${}_j\boldsymbol{\xi}$ in the above expression. The need to use the smaller of two values in the right-hand side of the expression is due to the fact that, with the usual normalization of eigenvectors to make them of unit length, the elements of the eigenvector are the coordinates of either of a pair of points constituting the ends of a diameter of a unit sphere. The squared distances, $\{{}_j\delta_i^2\}$ or $\{{}_j\hat{\delta}_i^2\}$, cannot exceed the value 2 since there is a relationship between them and the cosine of the angle between two vectors of unit length each, namely, $c^2 = a^2 + b^2 - 2ab\cos\theta = 2(1 - \cos\theta)$ for a triangle with sides of lengths a, b, and c and the included angle between the sides of lengths a and b is θ.

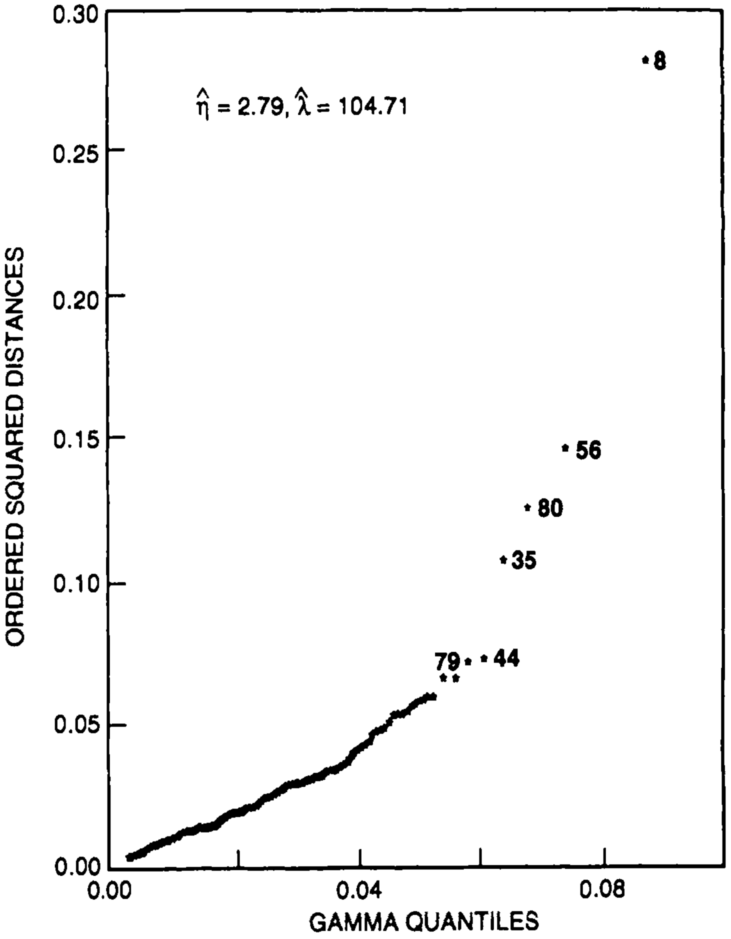
There remains the question of how to obtain a data-determined typical vector, ${}_j\tilde{\mathbf{x}}$, if one chooses to use that as the standard of comparison. Some obvious choices would be a location estimate calculated from $\{{}_j\mathbf{x}_i\}$, $i = 1, \ldots, k$, such as the mean or one of the many robust estimators described in Section 5.2.3. A different alternative would be to choose as the typical vector one that minimizes the angles between itself and the set $\{{}_j\mathbf{x}_i\}$. If the angle between a typical vector and ${}_j\mathbf{x}_i$ is denoted ${}_j\theta_i$, then an explicit criterion would be to choose that unit-length vector which minimizes $\Sigma_{i=1}^k \cos^2{}_j\theta_i$. The required vector turns out to be (see Keramidas et al., 1987) the eigenvector associated with the largest eigenvalue of the matrix, $\mathbf{E} = \Sigma_{i=1}^k {}_j\mathbf{x}_i\,{}_j\mathbf{x}_i'$. If one suspects that there are outliers in the data which might have, in turn, distorted the eigenvectors $\{{}_j\mathbf{x}_i\}$, the typical vector can be defined as the eigenvector corresponding to the largest eigenvalue of a robust version of $\mathbf{E}$ by iteratively trimming a fraction of the eigenvectors whose angles with the typical vector are among the largest. The spirit of the scheme here is similar to the ideas discussed in Section 5.2.3 for robust estimation of multivariate dispersion (see discussion associated with Eq. 74) but the criterion on which the trimming is based is different.

The graphical procedure for assessing the similarity of $\{{}_j\mathbf{x}_i\}$, $i = 1, \ldots, k$, is to make a gamma probability plot of the k ordered values of either ${}_j\delta_i^2$ (in the case of comparing against a prespecified standard), or ${}_j\hat{\delta}_i^2$ (when the standard is data determined), against the corresponding quantiles of a gamma distribution. The scale and shape parameters of the gamma distribution may be estimated using an order statistics formulation, exactly as was done in the case of the d_i's in Section 6.3.1 and of the a_i's and the g_i's in assessing the similarity in sizes of several covariance matrices described earlier in this section. Once again, to minimize the influence of possible "outliers" among the set of eigenvectors under comparison, the maximum likelihood estimation can be based on the smallest order statistics of the observed dissimilarity values.

Keramidas et al. (1987) illustrate their graphical method and its properties using both simulated and real data. The gamma probability plotting approach to assessing the similarities of sizes and orientations of covariance matrices may turn out to be more robust indicators than the more classical formal tests for equality of covariance matrices, which are known to be generally quite nonrobust in the sense that they tend to be more sensitive to nonnormality of data than to heteroscedasticity. Also, these techniques may be extensible in a straightforward manner to correlation matrices. Both the issue of robustness and the development of such extensions of these methods, however, need further study.

Example 49. To illustrate the gamma probability plotting technique for comparing eigenvectors, a real data example from Keramidas et al. (1987) is considered. The data resulted from student evaluations of instructors at a large university. Students rated the instructors on a seven-point scale, ranging from

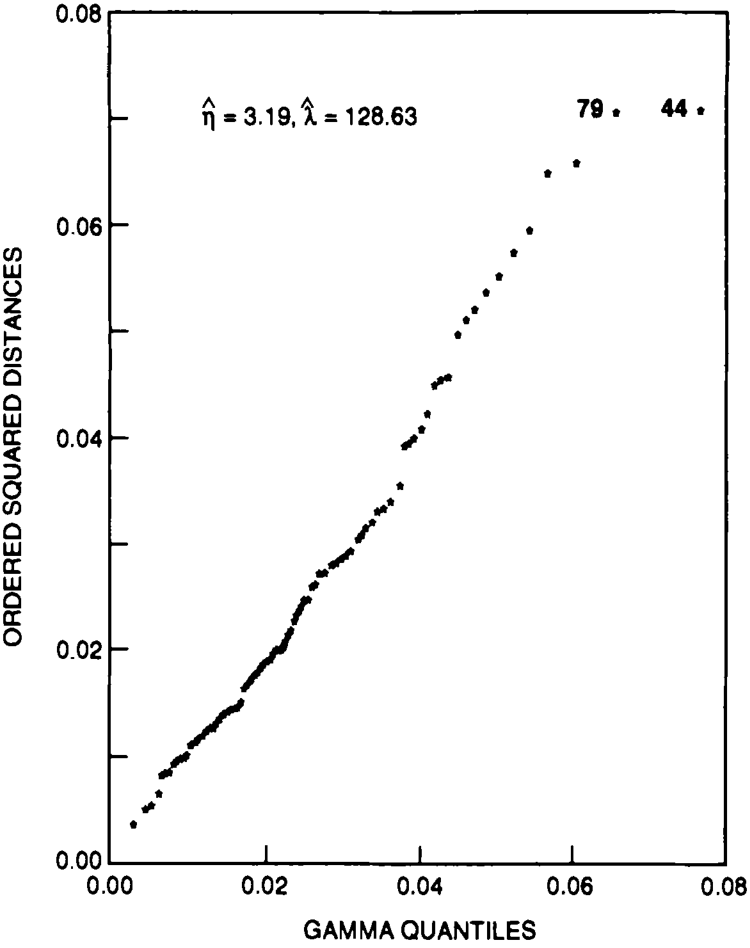
Exhibit 49a. Gamma $Q-Q$ plot for comparing the eigenvectors defining the first principal component with the data-determined typical vector



unsatisfactory to exceptional, for each of 18 $(=p)$ items, including such things as "suitability of the textbooks and/or materials," "fairness of grading," and "overall rating of the instructor." For preliminary analysis, only classes for which at least 75% of the enrolled students completed the questionnaire were included. Moreover, to help in distinguishing among the eigenvalues and to minimize difficulties of varying sample sizes, classes with less than 30 or more than 50 students were excluded. This left data from 117 $(=k)$ classes.

One question of interest in analyzing these data was if there was any dominant dimension (e.g., a prime principal component) of instructor evaluation underlying the data. To answer this question, 18 side-by-side box plots, each consisting of 117 values of an ordered eigenvalue, were made and the conclusion was that the largest eigenvalue was well separated from the remaining 17 smaller ones. Given this finding that the first principal compo-

Exhibit 49*b*. Replot of Exhibit 49*a* omitting classes 8, 56, 80, 35



nent was likely to be dominant, a natural second question in analyzing the data was if this index remains stable across the 117 classes. A third interesting question is if such a stable index gives equal weight to all 18 items so that a simple summary such as the student mean across all 18 items might suffice. The second question can be addressed by making a gamma probability plot of the dissimilarities between the eigenvectors defining the first principal component of the covariance matrix of each of the 117 classes and a data-determined typical value computed as described above (see discussion of Exhibit 49*a* below). The third question can be addressed either by comparing the typical vector with the 18-dimensional vector all of whose elements are equal to $1/\sqrt{18}$, or by making a gamma probability plot of the dissimilarities between the latter vector and the eigenvectors defining the first principal component of the covariance matrix of each of the 117 classes. Keramidas et al. (1987) using both approaches concluded that the index based on the typical vector is not very different from the simple equally-weighted index of all 18 items.

Exhibit 49*a* (see page 290) shows a gamma probability plot of the dissimilarities between the eigenvectors defining the first principal component of the covariance matrix of each of the 117 classs and the typical value. The maximum likelihood estimates, $\hat{\lambda}$ and $\hat{\eta}$, of the scale and shape parameters needed for this plot were obtained from the full set of 117 values thus ignoring the possible presence of "aberrant" eigenvectors.

Class #8 is clearly an "outlier," that is, its first principal component is quite different from a possibly common typical vector (and by implication from an equally weighted combination of the 18 items). Looking into what this class was, it was found that the course dealt with speech communications and was atypical in that it involved audio-visual aids for recording and then evaluating the students' talks, with the instructor acting primarily as a moderator. From Exhibit 49*a*, classes 56, 80 and 35 are also marginally suspect of being different. A comparison of the dissimilarities between the first principal components of the 117 classes and the vector assigning equal weights to the 18 items, carried out by Keramidas et al. (1987) by using a gamma probability plot, confirmed the same deviants but indicated that class 79 may also be different. Because of the consistency of the findings of these two analyses, they omitted classes 8, 35, 56 and 80 and replotted the remaining 113 dissimilarities using all of these to recalculate the maximum likelihood estimates of the scale and shape parameters of the gamma distribution. Exhibit 49*b* (see page 291) shows the resulting plot. The points form a strong linear pattern through the origin, suggesting that the first principal component for the 113 classes is essentially the same and, by implication from other analyses, they are basically an equally weighted average of the 18 items. In comparing the first principal component of class 8 with the vector of equal weights, it was found that in particular this class gave noticeably greater weight to three questions which asked the student to rate course materials, assignments and examinations. The first principal component for class 8 also gave noticeably smaller weights to four other questions that addressed the instructor's involvement in the class. Thus, an interpretation of the "aberrant" behavior of class 8 is that the speech communications course was different in that its focus was on special materials and class feedback, with only indirect participation of the instructor. None of the other 116 classes happened to have a comparable format.

6.4. MULTIDIMENSIONAL RESIDUALS AND METHODS FOR DETECTING MULTIVARIATE OUTLIERS

With large bodies of data, although models are appealing as parsimonious representations that may lead to simple interpretations of the data, it is very important to have means of gauging the appropriateness and sensitivities of the models under consideration. The useful role of residuals in exposing any inadequacies of a fitted model in the analysis of uniresponse problems has come to be widely recognized (see, for example, Terry, 1955; Anscombe, 1960, 1961; Draper & Smith, 1981).

One use of univariate residuals is to detect so-called outliers or extremely deviant observations, which are not uncommon in large data sets. Robust fitting of models or robust estimation (see Section 5.2.3) is one approach for handling outliers, namely, by minimizing the influence of such outliers on the fitted model. Often, however, pinpointing an outlier for further investigation and pursuit can be a valuable outcome of the statistical analysis of the data, and procedures directed specifically at detecting outliers can be useful (see, for example, Grubbs, 1950, 1969; Dixon, 1953; Barnett & Lewis, 1994).

The purpose of this section is to discuss multiresponse residuals and describe some techniques for identifying multivariate outliers. The discussion here draws on the work of Gnanadesikan & Kettenring (1972) and Devlin et al. (1975).

6.4.1. Analysis of Multidimensional Residuals

Given some summarizing fit to a body of multiresponse data, there exists, in principle, a vector of multivariate residuals between the data and the fit; but, more than in the univariate case, the important issue arises of how to express these multivariate residuals. Although experience is still rudimentary on these matters, some things can be done, and the discussion in this section will be concerned with some statistical methods for analyzing multivariate residuals.

For the discussion here and in Section 6.4.2, it is convenient to distinguish two broad categories of statistical analyses of multiresponse problems: (i) the analysis of internal structure, and (ii) the analysis of superimposed or extraneous structure (see also Section 3.1). The first category includes techniques, such as principal components, factor analysis, and multidimensional scaling (see Chapter 2), that are useful for studying internal dependencies and for reduction of the dimensionality of response. Multivariate multiple regression and multivariate analysis of variance (see Section 5.2.1), which are the classical techniques for investigating and specifying the dependence of multiresponse observations on design characteristics or extraneous independent variables, are examples of the second category.

Each category of analysis gives rise to multivariate residuals. For instance, as discussed in Chapter 2 (see pp. 56–57), linear principal components analysis may be viewed as fitting a set of mutually orthogonal hyperplanes by minimizing the sum of squares of orthogonal deviations of the observations from each plane in turn. At any stage, therefore, one has residuals that are perpendicular deviations of data from the fitted hyperplane. On the other hand, in analyzing superimposed structure (i.e., the second category above) by multivariate multiple regression, one has the well-known least squares residuals, namely (observations)–(predictions from a least squares fit). For purposes of data analysis it is often desirable to use the least squares residuals as input to a principal components analysis, which, in turn, will lead to the orthogonal residuals mentioned earlier. Augmenting multivariate multiple regression fitting by a principal components transformation of the residuals from fit may help in describing statistical correlations in the errors of the combined original variables, or in indicating inadequacies in the fit of the response variables by the design variables. For present purposes principal components residuals and least squares residuals are considered separately.

Principal Components Residuals. Equation 6 (in Chapter 2) defines the linear principal components transformation of the data in terms of the eigenvectors of the sample covariance matrix, S. Each row, $\mathbf{a}_j'$ $(j = 1, \ldots, p)$, of $\mathbf{A}'$ provides a principal component coordinate, and each row of $\mathbf{Z}$ gives the deviations of the projections of the original sample from the projection of the sample centroid, $\bar{\mathbf{y}}$, onto a specific principal component coordinate. Using standardized variables as the starting point would lead to corresponding interpretations of the principal components analysis of $\mathbf{R}$, the sample correlation matrix.

When the principal components analysis is viewed as a method of fitting linear subspaces, or as a statistical technique for detecting and describing possible linear singularities in the data, interest lies especially in the projections of the data onto the principal component coordinates corresponding to the small eigenvalues (i.e., the last few rows of $\mathbf{Z}$). Thus, for instance, with $p = 2$

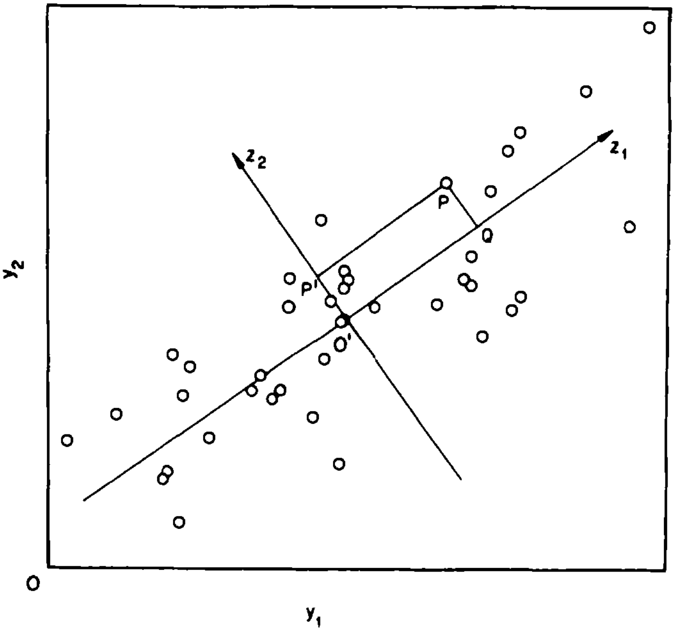


Fig. 17. Illustration of principal components residuals.

the essential concepts are illustrated in Figure 17, where y_1 and y_2 denote the original coordinates and z_1 and z_2 denote the two principal components derived from the covariance matrix of the bivariate data. The straight line of closest fit to the data (where closeness is measured by the sum of squares of perpendicular deviations) is the z_1-axis. The orthogonal residual of a typical data point, P, as shown in the figure, is the vector $\overline{QP}$, which is seen to be equivalent to the vector $\overline{O'P'}$, where P' is the projection of P onto the z_2-axis, the second principal component. More generally, with p-dimensional data, the projection onto the "smallest" principal component (i.e., the one with least variance) will be relevant for studying the deviation of an observation from a hyperplane of closest fit, while projections on the "smallest" q principal component coordinates will be relevant for studying the deviation of an observation from a fitted linear subspace of dimensionality $(p - q)$.

For detecting lack of fit of individual observations, one method suggested by Rao (1964) is to study the sum of squared lengths of the projections of the observations on the last few, say q, principal component coordinates. For each initial observation, $\mathbf{y}_i$ $(i = 1, \ldots, n)$, the procedure consists of computing

$$\begin{aligned} d_i^2 &= \sum_{j=p-q+1}^{p} [\mathbf{a}_j'(\mathbf{y}_i - \bar{\mathbf{y}})]^2 \\ &= (\mathbf{y}_i - \bar{\mathbf{y}})'(\mathbf{y}_i - \bar{\mathbf{y}}) - \sum_{j=1}^{p-q} [\mathbf{a}_j'(\mathbf{y}_i - \bar{\mathbf{y}})]^2, \end{aligned}$$

and considering inappropriately large values of d_i^2 as indicative of a poor $(p - q)$-dimensional fit to the observation (or, equivalently, that the observation is possibly an aberrant one). An informal graphical technique, which might have value as a tool for exposing other peculiarities of the data in addition to assessing the fit, would be to make a gamma probability plot of the d_i^2's, using an appropriately chosen or estimated shape parameter. One method of obtaining a suitable estimate of the shape parameter would be to base it on a collection of the smallest observed d_i^2's.

In addition to looking at a single summary statistic, such as d_i^2 above, it may often be useful to study the projections of the data on the last few principal component coordinates (i.e., the last few rows of $\mathbf{Z}$ in Eq. 6) in other ways. These might include the following:

1. Two- and three-dimensional scatter plots of bivariate and trivariate subsets of the last few rows of $\mathbf{Z}$ with points labeled in various ways, such as by time if it is a factor.

2. Probability plots of the values within each of the last few rows of $\mathbf{Z}$. Because of the linearity of the transformation involved, it may not be unreasonable to expect these values to be distributed more nearly normally than the original data, and normal probability plotting will provide a reasonable starting point for the analysis. This analysis may help in pinpointing

specific "smallest" principal component coordinates, if any, on which the projection of an observation may look abnormal, and thus may augment the earlier-mentioned gamma probability plotting analysis of the d_i^2's.

3. Plots of the values in each of the last few rows of **Z** against certain distances in the space of the first few principal components. If, for example, most of the variability of a set of five-dimensional data is associated with the first two principal components, it may be informative to plot the projections on each of the three remaining principal component axes against the distance from the centroid of each of the projected points in the two-dimensional plane associated with the two largest eigenvalues. This may show a certain kind of multidimensional inadequacy of fit—namely, if the magnitude of the residuals in the coordinates associated with the smaller eigenvalues is related to the clustering of the points in the two-dimensional space of the two eigenvectors corresponding to the largest two eigenvalues.

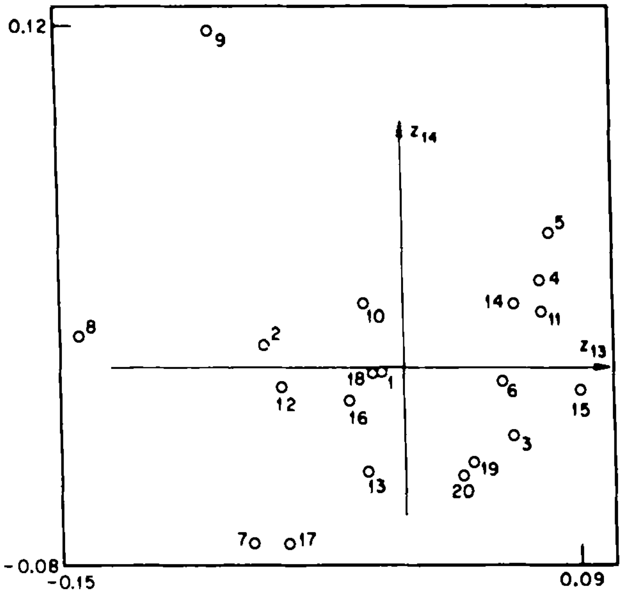
An important issue concerning the analyses suggested above is their robustness. Clearly, if an aberrant observation is detected, one may want to exclude it from the initial estimate of **S** (or **R**) and then repeat the process of obtaining and analyzing the principal components residuals. In some circumstances one may also decide to use a robust estimate of the covariance (or correlation) matrix, such as the ones considered in Section 5.2.3, even for the initial analysis, in the hope that the aberrant observations will become even more conspicuous in the subsequent analysis of residuals. (See Example 29.)

Example 50. To illustrate the use of some of the methods for analyzing principal components residuals, two sets of data are taken from a study by Chen et al. (1970, 1974) concerned with grouping corporations. (See also Examples 17, 18, 23, and 43.) As part of the study, the appropriateness of prespecified groupings (e.g., chemicals, oils, drugs) was examined initially, and, as mentioned in the discussion of Example 17, a preliminary attempt was made to develop core groups of companies from an internal analysis of each prespecified category. One approach for forming core groups was to identify and eliminate outliers by studying the principal components residuals.

There were 14 variables per company per year in the study. Specifically, for 1963 data were available for 20 drug companies, and Exhibit 50*a* shows a scatter plot of all the drug companies in the space of the last two principal components of the 14 × 14 correlation matrix derived from these data. Companies 8 and 9 are indicated as possible outliers with respect to the configuration of the remaining companies in this plot. Company 9 appears to be an outlier with respect to the last principal component in particular, while company 8 seems to be a moderate outlier on the penultimate principal component.

The second set of data is from 23 drug companies for the year 1967 and illustrates the use of probability plotting of the elements in each of the last few rows of **Z**. Exhibit 50*b* (see page 298) shows a normal probability plot of the

Exhibit 50a. Plot of 20 drug companies in the space of the last two principal components



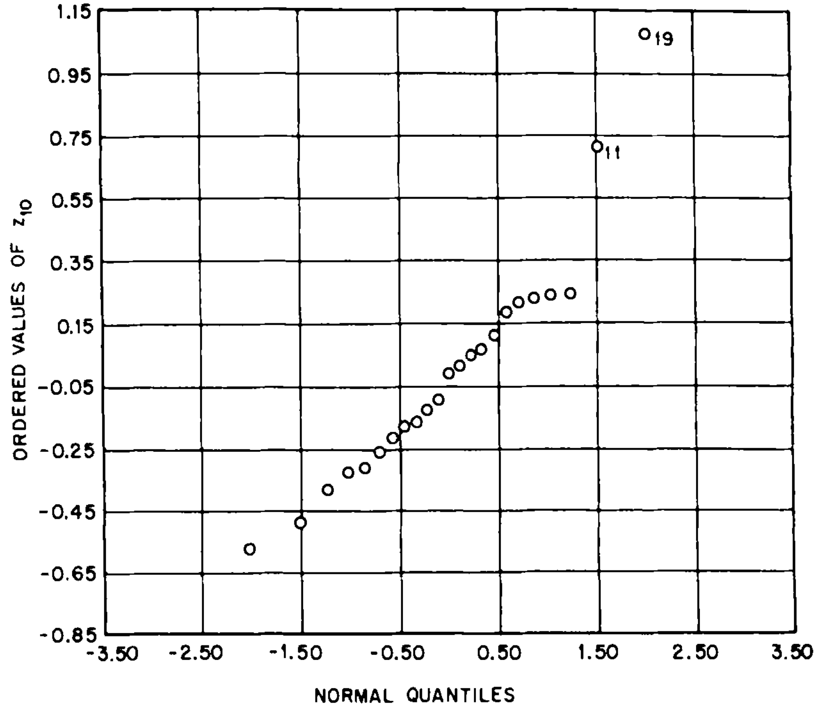
tenth principal component of the sample correlation matrix. The points corresponding to companies 11 and 19 are seen to deviate at the top right-hand end of the plot from the reasonably good linear configuration of the remaining points. The original data in this example exhibited considerable nonnormality (see Examples 37 and 43), and the earlier-mentioned aspect of improved normality induced by the principal components transformation is evident in Exhibit 50*b* by the linearity of the configuration of most of the points, with just a mild indication of a distribution with shorter tails than the normal.

Least Squares Residuals. In the notation employed earlier (see Eqs. 49, 50, and 63) in discussing the multivariate multiple regression model, the n multivariate least squares residuals (called residuals hereafter) are the p-dimensional rows (denoted as $\mathbf{e}_1', \mathbf{e}_2', \ldots, \mathbf{e}_n'$) of

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y}' - \mathbf{X}\hat{\boldsymbol{\Theta}}. \qquad (91)$$

Depending on the structure of $\mathbf{X}$, there will be certain singularities among the residuals in that certain linear combinations of the rows of $\hat{\boldsymbol{\varepsilon}}$ will be $\mathbf{0}'$. Depending on the correlational structure and functional dependencies among the p responses, there could be singularities in the other direction (viz., the columns of $\hat{\boldsymbol{\varepsilon}}$), and the existence and nature of such singularities may be investigated by principal components transformations of the p-dimensional residuals.

Exhibit 50*b*. Normal probability plot of the tenth principal component of the 14 × 14 correlation matrix for 23 drug companies



In some applications there may be a natural ordering among the responses, which may lead one to consider the use of a step-down analysis (see Section 4.c of Chapter IV in Roy et al., 1971). The analysis at each stage is a univariate analysis of a single response, utilizing all the responses that have been analyzed at the preceding stages as covariates. At each stage, therefore, step-down residuals may be obtained from this approach and studied by any of the available techniques for analyzing univariate least squares residuals.

Larsen & McClearly (1972) have proposed the concept of partial residuals and ways of using them. Entirely analogous definitions of multivariate partial residuals and methods of analyzing them may be suggested.

As a first approach to analyzing the residuals defined in Eq. 91, one may wish to consider the entire collection of them as an unstructured multivariate sample. Sometimes such a view may be more appropriate for subsets of the residuals than for the totality of them. For instance, in a two-way table the residuals within a particular row (or column) may be considered as an unstructured sample. At any rate, with such a view one can then employ methods applicable to the study of unstructured multivariate samples (see

Section 6.2), including the following:

1. Separate plotting of uniresponse residuals, perhaps against values of certain independent or extraneous variables (e.g., time) or against the predicted values. Augmenting such scatter plots with curves of locally smoothed quantiles (e.g., moving median and quartiles) can be very useful (see Cleveland & Kleiner, 1975).

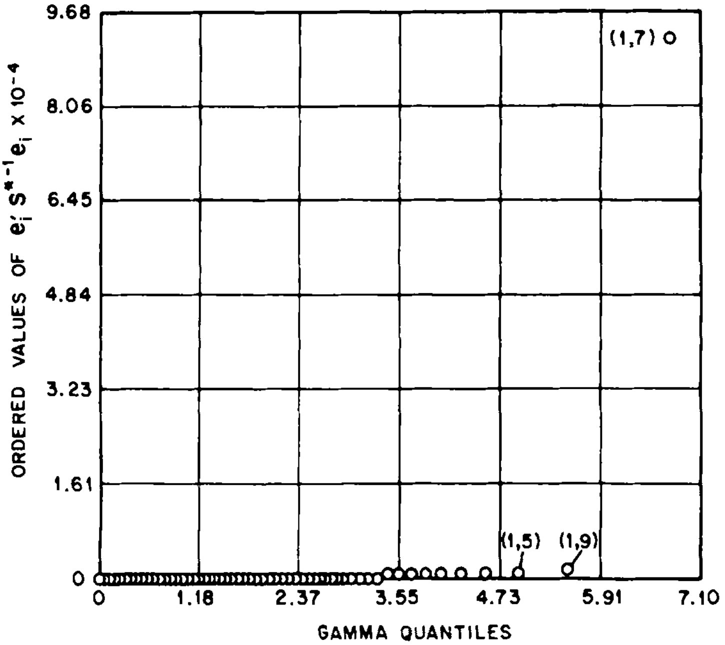
2. One-dimensional probability plotting of the uniresponse residuals. Full-normal plots of the uniresponse residuals or half-normal plots of their absolute values (or, equivalently, $\chi^2_{(1)}$ plots of squared residuals) provide natural starting points. Residuals generally seem to tend to be "supernormal" or at least more normally distributed than original data, and such probability plots may be useful in delineating outliers or other peculiarities in the data.

3. The use of one, or preferably several, distance functions to convert the multiresponse residuals to single numbers, followed by the probability plotting of these. The idea here is simply to treat the residuals (the **e**'s defined in Eq. 91) as single-degree-of-freedom vectors and to use the gamma probability plotting methodology described in Section 6.3.1 for analyzing them. The presence of outliers and of heteroscedasticity will be revealed by departures from linearity of the configuration on an appropriately chosen gamma probability plot of the values of a quadratic form in the **e**'s. For instance, an aberrant observation may be expected to yield a residual for which the associated quadratic form value will be unduly large, thus leading to a departure of the corresponding point from the linearity of the other points on the gamma probability plot (see Example 51). Heteroscedasticity will be indicated by a configuration that is piecewise linear, with the points corresponding to the residuals derived from observations with the same covariance structure belonging to the same linear piece.

The approach of gamma plotting quadratic forms of the residuals assumes a particularly simple, and already encountered, form for an unstructured sample. The residuals in this case are just deviations of the individual multiresponse observations from the sample mean vector, $\mathbf{e}_i = \mathbf{y}_i - \bar{\mathbf{y}}$ $(i = 1, \ldots, n)$. The study of the generalized squared distance of the observations from the sample mean (see Example 7 in Chapter 2 and the procedure for plotting radii described on pp. 197–200) is thus a special case. (See also Cox, 1968; Healy, 1968.)

Example 51. The data derive from an experiment on long-term aging of a transistor device used in submarine cable repeaters (see Abrahamson et al., 1969). Sets of 100 devices, in a configuration of 10 rows by 10 columns, were aged, and a characteristic called the *gain* of each device was obtained at each

Exhibit 51a. Gamma probability plot derived from three-dimensional residuals scaled by a robust covariance matrix



of several test periods. An initial transformation to logarithms was made, and the aging phenomenon of interest was then the behavior of the log gain as a function of time. One approach to studying the aging behavior for purposes of identifying devices with peculiar aging characteristics was to fit a polynomial (specifically, a cubic was used) to the data on log gain versus time for each device, and to study the fitted coefficients by analysis of variance techniques. A separate univariate analysis of variance of each coefficient, as well as a multivariate analysis of variance of the four coefficients simultaneously, was performed. The multivariate approach was employed partly because of the high intercorrelations observed among the fitted coefficients. It was not used as a substitute for the separate univariate analyses of the individual coefficients. For present purposes attention is confined to the multivariate approach.

A simple one-way (i.e., rows and columns-within-rows were the sources of variation) multivariate analysis of variance (MANOVA), when used as a means for obtaining formal tests of hypotheses, revealed very little. None of the usual MANOVA tests of the null hypothesis of no row effects (see Section 5.2.1 and also Chapter IV of Roy et al., 1971) had an associted p-value smaller than 0.3. The danger in basing an analysis solely on such tests, which are based on single summary statistics, is revealed by the use of the informal gamma plotting technique described above.

Exhibit 51*b*. Replot obtained from Exhibit 51*a* after omitting point (1,7)

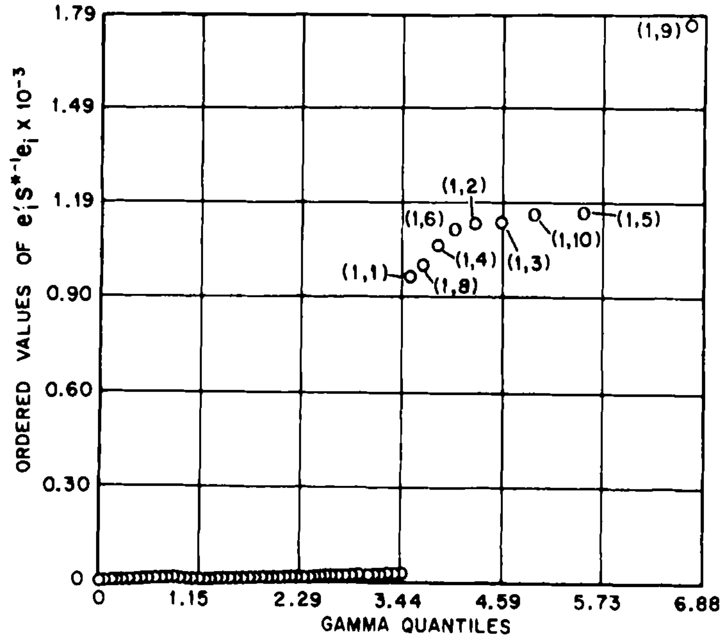


Exhibit 51*a* shows a gamma probability plot of the 100 values of a quadratic form, $e_i'S^{*-1}e_i$ $(i = 1, \ldots, 100)$, in the four-dimensional residuals. The covariance matrix, S^*, of the residuals is a robust estimate (of the type discussed in Section 5.2.3) obtained from the residuals themselves. [*Note*: Since S^{*-1} is common to all 100 values of the quadratic form being analyzed, it is not necessary to multiply S^* by the "unbiasing" constant for the present application.] The shape parameter required for the plot was estimated by maximum likelihood based on the 50 smaller values of the quadratic form, considered as the 50 smallest order statistics in a random sample of size 100.

The point that stands out clearly from the configuration of the others in Exhibit 51*a* corresponds to the seventh device in the first row, and the implication is that the four-dimensional residual for this device is inordinately "large," that is, a possibly aberrant observation has been pinpointed! This residual (and other such if they exist) has, of course, contributed to the estimate of the columns-within-rows dispersion matrix that was employed as the error dispersion matrix in the formal tests of significance mentioned earlier. The effect would be to inflate the error dispersion inappropriately, and it is not surprising, therefore, that the tests revealed no significant departures from the null hypothesis. Upon verification, the aging configuration of device 7 in row 1 was found to be indeed abnormal in relation to the behavior of the majority of devices.

To facilitate further study of the residuals, a replot, shown in Exhibit 51*b* (see page 301) may be made of the 99 points left after omitting the point corresponding to the aberrant device. The configuration on this plot may lead one to conclude that device 9 (the one from which the top right-hand corner point derives) and also the other devices (1–6 and 8) in row 1 are suspect, that is, all 10 devices in row 1 are associated with peculiar residuals. Such a conclusion, however, may not be warranted, and the discussion that follows will clarify the issue involved. The analysis of the data is then continued in Example 52.

When the matrix **X** in Eq. 49, the so-called design matrix or matrix of values of the regressors, corresponds to more structured situations (e.g., a multiway classification), there are at least two sources of statistical difficulty in analyzing the residuals. First, there are constraints on subsets of the residuals (e.g., the sum of the residuals in a row of a two-way table is the null vector), which imply correlations among the residuals. Second, the presence of outliers may seriously bias the usual effects which are subtracted from an observation (e.g., row, column, and overall mean vectors in a two-way classification) so as to mask the local effect of an outlier on the corresponding residual. The first source of difficulty (viz., the singularities among residuals) is especially critical when the numbers of levels of the factors involved (e.g., the number of rows or columns in a two-way table) are small, but the second source can be important even when each of the factors has a moderate number of levels.

Thus in Example 51 the extreme outlier (viz., the observation for device 7 in row 1) may have so badly biased the mean vector for the first row that all the residuals [=(observation vector)–(row mean vector)] in that row have been unduly biased. If the outler is extreme enough, this can indeed happen, and a method is needed for insuring against such masking effects of the outliers on the residuals.

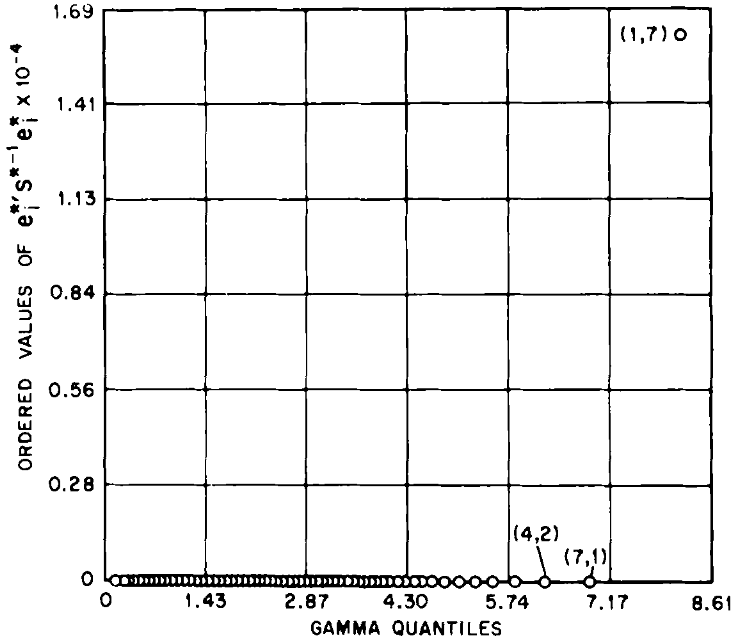
One way of accomplishing this is to combine the ideas and methods of robust estimation discussed in Section 5.2.3 with the desirability of analyzing the residuals. Specifically, instead of using the usual least squares estimates of the elements of $\boldsymbol{\Theta}$ in the linear model (Eq. 49), one could use robust estimates of them, thus obtaining $\hat{\boldsymbol{\Theta}}^*$, and then define a set of *robustified residuals* (see also the discussion at the end of Section 5.2.3) as the rows of

$$\hat{\boldsymbol{\varepsilon}}^* = \mathbf{Y}' - \mathbf{X}\hat{\boldsymbol{\Theta}}^*. \tag{92}$$

If one were to utilize the simplest direct approach to developing $\hat{\boldsymbol{\Theta}}^*$, which was described toward the end of Section 5.2.3, $\hat{\boldsymbol{\Theta}}^*$ would just be a matrix each of whose elements, $\hat{\theta}^*_{ij}$, is a uniresponse robust estimator of a univariate location-type parameter.

Example 52. To illustrate the use of robustified residuals, the data used in Example 51 are employed again. Instead of using the row mean vectors for defining the residuals, the vector of midmeans, $\mathbf{y}^*_{T(.25)}$, discussed in Section 5.2.3, for each row is used, and the robustified four-dimensional residuals are

Exhibit 52*a*. Gamma probability plot derived from three-dimensional robust residuals scaled by a robust covariance matrix

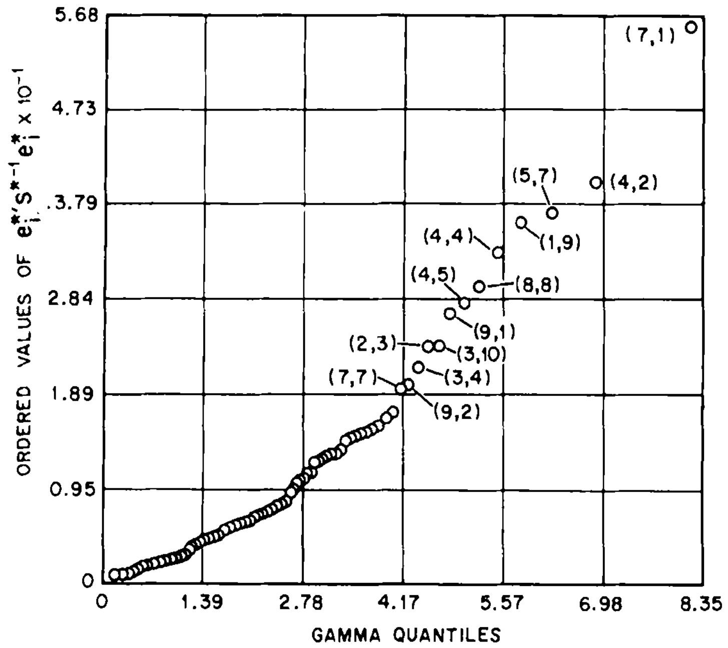


obtained as the difference between the four-dimensional observation (viz., the four coefficients of the aging curve for a device) and the vector of midmeans for the row in which the observation appears.

The 100 four-dimensional robustified residuals thus obtained in this example can then be analyzed by the gamma probability plotting technique described and illustrated earlier in the context of analyzing the regular residuals. Exhibit 52*a* shows a gamma probability plot of the 100 values of a quadratic form in the modified residuals, $e_i^{*\prime}S^{*-1}e_i^*$ ($i = 1, \ldots, 100$), where S^* as before is a robust estimate of the covariance matrix, and the shape parameter required for the plot is estimated once again using the smallest 50 observed values of the quadratic form. In Exhibit 52*a* the point corresponding to device 7 in row 1 again stands out, and Exhibit 52*b* (see page 304) shows a replot obtained after omitting this point. Comparing Exhibits 52*b* and 51*b*, it is seen that the biasing effect on all the residuals in the first row caused by the extremely deviant observation for device 7 in that row is no longer evident. The configuration in Exhibit 52*b* may be used to delineate additional outliers, such as device 1 in row 7, by looking for points in the top right-hand corner that deviate noticeably from the linear configuration of the points in the lower left-hand portion of the picture.

Whether $\hat{\varepsilon}^*$ as defined in Eq. 92 is the most appropriate set of robust residuals for purposes of analysis, or whether one needs to modify them (e.g.,

Exhibit 52*b*. Replot obtained from Exhibit 52*a* after omitting point (1,7)



by weighting them), is a question for further investigation. They do at least constitute a simple starting point. The robustified residuals defined by Eq. 92 will not necessarily satisfy the constraints satisfied by the usual residuals. For example, in a two-way classification they will not necessarily add up to the null vector, either by rows or by columns, or even across all cells. The robustified residuals do not form a cohesive group unless there are no outliers in the data, and in the latter case the usual least squares estimator, $\hat{\Theta}$, and the robust estimator, $\hat{\Theta}^*$, will not be very different, so that the usual residuals, $\hat{\varepsilon}$, and the robustified residuals, $\hat{\varepsilon}^*$, will also be expected to be very similar when there are no outliers. The main use of the robustified residuals is, in fact, to accentuate the presence of outliers, and hence the fact that they do not satisfy the same constraints as the usual residuals is perhaps unimportant. If, however, one desires to have modified residuals satisfy these constraints as nearly as possible, then iterating the analysis in certain ways may help. Tukey (1970) has suggested such a scheme for using midmeans in analyzing multiway tables with uniresponse data, and an extension of this approach to the multiresponse case may be feasible. For data-analytic purposes, robustified residuals are useful because of their ability to "localize" the effects of outliers. This illustrates the importance of such residuals for diagnostics and demonstrates the value of robust estimation for both summarization and exposure.

6.4.2. Other Methods for Detecting Multivariate Outliers

In the preceding section ways of pinpointing maverick observations through an analysis of multivariate residuals were discussed. In this section some additional techniques are suggested for detecting multivariate outliers.

The consequences of having defective responses are intrinsically more complex in a multivariate sample than in the much-discussed univariate case. One reason is that a multivariate outlier can distort not only measures of location and scale but also those of orientation (i.e., correlation). A second reason is that it is much more difficult to characterize a multivariate outlier. A single univariate outlier may typically be thought of as "the one that sticks out on the end," but no such simple concept suffices in higher dimensions. A third reason is the variety of types of multivariate outliers that may arise: a vector response may be faulty because of a gross error in one of its components or because of systematic mild errors in all of its components.

The complexity of the multivariate case suggests that it would be fruitless to search for a truly omnibus outlier detection procedure. A more reasonable approach seems to be to tailor detection procedures to protect against specific types of situations, for example, correlation distortion, thus building up an arsenal of techniques with different sensitivities. This approach recognizes that an outlier for one purpose may not necessarily be one for another purpose! However, if several analyses are to be performed on the same sample, the result of selective segregation of outliers can be a more efficient and effective use of the available data.

It is essential that the procedures be computationally inexpensive enough to allow for routine screening of large data sets. Those that can simultaneously expose other features of the data, such as distributional peculiarities, have added economic appeal.

Following the dichotomy of multivariate methods mentioned in Section 6.4.1, the proposed procedures will be presented under the general headings of internal and external analysis techniques. In the former category are the techniques, such as principal components analysis, that are appropriate for examining an unstructured sample of data; in the latter category are techniques, such as canonical correlation analysis, that are applicable in the presence of some superimposed structure.

An approach that can lead to outlier-detection methods for both categories of problems is one which exploits the feature that outliers tend to have an unduly large or distorting influence on summary statistics. Gnanadesikan & Kettenring (1972) propose a variety of statistics, addressed to different multivariate problems, for assessing the influence of each observation on several standard multiresponse analyses. A few of these will be described later in this section.

The influence function, advocated by Hampel (1968, 1973, 1974), is a useful device for considering the effect of observations on a statistic. As such it can be useful not only as a tool for motivating and designing specific types of robust estimators but also as a means for developing methods for outlier detection (see Devlin et al., 1975).

For a general parameter $\theta = T(F)$, expressed as a functional of the distribution function, F, the influence function $I(\mathbf{y}; \theta)$ at $\mathbf{y}$ is defined (see Hampel, 1974; Hampel et al., 1986) as

$$I(\mathbf{y}; \theta) = \lim_{\varepsilon \to 0} \left(\frac{\tilde{\theta} - \theta}{\varepsilon} \right),$$

where $\tilde{\theta} = T(\tilde{F})$ and $\tilde{F} = (1 - \varepsilon)F + \varepsilon\delta_{\mathbf{y}}$ is a "perturbation" of F by $\delta_{\mathbf{y}}$, the distribution function for a point mass of 1 at $\mathbf{y}$. The essential concept in this "theoretical" influence function is that one can use it to assess the influence of the point $\mathbf{y}$ on the parameter θ.

Three finite sample versions of the influence function may be distinguished. The first, termed the *empiric* influence function by Mallows (1973), is obtained by replacing F in the above definition by the empirical cumulative distribution function, F_n, which is a step-function with a step of height $1/n$ at each of the observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$.

In the second finite sample version, the desire is to study the difference between $\hat{\theta}$ (an estimator of θ obtained from the n observations in the sample) and $\hat{\theta}_+$, an estimator of the same form as $\hat{\theta}$ obtained from the n original observations plus a conceptualized additional observation, $\mathbf{y}$. Specifically, this version of the influence function is defined as

$$I_+(\mathbf{y}; \hat{\theta}) = (n + 1)(\hat{\theta}_+ - \hat{\theta}),$$

and it is essentially the so-called sensitivity curve used by Andrews et al. (1972) for studying the properties of various robust estimates of location.

The third version focuses on the individual effects of the actual observations in the sample and is particularly suited to assessing the influence of individual observations on the estimator $\hat{\theta}$. This version, called the *sample* influence function by Devlin et al. (1975), is defined as

$$I_-(\mathbf{y}_i; \hat{\theta}) = (n - 1)(\hat{\theta} - \hat{\theta}_{-i}), \qquad i = 1, \ldots, n, \tag{93}$$

where $\hat{\theta}_{-i}$ is an estimator of the same form as $\hat{\theta}$ but is calculated by omitting the ith observation, $\mathbf{y}_i$. The quantity $(\hat{\theta} + I_-)$ is the ith pseudo-value in Tukey's (1958) jackknife technique (see also Miller, 1974, and references therein).

[*Note*: Both I_+ and I_- can be considered as approximations to the empiric influence function by taking ε in the latter to be $1/n + 1$ and $-1/n - 1$, respectively (see Mallows, 1973).]

Hampel (1968) discusses the use of the influence function in the contexts of estimating univariate location and scale. Devlin et al. (1975) describe its application in the context of bivariate correlation. Specifically, it can be established that the influence function of ρ, the population product moment correlation coefficient, for any bivariate distribution for which ρ is defined (viz., second moments are finite) is

$$I(y_1, y_2; \rho) = -\tfrac{1}{2}\rho(\tilde{y}_1^2 + \tilde{y}_2^2) + \tilde{y}_1\tilde{y}_2,$$

where $\tilde{y}_j$ is the standardized form of y_j [i.e., $\tilde{y}_j = (y_j - \mu_j)/\sqrt{\sigma_{jj}}$, $j = 1, 2$]. Furthermore, if z_1 and z_2 denote, respectively, the standardized sum of and difference between $\tilde{y}_1$ and $\tilde{y}_2$, and if $u_1 = (z_1 + z_2)/\sqrt{2}$, $u_2 = (z_1 - z_2)/\sqrt{2}$, the above equation may be rewritten as

$$I(y_1, y_2; \rho) = (1 - \rho^2)u_1u_2.$$

Also, the influence function of $z(\rho) = \tanh^{-1}\rho$, Fisher's z-transform of ρ, may be shown to be free of ρ:

$$I(y_1, y_2; z(\rho)) = u_1u_2,$$

where u_1 and u_2 are as above. With the additional assumption that (y_1, y_2) has a bivariate normal distribution, it follows that the influence function of $z(\rho)$ has a *psn* (product of two independent standard normal variables) distribution.

The analogous sample influence function of r, the sample product moment correlation coefficient, is

$$I_{-}(y_{i1}, y_{i2}; r) = (n - 1)(r - r_{-i}) \approx (1 - r^2)u_{i1}u_{i2}, \qquad i = 1, \ldots, n, \qquad (94)$$

wherein the first equality follows from the definition of I_{-} given in Eq. 93, and the expression on the right is the value of the empiric influence function at the ith observation, $\mathbf{y}_i' = (y_{i1}, y_{i2})$. The quantity r_{-i} in Eq. 94 denotes the correlation coefficient based on all but the ith observation, and u_{i1}, u_{i2} are sample analogues of u_1 and u_2:

$$u_{i1} = \frac{\sqrt{n}}{2}\left(\frac{d_{i1} + d_{i2}}{\sqrt{1 + r}} + \frac{d_{i1} - d_{i2}}{\sqrt{1 - r}}\right),$$

$$u_{i2} = \frac{\sqrt{n}}{2}\left(\frac{d_{i1} + d_{i2}}{\sqrt{1 + r}} + \frac{d_{i1} - d_{i2}}{\sqrt{1 - r}}\right),$$

where

$$d_{ij} = \frac{y_{ij} - \bar{y}_j}{\sqrt{a_{jj}}}, \qquad \bar{y}_j = \frac{\sum_{i=1}^{n} y_{ij}}{n}, \qquad a_{jj} = \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2.$$

For the Fisher transform, $z(r)$, the analogous approximate result is

$$I_{-}(y_{i1}, y_{i2}; z(r)) = (n-1)[z(r) - z(r_{-i})] \approx u_{i1}u_{i2}, \tag{95}$$

which is free of r. From the forms of Eqs. 94 and 95 it follows that the contours of the sample influence functions of both r and $z(r)$ may be approximated by hyperbolas with axes oriented along the principal axes of the sample correlation matrix. Also, from Eq. 95 it follows that for a reasonably large sample of bivariate normal data one can approximate the distribution of $I_{-}(y_{i1}, y_{i2}; z(r))$ by a psn distribution. These properties of the sample influence function are used below to develop informal graphical tools for detecting bivariate observations that may unduly distort r.

Internal Analysis Techniques for Outlier Detection. A basic and widely used approach to displaying multiresponse data is through two- and three-dimensional scatter plots of the original and the principal component variables. Of the principal components the first and last few are usually of greatest interest to study. The first few principal components are especially sensitive to outliers which are inappropriately inflating variances and covariances (if one is working with **S**) or correlations (if one is working with **R**). Motivation, in terms of residuals, for looking at the last few principal components was discussed in Section 6.4.1. The kind of outlier which can be detected along these axes is one that is adding unimportant dimensions to, or obscuring singularities in, the data.

Probability plots (e.g., normal plots) and standard univariate outlier tests (such as those due to Grubbs, 1950, 1969, and to Dixon, 1953) may be carried out on each row of the observation matrix, **Y**, or of the derived principal components, **Z** (see Eq. 6 in Chapter 2). Outliers that distort location, scale, and correlation estimates may be uncovered in this manner.

Two graphical methods, specifically addressed to detecting observations that may have a distorting influence on the correlation coefficient, r, have been proposed by Devlin et al. (1975). The first of these is to augment a simple x–y scatter plot of the data with contours of the sample influence function of r (see Eq. 94) so as to facilitate the assessment of the effect of individual bivariate observations on the value of r. Thus, treating the approximation to $I_{-}(y_{i1}, y_{i2}; r)$ in Eq. 94 as a function of two variables, y_1 and y_2, one superimposes selected contours (which will be hyperbolas) of the function directly onto the scatter plot. The contour levels chosen for display purposes will depend on the sample size and other considerations relevant to the particular application.

The second proposal is to make a suitable Q-Q probability plot of the n values, $I_{-}(y_{i1}, y_{i2}; z(r))$, for $i = 1, \ldots, n$, utilizing the approximation in Eq. 95. From a data-analytic viewpoint, it is appropriate to use distributional assumptions (such as normality) for developing methodology as long as the effects of departures from such assumptions are themselves assessable in specific applications. With this in mind, it is proposed that bivariate normality of the data be assumed as a null background and a *Q-Q* plot be made of the n ordered values of the sample influence function of $z(r)$ against the corresponding quantiles of the distribution of the product of two independent standard normal deviates (see Eq. 95 and the discussion following it). Despite the facts that these sample influence function values, by definition, are just $(n-1)$ times the differences between $z(r)$ and $z(r_{-i})$, and that one would intuitively expect the appropriate null distribution to be normal [since $z(r)$ and $z(r_{-i})$ are themselves approximately normal], algebraic manipulations of the sample influence function reveal that the relevant distribution is psn rather than normal. In fact, initially Gnanadesikan & Kettenring (1972) proposed a normal probability plot for the $z(r_{-i})$ values, only to discover later that the more appropriate procedure would be to utilize a psn distribution, which is also symmetric but has much thicker tails than the normal.

However, there is an "equivalent" normal probability plot that can be made in place of the psn plot. This may be preferred for purposes of interpretation because of the greater familiarity of normal plots for many people. The idea follows from recognizing that the psn distribution is parameter free, so that one can transform the n sample influence function values involved to their equivalent standard normal deviates and then make a normal plot of these transformed quantities. Thus, if $i_{(1)} \leqslant i_{(2)} \leqslant \cdots \leqslant i_{(n)}$ denote the n ordered sample influence function values of $z(r)$ and if G denotes the distribution function of the psn distribution, the transformed values needed for the normal plot are the v's defined by

$$\Phi(v_{(l)}) = G(i_{(l)}), \qquad l = 1, \ldots, n,$$

where Φ is the distribution function of the standard normal distribution.

The configuration of the psn probability plot of the $i_{(l)}$, or the normal plot of the $v_{(l)}$, may be used for checking on possible departures from the assumed null conditions, such as the presence of outliers that distort r, or smoother departures of the data distribution from bivariate normality. For instance, if most of the data are reasonably well behaved, with the exception of a few outliers that have disproportionate effects on r, one would expect most of the points on the *Q-Q* plot to conform to a linear configuration, while the points that derive from omission of the outlying observations will depart from such a linear configuration by being either "too big" or "too small." On the other hand, if the entire data distribution is distinctly nonnormal, one will expect to see departures from linearity in most regions of the plot (see Examples 53 and 54).

The differences, $(r - r_{-i})$ and $[z(r) - z(r_{-i})]$, are two examples of unidimensional statistics that can aid in pinpointing the effects of individual observations on familiar summary statistics. A variety of others, including ones for detecting observations that distort eigenanalyses such as principal components analysis, are described by Gnanadesikan & Kettenring (1972). Also, Wilks

(1963) proposed that a test for a single outlier in an unstructured sample be based on the statistic

$$w = \max_i \left\{ \frac{|\mathbf{A}_{-i}|}{|\mathbf{A}|} \right\},$$

where $\mathbf{A} = (n-1)\mathbf{S}$ denotes the sum-of-products matrix based on all n observations in the sample, and $\mathbf{A}_{-i}$ again denotes a matrix computed just like $\mathbf{A}$ but without the ith observation. The statistic w turns out to be equivalent to the maximum observed generalized squared distance in the sample, that is, $\max_i (\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{S}^{-1}(\mathbf{y}_i - \bar{\mathbf{y}})$, thus establishing a connection with the procedure of studying these generalized squared distances as described in Section 6.4.1 in the context of least squares residuals. Focusing on the largest observed generalized squared distance would be natural for developing a formal single-statistic test, but studying a gamma probability plot of the collection of n generalized squared distances may be more revealing for data-analytic purposes. For carrying out the formal test, the work of Siotani (1959) on the asymptotic distribution of the maximum generalized squared distance in multivariate normal samples provides some useful results and tables of percentage points.

The cluster analysis techniques discussed in Chapter 4 provide a different type of tool for identifying outliers. If the outliers constitute a distinct group of observations that are far removed from the majority of the data, one would expect them to be delineated as a cluster of observations. For instance, in a hierarchical clustering scheme, if one uses interpoint distances [i.e., $(\mathbf{y}_i - \mathbf{y}_k)'\mathbf{S}^{-1}(\mathbf{y}_i - \mathbf{y}_k)$] as the input, the expectation is that outlier clusters, if any exist, will join the main body of points near or at the final level of clustering (see Example 56).

Example 53. The computer-generated data shown in Exhibit 53*a* are a sample of 60 observations, 58 of which are from a bivariate normal distribution with $\rho = 0.9$ and the remaining 2 observations simulate moderate outliers with opposite (i.e., inflation vs. deflation) effects on r. Also shown on the scatter plot are selected contours of the sample influence function of r.

The interpretation of the contours would be that the point labeled 1 would increase the value of r by about 0.2, while observation 2 would decrease r by about the same amount. These separate inferences concerning the two outliers are reasonably accurate in these data since the actual computed values are $r = 0.026$, $r_{-1} = -0.201$, and $r_{-2} = 0.253$. The value of the correlation coefficient when both outliers are omitted is, however, 0.029.

Exhibit 53*b* shows a psn probability plot of the sample influence function values of $z(r)$ for the same data, and Exhibit 53*c* (see page 312) shows the equivalent normal plot of the associated transformed quantities. Although the outliers stand out clearly on both these plots, the middle of the configuration (i.e., the linear part) in Exhibit 53*c* is stretched out more than the corresponding part of Exhibit 53*b*.

Exhibit 53a. Scatter plot with influence function contours for sample of bivariate normal data with two outliers added; $n = 60$, $\rho = 0$, $r = 0.026$

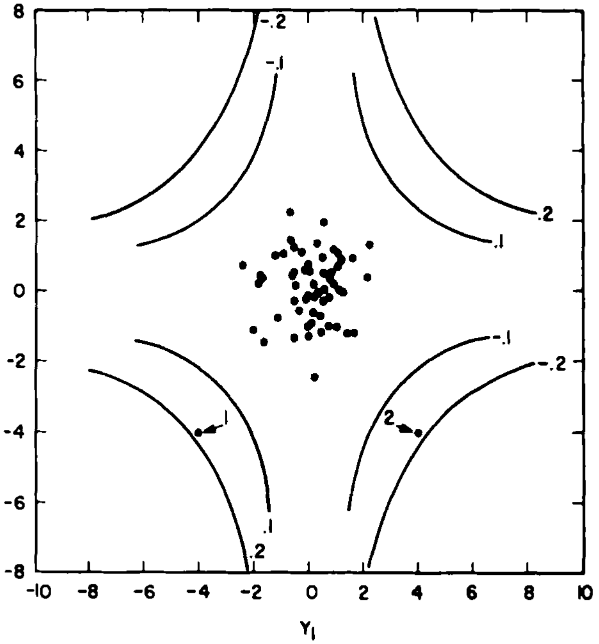


Exhibit 53b. PSN probability plot of values of z_{-i} for data of Exhibit 53a

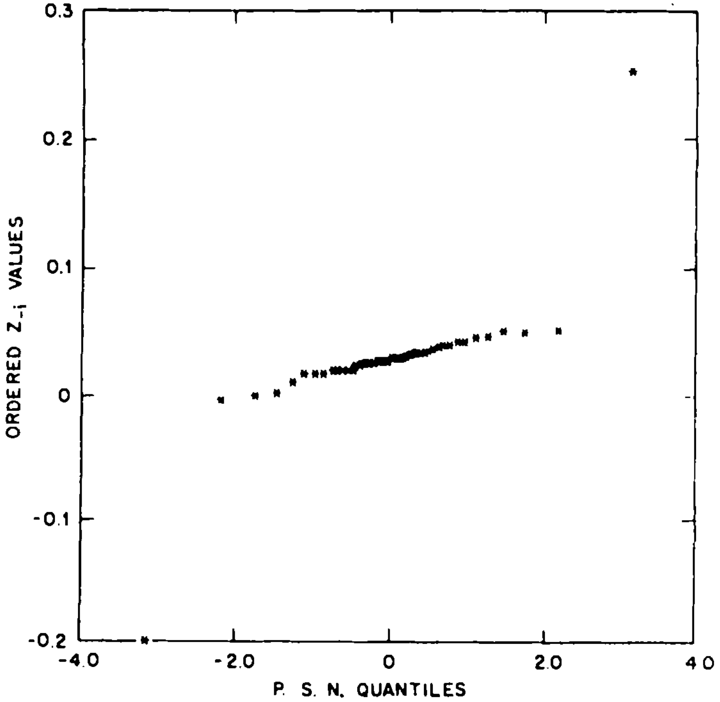
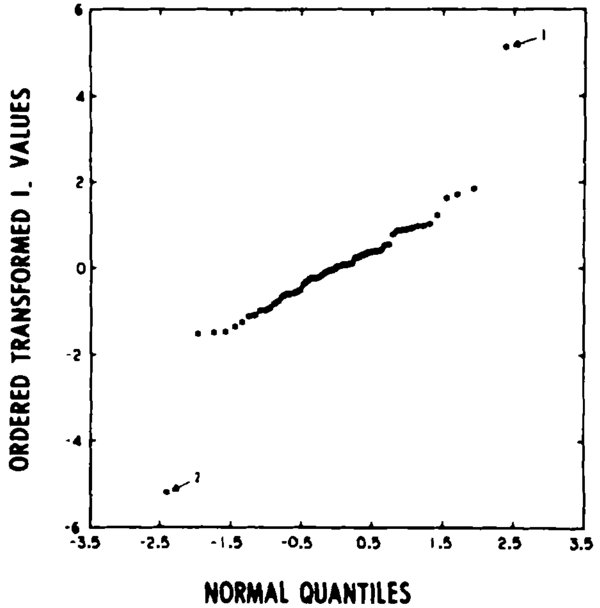


Exhibit 53c. Normal probability plot of transformed influence function values for data of Exhibit 53*a*



Example 54. The iris data (Anderson, 1935; Fisher, 1936) employed in Example 42 are used here to illustrate the issues and methods of outlier detection. Specifically, Exhibit 54*a* shows a scatter plot of values of the natural logarithms of 10 times the sepal lengths and widths for the 50 specimens of *Iris setosa*. Also shown in Exhibit 54*a* are contours of the sample influence function of r. Although the 42nd observations stands out clearly from the rest of the data, its location with respect to the contours suggests that it does not have a distorting influence on the value of r. Indeed, $r_{-42} = 0.723$, whereas $r = 0.730$.

Exhibit 54*b* which shows a normal plot of the transformed sample influence function values of $z(r)$, provides further confirmation of this. There ae 11 points with more extreme influence on r than the 42nd observation. The most striking feature of Exhibit 54*b*, however, is the nonlinearity of the configuration even in the middle region, indicating that the logarithmically transformed data may be quite nonnormal, at least with respect to the two variables considered here.

Exhibit 54*c* (see page 314) shows a $\chi^2_{(2)}$ probability plot of the 50 generalized squared distances in these data. As expected from the scatter plot (Exhibit 54*a*), in terms of the elliptical distance measured by the generalized squared distance, the 42nd observation is indeed an outlier (see also Example 42). This example thus illustrates the point that a multiresponse observation that is judged to be an outlier for one purpose may be quite a reasonable observation for other purposes.

Exhibit 54a. Scatter plot with influence function contours for natural logarithms of sepal length and width of 50 *Iris setosa*

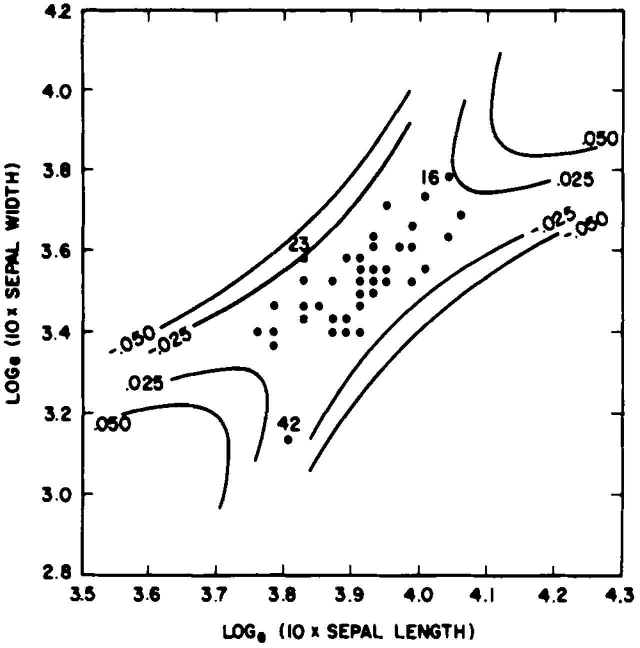


Exhibit 53b. PSN probability plot of values of z_{-i} for data of Exhibit 53a

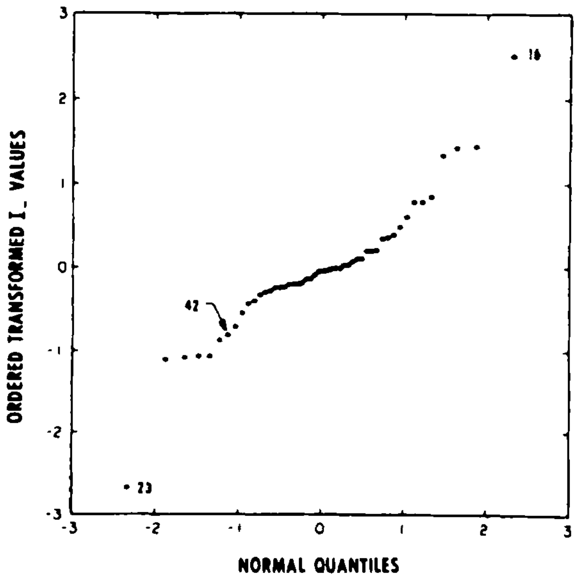
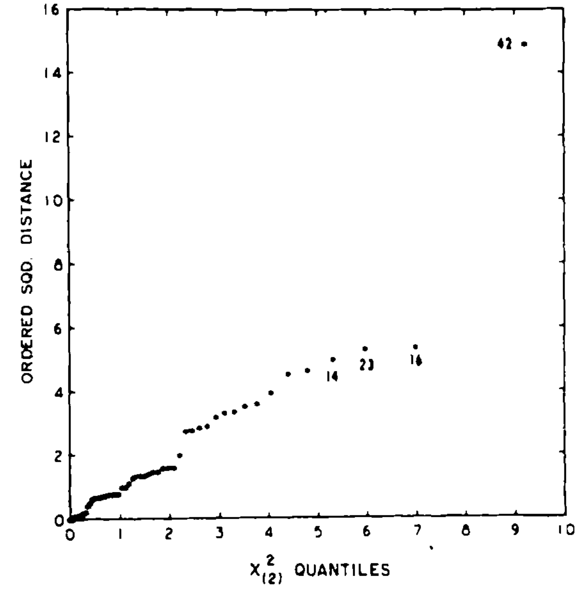


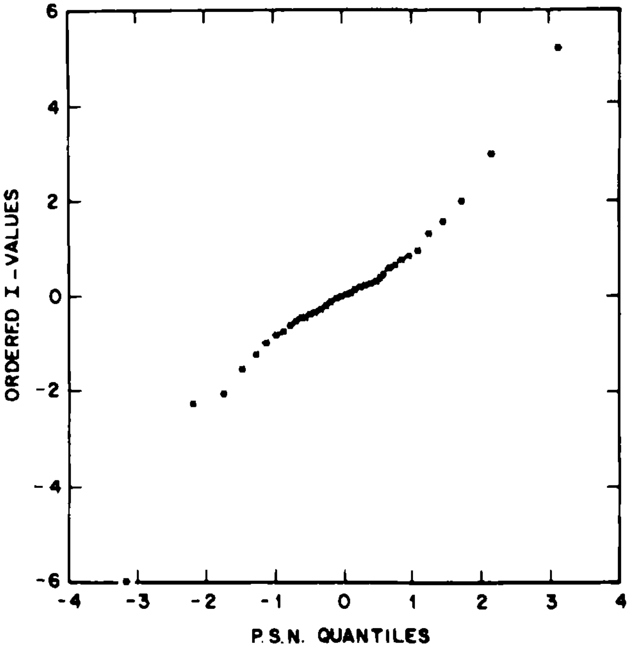
Exhibit 54c. Chi-squared ($df = 2$) probability plot of the generalized squared distances for the *Iris setosa* data of Exhibit 54*a*



Example 55. This example, taken from Devlin et al. (1976), illustrates the use of the psn probability plot (or the equivalent normal plot) of the sample influence function values of $z(r)$ to detect relatively smooth departures of the data distribution from normality. Five random samples, each of size 200, were generated from a bivariae t distribution with 5 degrees of freedom. Each sample yielded 200 sample influence function values of $z(r)$. By averaging the five corresponding ordered sample influence function values (i.e., average of the smallest in each set of 200, average of the second smallest, etc.), a smoothed set of ordered values was obtained. Exhibit 55 shows a psn probability plot of these, and the smoothly nonlinear configuration obtained indicates that the effect of the data having a bivariate t distribution rather than a bivariate normal distribution is to induce a longer-tailed (although still symmetric) distribution for the influence function values. Such an implication is accurate since it can also be established theoretically (Devlin et al., 1976).

Example 56. For illustrating the use of hierarchical clustering in identifying outliers, 14-dimensional data for 32 chemical companies for the year 1965 are taken from the study on grouping of corporations by Chen et al. (1970, 1974). Data from this study have been used repeatedly in earlier examples, and the present illustration is taken from Gnanadesikan & Kettenring (1972). The generalized squared intercompany distances in the 14-dimensional space were used as input to the minimum method of hierarchical clustering described in Section 4.3.2*a*. The results, along with clustering strength values, are displayed in Exhibit 56 (see page 316). Company 14, which joins the cluster

Exhibit 55. PSN probability plot of averaged z_{-i} values from samples of bivariate t distribution ($df = 5$)

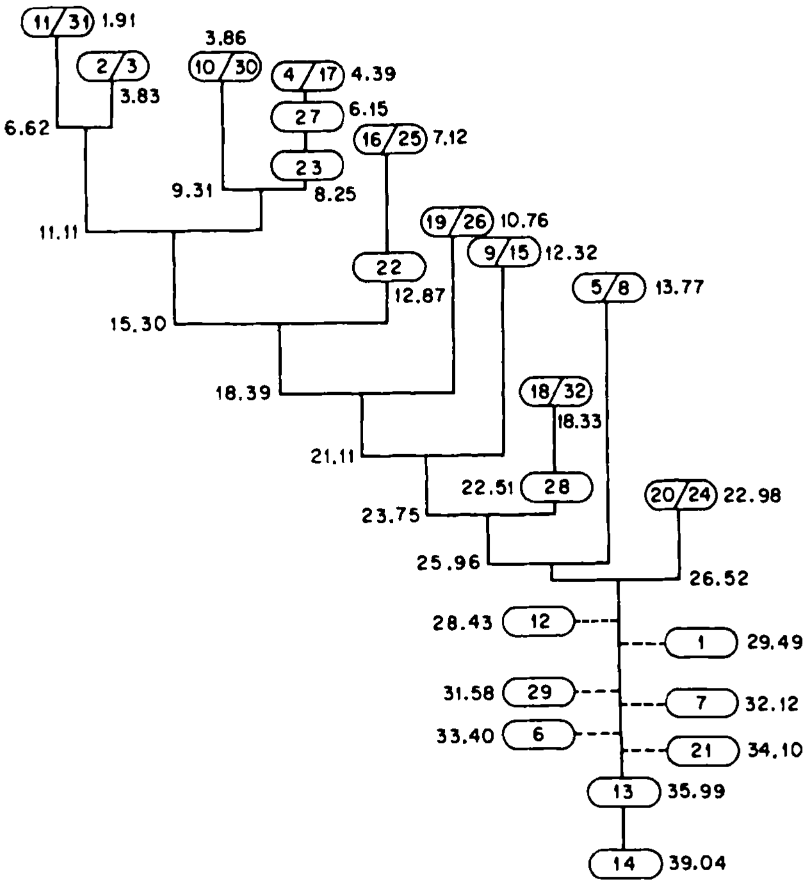


at the very end at a substantially higher clustering strength than the preceding value, appears to be an outlier, a finding that was corroborated by a variety of other analyses.

External Analysis Techniques for Outlier Detection. Discriminant analysis of two or more groups of multiresponse observations (see Section 4.2) and canonical analysis of two or more sets of variables (see Section 3.3) are among the basic multivariate external analysis techniques.

Valuable insight can be gleaned from two- and three-dimensional displays of the discriminant and canonical variables. Such views of the discriminant space, as illustrated in Examples 16 and 17, show the relative sizes, shapes, and locations of the groups, as well as possible peculiarities in the positions of individual points. The discriminant analysis may be preceded by internal analyses of the individual groups for outliers, with the hope of making the dispersions within the individual groups similar, as is required for the validity of the standard multigroup discriminant analysis procedure (see the discussion in Example 17). The remaining observations can then be used to derive the discriminant coordinates, but the visual displays may profitably include the positions of all of the data in the transformed space. The canonical variable plot, another mechanism for data exposure, can reveal outliers that are inducing an artificial linear relationship among the sets. Plots of the principal components (or other appropriate linear functions) of the canonical variables,

Exhibit 56. Hierarchical clustering tree for 32 chemical companies



as discussed in Kettenring (1971), are alternative summaries that have special appeal when the number of sets is large.

Normal probability plots and univariate outlier procedures can be applied to the canonical variables or to linear functions of them, and to the discriminant variables, making a separate plot for each group. The slopes of the configurations on the last-mentioned of these plots provide a partial check on the homogeneity of dispersion among the groups.

Gnanadesikan & Kettenring (1972) propose two examples of univariate statistics that are sensitive to the type of multivariate effects of interest in discriminant and canonical analyses. The first is

$$
\begin{aligned}
w_{ki}^2 &= \sum_i c_i\{\mathbf{a}_i'(\mathbf{y}_{ki} - \bar{\mathbf{y}}_k)\}^2 \\
&= (\mathbf{y}_{ki} - \bar{\mathbf{y}}_k)'\mathbf{W}^{-1}(\mathbf{y}_{ki} - \bar{\mathbf{y}}_k), \qquad k = 1,\ldots,g; i = 1,\ldots,n_k,
\end{aligned}
$$

where $\mathbf{y}_{ki}$ is the ith observation in the kth group, $\bar{\mathbf{y}}_k$ is the kth group mean, n_k is the number of observations in the kth group, and the eigenvalue c_t, the eigenvector $\mathbf{a}_t$, and the matrices $\mathbf{B}$ and $\mathbf{W}$ are as defined in Section 4.2 (see Eqs. 51 and 52). The statistic, w_{ki}^2, is a weighted sum of squares of the projections of $(\mathbf{y}_{ki} - \bar{\mathbf{y}}_k)$ onto the discriminant axes, and $\Sigma\Sigma\, w_{ki}^2 = (n - g)\,\Sigma\, c_t$, where $n = \Sigma\, n_k$.

For the case of canonical analysis of two sets of variables, the proposed statistic is

$$x_i^2 = \frac{\prod_t (1 - r^{(t)^2})}{\prod_t (1 - r_{-i}^{(t)^2})}$$

$$= \frac{\{1-(n/n-1)(\mathbf{y}_{1i}-\bar{\mathbf{y}}_1)'\mathbf{A}_{11}^{-1}(\mathbf{y}_{1i}-\bar{\mathbf{y}}_1)\}\{1-(n/n-1)(\mathbf{y}_{2i}-\bar{\mathbf{y}}_2)'\mathbf{A}_{22}^{-1}(\mathbf{y}_{2i}-\bar{\mathbf{y}}_2)\}}{1-(n/n-1)(\mathbf{y}_i-\bar{\mathbf{y}})'\mathbf{A}^{-1}(\mathbf{y}_i-\bar{\mathbf{y}})},$$

$$i = 1, \ldots, n,$$

where $r^{(t)}$ is the tth canonical correlation computed from all n observations, while $r_{-i}^{(t)}$ is based on all but the ith observation, and where

$$\mathbf{y}_i' = (\mathbf{y}_{1i}' \mid \mathbf{y}_{2i}'), \qquad \bar{\mathbf{y}}' = (\bar{\mathbf{y}}_1' \mid \bar{\mathbf{y}}_2'), \qquad \text{and} \qquad \mathbf{A} = (n-1)\mathbf{S} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

are partitioned in accordance with the dimensions of the two sets. (The subscript k, which designated the group in the definition of w_{ki}^2, now refers to the set.)

As aids for examining the collection of w_{ki}^2 and of x_i^2, it would seem reasonable to use gamma probability plots of the w_{ki}^2 and normal probability plots of the $\log x_i^2$. These choices for the null distributions, however, need to be investigated more carefully for their appropriateness.

REFERENCES

Section 6.1 Cox (1973), Cox & Hinkley (1974).

Section 6.2 Anderson (1935), Andrews (972), Ashton et al. (1957), Azimov et al. (1988), Buja & Eyuboglu (1993), Buja & Hurley (1990), Catell (1966), Chen et al. (1970, 1974), Cleveland et al. (1975), Cook et al. (1993), Devlin et al. (1975), Fisher (1936), Fisherkeller et al. (1974), Friedman & Rubin (1967), Gnanadesikan (1968, 1973), Gnanadesikan et al. (1967), Gnanadesikan & Wilk (1969), Hartigan (1973), Hastings (1970), Mallows & Wachter (1970), Rao (1960, 1962), Swayne et al. (1991), Stein (1969), Tukey (1970), Wilk et al. (1962a), Wilk & Gnanadesikan (1968), Zimmer & Larsen (1965).

Section 6.3 Daniel (1959), Dempster (1958), Gnanadesikan (1980), Gnanadesikan & Lee (1970), Gnanadesikan & Wilk (1970), Roy et al. (1971), Wilk et al. (1962a), Wilk & Gnanadesikan (1961, 1964).

Section 6.3.1 Andrews et al. (1971), Box (1954), Box & Cox (1964), Gnanadesikan (1980), Patnaik (1949), Roy et al. (1971), Satterthwaite (1941), Wilk et al. (1962, 1962b), Wilk & Gnanadesikan (1961, 1964).

Section 6.3.2 Flury (1984), Gnanadesikan & Lee (1970), Hoel (1937), Keramidas et al. (1987), Krzanowski (1979), Roy et al. (1971), Wilk et al. (1962b).

Section 6.4 Anscombe (1960, 1961), Barnett & Lewis (1994), Devlin et al. (1975), Dixon (1953), Draper & Smith (1981), Gnanadesikan & Kettenring (1972), Grubbs (1950, 1969), Terry (1955).

Section 6.4.1 Abrahamson et al. (1969), Chen et al. (1970, 1974), Cleveland & Kleiner (1975), Cox (1968), Healy (1968), Larsen & McCleary (1972), Rao (1964), Roy et al. (1971), Tukey (1970).

Section 6.4.2 Anderson (1935), Andrews et al. (1972), Chen et al. (1970, 1974), Devlin et al. (1975, 1976), Dixon (1953), Fisher (1936), Gnanadesikan & Kettenring (1972), Grubbs (1950, 1969), Hampel (1968, 1973, 1974), Hampel et al. (1986), Kettenring (1971), Mallows (1973), Miller (1974), Siotani (1959), Tukey (1958), Wilks (1963).

References

Abelson, R. P. & Tukey, J. W. (1959). Efficient conversion of non-metric information into metric information. *Proc. Soc. Stat. Sect. Am. Stat. Assoc.*, 226–30.

Abrahamson, I. G., Gentleman, J. F., Gnanadesikan, R., Walcheski, A. F., & Williams, D. E. (1969). Statistical methods for studying aging and for selecting semiconductor devices. *ASQC Tech. Conf. Trans.*, 533–40.

Aitchison, J. & Brown, J. A. C. (1957). *The Lognormal Distribution*. Cambridge University Press, Cambridge, UK.

Aitkin, M. A. (1972). A class of tests for multivariate normality based on linear functions of order statistics. Unpublished manuscript.

Anderson, E. (1935). The irises of the Gaspe Peninsula. *Bull. Am. Iris Soc.* **59**, 2–5.

Anderson, E. (1954). Efficient and inefficient methods of measuring specific differences. In *Statistics and Mathematics in Biology* (O. Kempthorne, ed.), Iowa State College Press, Ames, pp. 98–107.

Anderson, E. (1957). A semi-graphical method for the analysis of complex problems. *Proc. Nat. Acad. Sci. USA* **43**, 923–7. [Reprinted in *Technometrics* **2** (1960), 387–92.]

Anderson, T. W. (1966). Some nonparametric multivariate procedures based on statistically equivalent blocks. In *Multivariate Analysis* (P. R. Krishnaiah, ed.), Academic Press, New York, pp. 5–27.

Anderson, T. W. (1969). Statistical inference for covariance matrices with linear structure. In *Multivariate Analysis II* (P. R. Krishnaiah, ed.), Academic Press, New York, pp. 55–66.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. Second Edition. Wiley, New York.

Anderson, T. W. & Rubin, H. (1956). Statistical inference in factor analysis. *Proc. 3rd Berkeley Symp. Math. Stat. Probab.* **5**, 11–50.

Andrews, D. F. (1971). A note on the selection of data transformations. *Biometrika* **58**, 249–54.

Andrews, D. F. (1972). Plots of high-dimensional data. *Biometrics* **28**, 125–36.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust Estimates of Location—Survey and Advances*. Princeton University Press.

Andrews, D. F., Gnanadesikan, R., & Warner, J. L. (1971). Transformations of multivariate data. *Biometrics* **27**, 825–40.

Andrews, D. F., Gnanadesikan, R., & Warner, J. L. (1972). Methods for assessing multivariate normality. Bell Laboratories Memorandum.

Andrews, D. F., Gnanadesikan, R., & Warner, J. L. (1973). Methods for assessing multivariate normality. In *Multivariate Analysis III* (P. R. Krishnaiah, ed.), Academic Press, New York, pp. 95–116.

Andrews, D. F. & Mallows, C. L. (1974). Scale mixtures of normal distributions. *J. R. Stat. Soc.* **B36**, 99–102.

Anscombe, F. J. (1960). Rejection of outliers. *Technometrics* **2**, 123–47.

Anscombe, F. J. (1961). Examination of residuals. *Proc. 4th Berkeley Symp. Math. Stat. Probab.* **1**, 1–36.

Arabie, P. & Carroll, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika* **45**, 211–35.

Art, D., Gnanadesikan, R., & Kettenring, J. R. (1982). Data-based metrics for cluster analysis. *Utilitas Mathematica* **21 A**, 75–99.

Ashton, E. H., Healey, M. J. R., & Lipton, S. (1957). The descriptive use of discriminant functions in physical anthropology. *Proc. R. Soc.* B **146**, 552–72.

Azimov, D., Buja, A., Hurley, C. B., & MacDonald, J. A. (1988). Elements of a viewing pipeline for data analysis. In *Dynamic Graphics for Statistics* (W. S. Cleveland & M. E. McGill, eds.), Wadsworth, Belmont, CA, pp. 277–97.

Baker, F. B. & Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *J. Am. Stat. Assoc.* **70**, 31–8.

Ball, G. H. (1965). Data analysis in the social sciences—what about details? *AFIPS Conf. Proc., Fall Joint Comput. Conf.* **27**, 533–60.

Ball, G. H. & Hall, D. J. (1965). ISODATA, a novel method of data analysis and pattern classification. Stanford Research Institute Report.

Barnett, V. & Lewis, T. (1994). *Outliers in Statistical Data.* Third Edition. Wiley, New York.

Bartholomew, D. J. (1959). A test of homogeneity for ordered alternatives. *Biometrika* **46**, 36–48.

Bartlett, M. S. (1951). The effect of standardization on an approximation in factor analysis. *Biometrika* **38**, 337–44.

Barton, D. E. & Mallows, C. L. (1961). The randomization bases of the amalgamation of weighted means. *J. R. Stat. Soc.* **B23**, 423–33.

Becker, M. H., Gnanadesikan, R., Mathews, M. V., Pinkham, R. S., Pruzansky, S., & Wilk, M. B. (1965). Comparison of some statistical distance measures for talker identification. Bell Laboratories Memorandum.

Bennett, R. S. (1965). The intrinsic dimensionality of signal collections. Ph.D. thesis, Johns Hopkins University.

Bickel, P. J. (1964). On some alternative estimates for shift in the p-variate one-sample problem. *Ann. Math. Stat.* **35**, 1079–90.

Blackith, R. E. (1960). A synthesis of multivariate techniques to distinguish patterns of growth in grasshoppers. *Biometrics* **16**, 28–40.

Blackith, R. E. & Roberts, M. I. (1958). Farbenpolymorphismus bei einigen Feldheuschrecken. *Z. Vererbungsl.* **89**, 328–37.

Blake, I. F. & Thomas, J. B. (1968). On a class of processes arising in linear estimation theory. *IEEE. Trans. on Inf. Theory IT*-**14**, 12–6.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems—I. *Ann. Math. Stat.* **25**, 290–302.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc.* **B26**, 211–52.

Boynton, R. M. & Gordon, J. (1965). Bezold-Brüke hue shift measured by color-naming technique. *J. Opt. Soc. Am.* **55**, 78–86.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, CA.

Bricker, P. D., Gnanadesikan, R., Mathews, M. V., Pruzansky, S., Tukey, P. A., Wachter, K. W., & Warner, J. L. (1971). Statistical techniques for talker identification. *Bell. Syst. Tech. J.* **50**, 1427–54.

Bruntz, S. M., Cleveland, W. S., Kleiner, B., & Warner, J. L. (1974). The dependence of ambient ozone on solar radiation, wind, temperature, and mixing height. *Proc. Symp. Atmos. Diffus. Air Pollution, Am. Meteorol. Soc.* 125–8.

Buja, A. & Eyuboglu, N. (1993). Remarks on parallel analysis. *Multivariate Behavioral Research* **27**, 509–40.

Buja, A. & Hurley, C. B. (1990). Analyzing high-dimensional data with motion graphics. *SIAM J. on Scientific and Statistical Computing* **11**, 1193–1211.

Burnaby, T. P. (1966). Growth invariant discriminant functions and generalized distances. *Biometrics* **22**, 96–110.

Businger, P. A. (1965). Algorithm 254. Eigenvalues and eigenvectors of a real symmetric matrix by the QR method. *Commun. ACM* **8**, 218–9.

Businger, P. A. & Golub, G. H. (1969). Algorithm 358. Singular value decomposition of a complex matrix. *Commun. ACM* **12**, 564–5.

Carroll, J. D. (1969). Polynomial factor analysis. *Proc. 77th Ann. Conv. Am. Psych. Assoc.*, 103–4.

Carroll, J. D. & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an *N*-way generalization of "Eckart-Young" decomposition. *Psychometrika* **35**, 283–319.

Catell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research* **1**, 245–76.

Chambers, J. M. (1973). Fitting nonlinear models: numerical techniques. *Biometrika* **60**, 1–15.

Chambers, J. M. (1977). *Computational Methods for Data Analysis.* Wiley, New York.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical Methods for Data Analysis.* Wadsworth International Group, Belmont, CA.

Chen, H., Gnanadesikan, R., & Kettenring, J. R. (1974). Statistical methods for grouping corporations. *Sankhyā* **B36**, 1–28.

Chen, H. J., Gnanadesikan, R., Kettenring, J. R., & McElroy, M. B. (1970). A statistical study of groupings of corporations. *Proc. Bus. Econ. Stat. Sect. Am. Stat. Assoc.*, 447–51.

Chen, H. J. & Kettenring, J. R. (1972). CANON: A computer program package for the multi-set canonical correlation analysis. Bell Laboratories Technical Memorandum.

Chernoff, H. (1973). The use of faces to represent points in *k*-dimensional space graphically. *J. Am. Stat. Assoc.* **68**, 361–8.

Chu, K. C. (1973). Estimation and decision for linear systems with elliptical random processes. *IEEE Trans. Automatic Control AC*-**18**, 499–505.

Cleveland, W. S. & Kleiner, B. (1975). A graphical technique for enhancing scatterplots with moving statistics. *Technometrics* **17**, 447–54.

Cleveland, W. S., Kleiner, B., McRae, J. E., Warner, J. L., & Pasceri, R. E. (1975). The analysis of ground-level ozone data from New Jersey, New York, Connecticut, and Massachusetts: data quality assessment and temporal and geographical properties. Paper presented at the 68th annual meeting of the Air Pollution Control Association.

Cohen, A., Gnanadesikan, R., Kettenring, J. R., & Landwehr, J. M. (1977). Methodological developments in some applications of clustering. In *Applications of Statistics* (P. R. Krishnaiah, ed.), North-Holland Publishing Co., New York, pp. 141–62.

Cook, D., Buja, A., & Cabrera, J. (1993). Projection pursuit indices based on orthogonal function expansions. *J. Computational and Graphical Statistics* **2**, 225–50.

Coombs, C. H. (1964). *A Theory of Data.* Wiley, New York.

Cormack, R. M. (1971). A review of classification. *J. R. Soc.* **A134**, 321–67.

Cox, D. R. (1968). Notes on some aspects of regression analysis. *J. R. Stat. Soc.* **A131**, 265–79.

Cox, D. R. (1970). *The Analysis of Binary Data.* Methuen, London.

Cox, D. R. (1972). The analysis of multivariate binary data. *Appl. Stat.* **21**, 113–20.

Cox, D. R. (1973). Theories of statistical inference. Rietz Lecture, Institute of Mathematical Statistics meetings in New York.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics.* Chapman & Hall, London.

D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika* **58**, 341–8.

D'Agostino, R. B. & Pearson, E. S. (1973). Tests for departure from normality. Empirical results for the distributions of b_2 and $\sqrt{b_1}$. *Biometrika* **60**, 613–22.

Daniel, C. (1959). The use of half-normal plots in interpreting factorial two level experiments. *Technometrics* **1**, 311–41.

Das Gupta, S., Eaton, M. L., Olkin, I., Perlman, M., Savage, L. J., & Sobel, M. (1972). Inequalities on the probability content of convex regions for elliptically contoured distributions. *Proc. 6th Berkeley Symp. Math. Stat. Probab.* **2**, 241–65.

David, F. N. & Johnson, N. L. (1948). The probability integral transformation when parameters are estimated from the sample. *Biometrika* **35**, 182–90.

Dempster, A. P. (1958). A high dimensional two sample significance test. *Ann. Math. Stat.* **29**, 995–1010.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62**, 531–45.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1976). Some multivariate applications of elliptical distributions. In *Essays on Probability and Statistics* (S. Ikeda et al., eds.), Shinko Tsusho Co. Ltd., Tokyo, pp. 365–93.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. *J. Am. Stat. Assoc.* **76**, 354–62.

Dixon, W. J. (1953). Processing data for outliers. *Biometrics* **9**, 74–89.

Donnell, D. J., Buja, A., & Stuetzle, W. (1994). Analysis of additive dependencies and concurvities using smallest additive principal components. (with discussion) *Ann. Stat.* **22**, 1635–73.

Downton, F. (1966). Linear estimates with polynomial coefficients. *Biometrika* **53**, 129–41.

Draper, N. R. & Smith, H. (1981). *Applied Regression Analysis.* Second Edition. Wiley, New York.

Ekman, G. (1954). Dimensions of color vision. *J. Psych.* **38**, 467–74.

Everitt, B. (1974). *Cluster Analysis.* Wiley, New York.

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. *Technometrics* **17**, 111–7.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–88.

Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Ann. Eugen.* **8**, 376–86.

Fisherkeller, M. A., Friedman, J. H., & Tukey, J. W. (1974). PRIM-9. An interactive multidimensional data display and analysis system. [Also a film "PRIM-9," produced by Stanford Linear Accelerator Center (S. Steppel, ed.).] Stanford Linear Accelerator Center Pub. 1408.

Fletcher, R. & Powell, M. J. D. (1963). A rapidly convergent descent method for minimization. *Comput. J.* **2**, 163–8.

Flury, B. H. (1984). Common principal components in K groups. *J. Am. Stat. Assoc.* **79**, 892–8.

Fowlkes, E. B., Gnanadesikan, R., & Kettenring, J. R. (1987). Variable selection in clustering and other contexts. In *Design, Data & Analysis* (C. L. Mallows, ed.), Wiley, New York, pp. 13–34.

Fowlkes, E. B., Gnanadesikan, R., & Kettenring, J. R. (1988). Variable selection in clustering. *J. Classification* **5**, 205–28.

Fowlkes, E. B. & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.* **78**, 553–84.

Friedman, H. P. & Rubin, J. (1967). On some invariant criteria for grouping data. *J. Am. Stat. Assoc.* **62**, 1159–78.

Friedman, J. H. & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. on Computers* **23**, 881–9.

Gentleman, W. M. (1965). Robust estimation of multivariate location by minimizing *p*th power deviations. Ph.D. thesis, Princeton University.

Gnanadesikan, R. (1968). The study of multivariate residuals and methods for detecting outliers in multiresponse data. Invited paper presented at American Society for Quality Control, Chemical Division, meetings at Durham, NC.

Gnanadesikan, R. (1972). Methods for evaluating similarity of marginal distributions. *Stat. Neerl.* **26**, No. 3, 69–78.

Gnanadesikan, R. (1973). Graphical methods for informal inference in multivariate data analysis. *Bull. Int. Stat. Inst. Proc. 39th Sess. ISI at Vienna* **45**, Book 4, 195–206.

Gnanadesikan, R. (1980). Graphical methods for internal comparisons in ANOVA and MANOVA. In *Handbook of Statistics* **1** (P. R. Krishaiah, ed.), North-Holland, Amsterdam, pp. 133–77.

Gnanadesikan, R., Harvey, J. W., & Kettenring, J. R. (1993). Mahalanobis metrics for cluster analysis. *Sankhyā* **A55**, 494–505.

Gnanadesikan, R. & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **28**, 81–124.

Gnanadesikan, R. & Kettenring, J. R. (1989). Discriminant analysis and clustering. (Report of NRC Panel on Discriminant Analysis and Clustering.) *Statistical Science* **4**, 34–69.

Gnanadesikan, R., Kettenring, J. R., & Landwehr, J. M. (1977). Interpreting and assessing the results of cluster analyses. *Bull Int. Stat. Stat. Inst., Proc. 41st Sess. ISI at New Delhi* **47**, Book 2, 451–63.

Gnanadesikan, R., Kettenring, J. R., & Landwehr, J. M. (1982). Projection plots for displaying clusters. In *Statistics and Probability: Essays in Honor of C. R. Rao* (G. Kallianpur, P. R. Krishnaiah, & J. K. Ghosh, eds.), North-Holland, Amsterdam, pp. 269–80.

Gnanadesikan, R., Kettenring, J. R., & Tsao, S. L. (1995). Weighting and selection of variables for cluster analysis. *J. Classification* **12**, 113–36.

Gnanadesikan, R. & Lee, E. T. (1970). Graphical techniques for internal comparisons amongst equal degree of freedom groupings in multiresponse experiments. *Biometrika* **57**, 229–37.

Gnanadesikan, R., Pinkham, R. S., & Hughes, L. P. (1967). Maximum likelihood estimation of the parameters of the beta distribution from smallest order statistics. *Technometrics* **9**, 607–20.

Gnanadesikan, R. & Wilk, M. B. (1966). Data analytic methods in multivariate statistical analysis. General Methodology Lecture on Multivariate Analysis, 126th Annual Meeting of the American Statistical Association, Los Angeles.

Gnanadesikan, R. & Wilk, M. B. (1969). Data analytic methods in multivariate statistical analysis. In *Multivariate Analysis II* (P. R. Krishnaiah, ed.), Academic Press, New York, pp. 593–638.

Gnanadesikan, R. & Wilk, M. B. (1970). A probability plotting procedure for general analysis of variance. *J. R. Stat. Soc.* **B32**, 88–101.

Goldman, J. (1974). Statistical properties of a sum of sinusoids and gaussian noise and its generalization to higher dimensions. *Bell Syst. Tech. J.* **53**, 557–80.

Golub, G. H. (1968). Least squares, singular values and matrix approximations. *Apl. Mat.* **13**, 44–51.

Golub, G. H. & Reinsch, C. (1970). Handbook series linear algebra: singular value decomposition and least squares solutions. *Numer. Math.* **14**, 403–20.

Gower, J. C. (1967). A comparison of some methods of cluster analysis. *Biometrics* **23**, 623–37.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Ann. Math. Stat.* **21**, 27–58.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics* **11**, 1–21.

Hampel, F. R. (1968). Contributions to the theory of robustness. Ph.D. thesis, University of California at Berkeley.

Hampel, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Stat.* **42**, 1887–96.

Hampel, F. R. (1973). Robust estimation: a condensed partial survey. *Z. Wahr. verw. Geb.* **27**, 87–104.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.* **69**, 383–93.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

Hand, D. J. (1981). *Discrimination and Classification.* Wiley, New York.

Harman, H. H. (1967). *Modern Factor Analysis,* second edition (revised). University of Chicago Press.

Hartigan, J. A. (1967). Representation of similarity matrices by trees. *J. Am. Stat. Assoc.* **62**, 1140–58.

Hartigan, J. A. (1973). Printer graphics for clustering. Unpublished manuscript. (Also see section 1.7.6. of Hartigan, 1975.)

Hartigan, J. A. (1975). *Clustering Algorithms.* Wiley, New York.

Hastie, T. J., Tibshirani, R., & Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *J. Am. Stat. Assoc.* **89**, 1255–70.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

Healy, M. J. R. (1968). Multivariate normal plotting. *Appl. Stat.* **17**, 157–61.

Hoel, P. G. (1937). A significance test for component analysis. *Ann. Math. Stat.* **8**, 149–58.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Horst, P. (1965). *Factor Analysis of Data Matrices.* Holt, Rinehart & Winston, New York.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.* **24**, 417–41, 498–520.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321–77.

Hotelling, H. (1947). Multivariate quality control, illustrated by the air testing of sample bombsights. In *Selected Techniques of Statistical Analysis* (C. Eisenhart et al., eds.), McGraw-Hill, New York, pp. 111–84.

Howe, W. G. (1955). Some contributions to factor analysis. Oak Ridge National Laboratory, ORNL-1919, Oak Ridge, TN.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 73–101.

Huber, P. J. (1970). Studentizing robust estimates. In *Nonparametric Techniques in Statistical Inference* (M. L. Puri, ed.), Cambridge University Press, pp. 453–63.

Huber, P. J. (1972). Robust statistics: a review. *Ann. Math. Stat.* **43**, 1041–67.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Stat.* **1**, 799–821.

Huber, P. J. (1977). Robust covariances. In *Statistical Decision Theory and Related Topics 2* (S. S. Gupta & D. S. Moore, eds.), Academic Press, New York, pp. 165–91.

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

Hubert, L. J. (1974). Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *J. Am. Stat. Assoc.* **69**, 698–704.

Imbrie, J. (1963). Factor and vector analysis programs for analysing geological data. Tech. Rept. 6, ONR Task No. 389–135.

Imbrie, J. & Kipp, N. G. (1971). A new micropaleontological method for quantitative paleoclimatology: application to a late pleistocene Caribbean core. In *Late Cenozoic Glacial Ages* (K. K. Turekian, ed.), Yale University Press, New Haven, CT, pp. 71–181.

Imbrie, J. & Van Andel, T. H. (1964). Vector analysis of heavy-mineral data. *Bull Geol. Soc. Am.* **75**, 1131–55.

Jackson, J. E. (1956). Quality control methods for two related variables. *Ind. Qual. Control* **12**, 2–6.

Johnson, N. L. & Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* Wiley, New York.

Johnson, N. L. & Leone, F. C. (1964). *Statistics and Experimental Design in Engineering and the Physical Sciences*, Vol. I. Wiley, New York.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika* **32**, 241–54.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32**, 443–82.

Jöreskog, K. G. (1973). Analysis of covariance structures. In *Multivariate Analysis III* (P. R. Krishnaiah, ed.), Academic Press, New York, pp. 263–85.

Jöreskog, K. G. & Lawley, D. N. (1968). New methods in maximum likelihood factor analysis. *Br. J. Math. Stat. Psych.* **21**, 85–96.

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data.* Wiley, New York.

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā* **A32**, 419–30.

Kempthorne, O. (1966). Multivariate responses in comparative experiments. In *Multivariate Analysis* (P. R. Krishnaiah, ed.), Academic Press, New York, pp. 521–40.

Kendall, M. G. (1968). On the future of statistics—a second look. *J. R. Stat. Soc.* **A131**, 182–92.

Keramidas, E. M., Devlin, S. J., & Gnanadesikan, R. (1987). A graphical procedure for comparing the principal components of several covariance matrices. *Commun. Stat.-Simula.* **16**, 161–91.

Kessell, D. L. & Fukunaga, K. (1972). A test for multivariate normality with unspecified parameters. Unpublished report, Purdue University School of Electrical Engineering.

Kettenring, J. R. (1969). Canonical analysis of several sets of variables. Ph.D. thesis, University of North Carolina.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika* **58**, 433–51.

Kettenring, J. R., Rogers, W. H., Smith, M. E., & Warner, J. L. (1976). Cluster analysis applied to the validation of course objectives. *J. Educ. Stat.* **1**, 39–57.

Kingman, J. F. C. (1972). On random sequences with spherical symmetry. *Biometrika* **59**, 492–94.

Krasker, W. S. & Welsch, R. E. (1982). Efficient bounded-influence regression estimation. *J. Am. Stat. Assoc.* **77**, 595–604.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**, 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika* **29**, 115–29.

Krzanowski, W. J. (1979). Between-groups comparisons of principal components. *J. Am. Stat. Assoc.* **74**, 703–7.

Larsen, W. A. & McCleary, S. J. (1972). The use of partial residul plots in regression analysis. *Technometrics* **14**, 781–90.

Laue, R. V. & Morse, M. F. (1968). Simulation of traffic distribution schemes for No. 5 ACD. Bell Laboratories Memorandum.

Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proc. R. Soc. Edin.* **A60**, 64–82.

Lawley, D. N. (1967). Some new results in maximum likelihood factor analysis. *Proc. R. Soc. Edin.* **A67**, 256–64.

Lawley, D. N. & Maxwell, A. E. (1963). *Factor Analysis as a Statistical Method.* Butterworth, London. (Second edition, 1971.)

Lax, D. A. (1975). An interim report of a Monte Carlo study of robust estimators of width. Technical Report 93, Series 2, Department of Statistics, Princeton University, Princeton, N.J.

Lindley, D. V. (1972). Book review (*Analysis and Design of Certain Quantitative Multiresponse Experiments* by Roy et al.) *Bull. Inst. Math. Appl.* **8**, 134.

Ling, R. F. (1973). A probability theory for cluster analysis. *J. Am. Stat. Assoc.* **68**, 159–64.

MacQueen, J. (1965). Some methods for classification and analysis of multivariate observations. *Proc. 5th Berkeley Symp. Math. Stat. Probab.* **1**, 281–97.

Malkovich, J. F. & Afifi, A. A. (1973). On tests for multivariate normality. *J. Am. Stat. Assoc.* **68**, 176–9.

Mallows, C. L. (1973). Influence functions. Unpublished talk presented at the Working Conference on Robust Regression at National Bureau of Economics Research in Cambridge, MA.

Mallows, C. L. (1983). Robust methods. In *Statistical Data Analysis* (R. Gnanadesikan, ed.), *Proceedings of Symposia in Applied Math.* **28**, American Mathematical Society, Providence, RI, pp. 49–74.

Mallows, C. L. & Wachter, K. W. (1970). The asymptotic configuration of Wishart eigenvalues. Abstract 126–5, *Ann. Math. Stat.* **41**, 1384.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57**, 519–30.

Mardia, K. V. (1975). Assessment of multinormality and the robustness of Hotelling's T^2 test. *Appl. Stat.* **24**, 163–71.

Mardia, K. V. (1980). Tests of univariate and multivariate normality. In *Handbook of Statistics* **1** (P. R. Krishnaiah, ed.), North-Holland, Amsterdam, pp. 279–320.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Stat.* **4**, 51–67.

McDonald, R. P. (1962). A general approach to nonlinear factor analysis. *Psychometrika* **27**, 397–415.

McDonald, R. P. (1967). Numerical methods for polynomial models in non-linear factor analysis. *Psychometrika* **32**, 77–112.

McLaughlin, D. H. & Tukey, J. W. (1961). The variance of means of symmetrically trimmed samples from normal populations, and its estimation from such trimmed samples. (Trimming/Winsorization I.) Tech. Rept. 42, Statistical Techniques Research Group, Princeton University.

Miles, R. E. (1959). The complete amalgamation into blocks, by weighted means, of a finite set of real numbers. *Biometrika* **46**, 317–27.

Miller, G. A. & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.* **27**, 338–52.

Miller, R. G. (1974). The jackknife—a review. *Biometrika* **61**, 1–15.

Mood, A. M. (1941). On the joint distribution of the median in samples from a multivariate population. *Ann. Math. Stat.* **12**, 268–78.

Moore, P. G. & Tukey, J. W. (1954). Answer to query 112. *Biometrics* **10**, 562–8.

Patnaik, P. B. (1949). The non-central χ^2 and F-distributions and their approximations. *Biometrika* **36**, 202–32.

Pearson, E. S. & Hartley, H. O. (1966). *Biometrika Tables for Statisticians*, Vol. I. Cambridge University Press.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Phil. Mag.* [6] **2**, 559–72.

Picinbono, B. (1970). Spherically invariant and compound Gaussian stochastic processes. *IEEE Trans. Information Theory, IT*-**16**, 77–9.

Pillai, K. C. S. & Jayachandran, K. (1967). Power comparisons of tests of two multivariate hypotheses based on four criteria. *Biometrika* **54**, 195–210.

Puri, M. L. & Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. Wiley, New York.

Rao, C. R. (1960). Multivariate analysis: an indispensable statistical aid in applied research. *Sankhyā* **22**, 317–38.

Rao, C. R. (1962). Use of discriminant and allied functions in multivariate analysis. *Sankhyā* **A24**, 149–54.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā* **A26**, 329–58.

Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*. Wiley, New York. (Second edition, 1973.)

Rao, C. R. (1966). Discriminant function between composite hypotheses and related problems. *Biometrika* **53**, 339–45.

Rogers, W. H. & Tukey, J. W. (1972). Understanding some long tailed symmetrical distributions. *Stat. Neerl.* **26**, 211–26.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Stat.* **23**, 470–2.

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *J. Exp. Psych.* **53**, 94–101.

Rousseeuw, P. J. (1983). Multivariate estimation with high breakdown point. Fourth Pannonian Symposium on Mathematical Statistics, Bad Tatzmannsdorf, Austria.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Stat.* **24**, 220–38.

Roy, S. N. (1957). *Some Aspects of Multivariate Analysis.* Wiley, New York.

Roy, S. N. & Bose, R. C. (1953). Simultaneous confidence interval estimation. *Ann. Math. Stat.* **24**, 513–36.

Roy, S. N. & Gnanadesikan, R. (1957). Further contributions to multivariate confidence bounds. *Biometrika* **44**, 399–410.

Roy, S. N. & Gnanadesikan, R. (1962). Two-sample comparisons of dispersion matrices for alternatives of intermediate specificity. *Ann. Math. Stat.* **33**, 432–7.

Roy, S. N., Gnanadesikan, R., & Srivastava, J. N. (1971). *Analysis and Design of Certain Quantitative Multiresponse Experiments.* Pergamon Press, Oxford.

Sarhan, A. E. & Greenberg, B. (1956). *Contributions to Order Statistics.* Wiley, New York.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika* **6**, 309–16.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika* **40**, 87–104.

Scheffé, H. (1959). *The Analysis of Variance.* Wiley, New York.

Scott, A. J. & Symons, M. J. (1971). On the Edwards and Cavalli-Sforza method of cluster analysis. Note 297, *Biometrics* **27**, 217–9.

Seal, H. L. (1964). *Multivariate Statistical Analysis for Biologists.* Methuen, London.

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality. *Biometrika* **52**, 591–611.

Shapiro, S. S., Wilk, M. B., & Chen, H. (1968). A comparative study of various tests for normality. *J. Am. Stat. Assoc.* **63**, 1343–72.

Shepard, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function—I. *Psychometrika* **27**, 125–40.

Shepard, R. N. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function—II. *Psychometrika* **27**, 219–46.

Shepard, R. N. (1963). Analysis of proximities as a study of information processing in man. *Human Factors* **5**, 33–48.

Shepard, R. N. & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review* **86**, 87–123.

Shepard, R. N. & Carroll, J. D. (1966). Parametric representation of nonlinear data structures. In *Multivariate Analysis* (P. R. Krishnaiah, ed.), Academic Press, New York, pp. 561–92.

Siotani, M. (1959). The extreme value of the generalized distances of the individual points in the multivariate sample. *Ann. Inst. Stat. Math.* **10**, 183–203.

Sneath, P. H. A. (1957). The application of computers to taxonomy. *J. Gen. Microbiol.* 17, 201–26.

Sokal, R. R. & Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*. W. H. Freeman & Co., San Francisco.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biol. Skr.* **5**, No. 4, 1–34.

Srivastava, J. N. (1966). On testing hypotheses regarding a class of covariance structures. *Psychometrika* **31**, 147–64.

Steel, R. G. D. (1951). Minimum generalized variance for a set of linear functions. *Ann. Math. Stat.* **22**, 456–60.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. 3rd Berkeley Symp. Math. Stat. Probab.* **1**, 197–206.

Stein, C. (1965). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Ann. Inst. Stat. Math.* **16**, 155–6.

Stein, C. (1969). Mimeographed lecture notes on multivariate analysis as recorded by M. Eaton. Stanford University.

Swayne, D. F., Cook, D., & Buja, A. (1991). XGobi: Interactive dynamic graphical displays in the X Window System with a link to S. *Proc. Statist. Graphics Sec. Am. Stat. Assoc.*, 1–8.

Teichroew, D. (1956). Tables of expected values of order statistics and products of order statistics for samples of size twenty and less from the normal distribution. *Ann. Math. Stat.* **27**, 410–26.

Terry, M. E. (1955). On the analysis of planned experiments. *Nat. Conv. ASQC Trans.*, 553–6.

Theil, H. (1963). On the use of incomplete prior information in regression analysis. *J. Am. Stat. Assoc.* **58**, 401–14.

Thomson, G. H. (1934). Hotelling's method modified to give Spearman's g. *J. Educ. Psych.* **25**, 366–74.

Thurstone, L. L. (1931). Multiple factor analysis. *Psych. Rev.* **38**, 406–27.

Thurstone, L. L. & Thurstone, T. G. (1941). Factorial studies of intelligence. *Psychometric Monogr.* **2**.

Tukey, J. W. (1957). On the comparative anatomy of transformations. *Ann. Math. Stat.* **28**, 602–32.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples (Abstract). *Ann. Math. Stat.* **29**, 614.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics* (I. Olkin et al., eds.), Stanford University Press, pp. 448–85.

Tukey, J. W. (1962). The future of data analysis. *Ann. Math. Stat.* **33**, 1–67.

Tukey, J. W. (1970). *Exploratory Data Analysis*, limited preliminary edition, Addison-Wesley, Reading, Mass.

Tukey, J. W. & McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: trimming/Winsorization 1 *Sankhyā* **A25**, 331–52.

Tukey, J. W. & Wilk, M. B. (1966). Data analysis and statistics: an expository overview. *AFIPS Conf. Proc., Fall Joint Comput. Conf.* **29**, 695–709.

Tyler, D. E. (1983). Robustness and efficiency properties of scatter matrices. *Biometrika* **70**, 411–20.

Van De Geer, J. P. (1968). Fitting a quadratic function to a two-dimensional set of points. Unpublished research note, RN 004-68, Department of Data Theory for the Social Sciences, University of Leiden, The Netherlands.

Van Eeden, C. (1957a). Maximum likelihood estimation of partially or completely ordered parameters, I. *Proc. Akad. Wet.* **A60**, 128–36.

Van Eeden, C. (1957b). Note on two methods for estimating ordered parameters of probability distributions. *Proc. Akad. Wet.* **A60**, 506–12.

Vershik, A. M. (1964). Some characteristic properties of Gaussian stochastic processes. *Theor. Probability Appl.* **9**, 353–6.

Warner, J. L. (1968). An adaptation of ISODATA-POINTS, an *i*terative *s*elf-*o*rganizing *d*ata *a*nalysis *t*echnique *a*. Bell Laboratories Memorandum.

Warner, J. L. (1969). Hierarchical clustering schemes. Bell Laboratories Memorandum.

Weiss, L. (1958). A test of fit for multivariate distributions. *Ann. Math. Stat.* **29**, 595–9.

Wilk, M. B. & Gnanadesikan, R. (1961). Graphical analysis of multi-response experimental data using ordered distances. *Proc. Nat. Acad. Sci. USA* **47**, 1209–12.

Wilk, M. B. & Gnanadesikan, R. (1964). Graphical methods for internal comparisons in multiresponse experiments. *Ann. Math. Stat.* **35**, 613–31.

Wilk, M. B. & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika* **55**, 1–17.

Wilk, M. B., Gnanadesikan, R., & Huyett, Miss M. J. (1962a). Probability plots for the gamma distribution. *Technometrics* **4**, 1–20.

Wilk, M. B., Gnanadesikan, R., & Huyett, M. J. (1962b). Estimation of the parameters of the gamma distribution using order statistics. *Biometrika* **49**, 525–45.

Wilk, M. B., Gnanadesikan, R., Huyett, M. J., & Lauh, Miss E. (1962). A study of alternate compounding matrices used in a graphical internal comparisons procedure. Bell Laboratories Memorandum.

Wilkinson, G. N. (1970). A general recursive procedure for analysis of variance. *Biometrika* **57**, 19–46.

Wilks, S. S. (1963). Multivariate statistical outliers. *Sankhyā* **A25**, 407–26.

Yao, K. (1973). A representation theorem and its applications to spherically-invariant random processes. *IEEE Trans. Information Theory IT*-**19**, 600–8.

Zimmer, C. E. & Larsen, R. I. (1965). Calculating air quality and its control. *J. Air Pollution Control Assoc.* **15**, 565–72.

APPENDIX

Software

The development of software for multivariate data analysis in the last two decades has seen a veritable explosion of implementations ranging from specialized packages for specific techniques to inclusion of the methods in widely-used software systems with broad capabilities for data analysis (e.g., S, S-plus, SAS, SPSS). Even with the latter type, there is a variety of platforms on which they run resulting in some differences in capabilities and features, although general characteristics such as the names of the functions remain the same across platforms. In this appendix, with few exceptions, the emphasis is on two of the most widely-used systems, S and SAS.

A perennial peril with advice on software is that it is out of date almost as soon as it is given! The material in this appendix is no exception. Despite this state of affairs, the practical value and impact of methods such as those discussed in this book depend critically on the ability to use them as implemented in some software. With this in mind, all that is intended here is to provide some pointers to software and references which will enable the reader to use several, though not all, of the techniques described in the earlier chapters. Specifically, the main aim is to list the function names for the techniqus as implemented in S (and S-plus if different) and SAS. In a few cases, where specialized computer programs not directly available in either of these systems are needed, references to other sources are mentioned. The order followed in the discussion and listing will closely parallel the one in which the techniques were described in the earlier chapters.

Before listing the functions, or providing references to sources for specific software, a few general comments are appropriate. As documentation sources, of course, the most complete are the hardcopy and online "manuals" for S, S-plus and SAS. The appropriate references for the hardcopy versions of these manuals are:

For S,

(1) Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The New S Language*. Wadsworth & Brooks/Cole, Pacific Grove, CA.
(2) Chambers, J. M. & Hastie, T. J. (eds.) (1992). *Statistical Models in S*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

For S-plus, of the many manuals available for it, the following form a useful subset:

(1) Statistical Sciences (1993). *A Gentle Introduction to S-PLUS, Version 3.2*. StatSci, Seattle, WA.
(2) Statistical Sciences (1993). *A Crash Course in S-PLUS, Version 3.2.* StatSci, Seattle, WA.
(3) Statistical Sciences (1993). *S-PLUS User's Manual, Version 3.2.* StatSci, Seattle, WA.
(4) Statistical Sciences (1993). *S-PLUS Reference Manual, Volume 1, Version 3.2.* StatSci, Seattle, WA.
(5) Statistical Sciences (1993). *S-PLUS Reference Manual, Volume 2, Version 3.2.* StatSci, Seattle, WA.

For SAS:

(1) SAS Institute (1985). *SAS User's Guide: Basics, Version 5.* SAS Institute, Cary, NC.
(2) SAS Institute (1985). *SAS User's Guide: Statistics, Version 5.* SAS Institute, Cary, NC.

For users who may have access to any of these systems, the online documentation of them may be easier to use. Given the name of a function, the online documentation is obtained either by typing in the command "help('function name')" or by an item in a pulldown menu depending on the particular implementation. Thus the function names listed in this appendix may be useful for such users as a starting point.

In addition to documentation provided by the developers/distributors of these systems, there are a number of books written by others for guiding users with different levels of statistical and computing expertise. For instance, for users of S-plus, a valuable reference for a whole range of data analyses that one may wish to undertake is:

Venables, W. N. & Ripley, B. D. (1994). *Modern Applied Statistics with S-Plus.* Springer-Verlag, New York.

Similarly, for SAS an example is the book:

Hatcher, L. & Stepanski, E. (1994). *A Step-by-Step Approach to Using SAS for Univariae and Multivariate Statistics.* SAS Institute, Cary, NC.

A list of function names in S/S-plus and in SAS, some comments on details of usage in a few cases, and pointers regarding additional software sources, follow.

Matrix Computations

Since virtually all of the techniques described in the earlier chapters consist of building blocks entailing matrix operations, the functions for basic matrix computations are mentioned first. A particular strength of S/S-plus is the inclusion of basic matrix operators and functions for the more widely-used matrix computations. In the following list, where functions are involved their names are shown in quotes while operations are indicated by their symbols:

Addition and subtraction—**A** + **B**, **A** − **B**
Multiplication—**A%*%B**
Transpose—"t(A)"
Inverse—"solve(A)"
Choleski decomposition—"chol(A)" [Note that what is returned in S/S-plus by this function is the upper triangular matrix, **T**′, where **A** = **TT**′].
Singular Value Decomposition—"svd(A)"
Eigenanalysis—"eigen(A)".

In Version 5 of SAS, PROC MATRIX embodies a number of functions for matrix manipulations. However, Version 6 of SAS includes a software system called SAS/IML (where IML stands for Interactive Matrix Language) and PROC IML appears to have some significant advantages over PROC MATRIX. The interested reader is referred to: "SAS/IML User's Guide, Release 6.03 Edition," SAS Institute, Cary, NC.

Plotting Multivariate Data

A major strength of S/S-plus is their graphical capabilities. The functions mentioned here are just a few of the more basic ones available in these two systems. For obtaining the most widely-used display, an x-y scatter plot, the function is "plot(x, y)", where "x" and "y" are vectors containing the x- and y-coordinates of the points to be plotted. [*Note*: There are a number of arguments and graphical parameters that enable the user to manipulate the displays in a variety of ways. However, for present purposes, these details are left out.] The function, "pairs(x)", where "x" is an $n \times p$ matrix of data, will produce scatter plots of all possible pairs of the p variables. The function, "stars(x)", will generate a star plot (or snowflake plot) of the data with a star representing each of the n observations. Also, "faces(x)" is the function to use in S/S-plus for getting a Chernoff's faces display of the data. (See Section 3.2.) At the moment, the reader interested in Andrews's Curves (see Section 6.2) can obtain S functions developed by Chris Rogers at Rutgers upon request. Soon these may be made available through the statlib facility at Carnegie Mellon University (see discussion of statlib below).

In SAS, PROC PLOT is the basic plotting tool. A reference for the reader interested in using SAS for graphics more generally is: Friendly, M. (1991). "SAS System for Statistical Graphics," SAS Institute, Cary, NC. This book describes ways of getting a large number of graphical dispays in SAS, including univariate displays (e.g., standard things such as histograms and box plots as well as Q-Q probability plots) and multivariate displays (e.g., glyphs, Andrews's curves, star plots, and plots for assessing multivariate normality). SAS macro programs which package together the pieces needed for the various displays are described and illustrated.

As mentioned in various places in the earlier chapters (e.g., Sections 4.3.2, 6.2), graphical displays that incorporate motion, linking, and interaction are now widely available in a variety of systems. In particular, one of the more recent implementations is the XGobi system which is linked to S. The interested reader will find more details in: Swayne, D. F., Cook, D., & Buja, A. (1991). "XGobi: Interactive Dynamic Graphical Displays in the X Window System Linked to S," *Proc. Statist. Graphics Sec. Am. Statist. Assoc.*, 1–8. An S function called "xgobi( )" is available through statlib. (See discussion of statlib at the end of this appendix.)

Principal Components Analysis (Sections 2.2.1 and 2.4)

In S/S-plus, "prcomp(*x*)"; [*Notes*: (a) "*x*" is the $n \times p$ matrix of data and the function returns quantities associated with the principal components of the covariance matrix. If the principal components of the correlation matrix are desired then the user will have to specify the standardized form of the data as the matrix, *x*. (b) If the user has, or wishes, to start with either a covariance matrix or a correlation matrix, or robust versions of these, then the function, "eigen", mentioned above would be the appropriate one to use and not "prcomp".]

In SAS, PROC PRINCOMP, incorporates user specification of "data type" and yields the principal components of either the covariance or the correlation matrix of the data.

Factor Analysis (Section 2.2.2)

To carry out a principal factor analysis in S/S-plus, the user will need to first compute the reduced correlation matrix, $\mathbf{R}^*$, or perhaps a robust version of it, and then use the function, "eigen", mentioned above. SAS provides PROC FACTOR which has options, including the principal factor method and the maximum likelihood method as well as facilities for rotation. As for specialized software, LISREL, a computer-aided system for fitting so-called linear structural equations models developed by Jöreskog and his collaborators is a rich source [see, for example, Jöreskog, K. G. & Sörbom, D. (1984). "LISREL VI Analysis of Linear Structural Relations by Maximum Likelihood, Instrumental Variables, and Least Squares Methods," User's Guide, Department of Statistics, University of Uppsala, Uppsala, Sweden].

Multidimensional Scaling (Section 2.3)

Specialized software seems to be the route. For nonmetric multidimensional scaling, the best currently available system appears to be one called KYST which is distributed by AT&T Bell Labs on request. For the three-way scaling method, INDSCAL, mentioned in Section 2.3, a program by that name is also available on request from AT&T Bell Labs. [See Kruskal, J. B. & Wish, M. (1978). "Multidimensional Scaling," Sage Publications, Beverly Hills, CA. (See, in particular, their pages 78–82).]

Canonical Correlation Analysis (Section 3.2)

For the classical case of two sets of variables, in S/S-plus the function is "cancor(x, y)". [*Note*: If the starting point is a correlation matrix, or a robust version of it, instead of the original data, then the user will have to compute the matrix, $\mathbf{R}_{12}$, described in Section 3.2, initially and then use the function, "svd($\mathbf{R}_{12}$)", to obtain the results needed. (See discussion of computations via the singular value decomposition in Section 3.2.)] In SAS, PROC CANCORR is the one to use with an option to specify the data type as either the original data or the correlation matrix.

For the extension to more than two sets of variables, as of now it seems that the appropriate reference to specialized programs is still the one mentioned in the first edition of this book [Chen, H. J. & Kettenring, J. R. (1972). "CANON: A Computer Program Package for the Multi-set Canonical Correlation Analysis," Bell Laboratories Technical Memorandum].

Discriminant Analysis (Section 4.2)

In S/S-plus, the function for carrying out a multi-group analysis is "discr(x, k)", where x is the $n \times p$ matrix of the data from all groups and k is either the number of groups if they are all of the same size or the vector of group sizes if they are of different sizes. In SAS, the function is PROC DISCRIM with a variety of options and formal tests of significance included.

For the new paradigm of classification trees, the reader is referred to: (a) Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). "Classification and Regression Trees," Wadsworth International Group, Belmont, CA; and (b) Clark, L. A. & Pregibon, D. (1992). "Tree-based Models," Chapter 9 of the book edited by J. M. Chambers & T. J. Hastie listed earlier in this appendix.

Cluster Analysis (Section 4.3)

In S/S-plus the function for performing hierarchical cluster analysis is "hclust( )". As one of the input arguments, the user is expected to provide either an $n \times n$ matrix of interobject distances or similarities. [*Note*: S/S-plus have a function called "dist(x)" which will compute either Euclidean distances or Manhattan distances between every pair of rows of the $n \times p$ matrix, x. While the $\mathbf{W}^*$-algorithm described in Section 4.3.1 is not yet available in the standard forms of S and S-plus, it is expected that S functions developed by Joann Harvey will be made available through statlib (see discussion of statlib below).] Another argument in hclust specifies the method of hierarchical clustering to be used. The output of hclust is the entire hierarchical tree structure. The function, "plclust( )", can be used to plot the tree. Also, the tree can be cut at different levels to produce mutually exclusive clusters by using the function, "cutree( )", with the user specifying either the number of clusters desired, or the height of the tree in terms of the strength (see definition of strength in Section 4.3.2a), for producing the clusters. S-plus has the function, "kmeans( )", which implements the non-hierarchical clustering algorithm of k-means described in Section 4.3.2b. [*Note*: This function expects the user to provide not only a choice of the number of clusters desired but also a starting point for the cluster centers.] As to the aids for interpreting the results of a cluster analysis discussed in Section 4.3.3, functions for many exist in private S libraries which will hopefully be made available more widely through statlib before long. For example, in a Supplement to S developed by Elaine Keramidas and Karen Bogucz at Bellcore, there are functions available for getting a cluster profile plot ("clprofile( )") and the plot of the B_k-statistic vs. k ("bkplot( )") described in Section 4.3.3.

SAS has a large variety of clustering algorithms to choose from including all the hierarchical methods discussed in Section 4.3.2a, k-means and even methods for overlapping clusters. PROC CLUSTER is for hierarchical clustering and will handle either the data matrix or inter-object distances matrix as input to it. PROC FASTCLUS with the data matrix as the input deals with a number of non-hierarchical clustering methods. PROC OVERCLUS is the one to use if one wishes to allow for overlapping clusters. PROC ACECLUS implements a version of the W*-algorithm for computing an estimate of the within-cluster covariance matrix. ACECLUS is a preprocessing step to either PROC CLUSTER or PROC FASTCLUS in that it provides a means for either sphericizing the data prior to PROC FASTCLUS or computing the inter-object distances as inputs to PROC CLUSTER.

The reader interested in ISODATA is referred to the specialized package mentioned in the first edition of this book. [See, Warner, J. L. (1968). "An Adaptation of ISODATA POINTS, an *I*terative *S*elf-*O*rganizing *D*ata *A*nalysis *T*echnique *A*," Bell Laboratories Technical Memorandum.]

There are a number of special packages of computer programs directed towards the vast array of clustering algorithms that are currently available. A good reference source for many of these is: Kaufman, L. & Rousseeuw, P. J. (1990). "Finding Groups in Data," Wiley, New York.

General Linear Model and Analysis of Variance (Sections 5.2.1 and 5.2.2)

There is an abundance of software for the classical least squares fits and formal tests of hypotheses associated with this topic. In S/S-plus the functions "lsfit( )" and "aov( )" are the key ones for univariate analyses of observations on each variable separately. In SAS, the corresponding procedures are PROC GLM and PROC ANOVA. PROC GLM is the one to use for fitting univariate general linear models, including multiple regression and unbalanced designs. PROC ANOVA is the efficient method for carrying out univariate analyses of variance for each variable. When it comes to carrying out multiresponse analyses, S/S-plus have the feature "maov( )" which collects together the fitted effects from the separate analyses together in vectors but there is no provision for obtaining a partitioning of the total sum-of-cross-products matrix into component matrices or carrying out the standard tests of significance or other so-called MANOVA procedures. SAS, however, includes a statement MANOVA in PROC GLM which returns the results of a MANOVA including all of the standard tests of significance.

Robust/Resistant Estimates (Section 5.2.3)

Both S/S-plus and SAS contain standard univariate measures of location and dispersion that are based on order statistics (e.g., median, inter-quartile range, etc.). When it comes to the currently preferred robust estimates, such as *m*-estimates, however, the choices are quite limited. This is especially so in dealing with the multivariate case. As of now, S/S-plus have an edge over SAS in this area but undoubtedly this will change in the future. The discussion here is confined to S/S-plus.

Location and Dispersion

To collect together univariate estimates into a vector estimate, S/S-plus have a handy function called "apply(*x*,2,'name of univariate function')", where '*x*' is the $n \times p$ matrix of data, the '2' specifies that the univariate function applies to the columns of *x*, and the name of the univariate function for the third argument can be 'mean' if the usual mean, $\bar{y}$, is desired, or 'median', if the vector of univariate medians is desired. In S-plus, one can use 'robloc' for the third argument of apply to obtain a vector of univariate *m*-estimates of location such as Huber's or Tukey's bisquare (see Section 5.2.3). The function, "var(*x*)", with the argument '*x*' specified as the $n \times p$ matrix of data will yield the usual non-robust $p \times p$ covariance matrix of the data.

For the full-fledged multivariate robust estimates of location and dispersion, such as the ellipsoidally trimmed estimator or the *m*-estimates proposed by Huber and Maronna described in Section 5.2.3, at the moment, S functions are available in private libraries that the author has access to but hopefully soon these will be made widely available through the statlib facility. [*Note*: S-plus has a function, "cov.mve( )", which calculates the minimum volume ellipsoid estimate of a covariance matrix mentioned in Section 5.2.3. As part of the returned information by this function, one can obtain a "center" which is a robust estimate of location.]

Correlation

For obtaining the robust estimator, r^*(SSD), of bivariate correlation described in Section 5.2.3, S/S-plus have the function "cor( )" with an argument called, 'trim', that can be used for specifying the proportion of the observations the user wishes to trim. In this usage the first two arguments of "cor( )" are two vectors containing the observations on the two variables of interest. To obtain, $\mathbf{R}^*$(SSD), the $p \times p$ matrix of bivariate correlations each of which is itself an r^*(SSD)-type of estimate, the same function "cor( )" with one argument 'x' specifying the $n \times p$ matrix of multivariate observations and a second one specifying a value of 'trim' will suffice. [*Note*: Recall the discussion in Section 5.2.3 about the need to shrink this estimate if one wants to guarantee positive definiteness of the estimate. At the moment, there is no S function for producing $\mathbf{R}^*_+$(SSD) from $\mathbf{R}^*$(SSD) using the scheme in Eq. (72a).]

Assessment of Normality

PROC UNIVARIATE in SAS contains a number of things including histograms, normal probability plots, and an array of formal tests of significance for normality. S/S-plus are well equipped to provide a wide range of probability plots. The function, "qqnorm( )", produces a normal probability plot. However, it is a relatively simple matter in S/S-plus to obtain probability plots such as those for the "radius-and-angles" representation for checking multivariate normality described in Section 5.4.2. The next item pertaining to probability plotting clarifies what is involved. [*Note*: For computing the squared radii, which are just the squared Mahalanobis distances of the multivariate observations from the mean vector using the inverse of the covariance matrix as the metric, S-plus has a handy function called, "mahalanobis()". This function requires the user to specify not only the data matrix, but a 'center' and a 'cov' matrix whose inverse is used as the metric. These specifications enable the user to obtain not only the usual squared Mahalanobis distances in a set of data but also robust versions of these by specifying robust estimates of location and dispersion in place of the mean vector and the usual covariance matrix.]

Probability Plotting (Section 6.2 and throughout the book)

For SAS, the book by M. Friendly referred to above under plotting of multivariate data is a source for descriptions of macros. The discussion here is, however, confined to S/S-plus and pertains to Q-Q plots as defined in Section 6.2. Basically, what enables the ease of obtaining Q-Q plots of a set of initial observations, or of things derived from them (such as radii, angles, projections onto a specific principal component, etc.), is the availability of a quantile function in S/S-plus that enables the computation of quantiles corresponding to given values of the cumulative proportion for a number of well-known distributions, including the uniform, chisquare, gamma and beta distributions. For generating the vector of appropriate cumulative fractions, $(i-a)/(n-2a+1)$ for $i = 1, \ldots, n$, there is the function "ppoints()" with an option for specifying a choice for 'a'. Next the function, "qxxx(ppoints())", where 'xxx' is to be specified by a conventional name or abbreviation for a distribution, produces the desired quantiles. The choices for 'xxx' include 'unif' for the uniform distribution, 'chisq' for the chi-squared distribution, 'gamma' for the gamma distribution, 'beta' for the beta distribution, and so on. The quantile function, "qxxx( )", also expects the user to specify any shape parameters that may be involved in order to calculate the quantiles of the particular member of the class of distributions. Thus, values of the degrees of freedom for a chi-squared distribution, or of the single shape parameter for the gamma, or of the two shape parameters for a beta distribution will need to be specified by the user. Finally, to obtain a *Q-Q* plot of a set of values, *z* (either of initial observations or of things derived from data), one would use the scatter plot function, "plot('qxxx( )',sort(z))".

As mentioned in the preceding paragraph, for *Q-Q* plotting against the quantiles of a distribution involving shape parameters, the user has to specify values of these parameters. Some of the techniques described in the earlier chapters of this book (e.g., the gamma, probability plotting methods discussed in Section 6.3), depend on estimating such parameters from the data at hand. At the moment, there are S functions available in private libraries (e.g., the functions, "egamma( )" and "ebeta( )", for maximum likelihood estimation of the parameters of the gamma and beta distributions, respectively, in a Supplement to S developed by Elaine Keramidas and Karen Bogucz of Bellcore) for calculating such estimates and hopefully these will be made available through statlib.

A *Q-Q* plot for comparing the distributions of two sets of data without specifying the common distribution, is obtained in S/S-plus by the function, "qqplot(*x*, *y*)", where '*x*' and '*y*' are the two sets of data. [*Note*: In this connection, the term data includes both raw observations and derived summaries.] This function is useful, for example, in obtaining the component probability plot and the standardized component probability plot described in Section 5.4.1 for assessing the similarity of the marginal distributions of a pair of variables. [*Note*: This distribution free *Q-Q* plot can also be used for comparing the distributions of two sets of data which are of unequal size. If '*x*' and '*y*' are not of the same size, qqplot(*x*, *y*) results in a plot of the ordered observations in the smaller set against the corresponding quantiles extracted from the larger set.]

statlib

This is an archive of S functions supplied by a large population of users and made available to the profession at large as a service by Michael Meyer at Carnegie Mellon University. Information on the contents of statlib at any given time can be obtained by sending electronic mail to the Internet email address, statlib@lib.stat.cmu.edu, with the two-line message,

```
send index
send index from S
```

as the body of the message. One can also obtain the sources for any of the S functions in statlib by using the file transfer program, "ftp". (See also Appendix D of the book by W. N. Venables & B. D. Ripley listed above.)

Methods for Statistical Data Analysis of Multivariate Observations, Second Edition
by R. Gnanadesikan

Author Index

Subject Index

WILEY SERIES IN PROBABILITY AND STATISTICS

Probability and Statistics

ANDERSON · An Introduction to Multivariate Statistical Analysis, *Second Edition*
*ANDERSON · The Statistical Analysis of Time Series
ARNOLD, BALAKRISHNAN, and NAGARAJA · A First Course in Order Statistics
BACCELLI, COHEN, OLSDER, and QUADRAT · Synchronization and Linearity: An Algebra for Discrete Event Systems
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ · Probability and Statistical Inference
BERNARDO and SMITH · Bayesian Statistical Concepts and Theory
BHATTACHARYYA and JOHNSON · Statistical Concepts and Methods
BILLINGSLEY · Convergence of Probability Measures
BILLINGSLEY · Probability and Measure, *Second Edition*
BOROVKOV · Asymptotic Methods in Queuing Theory
BRANDT, FRANKEN, and LISEK · Stationary Stochastic Models
CAINES · Linear Stochastic Systems
CAIROLI and DALANG · Sequential Stochastic Optimization
CHEN · Recursive Estimation and Control for Stochastic Systems
CONSTANTINE · Combinatorial Theory and Statistical Design
COOK and WEISBERG · An Introduction to Regression Graphics
COVER and THOMAS · Elements of Information Theory
CSÖRGŐ and HORVÁTH · Weighted Approximations in Probability Statistics
*DOOB · Stochastic Processes
DUDEWICZ and MISHRA · Modern Mathematical Statistics
DUPUIS · A Weak Convergence Approach to the Theory of Large Deviations
ETHIER and KURTZ · Markov Processes: Characterization and Convergence
FELLER · An Introduction to Probability Theory and Its Applications, Volume 1, *Third Edition,* Revised; Volume II, *Second Edition*
FREEMAN and SMITH · Aspects of Uncertainty: A Tribute to D. V. Lindley
FULLER · Introduction to Statistical Time Series, *Second Edition*
FULLER · Measurement Error Models
GHOSH · Sequential Estimation
GIFI · Nonlinear Multivariate Analysis
GUTTORP · Statistical Inference for Branching Processes
HALD · A History of Probability and Statistics and Their Applications before 1750
HALL · Introduction to the Theory of Coverage Processes
HANNAN and DEISTLER · The Statistical Theory of Linear Systems
HEDAYAT and SINHA · Design and Inference in Finite Population Sampling
HOEL · Introduction to Mathematical Statistics, *Fifth Edition*
HUBER · Robust Statistics
IMAN and CONOVER · A Modern Approach to Statistics
JUREK and MASON · Operator-Limit Distributions in Probability Theory
KASS and VOS · Geometrical Foundations of Asymptotic Inference: Curved Exponential Families and Beyond
KAUFMAN and ROUSSEEUW · Finding Groups in Data: An Introduction to Cluster Analysis

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Probability and Statistics (Continued)

KOTZ · Leading Personalities in Statistical Sciences from the Seventeenth Century to the Present
LAMPERTI · Probability: A Survey of the Mathematical Theory, *Second Edition*
LARSON · Introduction to Probability Theory and Statistical Inference, *Third Edition*
LESSLER and KALSBEEK · Nonsampling Error in Surveys
LINDVALL · Lectures on the Coupling Method
MANTON, WOODBURY, and TOLLEY · Statistical Applications Using Fuzzy Sets
MARDIA · The Art of Statistical Science: A Tribute to G. S. Watson
MORGENTHALER and TUKEY · Configural Polysampling: A Route to Practical Robustness
MUIRHEAD · Aspects of Multivariate Statistical Theory
OLIVER and SMITH · Influence Diagrams, Belief Nets and Decision Analysis
*PARZEN · Modern Probability Theory and Its Applications
PRESS · Bayesian Statistics: Principles, Models, and Applications
PUKELSHEIM · Optimal Experimental Design
PURI and SEN · Nonparametric Methods in General Linear Models
PURI, VILAPLANA, and WERTZ · New Perspectives in Theoretical and Applied Statistics
RAO · Asymptotic Theory of Statistical Inference
RAO · Linear Statistical Inference and Its Applications, *Second Edition*
*RAO and SHANBHAG · Choquet-Deny Type Functional Equations with Applications to Stochastic Models
RENCHER · Methods of Multivariate Analysis
ROBERTSON, WRIGHT, and DYKSTRA · Order Restricted Statistical Inference
ROGERS and WILLIAMS · Diffusions, Markov Processes, and Martingales, Volume I: Foundations, *Second Edition;* Volume II: Îto Calculus
ROHATGI · An Introduction to Probability Theory and Mathematical Statistics
ROSS · Stochastic Processes
RUBINSTEIN · Simulation and the Monte Carlo Method
RUBINSTEIN and SHAPIRO · Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method
RUZSA and SZEKELY · Algebraic Probability Theory
SCHEFFE · The Analysis of Variance
SEBER · Linear Regression Analysis
SEBER · Multivariate Observations
SEBER and WILD · Nonlinear Regression
SERFLING · Approximation Theorems of Mathematical Statistics
SHORACK and WELLNER · Empirical Processes with Applications to Statistics
SMALL and McLEISH · Hilbert Space Methods in Probability and Statistical Inference
STAPLETON · Linear Statistical Models
STAUDTE and SHEATHER · Robust Estimation and Testing
STOYANOV · Counterexamples in Probability
STYAN · The Collected Papers of T. W. Anderson: 1943–1985
TANAKA · Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON and SEBER · Adaptive Sampling
WELSH · Aspects of Statistical Inference
WHITTAKER · Graphical Models in Applied Multivariate Statistics
YANG · The Construction Theory of Denumerable Markov Processes

Applied Probability and Statistics

ABRAHAM and LEDOLTER · Statistical Methods for Forecasting
AGRESTI · Analysis of Ordinal Categorical Data

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Applied Probability and Statistics (Continued)

AGRESTI · Categorical Data Analysis
AGRESTI · An Introduction to Categorical Data Analysis
ANDERSON and LOYNES · The Teaching of Practical Statistics
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG · Statistical Methods for Comparative Studies
ARMITAGE and DAVID (editors) · Advances in Biometry
*ARTHANARI and DODGE · Mathematical Programming in Statistics
ASMUSSEN · Applied Probability and Queues
*BAILEY · The Elements of Stochastic Processes with Applications to the Natural Sciences
BARNETT and LEWIS · Outliers in Statistical Data, *Second Edition*
BARTHOLOMEW, FORBES, and McLEAN · Statistical Techniques for Manpower Planning, *Second Edition*
BATES and WATTS · Nonlinear Regression Analysis and Its Applications
BECHHOFER, SANTNER, and GOLDSMAN · Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons
BELSLEY · Conditioning Diagnostics: Collinearity and Weak Data in Regression
BELSLEY, KUH, and WELSCH · Regression Diagnostics: Identifying Influential Data and Sources of Collinearity
BERRY · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
BERRY, CHALONER, and GEWEKE · Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner
BHAT · Elements of Applied Stochastic Processes, *Second Edition*
BHATTACHARYA and WAYMIRE · Stochastic Processes with Applications
BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN · Measurement Errors in Surveys
BIRKES and DODGE · Alternative Methods of Regression
BLOOMFIELD · Fourier Analysis of Time Series: An Introduction
BOLLEN · Structural Equations with Latent Variables
BOULEAU · Numerical Methods for Stochastic Processes
BOX · R. A. Fisher, the Life of a Scientist
BOX and DRAPER · Empirical Model-Building and Response Surfaces
BOX and DRAPER · Evolutionary Operation: A Statistical Method for Process Improvement
BOX, HUNTER, and HUNTER · Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building
BROWN and HOLLANDER · Statistics: A Biomedical Introduction
BUCKLEW · Large Deviation Techniques in Decision, Simulation, and Estimation
BUNKE and BUNKE · Nonlinear Regression, Functional Relations and Robust Methods: Statistical Methods of Model Building
CHATTERJEE and HADI · Sensitivity Analysis in Linear Regression
CHATTERJEE and PRICE · Regression Analysis by Example, *Second Edition*
CLARKE and DISNEY · Probability and Random Processes: A First Course with Applications, *Second Edition*
COCHRAN · Sampling Techniques, *Third Edition*
*COCHRAN and COX · Experimental Designs, *Second Edition*
CONOVER · Practical Nonparametric Statistics, *Second Edition*
CONOVER and IMAN · Introduction to Modern Business Statistics
CORNELL · Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Second Edition*
COX · A Handbook of Introductory Statistical Methods
*COX · Planning of Experiments

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Applied Probability and Statistics (Continued)

COX, BINDER, CHINNAPPA, CHRISTIANSON, COLLEDGE, and KOTT · Business Survey Methods

CRESSIE · Statistics for Spatial Data, *Revised Edition*

DANIEL · Applications of Statistics to Industrial Experimentation

DANIEL · Biostatistics: A Foundation for Analysis in the Health Sciences, *Sixth Edition*

DAVID · Order Statistics, *Second Edition*

*DEGROOT, FIENBERG, and KADANE · Statistics and the Law

*DEMING · Sample Design in Business Research

DILLON and GOLDSTEIN · Multivariate Analysis: Methods and Applications

DODGE and ROMIG · Sampling Inspection Tables, *Second Edition*

DOWDY and WEARDEN · Statistics for Research, *Second Edition*

DRAPER and SMITH · Applied Regression Analysis, *Second Edition*

DUNN · Basic Statistics: A Primer for the Biomedical Sciences, *Second Edition*

DUNN and CLARK · Applied Statistics: Analysis of Variance and Regression, *Second Edition*

ELANDT-JOHNSON and JOHNSON · Survival Models and Data Analysis

EVANS, PEACOCK, and HASTINGS · Statistical Distributions, *Second Edition*

FISHER and VAN BELLE · Biostatistics: A Methodology for the Health Sciences

FLEISS · The Design and Analysis of Clinical Experiments

FLEISS · Statistical Methods for Rates and Proportions, *Second Edition*

FLEMING and HARRINGTON · Counting Processes and Survival Analysis

FLURY · Common Principal Components and Related Multivariate Models

GALLANT · Nonlinear Statistical Models

GLASSERMAN and YAO · Monotone Structure in Discrete-Event Systems

GNANADESIKAN · Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*

GREENWOOD and NIKULIN · A Guide to Chi-Squared Testing

GROSS and HARRIS · Fundamentals of Queueing Theory, *Second Edition*

GROVES · Survey Errors and Survey Costs

GROVES, BIEMER, LYBERG, MASSEY, NICHOLLS, and WAKSBERG · Telephone Survey Methodology

HAHN and MEEKER · Statistical Intervals: A Guide for Practitioners

HAND · Discrimination and Classification

*HANSEN, HURWITZ, and MADOW · Sample Survey Methods and Theory, Volume 1: Methods and Applications

*HANSEN, HURWITZ, and MADOW · Sample Survey Methods and Theory, Volume II: Theory

HEIBERGER · Computation for the Analysis of Designed Experiments

HELLER · MACSYMA for Statisticians

HINKELMAN and KEMPTHORNE: · Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design

HOAGLIN, MOSTELLER, and TUKEY · Exploratory Approach to Analysis of Variance

HOAGLIN, MOSTELLER, and TUKEY · Exploring Data Tables, Trends and Shapes

HOAGLIN, MOSTELLER, and TUKEY · Understanding Robust and Exploratory Data Analysis

HOCHBERG and TAMHANE · Multiple Comparison Procedures

HOCKING · Methods and Applications of Linear Models: Regression and the Analysis of Variables

HOEL · Elementary Statistics, *Fifth Edition*

HOGG and KLUGMAN · Loss Distributions

HOLLANDER and WOLFE · Nonparametric Statistical Methods

HOSMER and LEMESHOW · Applied Logistic Regression

HØYLAND and RAUSAND · System Reliability Theory: Models and Statistical Methods

HUBERTY · Applied Discriminant Analysis

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Applied Probability and Statistics (Continued)

OKABE, BOOTS, and SUGIHARA · Spatial Tesselations: Concepts and Applications of Voronoi Diagrams
OSBORNE · Finite Algorithms in Optimization and Data Analysis
PANKRATZ · Forecasting with Dynamic Regression Models
PANKRATZ · Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
PORT · Theoretical Probability for Applications
PUTERMAN · Markov Decision Processes: Discrete Stochastic Dynamic Programming
RACHEV · Probability Metrics and the Stability of Stochastic Models
RÉNYI · A Diary on Information Theory
RIPLEY · Spatial Statistics
RIPLEY · Stochastic Simulation
ROSS · Introduction to Probability and Statistics for Engineers and Scientists
ROUSSEEUW and LEROY · Robust Regression and Outlier Detection
RUBIN · Multiple Imputation for Nonresponse in Surveys
RYAN · Modern Regression Methods
RYAN · Statistical Methods for Quality Improvement
SCHOTT · Matrix Analysis for Statistics
SCHUSS · Theory and Applications of Stochastic Differential Equations
SCOTT · Multivariate Density Estimation: Theory, Practice, and Visualization
SEARLE · Linear Models
SEARLE · Linear Models for Unbalanced Data
SEARLE · Matrix Algebra Useful for Statistics
SEARLE, CASELLA, and McCULLOCH · Variance Components
SKINNER, HOLT, and SMITH · Analysis of Complex Surveys
STOYAN, KENDALL, and MECKE · Stochastic Geometry and Its Applications, *Second Edition*
STOYAN and STOYAN · Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
THOMPSON · Empirical Model Building
THOMPSON · Sampling
TIERNEY · LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TIJMS · Stochastic Modeling and Analysis: A Computational Approach
TITTERINGTON, SMITH, and MAKOV · Statistical Analysis of Finite Mixture Distributions
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume I: Point Pattern and Quantitative Data
UPTON and FINGLETON · Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
VAN RIJCKEVORSEL and DE LEEUW · Component and Correspondence Analysis
WEISBERG · Applied Linear Regression, *Second Edition*
WESTFALL and YOUNG · Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment
WHITTLE · Optimization Over Time: Dynamic Programming and Stochastic Control, Volume I and Volume II
WHITTLE · Systems in Stochastic Equilibrium
WONNACOTT and WONNACOTT · Econometrics, *Second Edition*
WONNACOTT and WONNACOTT · Introductory Statistics, *Fifth Edition*
WONNACOTT and WONNACOTT · Introductory Statistics for Business and Economics, *Fourth Edition*
WOODING · Planning Pharmaceutical Clinical Trials: Basic Statistical Principles
WOOLSON · Statistical Methods for the Analysis of Biomedical Data
*ZELLNER · An Introduction to Bayesian Inference in Econometrics

*Now available in a lower priced paperback edition in the Wiley Classics Library.

Tracts on Probability and Statistics

BILLINGSLEY · Convergence of Probability Measures
KELLY · Reversability and Stochastic Networks
TOUTENBURG · Prior Information in Linear Models